2024

# An Exploration of Misconceptions in Introductory Physics

Christopher Mattthew Wheatley
*West Virginia University*

# An Exploration of Misconceptions in Introductory Physics

Christopher M. Wheatley

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Physics

John Stewart, Ph.D., Chair
Marjorie Darrah, Ph.D.
Paul Miller, Ph.D.
Joonhee Lee, Ph.D.

Department of Physics and Astronomy

Morgantown, West Virginia
2024

Keywords: physics, education, conceptual inventories, network,

# ABSTRACT

## An Exploration of Misconceptions in Introductory Physics

**Christopher M. Wheatley**

The study of student misconceptions about physics concepts has long been an important area of inquiry in physics education research (PER). The research discussed in this dissertation builds upon the developments in PER by exploring the prevalence of consistently held undergraduate student misconceptions in introductory calculus-based physics. This thesis explores the nature of student misconceptions, mistakes, and naive answering patterns in both introductory undergraduate Newtonian mechanics and electromagnetism by applying a network analytic technique called module analysis to student responses to different concept inventories from institutions of various levels of incoming physics preparation. Each study applying these methods also demonstrates how they can also be used to inform future inventory development. Network analysis was also used to study the growth and evolution of the First2 Network, a project with the goal of doubling the retention rate of STEM students in West Virginia, with a particular emphasis on rural and first-generation students. The final part of this thesis compares students' performance and attendance in an introductory electricity and magnetism course before and after the COVID-19 pandemic.

# ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor and mentor, Dr. John Stewart. As a research advisor, you have shown me tremendous patience as I learned how to perform, write, and present academic research. Both when I was a laboratory TA and a lecturer for your class, you taught me innumerable lessons leading by example as an exemplary educator and leader. As a mentor, you indulged my countless trivial questions and showed me endless grace and kindness during the many major life events that I encountered in graduate school. Without your dedication to my success as a student, researcher, and educator, the work within this dissertation would not have been possible. I count myself beyond blessed to have worked with you.

I would also like to thank the rest of my dissertation committee: Dr. Marjorie Darrah, Dr. Joonhee Lee, and Dr. Paul Miller. Each of you have invested additional time and effort to improve this dissertation. I appreciate the thoughtful questions you have posed about my own research and during my oral qualification process. A special thanks to Marjorie for your mentorship and support during our collaborations and presentations.

In addition, thank you to my high school physics teacher Greg Reger for sparking a passion for physics in me and to Dr. Daniel Cross for fueling and sustaining that passion as an undergraduate. I have had countless educators over my life who have inspired me with their kindness, their dedication to students, and their love for their subject that I could not possibly name them all. But to name a few, Dr. Eric Miller, Dr. Robert Frazier, Dr. Esther Meek, Dr. Phil Holladay, Dr. John Stahl, Dr. Frederick Neikirk, and Dr. Aldo Romero; each of you have fundamentally shaped my understanding of the universe and my place in it, and I am eternally grateful for your service as educators and mentors.

I am sincerely grateful to my physics education research group, Dr. Dona Hewagallage, Dr. John Hansen, Elaine Christman, John Pace, Amanda Nemeth, and Danielle Maldonado, for their camaraderie, collaboration, and guidance throughout my time in graduate school. As classmates, research collaborators, and friends, you were each a pleasure to get to know and learn from over the last five years. Attending conferences with you all of you helped calm the nerves of presentation and made each experience a lasting memory. To my $John^2$ office-mates, our many conversations from physics to politics to video-games and Garfield really broke the daily monotony of research, so I appreciate your unique contribution to my sustained sanity throughout graduate school.

I owe heartfelt thanks to my family for their boundless love and support, without each of you I would not be where I am today. To my mother and father, Elisabeth and John Wheatley, thank you both for instilling in me a passion for learning and a drive to reach my goals. Thank you to my siblings, Nathan Wheatley and Sarah Yates, for being some of my closest friends and confidants. Thank you all for believing in me and encouraging me in all my endeavors.

I would also like to thank my friends for being a valuable outlet from studying and research. In particular, thank you to William Eshbaugh for the hours we spent studying together for classes and written qualifiers and our many lunches and long walks discussing our research and musing over physics we didn't understand. Your friendship greatly eased the stresses and burdens that come with graduate student life. Also, a special thanks is owed to my college roommate and childhood best-friend Kevin Donaldson for making my transition into graduate school much easier my first year by housing me every weekend while visiting my fiancé. Your friendship has been an invaluable part of my life and crucial to my success as a graduate student.

Finally, I would like to thank my wife Julianna. You have commuted over two hours a day for years so that I could live in Morgantown to finish my graduate work. You have endured countless ramblings about research projects, coding, education, teaching, and physics. Since high school you have motivated and encouraged me to go to college in the first place, helping me with every step of the way, from scholarships to undergraduate applications and interviews to summer internships and graduate school applications, I genuinely would not be anywhere without you. I owe any success I might find in my career and my life to your faithfulness, your patience, your love, and your friendship.

*For Julianna,*
*my raison d'être.*

# Contents

# List of Tables

# List of Figures

# Chapter 1

## Introduction to Physics Education Research

In their 2014 review of physics education research (PER), Docktor and Mestre [1] separate physics education research into six broad categories; conceptual understanding, problem solving, curriculum and instruction, assessment, cognitive psychology, and attitudes and beliefs about teaching and learning. The research discussed in this dissertation would primarily fit into conceptual understanding, as such, this chapter provides a brief introduction to discipline-based education research (DBER) in physics focused on the study of conceptual understanding and how those studies led to the development and implementation of research-based instructional strategies (RBIS).

## 1.1 Conceptual Understanding of Physics

The study of the existence and the cause of common difficulties in student conceptual understanding in physics (e.g. that heavier objects do not fall faster) in the late 1970s and early 1980s could be considered the beginning of modern physics education research (PER) [2–5]. Difficulties in student conceptual understanding have many names in PER; alternative conceptions, naive conceptions, preconceptions, and misconceptions are the most commonly used terms, each with slightly different definitions. The modern accepted definition of misconception generally includes at least some of the following four attributes; they are deeply rooted cognitive structures that are stable in time, they affect how students interpret scientific explanations, they differ from expert explanations of concepts, and they must be eliminated to master the subject [6].

Early studies identified and compiled the most common misconceptions in introductory physics [2, 3]. These studies also began to introduce new instructional strategies to better

target these misconceptions and move students towards more rigorous scientific reasoning. One of the most popular ways to measure a student's change in conceptual understanding from the beginning to the end of a physics course was through conceptual inventories.

### 1.1.1 Concept Inventories

In 1985, Halloun and Hestenes [7] developed an instrument called the mechanics diagnostic test (MD), to determine students' initial knowledge of mechanics when entering an introductory mechanics course. This test measured student conceptual understanding of basic dynamics and kinematics. Halloun and Hestenes found that students' MD scores predicted their final course grades and that students gained very little in their conceptual understanding by the end of a traditional lecture course. Immediately after this study, Halloun and Hestenes conducted another study where they interviewed students as they were taking the mechanics diagnostic test. They separated student comments into the categories "principles of motion" and "influences on motion" [8].

These two studies, along with the misconception research by Clement [2] and McDermott [3], identified the need for the development of research-based assessments in physics education. Research-based assessments were designed both to study misconceptions about the physical world that students held when entering introductory mechanics courses and to study the misconceptions retained by the end of the course.

In 1992, Hestenes *et al.* developed the first widely distributed physics concept inventory, the Force Concept Inventory (FCI) [9], by heavily revising the mechanics diagnostic test using data from student interviews. The FCI uses 30 items (questions) to test students' conceptual understanding of kinematics in one and two dimensions and Newton's three laws.

Each item in the FCI includes four incorrect responses and one correct response. Many incorrect responses are intended to be attractive responses (distractors) that correspond to common misconceptions. The FCI remains the most popular, transformative, and well-studied physics assessment today, 32 years after its inception. The version of the FCI used today, after its revision in 1997 [10], can be found at PhysPort [11].

The introduction of the FCI into PER led to the development of many new research-based assessments in physics. Many well-vetted instruments have been made to test student conceptual understanding of Newtonian mechanics [12], electricity and magnetism [13, 14], quantum mechanics [15, 16], thermodynamics [17, 18], waves [19], and astronomy [20–22]. There have also been physics instruments developed to test students' mathematical reasoning ability [23, 24], beliefs/attitude about science [25, 26], and scientific reasoning [27]. Other than the FCI, some of the most popular physics conceptual inventories include the Force and Motion Conceptual Evaluation (FMCE) [12], the Conceptual Survey of Electricity and Magnetism (CSEM) [13], and the Brief Electricity and Magnetism Assessment (BEMA) [14].

## 1.2 Reformed Instruction

One of the most influential uses of a physics concept inventory was in Hake's [28] analysis of FCI scores to compare modes of instruction between 62 different courses from various institutions. Hake introduced the normalized gain, Equation 1.1, to compare the effectiveness of a given course at encouraging conceptual understanding.

$$\langle g \rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{100 - \langle S_i \rangle} \tag{1.1}$$

The normalized gain $\langle g \rangle$ utilizes the average post test score over the class $\langle S_f \rangle$ and the average pretest score over the class $\langle S_i \rangle$, both scored out of 100, to calculate a score that represents the actual improvement students made in a course over the maximum possible improvement that could have been made in the course given the class average pretest scores. Although the normalized gain was introduced using FCI responses, it can be calculated for any instrument. Like the effect size in many significance tests [29], a low normalized gain is described as $\langle g \rangle < 0.3$, a medium gain as $0.3 \leq \langle g \rangle < 0.7$, and a large gain as $\langle g \rangle \geq 0.7$.

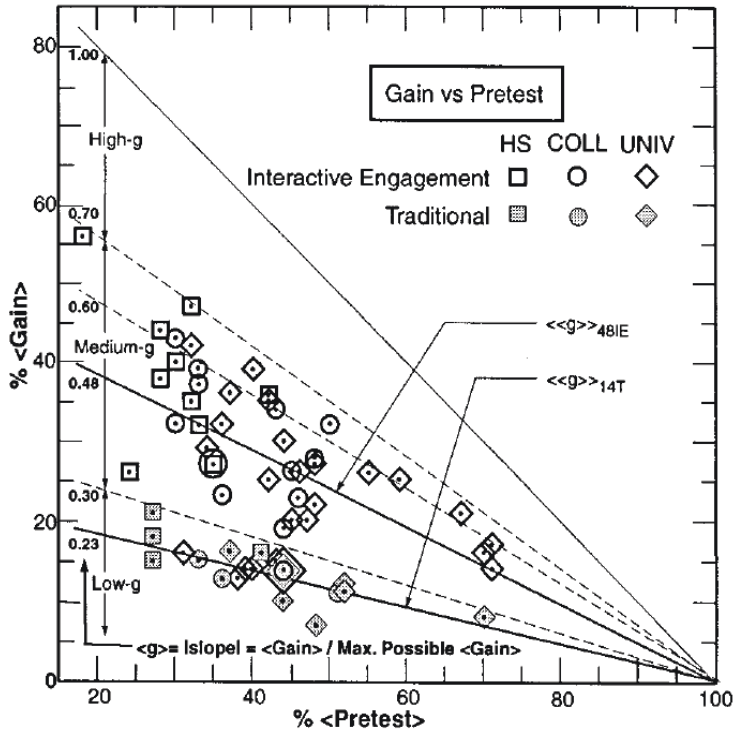


Figure 1.1: Gain vs. pretest scores for 62 courses comparing traditional instruction with interactive engagement. [28]

Figure 1.1 compares traditional lecture instruction with interactive engagement (IE) instruction, sometimes called active learning. Data from 62 courses were plotted, with 14 of them utilizing traditional lecture instruction and 48 utilizing interactive engagement. The y-axis displays the average gain over a class $\langle Gain \rangle$, which is just the numerator of the

normalized gain shown in Equation 1.1, while the x-axis displays average pretest scores over a class. Any slope drawn on this figure represents a normalized gain. Dashed lines are drawn to show the thresholds for small, medium, and large normalized gains and a solid line is drawn to show the upper limit. Two darker solid lines are drawn to compare the average normalized gains from institutions taught with interactive engagement to institutions taught in a traditional lecture environment. Hake's work provided clear evidence that interactive engagement instruction had measurable benefits over traditional lecture instruction, and it motivated many instructors to adopt various forms of reformed instruction, often referred to as research-based instructional strategies (RBIS).

### 1.2.1 Research-Based Instructional Strategies

In their 2012 resource letter on active learning instruction in physics, Meltzer and Thornton [30] classified "research-based active learning instruction" as any teaching method that is derived from research in physics education, that integrates interactive activities requiring students to express their reasoning about physics, and that has evidence of improved student learning in real classroom environments. Research-based active learning instruction, usually shortened to active learning, will be used synonymously with interactive engagement instruction, research-based instructional strategies, and reformed instruction for the remainder of this document. Modern active learning instruction can be very broadly placed into three categories; instruction that incorporates active learning into a traditional lecture environment, instruction that incorporates active learning strategies into the laboratory, and instruction that either incorporates active learning into recitations/discussion sections or transforms the structure of the learning environment to be better fit for student engage-

ment.

### Reformed Lecture

The most popular type of active learning instruction in physics has been the reform of the traditional lecture. Traditional lecture reforms retain the classroom structure of a traditional lecture, often with an instructor lecturing and students taking notes, but it transitions the passive environment to an active one [31]. Actively involving students in the lecture usually includes some type of small group work or polling system that can be used both to gauge students' level of understanding in real time and to increase the level of focus and participation in the classroom. The most popular way of increasing student engagement in physics lectures has been through the use of clickers [32, 33]. Clickers are small, portable devices that allow students to select a multiple-choice response posed to the entire classroom. Many courses that utilize clickers have recently transitioned to phone applications that serve the same purpose rather than providing or requiring separate clicker devices.

In 1997, Eric Mazur published a user manual to Peer Instruction, an active learning method that heavily involved the use of clickers [10]. Peer Instruction utilizes cycles of about ten minutes lecturing followed by a clicker question. For each of these clicker questions, students are given time to think and respond to the question individually, then they are prompted to discuss their responses with neighboring students, explaining their reasoning and coming to some common ground about the principles applied. Students then report their revised responses and the instructor elicits explanations from students, reasoning with them and pointing them in the direction of the correct response until a general consensus of understanding is met across the class. Significantly higher normalized gains have been mea-

sured from courses taught with Peer Instruction relative to a traditional lecture environment [34].

**Reformed Laboratory**

Laboratory experiments have long been conducted outside of lectures for introductory physics courses. In a 1991 review of RBIs in physics [35], Alan Van Heuvelen emphasized that traditional labs were having no measurable effects on student learning outcomes. Traditional labs focus on knowledge acquisition through formulaic experiments. Reformed labs broadly reorient the focus to an exploration of the physical processes in an experiment, rather than providing students with the requisite physics knowledge then having them conduct the experiment themselves. One particularly well-established reformed lab curriculum is Investigative Science Learning Environment (ISLE) [36]. The purpose of ISLE is to treat students as novice scientists; prompting them to observe, explain, design, and test their own experiments instead of going through a structured lab manual. For ISLE-based labs, students design experiments before the concept is taught in lecture. Guiding questions are included to aid the process of experimental design. Further experiments are then designed to test the explanation developed from earlier experiments, or from models provided by an instructor. These experiments are performed in small groups and findings are discussed with the whole class toward the end of the lab.

Labs taught from an ISLE framework have been shown to improve students' facilities with data collection and analysis, experimental design, and science communication [37, 38]. Some educators argue for utilizing labs as a means to increase scientist-like skills and reasoning abilities, rather than to improve content knowledge, after multiple studies demonstrated

that many labs provide no added value for learning course content [39, 40]. Holmes *et al.* [40] evaluated nine labs from three different institutions with varying instructional techniques. For each of these courses, students could opt out of taking the lab portion of the class without penalty to their grade. The researchers compared performance between those who opted in to the labs and those who did not and they found no measurable difference between the two samples. This study implies that non-laboratory environments may be better suited for instructional interventions if the goal is explicitly improving course outcomes.

### Reformed Environment

One particularly successful type of active learning instruction in physics has been modifying the environment to allow for more interactive engagement. Designing physics classrooms in a way that fosters teamwork and small-group cooperation allows for integrating different aspects of lectures, labs, and recitations into a single classroom setting [1]. There are many curricula designed around implementing small-group cooperative work, reducing in-class lecture time, and increasing the level of faculty-student interaction within small classrooms, but far fewer for large introductory classes. One curriculum designed specifically for large classroom settings is the student-centered active learning environment for undergraduate programs (SCALE-UP). Developed by Beichner and collaborators at North Carolina State University in 2007, SCALE-UP is an integrated learning environment designed for large-enrollment physics classes with up to 100 students [41]. A typical SCALE-UP course includes very little time for lecture, relegating information transfer to assigned readings outside of class. This leaves class time for cooperative group problem solving, experiments, and answering questions. Students work in groups of three, each with access to a laptop, a

whiteboard, and any other material needed for assignments or experiments in a given class period. SCALE-UP, and other curricula that combine labs into the lecture by transforming the educational environment, allows for more classroom time collaboratively constructing knowledge of the material, rather than taking notes. Studies have shown improved exam scores, improved concept inventory gains, more positive attitudes toward the class, and reduced course attrition in classes that use SCALE-UP relative to traditional lecture settings [41, 42].

### 1.2.2 Adoption

In 2012, Henderson *et al.* [43] showed that out of 722 surveyed physics faculty across the United States, less than 50% of them were currently implementing some research-based instructional strategy. While the vast majority of surveyed individuals knew about some active learning method, over a third of those who had ever attempted to implement an RBIS discontinued it. Henderson *et al.* found that the knowledge or use of RBIS was correlated with factors such as reading teaching-related journals and attending talks and workshops about teaching. However, they also found that during these talks and workshops, the success and ease of implementation of these teaching methods were often exaggerated and their difficulties not fully explained. This led to a misalignment of expectations with the reality of implementing these new strategies for many faculty members. Another commonly quoted reason for abandonment was a lack of support after initial implementation. Many RBIS have nuances or complexities that faculty members have no ability to consult experts about. Another difficulty with implementing RBIS at a university level is the cost [44]. Nearly any method that increases interactive engagement will come at some cost to the

university, and reformed environment methods can come at a substantial cost.

### 1.2.3 Conclusion

Since Hake's initial study demonstrated the efficacy of reformed instruction, many new research based instructional strategies have been developed, validated, and analyzed in the classroom. The effects of reformed instruction in different STEM (science, technology, engineering, and mathematics) disciplines were compiled by Freeman *et al.* in 2014 [45]. They analyzed 225 studies across Biology, Chemistry, Computer Science, Engineering, Geology, Mathematics, Physics, and Psychology and found that reformed instruction had statistically significant increases in exam scores and concept inventory scores and decreases in failure rates across the board. The effect size varied significantly between disciplines, but most disciplines had at least a medium effect size when comparing student performance in reformed instruction courses with traditional lectures.

The research discussed in this dissertation builds upon the developments in PER described in this chapter by exploring the prevalence of consistently held student misconceptions in introductory physics. Chapters 3-7 explore the nature of student misconceptions, mistakes, and naive answering patterns in both introductory undergraduate Newtonian mechanics and electromagnetism by applying a network analytic technique called module analysis to student responses from multiple concept inventories. The methodology for this exploration is refined in each successive study. These methods can also be used to inform future inventory development. The remainder of the thesis covers other projects like studying the growth and evolution of a West Virginia STEM education network with network analysis (Chapter 8) and comparing students' performance in an introductory electromagntism course

before and after the COVID-19 pandemic (Chapter 9).

# Chapter 2

## Statistical Methods

Statistics provide a framework to understand the underlying properties of datasets. Statistics are necessary to infer information about a population from a randomly selected sample within that population. A population describes the entirety of a group of interest, while a sample describes a subset of that population from which data can be collected. Samples are often drawn from a population in such a way (large enough sample size and random selection) that statistics calculated from the sample serve as an acceptable approximation of the population statistic.

Statistics will be used to characterize and analyze samples of student data throughout this work. This chapter summarizes some of the most common statistical methods used throughout the document. Additional statistical methods will be introduced as needed.

## 2.1 Descriptive Statistics

Descriptive statistics are a set of techniques used to summarize general properties of datasets. These techniques generally include measures of central tendency and variability.

### 2.1.1 Central Tendency

Measures of central tendency describe a distribution of data by its most "average" value. The three most frequently applied measures of central tendency are median, mean, and mode. The mode measures the observation that appears the most frequently within the distribution; describing the "average" as the value most likely to be chosen by random sampling of the dataset. The median measures the middle point of an ordered distribution; describing the "average" as the value with an equal number of values greater than it and an equal number less than it. The mean, given by Equations 2.1 and 2.2, is computed by

summing each data point and dividing by the number of data points summed; describing the "average" as the value arithmetically closest to every value in the distribution.

Depending on the distribution of the data, these measures can vary substantially. For many distributions, each statistic provides unique descriptive information. The mean is particularly susceptible to outliers or distributions that are substantially non-symmetric [46]. In which case, it may be more appropriate to report the median instead. In this document, however, the mean is frequently applied to samples of student data. The sample mean, shown below as Equation 2.1, sums over all data in the sample. The population mean is also provided in Equation 2.2,

$$\bar{x} = \frac{\sum_i x_i}{n} \tag{2.1}$$

$$\mu = \frac{\sum_i x_i}{N} \tag{2.2}$$

where $x_i$ is a value for individual $i$, n is the size of the sample, and N is the size of the population. The sample mean is almost always reported because collecting data from every member of the population is often impossible.

### 2.1.2 Variability

The most central value provides incomplete information about a distribution without some description of the width of the distribution. The variance of a sample, Equation 2.3, provides a description of the spread of the distribution.

$$s^2 = \frac{\sum_i (x_i - \bar{x}^2)^2}{n-1} \tag{2.3}$$

The square root of this value, $s$, is called the standard deviation and is often reported alongside the mean to further describe the distribution. For normally distributed data, if $s$ is small, values within the distribution tend to be close to the mean and the distribution is narrow. If $s$ is large, some values are far from the mean and the distribution is wide.

## 2.2 Inferential Statistics

Inferential statistics describes a set of methods used to infer information about populations based on samples taken from those populations.

### 2.2.1 Hypothesis testing

A hypothesis test is a method of statistically deciding whether the studied data supports a hypothesis. Hypothesis testing involves presenting two competing hypotheses; the null hypothesis, which asserts that the effect in question does not exist, and the alternate hypothesis, which asserts that it does exist. Significance tests are then applied to the data to determine if an apparent effect would better be attributed to random chance; the null hypothesis, or to a real relationship between two variables; the alternate hypothesis. If every observation is independent and equally likely to occur, then any measured effect would be due to random chance. The first step in significance testing is to calculate the test statistic. Test statistics compare the distribution of data with the distribution predicted under the null hypothesis, which differs by the significance test being used. The result of a significance

test provides the likelihood that the calculated test statistic exists under the null hypothesis, so if that chance is sufficiently small, then the null hypothesis can be rejected and the level of confidence in the alternate hypothesis can be determined.

This manuscript makes use of Welch's $t$-test [47] for hypothesis testing. The independent two-sample $t$-test is a significance test used to evaluate the null hypothesis that there is no difference between the mean of two independendt groups. The $t$-test is a parametric statistic, a statistic that requires certain distributional conditions should be satisfied for its use. Non-parametric hypothesis tests, tests that make no assumptions about the distribution of the data, are used and explained in Chapter 9. The conditions that must be met for a $t$-test are as follows; the data should be normally distributed and each observation within the data should be independent from one another. Welch's $t$-test is an adaptation of Student's $t$-test [48] that does not require the variance of the samples to be the same.

The $t$-statistic and its associated degrees of freedom can be calculated and compared to the t-distribution to determine a $p$-value. The equation for calculating Welch's $t$-statistic is shown in Equation 2.4,

$$t = \frac{\Delta\bar{x}}{s_{\Delta\bar{x}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}} \tag{2.4}$$

where $s_{\bar{x}_j} = \frac{s_j}{\sqrt{n_j}}$ is the standard error of sample $j$'s mean. The standard error of the sample mean gives an indication of how accurately the sample data represents the population. The degrees of freedom of the $t$-statistic can be calculated as in Equation 2.5.

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \tag{2.5}$$

A $p$-value for Welch's $t$-test between two independent samples represents the probability of observing a difference in means between the samples at least as large as the observed difference assuming the null hypothesis is correct. Given that the $p$-value represents a probability, researchers generally choose a $p$-value threshold that allows them to be reasonably confident that the results are not a product of chance. In 1925, Fisher suggested a $p$-value threshold of $p < 0.05$, deviations greater than two standard deviations from the mean, as a convenient significance threshold [49]. He argued that with this threshold, the null hypothesis would falsely be rejected only 1 in 22 experiments. This threshold has widely been adopted as the standard threshold of significance, though $p$-value threshold, usually referred to as $\alpha$, of 0.01 or 0.1 are sometimes used as well [50]. Hypothesis testing allows for the identification of statistically noteworthy effect, but it does not quantify the magnitude of those effects.

### 2.2.2 Effect Size

An effect size is often reported alongside a $p$-value to indicate the magnitude of a significant effect. Cohen's $d$ [29] is an appropriate effect size when measuring the difference between two independent sample means. Cohen's $d$ is defined in Equation 2.6

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \tag{2.6}$$

where $s_p$ is the pooled standard deviation between the two independent samples defined in Equation 2.7.

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \tag{2.7}$$

The criteria for characterizing effect sizes is that $d > 0.2$ is a small effect, $d > 0.5$ a medium effect, and $d > 0.8$ a large effect. There are many other effect sizes used to quantify the magnitude of effects, but the only other one relevant to this manuscript will be discussed with non-parametric hypothesis testing in Chapter 9.

### 2.2.3 Error and Correction

Statistical hypothesis testing categorizes errors as Type I, false positives, and Type II, false negatives. For Type I errors, the null hypothesis is incorrectly rejected. In other words, a measured effect resulting from chance is falsely claimed to be a real effect. By choosing a $p$-value threshold, $\alpha = 0.05$, researchers accept that there is a 5% chance of falsely claiming a meaningful result. When an experiment makes multiple statistical inferences, the probability of making a Type I error in at least one of those tests increases. A common correction to this is to decrease the threshold on each subsequent hypothesis test. This process is called the Bonferroni correction [51]. A Bonferroni correction reduces the threshold value by the number of statistical tests. For a study applying $m$ statistical tests, the original $\alpha$ value becomes $\alpha/m$ for the $m$th test. Though it is a solution to the problem of Type I error inflation, the Bonferroni correction is sometimes criticized for overcorrecting, or being too conservative for large numbers of statistical tests [52].

A Type II error occurs when the null hypothesis is incorrectly not rejected. In other words, a real effect is falsely claimed to be the result of random fluctuation. The probability of making a Type II error is directly related to the statistical power of the dataset. Statistical power is given by $1-\beta$, where $\beta$ is the likelihood of making a Type II error. Statistical power is heavily dependent on sample size, so for large samples, Type II errors are less likely to

occur. It is worth noting that if the Bonferroni correction overcorrects for Type I error, there will be a higher rate of Type II error, essentially turning false positives into false negatives [53].

### 2.2.4 Bootstrapping

Bootstrapping is a method of random sampling with replacement. In the same way that inferences are made about a population using sample data, bootstrapping can be used to make inferences about sample data with sub-samples. The benefit of this process is that the accuracy of the inference between the sub-sampled data and the sample data can be measured, while the accuracy of the inference between the sample data and the population is unknown. If the bootstrapped data offers a good approximation of the sample data, then the sample data should offer a good approximation of the population. Bootstrapping is used in this manuscript to account for random deviations in the sample by effectively increasing the sample size with bootstrapped replications. Bootstrapping is one of the steps in module analysis that is used throughout this manuscript and described in the next chapter.

# Chapter 3

## Introduction to Module Analysis

The following Chapter introduces the necessary information for the methods Modified Module Analysis and Modified Module Analysis - Partial that will be used throughout this document.

## 3.1 Background

Students' conceptual understanding of physics and coherently applied errors in that understanding, misconceptions, have long been important research areas within Physics Education Research (PER). This research has been fostered by the introduction of multiple-choice conceptual instruments such as the Force Concept Inventory (FCI) [9], the Force and Motion Conceptual Evaluation [12], the Conceptual Survey of Electricity and Magnetism (CSEM) [13], the Brief Electricity and Magnetism Assessment (BEMA) [54], and the Quantum Mechanics Concept Assessment (QMCA) [15]. Recently, network analytic techniques, called Modified Module Analysis (MMA) or Modified Module Analysis - Partial (MMA-P), have been applied to the FCI and FMCE [55–58] and have identified common student incorrect answering patterns as well as potential flaws in the instruments.

Chapters 4 and 6 discuss the application of MMA and MMA-P to the CSEM and the BEMA. Chapter 5 compares misconception structures across institutions with these same methods with the FCI. This chapter serves as an introduction to the method and provides the background for its application to physics conceptual inventories.

### 3.1.1 Theories of Knowledge

Physics Education Research has investigated common student difficulties with conceptual physics since its inception. These difficulties have often been conceptualized as "misconceptions" or "alternate conceptions/hypotheses." Early work [2, 59, 60] analyzed common alternate views of force and acceleration. Halloun and Hestenes [7, 8] extended these works by developing a taxonomy of "common sense concepts" representing incorrect

reasoning about Newtonian mechanics. The FCI was developed partially with the goal of measuring these incorrect reasoning patterns [9]. The authors of the FCI provide a detailed description of the misconceptions measured by the instrument. This description was refined by Hestenes and Jackson [61] to produce a complete taxonomy of misconceptions measured by the FCI. This taxonomy is also used in Chapters 4 and 5.

The misconception model of incorrect reasoning was important in the development of many conceptual physics instruments, particularly the FCI and FMCE. Other models of incorrect thinking have also been used to explain student reasoning in physics. Two of the most important models are ontological categories [62–64] and knowledge in pieces [65, 66]. The ontological categories theory explains incorrect student reasoning as the misclassification of some quantity [62–64], for example the misclassification of force as a substance. If that substance can be used up, then one would predict that an object in motion after the application of a force would come to a stop as the force was used up. The knowledge-in-pieces framework suggests that student incorrect and correct reasoning is composed of small pieces of reasoning that are activated either singly or collectively to address a problem [65, 66]. Many researchers have explored variations of this model, conceptualizing the reasoning fragments as phenomenological primitives (p-prims) [65, 66], facets of knowledge [67], or resources [6, 68, 69].

Scherr provides a definition that contrasts the knowledge-in-pieces and misconceptions views which we adopted in this work [70]. The misconception view is "a model of student thinking in which student ideas are imagined to be determinant, coherent, context-independent, stable, and rigid," [70] while knowledge-in-pieces models student conceptions "as being at least potentially truth-indeterminate, independent of one another, context-

dependent, fluctuating, and pliable" [70].

MMA and MMA-P are quantitative methods that identify consistently selected responses to a multiple-choice instrument. They cannot identify the correct theory of knowledge to represent the student thinking that generated the consistent responses. For brevity, communities of incorrect responses are often discussed as resulting from misconceptions. The communities identified in the FCI and FMCE were often associated with misconceptions from Hestenes and Jackson's taxonomy [61]. The identification of communities of incorrect responses as misconceptions in the CSEM or the BEMA is far less clear. A more nuanced discussion of the classification of the incorrect communities is provided in Chapters 4-6.

## 3.2 Network analysis

Network analysis is a versatile set of techniques that have been applied across many different research areas. A network is a series of nodes (vertices) interconnected by edges to form a graph. Numerical weights may be associated with the edges representing some feature of the relationship between the nodes. These techniques have been used in a variety of studies outside of education, such as mapping electrical signals in the brain as functional networks [71], the difference between passing patterns in different teams at the World Cup [72], plants' response to bacterial infection [73], and the probability of becoming a homicide victim when living within a disadvantaged neighborhood [74]. Network analysis has also been fruitful within education research to study the structure of classrooms through the social interactions of students and teachers [75], undergraduate student representations of the relatedness of physics concepts through concept maps [76], and the difference between high

school students' and interdisciplinary professionals' emotional perception and conceptual knowledge of STEM [77].

### 3.2.1 Social network analysis in physics education research

Analyzing social structures through social network analysis has been the primary application of network analysis in PER. In a social network, actors, usually students or educators, are represented by nodes in a network, with edges representing some social interaction between the actors. Social networks have been used in physics education to characterize and test active learning environments [78–80], to predict future performance [81, 82], to predict retention and persistence within a degree program [83, 84], to explore physics self-efficacy and anxiety [85, 86], to explore interactions between lab groups by gender [80], to study conceptual change in student responses and discussions [87, 88], to determine the effect of informal learning environments and out of class relationships on class involvement and commitment [89, 90], and to explore the change in co-authorship behaviors in PER over time [91]. For an overview of network analysis in PER, see the review by Brewe [92].

### 3.2.2 Module Analysis in physics education research

Module analyses are a set of network analytic techniques used to analyze multiple-choice instruments [55–58, 93]. Module analysis was introduced by Brewe *et al.* as module analysis for multiple choice responses (MAMCR); MAMCR was applied to the responses of 143 first year physics students' FCI post-test results at a university in Denmark [55]. A network was formed in which the nodes represented incorrect responses and the edges represented the frequency of selection of both incorrect responses by the same student. When

the correct responses were included in the network, a single community appeared that hid any interesting structure; as such, only incorrect responses were retained. Nine communities were identified in this analysis, but only three were found to represent a coherent, underlying incorrect concept.

MAMCR inspired a series of further studies of conceptual instruments with modifications to the algorithm. Wells *et al.* attempted to replicate the MAMCR analysis and found that in their case, the algorithm did not scale to large datasets [56]. To produce a scalable algorithm, the frequency of common selection was replaced by the correlation of selection. To calculate this correlation, the selection of each response to the instrument is dichotomously scored producing a vector of 150 values (the FCI has 30 items, each with 5 responses). Correct responses are removed leaving a vector with 120 entries. The correlation matrix of this vector forms the edge weights in the network. The modified algorithm was called modified module analysis (MMA) [56]. The communities extracted by MMA are generally small, which simplifies the identification of the reasoning which led to the responses being selected together. MMA was applied to 4500 responses to the FCI from an introductory calculus-based physics class [56]. The resulting communities were composed of blocked items and items consistently applying a variety of misconceptions: the circular impetus misconception, the largest force determines motion misconception, the motion implies active forces misconception, and two Newton's 3rd law misconceptions. All of these are described in detail in Hestenes and Jackson's [61] taxonomy of Newtonian misconceptions measured by the FCI.

As with other quantitative methods such as cluster analysis or factor analysis, the identification of the possible reasoning behind communities extracted in module analysis

relies upon the interpretation of the researchers. This process is greatly aided for the FCI by the detailed description of the instrument as it was introduced [9], the detailed description of misconceptions measured by the instrument provided by Jackson and Hestenes [61], and the detailed mapping of the granular knowledge measured by the instrument provided by Stewart *et al.* [94].

Like MAMCR, MMA was not productive in examining correct and incorrect responses in the same network. To remove this restriction, Modified Module Analysis-Partial (MMA-P) was developed by Yang *et al.* [58]; MMA-P replaces the correlation between the 120 dichotomously scored responses with the partial correlation correcting for overall instrument score for all 150 responses. Some responses may be correlated because only very high performing students choose them, and others may be correlated because only the lowest performing students choose them; the items are correlated through the overall instrument score. MMA-P corrects for these correlations by controlling for overall instrument score. The network produced by MMA-P includes communities of incorrect responses as identified by MMA, but also communities with a mix of correct and incorrect responses and communities with entirely correct responses. Yang *et al.* applied MMA-P to the same sample of FCI responses as used by Wells *et al.* and found very similar incorrect communities. The mixed communities indicated that some FCI items were not functioning as intended, and the completely correct communities were composed primarily of blocked items or isomorphic items. The module analysis algorithm applied in Chapters 4-6, MMA-P, is the same algorithm as developed by Yang *et al.* [58]; this algorithm will be used to construct the networks.

## 3.3 Methods

Chapters 4-6 are presented in chronological order and developments to MMA-P are made in each subsequent application. In this introduction, MMA-P will be presented as it is used in Chapter 4 and the developments to the method will be presented as they appear in subsequent chapters.

### 3.3.1 Correlation and Partial Correlation

The correlation, $r_{XY}$, between response $X$ and response $Y$, measures the degree of association of the responses and is calculated using Equation 3.1 for two continuous random variables $X$ and $Y$,

$$r_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{3.1}$$

where $E[X]$ is the expectation value, $\mu_i$ is the average of variable $i$, and $\sigma_i$ is the standard deviation of the same variable. The expectation value is defined as $E[X] = \sum_{j=1}^{N} \frac{X_j}{N}$, where $j$ indexes the observations of $X$ and $N$ is the number of possible item responses.

The partial correlation $r_{XY|Z}$ between response $X$ and response $Y$, controlling for the total instrument score $Z$ represents the degree of association between $X$ and $Y$ that does not result from $Z$. The partial correlation is defined in Equation 3.2:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}} \tag{3.2}$$

Partial correlation can be understood by considering linear regression. Linear regression can be used to control for the effect of $Z$ on $X$ or the effect of $Z$ on $Y$, where $Z$ is a

variable related to both $X$ and $Y$. Using $X$ as the dependent variable and $Z$ as the independent variable of the regression, the residuals of the regression represent the portion of $X$ not explained by $Z$. The partial correlation is the correlation between the residuals of the linear regression of $X$ and $Z$ and the residuals of a linear regression of $Y$ and $Z$.

### 3.3.2 Modified Module Analysis and Modified Module Analysis - Partial

Module analysis begins by forming a network of the responses to a multiple-choice instrument. Each response forms a node in the network. The responses for each student $i$ are formed into a vector $V_i$ of length $k \cdot n$ where $n$ is the number of items and $k$ is the number of responses per item. Each entry in this vector codes whether student $i$ selected response $l$ to item $j$; the entry is one if the response was selected, zero otherwise. The nodes in the network represent individual responses; response A to item 7 becomes node 7A. MMA and MMA-P differ in the way they construct the edges connecting the nodes in the network. In MMA, an edge connects two nodes if the correlation between the two responses $r$ is larger than some threshold. Only incorrect responses are analyzed in MMA. In MMA-P, an edge connects two nodes if the partial correlation $r_{XY|S}$ controlling for total CSEM score $S$ exceeds some threshold. In most previous studies utilizing modified module analysis, $r > 0.2$ was used. This threshold was selected to produce compact communities with theoretically understandable structure. The threshold also naturally removed the large negative correlations between different responses to the same item and ensured that each edge retained in the network represented a significant correlation. The application of techniques to reduce the complexity of a network, such as applying the threshold, is called sparsification in network analysis [95]. Some operations, such as requiring that nodes are selected by

some minimum number of students, directly remove nodes; most remove edges, but once all edges to a node are removed, the node itself is removed from the network. Item responses selected by fewer than 30 students were removed as statistically unreliable. Each edge was also checked for significance at $p > 0.05$ after a Bonferroni correction was applied. The sparsification process and its relation to sample size is discussed in more detail in Chapter 7.

Once the network is constructed, a community detection algorithm (CDA) is applied to identify communities within the network. In network analysis, a community is a set of nodes more closely related to each other than to nodes outside of the community. In this manuscript, very strong levels of sparsificaiton are used to produce compact disconnected subgraphs; a disconnected subgraph is called a "component" in network analysis. Different levels of sparsification would generate more connected structure; as such, we continue to use the term community. MMA and MMA-P use a global sparsification method which does not attempt to preserve structure resulting from responses selected by very few students. For networks with important structure on many levels, this may result in the removal of interesting structures [96]; however, for networks formed of conceptual inventory responses it seems likely this low level structure results from student mistakes when bubbling scantron sheets, unserious answering, and random noise. As such, global sparsification seems theoretically justified. The fast-greedy CDA was applied [97] to identify communities within the network. Wells *et al.* [56] showed that other community detection algorithms produced similar results to the fast-greedy CDA in most cases. The CDA was applied to 1000 bootstrap replications sampling the dataset with replacement. The community fraction $C$ is defined as the fraction of times any two nodes appeared in the same community. Communities were retained for

analysis when $C > 80\%$; the community was identified in 80% of the bootstrap replications. The boot package [98] in "R" was used for bootstrapping and the igraph package [99] in "R" was used for the community detection.

### 3.3.3 Misconception scores

Wells *et al.* [56] used the consistently selected incorrect responses identified by module analysis to define a misconception score which quantitatively captures the average fraction of misconceptions of each type selected by a student. This statistic measures the frequency of applying different misconceptions and should be related to how strongly they are held. Misconception scores represent the number of responses chosen that are associated with a misconception out of the total number of item responses that a student could possibly choose associated with the same misconception. For example, if responses 6A and 7A are associated with a misconception; a student can select either 0, 1, or 2 of these responses resulting in a misconception score of 0%, 50%, or 100% respectively. The score calculated for this misconception is the average of each student's misconception score.

| | | $R_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| $R_2$ | 0 | A | B |
| | 1 | C | D |

Table 3.1: $2 \times 2$ contingency table

Misconception scores can be calculated from the contingency table between two item responses, $R_1$ and $R_2$, as shown in above in Table 3.1. There are four possible combinations for two item responses: choosing neither, choosing $R_1$ and not $R_2$, choosing $R_2$ and not $R_1$, and choosing both. For example, if 600 students selected neither response ($R_1 = 0$ and

$R_2 = 0$), $A$ would equal 600. If 250 students selected response $R_2$, but not response $R_1$, then $C$ would equal 250.

The misconception score is calculated with Equation 3.3.

$$M_1 = \frac{B + C + 2D}{2N} \tag{3.3}$$

where $N = A + B + C + D$ is the total number of students responding to the instrument. The numerator of Equation 3.3 is derived by summing the "1's" in Table 3.1: $B + D$ for $R_1$ and $C + D$ for $R_2$. The $2N$ represents the total number of times either response could be selected. In other words, this provides the fraction of the misconception group that is chosen on average and should be associated with the likelihood for a student to apply that misconception.

The methods described in this chapter will be applied in Chapters 4, 5, and 6 and expounded upon in Chapter 7.

# Chapter 4

## Applying Module Analysis to the Conceptual Survey of Electricity and Magnetism*

*This chapter presents the work published in Physical Review Physics Education Research [93]. This work was constructed with collaborative efforts from James Wells, Rachel Henderson, and John Stewart. It was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY0108787. Data collection for this work was supported by National Science Foundation Grants No. EPS-1003907 and No. ECR-1561517. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 4.1 Introduction

This chapter applies Modified Module Analysis (MMA) to the CSEM to investigate coherent patterns of student responses to the instrument. MMA forms a network where the responses to the items in a multiple-choice instrument are the nodes and the edges represent the correlation between the responses. Network analysis identifies "communities" within the network of responses that are often selected together by a student. Two versions of MMA have been used to explore conceptual physics instruments, one using the correlation matrix (MMA) [56], the other the partial correlation matrix (MMA-P) [58]. The current work applied both versions to two large samples of CSEM responses from two different institutions. This chapter explored the following research questions:

**RQ1** What community structure is identified by network analysis of the CSEM? How are the communities associated with previously identified features of the instrument?

**RQ2** Does the community structure of the CSEM have communities related to Newtonian mechanics? If so, how do these communities compare to the communities identified in the FCI or the FMCE?

**RQ3** How do the communities identified by the two versions of Module Analysis, MMA and MMA-P, compare? How do the communities identified at different institutions compare?

This work identified a rich ecology of diverse types of incorrect reasoning, much broader than that identified in the FCI or FMCE. It used the community structure identified to calculate scores representing the relative strength of types of incorrect reasoning and offered

a modified scoring rubric for the CSEM which corrects for the relations found between items.

### 4.1.1 Prior Studies

This work makes extensive use of the results of four prior studies which will be referenced as Study 1 to Study 4.

#### Study 1

In Study 1, Maloney *et al.* [13] introduced the CSEM, provided a classification of items in the instrument, and discussed common errors made by students both pre- and post-instruction. The CSEM is an instrument with 32 items that measures a student's knowledge of concepts in electricity and magnetism. The instrument includes questions about topics commonly covered in introductory electricity and magnetism courses, such as conductors and insulators, Coulomb's law, superposition, electric fields, magnetic fields, and magnetic induction. Questions from two prior surveys about electricity and magnetism by Hieggelke and O'Kuma [100] were combined to create the CSEM. This study uses the version available at PhysPort [11].

Study 1 provided a general classification of the items in the CSEM: charge distribution on conductors/insulators (items 1, 2, 13), Coulomb's force law (items 3, 4, 5), electric force and field superposition (items 6, 8, 9), force caused by an electric field (items 10, 11, 12, 15, 19, 20), work, electric potential, field, and force (items 11, 16, 17, 18, 19, 20), induced charge and electric field (items 13, 14), magnetic force (items 21, 22, 25, 27, 31), magnetic field caused by a current (items 23, 24, 26, 28), magnetic field superposition (items 23, 28), Faraday's law (items 29, 30, 31, 32), and Newton's 3rd law (items 4, 5, 7, 24).

Study 1 also discussed common errors made by students; these errors were not referred to as misconceptions, but were often made on multiple items suggesting coherently applied incorrect knowledge. Students confuse the behavior of conductors and insulators (items 1 and 2). Students do not fully understanding the shielding of the electric field by conductors (items 13 and 14). Responses to item 14 also show a failure to understand Newton's 3rd law; this misunderstanding is also detected on items 4, 7, and 24. Students apply the larger object exerts more force misconception [61] on item 4; this misconception was also identified within the FCI in Study 3 and the FMCE [57]. On items 8 and 9, response D, students misunderstand how the addition of another charge affects the field. In the current study, response D to item 8 is abbreviated as response 8D. Students confuse the behavior of electric and magnetic fields in responses 23B, 23C, and 26B. Response 10B represents the force proportional to velocity misconception [61] also detected in the FMCE [57]. On items 19 and 20 students confuse the relation of changes in electric potential to the direction of the electric field.

**Study 2**

Prior network analytic studies have made extensive use of Constrained Multidimensional Item Response Theory (MIRT) models of the correct physical reasoning needed to solve the items in the FCI [94] and FMCE [101]. This work utilized a similar study of the CSEM [102] which is referenced as Study 2 in this work. All three studies identified the practice of "blocking" items into item groups as a source of correlations within the responses to the instrument. A group of items is blocked if all items in the group refer to a common stem describing the physical system or if one item explicitly refers to a prior item in the

block. The CSEM contains 3 item blocks {3, 4, 5}, {10, 11}, and {17, 18, 19}. In Study 2, items 4, 5, and 11 were eliminated from the analysis and only the first item in the block was retained because the answers to the latter items in the group directly depended on the earlier items. Items 18 and 19 were retained because it was felt that all items in the block could be answered independently. Study 2 also identified 3 groups of isomorphic items {6, 8}, {16, 17}, and {21, 27}. Isomorphic items all require the same solution process. In the prior network analytic studies of the FCI and the FMCE, responses to isomorphic items have often been detected both in the same correct communities and the same incorrect communities.

**Study 3**

In Study 3, Wells *et al.* introduced the modified module analysis (MMA) technique to study the FCI [56]. MMA uses the correlation between the responses to form edges in the network, which scales to larger networks better than than the original module analysis method used in PER, module analysis for multiple choice responses (MAMCR). The communities that result from MMA tend to only include a pair or a small number of responses, which allows the underlying idea that may lead a student to select those responses to be identified. MMA was applied to 4509 pretest and 4716 post-test FCI records of students in a introductory calculus-based physics class. Some of the structure found by MMA was due to blocked items within the instrument. Excluding these blocked items, there were five communities on the post-test that represented Newton's 3rd law misconceptions, the motion implies active forces misconception, the motion implies active forces for the centrifugal force misconception, the circular impetus misconception, and the largest force determines motion misconception, as described in Hestenes and Jackson's taxonomy of misconceptions

measured by the FCI [61].

**Study 4**

In Study 4, in order to include both correct and incorrect responses in a module analysis, modified module analysis using partial correlations (MMA-P) was developed by Yang *et al.* [58]. In MMA-P, the partial correlation matrix is used to connect each pair of responses, controlling for the effect of total score. MMA-P can identify communities that include only correct responses, only incorrect responses, or both correct and incorrect responses. As such, a richer set of communities are found with MMA-P than with MMA. MMA-P was applied to the same sample of FCI responses as in Study 3. The completely incorrect communities were very similar to those identified by MMA. The completely correct communities generally involved blocked items or were identified as isomorphic by MIRT [94]. The mixed communities suggested some items in the FCI were not functioning correctly.

### 4.1.2 Studies informing the construction of the CSEM

Mechanics instruments such as the FCI were written after a great deal of research had been performed on students' conceptual understanding of mechanics. Substantially less research had been performed on electricity and magnetism before the CSEM was published. This section summarizes some of the work that informed the construction of the CSEM.

Maloney [103] performed a study using activities that tested students' conception of magnetic poles post-instruction in a general physics course and found that the majority of students had an alternate conception based around the idea that "magnetic poles are charged." Guruswamy *et al.* [104] studied student understanding of the transfer of charge

between conductors through a small set of questions about simple charge transfer experiments. The vast majority of students tested, from 8th grade to physics majors in senior-level university physics courses, could not correctly explain or predict what happens when charged conductors come in contact with each other. Törnkvist *et al.* [105] investigated the understanding of electric fields of introductory college students. They concluded that the majority of students personify field lines as isolated entities in real space, rather than a set of curves that represent mathematical properties of space. Galili [106] studied high school student difficulties with the field concept in electricity and magnetism. Students often regressed in their understanding of mechanics concepts that were previously understood when learning about the concept of fields in electricity and magnetism. Studies on student reasoning about and understanding of the superposition of electric fields have shown that many students struggle with causality in electricity and magnetism [107, 108]. Most notably, some students do not recognize the existence of a field unless there is some motion caused by the field.

### 4.1.3 Prior studies of the CSEM

While not as thoroughly studied as the FCI, multiple studies have used the CSEM to explore student conceptual thinking about electricity and magnetism.

Planinic [109] compared Croatian students to American students in a study that introduced six overarching conceptual areas measured by the CSEM. In order to produce groups large enough for analysis, Planinic qualitatively grouped the eleven concepts reported in Study 1. These areas include electric charge and force (items 1, 2, 3, 5, 6, and 8), Newton's laws (items 4, 7, 10, 24, and 27), electric field and electric force (items 9, 12, 13, 14, 15), electric potential and energy (items 11, 16, 17, 18, 19, 20), magnetic field and magnetic force

(items 21, 22, 23, 25, 26, and 28), and induction (items 29, 30, 31, 32). The difficulty of the individual items in each group was very similar for the two populations [109].

Performance differences between men and women on the CSEM have also been explored. A difference in CSEM test performance by gender was measured by Kreutzer and Boudreaux [110] and was greatly reduced by pedagogical changes. Gender differences on the CSEM were also examined by Kohl and Kuo [111] through a transition to studio physics. Studio physics is a model of instruction where students take an active role in learning by doing hands-on activities and group work during instruction rather than in separate labs. They found the gap in normalized gain was reduced by this transition [112]. Henderson *et al.* examined gender difference in performance on the CSEM and compared these to gender differences in other multiple-choice problems in a university physics class [113]. A 5% difference in CSEM post-test scores was measured; however, the difference in conceptual test questions was only 3% with no difference observed in quantitative test questions. For a more complete summary of gender differences in conceptual understanding of physics see Madsen, McKagan, and Sayre [114].

Other studies have analyzed a subset of items in the CSEM. Leppävirta examined Newton's 3rd law using items 4, 5, 7, and 24 and showed that the number of students with an incorrect model of Newton's 3rd law decreased from 20% to 10% from pretest to post-test [115].

Changes from pretest to post-test were also explored by Meltzer using items 18 and 20 to investigate how electric field concepts intersect with potential concepts [116]. Meltzer reported that students' conflation of electric field magnitude with potential slightly increased post-instruction. Study 1 also reported the conflation of electric and magnetic fields. All

other incorrect response pairs between items 18 and 20 decreased post-instruction, while the correct response pairs significantly increased.

Karim *et al.* [117] used the CSEM to study the degree to which graduate teaching assistants (TA) could predict introductory physics students' alternate conceptions in electricity and magnetism. TAs were told to choose the response that they thought would be the most chosen incorrect response by the students in the introductory course. The TAs were likely to choose responses that included both correct and incorrect concepts, but their choices did not correspond to the most frequent incorrect responses by students in the course.

### 4.1.4 Misconceptions

In the current work, network analysis identified two types of Newtonian mechanics misconceptions described by Hestenes *et al.* [9]. Within the "Active Force" group of misconceptions, the "velocity proportional to applied force" misconception showed that Newton's 2nd law was not well understood. A student applying this misconception reasons that the velocity of a particle in motion will be equal or proportional to the force applied to the object. This misconception was also identified in the FMCE by Wells *et al.* [57] using MMA.

Two misconceptions involving a misunderstanding of Newton's 3rd law were identified as forming the "Action/Reaction Pairs" group of misconception by Hestenes *et al.* [9]: "greater mass implies greater force" and "most active agent produces greatest force." Students applying greater mass implies greater force misconception reason that the larger or heavier object exerts more force than the smaller or lighter object. Students applying the most active agent produces greatest force misconception assume an active object produces more force than an inactive object; for example, a small car pushing a large truck exerts more

force on the truck than the truck exerts on the car. Module analysis identified both misconceptions in the FCI (Study 3) and the FMCE [57]. In both studies, responses demonstrating both Action/Reaction Pairs misconceptions were found in the same community.

The FCI and FMCE were developed within the misconception framework and network analysis largely supported this framework identifying only one community which was better described as a phenomenological primitive within the knowledge-in-pieces model of student knowledge. The CSEM was not developed from a robust framework of misconceptions, but rather responses were taken from common open-response answers to the items. No robust taxonomy of misconceptions of electricity and magnetism similar to Hestenes and Jackson's taxonomy of misconceptions of mechanics [61] has been published. This offers the possibility that a broader set of structures not identifiable as misconceptions may be identified using MMA.

## 4.2 Methods

### 4.2.1 Sample

This study was performed on samples from two US institutions.

#### Sample 1

Sample 1 was collected from spring 2003 until spring 2012 at a southern land-grant university with a total enrollment of about 25,000 students. The demographics of the undergraduates at the university were 77% White, 8% Hispanic, 5% African American, 2% Asian with other groups composed of no more than 3% [118]. The overall undergraduate

population had ACT scores ranging from 23-29 (25th to 75th percentile) [118].

The CSEM post-test was given in the introductory calculus-based electricity and magnetism courses serving scientists and engineers. Only students with complete post-test responses were retained for the study ($N = 2538$). The CSEM was given as a quiz after instruction and the student's scores were recorded for part of their course grade. The same professor taught the course for the total period studied. The course implemented a number of interactive engagement instructional practices in both the lecture and laboratory.

**Sample 2**

Sample 2 was collected from fall 2015 to spring 2019 at an eastern land-grant university with a total enrollment of about 30,000 students. The overall undergraduate demographics were 80% White, 6% international, 4% Hispanic, 4% African American, 4% reporting two or more races, 2% Asian with other groups composed of no more than 1% [118]. The overall undergraduate population had ACT scores ranging from 21-26 (25th to 75th percentile) [118].

The CSEM post-test was given in the introductory calculus-based electricity and magnetism courses serving scientists and engineers. Only students with complete post-test responses were retained ($N = 3595$). The CSEM was given as a quiz post-instruction and graded for a small amount of course credit. The course was managed by a single lead instructor in the time studied who taught the majority of the lecture sections. Interactive engagement methods were applied in both the lecture and laboratory.

### 4.2.2 Modified Module Analysis

Modified Module Analysis (MMA) and Modified Module Analysis - Partial (MMA-P), as described in Study 3 and 4, were applied to the CSEM. These methods are described in Chapter 3.

## 4.3 Results

Modified Module Analysis and Modified Module Analysis - Partial were applied to the CSEM; Figure 4.1 shows the communities detected with their respective correlation thresholds. Three of the four analyses were performed with $r > 0.15$ instead of the $r > 0.20$ threshold used in Study 3 and 4. The threshold was adjusted to provide a fairly disconnected, but rich, set of communities. The nodes of the networks that correspond to correct responses are labeled with an asterisk (*). The magnitude of the correlation or partial correlation between nodes in Figure 4.1 is proportional to the line thickness.

The total scores of the two samples were quite different. For Sample 1, the CSEM post-test percentile score was $61.8 \pm 15$; for Sample 2, the post-test percentile score was $45.5 \pm 18$.

Table 4.1 provides a summary of the communities as well as a possible explanation of the common reasoning applied. A number of the communities identified were composed of items within item blocks. There is substantial evidence that the practice of blocking items produces correlations between the items that are not related to consistently applied reasoning [94, 101]. These groups are labeled "blocked items" and are discussed separately. The communities are divided into three classes: communities composed entirely of incorrect

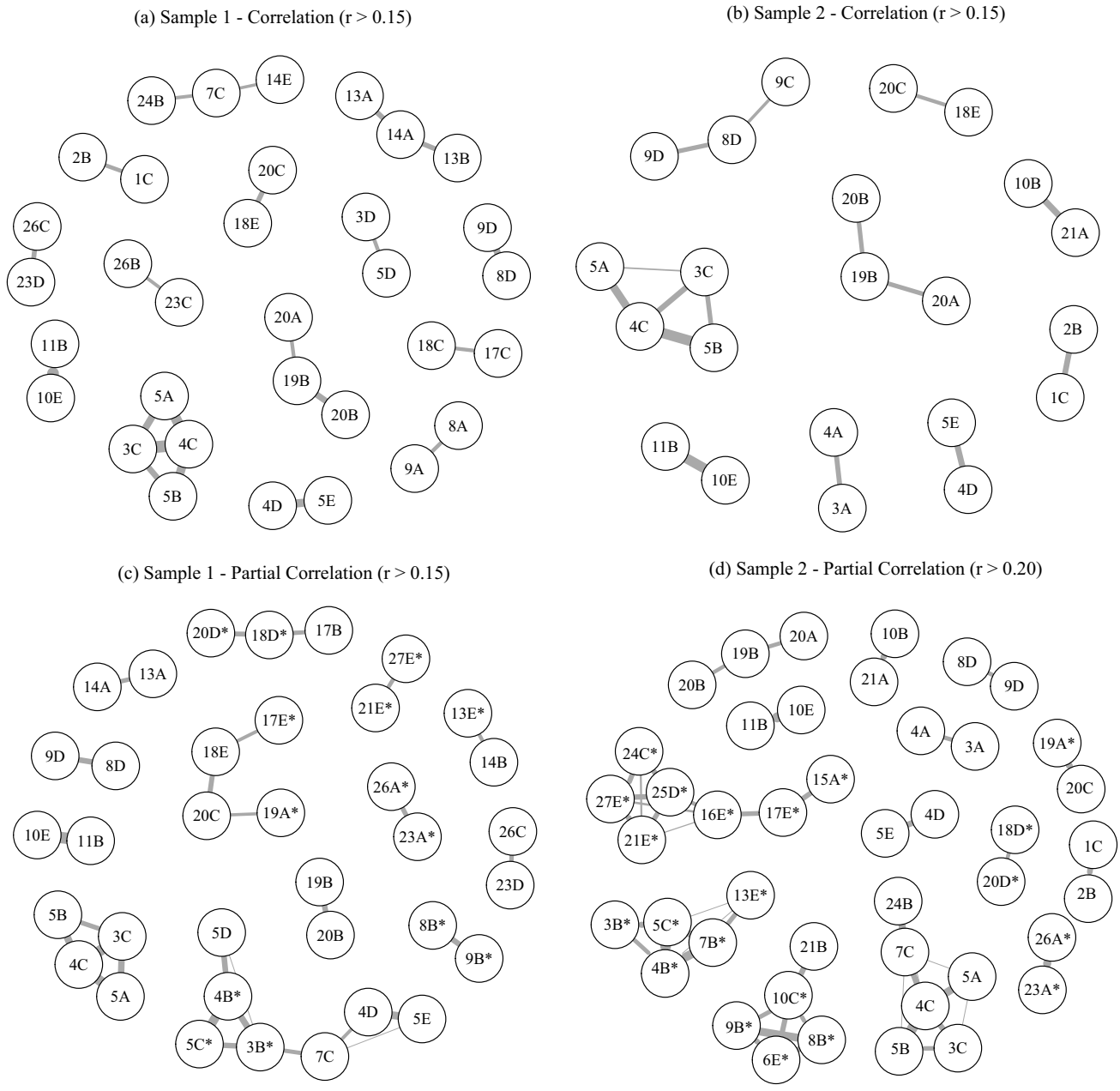


Figure 4.1: Communities detected in the CSEM. Figure (a) shows the correlation network for Sample 1. Figure (b) shows the correlation network for Sample 2. Figure (c) shows the partial correlation network for Sample 1. Figure (d) shows the partial correlation network for Sample 2. The strength of the correlation or partial correlation is represented by the line thickness.

responses, mixed communities composed of both correct responses and incorrect responses, and communities composed entirely of correct responses. Some smaller communities appear in one community but also appear as a part of a larger community. These sub-communities

Table 4.1: Communities identified in CSEM responses. Communities labeled × are sub-communities of another community. Sample 1 is abbreviated S1; Sample 2, S2. Responses in parenthesis are completely connected. Responses separated by dashes are only connected to each other. Blocked items are marked consistent if later items apply correct reasoning to an incorrect earlier item.

| Community | Correlation | | Partial Correlation | | Misconception/Principle/Explanation |
|---|---|---|---|---|---|
| | S1 | S2 | S1 | S2 | |
| Completely Incorrect Communities | | | | | |
| 1C, 2B | ⊗ | ⊗ | | ⊗ | Conductor and insulator misconceptions. |
| 3A, 4A | | ⊗ | | ⊗ | Blocked items - consistent. |
| 3D, 5D | ⊗ | | | | Blocked items - (5D) $E \propto 1/r$. |
| 3C, 4C, 5A, 5B | ⊗ | ⊗ | ⊗ | × | Blocked items - (3C, 4C) consistent - (5A) consistent - (5B) $E \propto 1/r$. |
| 5A - (3C, 5B, 4C) - 7C - 24B | | | | ⊗ | Blocked items - see text. (7C, 24B) - Newton's 3rd law misconceptions. |
| 4D, 5E | ⊗ | ⊗ | × | ⊗ | Blocked items - consistent. |
| 14E - 7C - 24B | ⊗ | | | | Newton's 3rd law misconceptions. |
| 8A, 9A | ⊗ | | | | Charge on axis produces zero field. |
| 8D, 9D | ⊗ | × | ⊗ | ⊗ | Interaction between charges modifies superposition. |
| 9C - 8D - 9D | | ⊗ | | | Interaction between charges modifies superposition. |
| 10B, 21A | | ⊗ | | ⊗ | (10B, 21A) Velocity proportional to applied force. (21A) Electric-magnetic field undiscriminated. |
| 10E, 11B | ⊗ | ⊗ | ⊗ | ⊗ | Blocked item - consistent. |
| 13A, 14A | × | | ⊗ | | Conductor does not shield electric field. |
| 13A - 14A - 13B | ⊗ | | | | Conductor does not shield electric field - (13B) Like charges attract. |
| 17C, 18C | ⊗ | | | | Equipotential spacing proportional to electric field (18C); Electric field proportional to work (17C). |
| 18E, 20C | ⊗ | ⊗ | × | | Electric field-potential undiscriminated. |
| 19B, 20B | × | × | ⊗ | × | Electric field points to higher potential. |
| 20A - 19B - 20B | ⊗ | ⊗ | | ⊗ | Electric field points to higher potential. (20A) Electric field-potential undiscriminated. (20A) Electric field proportional to equipotential spacing. |
| 23C, 26B | ⊗ | | | | Electric-magnetic field undiscriminated. |
| 23D, 26C | ⊗ | | ⊗ | | Left hand rule. |
| Mixed Correct and Incorrect Communities | | | | | |
| 5D - (4B*, 5C*, 3B*) - 7C - 4D - 5E | | | ⊗ | | Blocked items. See text. |
| (6E*, 8B*, 9B*, 10C*) - 21B | | | | ⊗ | (10C*, 21B) Electric-magnetic field undiscriminated. |
| 13E*, 14B | | | ⊗ | | Shielding works symmetrically. |
| 17B - 18D* - 20D* | | | ⊗ | | Work proportional to electric field. |
| 17E* - 18E - 20C - 19A* | | | ⊗ | | (18E, 20C) Electric field-potential undiscriminated. |
| 19A*, 20C | | | × | ⊗ | (20C) Electric field-potential undiscriminated. (19A*, 20C) Field points to lower potential. |
| Completely Correct Communities | | | | | |
| (3B*, 4B*, 5C*, 7B*) - 13E* | | | | ⊗ | Coulomb's law. |
| 8B*, 9B* | | | ⊗ | × | Coulomb's law and superposition. |
| 15A* - 17E* - 16E* - (25D*, 21E*, 24C*, 27E*) | | | | ⊗ | See text. |
| 18D*, 20D* | | | × | ⊗ | Relation of equipotential spacing to field. |
| 21E*, 27E* | | | ⊗ | × | Zero velocity implies zero magnetic force. |
| 23A*, 26A* | | | ⊗ | ⊗ | Magnetic field of wires. |

are marked by × to show the continuity between communities more accurately. Examination of Figure 4.1 shows that while many communities are completely connected, some are not. In order to display the connectivity of individual communities in Table 4.1, responses that are completely connected are shown in parenthesis, while responses that are only connected

to each other are separated by dashes.

The items in the CSEM can be broadly divided into 4 topics: electrostatics (items 1 to 15), electric potential (items 16 to 20), magnetostatics (items 21 to 28), and magnetic induction (items 29 to 32). No communities were detected that involved the magnetic induction items. Examination of these items suggests that items 29, 30, and 32 should require related reasoning; failure to find responses to these items in the same community may indicate that the items are not functioning as intended. Generally, communities were formed within these broad topics with a few notable exceptions. Communities were identified that involved misconceptions of Newtonian mechanics that involved both electrostatic and magnetostatic items. Some students also answered questions about the electric and magnetic field in the same way producing communities representing the electric-magnetic field undiscriminated misconception. Study 3 and 4 also identified both blocked items and isomorphic items as important in contributing to the formation of communities. The electric-magnetic field undiscriminated terminology was selected to mirror that of the misconception classifications of the FCI in Hestenes and Jackson's taxonomy [61].

### 4.3.1 Blocked Items

The CSEM contains 3 item blocks {3, 4, 5}, {10, 11}, and {17, 18, 19}. In the first two blocks, the answers to later items in the block are dependent on the answers to the earlier items in the block. Items 17, 18, and 19 can be answered independently, but could be correlated if a student misinterpreted the shared description of the physical system. The items in an item block are often found in the same community in Table 4.1. Both Study 3 and 4 also found that items within item blocks often were found in the same community

in the FCI and FMCE. In Table 4.1, communities largely composed of blocked items are labeled blocked items. When the responses to later items would have been correct if the response to the earlier item was correct, the community is labeled "consistent."

Many correct, mixed, and incorrect communities contain combinations of responses to items 3, 4, and 5. The common stem of this set of items presents the student with two objects each with charge $Q$ which exert a force $F$ on each other. The charge of one of the objects is then increased to $4Q$. Item 3 asks about the new force on the object of charge $Q$. Item 4 asks about the new force on the object of charge $4Q$. Item 5 asks about how the force on the $4Q$ charge changes if the objects are moved 3 times farther apart. Many of the later responses in these communities are the correct response if an earlier response were correct as in communities {3A, 4A}, {3C, 4C, 5A}, and {4D, 5E}. The community {4D, 5E} is found independently and as part of a larger mixed community. The responses to item 3 in these communities all represent a failure to understand the relation of electric charge to electric force. The correct answers to items 3, 4, 5 were found both in a mixed and a completely correct community. In the mixed community, 5D is also associated with the correct responses (3B*, 4B*, 5C*) demonstrating the $E \propto 1/r$ misconception. Response 5B in the community {3C, 4C, 5A, 5B} also applies the $E \propto 1/r$ misconception. Response 5D in community {3D, 5D} may also apply this misconception except the student mistakenly selected $F/4$ instead of $4F$ for item 3 (this is the most straightforward way to reconcile these two inconsistent responses). Communities mixing items 3, 4, and 5 with other items were also identified and are discussed later.

Blocked items 10 and 11 are only found together in one incorrect community, {10E, 11B}; response 11B is the correct response if response 10E were correct. As in prior work,

the practice of blocking items generates relations, correlations, between items that make the score of the items difficult to interpret. A simple correct or incorrect scoring of items 4, 5, and 11 almost certainly understates a student's understanding of the items; a modified scoring rubric that takes into account relations between the items is proposed in Sec. 4.3.10.

### 4.3.2 Mechanics misconceptions

Multiple communities were identified where the responses represented misconceptions about Newtonian mechanics. Students continue to apply some non-Newtonian misconceptions identified with the FCI and FMCE in introductory mechanics courses in introductory electricity and magnetism courses. Three responses, 7C, 14E, and 24B, as well as inconsistent responses to items 3 and 4 show that Newton's 3rd law is not well understood. Hestenes and Jackson [61] identified two Newton's 3rd law misconceptions in the FCI; greater mass implies greater force and most active agent produces greater force. While neither is completely appropriate for the CSEM, some responses to item 4 and responses 7C and 24B seem more aligned with the greater mass implies greater force misconception; response 7C applies a greater charge implies greater force reasoning, while response 24B applies greater current implies greater force reasoning. Response 14E, which involves an asymmetric application of the shielding by a conductor of the electric field, does not fit the larger implies more force model. All three responses indicate that the student does not apply Newton's 3rd law in a variety of contexts. The identification of these responses as misconceptions requires further study. It may be students' thinking is better modeled by the knowledge-in-pieces framework where the student is applying a p-prim such as "large implies large."

Response 7C was also associated with different combinations of the item 3, 4, 5 block in

completely incorrect community {5A - (3C, 5B, 4C) - 7C - 24B} and mixed community {5D - (4B*, 5C*, 3B*) - 7C - 4D - 5E}. The inclusion of 7C in these communities is unusual; item 4 can be answered either using Newton's 3rd law or Coulomb's law. Response combinations {3B*, 4B*} and {3C, 4C} represent a correct application of Newton's 3rd law, though in the latter Coulomb's law is not applied correctly. It is unclear why incorrect applications of Newton's 3rd law, 7C and 24B, should be associated with its correct application. This may indicate item 4 is being answered using some reasoning pattern other than Newton's 3rd law.

The responses in the community {10B, 21A} apply the force proportional to velocity misconception identified in Hestenes and Jackson's taxonomy [61].

### 4.3.3 Isomorphic items

Study 2 identified 3 groups of isomorphic items {6, 8}, {16, 17}, and {21, 27}. Isomorphic items require very similar reasoning for their solution. In both Study 3 and 4, isomorphic items often formed both incorrect and correct communities. The isomorphic items were less important in forming communities in the CSEM. All three sets of isomorphic items were identified together in a completely correct community or as correct items in a mixed community. Only items 21 and 27 were identified independently as a correct community; there was no corresponding incorrect community. These items ask for the magnetic force on a stationary charge; students are consistently reasoning correctly, but do not apply consistent incorrect reasoning. The students may not hold misconceptions for these items.

While items 6E* and 8B* were identified in a mixed correct community, item 8 was more consistently associated with item 9 in incorrect communities. These two items may

be described as nearly isomorphic; item 8 involves the change in electric force on a point charge as a third charge is added to the system; item 9 also uses two point charges and adds a third asking about the electric field. In Study 2, the solution of the items differ only by the application of the relation between the force and field ($\vec{F} = q\vec{E}$). The point charge experiencing the force is a positive charge, and therefore force and field are parallel. Responses 8D, 9C, and 9D were found in incorrect communities in all samples; all explicitly test the misconception that the addition of an additional charged object somehow modified the total field beyond simply adding the field of the new object. It seems likely this misconception is responsible for the formation of the communities that include items 8 and 9.

Items 23 and 26 differ only by the principle of superposition. Item 23 asks about the addition of the fields of two wires; item 26 asks for the direction of the field of a single wire. Study 2 found item 23 discriminates very weakly on the principle of superposition; as such, these items may be effectively isomorphic explaining their identification as a correct community {23A*, 26A*} and as an incorrect community {23D, 26C} where the students incorrectly apply the right hand rule.

Items 16 and 17 are only found in one correct community connecting disparate items. Both are coded as applying the definition of electric potential in Study 2. While this is true, the items are fairly different with item 16 applying the principle that only differences in electric potential are physically important while item 17 asks to rank the work needed to move through a field where the equipotentials are given. Study 2 did not test the principle that only differences in electric potential are important; it seems likely, given the failure to find items 16 and 17 in the same communities, that this was an oversight and they may not actually be isomorphic.

### 4.3.4 Electric-magnetic field undiscriminated

Study 1 reported that students often conflated electric fields and magnetic fields in items 23 and 26. The community {23C, 26B} is consistent with their analysis. The wires in these items seemed to be viewed as charges with magnetic field pointing either away from or toward the wire in the incorrect community. There could be a reverse conflation of electric fields and magnetic fields in the community {10E, 11B} where students seem to apply the Lorentz force law to electric fields. This connection is less clear; response 10E states that a charged particle released in an electric field remains at rest. There could be many reasons for the selection of this response. Mixed community {10C*, 21B} also contains responses that were answered symmetrically for the electric and magnetic fields, where a uniform field produces a constant acceleration on a charge released in the field.

### 4.3.5 Electrostatic communities

Items 13 and 14 appear in two completely incorrect communities, {13A, 14A} and {13A - 14A - 13B}, and one mixed community {13E*, 14B}; these communities demonstrate different incorrect ideas about conductors and shielding. The two completely incorrect communities show the student does not understand that conductors shield the electric field. The mixed community applies the principle of shielding symmetrically reasoning that conductors not only shield their interior from external electric fields, but their exterior is also shielded from internal electric fields.

The completely correct community {(3B*, 4B*, 5C*, 7B*) - 13E*} is curious. Items 3 to 7 all apply Coulomb's force law; however, the connection of this group with item 13

is unclear. Item 13 correctly applies the principle of the shielding of a conductor's interior from electric fields and forces.

### 4.3.6 Electric potential communities

Items 16 to 20 all require the definition of electric potential. Item 16 was discussed as part of the isomorphic group formed of items 16 and 17 and is only identified in one community which mixed electric potential and magnetostatic items. The remaining electric potential items are found in a large number of communities. The composition of these communities can shed light on how these items are functioning. The completely correct community {18D*, 20D*} includes responses that correctly represent the relation of equipotential spacing and field strength, 18D* and 20D*, and the relation of equipotential magnitude and electric field direction, 20D*. Mixed community {17B - 18D* - 20D*} connects 17B to 18D* which may indicate the student believes work is proportional to electric field independent of distance. The mixed community {19A*, 20C} is identified independently and as part of the larger community {17E* - 18E - 20C - 19A*}. Both responses 19A* and 20C correctly capture the concept that electric field points toward lower electric potential; response 20C applies the electric field-potential undiscriminated misconception. Responses consistent with the electric field proportional to the electric potential may indicate a student applied this misconception. This implies a student being scored as incorrect on 20C is demonstrating some of the knowledge the correct students are demonstrating. This pair is connected with response 18E in mixed community {17E* - 18E - 20C - 19A*}; 18E also represents the electric field-potential undiscriminated misconception. The combination {18E - 20C} is also identified as an independent community. Response 17E* is the correct response that work is the

difference in potential, but the connection to 18E suggests the student does not understand the relation of electric field to potential and may be reasoning that work is proportional to electric field. The connection of the correct response to the incorrect response suggests the item 17 may not be functioning as intended.

Many completely incorrect communities that are not part of mixed communities were also identified. In community {17C, 18C}, 18C applies the misconception that electric field magnitude increases with equipotential spacing; this suggests 17C may apply the work is proportional to electric field misconception. All items in the community {20A - 19B - 20B} suggest the student believes electric field points to higher potential. Item 20A could also represent the electric field magnitude increases with equipotential spacing or the electric field-potential undiscriminated misconception.

### 4.3.7 Magnetostatic communities

Incorrect communities {23D, 26A} and correct communities {21E*, 27E*}, {23A*, 26A*} were discussed earlier as examples of isomorphic or nearly isomorphic item groups; response 24B was discussed as applying Newton's 3rd law misconceptions. Response 21B was identified as applying the electric-magnetic field undiscriminated misconception. The only other community formed of magnetostatic items combined these items with electric potential items and is discussed in the following section.

### 4.3.8 Other communities

The community {15A* - 17E* - 16E* - (25D*, 21E*, 24C*, 27E*)} mixes electric potential and magnetostatic items. The physical reasoning linking these items is unclear.

One possible explanation for this community is that many of the responses are of a type preferentially avoided by students [119] for non-physical reasons. Response 16E is a "none of the above" response, responses 17E and 24C reports all the quantities are equal, response 21E and 27E are "zero" responses, and response 25D have the quantities in ascending order.

### 4.3.9 Misconception scores

Table 4.2: Misconception Scores. Items not in parenthesis represent independent misconceptions. Items in parenthesis represent a misconception only if both items are selected and are counted as one response.

| Misconception | Responses | Misconception Score | |
|---|---|---|---|
| | | Sample 1 | Sample 2 |
| Conductor and insulator misconceptions | 1C, 2B | 10.0% | 18.1% |
| Electric field proportional to $1/r$ | (3A, 4D), (4B,5D), (4C, 5D) | 12.8% | 17.2% |
| Newton's 3rd law misconceptions | $(3 \neq 4)$, 7C, 14E, 24B | 17.6% | 26.5% |
| Charge on axis produces zero field | 8A, 9A | 2.5% | 4.4% |
| Interaction between charges modifies superposition | 8D, 9C, 9D | 17.3% | 31.1% |
| Velocity proportional to applied force | 10B, 21A | 12.9% | 24.5% |
| Shielding misconceptions | 13A, 13B,14A | 27.8% | 56.2% |
| Electric field-potential undiscriminated | 18E, 20C | 32.3% | 31.4% |
| Electric-magnetic field undiscriminated | (10B, 21A) (10C*, 21B), 23C, 26B | 2.2% | 13.3% |
| Electric field points to higher potential | 19B, 20A, 20B | 13.0% | 29.5% |
| Left hand rule | 23D, 26C | 4.6% | 8.7% |

Misconception scores are presented in Table 4.2. A misconception score is a percentage representing the number of responses a student selected that are identified with a given misconception out of the total number of CSEM responses identified with that misconception. This work refines the calculation of misconception scores by identifying combinations of items which taken together represent a misconception; these combinations are counted as a single instance of a misconception and represented by two items in parenthesis in Table 4.2. Multiple combinations of responses to items 3 and 4 represent a failure to apply Newton's 3rd law; one misconception is counted for each time the response to item 3 is different to the response to item 4.

Sample 2 had consistently higher misconception scores than Sample 1 for all categories

excluding electric field-potential undiscriminated, which was commensurate for both samples. This was expected given the average CSEM score for each sample. The high misconception score of 56.2% for Sample 2's shielding misconceptions category along with the lack of communities composed of items 13A, 13B, and 14A in Sample 2's networks could be explained by a general misunderstanding of shielding by conductors. The high misconception score for both samples' electric field-potential undiscriminated misconception as well as the community appearing in both samples' networks indicates that the conflation of electric fields and potentials could be a strongly held misconception among many students post-instruction.

Communities with a very low misconception score such as {8A, 9A} and {23D, 26C} contain responses chosen by very few students. The correlation coefficient, which is used to define the relation between responses, is fairly insensitive to sample size. The identification of these communities selected by few students implies those students who select one response in the community also selected the other response.

### 4.3.10 Alternate Scoring Rubric

The consistent answering patterns on blocked items and the existence of isomorphic groups of items suggest an alternate scoring rubric for the CSEM might be appropriate. Items 3, 4, and 5 are blocked items all measuring the understanding of Coulomb's force law. We propose item 4 be counted as correct if it is consistent with the response to item 3 or is itself correct. Likewise, item 5 should be counted as correct if it is answered consistently with item 4 or is itself correct. Repeated instances of the response E, "other," for items 3, 4, and 5 should not be considered consistent for the grading of items 3, 4, and 5. Items 10 and 11 are blocked items involving the motion of a charged particle in a uniform electric

field. Response 11B should be counted as correct if response 10E was selected; both assert the particle remains at rest. Items 21 and 27 both test whether the student understands the magnetic force on a stationary charge is zero; we suggest grading these two items as a block where one point is received if both are correct. This would also serve to lower the weighting of this relatively specific piece of knowledge on the overall score.

Item 4 could either be answered using the same reasoning as item 3 using Coulomb's law or by applying Newton's 3rd law to the result of item 3. The majority of communities identified involving item 4 do not consistently contain the other Newton's 3rd law items. Because the reasoning applied is ambiguous, it seems unfair to penalize the student for an answer which may have involve correct reasoning based on a prior incorrect answer.

An alternate rubric for the BEMA is available at PhysPort [11]. CSEM items 3 to 5 are very similar to items 1 to 3 on the BEMA; the alternate scoring of item 5 suggested is consistent with the suggestions for the BEMA. The modified rubric to the BEMA also suggests grading one pair of isomorphic items as a block where the student receives one point if both are correct.

With the modified scoring producing a score out of 31, the Sample 1 CSEM percentage score became $62.2 \pm 15$ and the Sample 2 CSEM percentage score became $46.4 \pm 18$. The modified scoring resulted in a 0.4 percentage point increase in average CSEM score for Sample 1 and a 0.9 percentage point increase in average score for Sample 2.

## 4.4 Discussion

### 4.4.1 Research questions

This work explored three research questions which will be addressed in the order proposed.

*RQ1: What community structure is identified by network analysis of the CSEM? How are the communities associated with previously identified features of the instrument?* The community structure identified by MMA and MMA-P was discussed in detail in the previous section. Communities of incorrect responses or mixed correct and incorrect responses were identified in electrostatics, electric potential, and magnetostatics but not in magnetic induction. Most communities connected items within the individual topics; however, communities of items incorrectly applying Newtonian mechanics and communities applying the same reasoning to electric and magnetic fields included items from multiple topics. This indicates Newtonian misconceptions continued to be applied to multiple contexts involving both electricity and magnetism. The isomorphic items identified in Study 2 were much less important to the community structure than in the FCI or FMCE in Studies 3 and 4. This indicates the consistent incorrect reasoning patterns were more independent from the correct reasoning in electricity and magnetism than they were in mechanics.

Blocked items were included in many communities offering additional evidence that the practice of blocking items makes an instrument difficult to interpret statistically. Unlike prior studies, analysis of patterns of blocked responses were useful in further understanding the answering patterns of students. This suggests a simple all or nothing scoring of each CSEM item may fail to correctly capture the student's understanding of electricity and magnetism.

An alternate scoring scheme that grades items contingently based on the responses to earlier items might produce a more accurate reflection of student understanding. One such scheme was proposed in Sec. 4.3.10. While some individual students scores were modified, the overall CSEM average changed little.

Many of the communities identified represented consistent application of incorrect reasoning identified in the introduction of the instrument in Study 1. Network analysis showed that students were applying similar incorrect reasoning across multiple contexts. There was a substantial variety of incorrect reasoning possibly indicating the CSEM is a superior instrument to the FCI for exploring the structure of incorrect physical understanding. Some communities represented failure to broadly understand general concepts: shielding misconceptions and conductor-insulator misconceptions. Some represented naive conceptions where two disparate quantities were assumed to behave in the same manner: electric-magnetic field undiscriminated, electric field-potential undiscriminated, or electric field proportional to work. Some seemed to represent simple mistakes: incorrectly applying the right-hand rule or believing the electric field points toward higher potential. There were also a few misconceptions that assumed additional properties of the electric field not consistent with physics: interaction between charges modifies superposition and a charge on the axis produces zero field. It seems likely that all these general types of consistent incorrect reasoning are also present in mechanics but are not detected by the FCI which was developed explicitly to detect common misconceptions.

Most completely correct communities involved items requiring either the same reasoning or very similar reasoning. Many of the mixed communities involved chains of items where members of the community are only connected in pairs. Many of the mixed communities

involved electric potential. Most of these communities were only identified in Sample 1. This may indicate that electric potential is not being effectively covered for the students in Sample 1; this is supported by a misconception score for the electric field-potential undiscriminated in Sample 1 that is much more similar to the weaker performing Sample 2 than other misconception scores. The mixing of correct and incorrect responses for electric potential in Sample 1 may indicate the correct responses are being selected without correct understanding and that the scores on these items may overstate student knowledge. Some mixed communities were also formed when an incorrect response required multiple reasoning steps; the student may reason correctly on one step and incorrectly on another step ultimately selecting the incorrect response. This can be seen in the community {19A*, 20C} where the student correctly reasons that electric field points to lower potential but does not understand the relation between equipotential spacing and field strength.

*RQ2: Does the community structure of the CSEM have communities related to Newtonian mechanics? If so, how do these communities compare to the communities identified in the FCI or the FMCE?* Two general groups of mechanics misconceptions were identified in the CSEM by module analysis. The first, responses {10B, 21A}, applied the velocity proportional to applied force misconception and were identified by both MMA and MMA-P, but only in the lower performing Sample 2. Communities testing this misconception were also identified in the FMCE by Wells *et al.* [57]. These responses had moderate misconception scores of 12.9% for Sample 1 and 24.5% for Sample 2. As such, while this misconception does persist after instruction in mechanics, it is not one of the most applied misconceptions in the electricity and magnetism class.

Four items are potentially related to an incorrect application of Newton's 3rd law:

7C, 14E, 24B, and inconsistent responses to items 3 and 4. These items were identified in communities in a number of combinations. All communities containing 24B also contain 7C. Two of the three communities containing 7C also contain 24B. The only community containing 14E also contains 7C and 24B. Response 7C is also found in communities containing incorrect responses to item 4, either response 4C or 4D. Responses 7C and 24B are consistent with a greater charge/current implies greater force misconception which is analogous to the "greater mass implies greater force" misconception in mechanics. Response 14E only states the forces are different; the relative charges of the two objects involved are not given and, therefore, it is not possible to determine if a modification of greater mass implies greater force is being applied. The inconsistent association of 7C to responses to item 4 and the failure to associate either responses 14E or 24B with these items seems to indicate inconsistent responses to items 3 and 4 are selected for reasons other than applying Newton's 3rd law incorrectly. The communities formed by all four items are very different than the communities of Newton's 3rd law items identified in Study 3 and Wells *et al.* [57] which were tightly interconnected. The Newton's 3rd law responses in the current study were sparsely connected. This seems to indicate that students' application of incorrect reasoning about Newton's 3rd law is far less consistent when applied to electricity and magnetism or that these items measure reasoning beyond Newton's 3rd law.

*RQ3: How do the communities identified by the two versions of Module Analysis, MMA and MMA-P, compare? How do the communities identified at different institutions compare?* A total of 20 completely incorrect communities were identified by MMA or MMA-P in the two samples; only 6 were identified by both techniques in both samples. Four of these were responses to blocked items. MMA identified 19 communities in the two samples, 8 were

identified in both samples. Sixteen communities were identified in the higher performing Sample 1, while only 11 were identified in the lower performing Sample 2. All MMA-P communities identified in Sample 1 were also identified as MMA communities. The 8 communities identified by MMA-P in Sample 1 were substantially fewer than the 16 identified by MMA suggesting many of the communities identified by MMA in Sample 1 resulted from correlations through total test score. This, and the facility to include correct answers, suggests MMA-P is a superior technique for exploring consistent student answering patterns.

For Sample 2, only one fewer community was identified by MMA-P than MMA. All MMA-P communities in Sample 2 were also identified by MMA except {5A - (3C, 5B, 4C) - 7C - 24B}. The additional community is curious because it mixed incorrect application of Newton's 3rd law in responses 7C and 24B with a correct application in the pair 3C and 4C. As such, the MMA-P algorithm produced a simplified community structure and removed many communities which were correlated through total CSEM score.

There was very little similarity between the mixed communities in the two samples; only one community was identified in both samples. Many more mixed communities were identified in Sample 1. Both mixed communities for Sample 2 involved confounding different concepts through the electric-magnetic field undiscriminated and the electric field-potential undiscriminated misconceptions. Application of naive conceptions is consistent with the overall low performance of the sample and shows students continue to have a fundamental misunderstanding of the central concepts of electricity and magnetism post-instruction.

Four of the six completely correct communities were consistent between samples. One of the inconsistent communities {15A* - 17E* - 16E* - (25D*, 21E*, 24C*, 27E*)} may be related to issues of test logic such as avoiding "none of the above responses" with little

relation to physical reasoning. The other inconsistent community is curious {(3B*, 4B*, 5C*, 7B*) - 13E*}; four of the responses are related to Coulomb's force law while response 13E* involves understanding that a conductor shields its interior from the electric field.

### 4.4.2 Other Observations

This work used the term misconception for the reasoning generating incorrect response communities. This is almost certainly too broad a classification. In the work introducing the FCI, Hestenes *et al.* differentiated naive conceptions from misconceptions and reported that some misconceptions were weakly held and some were strongly held [9]. They separate naive conceptions of kinematics where various combinations of position, velocity, acceleration, and force are not differentiated in student thinking from more robust incorrect models similar to medieval theories of motion. Many incorrect communities resulted from the naive application of reasoning associated with one concept to a different concept: electric-magnetic field undiscriminated and electric field-potential undiscriminated. A number of communities seem to result from simple mistakes which seem unlikely to represent strongly held beliefs: electric field points to higher potential and the left hand rule. Other communities do not seem to represent a single consistent fragment of incorrect reasoning, but more general topics that are not well understood: conductor and insulator misconceptions and that the conductor does not shield the electric field. Many communities could be explained by an alternate framework of student knowledge. The "larger implies larger" p-prim might explain the electric field-potential undiscriminated, the electric field points to higher potential, the electric field proportional to work, and the Newton's 3rd law misconceptions. However, if this was the case one might expect these items to appear together in more communities.

It seems unlikely that any of the incorrect reasoning about electricity and magnetism resulted from the students' lived experiences. It also seems likely that similar differences in strength and kind of incorrect reasoning are found in mechanics. It is important to understand these differences because effective and efficient methods of addressing incorrect thinking may differ depending on the origin and degree the student has internalized the misconception. For example, a misconception rooted in a robust naive theory applicable and productive across many contexts may require substantial instructional resources to overcome; however, a simple mistake such as using your left hand to do a cross-product might be corrected by simply pointing out the mistake.

The variety of coherent incorrect responses identified in this study suggests that an instrument like the FCI developed to measure strongly held misconceptions may miss a variety of other incorrect answering patterns. Evidence supporting this possibility can be found in the MMA analysis of the FMCE by Wells *et al.* [57]. Both the CSEM and FCI use five responses per item where the responses were developed from interviews and student responses to open-response applications of the instrument. The FMCE employs a different strategy offering the student up to nine responses per item where the items generally exhaust all possible responses. The present study and the MMA analysis of the FCI in Study 3 both discarded nodes selected by fewer than 30 students producing relatively compact disconnected communities with a correlation threshold of 0.15 or 0.20. This was not the case for the FMCE. When Wells *et al.* removed only nodes selected by fewer than 30 students, an exceptionally complex community structure was produced at the $r > 0.20$ correlation threshold; to reproduce the compact community structure of the FCI, only nodes selected by 20% of the students were retained. When this threshold was relaxed to 10% of the students, a

substantially more complex community structure was exposed where the communities generally could be identified with a pattern of incorrect reasoning. Often this incorrect reasoning was not represented as a misconception in Hestenes and Jackson's taxonomy [61]. The work on the FMCE and the present work on the CSEM suggest there is a much richer ecology of incorrect conceptions than those measured by the incorrect reasoning on an instrument such as the FCI which focuses on the most strongly held misconceptions. The relations identified between items within the CSEM suggested an alternate scoring rubric was needed to correctly represent student understanding; one such rubric was proposed.

## 4.5 Implications

Network analysis was successful in identifying consistently selected incorrect and correct answers in the CSEM. This provides support for the need for a taxonomy of misconceptions (common incorrect answering patterns) of electricity and magnetism similar to the taxonomy of mechanics misconceptions provided by Hestenes and Jackson [61]. The variety of misconception communities identified in the CSEM was commensurate with the variety of misconception communities measured by the FCI in Study 2 and larger than the number of communities found in the FMCE by Wells *et al.* [57]. This suggests the existence of a rich set of electromagnetic misconceptions; a systematic identification of the full set of misconceptions would benefit instructors. The calculation of misconception scores similar to those in Table 4.2 should allow instructors to determine what kinds of incorrect reasoning most need to be addressed.

## 4.6 Conclusion

The CSEM was constructed to assess students' conceptual understanding of electricity and magnetism. This study explored the use of MMA and MMA-P as productive ways to identify communities of correlated responses to individual items within the CSEM. In general, MMA-P produced a richer set of communities and eliminated communities of items correlated through total instrument score. Overall, a number of communities were identified for the two samples; however, the explanations for the reasoning represented by these communities varied. A substantial number of the identified incorrect communities consisted of blocked items providing continued support that the practice of blocking items can produce correlations that are not related to physical reasoning. The consistent identification of blocking as generating psychometric problems for the primary conceptual instruments used in PER may suggest the need for a new generation of conceptual inventories. Multiple communities were formed of responses where the response to later items would be correct if the response to an earlier item was correct. This suggests that the scoring rubric to the CSEM should be modified to include relations between responses. A modified scoring rubric was proposed, but changed overall CSEM post-test averages little. Most communities of completely incorrect responses and mixed correct and incorrect responses consisted of items with the same subtopic, either electrostatics, electric potential, or magnetostatics. Some of communities connected items in multiple subtopics including misconceptions about mechanics and a failure to differentiate the electric and the magnetic field. The results suggest the existence of a rich collection of misconceptions about electricity and magnetism which are consistently applied by students after instruction in introductory physics. This collection

was much more diverse than those identified by the FCI and FMCE which may indicate these instruments do not fully characterize the scope of coherent incorrect reasoning.

# Chapter 5

## Comparing Conceptual Understanding Across Institutions with Module Analysis*

*This chapter presents the work published in Physical Review Physics Education Research [120]. This work was constructed with collaborative efforts from James Wells, David E. Pritchard, and John Stewart. This work was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY0108787. Data collection for this work was supported by National Science Foundation Grants No. EPS-1003907, No. ECR-1561517, and No. HRD-1834569. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 5.1 Introduction

The Force Concept Inventory (FCI) is an instrument used to evaluate students' conceptual understanding of Newtonian mechanics using items which test Newton's three laws, one-dimensional kinematics, and two-dimensional kinematics [9]. The FCI has been one of the most commonly used and, accordingly, studied conceptual instruments in physics since its introduction in 1992. The FCI along with the catalog of common misconceptions it measures [61] has been transformative to PER. Hake collected FCI data from multiple institutions to show traditional teaching methods were broadly ineffective at improving student understanding [28]. The Hake study provided the impetus for the ongoing effort to move to active learning strategies in all physics classes. Recent studies across many institutions have continued to demonstrate the efficacy of these methods [45]. Eliminating misconceptions, stable context insensitive alternate scientific theories, continues to represent a significant challenge for PER. An overview of research using the FCI is provided in Sec. 5.1.3.

There have been many quantitative studies of the FCI, the Force and Motion Conceptual Evaluation (FMCE), and the Conceptual Survey of Electricity and Magnetism (CSEM). Many of these studies have applied factor analysis and have been largely ineffective at extracting meaningful substructure from the FCI [94, 121–124]. The authors of the FCI argue that the instrument was not constructed to factor [9, 125]. The reason for this is evident in the correlation matrices presented by Stewart *et al.* [94]; the FCI items are deeply interconnected often mixing different physical principles in different ways. Factor analysis also only considers the correct responses, not the incorrect responses representing misconceptions around which the FCI is built. Module analysis, which can identify complex substructures

and relations within both the correct and incorrect responses to an instrument, has been consistently productive at identifying theoretically explainable structures within the responses to multiple-choice instruments. Further, module analysis identifies consistently selected correct and incorrect responses allowing the determination of incorrect thinking that is applied across multiple contexts. These incorrect ideas, misconceptions, may indicate areas where instructional interventions may most productively be directed.

### 5.1.1 Research Questions

In recent research, patterns in students' incorrect responses to these instruments have been identified by network analytic techniques applied to the FCI, FMCE, CSEM, and QMCA [55–58, 93, 126]. The current work applied the network analytic technique called Modified Module Analysis-Partial (MMA-P) to five samples of FCI responses from five different U.S. institutions. Prior work using module analysis has been restricted to single samples in most cases and two samples in a study of the CSEM. These samples came from institutions with student populations with fairly commensurate levels of incoming high school preparation. The five samples used in the present study were drawn from institutions with a broad range of student high school academic preparation and prior knowledge of physics. As such, the present study should advance the understanding of whether module analysis results are fairly universal across institutions with differing student populations. Further, when two samples were available, comparison of the community structure was primarily qualitative. Network analysis offers a wealth of quantitative comparison metrics, some of which are applied in the current work. This work will apply some of these metrics to provide the quantitative comparisons of the five samples not available in prior studies. Module

analysis, like other network analysis methods, requires the setting of a number of parameters to control the density of the network. These parameters have been set qualitatively in past studies; the present study will investigate a possibly productive means of setting the primary parameter, the correlation threshold, more systematically.

The following research questions were explored in this study:

**RQ1** How does the community structure of the FCI identified through module analysis compare across multiple institutions? What does this community structure imply about student understanding of mechanics?

**RQ2** How can the primary parameter required by module analysis be selected quantitatively?

**RQ3** What quantitative network analytic metrics are productive for characterizing institutional differences and similarities identified by module analysis? What do these metrics imply about the student understanding of mechanics?

### 5.1.2 The Force Concept Inventory

The FCI contains 30 items, each with four incorrect responses and one correct response. Many of these incorrect responses were specifically constructed to be attractive to students applying common misconceptions. The version of the FCI used in this study was released in 1995 [127] and can be found at PhysPort [11].

### 5.1.3 Prior studies of the FCI

A thorough summary of the prior research using the FCI was presented in previous module analysis studies [56, 94, 128]. An overview is provided below.

#### The structure of the FCI

When formulating the FCI, Hestenes, Wells, and Swackhammer separated the introductory physics curriculum on forces into six unique conceptual dimensions and described which concepts each FCI item was created to measure. Soon after the introduction of the instrument, other researchers challenged whether this internal division was actually measured by the FCI.

Several studies have applied Exploratory Factor Analysis (EFA) to understand the structure of the FCI. These studies dichotomously scored each item as correct or incorrect. Huffman and Heller applied EFA to 145 high school student responses to the FCI [121]. This analysis identified only two out of the six factors described by Hestenes *et al.*: "Kinds of Forces" and Newton's 3rd law. When EFA was applied to 750 students at the university level, the only factor identified was "Kinds of Forces" [121]. Scott *et al.* performed a factor analysis of 2150 college student post-test responses and found five factors were required for the optimal model; one factor explained much of the variance [123]. Using a related dataset, Scott and Schumayer repeated the factor analysis using multidimensional item response theory (MIRT) also identifying the five factor model as optimal [124]. Semak *et al.* also reported an EFA of the FCI using 427 pretest and post-test responses finding six factors were required for the optimal model [122]. Stewart *et al.* also performed a factor analysis

using MIRT on 4716 post-test responses [94] showing a nine factor model was optimal. The factors identified were strongly related to the practice of item blocking or chaining and the existence of a small number of isomorphic groups of items in the instrument. An item block is a group of items that all refer to a common stem. Two problems are isomorphic if they both can be solved by the same reasoning. None of these analyses recovered the structure proposed by the authors of the FCI; many of the extracted factor structures mixed items requiring different reasoning for their solution. These factor analyses examined only the correct answer structure of the FCI; additional techniques are required to examine the incorrect answers along with the correct answers. Module analysis is one such technique.

### Misconceptions

The FCI was created within a misconceptions framework. The misconceptions framework holds that students have a belief system of commonsense alternative ideas that are stable, largely context-independent, and resistant to change. Misconceptions are fundamentally scientific hypotheses that happen to be false and not errors in reasoning. Examples of misconceptions identified by Hestenes, Wells, and Swackhamer include impetus dissipation and active forces [9].

Impetus is an internal motive force that continues to carry an object forward after the initial external force no longer acts. Impetus dissipation is the idea that this impetus will dissipate and the object will stop unless it is replenished, somewhat analogous to gasoline in a car. When students apply this to circular motion (circular impetus) the students are applying the idea of "training," where objects continue to do what they "learned" when given the initial impetus [9]; the object remembered it was traveling in a circle.

The active force misconception is the idea that only active agents, usually living or in motion under their own power, can exert forces and cause motion. This explains not only the motion of objects (a couch moves because a person pushes it), but also the interactions between objects (a moving car exerts a force on a parked car, but not vice versa) [9].

## 5.2 Methods

### 5.2.1 Sample

This work examined FCI pretest and post-test responses from five U.S. institutions. These will be denoted as Samples 1 to 5 in what follows. Demographic data, undergraduate populations, and ACT 25th-75th percentiles for all institutions in these samples were obtained from the National Center of Education Statistics [129]. All samples contained only matched pretest/post-test responses with no missing responses.

Sample 1: 49% White, 22% Hispanic/Latino, 9% non-resident alien, 8% Asian, 5% two or more races, 4% Black or African American, and 1% American Indian or Alaska Native.

Sample 2: 75% White, 9% Hispanic/Latino, 4% two or more races, 4% Black or African American, 3% non-resident alien, 3% Asian, and 1% American Indian or Alaska Native.

Sample 3: 73% White, 18% Black or African American, 3% Hispanic/Latino, 2% two or more races, 1% non-resident alien, 1% Asian, and 1% American Indian or Alaska Native.

Sample 4: 32% White, 26% Asian, 16% Hispanic/Latino, 12% non-resident alien, 7% Black or African American, and 5% two or more races.

Sample 5: 38% White, 18% Asian, 12% Hispanic/Latino, 12% non-resident alien, 8% Black or African American, and 6% two or more races.

The size of the sample $N$, the total undergraduate population of the institution, and the 25th to 75th percentile range for the ACT scores of the institution are shown in Table 5.1.

These samples among others were collected by Pritchard as part of a work to improve item response theory analysis of the FCI [130]. While largely a convenience sample, these five were used because of both the size of three of the samples and the range of selectivity of all five institutions measured by ACT score range.

Table 5.1: Sample Description

| Sample | N | Undergraduate Population | ACT 25th-75th Percentile |
|---|---|---|---|
| 1 | 9606 | 44,000 | 22-29 |
| 2 | 4360 | 23,000 | 23-30 |
| 3 | 1496 | 19,000 | 22-30 |
| 4 | 466 | 4,000 | 33-35 |
| 5 | 213 | 10,000 | 33-35 |

### 5.2.2 Modified Module Analysis - Partial

Modified Module Analysis - Partial was applied to pretest and post-test responses to the FCI. An overview of MMA-P is provided in Chapter 3. Improvements on the method as it was applied in Chapter 4 are provided below. Note, for the remainder of this chapter, we will shorten partial correlation to correlation and use $r$ as the partial correlation coefficient for brevity.

### 5.2.3 Partial correlation threshold

In MMA-P, a threshold value for the partial correlation coefficient $r$ was used to sparsify the network. In previous module analysis studies, the sparsification criteria was selected

qualitatively [55–57, 93, 101]; the minimum value of $r$ was selected which produced networks with sufficiently small communities that the common reasoning required by items in the community could be identified. In this work, a more quantitative method was used to choose the threshold. The MMA-P networks were calculated using a range of $r$ thresholds; for these networks, the average community size (ACS) was plotted against the total number of communities (NC). The ACS is the average number of nodes in a community. The correlation threshold was chosen as the $r$ value for which this plot was changing most rapidly. At this correlation threshold, the community structure is simplifying rapidly with changing $r$. This is similar to selecting the optimal number of factors in an exploratory factor analysis by examining the scree plot and choosing the number of factors at the "knee" in the plot.

### 5.2.4 Sparsification and statistical power

Prior MMA and MMA-P studies used single large datasets or two large datasets of commensurate size. The current study uses five datasets of very different sizes; some elements of sparsification interact with sample size and need to be considered if the goal is to compare networks across institutions.

In prior MMA and MMA-P studies, one of the sparsification operations was to remove nodes selected by fewer than 30 students. These studies all used large samples of at least 2500 students; as such, the 30 student threshold removed only responses selected by less than approximately 1% of students. This threshold was introduced to remove the inevitable small background of students who misread questions or bubble scantron sheets incorrectly; these errors introduce responses not related to physics reasoning. Three of the five samples used in this work are smaller than in previous studies; two substantially smaller. The purpose of

this study is to compare MMA-P results across institutions; applying a 30-student response threshold would represent a substantially different percentage of total responses removed at the five institutions studied. To allow fair comparison, a response threshold of 5% was used in this study. This was selected to allow the retention of at minimum nodes with 10 responses in Sample 5. Analysis in Chapter 7 suggests that at even this small sample size, MMA-P can identify statistically significant structure.

The sparsification operations applied in this study are the minimum student response threshold (5% in this study), requiring edges represent correlations between nodes with significance of $p < 0.05$ after a Bonferroni correction is applied, requiring edges to have positive correlations, requiring those correlations to be above a correlation threshold (generally around $r > 0.2$ where $r$ is the partial correlation coefficient between nodes), and requiring the edge be detected in the same community in 80% of bootstrap replications. Because the Bonferroni correction depends on the number of statistical tests performed, the order of these operations should be investigated. In this study, we chose to apply the Bonferroni corrected significance threshold first because we felt the highest priority should be to eliminate the consideration of statistically insignificant structures; however, we acknowledge an argument can be made for applying the student response threshold first to minimize the number of statistical tests performed. Chapter 7 present a comparison of the resulting structure if the student response threshold is applied first or after the Bonferroni corrected significance threshold. For all samples, the order of the response threshold and the significance threshold does not change the number of nodes in the final network for the post-test; some small differences are found in the pretest network for Samples 1 to 4. The pretest differences were more pronounced for Sample 5. As such, MMA-P is generally not sensitive to the order of

applying the response threshold and the significance threshold. The reason for this is likely that the $r > 0.2$ correlation threshold is a very strong criteria ($r = 0.1$ represents a small effect and $r = 0.3$ a medium effect) making the significance threshold unimportant. Even at the size of Sample 5, a correlation of $r > 0.2$ is significant with a small $p$ value. The difference in the number of final nodes between the response threshold of 30 in prior studies and 5% in this study was also examined. There was little effect for Samples 1 to 4 for the post-test; however, the number of nodes in Sample 5 changed from 14 with the 5% threshold to 8 with the 30 threshold. Differences were smaller in the pretest networks.

Naturally, nodes removed by either the 30 or 5% response threshold are selected by a few students. Chapter 7 also presents an analysis of the correlation of small occupation nodes; with sufficiently consistent answering, even infrequently selected nodes, can have statistically significant correlations. This analysis also revealed that, for Sample 5, correlations between nodes needed to be at least 0.35 to pass the significance threshold test. This and the inconsistencies observed above suggest that Sample 5 is too small to resolve any but the most correlated network structure; as such, we focus on comparisons of Samples 1 to 4 and discuss Sample 5 only as a partially resolved network structure and as an example of the information which can be extracted by MMA-P even for smaller samples. This is fundamentally an issue of statistical power; at the size of Sample 5 there is insufficient statistical power to resolve structure with the same detail as other samples.

### 5.2.5 Multiplex networks

Multiplex networks are networks composed of multiple layers where each layer is itself a network. The same node may be present in many layers; nodes in multiplex networks

are called "actors" [131]. In general, actors may be connected through edges that represent different types of relations in different layers of the network. As an example, a multiplex network could be used to represent social and professional connections where actors are people and different layers represent different mediums in which people interact (work, home, social media, etc.). For a more complete explanation of multiplex networks consult Dickison *et al.* [131] or Kivelä *et al.* [132].

A multiplex network was formed applying MMA-P to the FCI response data from each institution studied individually creating 5 distinct networks. These networks were then added as layers forming a multiplex network. As the networks were computed independently using the MMA-P algorithm of Yang *et al.* [58], the different sample sizes did not restrict their use in a multiplex network. The multiplex network framework is used for the depth of layer comparison tools available. The actors in this context are item responses and the edges are the partial correlations between pairs of item responses. While we propose no explicit interaction between the layers, each layer represents features of the structure of Newtonian thinking measured by the FCI at a single institution. We will find this thinking extremely similar across institutions leading to an implicit interaction between the layers in the form of the general structure of conceptual Newtonian reasoning. The "R" package multinet [133] was used to construct the multiplex network.

### 5.2.6 Network comparison metrics

In this work, the primary benefit of combining the five networks into a single multiplex network is the availability of a rich set of tools to identify common structure in multiplex networks and metrics to characterize those networks. This work will utilize only a small

subset of the available analysis methods.

The Clique Percolation Method (CPM) is an efficient means of identifying overlapping communities in multiplex networks [134]. The CPM identifies communities which share $k$ edges in $m$ layers. Figures 5.1 and 5.2 show an example of the clique percolation method with $k = 1$ and $m = 3$. Cliques with one edge that appear in the networks of at least 3 of the 4 largest samples are shaded with the same color. Clique percolation can also be used to simplify the process of identifying sets for further network comparison metrics, such as the set of triangle communities with $m = 1$ and $k = 3$ [135]. A triangle is a completely connected sub-network with 3 nodes. As an example, consider the Sample 4 pretest network in Figure 5.1. The completely connected 3-node communities are {4E*, 15A*, 28E*} and {17B*, 25C*, 26E*} which each count as one triangle. The completely connected 4 node community {5B*, 11D*, 13D*, 18B*} contains 4 completely connected 3-node groups and counts as four triangles; therefore, the Sample 4 pretest network contains 6 total triangles.

Many network comparison metrics are available for multiplex networks. In this work, we report the Coverage Index (CI) [136] for a variety of structures found in the networks. The CI measures the similarity between two sets by dividing the size of the intersection of the sets by the size of each set. The intersection of sets $A$ and $B$ is the set containing all elements that are found in both sets. For two sets $A$ and $B$, two coverage indexes can be calculated: $\mathrm{CI}_A = N(A \cap B)/N(A)$ and $\mathrm{CI}_B = N(A \cap B)/N(B)$ where the function $N(X)$ computes the size of the set $X$. The CI provides a natural measure of the degree to which one set has members in common with another set. CI is calculated for three network structures: actors (nodes), edges, and triangles. CI results are represented using the corrplot package [137] in "R" as shown in Figure 5.4.

## 5.3 Results

Table 5.2 shows the sample size, pretest average, post-test average (mean ± standard deviation), normalized gain, and Cohen's $d$ between pretest and post-test.

Table 5.2: Descriptive statistics

| Sample | $N$ | Pretest Average % | Post-test Average % | Normalized Gain | $d$ |
|---|---|---|---|---|---|
| 1 | 9606 | 26.7 ± 13.2 | 54.1 ± 22.5 | 0.37 | 1.49 |
| 2 | 4360 | 40.9 ± 18.1 | 71.4 ± 17.9 | 0.52 | 1.69 |
| 3 | 1496 | 31.6 ± 16.4 | 43.3 ± 20.2 | 0.17 | 0.64 |
| 4 | 466 | 42.7 ± 18.9 | 61.5 ± 19.3 | 0.33 | 0.98 |
| 5 | 213 | 68.0 ± 19.9 | 88.5 ± 11.9 | 0.64 | 1.25 |

### 5.3.1 The networks

The community structure identified by MMA-P is shown for the pretest in Figure 5.1 and for the post-test in Figure 5.2. The figures shade like communities in multiple networks with the same color. Only communities identified in three of the four largest post-test samples are shaded. Various combinations of 4E*, 15A*, and 28E* were found in either the pretest or post-test networks; these have also been shaded. The asterisk indicates that the response is the correct response. Items 4, 15, and 28 require Newton's 3rd law for their solution. For the pretest in Figure 5.1, the communities have been colored consistently with the post-test in Figure 5.2 to allow comparison. Shaded communities that were not found in at least three of the four largest samples on the pretest have been outlined in red.

The communities that appear in at least three out of eight pretest or post-test networks of the four largest samples are summarized in Table 5.3. Only a subset of all communities are presented to highlight structures that were common across many institutions and to

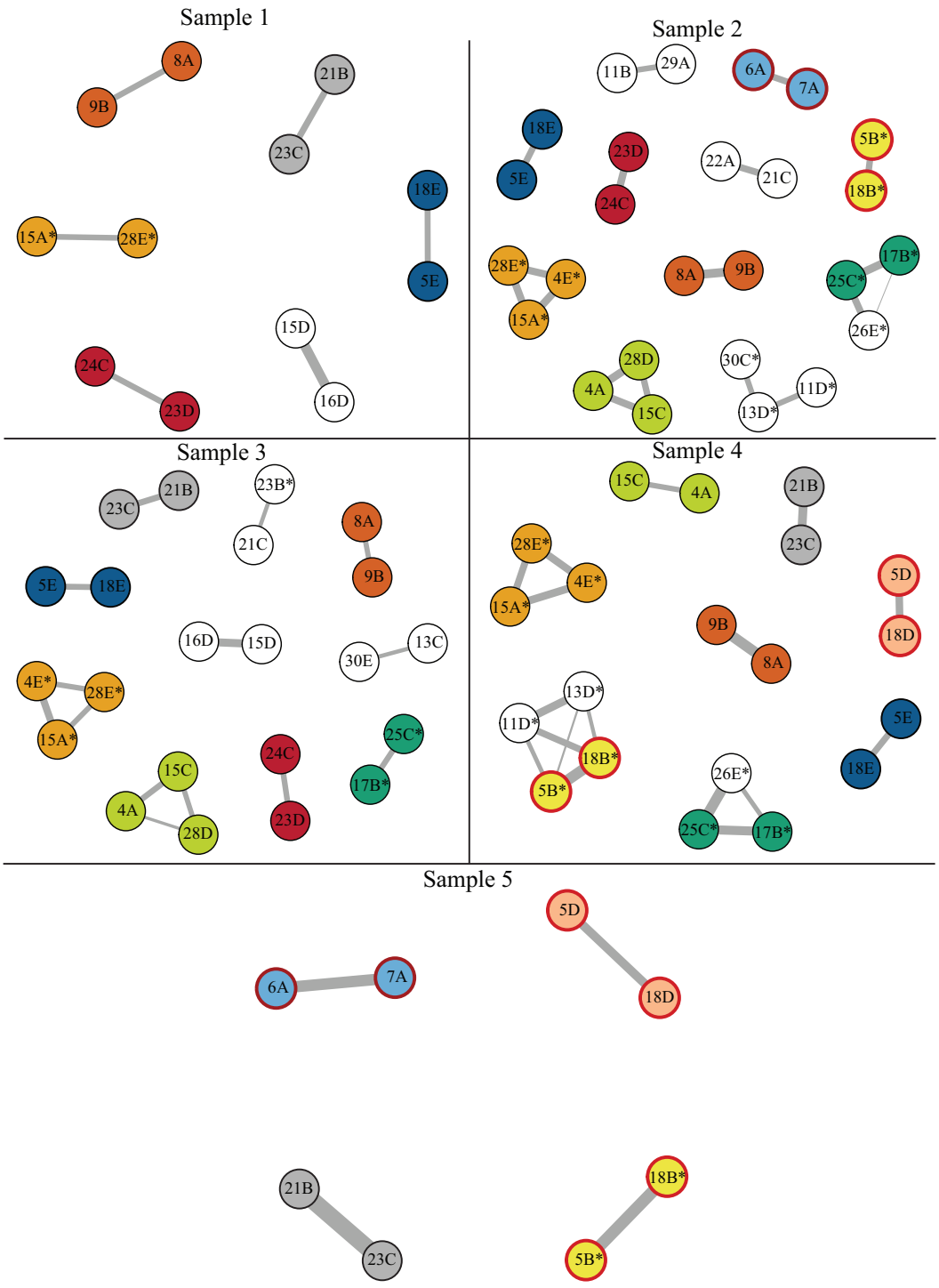


Figure 5.1: Pretest networks. Communities are shaded consistently with the post-test networks to allow comparison. Shaded communities not found in at least three of the four largest pretest samples are outlined in red. Correct responses are marked with an asterisk. The size of the partial correlation between the responses is proportional to the edge width.

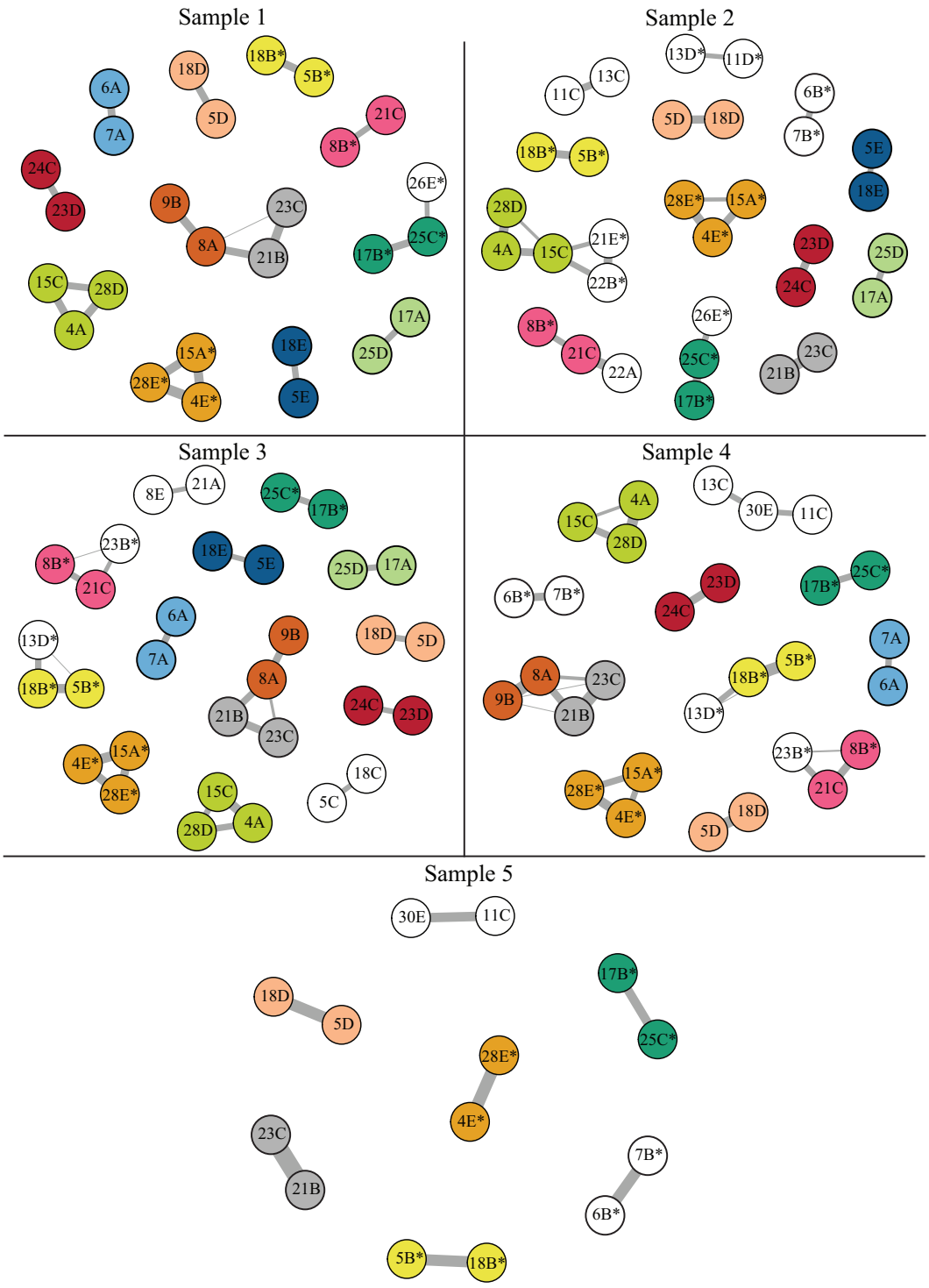


Figure 5.2: Post-test networks. Communities found in three of the four largest samples are shaded with the same color. Correct responses are marked with an asterisk. The size of the partial correlation between the responses is proportional to the edge width.

suppress communities that differ by a single edge. To partially capture the rich morphologies shown in the figures, completely connected communities are shown in parentheses separated by commas. A community is completely connected when each node in the community is connected by an edge to every other node in the community. A node that is only connected to one other node is indicated by a dash. For example, the Sample 1 post-test community 9B-(8A, 21B, 23C) contains a completely connected subgroup (8A, 21B, 23C) and one node, 9B, that is only connected to node 8A. Communities containing only two nodes must be completely connected and, therefore, the communities 8A-9B and (8A, 9B) are equivalent.

Table 5.3: Communities of FCI responses identified in at least 3 out of 8 pretest or post-test networks of the four largest samples. Cells with the label × are sub-communities of a larger community or are found with a different edge structure, while cells labeled ⊗ are explicitly found in the network. Sample 1 is abbreviated as S1, Sample 2 S2, etc. Responses that are separated by dashes are connected to each other, but not to other responses in the community. Responses that are in parenthesis are completely connected.

| Community | Pretest | | | | | Post-test | | | | | Explanation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 | S1 | S2 | S3 | S4 | S5 | |
| Completely Incorrect Communities | | | | | | | | | | | |
| 4A - 15C | | × | × | ⊗ | | × | × | × | × | | Newton's 3rd law misconceptions. |
| (4A, 15C, 28D) | | ⊗ | ⊗ | | | ⊗ | × | ⊗ | ⊗ | | Newton's 3rd law misconceptions. |
| 5D - 18D | | | | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | Motion implies active forces. |
| 5E - 18E | ⊗ | ⊗ | ⊗ | ⊗ | | ⊗ | ⊗ | ⊗ | | | Motion implies active forces: centrifugal force. |
| 6A - 7A | | ⊗ | | | ⊗ | ⊗ | | ⊗ | ⊗ | | Circular impetus. |
| 8A - 9B | ⊗ | ⊗ | ⊗ | ⊗ | | × | | × | × | | Blocked items: Last force to act determines motion. |
| 9B - (8A, 21B, 23C) | | | | | | ⊗ | | ⊗ | × | | 8A-9B: Blocked items. 21B-23C: Blocked Items. Both: Last force to act determines motion. |
| 17A - 25D | | | | | | ⊗ | ⊗ | ⊗ | | | Largest force determines motion. |
| 21B - 23C | ⊗ | | ⊗ | ⊗ | ⊗ | × | ⊗ | × | × | ⊗ | Blocked items: Last force to act determines motion. |
| 23D - 24C | ⊗ | ⊗ | ⊗ | | | ⊗ | ⊗ | ⊗ | ⊗ | | Impetus dissipation. |
| Mixed Correct and Incorrect Communities | | | | | | | | | | | |
| 8B* - 21C | | | | | | ⊗ | × | × | × | | 8B* and 21C share a similar trajectory. |
| 21C - 23B* | | | ⊗ | | | | | × | × | | Blocked Items: 21C and 23B* share a similar trajectory. |
| Completely Correct Communities | | | | | | | | | | | |
| 4E* - 28E* | | × | × | × | | × | × | × | × | ⊗ | Newton's 3rd law. |
| 15A* - 28E* | ⊗ | × | × | × | | × | × | × | × | | Newton's 3rd law. |
| (4E*, 15A*, 28E*) | | ⊗ | ⊗ | ⊗ | | ⊗ | ⊗ | ⊗ | ⊗ | | Newton's 3rd law. |
| 5B* - 18B* | | ⊗ | | × | ⊗ | ⊗ | ⊗ | × | × | ⊗ | Centripetal acceleration in a curved trajectory. |
| (5B*, 13D*, 18B*) | | | | × | | | | ⊗ | ⊗ | | Motion under gravity; a force in the direction of motion is not necessary. |
| 11D* - 13D* | | × | | × | | | ⊗ | | | | Motion under gravity; a force in the direction of motion is not necessary. |
| 17B* - 25C* | | × | ⊗ | × | | × | × | ⊗ | ⊗ | ⊗ | Newton's 1st law; Addition of forces. |
| 17B* - 25C* - 26E* | | × | | × | | ⊗ | ⊗ | | | | Newton's 1st and 2nd law; Addition of forces; (26E*) 1D acceleration. |

Some communities appear as independent communities not connected to other nodes in

some samples and as subgroups of larger communities in other samples. Some communities also share the same nodes but have different edges in different samples. A community is marked with an $\times$ to indicate it is also contained in a larger community or that it is also found with an alternate edge structure. For example, the community formed of nodes 8A, 9B, 21B, and 23C is found with two different structures. In the Sample 4 post-test, the community is completely connected. In the Sample 1 and Sample 3 post-test, the edges connecting 21B and 23C with 9B are missing. It is also found as two distinct communities in the Sample 1, 3, and 4 pretest as 8A-9B and 21B-23C.

Table 5.3 also includes a descriptive phrase explaining either the misconception or correct reasoning principle represented by the community. For incorrect communities, these were drawn from the taxonomy of Jackson and Hestenes [61] while incorporating changes to this taxonomy suggested in the original MMA paper [56]. Correct answers are classified using the detailed model of the FCI constructed by Stewart *et al.* [94]. The original model proposed with the publication of the FCI [9] divided the items into six broad categories. The model by Stewart *et al.* classifies each item by the set of reasoning principles needed to solve the item produces a much more detailed model of each item.

A table of all communities that appear in either the pretest or post-test networks is presented in Chapter 7. On the post-test, the communities identified in the networks of only one or two institutions differ from those in Table 5.3 by the addition or subtraction of a single edge. The communities on the pretest found only at 1 or 2 institutions were generally communities formed of only two nodes.

Figure 5.2 indicates a strong similarity between student responses post-instruction with most communities identified in three of the four largest samples. All 12 shaded communities

were identified in Samples 1 and 3; Sample 2 is missing 8A-9B and 6A-7A while Sample 4 is missing 5E-18E and 17A-25D. There was also substantial consistency in those nodes identified in fewer than three of the four samples. The combination 6B*-7B* was identified in two of the four samples. Different combinations of responses to items 11, 13, and 30 were sporadically identified; these items involve the identification of the forces acting on an object in motion. Blocked responses 26E* and 23B* were also sometimes found attached to other responses in their item block. As such, the consistency of completely correct, completely incorrect, and mixed communities was striking at these very different institutions.

The communities identified post-instruction in three of the four samples include all communities identified in three of the four largest pre-instruction samples; however, generally less community structure was identified in the pretest networks. Samples 1 to 4 contain only 5 to 8 of the 12 consistently identified (shaded) post-test communities. The structure that was identified was also less consistent between the 4 largest pretest samples. Figure 5.1 indicates shaded communities not found in at least three of the four pretest samples by outlining the nodes in red. The pretest networks contain all consistently identified completely correct communities identified in the post-test. Communities 17B*-25C* and (4E*, 15A*, 28E*) were identified in three of the four largest pretest samples; community 5B*-18B* was only identified in two of the four largest samples.

Interestingly, the post-test networks also contained completely incorrect communities not consistently found in the pretest: 6A-7A, 5D-18D, and 17A-25D. As the students correct thinking improved, those still answering consistently incorrectly were those applying a consistent misconception. The pretest also contained no mixed correct and incorrect communities while 8B*-21C was consistently identified in all four largest post-test samples.

Communities formed of incorrect responses to items requiring Newton's 3rd law for their correct solution (items 4, 15, 16, and 28) are categorized as "Newton's 3rd law misconceptions." These responses apply either the "greater mass implies greater force" or the "most active agent produces greatest force" misconceptions from Hestenes and Jackson's taxonomy [61]. The Newton's 3rd law items do not allow these misconceptions to be disentangled. This may explain the mixing of items with different combinations of Newton's 3rd law items shown in Table 5.3.

Most communities identified and described in Table 5.3 have been identified previously in the FCI by either Wells *et al.* [56, 57] or Yang *et al.* [58]; communities 21C-23B* and 5B*-18B*-13D* had not been reported in prior studies. These will be discussed with the mixed and completely correct communities.

To understand the communities identified by MMA-P, a detailed understanding of the structure of concepts measured by the FCI is needed. Stewart *et al.* identified four groups of isomorphic items [94]: {4, 15, 16, 28}, {5, 18}, {6, 7}, and {17, 25}. Isomorphic items can all be answered correctly by the same reasoning process. The FCI also contains 5 item blocks: {5, 6}, {8, 9, 10, 11}, {15, 16}, {21, 22, 23, 24}, and {25, 26, 27}. The blocking of items can produce correlations between items not related to the physical principles tested by the items and make the items difficult to interpret statistically [56]. For example, the correlations between items in an item block may be generated by the consistent misinterpretation of the item stem; thus producing a nested structure for the item correlations.

The completely incorrect communities are often formed by incorrect responses to isomorphic items. In general, when the same correct reasoning process is needed to solve two items, the misconceptions related to those items are also similar. The two-node commu-

nities not formed of responses to isomorphic items (21B-23C and 23D-24C) are both part of item blocks and both responses in each community share the same misconception based on Hestenes and Jackson's taxonomy [61]. It is not possible to separate the contribution of the blocked structure of the FCI from the effect of holding the "last force to act determines motion" misconception on students' selection of these responses together.

The only completely incorrect community with four nodes combines the communities from two different sets of blocked items: 8A-9B and 21B-23C. All four responses share the last force to act determines motion misconception [61]. Items 8 and 9 are blocked and ask the students about the trajectory and velocity of a hockey puck after it is struck at a right angle to its direction of motion. Items 21 and 23 are also blocked and involve the trajectory of a rocket; in item 21 the rocket experiences a thrust at a right angle to its trajectory; in item 23 the rocket continues after the thrust is removed. Responses 8A, 21B, and 23C present straight trajectories at right angles to initial direction of motion.

One community which mixes correct and incorrect responses was identified in each of the four largest post-test samples, 8B*-21C. Responses 8B* and 21C both present the students with straight line trajectories: this trajectory is correct for item 8 and incorrect for item 21. One mixed correct and incorrect community, 21C-23B*, appears in two post-test networks and one pretest network. Items 21 and 23 are part of an item block which asks about a rocket drifting in space which then fires its engine; the responses 21C and 23B* present the same trajectory, a diagonal line. This trajectory is correct for item 23 and incorrect for item 21. These two communities may show that the selection of the correct responses 8B* and 23B* does not indicate an understanding of the underlying mechanics concepts.

The completely correct communities were generally composed of responses to isomorphic items. The identification of these communities by MMA-P suggests that these correct responses are being selected together more often that one would predict based on the overall instrument score.

Some completely correct communities were not formed solely of isomorphic items. The community 11D*-13D* is formed of two items asking about the forces on an object moving under gravity: item 11 asks about a hockey puck sliding along a frictionless surface and item 13 about an object thrown directly upward. Both items have correct answers that gravity is one force acting on the object and both present the students with incorrect responses indicating a force in the direction of motion. In community 17B*-25C*-26E*, the isomorphic item pair 17 and 25 is joined by item 26; this item only has an edge with item 25. This community is found in two post-test networks and one pretest network. Items 25 and 26 are part of an item block which may explain the correlation. The community (5B*,18B*,13D*) was found in one pretest and one post-test network while 5B*-18B*-13D* was found in one post-test network. Item 13 asks about the forces on a ball thrown vertically in the air and has the correct response that only the force of gravity acts on the ball. Items 5 and 18 are isomorphic and ask about the forces acting on an object traveling in a curved trajectory. A downward force of gravity is one of the correct forces for both items. The three items may be selected together because of a correct understanding of the force of gravity. All three items have incorrect responses which posit a force in the direction of motion; the responses may also be selected together because the student does not hold the force in the direction of motion misconception.

Sample 5 shows the kind of information that can be extracted using MMA-P for smaller

samples. Both the Sample 5 pretest and post-test networks were smaller than the other samples likely because the lower statistical power prevented the resolution of more detailed structure. These networks did contain consistently selected correct and incorrect response identified in other networks suggesting that while not all structure may be resolvable at this sample size, the structure that is resolved is reliable. We note that some of failure to identify more structure may result from the very high general performance of this sample on the FCI.

### 5.3.2 Partial correlation threshold

The partial correlation threshold for each network was selected by plotting the average community size (ACS) against the number of communities (NC). The average community size is the total number of nodes in the network divided by the number of communities. An example ACS vs. NC graph for the Sample 1 post-test network is shown in Figure 5.3. Each point is calculated at a different $r$ threshold value. The plot changes slope quickly near the point with $r = 0.21$ which was used as the threshold for calculating the Sample 1 communities in Figure 5.2. Plots for other networks are included in Chapter 7.

The correlation thresholds selected using this method for the pretest and post-test are shown in Table 5.4. The Sample 5 network was independent of $r$ and, therefore, no threshold was required for this sample. As shown in Chapter 7, this behavior is the result of the sample size making the resolution of correlations below 0.35 unlikely.

Table 5.4: Partial correlation threshold coefficients used for each sample.

| Pre/Post | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Pretest | 0.20 | 0.20 | 0.16 | 0.21 | N/A |
| Post-test | 0.21 | 0.20 | 0.17 | 0.22 | N/A |

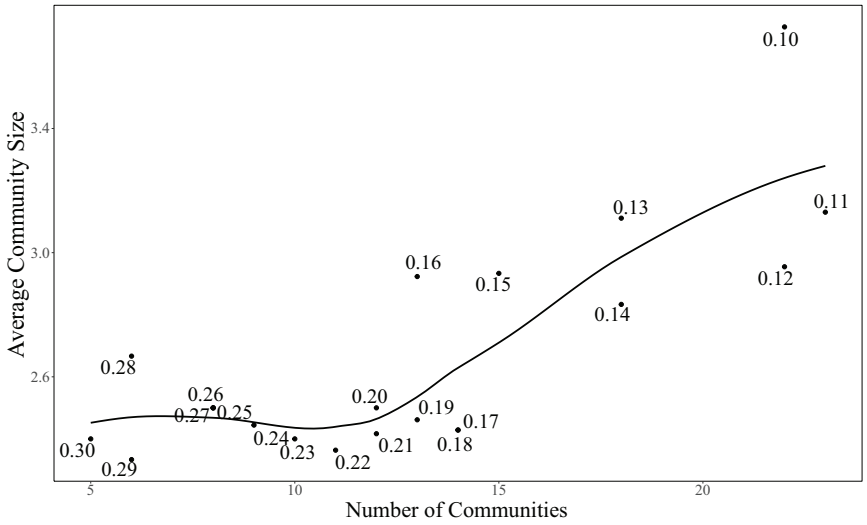


Figure 5.3: Plot used to determined the correlation threshold $r$ for the Sample 1 post-test network. Each point represents a network calculated at the labeled $r$ value.

### 5.3.3 Layer comparison results

A wealth of network comparison metrics have been developed for multiplex networks. For this work, we use the coverage index (CI) of the actors (nodes), edges, and triangles (completely connected 3 node sub-networks) to quantitatively characterize network similarity. Plots of these quantities are shown in Figure 5.4. These plots make use of pie charts where a completely filled circle represents an index of 1, an empty cell represents an index of 0, and a half filled circle an index of 0.5. For samples $i$ and $j$ with $i < j$, the plot below the diagonal represents $\mathrm{CI}_i = N(X_i \cap X_j)/N(X_i)$ and the plot above the diagonal $\mathrm{CI}_j = N(X_i \cap X_j)/N(X_j)$ where $X$ is the set of actors, edges, or triangles. For example, consider the plot of the pretest coverage edges of Sample $i = 1$ and Sample $j = 2$, the circle below the diagonal plots $\mathrm{CI}_1 = N(X_1 \cap X_2)/N(X_1)$ the fraction of the total number of edges in Sample 1 that are also in Sample 2. Approximately 67% of the edges in Sample 1 are also in Sample 2. The circle above the diagonal plots $\mathrm{CI}_2 = N(X_1 \cap X_2)/N(X_2)$, the fraction of

edges in Sample 2 that are also in Sample 1. The circle is 22% full; therefore, 22% of the edges in Sample 2 are also in Sample 1.

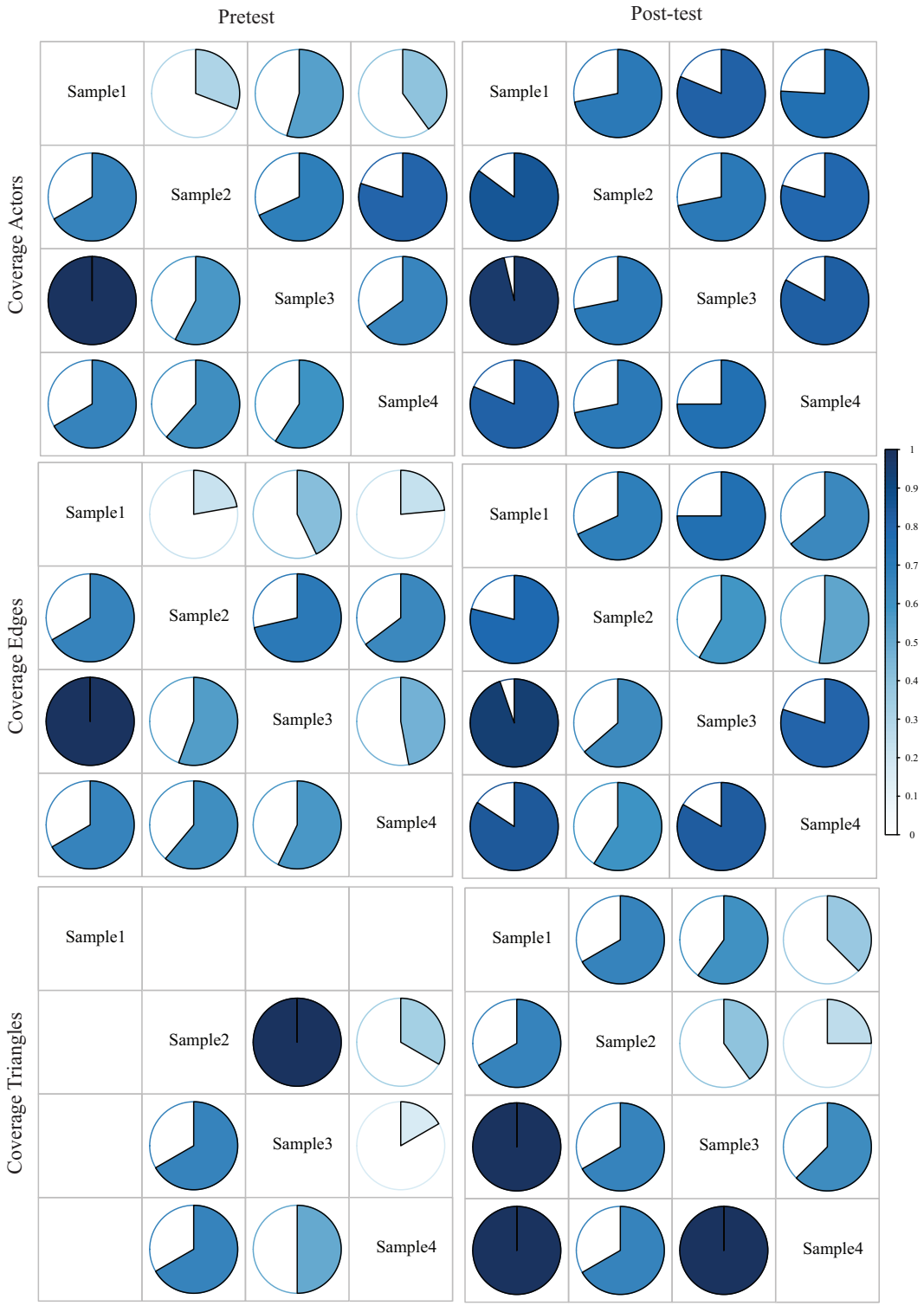


Figure 5.4: Coverage actors, edges, and triangles between samples.

Communities composed of two and three nodes form the majority of communities identified in all networks; as such, coverage edges and triangles are natural structures to investigate to characterize similarity. Figure 5.4 shows substantial similarity between networks for Samples 1 to 4 in actors, edges, and triangles on the post-test. The plots also illustrate the lower similarity of the pretest networks compared to the post-test networks. There are no triangles in the Sample 1 pretest.

The CI allows the quantitative exploration of the change in similarity between the networks from the pretest to the post-test. The average CI, $\langle CI \rangle$, of the four largest samples shows an increase in similarity from pretest to post-test (pretest $\langle CI \rangle = 0.63$ actors, 0.58 edges, and 0.28 triangles; post-test $\langle CI \rangle = 0.79$ actors, 0.72 edges, and 0.66 triangles). As such, on average, half of the actors and edges found in the pretest network of one sample are also found in the pretest network of another sample. These averages grow to 79% and 72% on the post-test indicating the structure of consistently selected responses is greater on the post-test. This is to be expected as physics instruction serves to even out differences in incoming student preparation. Only Samples 2, 3, and 4 contain triangles in both the pretest and post-test networks. Averaging the CI for these samples only shows the triangles change little from pretest to post-test ($\langle CI \rangle_{pre} = 0.56$, $\langle CI \rangle_{post} = 0.60$). This stability is partially the result of correct and incorrect responses to Newton's 3rd law items forming the majority of the triangles.

### 5.3.4 Misconception scores

Wells *et al.* [56] used the consistently selected incorrect responses identified by module analysis to define a misconception score which quantitatively captures the average fraction

of misconceptions of each type selected by a student. Misconception scores measure the frequency of applying different misconceptions and should be related to how strongly they are held. Misconception scores represent the number of responses chosen that are associated with a misconception out of the total number of item responses that a student could possibly choose associated with the same misconception. Table 5.5 presents the misconception scores for completely incorrect communities found in most post-test networks in Figures 5.1 and 5.2. The misconceptions are classified using the modifications proposed in Wells *et al.* [56] to the Hestenes and Jackson taxonomy [61]. Misconception scores are explained in greater detail in Chapter 7.

Table 5.5: Percentage of students selecting each incorrect response associated with a misconception for the FCI post-test.

| | | Misconception Scores | | | | |
|---|---|---|---|---|---|---|
| Misconception | Responses | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
| Largest force determines motion | 17A,25D | 40.0% | 38.3% | 51.3% | 45.9% | 13.9% |
| Newton's 3rd law misconceptions | 4A,15C,28D | 31.9% | 33.5% | 42.7% | 46.4% | 8.6% |
| Motion implies active forces | 5D,11C,13C,18D,30E | 31.4% | 20.7% | 39.3% | 33.6% | 10.6% |
| Circular impetus | 6A,7A | 12.9% | 6.7% | 15.2% | 8.7% | 2.8% |
| Motion implies active forces; centrifugal force | 5E,18E | 16.9% | 7.0% | 19.3% | 11.5% | 0.5% |
| Impetus dissipation | 23D,24C | 18.5% | 8.5% | 17.4% | 14.9% | 4.5% |
| Last force to act determines motion | 8A,9B,21B,23C | 21.4% | 6.4% | 21.3% | 13.6% | 4.7% |

The misconception score can be converted into the average number of misconception responses of each type selected by students by multiplying the score by the number of responses in the group. For example, the Newton's 3rd law misconception group contains 3 responses, the 31.9% misconception score for Sample 1 indicates that on average students in this sample select $3 \cdot 0.319 = 0.96$ of these responses from each application of the FCI. That is, even post-instruction, students are on average answering one Newton's 3rd law item incorrectly using a common misconception.

The largest force determines motion, Newton's 3rd law, and motion implies active forces

misconceptions consistently have the highest score across all five institutions. The average misconception scores for the four largest institutions are 44%, 39%, and 31% respectively. These misconceptions are also the most commonly selected, but at a lower rate, by the very highly performing Sample 5. Students on average select responses related to these misconceptions $2 \cdot 0.44 = 0.88$, $3 \cdot 0.39 = 1.2$, and $5 \cdot 0.31 = 1.6$ times each time the FCI is given. These misconceptions are likely some of the most widely held and consistently applied in mechanics and remain post-instruction at institutions with a broad spectrum of student populations. The rest of the communities vary greatly between about 1% and 20%. Misconception score is highly negatively correlated with post-test score, which explains why Sample 3 consistently has the highest scores (more students selecting responses indicating misconceptions) and Sample 5 consistently has the lowest scores for each community. Note, while Sample 5 had insufficient statistical power to fully resolve its network structure, this should not restrict the validity of its misconception scores.

## 5.4 Discussion

This work posed three research questions; they will be explored in what follows. Many of the findings were discussed in the prior section; this section will provide a summary.

*RQ1: How does the community structure identified through module analysis of the FCI compare across multiple institutions? What does this community structure imply about student understanding of mechanics?* Across four U.S. institutions with a range of ACT, pretest, and post-test scores as well as demographically different undergraduate populations, the community structure in the pretest and the post-test was very similar. As Table 5.3

shows, misconceptions related to Newton's 3rd law, circular impetus, impetus dissipation, motion implies active forces, last force to act determines motion, and motion implies active forces - centrifugal force appear in most of the post-test networks and many of the pretest networks. A large majority of the communities for both the pretest and post-test are found in at least three samples; the majority of post-test communities in four samples. These results imply that misconceptions measured by the FCI are coherently applied at a broad spectrum of U.S. institutions both pre-instruction and post-instruction. The misconception scores presented in Table 5.5 suggest the largest force determines motion, Newton's 3rd law, and motion implies active forces misconceptions are the most prevalent post-instruction.

The incorrect response communities corresponding to misconceptions are also largely found in Brewe *et al.*'s original module analysis work, which was applied to 143 FCI responses from first-year physics majors in Denmark. The modules in that work were much larger and were interpreted somewhat differently, but the following incorrect response communities identified in the current work were each found in one of their modules as well: 17A-25D related to the largest force determines motion misconception, (4A-15C-28D) related to Newton's third law misconceptions, 9B-(8A-21B-23C) related to the last force to act determines motion misconception, and 5D-18D the related to motion implies active forces misconception.

Community 5D-18D, corresponding to the motion implies active forces misconception, community (4A, 15C, 28D), corresponding the Newton's 3rd law misconceptions, and community 17A-25D, corresponding to the largest force determines motion misconception, stood out as particularly problematic post-instruction. All were identified in three of the four largest samples post-instruction and had the three highest misconception scores indicating

the misconceptions were still frequently applied post-instruction across a broad spectrum on institutions. Communities representing the circular impetus, last force determines motion, impetus dissipation, and motion implies active forces - centrifugal force were also identified in three of the four largest samples; however, these responses had generally lower misconceptions scores indicating they are applied less frequently post-instruction. Many of the incorrect communities identified post-instruction with high misconception scores were identified pre-instruction much less consistently. This may be because many students answer incorrectly pre-instruction because they have little knowledge of the correct physics and thus are not consistent, but students with consistently applied misconceptions retain these post-instruction.

The similarity of the community structure across the institutions studied suggests that the sets of consistently applied misconceptions present pre-instruction and remaining post-instruction may be very consistent across many institutions. Misconception scores suggest many of these misconceptions are still selected by many students post-instruction at all but the most highly performing institutions. This observation identifies a group of misconceptions which may be the most important to target to improve student understanding of Newtonian physics.

Both the completely correct and completely incorrect community structure was primarily related to groups of isomorphic and blocked items. The isomorphic item communities show that there are groups of items testing the same concept and generally the same misconception which are answered together more often than one would predict based on total instrument score; this indicates the FCI measures some more fine grained structure beyond a single Newtonian force concept. This is consistent with factor analysis work showing that

between 5 and 9 factors are optimal [94, 121–124]. The practice of the blocking of items continues to make correlations found in these samples difficult to reliably identify as consistently applied misconceptions.

The two mixed correct and incorrect communities are of particular interest. For both communities, the student is selecting responses representing qualitatively similar straight line trajectories. In both cases, the student selecting the same trajectory for both correct and incorrect responses may indicate the item being answered correctly is not functioning properly.

The communities identified in Sample 5 were substantially different than all other samples both pre-instruction and post-instruction. Far fewer communities were identified than in the other four samples which was likely the result of the lower statistical power requiring larger correlations for statistical significance. The smaller networks contained both completely correct and completely incorrect communities identified in the other samples. It seems quite likely that, with a larger dataset, the Sample 5 community structure would resemble that of other institutions, but additional research would be needed to establish this. The misconception scores of Sample 5 were dramatically lower than those of all four other samples suggesting that even if the networks were similar at higher sample size, many fewer students were left consistently applying common misconceptions in the classes from which Sample 5 was drawn.

*RQ2: How can the parameters required by module analysis be selected quantitatively?* This work proposed a new quantitative method to select the correlation threshold $r$; the correlation at which edges in the network are retained. In past module analyses, $r > 0.2$ was most commonly chosen as the correlation threshold [56–58, 93]. In some works, $r > 0.2$

yielded a network far too sparsified and $r > 0.15$ was chosen instead [93]. These values were chosen by examining the networks at multiple $r$ thresholds and qualitatively determining a threshold by choosing a network that had theoretically explainable structure while minimizing $r$.

To partially eliminate the uncertainty of this method, a number of quantitative approaches for choosing $r$ were explored with the goal of yielding similar results to the qualitative approach. The global clustering coefficient, the number of triangles divided by the number of triples in a graph [138], and other local transitivity measures within the graph were examined. A triple is a set of three nodes that are not fully connected; differing from a triangle by a single edge. These were not productive because of the low number of triangles in the networks. Graphing the average community size (ACS) against the number of communities (NC) yielded the most promising results out of the metrics tested. Both the ACS and the NC are calculable for small networks such as those identified by MMA-P for the FCI. For more complex networks other metrics may be more appropriate.

*RQ3: What quantitative network analytic methods are productive for characterizing institutional differences and similarities identified by module analysis? What do these measures imply about the student understanding of mechanics?* The coverage index for actors, edges, and triangles proved to be a useful metric for comparing institutional differences and similarities. Figure 5.4 shows the coverage index for both the pretest and the post-test for the four largest samples. The coverage indices identified substantial similarity in Samples 1 to 4 in the actors, edges, and triangles identified in the post-test. This is consistent with the fairly uniform number of communities identified, from 11 to 13 communities. The pretest networks were smaller and more variable with from 6 to 11 communities in the four largest

samples. This variability was captured by the coverage index. For the pretest, the Sample 2 to 4 networks, while less consistent than the post-test, were often not substantially less consistent. The Sample 1 pretest network was qualitatively different with fewer communities than the other large samples; this difference was clearly shown in the coverage index plots of the pretest (the first row).

The coverage index allowed the change from pretest to post-test to be quantitatively characterized with average coverage index of $\langle CI \rangle = 0.63$ for actors and 0.58 for edges on the pretest which increased to $\langle CI \rangle = 0.79$ for actors, 0.72 for edges on the post-test, an increase but not an overwhelming increase. Many other network comparison metrics are available for multiplex networks and may be useful in future research.

The average value of the CI for the actors and edges over all samples showed the similarity of the networks increased from pretest to post-test. As such, both the consistently selected correct responses and consistently applied misconceptions became more similar across four institutions with very different undergraduate populations. This indicates both correct knowledge that can be applied in multiple contexts and incorrect knowledge that is consistently applied in multiple contexts is fairly similar across U.S. institutions with very different undergraduate populations, FCI pretest scores, and FCI post-test scores.

The misconception scores show that students are on average selecting about one response indicting the application of the largest force determines motion, Newton's 3rd law, or motion implies active forces misconception post-instruction each time they take the FCI. The rate of consistently applying these misconceptions was much lower at the highest performing institution.

## 5.5 Implications

Module analysis was successful in identifying the same communities of consistently selected correct and incorrect responses within the FCI across a wide variety of institutions. This suggests that consistently applied Newtonian misconceptions exist prior to and after instruction in college physics classes that span the spectrum of incoming student preparation. These misconceptions persist post-instruction, despite each sample having an improvement in FCI scores of medium or large effect size from pretest to post-test. The primary misconceptions held by a substantial number of students post-instruction were misconceptions related to Newton's 3rd law, largest force determines motion, and motion implies active forces. It might be productive to focus on this group of misconceptions out of the broad catalog of FCI misconceptions tabulated by Hestenes and Jackson [61] for targeted instructional interventions.

The FCI contained a number of completely correct communities formed of isomorphic items. These items are selected together more than would be predicted based on the overall instrument score. This suggests that, if additional items measuring these concepts were developed, it might allow the measurement of sub-dimensions of these Newtonian force concepts. This would provide instructors with a more fine-grained measurement of student knowledge.

## 5.6 Future Work

Module analysis has been productively applied to the FCI, FMCE, and CSEM. These instruments are traditionally scored where each item has a single correct response. Module

analysis should also be productive for instruments with more complex scoring rules. For example, an instrument where students could select multiple responses to a single item. It might also be productive for more complex instrument structures such as contingent items where an item is only presented to the student if some response to a prior item is selected. Module analysis should also be able to be extended to Likert scale survey items and may provide additional insight into the relations of non-cognitive constructs such as self-efficacy, belonging, and identity.

The current study is part of a long history of quantitative studies of now venerable conceptual physics instruments. This work has accelerated in recent years with many new quantitative methods applied. It seems likely that this burst of quantitative research effort is nearing the limit of new findings which can be teased from these instruments. This research has an important secondary effect which may ultimately be more important than the findings of the studies themselves. These research efforts have lead to the identification of structural issues within the instruments including a lack of factor structure [121], items which would be in the range of problematic item functioning in Classical Test Theory (CTT) [139], and the effects of the practice of blocking or chaining items [94]. Beyond these, substantial issues of item fairness for some demographic groups have been identified in some of the instruments [139]. The growing list of concerns makes it imperative that a new generation of conceptual instruments be constructed and validated in the near future to allow our understanding of physics instruction to continue to improve and to provide insights that help all students. The quantitative methods used in recent studies establish a set of expectations that these new instruments will be expected to meet before broad deployment should considered. The new instruments should have a reproducible factor structure, have items that are well functioning

in CTT, not use item chaining or blocking, and have items that pass a quantitative fairness test for groups of students underrepresented in physics classes. Module analysis adds to these criteria by implying any new instruments should have community structures which are theoretically supportable and should be constructed to allow the calculation of misconception scores for the misconceptions most commonly applied in the topic covered.

## 5.7 Conclusion

The FCI was constructed under the misconception framework with the goal of measuring students' conceptual understanding of Newtonian mechanics. This study compared the structure of consistently applied student misconceptions to responses to the FCI across five institutions with student populations with differing levels of high school preparation using MMA-P. The networks identified had substantial similarity for four largest samples in both communities formed of correct responses and of communities associated with misconceptions. The study concluded that the smallest sample had insufficient statistical power to fully resolve the network structure. The cross-institutional similarities found in this work could motivate the application of module analysis to other multi-institutional datasets to investigate the similarity of the community structure of other conceptual instruments.

The largest force determines motion, Newton's 3rd law, and motion implies active forces misconceptions consistently had the highest misconception scores across all five institutions. On average, students select a response applying each of these misconceptions each time they complete the FCI showing they are a substantial part of student reasoning about mechanics at institutions with very different student populations and FCI outcomes. The

large number of students still applying misconceptions post-instruction supports a continued need to transition to research-based instructional methods and to continuously improve those methods.

# Chapter 6

## Applying Module Analysis to the Brief Electricity and Magnetism Assessment*

*This chapter presents the work published in Physical Review Physics Education Research [140]. This work was constructed with collaborative efforts from James Wells and John Stewart. We thank Steven Pollock for the collection and curation of this exceptional dataset and his helpful commentary.

## 6.1 Introduction

This study explored consistently selected patterns of student responses to the Brief Electricity and Magnetism Assessment (BEMA) with MMA-P [58].

This study investigated the following research questions:

**RQ1** What community structure is identified by network analysis of the BEMA? What underlying reasoning could explain these response patterns?

**RQ2** How does the community structure of the BEMA relate to the community structure of the CSEM?

### 6.1.1 The Brief Assessment of Electricity and Magnetism

The BEMA consists of 31 multiple-choice items that test students' understanding of electrostatics, magnetostatics, electric circuits, electric potential, and magnetic induction [141, 14]. Each item has between 3 and 10 possible responses; 23 items have at least 7 responses. Twenty-nine items include some combination of "none of the above" or "zero" responses, which may cause psychometric problems [119]. The BEMA is a highly blocked instrument, where the same physical system is used for multiple sequential items. Many studies have shown that blocking items may produce correlations which make the instrument difficult to interpret [56, 94, 101, 102]. The BEMA item blocks are: {1, 2, 3}, {4, 5}, {8, 9}, {14, 15, 16}, {21, 22}, {26, 27}, and {28, 29}. The BEMA also includes a number of semi-quantitative items where the responses are symbolic formulae.

A scoring rubric is provided for the BEMA that accounts for correct student reasoning beyond a traditional scoring rubric [11]. Item 3 is counted as correct if the response chosen

follows the inverse square law for Coulomb forces from the response to item 2, regardless of whether response 2 was correct. Item 16 is counted as correct if item 15 is answered correctly and the response to item 16 is identical to the response to item 14. Both items 28 and 29 must be correct to receive credit for a single item, resulting in a total instrument score of 30 instead of 31.

### 6.1.2 Prior studies of the BEMA

The BEMA has been used in multiple large studies to test students' conceptual knowledge of electricity and magnetism [142–144]. The reliability of the BEMA has been examined by various studies. Ding *et al.* tested individual items as well as the instrument as a whole [54] with five reliability statistics. The difficulty of items 9, 11, 12, 27, and 28 were all below the threshold of 0.3 for good item functioning in classical test theory (CTT) [145]; difficulty in CTT is the average score on the item. The discrimination of items 9, 11, 26, and 27 were well below the threshold of 0.3 for good item functioning; discrimination determines whether the item distinguishes between high performing and low performing students. However, the average value of each statistic for the entire instrument suggested that the BEMA reliably tests both higher and lower scoring students. Rasch theory was also used by Ding [146] to test the BEMA's construct validity which showed that the BEMA does measure a unidimensional construct.

The BEMA has also been used to compare gains in two different introductory undergraduate electricity and magnetism curricula [143] implemented at multiple institutions. The first curriculum used traditional electricity and magnetism instruction and similar textbooks; the second used the Matter and Interactions electricity and magnetism curriculum

[147]. At each of the four institutions studied, students following the Matter and Interactions curriculum consistently had higher BEMA post-test scores. In a separate study, Ding [148] compared BEMA responses from students taking two different electricity and magnetism courses at the same institution; one a traditional course, the other a Matter and Interactions course. Students in the Matter and Interactions course performed significantly higher on two items and significantly worse on three items. The study also showed that items 9 and 17 did not discriminate between high performing and low performing students effectively.

Item response theory was applied to student responses to the BEMA by Xiao *et al.* [149]. This study showed that items could be removed from the BEMA without reducing test validity or reliability and that the resulting concept inventory still measured the same latent electricity and magnetism constructs. Multidimensional item response theory (MIRT), exploratory factor analysis, and correlation analysis were also applied to nearly 10,000 student responses to the BEMA by Hansen and Stewart [150] to explore its structure and determine a model of student knowledge that it measures. Correlation analysis found that the majority of the instrument's substructure could be attributed to item blocks. A five-factor model optimized the model fit for the BEMA. A five-factor model was also reported by Eaton *et al.* [151].

### 6.1.3 The BEMA and The CSEM

The BEMA is one of two popular electricity and magnetism conceptual inventories often used in PER; the other is the CSEM. The CSEM has 32 items, each with 5 responses. It covers electrostatics, magnetostatics, electric potential, and magnetic induction. Unlike the BEMA, it has no coverage of electric circuits and has no items that ask the students to

select mathematical formulae in their responses.

The problem statement of items 1, 2, and 3 of the BEMA are identical to items 3, 4, and 5 of the CSEM; however, the responses are somewhat different. All CSEM items have 5 responses, while items 1 and 2 of the BEMA have 7 responses and item 3 has 9 responses. Items 30 and 31 of the BEMA are nearly identical to items 31 and 32 of the CSEM except that the BEMA items have 7 and 6 responses, respectively.

The BEMA and the CSEM were compared using item response theory and CTT by Eaton *et al.* [152]. They found that both tests had very similar difficulty with differences indistinguishable from the differences in the samples studied. Pollock [153] compared student performance on both instruments and found them to be roughly equivalent in measuring student learning with some differences in the content covered. In another study by Eaton *et al.* [151] using exploratory factor analysis, the CSEM was fit with a six-factor model and the BEMA with a five-factor model. Despite the difference in the number of factors, the two factor models showed substantial similarity in conceptual coverage between the two instruments. A recent study developed a method to compare student scores on the CSEM and the BEMA by linking and transforming assessment scales [149].

The models of student knowledge identified through MIRT of both the CSEM [102] and the BEMA [150] differed in both coverage and complexity. The best-fitting model of the CSEM required fewer logical principles than the BEMA. Much of the differences in coverage of the two instruments can be attributed to the use of electric circuit items in the BEMA and some differences in electrostatics coverage.

## 6.2 Methods

### 6.2.1 Sample

The sample for this study consists of data collected from Fall 2004 until Spring 2019 from a university in the U.S. with a total enrollment of about 35,000 students. The demographics of this institution in Spring 2019 were 67% White, 12% Hispanic, 6% Asian, 6% two or more races, 2% African American/Black, and less than 1% each Native Hawaiian or Pacific Islander and American Indian or Alaska Native [129]. International students comprised 7% of the students enrolled at this institution. The undergraduate population for this university had ACT scores ranging from 25–31 (25th to 75th percentile).

The BEMA was given in introductory calculus-based electricity and magnetism courses serving primarily scientists and engineers as a post-test. Only students with complete post-test responses were included in the study ($N = 12,214$).

### 6.2.2 Modified Module Analysis - Partial

Modified Module Analysis - Partial was applied to the BEMA; an introduction to the method is provided in Chapter 3. Chapter 5 introduced a method to select the correlation threshold and a change to the minimum number of required responses before removing an item response. We employ the same changes to MMA-P in this study. The appendix at the end of this chapter includes the plot that identified a partial correlation threshold of $r_{XY|S} > 0.17$ as optimal. Note, for the remainder of this chapter, we will shorten partial correlation to correlation and use $r$ as the partial correlation coefficient for brevity.

## 6.3 Results

### 6.3.1 The network

The average BEMA post-test score for the 12,214 students was $53.4\% \pm 17\%$ which increased to $55.7\% \pm 17\%$ after adjusting scores based on the suggested grading rubric. A table with the average score on each item, as well as the number of times each item response was selected, is included in the Appendix to this chapter. The difficulty (average score) of items 12, 27, 28, 29, and 31 were all below the 30% threshold for a well functioning item suggested in CTT [145].

Figure 6.1 shows the communities identified in the BEMA by MMA-P. Communities formed of responses from the same item block have been similarly colored. Correct response nodes are labeled with an asterisk (*). Nodes are connected by edges weighted by the partial correlation between item responses; larger partial correlations correspond to thicker lines. The network representation in Figure 6.1 was constructed using the Fruchterman-Reingold network layout algorithm [154] using the igraph [99] package in R. In this algorithm, Hooke's-law-like attractive forces are introduced between connected nodes, where the strength of attraction is proportional to edge weight. Repulsive Coulomb's-law-like forces are then introduced between all nodes. The nodes are randomly placed, then moved until the system's energy is minimized. The specific positions of nodes in the network convey no information; however, nodes more strongly related to each other are placed closer together. The partial correlations were calculated with a conventional grading rubric. When adjusted for the suggested scoring rubric, the resulting network differed by one community.

Each community that appears in Figure 6.1 is a part of an item block. Table 6.1

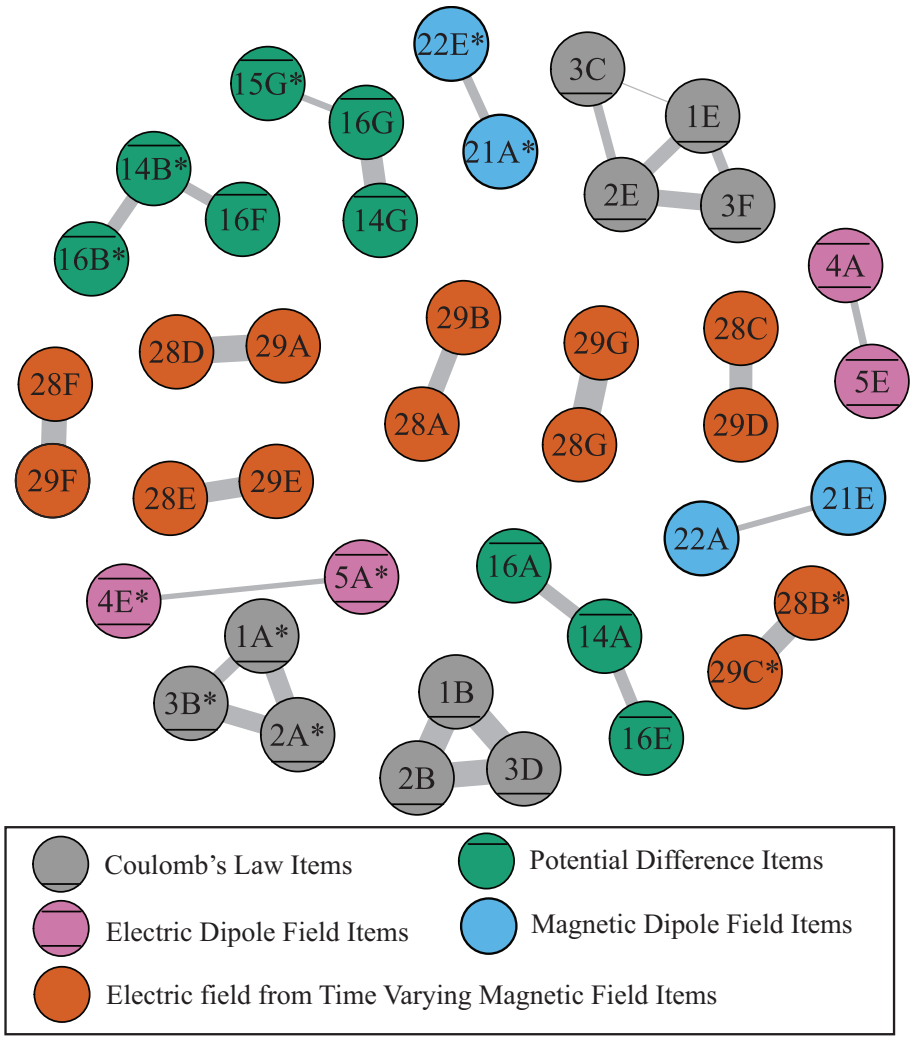


Figure 6.1: Post-test BEMA network. Communities formed of responses from the same item group have been similarly colored. Lines have been added to some nodes to help distinguish the nodes when viewed in grey scale. Correct responses are indicated by an asterisk (*). Thicker lines represents larger partial correlation between item responses.

provides a description of the possible common reasoning leading to each community. The first column in Table 6.1 uses a notation that distinguishes between partially and completely connected communities. A completely connected community has an edge between every node in the community. Item responses surrounded by parenthesis represent completely connected communities. Item responses separated by dashes (-) represent edges between those two responses. A community of size two is always completely connected. Communities

with multiple responses derived from the same item cannot be completely connected.

Table 6.1: Communities identified in BEMA responses. Responses in parentheses are completely connected. Responses separated by dashes are only connected to each other. Responses 1E, 2E, 3C, and 3F are part of a single community 3C-(1E,2E)-3F, which has been split between two lines.

| Community | Explanation |
|---|---|
| | **Coulomb's law item block** |
| (1B,2B,3D) | 1B-2B - Averaged forces on two charges. 3D - Consistently applies ($F \propto 1/r^2$). |
| (1E,2E,3C) | 1E-2E - Force not changed when charge changes. 3C - ($F \propto 1/r$). |
| (1E,2E,3F) | 1E-2E - Force not changed when charge changes. 3F - Consistently applies ($F \propto 1/r^2$). |
| (1A*,2A*,3B*) | Coulomb's law for the electric force. |
| | **Electric dipole field item block** |
| 4A-5E | Electric field lines point out of negative charges and into positive charges: Reversed electric dipole shape. |
| 4E*-5A* | Correct electric dipole field shape. |
| | **Potential difference item block** |
| 16E-14A-16A | Electric field lines point in the direction of increasing potential. 16E - Potential is path dependent. |
| 14G-16G-15G* | All zero responses – potential is not proportional to electric field. |
| 16F-14B*-16B* | Electric field lines point in the direction of decreasing potential. 16F - Potential is path dependent. |
| | **Magnetic dipole field item block** |
| 21E-22A | Reversed magnetic dipole field shape. |
| 21A*-22E* | Correct magnetic dipole field shape. |
| | **Electric field from time varying magnetic field item block** |
| 28A-29B | Solenoid as negative linear charge. |
| 28C-29D | Solenoid as positive linear charge. |
| 28D-29A | Electric field lines in opposite direction of the correct direction. |
| 28E-29E | Magnetic field of solenoid, opposite direction. |
| 28F-29F | Magnetic field of solenoid, correct direction. |
| 28G-29G | All zero response. Solenoid with changing current creates no electric field. |
| 28B*-29C* | Correct electric field from changing magnetic field. |

### Coulomb's law communities

The first item block {1, 2, 3} involves two small objects each with a net charge of $+Q$ which exert a force $F$ on each other. The charge on one of the objects is then changed to $+4Q$. Item 1 asks about the change to the force on the $+Q$ object; item 2 the change to the force on the $+4Q$ object. Item 3 asks about the resulting force on the $+4Q$ object if the the objects are moved 3 times farther apart. In the recommended grading rubric, item 3 is counted as correct if it is 1/9 the value reported for item 2. This relation is shown in Table 6.1 as "Consistently applied ($F \propto 1/r^2$)" when the answer is consistent with a $1/r^2$ distance dependence contingent on the response to item 2. For the communities identified, response 3D is 1/9 the value in response 2B and response 3F is 1/9 the value in response 2E.

The community (1B, 2B, 3D) represents students choosing the response $5F/2$ (responses 1B and 2B) for items 1 and 2 and the response consistent with $1/r^2$ for item 3. Students who choose 1B and 2B are reporting the average force on the two objects before and after the modification of the charge. The community 3C-(1E,2E)-3F includes two fully connected sub-communities of size 3 centered around 1E and 2E. This community has been split across two lines in the table as (1E, 2E, 3C) and (1E, 2E, 3F) for easier table readability. Both responses 1E and 2E indicate that the force remains $F$, that no change occurred when increasing one of the charges to $+4Q$. Response 3F is consistent with the inverse square law response, while response 3E is the response 1/3 the value in response 2E and represents a $1/r$ force. Both sub communities indicate a different possible outcome of making the mistake that charge is not proportional to the force between particles (1E,2E). The network also includes a completely correct community for this item block, (1A*, 2A*, 3B*).

Two separate types of incorrect reasoning are captured by these communities. The first involves how electric force scales with electric charge. One community represents the incorrect reasoning that force and charge are independent while the other community uses the average of the forces on the charge before and after the modification of one of the charges. The second type of reasoning involves how force scales with distance. Some communities use the correct $1/r^2$ scaling based on a prior incorrect answer while others use the incorrect $1/r$ distance dependence. This could indicate the students are failing to discriminate electric force from electric potential.

### Dipole field communities

Two item blocks ask the students about dipole fields: items {4, 5} and {21, 22}. Items 4 and 5 present the students with two charged objects with charge $\pm q$; the objects are spaced a small distance apart. Item 4 asks about the field along the axis of the dipole; item 5 at a point perpendicular to the axis of the dipole. Items 21 and 22 present the students with a bar magnet. The magnet's north and south poles are not labeled. The direction of the magnetic field is given by a vector at a point on the axis of the magnet. Item 21 asks for the direction of the magnetic field at a point along the axis of the dipole; item 22 at a point on the perpendicular bisector of the dipole. Both item blocks use the same set of responses. The network includes one incorrect community and one correct community for both item blocks. Both incorrect communities, 4A-5E and 21E-22A, contain responses where each are in the opposite direction to the correct response; the field is reversed. The correct community 4E*-5A* was barely above the partial correlation threshold, and when accounting for the suggested scoring rubric, this community falls below the threshold and is removed from the network.

### Potential difference communities

The potential difference item block {14, 15, 16} presents the students with a uniform electric field whose direction is indicated by a series of evenly spaced vectors. Four points labeled 1 to 4 are placed in the field in a rectangle; the width $w$ and a height $h$ of the rectangle is indicated on the figure. Item 14 asks students for the potential difference between two points across the width of the rectangle; the electric field points from the first point to the

second point. The shortest path between the points lies in the direction of the field. Item 15 asks for the potential difference between two points across the height of the rectangle; the shortest path between these points is perpendicular to the field. Item 16 asks for the potential difference across the diagonal of the rectangle.

The network includes one completely incorrect community and two mixed correct and incorrect communities. The completely incorrect community 16E-14A-16A includes all positive responses, representing the incorrect reasoning that electric fields point in the direction of increasing electric potential. Conversely, the mixed correct and incorrect community 16F-14B*-16B* includes all negative responses, which require the correct reasoning that electric field lines point in the direction of decreasing electric potential. Both communities contain an incorrect response, 16E (positive) or 16F (negative), indicating the potential difference is proportional to the length of the path $\sqrt{h^2 + w^2}$. Students who choose these responses calculate the potential difference using the total path traveled, rather than the projection of the path in the direction of the field.

As in the Coulomb's law communities, two types of incorrect reasoning are present. The first uses the incorrect relationship between electric field direction and lower electric potential. The second incorrectly reasons that the electric potential difference is proportional to the distance between the endpoint regardless of the field direction. The two reasoning errors seem fairly independent and, as such, it seems reasonable different combinations of the two are found.

The mixed correct and incorrect community 14G-16G-15G* includes the zero response for each item; the potential is the same everywhere in a uniform electric field. It is unclear what reasoning lies behind this community. Many explanations are possible including that

the students believe the potential is constant or zero when the field is constant. Using the suggested grading rubric, students choosing zero potential difference for each item in this block, 14G, 15G*, and 16G, are graded as correct for both items 15 and 16.

There was no community of completely correct responses. This was the result of most students choosing the correct response to 15 and the incorrect responses to 14 and 16.

**Electric field from time varying magnetic field communities**

Items 28 and 29 present the students with a fairly complicated problem. The items involve a solenoid carrying a current increasing in magnitude with time. The current direction is given and two equidistant points outside the solenoid are indicated. Each item asks for the direction of the electric field at one of those points. Each item gives the student eight response choices including zero and "none of the above." The students score quite poorly on the two items, getting only 18% of the items correct on average. It should be noted that it is possible to develop a right-hand-rule which substantially simplifies the reasoning required to answer the item correctly [147], but this rule is not included in most widely adopted physics textbooks.

The network includes six completely incorrect communities and one completely correct community. The first two incorrect communities 28A-29B and 28C-29D are composed of responses where all field lines point toward the solenoid or out from the solenoid, respectively. These students seem to be trying to model the system as a linear charge; they have no correct way to determine the sign of the charge. The community 28D-29A includes responses consistent with the electric field of the solenoid, but in a direction opposite to the correct direction. Communities 28E-29E and 28F-29F are composed of responses consistent with the

magnetic field from the solenoid either in the same direction as the solenoid's field, 28E-29E, or opposite that direction, 28F-29F. The community 28G-29G includes responses that both indicate zero electric field around the solenoid.

This rich set of incorrect communities give a picture of the reasoning process when students are given a situation they have no idea how to solve (as evidenced by the 18% success rate on the items). The incorrect communities see the students trying an electric field model (linear charge), a magnetic field model (the electric field in the same direction or the opposite direction as the magnetic field in the solenoid), and that there is no electric field. It is also impossible to determine if the item is answered correctly because the students apply correct physical reasoning or because they apply the incorrect model of the magnetic field in the solenoid being equivalent to a current in the solenoid (then applying the right hand rule).

### 6.3.2 Completely correct communities

Only three completely correct communities were identified. Items 1 to 3 ask questions about two point charges; the completely correct community includes responses 1A*, 2A*, and 3B*. Item 1 requires that the student understand that force is proportional to charge. Item 2 either requires a knowledge of Newton's 3rd law or the application of Coulomb's law a second time to infer that the force on each charge is equal but opposite. Item 3 is dependent on the answer to Item 2 and requires the student to apply the $1/r^2$ dependence of the electric force.

Items 4 and 5 present the student with two equal but opposite charges spaced a small horizontal distance apart; the completely correct community contains responses 4E* and

5A*. The student must correctly select the dipole field direction both along the line of the dipole and along a perpendicular bisector of the dipole. Items 28 and 29 present the student with a solenoid which carries a current increasing in magnitude with time. The students are asked about the induced electric field direction outside the solenoid. The completely correct community contains items 28B* and 29C*, which represent a circular electric field co-axial with the solenoid pointing in the correct direction given by Faraday's law.

The completely correct communities show the instructor places where correct conceptual reasoning is being consistently applied. These areas of well developed conceptual reasoning may represent resources which could be leveraged to address areas of weakness.

### 6.3.3 Quantifying common mistakes

Table 6.2: Mistake Scores

| Mistake | Responses | $M_1$ | $M_2$ | $M_3$ |
|---|---|---|---|---|
| **Coulomb's law item block (75%)** | | | | |
| Averaged force of two charges. | 1B,2B | 7% | 8% | 7% |
| Force is not proportional to charge. | 1E,2E | 9% | 14% | 5% |
| **Electric dipole field item block (68%)** | | | | |
| Reversed electric field direction. | 4A,5E | 9% | 14% | 3% |
| **Potential difference item block (49%)** | | | | |
| Field lines point in the direction of increasing potential. | 14A,16A,16E | 29% | 37% | 22% |
| Potential difference is not related to electric field. | 14G,16G | 17% | 21% | 13% |
| Potential difference is path dependent. | 16E,16F | 43% | 43% | NA |
| **Magnetic dipole field item block (78%)** | | | | |
| Reversed magnetic field direction. | 21E,22A | 10% | 17% | 3% |
| **Electric field from time varying magnetic field item block (18%)** | | | | |
| Solenoid as point charge (negative). | 28A,29B | 6% | 9% | 4% |
| Solenoid as point charge (positive). | 28C,29D | 16% | 20% | 13% |
| Electric field lines (reversed). | 28D,29A | 23% | 26% | 20% |
| Magnetic field of solenoid (correct). | 28F,29F | 11% | 13% | 9% |
| Magnetic field of solenoid (reversed). | 28E,29E | 7% | 10% | 5% |
| Changing magnetic field creates no electric field. | 28G,29G | 17% | 18% | 15% |

MMA-P reveals a set of mistakes that students consistently make when taking the

BEMA. In order to compare how often students make these mistakes, and, consequently, form a hierarchy of the most and least common mistakes students make on the instrument, three scores were calculated to show both how often students make each mistake, and how consistently each mistake is made. These scores measure the general frequency of making the mistake ($M_1$), the fraction of students making the mistake at least once ($M_2$), and the fraction of students making the mistake consistently ($M_3$).

Table 6.2 presents a list of common mistakes in introductory electricity and magnetism detected by MMA-P; item responses that correspond to those mistakes are also presented. The mistakes are organized by item block. The average score of the item block is provided in parenthesis; the rate at which items in the block are answered correctly. Three "mistake" scores are provided to quantify the consistency and the frequency of making these mistakes. These scores are called mistake scores to differentiate them from the misconception scores that appear in previous module analysis works [56, 57, 93, 120]. A misconception is generally defined as a stable, alternative schema [6]; the incorrect responses identified seem to more likely result from simple mistakes or just generally not understanding the concept. Details of the mistake score calculation are presented in more detail in the Appendix at the end of this chapter.

The first mistake score, $M_1$, is the same score calculated in previous module analysis works, called the misconception score in those works. This score represents the average fraction of responses that are selected out of the mistake response group per exam. $M_1$ gives insight into the frequency of selection of each mistake per exam.

The second mistake score, $M_2$, measures the fraction of students selecting at least one mistake in the response group. This statistic measures the percentage of students making

the mistake at least once on the exam and gives the instructor a better measure of how prevalent the mistake is than $M_1$. $M_1$ and $M_2$ are very similar metrics. $M_1$ is included to be consistent with past module analysis works, but $M_2$ is a superior metric for how many students are making the mistake. For the reversed electric field of a dipole, 14% of these students select at least one of the responses representing the mistake.

The third mistake score, $M_3$, measures the fraction of students that chose every possible mistake within a mistake group. $M_3$ represents the percentage of students consistently making the mistake. For the same example above, $M_3 = 3\%$; therefore, 3% of students chose every response in the mistake group. The "potential difference is path dependent" $M_3$ score is marked Not Applicable (NA) because the two responses in the group cannot be selected at the same time.

The ratio of $M_3$ to $M_2$ shows the percentage of students choosing all responses in the group out of the students who chose at least one response in the group. For the reversed electric field direction mistake, only 3 out of 14 students, who chose one response, chose both responses associated with that mistake and 11 out of 14 chose only one response. In contrast, 7 out of 8 students who make the "averaged force of two charges" mistake choose both responses in the mistake group.

In summary, $M_1$ measures general frequency of mistake selection, $M_2$ measures frequency of selecting at least one mistake in a mistake group, and $M_3$ measures the consistency of mistake selection. The original mistake score, $M_1$, may be large through two separate mechanisms: (1) many students could select one of the mistakes in the group or (2) a smaller set of students could select all the mistakes in the group. $M_2$ measures the first effect and $M_3$ the second. Comparing $M_2$ to $M_3$ allows an instructor to decide which

mode is prevalent in their class. If $M_2 \gg M_3$ then most students are only selecting a few, possibly one, of the responses in the group. A high number of students are confused about that concept, and the general topic might need covered for a longer amount of class time with some minor intervention such as a group problem. If $M_2 \approx M_3$, then students who are making the mistake are doing so consistently. The concept may require more targeted instructional time with different instructional methods.

## 6.4 Discussion

*RQ1: What community structure is identified by network analysis of the BEMA? What underlying reasoning could explain these response patterns?* The structure identified by the application of MMA-P to the BEMA was discussed at length in Sec. 6.3. All 17 communities identified are associated with 1 of 5 item blocks. The communities of 4 out of the 5 item blocks include 1 completely correct community containing all correct responses to the items in the block; the rest of the communities associated are completely incorrect communities. The other block, the potential difference item block, has 1 completely incorrect community, 2 mixed correct and incorrect communities, and 0 completely correct communities.

The community structure of item blocks found in this work is consistent with the correlation analysis performed by Hansen and Stewart, [150] which found that the majority of the instrument's substructure was centered around item blocks. Two separate studies [150, 151] applied exploratory factor analysis to the BEMA; each identified a five-factor model as optimal. Every community identified with MMA-P was composed of items that fit into one of those five factors: electrostatics, electric potential, magnetostatics, and magnetic

induction. The fifth factor was electric circuits; no community was identified for the electric circuit items.

Two item blocks did not have associated mistake communities: item block {8, 9} and {26, 27}. Item block {8, 9} is related to electric circuits and item block {26, 27} asks about the magnitude and direction of the external electric field for an electron moving in an electric field and a magnetic field. The first item in each block is a fairly straightforward qualitative item; the second item (on which the students score poorly) is a semi-quantitative item asking the students to select a formula. This combination of a qualitative and a semi-quantitative item may explain why no correct or incorrect communities are identified for these blocks. This is consistent with the analysis by Ding *et al.* [54], which found that items 9 and 27 were both problematic in difficulty and discrimination. Item 27 fell well below the 30% CTT item difficulty threshold for a well functioning item [155].

Item block {28, 29} stood out as potentially problematic. The average score on both of these two items was only 18%, also below the threshold suggested by CTT. The mistake communities show the students applying a variety of inappropriate models, both electric and magnetic, to the item. It seems likely that the students sampled simply have no idea how to answer these items. In most physics classes, the topic tested, the induced electric field caused by a changing magnetic field, receives little class time as compared to other more common topics in magnetic induction such as the direction of current induced in a coil of wire. As the next generation of conceptual instruments is constructed, there is likely a better choice than the inclusion of these or similar items. This item provides an interesting window into student reasoning when presented with an unfamiliar situation.

The frequency and consistency of selection of the communities identified were charac-

terized with mistake scores. For items blocks {1, 2, 3}, {4, 5}, and {21, 22}, the fraction of students selecting at least one mistake from each community ($M_2$) was small, less than 20%, while those selecting both mistake responses ($M_3$) was very small, less than 10%. For the course studied, it is unlikely additional instructional time should be directed to eliminating these mistakes. Mistakes associated with item block {14, 15, 16} were far more common. Thirty-seven percent of the students selected a response representing the electric field points to higher electric potential mistake; twenty-two percent consistently selected this mistake. Forty-three percent also made the mistake of using the overall path length to calculate the potential difference instead of the path length in the direction of the electric field. It seems likely the class studied would benefit from investing some additional time on electric potential.

*RQ2: How does the community structure of the BEMA relate to the community structure of the CSEM?* The community structure identified for the CSEM was substantially richer than that identified for the the BEMA in this work with 13 completely incorrect communities, 6 mixed correct and incorrect communities, and 6 completely correct communities. Many completely incorrect communities involved concepts not covered on the BEMA such as the properties of conductors and insulators, the field direction of an infinite straight wire, the linear superposition of the electric field, and the shielding of the electric field by a conductor. Seven of the communities were formed because students failed to discriminate between electric and magnetic fields or between electric potential and electric field; this type of incorrect response was not available on the BEMA. Multiple items could involve Newton's 3rd law in the solution; an incorrect community of these items was identified. The correct communities were disparate, including the shared Coulomb's law items (BEMA items 1, 2,

and 3), the force on a stationary charge is zero, the field of an infinite straight wire, and correctly interpreting the meaning of the spacing of electric potential.

Most of the overlap in community structure between the two instruments came from the first item block. Item block {1, 2, 3} in the BEMA uses the same questions as item block {3, 4, 5} in the CSEM, but the possible responses are different. In the BEMA, items 1 and 2 have 7 possible responses and item 3 has 9, while each CSEM item has 5 possible responses. Table 6.3 shows the responses to items 1, 2, and 3 on the BEMA and the corresponding items 3, 4, and 5 on the CSEM. The final response for each item in the item block in both instruments has the same meaning, but is worded differently: the BEMA uses "None of the above," while the CSEM uses "other." BEMA item 3 responses were written such that every response from item 2 had a consistent inverse square law response in item 3. CSEM item 5 does not include every inverse square response to CSEM item 4; both $F/4$ and $16F$ do not have an inverse square response in item 5.

| BEMA items 1, 2 | | | BEMA item 3 | | |
|---|---|---|---|---|---|
| CSEM items 3, 4 | | | CSEM item 5 | | |
| Response | BEMA | CSEM | Response | BEMA | CSEM |
| $4F$ | A* | B* | $4F/9$ | B* | C* |
| $5F/2$ | B | | $5F/18$ | D | |
| $3F$ | C | | $F/3$ | C | B |
| $2F$ | D | | $2F/9$ | E | |
| $F$ | E | C | $F/9$ | F | A |
| $F/4$ | F | D | $F/36$ | G | |
| NOTA | G | | NOTA | I | |
| other | | E | other | | E |
| $16F$ | | A | $4F$ | H | |
| | | | $4F/3$ | A | D |

Table 6.3: Responses to items 1, 2, and 3 in the BEMA and the corresponding items 3, 4, and 5 in the CSEM. Responses to items 1 and 2 are ordered by their appearance in the BEMA, while item 3 is ordered to show the consistent $1/r^2$ response to items 1 and 2. NOTA denotes the "None of the above." response.

Applying MMA-P to the CSEM found 5 communities including items from this item

block. CSEM community 5A-(3C, 4C)-5B appeared in both samples and is identical to BEMA community 3C-(1E,2E)-3F. This community was discussed in Sec. 6.3.1. The completely correct community for this item block appeared in both the BEMA and CSEM networks.

Multiple communities were identified in the CSEM involving items 3, 4, and 5 which were not found in the BEMA. CSEM community 3A-4A contained responses that were not included on the BEMA. Both 3A and 4A correspond to a quadratic relation between charge and force (16F). CSEM community 4D-5E may have resulted from students trying to choose the consistent inverse square law response $F/36$ based on their answer to item 4 ($F/4$), which was not a possible response on the CSEM, therefore "other" was chosen instead.

Additional communities which appeared in the CSEM networks mixed items within the 3, 4, and 5 item block with items outside the block. Students who chose the same response for both BEMA items 1 and 2 or CSEM items 3 and 4 could be using Newton's 3rd law to determine that the forces on the charges must be equal and opposite. The CSEM networks included other item responses related to Newton's 3rd law outside of this block which appeared in some communities related to items 3, 4, and 5. The BEMA does not contain additional items applying Newton's 3rd law and, therefore, these communities were not available.

The CSEM included a large community of "zero" or "none of the above" responses. The BEMA included two communities of zero responses, 14G-16G-15G* and 28G-29G, which were not linked into the same community. Wheatley *et al.* [93] speculated that the CSEM community resulted from the reluctance of students to select zero or none of the above responses. Devore *et al.* demonstrated that these responses are selected by students at

a lower rate than would be expected based on the distribution of selection of the other responses [119].

The community structure of the two instruments differed because the instruments probe different detailed skills in electromagnetism. For example, the communities 4A-5E and 21E-22A represent incorrect reasoning about the field of a dipole. The CSEM includes no items asking about dipole fields. The CSEM contains two items asking for the magnetic force on a stationary charge; no such items appear in the BEMA.

The CSEM networks, unlike those of the BEMA, included communities of responses that indicated students were conflating electric fields with magnetic fields as well as electric fields with electric potential. This failure to distinguish between important physical concepts was described as a naive conception in the FCI through the "velocity-acceleration undiscriminated" or the "position-velocity undiscriminated" responses [9, 61]. The CSEM has items that ask the same question with a background electric or magnetic field. The BEMA does not include similar items, therefore, no communities formed indicating students conflated the two fields. One can see some evidence of this kind of naive reasoning in the responses to items 28 and 29 where students use a variety of different models including a linear charge electric field for the induced electric field of a solenoid.

### 6.4.1 Other Observations

Many items did not have a response included in any community (items 6-13, 17-20, 23-27, 30, and 31, a total of 20 items). To be included in the network, a response must have be selected by at least 5% of the students and must have partial correlation with another response largest than $r = 0.17$. This represents a small effect size correlation by Cohen's

criteria [29]. At the sample size of this study, no edges meeting the correlation threshold are eliminate by the significance threshold with Bonferroni correction. Items not included have responses that are not consistently selected with other responses. Given the broad coverage of the instrument with items measuring many different constructs, the exclusion of many items from the network communities is not surprising. This was also found in other MMA works [56–58, 93].

To understand why items may not be included in the network, it is productive to consider the electric circuit items. No electric circuit item is included in a community. The six items measure a broad set of generally non-overlapping concepts. Items 8 and 9 ask the student about an ion channel with item 8 asking for the conventional current direction and item 9 asking for a formula for the current in the channel. Item 10 includes a circuit containing a single light bulb and an ammeter and asks about the ammeter reading if the meter is before the light bulb, after the light bulb, and if the bulb is removed from the circuit. Item 11 gives the students three circuits which include (1) a single light bulb, (2) two bulbs in series, and (3) two bulbs in parallel; the item asks the student to rank the brightness of the bulbs. Item 12 gives the student a single bulb and asks about the electric field in the filament. Items 13 is an RC circuit item. These items require both different correct conceptual knowledge to be answered correctly and involve different mistakes generating small correlations between both the correct answers and the incorrect answers. This is supported by the results of Hansen and Stewart [150] applying MIRT to the BEMA, which showed none of the items from 6 to 13 substantially loaded onto the same factor. The number of missing items is larger than found in other MMA studies; FCI (7 missing [58]), FMCE (7 missing [57], and CSEM (7 total missing between two samples [93]). This may

indicate that several items in the BEMA test very specific concepts not well related to a general conceptual understanding of electromagnetism and that these concepts are tested by a single item, rather than pairs of items to allow for correlations between the responses. This would also serve to explain why few completely correct communities were identified.

Both the BEMA and the CSEM were designed to broadly test a student's conceptual knowledge of electromagnetism. Module analysis of both the BEMA and CSEM identified groups of items which allowed the identification of consistently applied mistakes. Except for an item block shared by both instruments, the consistently applied mistakes differed between the two instruments. While not the primary purpose of either instrument, the identification of consistently applied incorrect reasoning either pre- or post-instruction and the ability to quantify how consistently that reasoning is applied is extremely valuable to physics instructors. It is very unlikely that the consistently applied mistakes identified in the two instruments represent a complete catalog of mistakes in introductory electromagnetism.

The MIRT models for the BEMA [150] and the CSEM [102] showed the detailed coverage of the two instruments were quite different. The module analysis studies of the instruments [93] showed the consistently applied incorrect reasoning was also different. It seems likely that these detailed differences will make the two instruments sensitive to the coverage of classes in which they are applied; evidence for this can be seen in the performance on item block {28, 29}. When viewed through the detailed lens of MIRT or MMA-P, some item choices on both instruments seem ill-advised if the goal is a broad measure of electromagnetism, producing scores comparable across a variety of classrooms and instructional contexts.

## 6.5 Implications

The differences in conceptual coverage and the kinds of common mistakes measured identified above have potential implications for research using these instruments. Educational interventions often change curricula in very local ways, modifying specific laboratory activities, adding specific group activities. If the conceptual coverage of the intervention does not align with the instrument used, then the instrument may not measure the efficacy of the intervention properly. Further, an intervention characterized using, for example, the CSEM may show performance differences when replicated at an institution monitoring learning with the BEMA simply because of the coverage difference in the two instruments. Further, because the instruments measure different consistently applied mistakes, an intervention addressing one of these mistakes may be completely mis-characterized by applying the instrument which does not measure the common mistake.

All these observations suggest the need for a new generation of electromagnetic conceptual instruments which feature both subscales with broad general coverage of the major subtopics of electromagnetism, but also subscales that capture common mistakes or allow measurement of finer details of conceptual knowledge. It seems impossible that a single instrument could meet these goals with an acceptable number of items, which may suggest a move away from the long, single-construct instruments which have been used in PER, and onto new testing structures, possibly involving flexible instruments where an instructor could assemble an instrument from a bank of subscales that target's their instructional needs.

Module analysis can be an important part of the validation process for these new instruments adding to more traditional psychometric analysis techniques. MMA augments

tradition methods by examining combinations of items. While MMA-P did not identify item combinations in the BEMA which were not well functioning, prior studies of the FCI, FMCE, and CSEM did identify items that might not be functioning as intended.

## 6.6 Conclusion

Module analysis of the BEMA identified a set of communities containing consistently selected correct and incorrect answers. All communities contained items restricted to a single item block. Mistake scores were introduced to further quantify the consistency and frequency of selection of these consistent mistakes. Instructors can use the mistake scores to target instruction to address the most common mistakes made by their students. For the course studied, the most commonly selected and consistently applied mistakes involved electric potential difference and the relation of electric potential difference to electric field: the mistakes that electric field lines point in the direction of increasing potential and that electric potential difference depends on the total path length not the length of the path in the direction of the field. Other consistent mistakes were identified, but were applied by a small number of students.

The BEMA and CSEM share three items with identical stems but different response choices. Communities were identified in both instruments which involved response choices not available in the other instrument. In the BEMA, a community where the student reported the average force on the two charges was identified; it was impossible for students on the CSEM to select this response. Likewise, the CSEM did not include all responses related to consistent reasoning about the distance dependence of the force, while the BEMA did. The

CSEM did include an "other" response to catch these forms of incorrect reasoning. Both instruments revealed unique aspects of student thinking because they allowed for different student responses.

## 6.7 Appendix

### 6.7.1 Partial Correlation Threshold

As discussed in detail in Chapter 5, a method to identify the optimal correlation threshold for network sparsification was used in this work. The community structure was calculated for a range of correlation thresholds from $r = 0.1$ to $r = 0.3$; the average community size (ACS) of each network was plotted against the total number of communities (NC). ACS is the average number of nodes per community. Figure 6.2 shows ACS plotted against NC; the points are labeled by the correlation threshold. Networks calculated through MMA-P tend to be too dense for theoretical description at low $r$ thresholds but become sparse at high $r$ thresholds. This method is similar to choosing the optimal number of clusters or factors by identifying the "knee" in a scree plot. From Figure 6.2, $r = 0.17$ where the curve bends upward was selected as the threshold applied in this work. A detailed exploration of the effects of sparsification techniques used in MMA-P is discussed in Chapter 7.

### 6.7.2 Mistake Scores

| | | $R_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| $R_2$ | 0 | A | B |
| | 1 | C | D |

Table 6.4: $2 \times 2$ contingency table

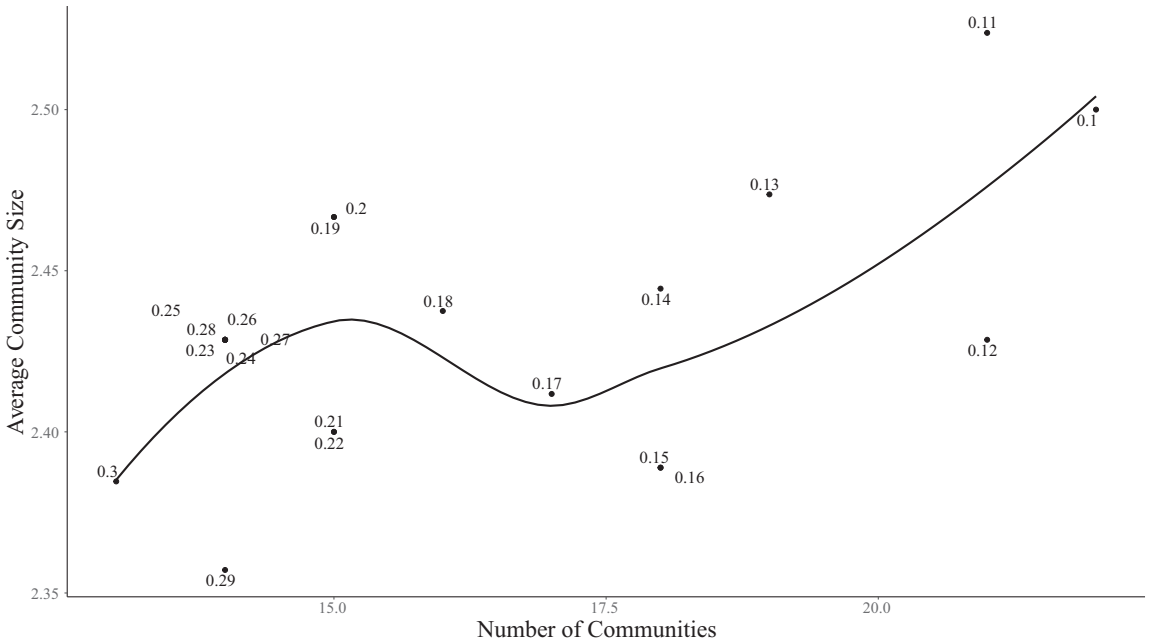


Figure 6.2: Plot used to determined the correlation threshold r. Each point represents a network calculated at the labeled r value.

The mistake scores shown in Table 6.2 can be calculated from the contingency table between two item responses, $R_1$ and $R_2$, as shown in above in Table 6.4. There are four possible combinations for two item responses: choosing neither, choosing $R_1$ and not $R_2$, choosing $R_2$ and not $R_1$, and choosing both.

Mistake score $M_1$ is calculated with Equation 6.1.

$$M_1 = \frac{B + C + 2D}{2N} \tag{6.1}$$

where $N = A + B + C + D$ is the total number of students. The numerator of Equation 6.1 is derived by summing the "1's" in Table 6.4: $B + D$ for $R_1$ and $C + D$ for $R_2$. The $2N$ represents the total number of times either response could be selected.

Mistakes that emerge from multiple responses to the same item, such as the "potential difference is path dependent" mistake and the "field lines point in the direction of increasing

potential" mistake, require some nuance in their calculation for $M_1$. $M_1$ is normalized by the number of items in the response group; the denominator in Equation 6.1 is 2N because a contingency table is built for a response group of size 2. For the "potential difference is path dependent" mistake, the only two responses that make up that mistake come from the same item, 16E and 16F, so the maximum value of the numerator in $M_1$ is $B + C = N$ because D is always 0 for multiple responses to the same item. For this mistake, $M_1$ is normalized by N rather than 2N. Likewise, with the "field lines point in the direction of increasing potential" mistake, the maximum number of responses made per student is two; 14A and 16A or 16E, so $M_1$ is normalized by 2N rather than 3N. Calculating the numerator in $M_1$ for this mistake is more easily understood as simply summing the number of students who chose each response, rather than utilizing multiple contingency tables for the possible combinations of responses.

The second mistake score, $M_2$, measures the fraction of students selecting at least one mistake in the response group and is calculated in Equation 6.2.

$$M_2 = \frac{B + C + D}{N} = 1 - \frac{A}{N} \tag{6.2}$$

The "potential difference is path dependent" mistake is calculated as shown in Equation 6.2 with $D = 0$. When $D = 0$ and the size of the response group is 1, both $M_1$ and $M_2$ measure the same thing because the fraction of students selecting at least one mistake in the response group (of size 1) is exactly the fraction of responses that are selected out of the response group; students either selected a the mistake or they didn't. The $M_2$ calculation for "field lines point in the direction of increasing potential" mistake can be more easily

understood with the $1 - A/N$ shown in Equation 6.2, where A is now the number of students who did not choose 14A, 16A, or 16E.

The third mistake score, $M_3$, measures the fraction of students that chose every possible mistake within a mistake group and is calculated in Equation 6.3.

$$M_3 = \frac{D}{N} \tag{6.3}$$

$M_3$ measures the percentage of students consistently making the mistake.

The mistake score for the "potential difference is path dependent" mistake is Not Applicable (NA) because both responses that demonstrate that mistake are responses to the same item and therefore cannot be chosen together. The calculation for the "field lines point in the direction of increasing potential" mistake is calculated with a contingency table between 14A and 16A and one between 14A and 16E. Choosing every possible mistake in the response group for this mistake means either 14A with 16A or with 16E, so the numerator in $M_3$ is calculated by summing the two Ds from both contingency tables. Regardless of the size of the response group, both $M_2$ and $M_3$ are normalized by N because the maximum value of the numerator is always N.

### 6.7.3 Item Responses and Scores

Table 6.5 presents the overall score on each item as well as the number of times each item's response was selected.

Table 6.5: Item response frequency ($N = 12,214$) and score for each item. The adjusted scores based on the suggested BEMA grading criteria for items 3, 16, and 28 and 29 are included in parenthesis. The 5% response threshold for this sample is 611. Item responses that do not appear on certain items are reported as NA for "Not Applicable."

| Item | Score | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 83.6% | 10212 | 874 | 111 | 149 | 750 | 95 | 22 | NA | NA | NA |
| 2 | 74.8% | 9141 | 902 | 124 | 156 | 1557 | 278 | 30 | NA | NA | NA |
| 3 | 67.6% (87.2%) | 887 | 8257 | 611 | 746 | 136 | 1228 | 123 | 91 | 135 | NA |
| 4 | 79.0% | 1239 | 107 | 263 | 131 | 9654 | 171 | 102 | 77 | 406 | 64 |
| 5 | 57.0% | 6959 | 448 | 648 | 170 | 855 | 446 | 100 | 75 | 2316 | 197 |
| 6 | 59.0% | 256 | 2362 | 303 | 7208 | 1096 | 801 | 182 | NA | NA | NA |
| 7 | 50.5% | 4492 | 480 | 397 | 564 | 6174 | 93 | NA | NA | NA | NA |
| 8 | 77.6% | 1937 | 9483 | 712 | NA | NA | NA | NA | NA | NA | NA |
| 9 | 30.4% | 4635 | 3720 | 1963 | 1383 | 471 | NA | NA | NA | NA | NA |
| 10 | 61.7% | 1780 | 829 | 1021 | 263 | 144 | 7539 | 482 | 94 | 59 | NA |
| 11 | 44.4% | 733 | 237 | 718 | 211 | 5429 | 590 | 1446 | 2784 | 65 | NA |
| 12 | 22.0% | 1484 | 554 | 996 | 5751 | 2683 | 322 | 415 | NA | NA | NA |
| 13 | 78.7% | 316 | 476 | 1264 | 9609 | 523 | NA | NA | NA | NA | NA |
| 14 | 45.6% | 3435 | 5567 | 264 | 209 | 185 | 76 | 2468 | NA | NA | NA |
| 15 | 77.1% | 195 | 254 | 1185 | 844 | 195 | 115 | 9420 | NA | NA | NA |
| 16 | 25.4% (48.2%) | 1536 | 3100 | 354 | 297 | 2227 | 3006 | 1692 | NA | NA | NA |
| 17 | 32.2% | 4536 | 253 | 3194 | 3929 | 274 | NA | NA | NA | NA | NA |
| 18 | 55.3% | 2589 | 6756 | 1350 | 1457 | NA | NA | NA | NA | NA | NA |
| 19 | 77.8% | 334 | 9507 | 336 | 998 | 318 | 710 | NA | NA | NA | NA |
| 20 | 61.7% | 1932 | 541 | 460 | 275 | 628 | 458 | 7533 | 383 | NA | NA |
| 21 | 86.6% | 10573 | 157 | 163 | 102 | 772 | 130 | 127 | 81 | 82 | 27 |
| 22 | 68.7% | 1625 | 389 | 328 | 509 | 8390 | 266 | 104 | 146 | 412 | 45 |
| 23 | 50.5% | 292 | 865 | 277 | 1095 | 6173 | 3350 | 153 | NA | NA | NA |
| 24 | 70.1% | 8563 | 201 | 638 | 261 | 283 | 206 | 1995 | 64 | NA | NA |
| 25 | 58.2% | 183 | 2069 | 218 | 7104 | 1908 | 239 | 441 | 46 | NA | NA |
| 26 | 40.3% | 2087 | 3534 | 569 | 4919 | 595 | 289 | 203 | NA | NA | NA |
| 27 | 12.8% | 1757 | 3260 | 1567 | 411 | 3737 | 520 | 404 | 552 | NA | NA |
| 28 | 18.0% | 727 | 2200 | 2032 | 2776 | 939 | 1350 | 2026 | 158 | NA | NA |
| 29 | 18.1% (15.2%) | 2816 | 823 | 2212 | 1946 | 858 | 1337 | 2066 | 147 | NA | NA |
| 30 | 40.8% | 905 | 1488 | 378 | 251 | 2531 | 4986 | 1667 | NA | NA | NA |
| 31 | 29.4% | 3894 | 1932 | 1069 | 3588 | 1019 | 698 | NA | NA | NA | NA |

# Chapter 7

## More on Module Analysis

This chapter explores module analysis examining the effects of sparsification order, discrete correlations, statistical power with varying sample sizes, and correlation thresholds. Examples from Chapter 5 are used throughout this chapter because the varying student performance provides a good range of realistic response patterns and the varying sample sizes help demonstrate the necessity of larger sample sizes for this method.

## 7.1 Sparsification Analysis

This section presents a summary of the sparsification process for each sample. The sparsification operations applied with MMA-P are the minimum student response threshold (5% in Chapter 5), requiring edges represent correlations between nodes with significance of $p < 0.05$ after a Bonferroni correction is applied, requiring edges to have positive correlations, requiring those correlations to be above a correlation threshold (generally around $r > 0.2$ where $r$ is the partial correlation coefficient between nodes), and requiring the edge be detected in the same community in 80% of bootstrap replications. For example, using Table 7.1 and focusing on Sample 4, one can follow the sparsification process. Initially there are 150 responses, 1A to 30E; removing edges representing correlations that are not significant after applying a Bonferroni correction disconnects 6 responses from the network leaving 144 responses. Removing edges representing negative correlations isolates another 29 responses; removing isolated nodes leaves 115 responses. Applying the correlation threshold requiring $r > 0.2$ (for example) removed additional edges isolating 7 more nodes; removing these left 108 nodes. A response threshold requiring at least 5% of the students select the response was then applied; this removed 65 responses leaving 43 in the network. The dataset was then bootstrapped 1000 times; a community detection algorithm was applied to each bootstrap replication. The number of times two nodes were identified in the same community was calculated. If the two nodes were not identified in the same community on 80% of the replications the edge was removed; this isolated 14 nodes. Isolated nodes were then removed leaving 29 responses.

Tables 7.1 and 7.4 present the nodes remaining after each step of the sparsification pro-

cess of MMA-P as it was applied in Chapter 5. Because the Bonferroni correction depends on the number of statistical tests performed, the order of these operations should be investigated. In this study, we chose to apply the Bonferroni corrected significance threshold first because we felt the highest priority should be to eliminate the consideration of statistically insignificant structures; however, we acknowledge an argument can be made for applying the student response threshold first to minimize the number of statistical tests performed. Tables 7.2 and 7.5 present a comparison of the resulting structure if the student response threshold is applied first or after the Bonferroni corrected significance threshold. For all samples, the order of the response threshold and the significance threshold does not change the number of nodes in the final network for the post-test; some small differences are found in the pretest network for Samples 1 to 4. The pretest differences were more pronounced for Sample 5. As such, MMA-P is generally not sensitive to the order of applying the response threshold and the significance threshold. The reason for this is likely that the $r > 0.2$ correlation threshold is a very strong criteria ($r = 0.1$ represents a small effect and $r = 0.3$ a medium effect) making the significance threshold unimportant. Even at the size of Sample 5, a correlation of $r > 0.2$ is significant with a small $p$ value.

The difference in the number of final nodes between the response threshold of 30 in prior studies and 5% in this study was also examined. Tables 7.3 and 7.6 show that there was little effect for Samples 1 to 4 for the post-test; however, the number of nodes in Sample 5 changed from 14 with the 5% threshold to 8 with the 30 threshold. Differences were even smaller in the pretest networks. Naturally, nodes removed by either the 30 or 5% response threshold are selected by a few students.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Bonferroni correction | 150 | 150 | 150 | 144 | 107 |
| Negative correlations | 149 | 148 | 133 | 115 | 87 |
| Correlation threshold | 30 | 50 | 72 | 108 | 87 |
| Response threshold (5%) | 28 | 35 | 48 | 43 | 28 |
| Community fraction | 27 | 32 | 32 | 29 | 14 |
| | | | | | |
| N | 9606 | 4360 | 1496 | 466 | 213 |

Table 7.1: The number of nodes retained for each sample at each stage of the sparsification process used in the current paper for the post-test.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Response threshold (5%) | 100 | 74 | 115 | 85 | 53 |
| Bonferroni correction | 100 | 74 | 115 | 84 | 51 |
| Negative correlations | 100 | 73 | 97 | 52 | 32 |
| Correlation threshold | 28 | 35 | 48 | 43 | 32 |
| Community fraction | 27 | 32 | 32 | 29 | 14 |
| | | | | | |
| N | 9606 | 4360 | 1496 | 466 | 213 |

Table 7.2: The number of nodes retained for each sample at each stage of the sparsification process exchanging the order of the response threshold and Bonferroni correction for the post-test.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Response threshold ($N = 30$) | 148 | 128 | 139 | 82 | 35 |
| Bonferroni correction | 148 | 128 | 139 | 81 | 29 |
| Negative correlations | 147 | 124 | 120 | 51 | 21 |
| Correlation threshold | 30 | 41 | 60 | 42 | 21 |
| Community fraction | 29 | 36 | 32 | 27 | 8 |
| | | | | | |
| N | 9606 | 4360 | 1496 | 466 | 213 |

Table 7.3: The number of nodes retained for each sample at each stage of the sparsification process using the 30 item response threshold from previous studies for the post-test.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Bonferroni correction | 150 | 150 | 149 | 140 | 129 |
| Negative correlations | 149 | 146 | 127 | 100 | 93 |
| Correlation threshold | 12 | 39 | 65 | 100 | 93 |
| Response threshold (5%) | 12 | 31 | 48 | 54 | 35 |
| Community fraction | 12 | 26 | 22 | 20 | 8 |
| N | 9606 | 4360 | 1496 | 466 | 213 |

Table 7.4: The number of nodes retained for each sample at each stage of the sparsification process used in the current paper for the pretest.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Response threshold (5%) | 125 | 109 | 125 | 103 | 85 |
| Bonferroni correction | 125 | 109 | 125 | 101 | 81 |
| Negative correlations | 123 | 105 | 102 | 59 | 40 |
| Correlation threshold | 12 | 31 | 48 | 54 | 40 |
| Community fraction | 12 | 26 | 24 | 20 | 12 |
| N | 9606 | 4360 | 1496 | 466 | 213 |

Table 7.5: The number of nodes retained for each sample at each stage of the sparsification process exchanging the order of the response threshold and Bonferroni correction for the pretest.

| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| Response threshold ($N = 30$) | 149 | 144 | 141 | 95 | 51 |
| Bonferroni correction | 149 | 144 | 141 | 95 | 44 |
| Negative correlations | 148 | 139 | 118 | 57 | 26 |
| Correlation threshold | 12 | 37 | 57 | 51 | 26 |
| Community fraction | 12 | 28 | 24 | 20 | 10 |
| N | 9606 | 4360 | 1496 | 466 | 213 |

Table 7.6: The number of nodes retained for each sample at each stage of the sparsification process using the 30 item response threshold from previous studies for the pretest.

## 7.2 Exploring Discrete Correlations

### 7.2.1 Definition of $\phi$

To gain an intuitive understanding of the working of the various module analysis algorithms, it is useful to consider the correlation between dichotomously scored responses, $R_1$ and $R_2$. Each response may have values of 0 or 1 for each student. The correlation between

the two responses is given by the $\phi$ coefficient which is calculated from the $2 \times 2$ contingency table shown in Table 7.7.

| | | Response $R_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| Response $R_2$ | 0 | A | B |
| | 1 | C | D |

Table 7.7: $2 \times 2$ contingency table

The contingency table captures how many students select the two responses in each of the four possible combinations. For example, if 400 students selected both responses ($R_1 = 1$ and $R_2 = 1$), $D$ would equal 400. If 300 students selected response $R_1$, but not response $R_2$, then $B$ would equal 300. The $\phi$ coefficient, the correlation between the responses, is calculated from the contingency table using Eq. 7.1.

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \tag{7.1}$$

The odd form of the denominator results from taking all combinations of the marginal sums of the rows and columns.

To measure a large correlation, most students must either select both responses ($D$) or neither response ($A$), while also not selecting one and not the other ($B$ and $C$). From the form of $\phi$, one can see that consistency is prioritized by the correlation coefficient.

### 7.2.2 Relation of MAMCR and MMA

The contingency table (Table 7.7) can be used to understand the relation between the MAMCR algorithm and the MMA algorithm which sought to extend MAMCR to large

samples. MAMCR initially constructed a bipartite network containing both students and responses; it then projected this onto a unipartite network containing only responses where edge weights are the number of times a student selected both responses. The edge weight in MAMCR is then the coefficient $D$ in Table 7.7 for $R_1$ and $R_2$. The edge weight in MMA is the $\phi$ coefficient. Two example $2 \times 2$ tables illustrate the difference in these two algorithms; both contain 1000 total responses. Table 7.8 shows the table that results if the two responses are answered perfectly consistently with 25% of the students selecting both responses.

| | | Response $R_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| Response $R_2$ | 0 | 750 | 0 |
| | 1 | 0 | 250 |

Table 7.8: $2 \times 2$ contingency for perfectly consistent answering; $\phi = 1$.

Table 7.9 shows the table that results if students select the two responses completely at random. For the perfectly consistent table (Table 7.8), MAMCR would use an edge

| | | Response $R_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| Response $R_2$ | 0 | 250 | 250 |
| | 1 | 250 | 250 |

Table 7.9: $2 \times 2$ contingency for random answering; $\phi = 0$.

weight of $D = 250$ between the responses and MMA would use an edge weight of $\phi = 1$. For the perfectly random guessing responses in Table 7.9, MAMCR would also use an edge weight of $D = 250$ while MMA would use an edge weight of $\phi = 0$. This serves to explain why MAMCR failed to scale to large datasets. If one has a sufficient number of two responses to two items, even if the selection of the responses together are completely random, a large number will be selected together by chance and thus the two responses will

be connected by an edge in the MAMCR network. As MAMCR is scaled to larger datasets, it progressively identifies more random structure as consistent student thinking. This explains why MAMCR was productive using the 143 student dataset in Brewe *et al.* [55] where it was introduced, but could not be scaled to the large 4500 student dataset by Wells *et al.* [56]. The difference in handling the contingency tables represents only one way MMA and MAMCR differ. MMA uses a global sparsification method which eliminates structure resulting from responses selected by few students, while MAMCR uses a sparsification process which attempts to retain meaningful structures at all response levels by applying the locally adaptive network sparsification (LANS) algorithm, which retains edges which weights are larger than a percentage of other local edge weights.

### 7.2.3 Example Contingency Tables

To further illustrate these contingency tables with real data, this section presents some example contingency tables from Sample 3. Table 7.10 shows the contingency table for a 2-response community identified in Sample 3 with the smallest $D$, community (8E,21A).

| | | Response $8E$ | |
|---|---|---|---|
| | | 0 | 1 |
| Response $21A$ | 0 | 1191 | 78 |
| | 1 | 173 | 54 |

Table 7.10: $2 \times 2$ contingency for responses 8E and 21A which were identified as a community in Sample 3.

Compare this with two responses not identified as a community also with fairly small $D$ in Table 7.11 (responses 3A and 28D). The off diagonal terms for (3A,28D) are much larger than those of 8E and 21A indicating 8E and 21A are being selected by the same student more consistently.

| | | Response $3A$ | |
|---|---|---|---|
| | | 0 | 1 |
| Response $28D$ | 0 | 729 | 406 |
| | 1 | 237 | 124 |

Table 7.11: $2 \times 2$ contingency table for responses for responses $3A$ and $28D$.

Most incorrect communities that appear look something like $5D$ and $18D$ shown in Table 7.12, where the D term is fairly large and plays a role in offsetting the negative effect of the B and C terms.

| | | Response $5D$ | |
|---|---|---|---|
| | | 0 | 1 |
| Response $18D$ | 0 | 790 | 300 |
| | 1 | 190 | 216 |

Table 7.12: $2 \times 2$ contingency table for responses for responses $5D$ and $18C$.

### 7.2.4 Parameterizing the Contingency Table

A $2 \times 2$ contingency table has four elements $A$, $B$, $C$, $D$ and can be naturally parameterized with the $\phi$ coefficient and three other parameters. The parameters $N$, $f_1$, $f_2$, and $\phi$ define any $2 \times 2$ table where $N$ is the sum of all elements $N = A + B + C + D$ representing the number of respondents, $f_1$ is the overall rate response $R_1$ is selected, and $f_2$ is the rate $R_2$ are selected; $f_1 = (B + D)/N$ and $f_2 = (C + D)/N$. The three parameters may be independently chosen where $N$ is a positive integer and $f_i$ are real numbers between 0 and 1. While appropriately selecting $\phi$, will generate all possible $2 \times 2$ tables for a selection of $N$ and $f_i$, not all choices of $\phi$ will produce positive values for $A$, $B$, $C$, and $D$.

Figure 7.1 simulates 1,000,000 random $2 \times 2$ tables. The figure has been subsampled to produce a constant density in all regions containing points. The $x$ axis plots $\bar{f} = (f_1 + f_2)/2$

and the $y$ axis the correlation coefficient. All positive $\phi$ values are available for all $\bar{f}$; however, a range of negative $\phi$ values are not available for either very uncommonly or very commonly selected responses. If the simulation had been run to sufficient $N$ without subsampling, the region $\phi \geq 0$ would have been entirely filled in. MMA filters out negative $\phi$ values; as such, the negative correlations are not important to the algorithm.

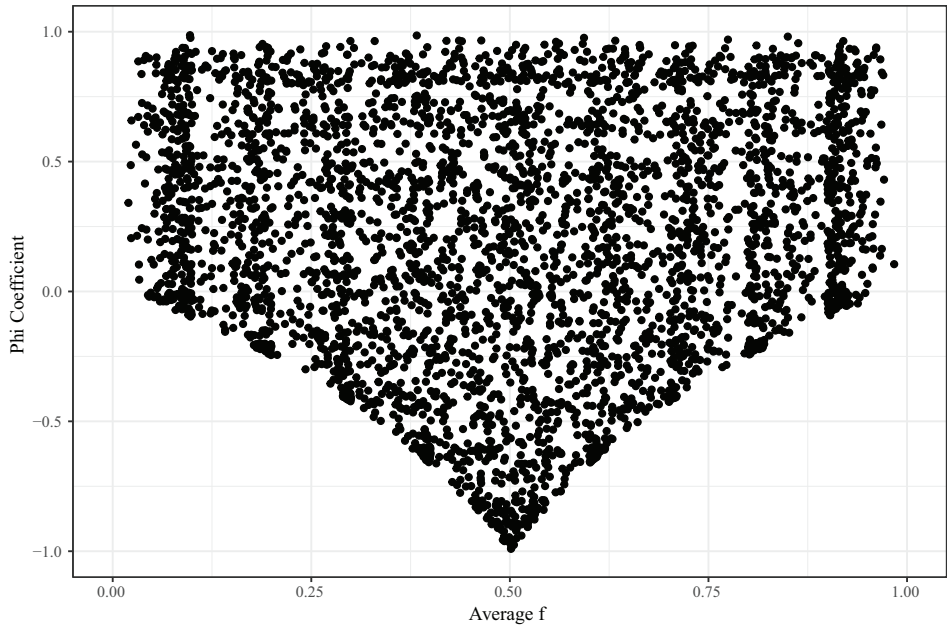


Figure 7.1: Plot of $\phi$ vs. $\bar{f} = (f_1 + f_2)/2$

### 7.2.5 Exploring Low Frequency Responses

The correlation between dichotomously scored items will be one if $A > 0$, $D > 0$ and $B = C = 0$. A correlation of one is very large, above a large effect in Cohen's categorization, and will be statistically significant for fairly small values of $D$. In MMA and MMA-P, restricting responses to those selected by a certain number of students, 30 in prior studies, 5% of $N$ in the present study, eliminates these very small $D$ significant correlations. To explore the properties of the contingency table which result in a significant correlation with

small $D$ values, assume the off diagonal terms in the $2 \times 2$ matrix as equal, $B = C$, and that the total number of responses is constrained to $N$. If $D$ is allowed to vary, this constrains $A$ to be $A = N - D - 2B$. Figure 7.2 plots the correlation coefficient $\phi$ against the size of the off diagonal terms, $B$, with $N = 1000$. This plot uses a significance threshold for $\phi$ of $p < 0.05$.

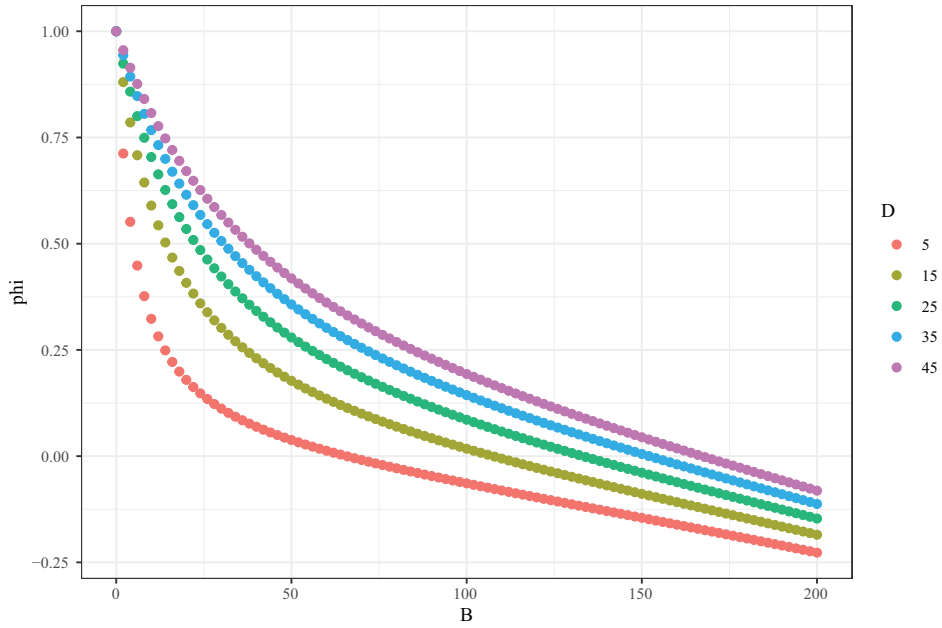


Figure 7.2: Plot of $\phi$ vs. $B$ for different levels of $D$, $N = 1000$.

Note, even for $D = 5$, one can meet the $\phi > 0.2$ threshold with sufficiently consistent answering, $B = C \approx 15$. As such, consistent structure selected by only a few students can meet the significance threshold.

### 7.2.6 Probability Threshold with Bonferroni Correction

Not all correlations shown in Figure 7.2 are significant after a Bonferroni correction is applied. There are 150 possible responses to the FCI; therefore, the correlation matrix contains $150 \cdot 149/2 = 11175$ unique entries (excluding the diagonal). If an $\alpha < 0.05$

significance threshold is used for a single statistical test, the Bonferroni correction adjusts this threshold to $\alpha_b = \alpha/11175 = 4.5 \times 10^{-6}$, a substantial correction. In general, larger samples will have smaller correlations pass this requirement. This can be seen by restricting Figure 7.2 to contain only combinations of $B$ and $D$ which pass the significance threshold with Bonferroni correction. Figure 7.3 shows the Bonferroni corrected significant correlations for Sample 1 ($N = 9606$), Figure 7.4 for Sample 4 ($N = 466$), and Figure 7.5 for Sample 5 ($N = 213$).

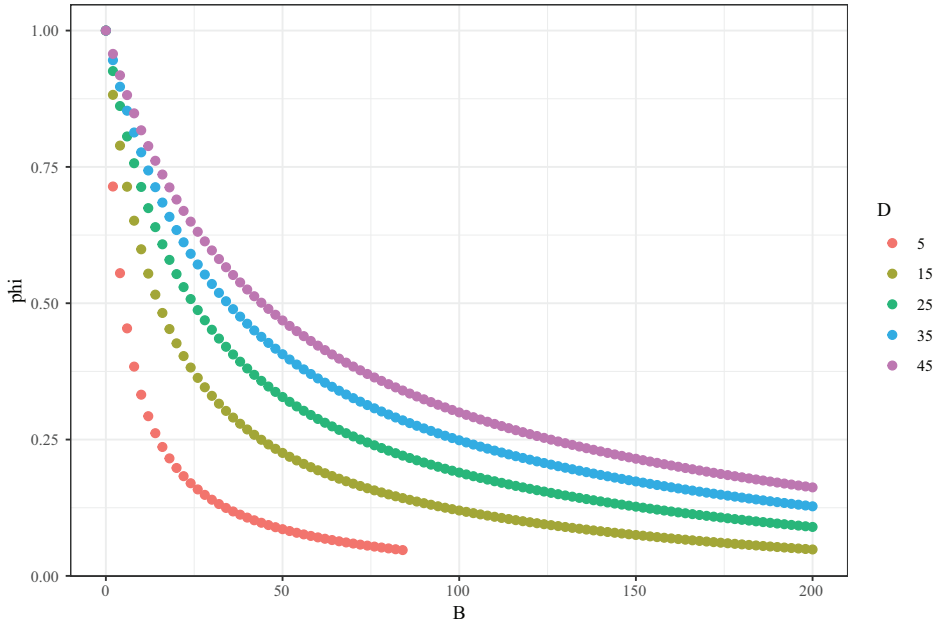


Figure 7.3: Plot of $\phi$ vs. $B$ for different levels of $D$ for Sample 1 applying the probability threshold with Bonferroni correction.

As one might expect, the much higher statistical power of Sample 1 allows the resolving of smaller occupation significant communities than in Sample 4 or 5. For Sample 4, the Bonferroni corrected probability threshold retains correlations of about 0.2 and as such has the same effect as the correlation threshold. This sample is near the minimum sample size for the adjustment of the correlation threshold to be productive and may explain some of

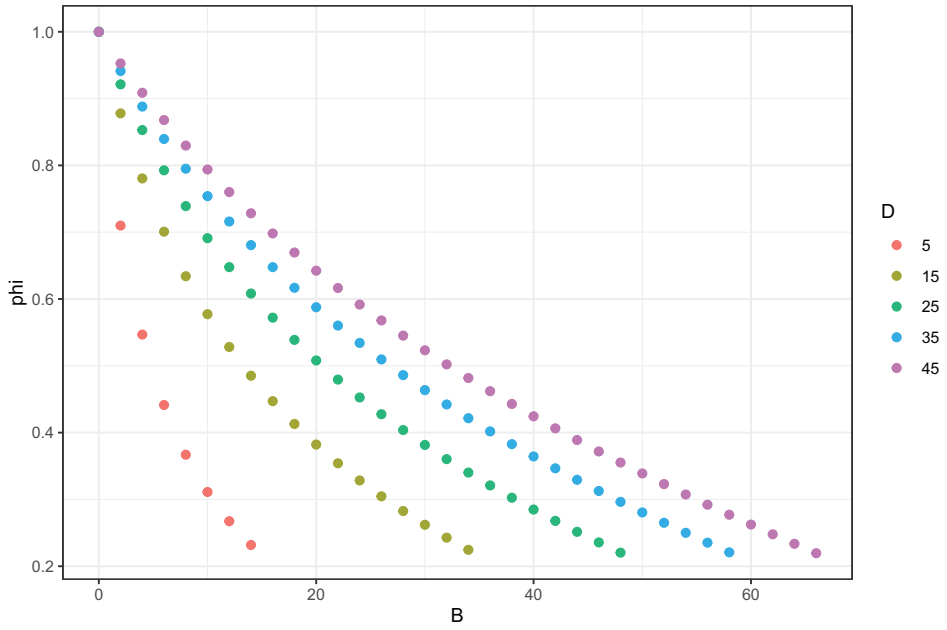


Figure 7.4: Plot of $\phi$ vs. $B$ for different levels of $D$ for Sample 4 applying the probability threshold with Bonferroni correction.

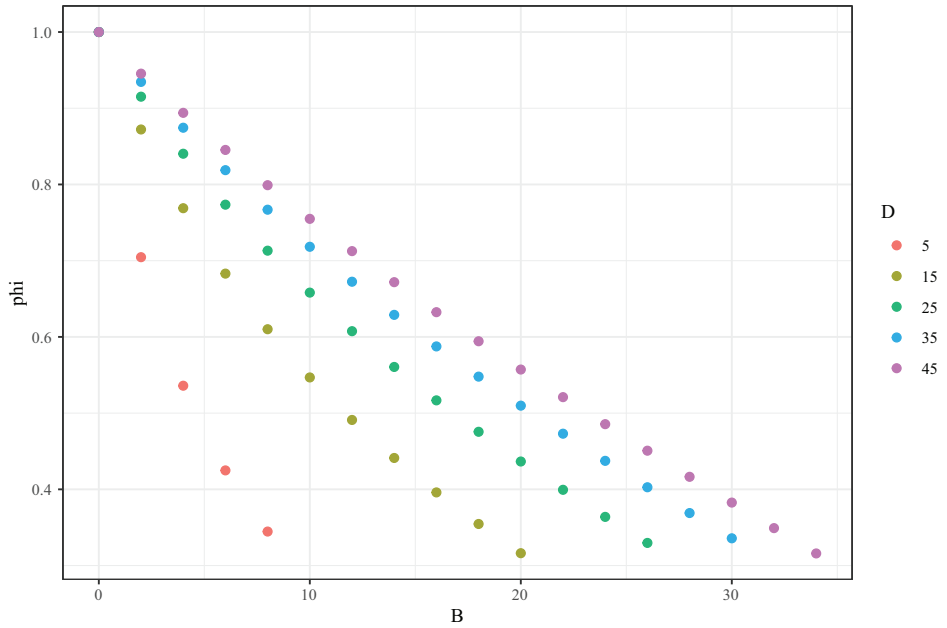


Figure 7.5: Plot of $\phi$ vs. $B$ for different levels of $D$ for Sample 5 applying the probability threshold with Bonferroni correction.

the unusual features of this sample's scree plots. For Sample 5, the Bonferroni corrected significance threshold removes $\phi < 0.35$ explaining why this samples scree plot was flat.

## 7.3 Correlation Threshold Plots

The figures which follow show the plots of average community size versus the number of communities used to select the correlation threshold for each network. The Sample 1 post-test graph is shown in the main paper.

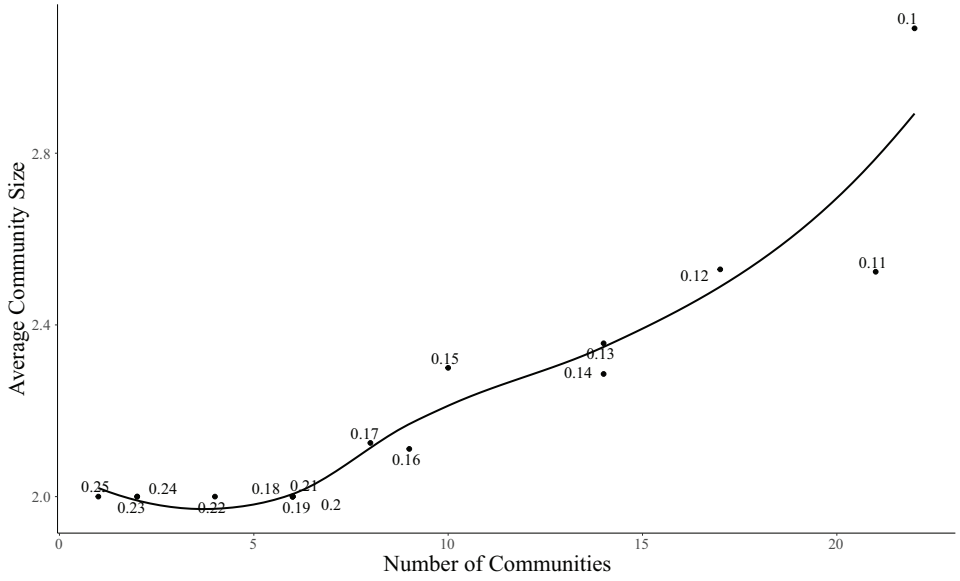


Figure 7.6: Plot used to determined the correlation threshold $r$ for the Sample 1 pretest network. Each point represents a network calculated at the labeled $r$ value.

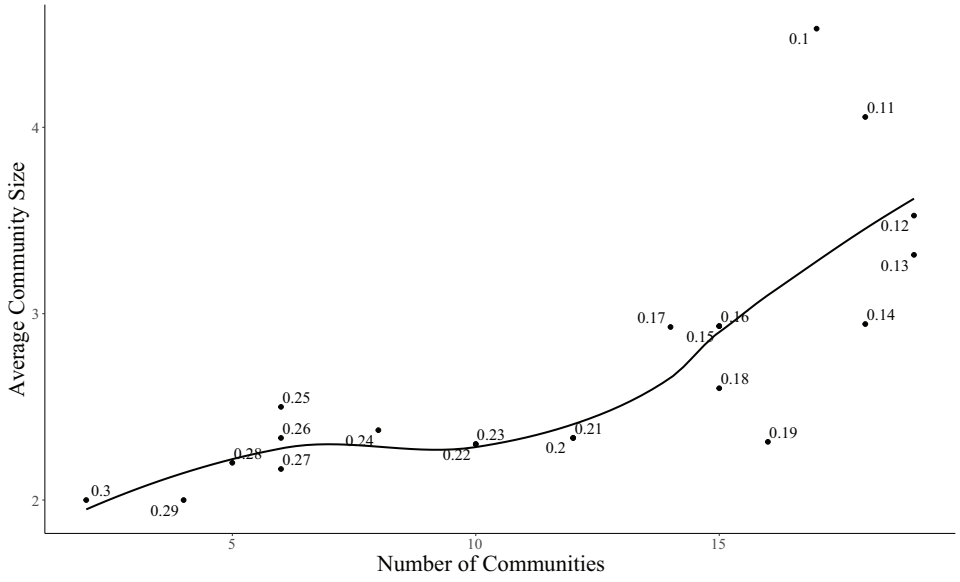


Figure 7.7: Plot used to determined the correlation threshold $r$ for the Sample 2 pretest network. Each point represents a network calculated at the labeled $r$ value.

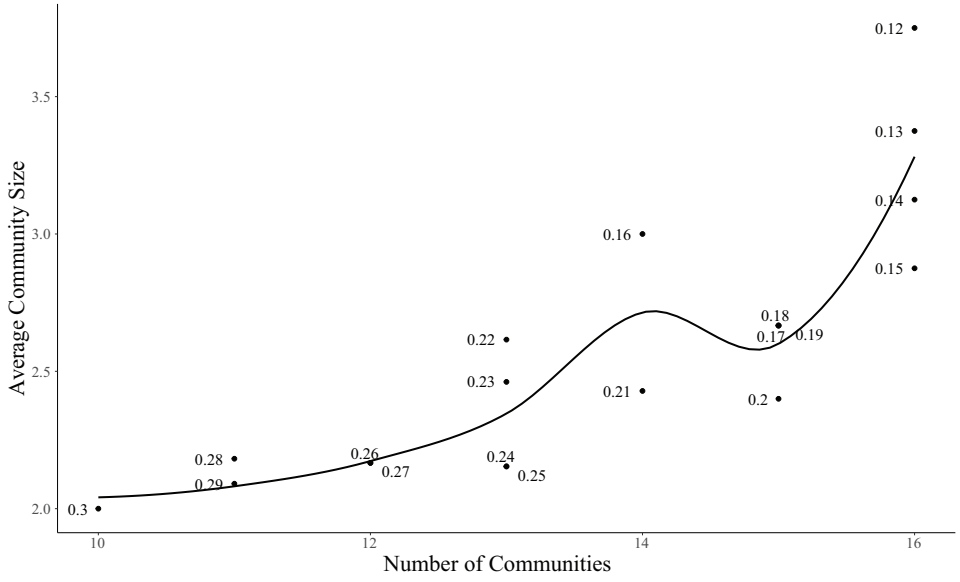


Figure 7.8: Plot used to determined the correlation threshold $r$ for the Sample 2 post-test network. Each point represents a network calculated at the labeled $r$ value.

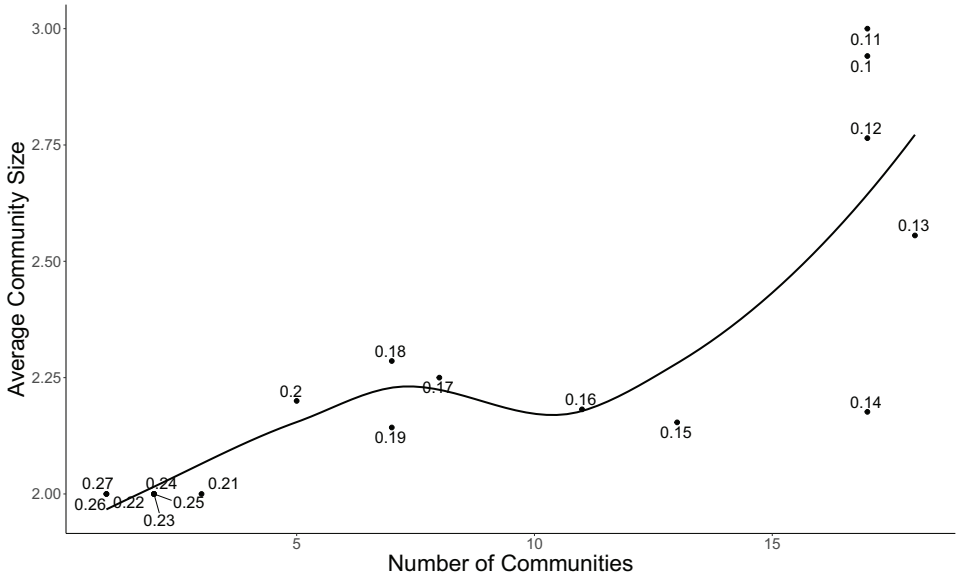


Figure 7.9: Plot used to determined the correlation threshold $r$ for the Sample 3 pretest network. Each point represents a network calculated at the labeled $r$ value.

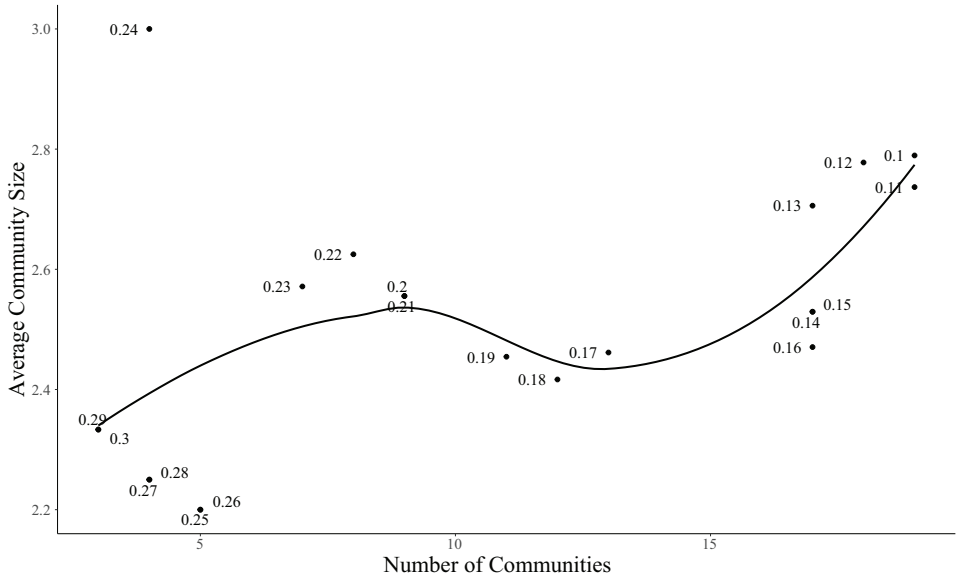


Figure 7.10: Plot used to determined the correlation threshold $r$ for the Sample 3 post-test network. Each point represents a network calculated at the labeled $r$ value.

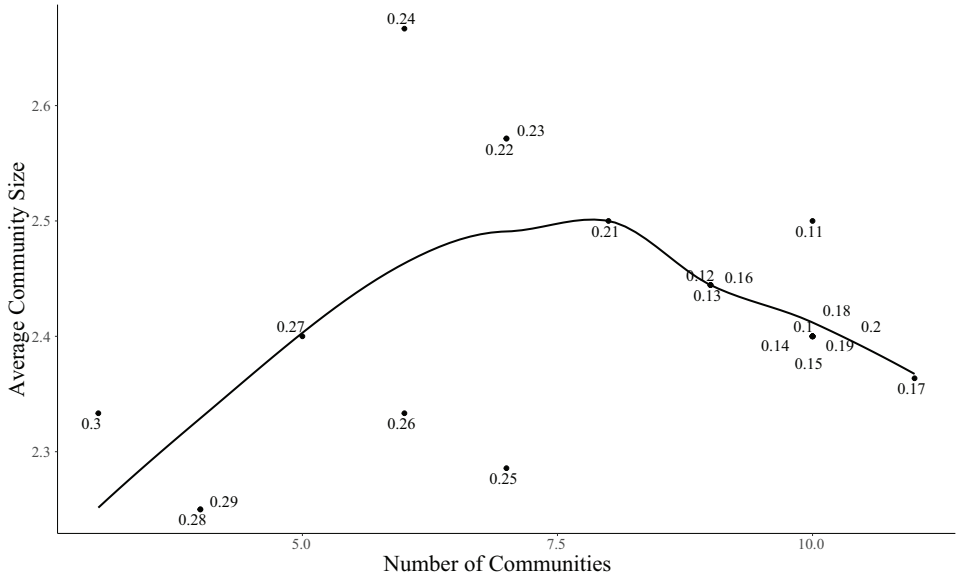


Figure 7.11: Plot used to determined the correlation threshold $r$ for the Sample 4 pretest network. Each point represents a network calculated at the labeled $r$ value.

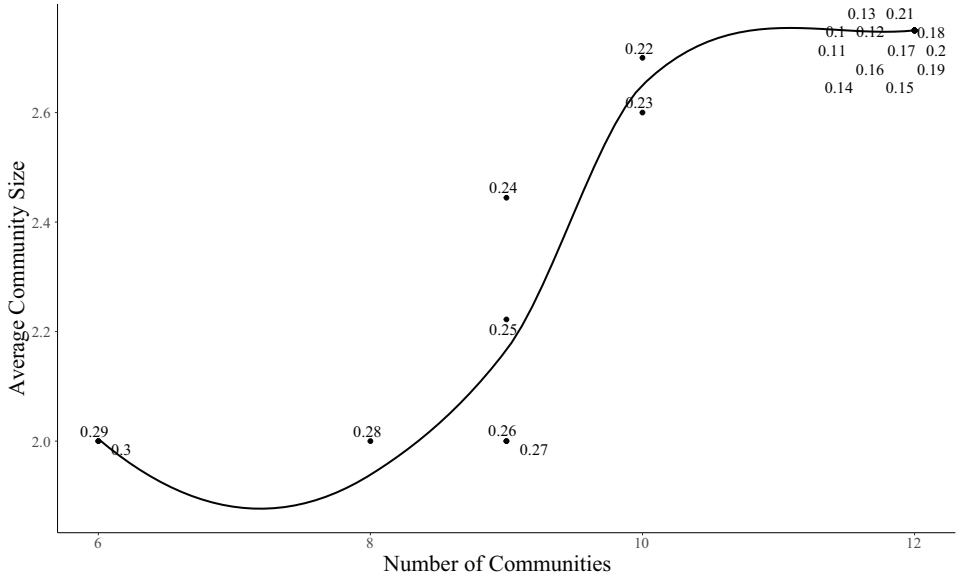


Figure 7.12: Plot used to determined the correlation threshold $r$ for the Sample 4 post-test network. Each point represents a network calculated at the labeled $r$ value.

## 7.4 Full Catalog of Communities

Table 7.13 shows all communities identified by MMA-P in Chapter 5.

Table 7.13: Communities of FCI responses identified in both the pretest and the post-test. Cells with the label × are sub-communities of a larger community or are found with a different edge structure, while cells labeled ⊗ are explicitly found in the network. Sample 1 is abbreviated as S1, Sample 2, etc. Responses that are separated by dashes are connected to each other, but not to other responses in the community, unlike responses that are in parenthesis, which are completely connected.

| Community | Pretest | | | | | Post-test | | | | | Explanation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 | S1 | S2 | S3 | S4 | S5 | |
| Completely Incorrect Communities | | | | | | | | | | | |
| 4A - 15C | | × | × | ⊗ | | × | × | × | × | | Newton's 3rd law misconceptions. |
| (4A, 15C, 28D) | | ⊗ | ⊗ | | | ⊗ | × | ⊗ | ⊗ | | Newton's 3rd law misconceptions. |
| 5C - 18C | | | | | | | | ⊗ | | | Motion implies active forces: centrifugal force. |
| 5D - 18D | | | | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | ⊗ | Motion implies active forces. |
| 5E - 18E | ⊗ | ⊗ | ⊗ | ⊗ | | ⊗ | ⊗ | ⊗ | | | Motion implies active forces: centrifugal force. |
| 6A - 7A | | ⊗ | | | ⊗ | ⊗ | | ⊗ | ⊗ | | Circular impetus. |
| 8E - 21A | | | | | | | | ⊗ | | | 8E and 21A share a similar trajectory. |
| 8A - 9B | ⊗ | ⊗ | ⊗ | ⊗ | | × | | × | × | | Blocked items: Last force to act determines motion. |
| 9B - (8A, 21B, 23C) | | | | | | ⊗ | | ⊗ | × | | 8A-9B: Blocked items. 21B-23C: Blocked Items.<br>Both: Last force to act determines motion |
| (9B, 8A, 21B, 23C) | | | | | | | | | ⊗ | | 8A-9B: Blocked items. 21B-23C: Blocked Items.<br>Both: Last force to act determines motion. |
| 11B - 29A | | ⊗ | | | | | | | | | Motion implies active forces. |
| 11C - 13C | | | | | | | ⊗ | | × | | Motion implies active forces. |
| 11C - 30E | | | | | | | | | × | ⊗ | Motion implies active forces. |
| 13C - 30E | | | ⊗ | | | | | | × | | Motion implies active forces. |
| 11C - 30E - 13C | | | | | | | | | ⊗ | | Motion implies active forces. |
| 15D - 16D | ⊗ | | ⊗ | | | | | | | | Newton's 3rd law misconceptions. |
| 17A - 25D | | | | | | ⊗ | ⊗ | ⊗ | | | Largest force determines motion. |
| 21B - 23C | ⊗ | | ⊗ | ⊗ | ⊗ | × | ⊗ | × | × | ⊗ | Blocked items: Last force to act determines motion. |
| 21C - 22A | | ⊗ | | | | | × | | | | Blocked items. |
| 23D - 24C | ⊗ | ⊗ | ⊗ | | | ⊗ | ⊗ | ⊗ | ⊗ | | Impetus dissipation. |
| Mixed Correct and Incorrect Communities | | | | | | | | | | | |
| 8B* - 21C | | | | | | ⊗ | × | × | × | | 8B* and 21C share a similar trajectory |
| 8B* - 21C - 22A | | | | | | | ⊗ | | | | 8B* and 21C share a similar trajectory.<br>22A blocked with 21C. |
| 8B* - 21C - 23B* | | | | | | | | ⊗ | × | | 8B* and 21C and 23B* share a similar trajectory.<br>21C-23B* Blocked items. |
| (8B*, 21C, 23B*) | | | | | | | | | ⊗ | | 8B* and 21C and 23B* share a similar trajectory.<br>21C-23B* Blocked items. |
| 21C - 23B* | | | ⊗ | | | | | × | × | | Blocked items: 21C and 23B* share a similar trajectory. |
| (21E*, 22B*, (15C), 4A, 28D) | | | | | | | ⊗ | | | | (4A, 15C, 28D): Newton's 3rd law misconceptions.<br>21E*-22B* Blocked items: 15C connection unknown. |
| Completely Correct Communities | | | | | | | | | | | |
| 4E* - 28E* | | × | × | × | | × | × | × | × | ⊗ | Newton's 3rd law. |
| 15A* - 28E* | ⊗ | × | × | × | | × | × | × | × | | Newton's 3rd law. |
| (4E*, 15A*, 28E*) | | ⊗ | ⊗ | ⊗ | | ⊗ | ⊗ | ⊗ | ⊗ | | Newton's 3rd law. |
| 5B* - 18B* | | ⊗ | | × | ⊗ | ⊗ | ⊗ | × | × | ⊗ | Centripetal acceleration in a curved trajectory. |
| (5B*, 18B*, 13D*) | | | | × | | | | ⊗ | ⊗ | | Motion under gravity;<br>a force in the direction of motion is not necessary. |
| 6B* - 7B* | | | | | | | ⊗ | | ⊗ | ⊗ | Instantaneous velocity is tangent to the trajectory. |
| 11D* - 13D* | | × | | × | | | ⊗ | | | | Motion under gravity;<br>a force in the direction of motion is not necessary. |
| 11D* - 13D* - 30C* | | ⊗ | | | | | | | | | Motion under gravity;<br>a force in the direction of motion is not necessary. |
| 11D* - 13D* - 5B* - 18B* | | | | ⊗ | | | | | | | Motion under gravity. |
| 17B* - 25C* | | × | ⊗ | × | | × | × | ⊗ | ⊗ | ⊗ | Newton's 1st law; Addition of forces. |
| 17B* - 25C* - 26E* | | × | | × | | ⊗ | ⊗ | | | | Newton's 1st and 2nd law; Addition of forces;<br>(26E*) 1D acceleration. |
| (17B*, 25C*, 26E*) | | ⊗ | | ⊗ | | | | | | | Newton's 1st and 2nd law; Addition of forces;<br>(26E*) 1D acceleration. |

# Chapter 8

## Social Network Analysis of West Virginia STEM Education Network *

*This chapter presents the work submitted for publication in Research in Higher Education. This work was constructed with collaborative efforts from Marjorie Darrah and John Stewart.

## 8.1 Introduction

The First2 Network is a collaboration of individuals from industry, higher education, K-12 schools, government, and federal research labs joined together for the purpose of increasing undergraduate STEM retention rates in West Virginia. This group has focused on classically underrepresented groups that are most prevalent in the state; rural and first-generation (FG) students. The First2 Network was formed with the intention of doubling the number of STEM graduates in West Virginia within 10 years by building sustainable collaborations that help support students in STEM as undergraduates and help connect them into the workforce.

Prior to the First2 Network, many groups within West Virginia were individually implementing programs to support student STEM persistence. The network formed because a greater impact could be achieved if each of these programs spread information and techniques to improve existing programs and implement new programs at more institutions. The network quickly expanded to include input from STEM-based industry members and from rural and FG STEM students.

Rural students who travel far from home for college are less likely to persist than non-rural students [156, 157]. These students often report feelings of isolation and homesickness when separated from the communities that they grew up in. These students may also value family commitment and community relationships more than their own achievement [158]. The First2 Network strives to build a sense of community for rural and FGS students within their institutions and within their STEM discipline.

The purpose of this chapter is to quantitatively assess the growth and development of

the First2 Network. The network has focused on building infrastructure to replicate, expand, and spread promising programs and practices across the state. We hope to understand the sustainability of these collaborations after the project ends. The following research questions were the primary focus of this research:

**RQ1** How is the overall structure of the network changing?

**RQ2** How are collaborative groups forming?

## 8.2 Background

### 8.2.1 Theoretical Framework

The First2 Network was formed with the intention of increasing student retention and persistence in STEM by involving rural and FG students in close-knit community groups that help develop their science identity and improve their self-efficacy. These methods are supported theoretically by both Tinto's Model of Student Departure [159, 160] and by Astin's Theory of Student Involvement [161].

Tinto states that the three primary reasons that students withdraw from an institution are difficulties becoming integrated into communities at the institution, difficulties with academics, and difficulties with career choice. The model emphasizes the importance of both formal and informal methods of student involvement in both academic systems and social systems for student retention. Tinto describes formal academic systems as those directly related to academic performance, informal academic systems as interactions between students and faculty/staff, formal social systems as any well-defined extracurricular activities, and informal social systems as interactions between students and their peer groups. According

to Tinto, students who have both social and academic connections are more likely to persist than those who do not.

In 1984, Astin performed a longitudinal study to determine the factors that were most related to student persistence. He found that students' level of involvement in their institution was not only related to persistence, but it was also linked to performance. Astin's Student Involvement theory was developed from this study. His theory states that the quantity and quality of energy that students invest into their college experience is directly proportional to the amount of personal development and learning a student experiences in college. Involvement is defined loosely as faculty/staff interactions and participation in academic work and extracurricular activities. Astin posits that student involvement should be the primary metric by which effective educational practices or policies are measured.

Both Tinto's Model of Student Departure and Astin's Theory of Student Involvement independently come to the conclusion that students who engage with extracurricular groups and on-campus activities are more likely to persist at their institutions than students who are more isolated. These models lay the groundwork for the current study by theoretically connecting student persistence with student connectedness. This work examines student-student groups forming as well as students' participation in the First2 Network as a whole. These models not only present the underlying argument for the existence of the First2 Network itself, but they also show the importance of the connections studied with network analysis.

It should be noted that both of these theories have been criticized for neglecting students' lived experiences and for encouraging students to separate themselves from past relationships in order to make room for new connections at their own institutions [162]. Tinto in

particular has been critiqued for drawing data from exclusively four-year, public institutions and generalizing it to all institutions, when it may not apply to smaller institutions or institutions that serve historically marginalized populations [163]. For our case however, recent research has shown that, for both rural and first-generation students, maintaining relationships between disconnected social networks (academic and hometown) allowed students to benefit from mentorship in both spheres [164]. Also, particularly for students who are both rural and first-generation, connecting students with professional mentors early on in their college career helps them better understand their interests and, subsequently, choose and persist in their major [164]. Tinto's and Astin's theories can be successfully applied in our context, then, assuming a rejection of the idea that students must be separated from past relationships.

### 8.2.2 Social Network Analysis

Social network analysis (SNA) describes the process of exploring social structures using principles and practices graph theory. A social network is composed of actors (nodes) who are connected through some kind of relationships (edges). SNA is often used to investigate the interchange of resources across a network of actors or to determine patterns of interactions apparent in the network [138]. SNA can also be used to measure the significance of individual network members on the size, number of connections, and information transfer within the network.

One of the benefits of SNA is the diversity of areas it can be used to explore. It has be used in many different research studies in academia alone. Lukacs and David considered how students' personal networks became unstable in the process of college transition [165].

They studied a group of Roman students and found significant differences between students in their reliance on certain groups in the process of academic adjustments. Eckles and Stradley determined relationships in a network by using archived data [166]. They found that the retention of students' friends had a greater impact on their retention than did the performance variables commonly believed to be associated with retention. Almeida *et al.* utilized SNA to study social capital in first-generation students' academic success [167]. They learned that, for this set of students, social capital with faculty and staff predicted grade point average. Poldin *et al.* studied how the achievements of students are influenced by the achievements of peers in their social network [168]. They discovered that this peer influence happens chiefly through relationships, such as study partners that share knowledge, and not as much through mere friendship connections. Another group that studied peer networks found that peer quality improves student performance and that the breadth and cohesion of a student's network positively affects a student's outcomes [169].

González Canché [170] has recently advanced the use of SNA in education research by bridging geographical and social network analysis to statistically model structures with education data. He has also shown how to reveal meaningful structure in qualitative data with these methods. González Canché and Rios-Aguilar applied SNA to institutional data from Calizona Community College to study the effects of peers and credit attainment on underrepresented minority students in community colleges [171]. They found that male Latino and male African American students benefited from interacting with peers in the same racial/ethnic group with higher amounts of credits accumulated. There was no effect found based on the variation of peers' credit attainment for female Hispanic or African American students.

## 8.3 Methods

### 8.3.1 Data Collection

A survey was developed internally by the research team of the First2 Network in order to study the composition of the network each year. The survey was assessed for its face validity by the First2 Network's research team and the leadership team before revision and redistribution to the entire network. Surveys were distributed by email to any individual who signed up on the First2 Network's website. Data were collected for five consecutive years, from 2018 to 2022. Surveys were completed online through the Qualtrics survey application. An IRB (institutional review board) approved consent form was provided at the beginning of the survey. Respondents then submitted basic demographic information such as name, organization, and role (student, faculty, administrator, etc.). Lastly, they were asked to name other individuals that they collaborated with on projects related to the First2 Network. Given that "collaboration" can mean very different things to different people, and the degree to which two people collaborate on projects differ, a numeric classification for levels of collaboration was provided to help participants understand the meaning of each term. The following scale developed by Hogue *et al.* [172] and Borden and Perkins [173] was used;

1. Networking: Aware of Organization, Little Communication, Loosely Defined Roles, Independent Decision Making

2. Cooperation: Share information, Formal Communication, Somewhat Defined Roles, Independent Decision Making

3. Coordination: Share Information Frequently, Defined Roles, Some Shared Decision Making,

4. Coalition: Frequent Communication, Shared Resources, Shared Decision Making

5. Collaboration: Frequent Communication, Shared Resources, and Mutual Trust. Coordination on Most or All Decision Making.

Over the five years of data collection, 249 individuals either responded or were named by someone who responded to the survey. One limitation of social network data collected through virtual surveys is that responding to the survey is entirely voluntary and moderately time consuming. In order to overcome this limitation, we decided to include any individual named by any respondent in the analysis as network members even if they did not fill out the survey themselves.

Network members consisted of undergraduate students, graduate students, student advisors, industry members, K-12 teachers, university staff members, faculty members, researchers, and administrators. The organizations that these network members collaborated with were also very diverse including colleges and universities, companies, state-level educational agencies, county school systems, nonprofits, and state research organizations. In addition to the survey data, student persistence and GPA data were collected from the largest institution in the First2 Network. A list of publications related to the First2 Network was also collected to form a publication network.

### 8.3.2 Network Construction

To explore the structure of connections between First2 Network members, survey responses were converted into an adjacency matrix. The adjacency matrix turns network members into nodes and the connections between members of the network into edges. The reported strength of collaboration forms the edge weight in the network. Individual cells in the adjacency matrix, $a_{ij}$, represent a weighted edge between two collaborators $i$ and $j$. If two survey respondents name each other in the survey, but they respond with a different strength of connection, then the average reported strength is used in both cell $a_{ii}$ and cell $a_{ji}$, so the resulting matrix is square and symmetric, and the corresponding graph is undirected. Consequently, if one person names a network member that did not fill out the survey, or did not name that person in their survey, then the reported strength of connection is averaged with 0. This allows for network members who did not fill out the survey to still be accounted for in the analysis, but with a lesser weight than two people who both filled out the survey. Using the exact reported strength of connection would create a directed adjacency matrix and would have the potential to reveal more salient and complex information about the network structure but would require a higher survey response rate for the direction of the edges to be meaningful.

Another network was constructed from the list of academic publications written about the First2 Network. Individuals who appeared in one or more publications were included as nodes and edges were formed between two individuals if they were coauthors on at least one of the papers. This resulted in an undirected graph. Edges were weighted by the number of papers two individuals collaborated on.

### 8.3.3 Network Statistics

Yearly network statistics were calculated after the adjacency matrix was formed. The number of nodes (people), the number of edges (collaborative connections), the graph density (the ratio of the number of edges in the network to the maximum possible number of edges), and the number of surveys completed was calculated.

In network analysis, a network is composed of a set of nodes and edges with any number of configurations. When nodes or groups of connected nodes are not connected to other nodes in the network at all, they form isolated "islands." These islands are called components. The component with the largest number of nodes is called the giant component. Many network statistics, primarily ones related to a distance across the network, are calculated over the giant component rather than over the entire network itself. This is because distances are measured by the number and weight of edges across a network, and, given that no edges are connecting smaller components to the main component, the distance between components is undefined. Any metric calculated for the giant component can be calculated for smaller components as well, but for networks where the giant component is significantly larger than other components in the network, these metrics are usually just calculated for the giant component. The size of the giant component and the number of components are both reported for this work. For a network changing in time, these two statistics can provide an indication of the growth of the main body of the network and how/if the network is splintering into smaller groups.

Two different centrality metrics were calculated in order to measure the connectedness of individuals to the larger network. These two metrics are strength (weighted degree)

and betweenness centrality. Strength is a local centrality measure that shows how closely connected a node is to its immediate surroundings within a network. Strength is found by adding the weight of each edge connected to an individual node. Betweenness centrality is a global centrality measure that accounts for a node's position relative to each other nodes' positions in the network. Betweenness centrality for a node v is found by computing the shortest path between each pair of nodes, finding the fraction of shortest paths that include the node v for each node pair, then summing this fraction over all node pairs [174]. Betweenness centrality scales with the number of node pairs in a network, so the statistic is normalized by dividing by the number of node pairs not including v itself; $\frac{(N-1)(N-2)}{2}$ for undirected networks. Including both a local and a global measure should capture most of the relevant information about the significance of each node in the structure of the network. Each node has its own strength and betweenness centrality and these are reported for the top actors in the network. The average strength and betweenness centrality are also computed for the entire network. A person with a higher betweenness centrality is someone who is the bridge between unconnected network members, whereas a person with a higher strength is someone who simply has many connections. People with higher strength tend to have higher betweenness centralities, and vice versa.

To study the similarity of the First2 Network over the years, the coverage index (CI) was calculated for the overlap in nodes each year. The CI quantifies this overlap by taking the ratio of the intersection of the set of nodes in two different years with the total number of nodes in each year. The intersection of two sets of nodes, set A and set B, is composed of the nodes that are in both set A and set B. The ratio for the CI involves the total number of nodes in each year; as such, two different coverage indexes are calculated for each two-year

pair:

$$CI_A = \frac{N(A \cap B)}{N(A)} \tag{8.1}$$

$$CI_B = \frac{N(A \cup B)}{N(B)} \tag{8.2}$$

where N (X) is the size of set X. A coverage index plot uses Equation 8.1 for the value below the diagonal and Equation 8.2 for the value above the diagonal. In other words, the lower diagonal represents the yearly overlap in people relative to the earlier year, while the upper diagonal represents the overlap in people relative to the later year. The CI is presented to give a natural degree of commonality between the years. The R package corrplot [175] was used to show yearly coverage indexes.

Maximal clique analysis was used to determine the close-knit groups forming in the network. A clique is a group of directly connected individuals within the larger graph, such that each pair of individuals in the clique has an edge connecting them. A clique is maximal if it includes the largest subgroup of individuals where everyone in the subgroup is connected. People can belong to more than one clique. The Bron and Kerbosch Algorithm [176] can be used to find all maximal cliques of each possible size along with the cliques' members. A clique strictly measures groups of individuals that each reference each other as a collaborator. Group structures are often more complicated than this. For example, an institution could have a committee of ten faculty members tasked with assessing the institution's social climate. In a group of that size, there is a reasonable chance that not every member of the group would collaborate with every other member, and they would not

be counted as a group in a clique analysis, even though they could meaningfully be defined as a collaborative group.

A community detection algorithm (CDA) was used to determine the grouping structure of the network beyond the strictly defined, well connected cliques. A community is a set of nodes such that pairs of nodes in the set are more likely to be connected if they are both members of the same community than if they were members of different communities. Since the clique analysis allowed for overlap, a CDA was chosen that did not allow for overlap in communities to determine if the network divided naturally into groups that were more connected internally than they were connected externally. The communities were then analyzed for similarities between network members. The fast-greedy CDA was used for this work [177]. This algorithm directly optimizes a modularity score, a score that represents how well a network has been divided into communities. The algorithm can be represented by a dendrogram (tree plot) where each level indicates a different number of possible communities. The modularity is calculated at each level in the dendrogram, and the maximum is then found by the peak in the modularity graph [97].

## 8.4 Results

In this section, we will examine the First2 Network's evolution using several social network analysis tools.

### 8.4.1 Network Structure and Evolution

Table 8.1 shows several statistics associated with the First2 Network. The number of individuals responding to the survey increased every year. From the survey responses,

people were added to the First2 Network in two ways, either they filled out the survey, or they were named by someone who filled it out. From the table, it can be seen that the number of connected members (nodes) also grew each year and likewise the total number of members to date continued to grow. From 2018 to 2021, the network was growing at relatively commensurate rates in all metrics, but in 2022 the number of connections (edges) drastically decreased. In 2022, network members were reporting fewer connections with lower collaboration levels relative to previous years. The average strength of network members is one of the most robust ways to test the connectivity of a network because it is less susceptible to the $N^2$ effect of edge density calculations, where smaller networks tend to have higher densities. The density, the ratio between the actual number of edges and the total possible number of edges, is included to illustrate this point.

Table 8.1: Network Statistics

| Statistic | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|
| Survey Responses | 25 | 30 | 44 | 62 | 83 |
| Nodes (Active Members) | 48 | 67 | 81 | 105 | 122 |
| Total Members to Date | 48 | 85 | 127 | 182 | 249 |
| Edges | 146 | 183 | 215 | 304 | 211 |
| Density | 0.129 | 0.083 | 0.066 | 0.056 | 0.029 |
| Average Strength | 10.85 | 10.57 | 11.62 | 12.50 | 7.00 |
| Average Betweenness | 0.041 | 0.037 | 0.038 | 0.031 | 0.032 |
| Giant Component | 48 | 67 | 75 | 96 | 102 |
| Number of Components | 1 | 1 | 4 | 4 | 11 |
| Number of Communities | 6 | 6 | 6 | 7 | 9 |

From 2018 through 2021, the evolution of the First2 Network was dominated by student growth, which can be seen in Figure 8.1. This changed in 2022 when the number of students in the network marginally increased, but the number of other types of network members substantially increased. The growth of active network members has been relatively constant, with an average of about 19 new members joining the First2 Network each year. Each year,

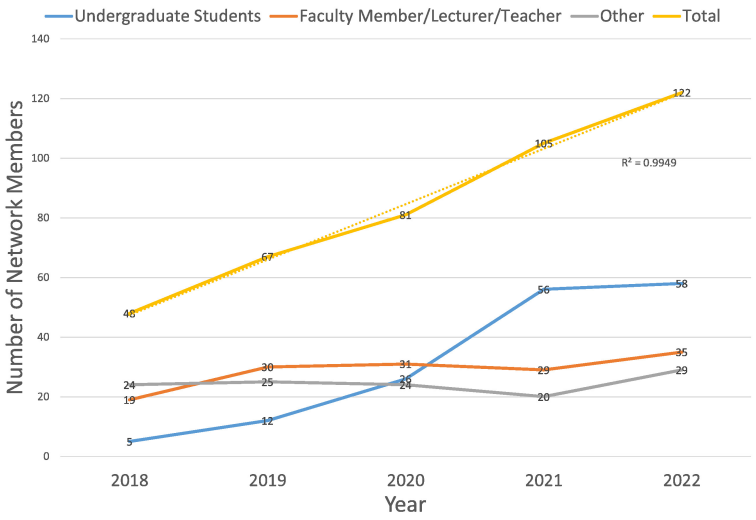


Figure 8.1: Network by Category.

new members were added while older members left or disengaged. Once a person was an active part of the network and a number was assigned to them, they kept this number even if they became inactive and were not part of the node count in subsequent years. From 2019 to 2021 the number of non-student active network members changed negligibly. However, many non-student network members were joining or leaving the network, leading to an equilibrium of active non-student members.

Figure 8.2 shows the First2 Network colored by members' self-reported role. The position of each node in Figure 8.2 (and in all future network graphs) is related to the strength of connection to other nodes in the network. The Fruchterman-Reingold force-directed layout algorithm [154] is utilized to place nodes closer together that have higher edge weights. Survey respondents are prompted by the question "Which of the following roles most accurately describes your role in the network?" They were provided with many options for their potential role, but for visualization purposes, the options in the survey were combined to the four roles displayed in Figure 8.2. Network members sometimes reported different

roles from year to year, depending on which role most accurately described them that year. For example, for the first three years, node 26 was part of the "Government/Industry contact role" because they worked for a state research organization, but by 2021 they identified as a Faculty Member/Lecturer/Teacher, and by 2022 they were put in the other category because they reported that they were now an administrator.

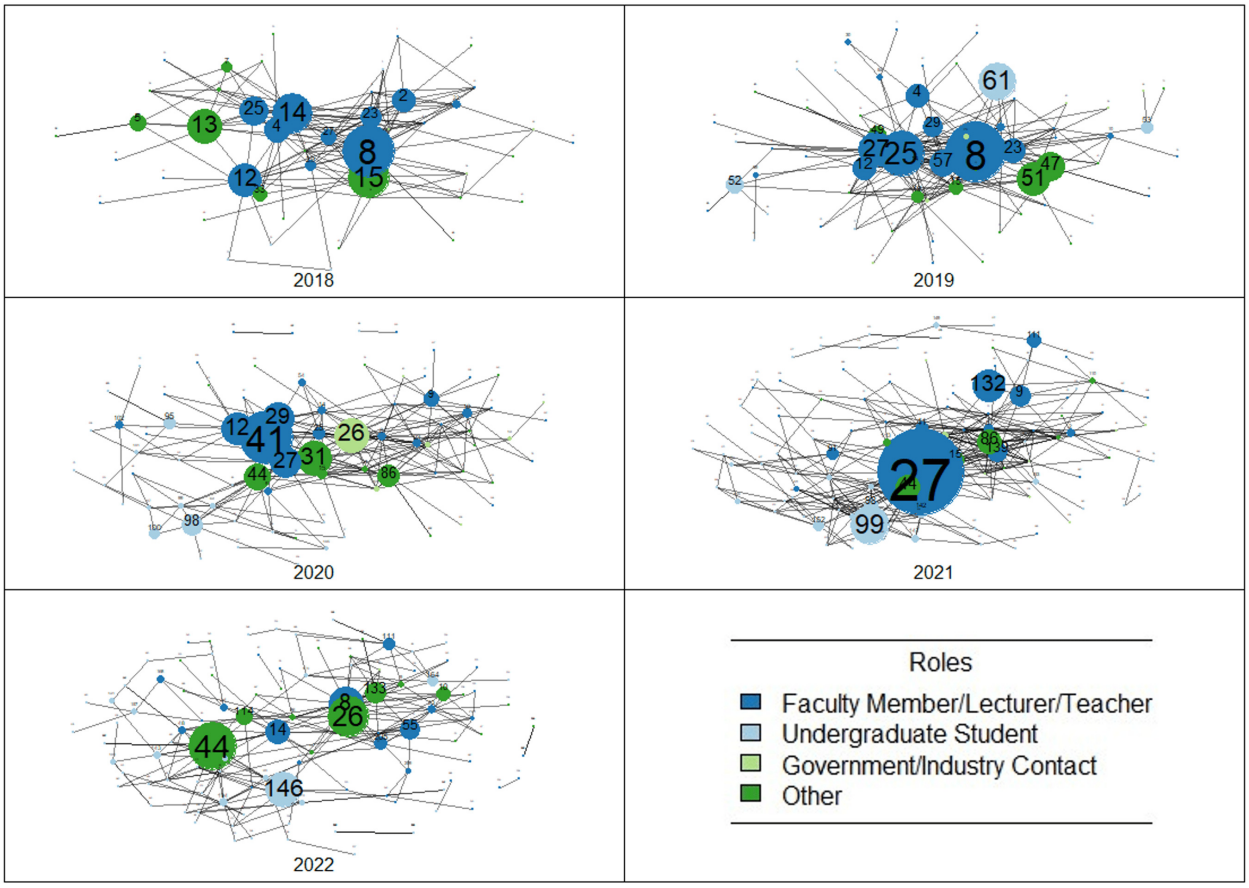


Figure 8.2: Network by role, sized by strength.

In Figure 8.2 nodes are sized proportional to network members' strength of connections, but the scaling factor is small to show the change in the network relative to the most prominent figures. These graphs make it clear that although students make up the dominant growth in the network in terms of new members, the large-scale structure of the network is dominated by non-student members. These networks also show some of the structural

changes that occur when very connected network members leave the network. The loss of members like 27 and 82 could, to some extent, help account for the decreasing network metrics in 2022.

Table 8.2: Yearly change in strength and betweenness of top actors in the network.

| ID | 2018 | | | | 2019 | | | | 2020 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S | SR | B | BR | S | SR | B | BR | S | SR | B | BR |
| 8 | 53.0 | 1 | 0.261 | 1 | 56.5 | 1 | 0.302 | 1 | 58.0 | 1 | 0.046 | 18 |
| 27 | 17.5 | 11 | 0.075 | 11 | 46.0 | 2 | 0.177 | 4 | 54.5 | 2 | 0.156 | 6 |
| 26 | 26.5 | 5 | 0.006 | 20 | 31.0 | 8 | 0.027 | 23 | 52.0 | 3 | 0.019 | 30 |
| 23 | 36.0 | 3 | 0.103 | 9 | 45.0 | 4 | 0.123 | 8 | 42.0 | 5 | 0.047 | 16 |
| 82 | NA | NA | NA | NA | NA | NA | NA | NA | 38.0 | 6 | 0.028 | 25 |
| ID | 2021 | | | | 2022 | | | | | | | |
| | S | SR | B | BR | S | SR | B | BR | | | | |
| 8 | 65.5 | 2 | 0.050 | 14 | 53.0 | 1 | 0.17 | 4 | | | | |
| 27 | 76.0 | 1 | 0.428 | 1 | NA | NA | NA | NA | | | | |
| 26 | 55.0 | 3 | 0.019 | 30 | 30.0 | 4 | 0.208 | 2 | | | | |
| 23 | 45.5 | 4 | 0.000 | 60 | 36.0 | 2 | 0.042 | 14 | | | | |
| 82 | 40.0 | 6 | 0.001 | 57 | NA | NA | NA | NA | | | | |

Table 8.2 illustrates the impact on network structure that members 27 and 82 had relative to other top network members. S stands for strength, SR for strength rank, B for betweenness, and BR for betweenness rank. Strength rank and betweenness rank represent the rank of an individual for a given centrality metric for a given year. For example, SR of 1 means that the individual had the highest strength for the entire year; BR of 5 means that individual had the fifth highest betweenness value for that year. The other top network members, 8, 26, and 23 were chosen by their strength rank; from 2018-2022 they consistently had some of the highest ranks, usually top 5. Top network members were chosen by SR and not BR because the top members by betweenness had more year-to-year variance. Table 8.2 makes clear that individuals who control the most unique information flow between connected groups (high BR) are often not the same individuals who simply have the most

connections (SR). Cells in Table 8.2 are labeled NA when a network member was not in the First2 Network that year, either because they had not joined yet, or because they left the network entirely.

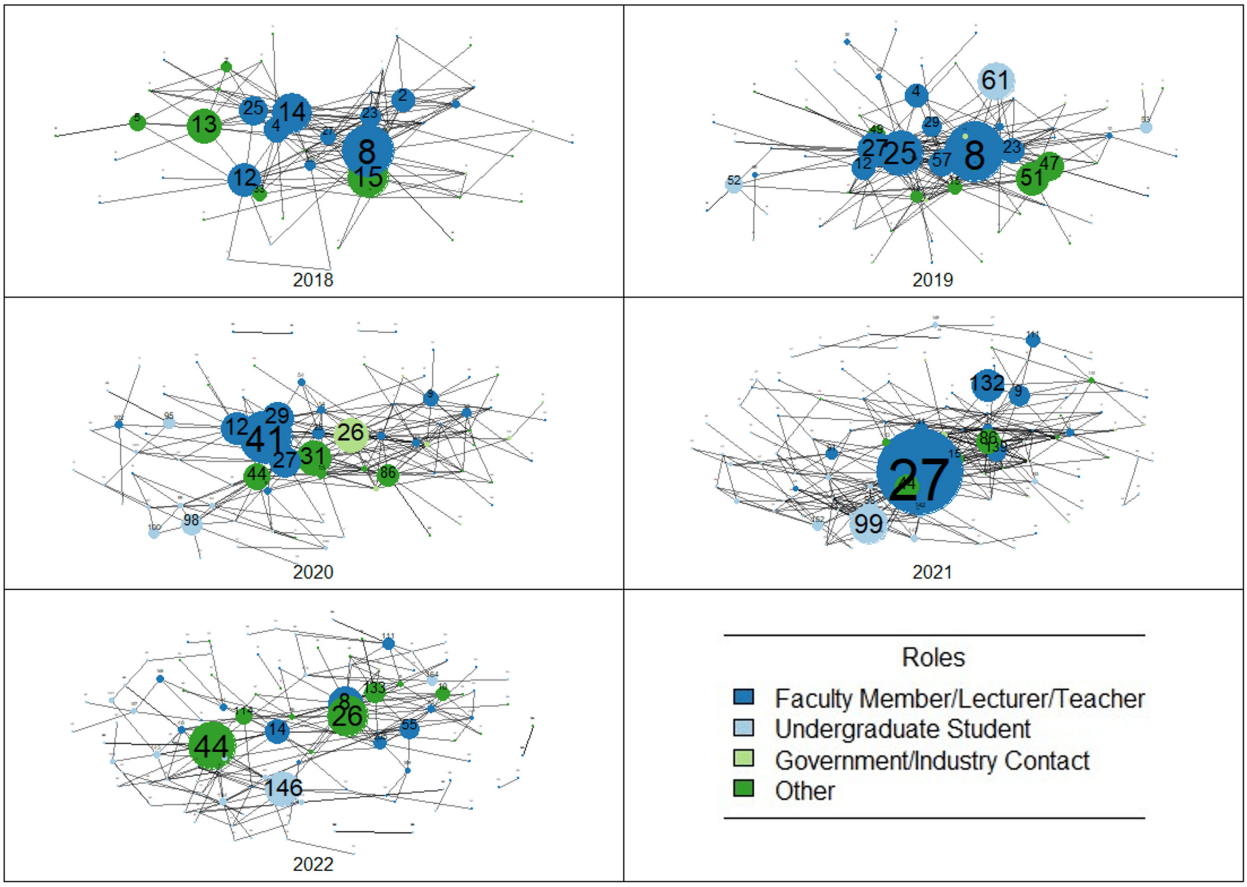


Figure 8.3: Network by role, sized by betweenness.

Figure 8.3 is colored the same as Figure 8.2, but it is sized by betweenness to show network members that are pivotal in retaining the structure of the network. Losing a network member with high betweenness is likely to split the network into smaller components, taking much of the possible information transfer with them. This graph clearly shows that the network is held together by members with various roles. Figure 8.2 emphasizes the dominance of faculty members in terms of the total number of connections in the network. Figure 8.3 shows that students, government/industry contacts, and other members (i.e., administrators

and non-profit employees) keep the network together as generally as a single component rather than separate, internally well-connected components. First2 Network members with high betweenness act as major information distributors within the network. For example, students like 61 in 2019, 99 in 2021, and 146 in 2022 are connected to many other students who are not connected to anyone else, so they seem to distribute information to these otherwise unconnected students.

Figure 8.4 displays the coverage index, the portion of overlap in specific First2 Network members between each year. The color scale and circle completeness both represent the same metric; a full circle represents complete overlap, while an empty space represents no overlap. The circles above the diagonal represent the overlap in network members divided by the later year, Equation 8.1, while circles below the diagonal represent the overlap in network members divided by the earlier year, Equation 8.2. A network growing in members will necessarily have more complete circles in the lower diagonal than in the upper diagonal. For example, $CI_{2018}$ is the overlap in members between 2018 and 2019 divided by the number of members in 2018. This comparison is shown in the second row, first column in Figure 8.4. In a growing network then, this will be a more complete circle than $CI_{2019}$, the overlap in members between 2018 and 2019 divided by members in 2019, which is found in the cell in the first row, second column in Figure 8.4.

Figure 8.4 quantifies the rate of individuals joining or leaving the network. The composition of the network changes by approximately 30-50% every year. Over 50% of individuals from each previous year remain in the network. Over time this flow of network members leads to the network changing drastically, with only about 35% of the individuals that were in the network in 2018 remaining in the network in 2022 (this is represented by the bottom

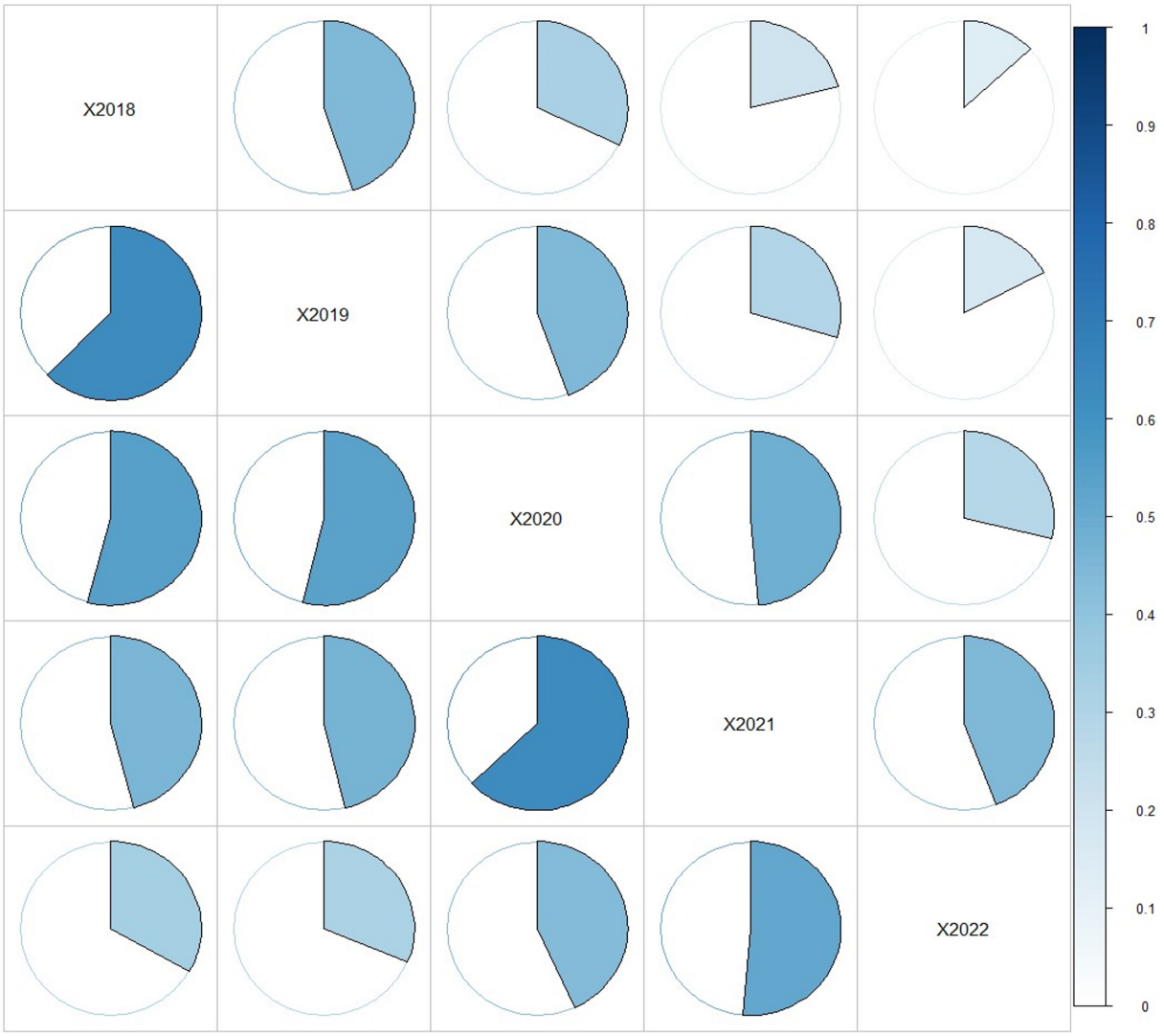


Figure 8.4: For years $i$ and $j$ with $i < j$, the plot above the diagonal represents $\text{CI}_i = N(X_i \cap X_j)/N(X_j)$ and the plot below the diagonal $\text{CI}_j = N(X_i \cap X_j)/N(X_i)$ where $X$ is the set of actors and the function N() computes the size of the set.

left-most circle). The change in network members is directly proportional to the change in network connections; when network members leave, their connections are taken with them.

Table 8.1 shows that the giant component size of the network was identical to the total network size until 2020. In 2020 and 2021 the network split into four components. In 2020 all three of the new components were of size two, while in 2021 one component was size four, one was size three, and one was size two. In 2022 many isolated sub-groups split off from

the main network, resulting in 11 components, all of which, outside of the giant component, were of size two. The number of components and the size of the giant component relative to the number of nodes in the network give some indication of the possibility of information flow in a network. Individuals who are part of the isolated components are less likely to be a part of, or even know about, many of the First2 Network's activities and events.

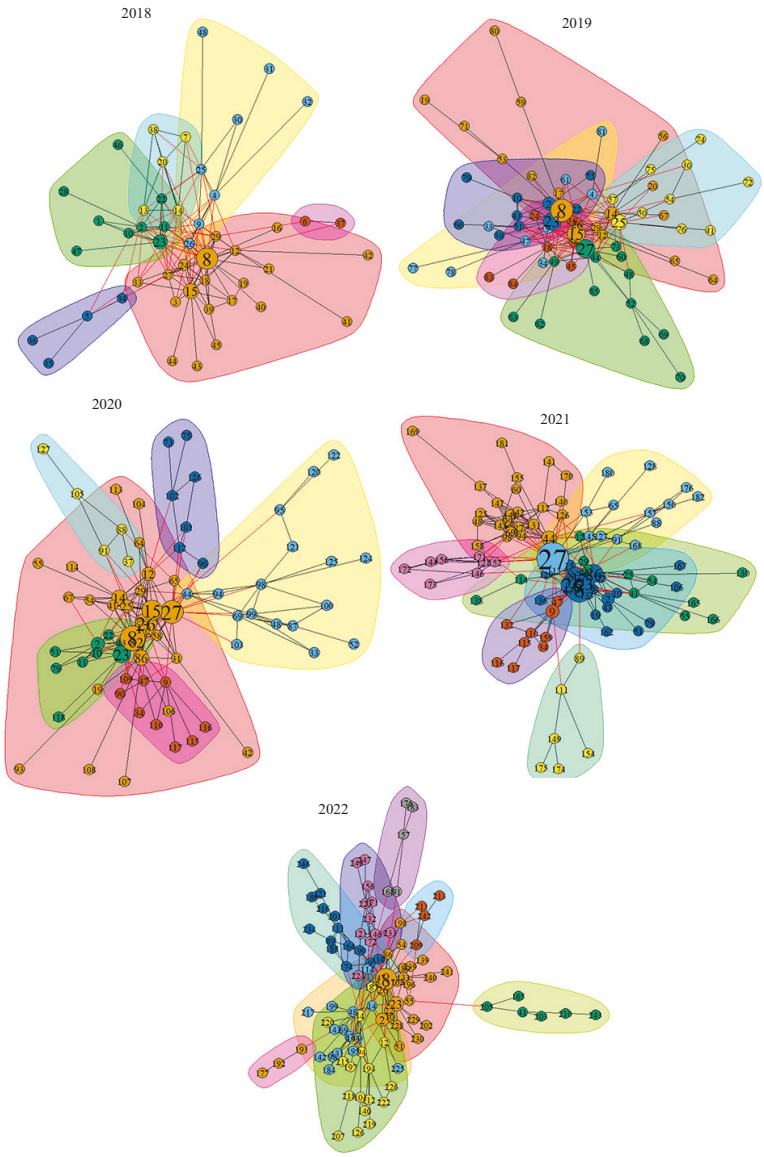


Figure 8.5: Communities identified in each year's giant component.

The giant component and the communities resulting from the application of the fast-

greedy community detection algorithm are shown in Figure 8.5. Nodes are colored by community membership and a shaded region is included over each community. Nodes are roughly sized by strength, with a minimum size threshold to make node color visible. Shaded regions overlap because of the force-directed graph node placement algorithm, but community membership does not actually overlap. Edges between communities are colored red and edges within communities are colored black. The CDA was applied only to the giant component because the maximum size of any other component in any year was four, so each component was its own community, and the community detection provided no new information for these smaller components. Communities that resulted from the application of the CDA to the networks were analyzed for institutional homogeneity. This gives a broad picture of the network, indicating which years had greater levels of inter-institutional collaboration and which years had collaborations forming within individual institutions. Cross-institutional collaboration should be indicative of information transfer between different programs and collaboration within a single institution could be indicative of institutions attempting to implement strategies and programs developed by the First2 Network.

In 2018, out of the six communities identified, the four smallest were almost entirely institutionally homogeneous, while the two largest communities had very little overlap in institutional membership. Out of the six communities in 2019, the smallest two communities were mostly composed of members from one or two institutions, while the other four communities were much more heterogeneous than 2018. In 2020, except for the largest community, the others were institutionally homogeneous. In 2021 the largest community, shaded red and colored orange, actually included members mostly from one institution, but was less central in the network than other smaller communities that were much more institutionally diverse.

In 2022, all eight of the communities, except for the largest and most central one, were each composed of members from single institutions. In the years 2018, 2020, and 2022, the CDA identified groups of individuals from the same institution that were more connected to each other than to other institutions. In 2019 and 2021, the grouping structures identified by CDA seemed to be much less dependent on the community member's institution, indicating a higher level of cross-institutional collaboration, and information transfer, during those two years.

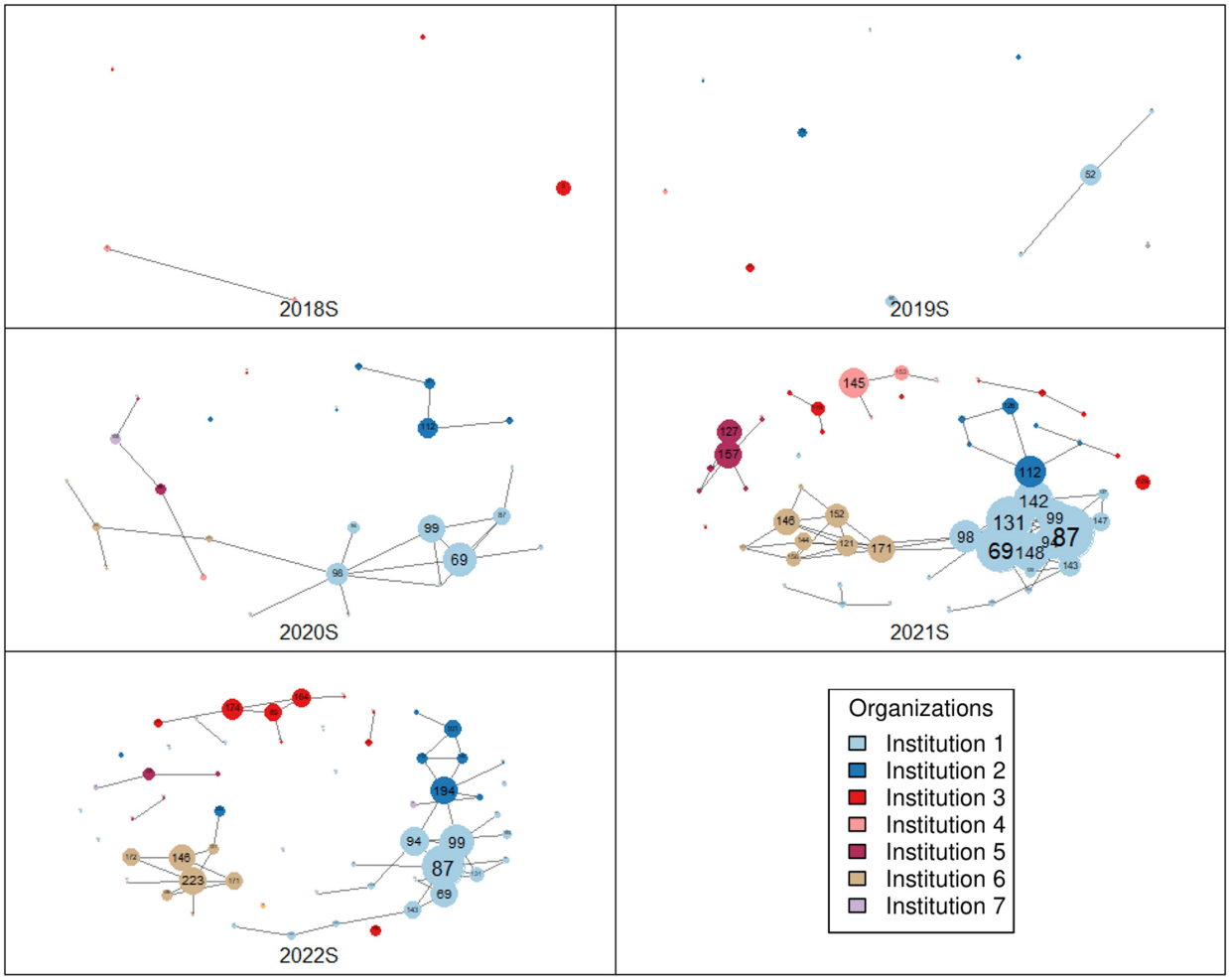


Figure 8.6: Graph of student to student connections.

Figure 8.6 shows connections between student members in the First2 Network from

2018 to 2022. Students who were only connected to non-student members are included as isolated nodes. Node size is proportional to the student's strength in Figure 8.2 to show which students were more connected to other students and which were more connected to non-student network members. For example, node 3 in 2018 is larger than the nodes in 2018 that actually have connections because node 3 is connected to more non-student members than the other students in 2018. Early in the project, student connections relied heavily on connections to faculty members. In 2020 and beyond, students seemed to form larger intra-institutional groups. The large drop in connectivity evident in the total network in 2022 is less evident here. It seems that student connections in 2022 either decreased or increased by institution, with the light blue and burgundy institutions having smaller student subgroups, and the red institution having larger student subgroups. On average there is still a decrease in average student connectivity, which can be seen in the number of tiny, isolated nodes in 2022 compared to 2021. These individuals are connected only to faculty members.

### 8.4.2 Faculty Collaboration and Productivity

Maximal clique analysis was applied as another way to examine the structure of the network to quantify the number of closely-knit collaboration groups. These close-knit groups consisted mainly of faculty members and, in the early years, were developing around the working group structures that had been set up by the Network. Table 8.3 grew in number and average size from 2018 through 2021, showing that more people were working closely in larger groups. It should be noted that the number of cliques can be larger than the number of nodes because they represent different combinations of connected nodes. The difference between this and the community analysis is that in clique analysis strictly includes fully

connected groups, where every member in the group has an edge between them, whereas communities are loosely connected by a closer proximity each other than those outside the group.

Table 8.3: Clique Structure. The yearly clique structure for cliques of size 3 or greater are included.

| Clique Size | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|
| 3 | 195 | 242 | 269 | 403 | 99 |
| 4 | 152 | 215 | 243 | 383 | 24 |
| 5 | 162 | 111 | 140 | 241 | 2 |
| 6 | 10 | 29 | 46 | 92 | 0 |
| 7 | 0 | 3 | 7 | 20 | 0 |
| 8 | 0 | 0 | 0 | 2 | 0 |
| Mean | 3.73 | 3.89 | 3.98 | 4.08 | 3.22 |

In 2018, the largest size cliques were of 10 cliques of size six. Each of these largest cliques were composed of individuals from different organizations, with either six, five, or three different organizations represented. In 2019, the largest cliques were 3 cliques of size seven and 29 cliques of size six. These cliques were made up of First2 Network members from seven different organizations. In 2020, the number of cliques of size six and seven increased to 46 and 7 respectively. These cliques contained First2 Network members from seven organizations. These cliques were mostly made up of faculty from different colleges and universities but also contained staff of research organizations or state education organizations. The pandemic years did not have much effect on the growth of the cliques, since most groups met on an online platform already because they were from all over the state. In 2021, there were 92 size six cliques, 20 size seven cliques, and 2 size eight cliques. The largest cliques in 2021 were again made up of five and six organizations. In 2022, the Network structure changed, and the number and size of cliques dropped off dramatically. The largest were 2 size five cliques. These two cliques were composed of only project leadership

and not composed of faculty at different institutions. The size five cliques went from 241 in 2021 to 2 in 2022.

An alternate measure of the connectivity of the First2 Network was developed by examining the network of academic publications, see Figure 8.7. This graph represents 25 publications and 44 authors, with 30 faculty and graduate students from the First2 Network and the other 14 faculty and graduate students from outside the network. There are 148 edges, indicating a minimum of 148 instances of publication-based collaboration. In the figure, the First2 Network members are colored in blue, while non-network members are in red. The nodes represent people, and the connections represent whether they are coauthors on a paper. The thickness of the edges corresponds to the number of co-authorships shared between a pair of nodes. The size of each node corresponds to the number of publications of each individual.

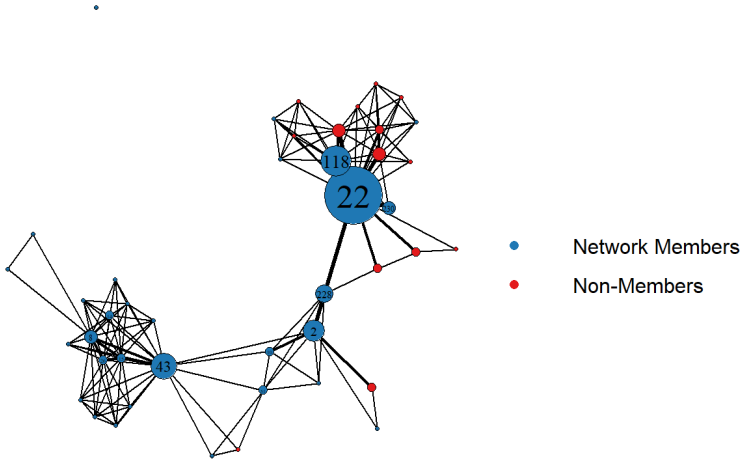


Figure 8.7: Network of publications.

The publication network naturally divides itself into several distinct clusters. The clusters are also connected to other clusters, showing that some members work with different groups. There is one isolated node, indicating a sole author publication. The non-members were typically graduate students who were working in a group with a faculty member who was part of the First2 Network. These members also show that there exist individuals outside of the network that are influencing the network. Non-network members participating in publications using data from the network or from students who participate in the network are influencing the structure of the main network by influencing network leaders' policy decisions with their data analysis. It is clear that some of the members of the First2 Network (22, 43, 118) are very productive (have many publications) and that they are working with others in the network to coauthor papers. The numbers on these nodes correspond to the numbers in the overall network graph, so when looking at the nodes in the publication network we can see that node 8 and node 26 also appear in Table 8.2, indicating that they were both also prominent members of the main network.

Precisely quantifying the edge overlap between the publication network and the main set of networks is difficult because the main network breaks down connections by year while the publication network includes papers published any year during the network's existence. However, any edge that exists in the publication network between two network members also exists at some point in time in the main network because publishing is a type of collaboration. Figure 8.7 is not included to show unique connections between faculty members not captured through the main network, it is included to reveal the structure of a single type of connection within the network. Publications represent particularly active network members who are collaborating through the peer-reviewed publication process rather than through informal

networking or formal committees within the First2 Network.

For the individuals with the highest degree in the network of publications, we looked at their strength and betweenness scores in the First2 Network in 2022. For the top ten individuals in the network of publications, seven were members of the First2 Network in 2022. The others had either left the First2 Network by 2022, did not take the survey, or were not named. The average strength for this group was 25 and the average betweenness score for this group was 582. This was much higher than the overall average in 2022, which was a strength of 7 and an average betweenness of 186. This shows that these individuals in the network of publications are highly connected and influential in the main Network.

## 8.5 Discussion

This section will answer the two main questions posed in the introduction.

Research Question 1 – *How is the overall structure of the network changing?* The overall structure of the First2 Network changed from year to year but had stable growth with more people joining than leaving each year. The number of connections and the average strength grew each year from 2018 to 2021; overall, network members were gaining connections and/or strength of connections. There was a reduction of both the number of edges and strength in 2022, which could be a result of changing the structure of the working groups. The giant component grew every year, which implies that network members were not isolated in small groups but were more centralized. This could be due to the First2 Network holding bi-annual conferences for all members and monthly meetings for leadership.

In 2022, the last year of a five-year grant, the leadership of the First2 Network started

focusing on the sustainability of changes at each organization. They discontinued most working groups and changed the focus and funding to institutional teams at each college and university. They also had leadership teams established around different areas such as sustainability and improvement science to support institutional team research efforts. Until 2022, collaboration was growing among network members and between network organizations, but due to some conscious policy changes, the inter-institutional working groups were replaced by institutional teams and thus the collaboration between institutions was reduced in 2022. The First2 Network made a conscious decision to change the way groups worked together in 2022. They moved away from working groups around project goals, where working group chairs were receiving funding, to working groups within institutions, where institutions were getting funding, and the groups focused on sustaining the work at their institution beyond the initial funding of the grant. These decisions had an impact on the structure of the network. These changes strongly affected the community structure, the number and sizes of the cliques, and the overall density of the network.

The First2 Network composition was dominated by student growth from 2018 until 2021. In 2022, network growth was dominated by non-student members; student numbers remained relatively constant. In the first few years of the network, student connections were primarily to faculty members. However, since 2020, robust groups of student clusters formed, largely within individual institutions, mainly due to practices that the network put in place including student campus clubs and student leadership groups. The drop off in network size and connectivity in 2022 did not impact student clustering as much, likely because these connections were made within individual institutions.

Although students played a major role in increasing the number of yearly First2 Net-

work members, non-student members were consistently the most connected and influential members of the network. Information flow in the network was not necessarily dominated by the most connected members, but by a mix of well-connected members and members who acted as bridges between well connected groups in the network. Students, faculty members, government/industry contacts, and other roles like administrators or K-12 educators were all found as important bridges connecting less connected groups in the network. Regardless, when particularly connected members left the network, their absence was clearly represented in the network structure. The loss of a few very connected members from 2021 to 2022 likely had at least some effect on the decrease of network statistics in 2022.

One interesting aspect of network change that is not obvious from the structure alone is that network members join and leave each year. The total composition of the network changes by about 30-50% each year. By 2022, only 35% of First2 Network members remained that started in the network in 2018, but the total number of First2 Network members increased every year. This highlights the flexibility of many of the interactions and characteristic roles that First2 Network members take; collaboration still increased even when many First2 Network members were replaced each year.

Research Question 2 – *How are collaborative groups forming?* The communities resulting from the application of the community detection algorithm revealed some interesting features in the First2 Network. In general, the largest, most connected communities included the most connected individuals in the network and included a diverse number of First2 Network members with many different roles and from many different institutions. The smaller communities were less centrally connected to the network and were most frequently composed of First2 Network members from one or two institutions. The CDA identified a natural

grouping structure in 2018, 2020, and 2022, where, more often than not, individuals in the same community were from the same institution, whereas in 2019 and 2021 that grouping structure appears much less frequently. Grouping by institution is reasonable in 2018 because the network was so new that most individuals that worked together already knew each other from their own institutions. The less central, more institutional grouping structure of 2020 could be seen as an effect of COVID, where there was some sense of necessity in institutional grouping to deal with the pandemic. In 2022 this grouping structure could be attributed to the development of institutional teams, which encouraged working within the institution to implement some of the strategies and programs developed by the First2 Network in the previous four years.

From the clique analysis perspective, from 2018 through 2021, there were working groups formed around the goals of the project including faculty/student engagement, industry connections, student readiness for college, and summer immersive research experiences to come up with ideas to make improvements in education in the state and in different colleges and universities. During these years, the First2 Network stressed working groups around the core goals of the network which were encouraged to develop best practices to share across organizations. From 2018 to 2021, the largest groups were composed of inter-organizational groups. However, in 2022, most of the fully connected groups were smaller and everyone in the group was from the same organization, reflecting the intentional change in working group structure made by the project leaders.

## 8.6 Conclusion

In this study, data were collected from 249 individuals over a five-year period measuring the network connections forming within the First2 Network project to promote the retention of West Virginia STEM majors. Social network analysis demonstrated that the structure of the network changed in time. The growth of the network in its first four years was relatively consistent; connections increased in number and strength, fully-connected groups increased in number and size, and key leaders appeared over the years to disseminate information and collaborate with many others. In the fifth year of the project, leaders began to transition the project to sustainability within state academic organizations rather than disseminating information across the network as in the first four years. The structure of the network reflected these changes with the number of edges, density, average strength, average betweenness, and number and size of cliques all decreasing. The number of components also increased, indicating a greater level of fracturing from the giant component of the network.

During the first four years, the project focused efforts on Networked Improvement Communities to implement, study, and revise replicable best practices for programs relevant to the First2 Network. Investing time and effort into these groups formed valuable social relationships that enhanced many of the desired outcomes of the project, particularly statewide connections between university faculty, students, and industry members. The First2 Network provided online working groups to foster collaboration anywhere in the state, as well as in-person conferences and leadership meetings to disseminate internal research about successful avenues for increasing student STEM retention. These meetings and groups led to a network of publications, increasing faculty productivity and collaboration with individuals

outside of the network as well.

The First2 Network provided multiple ways for students to be involved and to take leadership positions. From the beginning, student were encouraged to voice their opinions as equal network participants. This led to robust clusters of student to student connections in the network, with student leaders standing out as central network members that connected many otherwise unconnected students to the core of the network.

# Chapter 9

## Comparing introductory undergraduate physics learning and behavior before and after the COVID-19 pandemic*

*This chapter presents the work published in Physical Review Physics Education Research [178]. This work was constructed with collaborative efforts from Amanda Nemeth and John Stewart. This work was supported in part by the National Science Foundation under Grants No. ECR-1561517, No. HRD1834569, and No. DUE-1833694 and by a grant from the Howard Hughes Medical Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 9.1 Introduction

Many studies have explored the effects of the COVID-19 pandemic and the rapid transition to virtual modes of education on student attitudes and learning during this period of online instruction [179–181]. Now that most universities in the United States (US) have returned to in-person classes, the effects of this period of disruption on the return to in-person classes can be measured. This study explores differences in student behavior and achievement between the last two fully face-to-face semesters of a university physics class prior to the pandemic and the first two fully face-to-face semesters after the pandemic.

A recent study at a highly selective West Coast university in the US found that there was no evidence for a reduction in high-school physics learning after the pandemic using a physics diagnostic exam administered in the Fall 2019 and Fall 2021 semesters [182]. This institution is situated in a state with one of the highest per capita incomes in the US and with a high rate of residents with a bachelors degrees. The current study examines a broader collection of student achievement and behavior measures at an institution admitting students with lower levels of high school achievement than the West Coast university. This institution accepts 90% of its applicants. The institution is the flag-ship state university in a small eastern state with a state population with per capita income and rate of bachelors attainment among the lowest in the US. As such, it can provide context of the effect of the pandemic on the education of students coming from less resourced school systems.

## 9.2 Methods

### 9.2.1 Sample

Data were collected from the introductory calculus-based electricity and magnetism course at a large eastern US land-grant university with a total undergraduate enrollment in Fall 2021 of 19,600 students [129]. The demographics of the undergraduate population were 81% White, 4% Black or African American, 4% Hispanic or Latino, 3% non-resident alien, 6% two or more races, with other groups 2% or less. Student ACT scores ranged from 21 to 27 for the 25th to the 75th percentile. Pell grants are given to lower socioeconomic status students (SES) and are often used to measure the percentage of low SES students at a university: 23% of the undergraduate population was Pell eligible. The Fall 2021 enrollment was smaller than that of Fall 2019 when the university enrolled 21,000 students. The overall demographics of the Fall 2019 undergraduate population were 80% White, 4% Black or African American, 4% Hispanic/Latino, 5% non-resident alien, 4% two or more races, with other groups 2% or less.

The course studied enrolled primarily scientists and engineers. The course was taught in multiple lecture sections which were overseen by the same lead instructor for the entire period studied. This instructor had been managing the course for many years and oversaw general course content, homework assignments, tests, and the management of the laboratory segment of the class. The class had been offered in the same format for many years before the pandemic and returned to this format after the pandemic. Alternate modes of instruction were provided during the pandemic; these semesters are not considered in this study. As such the course represents an excellent laboratory to study changes before and after the

pandemic. This study focuses on the introductory electricity and magnetism course because the introductory mechanics course retained some pandemic course policy changes preventing comparison.

The class was presented with three 50-minute lecture sessions along with one 170-minute laboratory session per week. Each lecture session enrolled over 100 students. Both the lecture and laboratory utilized multiple active learning strategies. The lectures implemented Peer Instruction using clickers [10], while the labs featured a mixture of conceptual whiteboard questions, hands on inquiry activities, group problems, and traditional experiments. Two homework sets were collected each week; these were collected at the beginning of lecture and were turned in on paper. The first homework collected on Monday of each week, called the "short homework" in this study, consisted of ten multiple-choice questions. The second homework collected on Wednesday each week, called the "long homework" in this study, consisted of five multiple-choice questions and four open-response questions. Four tests and a final exam were given over the course of the semester. Conceptual learning was monitored by applying the Conceptual Survey of Electricity and Magnetism (CSEM) [13] as a pretest and post-test. The class was primarily taken by sophomores, who would have spent much of their freshman year in college and the last part of their senior year in high school receiving virtual instruction.

The total enrollment for the four semesters was $N = 1033$ (Spring 2019 $N = 217$, Fall 2019 $N = 327$, Fall 2021 $N = 327$, Spring 2022 $N = 162$). Assignment scores and lecture attendance were accessed from the course learning management system; students who withdraw from the class are automatically removed for course records. As such, the analysis, except the DFW percentage, includes only students who completed the class for a

grade.

### 9.2.2 Mann-Whitney $U$ test

Often in PER studies, t-tests are used to compare means; however, the t-test assumes the sample is normally distributed. Many of the quantities examined in this work (homework scores, lecture attendance rates, etc.) have distributions which are substantially non-normal. As such, this work applies the non-parametric Mann-Whitney $U$ test. The test calculates $U$ which is related to the sum of the ranks for one of the groups. An effect size, $r$, can be calculated from $U$; Cohen's criteria for $r$ are that 0.10 is a small effect, 0.30 a medium effect, and 0.50 a large effect [183, 29].

The Mann-Whitney $U$ test, sometimes referred to as the Wilcoxon rank-sum test, is a rank-sum test used to determine whether the total ranks of two independent groups significantly differ. This is achieved by combining the two groups and rank ordering the scores in numerical order. If the two groups are randomly distributed in rank, then the two samples do not differ statistically. The $U$ test statistic is calculated using Equation 9.1.

$$U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - \sum R_i \tag{9.1}$$

where for each group ($i = 1$ or 2), $U_i$ is the test statistic, $n_i$ is the sample size, and $\sum R_i$ is the sum of ranks. The smaller value of $U_i$ is reported as the test statistic [184, 183]. If both sample sizes $n_1$ and $n_2$ are small enough, significance is evaluated using the table provided by Milton [185] showing the critical values for the Mann-Whitney Two-Sample statistic. If $n_1$ or $n_2$ exceed these values, as they do in the present study, a large sample approximation

may be computed to obtain an effect size. The effect size for $U$ is $r$ [186] which can be calculated using Equation 9.2.

$$r = \frac{|z|}{\sqrt{n}} \tag{9.2}$$

where $z$ is the $z$-score of group $i$ and $n = n_1 + n_2$ [183, 186]. To obtain the z-score, one must first find the mean

$$\mu_U = \frac{n_1 n_2}{2} \tag{9.3}$$

and the standard deviation

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{9.4}$$

The $z$-score used in the effect size calculation is found as follows:

$$z = \frac{U_i - \mu_U}{\sigma_U} \tag{9.5}$$

where $U_i$ is the reported test statistic, $\mu_U$ is the mean, and $\sigma_U$ is the standard deviation calculated in Equations 9.3 and 9.4 [184, 183].

### 9.2.3 Two proportions $z$-test

Some quantities examined were semester-level frequencies such as the DFW percentage (percentage of students of earning a grade of D or F or withdrawing from the class); these were compared using the two proportions $z$-test. The two proportions $z$-test, which is equivalent to a chi-squared test for the equality of two proportions, is a test for checking if the difference between two proportions is statistically significant.

In order to apply a significance test to a binary distribution without a standard de-

viation (such as DFW rates), another difference in means method must be used. The two proportions $z$-test, which is equivalent to a chi-squared test for the equality of two proportions, is a test for checking if the difference between two proportions is statistically significant. Equation 9.6 shows the two proportions $z$-test (labeled as $\chi^2$ to differentiate it from Equation 9.5),

$$\chi^2 = \frac{P_1 - P_2}{\sqrt{\frac{P(1-P)}{N_1} + \frac{P(1-P)}{N_2}}} \tag{9.6}$$

where $P_1$ is the proportion of "successes" for group 1, $N_1$ is the sample size of group 1, and P is the pooled proportion, defined in Equation 9.7.

$$P = \frac{P_1 + P_2}{N_1 + N_2} \tag{9.7}$$

The corresponding effect size used with this significance test is Cohen's $h$, which can be calculated as follows using Equation 9.8 [29].

$$h = 2|\arcsin\sqrt{P_1} - \arcsin\sqrt{P_2}| \tag{9.8}$$

where $P_1$ and $P_2$ are the same proportions described in Equation 9.6. Cohen's $h$ measures the distance between two proportions and has effect size criteria: 0.2 corresponds to a small effect, 0.5 to a medium effect, and 0.8 to a large effect.

### 9.2.4 Holm–Bonferroni correction

This work applies many statistical tests and is, therefore, susceptible to the inflation of Type I error. The Holm–Bonferroni correction for the significance level is applied [187].

This method orders the $p$-values from smallest to largest, then progressively adjusts the significance level. If there are $m$ statistical tests, the significance threshold, $\alpha$, is adjusted to $\alpha/m$ for the smallest $p$ value, $\alpha/(m-1)$ for the second smallest, etc. The null hypothesis is rejected for all $p$ greater than the first $p$ which fails the test. This method provides the same Type I error correction as the Bonferroni correction with less risk of Type II error.

## 9.3 Results

Assignment scores and submission rates were compared between the last two completed semesters preceding the COVID-19 pandemic shutdowns (Spring and Fall 2019) and the first two semesters after in-person courses were resumed (Fall 2021 and Spring 2022). The assignment submission rate is the percentage of the assignments submitted for grading. The assignment percentage score is the average score on the assignment (zero if not submitted). The two fall semesters were compared against each other pre- and post-pandemic; as were the two spring semesters. The class studied has historically observed a substantial difference in student performance between the spring and fall semesters which is likely the result of differences in the high school preparation of the students in these semesters. For the class studied, students in the fall semester are "on-sequence" pursuing the plan of study suggested by the university's 4-year degree plans; these students were ready to enroll in Calculus 1 their first semester in college; largely students in the spring semester were not.

A summary of general descriptive statistics is shown in Table 9.1. The results are reported by semester with Fall 2019 abbreviated F19. Most quantities in Table 9.1 did not significantly change between pre- and post-pandemic semesters after applying the Holm-

| Variable | Semester | $N$ | $M \pm SD$ | $U$ | $p$ | $z$ | $r$ |
|---|---|---|---|---|---|---|---|
| Test Average | F19 | 307 | $76.1 \pm 14$ | 49217 | 0.188 | 1.32 | 0.05 |
| | F21 | 302 | $74.2 \pm 15$ | | | | |
| | S19 | 206 | $71.6 \pm 16$ | 15472 | 0.645 | 0.46 | 0.02 |
| | S22 | 146 | $70.3 \pm 18$ | | | | |
| High-school GPA | F19 | 298 | $3.89 \pm 0.45$ | 33781 | **0.000** | -4.94 | 0.20 |
| | F21 | 296 | $4.07 \pm 0.39$ | | | | |
| | S19 | 198 | $3.76 \pm 0.46$ | 10008 | **0.000** | -3.99 | 0.22 |
| | S22 | 136 | $3.96 \pm 0.43$ | | | | |
| Lecture Attendance Percentage | F19 | 307 | $82.6 \pm 23$ | 42690 | 0.082 | -1.74 | 0.07 |
| | F21 | 302 | $86.0 \pm 21$ | | | | |
| | S19 | 206 | $84.5 \pm 25$ | 17736 | **0.003** | 2.97 | 0.16 |
| | S22 | 146 | $78.0 \pm 27$ | | | | |
| ACT/SAT Mathematics Percentile Score | F19 | 285 | $84.9 \pm 15$ | 45610 | 0.046 | 2.00 | 0.08 |
| | F21 | 292 | $85.2 \pm 11$ | | | | |
| | S19 | 183 | $79.9 \pm 17$ | 11486 | 0.855 | 0.18 | 0.01 |
| | S22 | 124 | $79.0 \pm 18$ | | | | |
| Short Homework Percentage Score | F19 | 307 | $67.6 \pm 24$ | 42458 | 0.072 | -1.80 | 0.07 |
| | F21 | 302 | $71.3 \pm 22$ | | | | |
| | S19 | 206 | $67.7 \pm 25$ | 15165 | 0.893 | 0.14 | 0.01 |
| | S22 | 146 | $68.1 \pm 25$ | | | | |
| Long Homework Percentage Score | F19 | 307 | $62.4 \pm 23$ | 38187 | **0.000** | -3.76 | 0.15 |
| | F21 | 302 | $68.8 \pm 22$ | | | | |
| | S19 | 206 | $60.0 \pm 25$ | 13610 | 0.129 | -1.52 | 0.08 |
| | S22 | 146 | $63.5 \pm 25$ | | | | |
| Short Homework Submission Percentage | F19 | 307 | $83.8 \pm 23$ | 41773 | 0.025 | -2.25 | 0.09 |
| | F21 | 302 | $86.6 \pm 22$ | | | | |
| | S19 | 206 | $84.0 \pm 24$ | 16376 | 0.132 | 1.51 | 0.08 |
| | S22 | 146 | $80.6 \pm 27$ | | | | |
| Long Homework Submission Percentage | F19 | 307 | $82.7 \pm 23$ | 41611 | 0.020 | -2.33 | 0.09 |
| | F21 | 302 | $87.0 \pm 22$ | | | | |
| | S19 | 206 | $83.9 \pm 25$ | 16409 | 0.117 | 1.57 | 0.08 |
| | S22 | 146 | $79.8 \pm 28$ | | | | |
| CSEM Pretest Percentage | F19 | 298 | $27.7 \pm 10$ | 49027 | **0.002** | 3.08 | 0.13 |
| | F21 | 287 | $25.7 \pm 11$ | | | | |
| | S19 | 192 | $27.0 \pm 12$ | 13548 | 0.483 | 0.70 | 0.04 |
| | S22 | 135 | $26.0 \pm 11$ | | | | |
| CSEM Post-test Percentage | F19 | 247 | $59.4 \pm 17$ | 32266 | 0.042 | 2.04 | 0.09 |
| | F21 | 236 | $56.4 \pm 17$ | | | | |
| | S19 | 184 | $59.2 \pm 18$ | 11708 | 0.016 | 2.40 | 0.14 |
| | S22 | 109 | $54.0 \pm 18$ | | | | |

Table 9.1: Mean ± standard deviation by semester. Pairs of fall and spring semesters are compared with a Mann-Whitney $U$ test, the $U$ statistic; its $p$ value, $z$-score, and effect size $r$ are also reported. Bolded $p$-values are significant at the $p < 0.05$ level after a Holm–Bonferroni correction is applied. The $p$ value reported is the uncorrected value.

Bonferroni correction. The significant $p$-values are reported without correction in the table and bolded if they meets the adjusted significance threshold. For both semesters, high-school GPA (HSGPA) was significantly higher, a small effect, in post-pandemic semesters. This may have been a result of the changes imposed on high school instruction and grading by the pandemic. While lecture attendance in the fall semesters did not change through the pandemic, lecture attendance in the post-pandemic spring semesters was significantly lower, also a small effect. Although HSGPA increased after the pandemic, CSEM pretest scores significantly decreased in the fall semesters. This effect is functionally negligible, representing less than one additional pretest question answered correctly before the pandemic.

| Variable | Semester | $N$ | $M\%$ | $\chi^2$ | $p$ | $h$ |
|---|---|---|---|---|---|---|
| ACT/SAT | F19 | 274 | 97.5 | 0.48 | 0.490 | 0.08 |
| Reporting | F21 | 291 | 98.6 | | | |
| Percentage | S19 | 178 | 97.3 | 3.90 | 0.048 | 0.25 |
| | S22 | 121 | 91.7 | | | |
| | F19 | 47 | 14.4 | 0.42 | 0.516 | 0.06 |
| DFW | F21 | 54 | 16.5 | | | |
| Percentage | S19 | 42 | 19.4 | 0.49 | 0.485 | 0.09 |
| | S22 | 37 | 22.8 | | | |

Table 9.2: Difference in means between course-level variables where $M\%$ is the percentage of students reporting ACT/SAT scores or the percentage of DFW students, $p$ is the $p$-value comparing semesters, and $h$ is the effect size of the difference.

Table 9.2 shows the DFW rates - the percentage of students who received a D in the class, an F in the class, or withdrew from the class - and ACT/SAT score reporting rates for the class. Only domestic students are included in the ACT/SAT results. International students are less likely to submit ACT/SAT scores than US students. International attendance dropped in this class from 10.4% in Fall 2019 to 3.3% in Fall 2021. Neither the DFW rate (all students) nor the ACT/SAT reporting rate (domestic students) were significantly different pre- and post-pandemic.

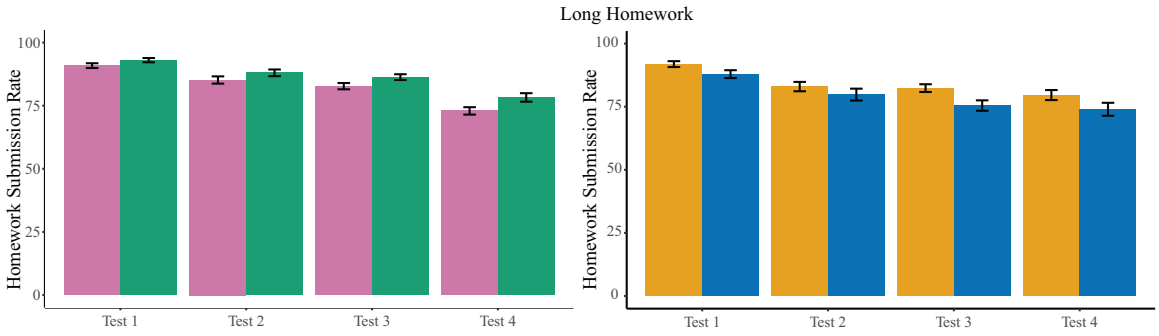


Figure 9.1: The average submission rates for the long homework. The rate is the percentage of the homework assignments submitted for grading.

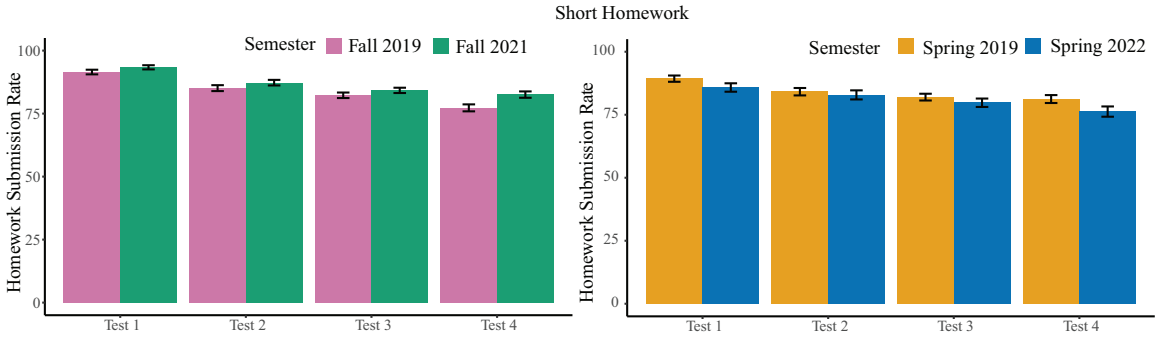


Figure 9.2: The average submission rates for the short homework. The submission rate is the percentage of the assignments submitted for grading.

To understand how student behavior changed over the course of a semester, this study looked at the submission rates of homework divided by the four in-semester examinations. The long homework submission rates are shown in Figure 9.1; the short homework rates are shown in Figure 9.2. All homework assignments which were due during the part of the course covered by each examination were included in the average for the test. For both types of homework, the rate at which students submitted homework assignments was higher in the post-pandemic Fall 2021 semester for every test, while for both types of homework that rate was lower in the post-pandemic Spring 2022 semester for all tests. These differences were, however, small; none of the differences were statistically significant after applying the Holm-Bonferroni correction.

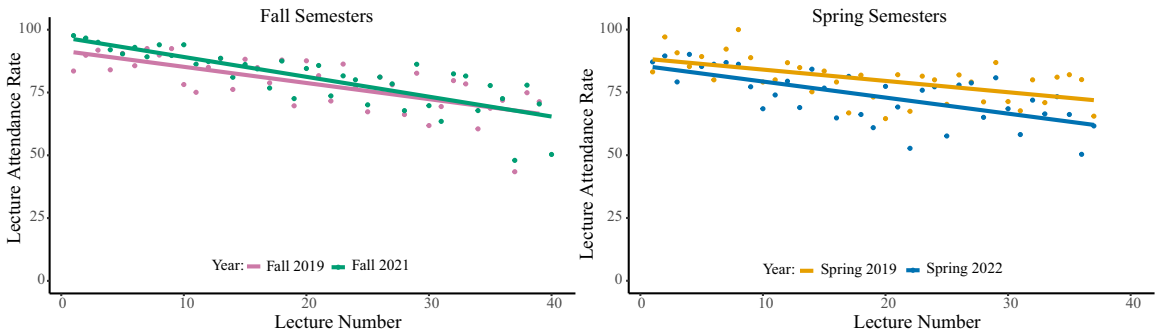


Figure 9.3: Average lecture attendance rates plotted against the order in which the lecture was given. The rate is the percentage of students attending each lecture section.

This study also examined the evolution of lecture attendance over the semester (Figure 9.3). In these plots, lecture number represents the order of the lecture in the semester and is therefore a rough measure of time. For the fall semesters, students attended lecture at a higher rate in Fall 2021 early in the semester, but the rates equalized late in the semester. For the spring semesters, students attended lecture at a lower rate in Spring 2022 throughout the semester and the difference in rates became larger later in the semester. In both cases, the rate of lecture attendance decreased at a larger rate in the post-pandemic semesters (simple linear regression slopes predicting submission rate with lecture number as an independent variable: Fall 2019: -0.013, Fall 2021: -0.016; Spring 2019: -0.009, Spring 2022: -0.013). These differences caused the attendance rates for the fall semesters to converge at the end of the semester while the attendance rates for the spring semesters diverged.

## 9.4 Discussion and Conclusion

The course studied is typically taken by sophomores; most students enrolled in the course are "on-sequence" if they enroll in the course in their fall sophomore semester. Those students likely had a pandemic-interrupted high school experience in their final senior

semester. They would have taken their ACT or SAT tests prior to the pandemic. High school physics is often taken in the senior year; as such, the transition to online instruction may have affected their high school physics class; this would explain differences in CSEM pretest scores. Most students in the class would have taken both Calculus 1 and the introductory mechanics physics class as fully online classes.

In general, few significant differences were measured between pre- and post-pandemic behavior and academic achievement; all significant differences were small effects. This suggests the results of Burkholder and Wieman [182] showing student physics preparation at a highly selective US institution were unchanged through the pandemic extend to students at less selective institutions and are fairly general across achievement on physics assignments and rates of turning in assignments and attending classes.

The few significant differences measured were consistent with the qualitative impression of course personnel, but smaller than they expected. Fall 2021 marked the first semester back to primarily in-person instruction, the course personnel reported that students were generally enthusiastic to return to in-person instruction. This enthusiasm might also explain the statistically significant increase found in long homework scores in Fall 2021. The initial higher lecture attendance in the fall semester post-pandemic decreased over the course of the semester until the fall lecture attendance became equal pre- and post-pandemic. This is consistent with an initial enthusiasm for a return to face-to-face instruction which declined over the semester. Course personnel also felt that student engagement was lower and declined over the semester in the Spring 2022, particularly after spring break. This is also supported by the growing gap between attendance rates for spring semesters pre- and post-pandemic. Course personnel were expecting larger differences from this study, possibly because they

were anticipating a substantial effect of the pandemic student performance and behavior. The actual differences observed were quite small.

# Chapter 10

## Conclusions and Future Work

**Module Analysis**

Student misconceptions in introductory physics have long been an important research area for physics education. This research applied modified module analysis to thousands of student responses to the CSEM, the FCI, and the BEMA to identify both frequently and consistently applied incorrect answering patterns.

The CSEM was constructed to assess students' conceptual understanding of electricity and magnetism. Chapter 4 used module analysis to identify communities of correlated responses to individual items within the CSEM. The resulting network revealed multiple communities formed of responses where the response to later items would be correct if the response to an earlier item was correct. This suggests that the scoring rubric to the CSEM should be modified to include relations between responses. A modified scoring rubric was proposed, but changed overall CSEM post-test averages little. Most communities of completely incorrect responses and mixed correct and incorrect responses consisted of items with the same subtopic, either electrostatics, electric potential, or magnetostatics. Some of communities connected items in multiple subtopics including misconceptions about mechanics and a failure to differentiate the electric and the magnetic field.

The FCI was constructed under the misconception framework with the goal of measuring students' conceptual understanding of Newtonian mechanics. Chapter 5 compared the structure of consistently applied student misconceptions to responses to the FCI across five institutions with student populations with differing levels of high school preparation using MMA-P. The networks identified had substantial similarity for four largest samples in both communities formed of correct responses and of communities associated with mis-

conceptions. The largest force determines motion, Newton's 3rd law, and motion implies active forces misconceptions were the most common misconceptions identified across all five institutions.

The BEMA was also constructed to assess students' conceptual understanding of electricity and magnetism. Chapter 6 used module analysis to identify communities of correlated responses to individual items within the BEMA and to compare this structure with that identified in the CSEM. The most commonly selected and consistently applied mistakes involved electric potential difference and the relation of electric potential difference to electric field.

The BEMA and CSEM share three items with identical stems but different response choices. Communities were identified in both instruments which involved response choices not available in the other instrument. In the BEMA, a community where the student reported the average force on the two charges was identified; it was impossible for students on the CSEM to select this response. Likewise, the CSEM did not include all responses related to consistent reasoning about the distance dependence of the force, while the BEMA did. The CSEM did include an "other" response to catch these forms of incorrect reasoning. Both instruments revealed unique aspects of student thinking because they allowed for different student responses. The results of both the CSEM and the BEMA module analysis suggest the existence of a rich collection of incorrect reasoning about electricity and magnetism which are consistently applied by students after instruction in introductory physics that may not be fully captured by either instrument.

A substantial number of the identified communities throughout these studies consisted of blocked items providing continued support that the practice of blocking items can produce correlations that are not related to physical reasoning. The consistent identification

of blocking as generating psychometric problems for the primary conceptual instruments used in PER, along with the limited coverage of the material covered by these instruments, suggest the need for a new generation of conceptual inventories. Module analysis can be an important part of the validation process for these new instruments adding to more traditional psychometric analysis techniques.

The large number of students still applying misconceptions or consistent mistakes post-instruction supports a continued need to transition to research-based instructional methods and to continuously improve and target those methods towards the mistakes identified as most frequently and consistently applied.

### West Virginia STEM Education Network

Social network analysis was used to study the growth and development of the First2 STEM education network from 2018 to 2022. The growth of the network over its first four years was relatively consistent; connections increased in number and strength, fully-connected groups increased in number and size, and key leaders appeared over the years to disseminate information and collaborate with many others. In the fifth year of the project, leaders began to transition the project to sustainability within state academic organizations rather than disseminating information across the network like the first four years. The structure of the network reflected these changes with the number of edges, density, average strength, average betweenness, and number and size of cliques all decreasing.

The network also provided multiple ways for students to be involved and to take leadership positions. From the beginning, student were encouraged to voice their opinions as equal network participants. This led to robust clusters of student to student connections in

the network, with student leaders standing out as central network members that connected many otherwise unconnected students to the core of the network.

### Student Learning and Behavior Before and After the COVID-19 Pandemic

Performance metrics, attendance rates, homework submission rates, and DFW rates were examined for students in an introductory calculus-based electromagnetism course both before and after the transition to virtual learning caused by the COVID-19 pandemic. This study was motivated by a recent study at a highly selective university that found that there was no evidence for a reduction in high-school physics learning after the pandemic using a physics diagnostic exam administered in the Fall 2019 and Fall 2021 semesters. Chapter 9 examined a broader collection of student achievement and behavior measures at an institution admitting students with lower levels of high school achievement. The institution is the flagship state university in a small eastern state with a state population with per capita income and rate of bachelors attainment among the lowest in the US. As such, it provided context of the effect of the pandemic on the education of students coming from less resourced school systems.

In general, few significant differences were measured between pre- and post-pandemic behavior and academic achievement; all significant differences were small effects. This suggests that the study that demonstrated that student physics preparation at a highly selective US institution were unchanged through the pandemic extends to students at less selective institutions and are fairly general across achievement on physics assignments and rates of turning in assignments and attending classes.

**Future Works**

This work has identified and categorized some of the most consistently and frequently applied misconceptions post-instruction in introductory calculus-based mechanics and electromagnetism in the U.S. However, much more work is needed to properly address these misconceptions in the classroom and to understand the breadth of misconceptions not captured by the instruments studied in this document. Future projects and ongoing research related to the work outlined in this thesis are provided below:

- Develop a new generation of mechanics conceptual instruments that have a reproducible factor structure, have items that are well functioning in Classical Test Theory, do not use item chaining or blocking, have items that pass a quantitative fairness test for groups of students underrepresented in physics classes, and have community structures which are theoretically supportable and which allow for the calculation of misconception scores for the misconceptions most commonly applied in the topic covered.

- Develop a new generation of electromagnetic conceptual instruments which feature both subscales with broad general coverage of the major subtopics of electromagnetism, but also subscales that capture common mistakes or allow measurement of finer details of conceptual knowledge.

- Develop a taxonomy of electromagnetic misconceptions identified post-instruction.

- Explore the difference between MMA applied to Likert scale surveys and network analysis for Likert-style surveys (NALS). Dalka *et al.* [188] introduced NALS in order

to analyze connections between items in Likert-style surveys. NALS uses the same sparsification process as MAMCR, but constructs edges with a metric related to the correlation used in MMA.

- Explore the context dependence of misconceptions identified in this manuscript. Misconceptions that are more or less likely to be applied in certain physical contexts may allow targeted interventions that make use of this context dependence.

# Bibliography


[1] J. L. Docktor and J. P. Mestre. Synthesis of discipline-based education research in physics. *Phys. Rev. Phys. Educ. Res.*, 10(2):020119, 2014.

[2] J. Clement. Students' preconceptions in introductory mechanics. *Am. J. Phys.*, 50(1):66–71, 1982.

[3] L.C. McDermott. Research on conceptual understanding in mechanics. *Phys. Today*, 37:24–32, 1984.

[4] L.C. McDermott and E.F. Redish. Resource letter: PER-1: Physics education research. *Am. J. Phys.*, 67(9):755–767, 1999.

[5] G.J. Posner, K.A. Strike, P.W. Hewson, and W.A. Gertzog. Accommodation of a scientific conception: Toward a theory of conceptual change. *Sci. Educ.*, 66(2):211–227, 1982.

[6] D. Hammer. Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions. *J. Learn. Sci.*, 5(2):97–127, 1996.

[7] I.A. Halloun and D. Hestenes. The initial knowledge state of college physics students. *Am. J. Phys.*, 53(11):1043–1055, 1985.

[8] I.A. Halloun and D. Hestenes. Common sense concepts about motion. *Am. J. Phys.*, 53(11):1056–1065, 1985.

[9] D. Hestenes, M. Wells, and G. Swackhamer. Force Concept Inventory. *Phys. Teach.*, 30:141, 1992.

[10] E. Mazur. *Peer Instruction: A User's Manual*. Prentice Hall, Upper Saddle River, NJ, 1997.

[11] Physport. `https://www.physport.org`. Accessed 8/8/2017.

[12] R.K. Thornton and D.R. Sokoloff. Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. *Am. J. Phys.*, 66(4):338–352, 1998.

[13] D.P. Maloney, T.L. O'Kuma, C. Hieggelke, and A. Van Huevelen. Surveying students' conceptual knowledge of electricity and magnetism. *Am. J. Phys.*, 69(S1):S12–S23, 2001.

[14] The BEMA itself was never published in an archival journal. Early references to the instrument use Chabay and Sherwood (1997) (Ref 1) to cite the instrument. This issue of the AAPT Announcer is not available electronically. The citation references the program to the Summer 1997 American Association of Physics Teachers meeting then published in the Announcer. The page referenced contains Chabay and Sherwood's contributed talk abstracts about research applying the instrument. Interestingly, Maloney, O'Kuma, Van Heuvelen, and Hieggelke discussed challenges to developing an electricity and magnetism instrument in the same session which lead to the CSEM.

[15] H. Sadaghiani and S. Pollock. Quantum mechanics concept assessment: Development and validation study. *Phys. Rev. ST Phys. Educ. Res.*, 11:010110, Mar 2015.

[16] S. B. McKagan, K. K. Perkins, and C. E. Wieman. Design and validation of the Quantum Mechanics Conceptual Survey. *Phys. Rev. ST Phys. Educ. Res.*, 6:020121, Nov 2010.

[17] S. Yeo and M. Zadnik. Introductory thermal concept evaluation: Assessing students' understanding. *The Physics Teacher*, 39(8):496–504, 11 2001.

[18] P. Wattanakasiwich, P. Taleab, M. Sharma, and I.D. Johnston. Development and implementation of a conceptual survey in thermodynamics. *International Journal of Innovation in Science and Mathematics Education*, 21:29–53, 01 2013.

[19] A. Tongchai, M. D. Sharma, I. D. Johnston, K. Arayathanitkul, and C. Soankwan. Developing, Evaluating and Demonstrating the use of a Conceptual Survey in Mechanical Waves. *International Journal of Science Education*, 31(18):2437–2457, 2009.

[20] B. Hufnagel. Development of the Astronomy Diagnostic Test. *Astronomy Education Review*, 1, 01 2002.

[21] J. Bailey, B. Johnson, E. Prather, and T. Slater. Development and Validation of the Star Properties Concept Inventory. *International Journal of Science Education*, 34:2257–2286, 09 2012.

[22] E. Bardar, E. Prather, and T. Slater. Development and Validation of the Light and Spectroscopy Concept Inventory. *Astronomy Education Review*, 5, 10 2006.

[23] J. Epstein. Development and Validation of the Calculus Concept Inventory. In *Ninth International Conference on Mathematics Education in a Global Community*, Charlotte, NC, 2007.

[24] M. Carlson, M. Oehrtman, and N. Engelke. The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cog. Instr.*, 28(2):113–145, 2010.

[25] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman. New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Phys. Rev. ST Phys. Educ. Res.*, 2:010101, Jan 2006.

[26] B. Zwickl, N. Finkelstein, and H. J. Lewandowski. Development and validation of the C]olorado Learning Attitudes about Science Survey for Experimental Physics, booktitle = Physics Education Research Conference 2012. volume 1513 of *PER Conference*, pages 442–445, Philadelphia, PA, 2012.

[27] S. W. Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman. Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 17:020129, Oct 2021.

[28] R.R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, 66:64–74, 1998.

[29] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, NY, 1977.

[30] D.E. Meltzer and R.K. Thornton. Resource letter ALIP–1: Active-learning instruction in physics. *Am. J. Phys.*, 80(6):478–496, 2012.

[31] R.J. Dufresne, W.J. Gerace, W.J. Leonard, J.P. Mestre, and L. Wenk. Classtalk: A classroom communication system for active learning. *J. Comput. High. Educ.*, 7(2):3–47, 1996.

[32] N.W. Reay, L. Bao, P. Li, R. Warnakulasooriya, and G. Baugh. Toward the effective use of voting machines in physics lectures. *Am. J. Phys.*, 73(6):554–558, 2005.

[33] L. Ding, N.W. Reay, A. Lee, and L. Bao. Are we asking the right questions? validating clicker question sequences by student interviews. *Am. J. Phys.*, 77(7):643–650, 2009.

[34] C.H. Crouch, J. Watkins, A.P. Fagen, and E. Mazur. Peer instruction: Engaging students one-on-one, all at once. *Research-based reform of university physics*, 1(1):40–95, 2007.

[35] A. Van Heuvelen. Learning to think like a physicist: A review of research-based instructional strategies. *Am. J. Phys.*, 59(891), 1991.

[36] E. Etkina, S. Murthy, and X. Zou. Using introductory labs to engage students in experimental design. *Am. J. Phys.*, 74(11):979–986, 2006.

[37] A. Karelina and E. Etkina. Acting like a physicist: Student approach study to experimental design. *Phys. Rev. ST Phys. Educ. Res.*, 3:020106, Oct 2007.

[38] E. Etkina, A. Karelina, M. Ruibal-Villasenor, D. Rosengrant, R. Jordan, and C.E. Hmelo-Silver. Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *J. Learn. Sci.*, 19(1):54–98, 2010.

[39] C. Wieman and N.G. Holmes. Measuring the impact of an instructional laboratory on the learning of introductory physics. *Am. J. Phys.*, 83(11):972–978, 2015.

[40] N.G. Holmes, J. Olsen, J.L. Thomas, and C.E. Wieman. Value added or misattributed? a multi-institution study on the educational benefit of labs for reinforcing physics content. *Phys. Rev. Phys. Educ. Res.*, 13(1):010129, 2017.

[41] R.J. Beichner, J.M. Saul, D.S. Abbott, J.J. Morse, D. Deardorff, R.J. Allain, S.W. Bonham, M.H. Dancy, and J.S. Risley. The student-centered activities for large enrollment undergraduate programs (scale-up) project. *Research-based reform of university physics*, 1(1):2–39, 2007.

[42] J. Gaffney, E. Richards, M. B. Kustusch, L. Ding, and R. Beichner. Scaling up education reform. *J. Coll. Sci. Teaching*, 37(5):48, May 2008.

[43] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj. Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Phys. Rev. Phys. Educ. Res.*, 8(2):020104, 2012.

[44] E. Brewe, R. Dou, and R. Shand. Costs of success: Financial implications of implementation of active learning in introductory physics courses for students and administrators. *Phys. Rev. Phys. Educ. Res.*, 14:010109, Feb 2018.

[45] S. Freeman, S. Eddy, M. McDonough, M. Smith, N. Okoroafor, H. Jordt, and M. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci.*, 111(23):8410–8415, 2014.

[46] F. J. Gravetter and L. B. Wallnau. Statistics for the behavioral sciences. 2002.

[47] B. L. Welch. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1-2):28–35, 01 1947.

[48] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[49] R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

[50] J. Welkowitz, H. B. Cohen, and B. R. Lea. *Introductory statistics for the behavioral sciences*. John Wiley & Sons, 2011.

[51] C.E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, Florence, Italy, 1936.

[52] M.D. Moran. Arguments for rejecting the sequential bonferroni in ecological studies. *Oikos*, 100(2):403–405, 2003.

[53] S. Nakagawa. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6):1044–1045, 11 2004.

[54] L. Ding, R. Chabay, B. Sherwood, and R. Beichner. Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment. *Phys. Rev. Phys. Educ. Res.*, 2:010105, Mar 2006.

[55] E. Brewe, J. Bruun, and I.G. Bearden. Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data. *Phys. Rev. Phys. Educ. Res.*, 12:020131, Sep 2016.

[56] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler. Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis. *Phys. Rev. Phys. Educ. Res.*, 15:020122, Sep 2019.

[57] J. Wells, R. Henderson, A. Traxler, P. Miller, and J. Stewart. Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis. *Phys. Rev. Phys. Educ. Res.*, 16(1):010121, April 2020.

[58] J. Yang, J. Wells, R. Henderson, E. Christman, G. Stewart, and J. Stewart. Extending modified module analysis to include correct responses: Analysis of the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 16:010124, Apr 2020.

[59] J. Clement, D.E. Brown, and A. Zietsman. Not all preconceptions are misconceptions: Finding 'anchoring conceptions' for grounding instruction on students' intuitions. *Int. J. Sci. Educ.*, 11(5):554–565, 1989.

[60] J. Clement. Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *J. Res. Sci. Teach.*, 30(10):1241–1257, 1993.

[61] Table II for the Force Concept Inventory (revised from 081695r). `http://modeling.asu.edu/R&E/FCI-RevisedTable-II_2010.pdf`. Accessed 3/17/2019.

[62] M.T.H. Chi and J.D. Slotta. The ontological coherence of intuitive physics. *Cogn. Instr.*, 10(2-3):249, 1993.

[63] M.T.H. Chi, J.D Slotta, and N. De Leeuw. From things to processes: A theory of conceptual change for learning science concepts. *Learn. Instr.*, 4(1):27, 1994.

[64] J.D. Slotta, M.T.H. Chi, and E. Joram. Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change. *Cogn. Instr.*, 13(3):373, 1995.

[65] A.A. diSessa. Toward an epistemology of physics. *Cogn. Instr.*, 10(2-3):105, 1993.

[66] A.A. diSessa and B.L. Sherin. What changes in conceptual change? *Int. J. Sci. Educ.*, 20(10):1155, 1998.

[67] J. Minstrell. Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, and H. Niedderer, editors, *Research in Physics Learning: Theoretical Issues and Empirical Studies*, page 110. IPN, Kiel, Germany, 1992.

[68] D. Hammer. More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research. *Am. J. Phys.*, 64(10):1316, 1996.

[69] D. Hammer. Student resources for learning introductory physics. *Am. J. Phys.*, 68(S1):S52, 2000.

[70] R.E. Scherr. Modeling student thinking: An example from special relativity. *Am. J. Phys.*, 75(3):272–280, 2007.

[71] F. De Vico, J. Richiardi, M. Chavez, and S. Achard. Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philos. T. R. Soc. Lon. B*, 369(1653), 2014.

[72] J. Lopéz Peña and H. Touchette. A network theory analysis of football strategies. In C. Clanet, editor, *Sports Physics: Proc. 2012 Euromech Physics of Sports Conference*, pages 517–528. Éditions de l'École Polytechnique, Abr 2012.

[73] Z. Zheng and Y. Zhao. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to "CandidatusLiberibacter asiaticus" infection. *BMC Genomics*, 14(1):27, Jan 2013.

[74] A.V. Papachristos and C. Wildeman. Network exposure and homicide victimization in an African American community. *Am. J. Public Health*, 104(1):143–150, 2014.

[75] D. Grunspan, B. Wiggins, and S. Goodreau. Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research. *CBE—Life Sci. Educ.*, 13(2):167–178, 2014.

[76] I. Koponen and M. Nousiainen. Concept networks in learning: Finding key concepts in learners' representations of the interlinked structure of scientific knowledge. *J. Complex Netw.*, 2:187–202, 05 2014.

[77] M. Stella, S. de Nigris, A. Aloric, and C. Siew. Forma mentis networks quantify crucial differences in stem perception between students and experts. *PLOS ONE*, 14(10):1–21, 10 2019.

[78] A. Traxler, T. Suda, E. Brewe, and K. Commeford. Network positions in active learning environments in physics. *Phys. Rev. Phys. Educ. Res.*, 16:020129, Oct 2020.

[79] K. Commeford, E. Brewe, and A. Traxler. Characterizing active learning environments in physics using network analysis and classroom observations. *Phys. Rev. Phys. Educ. Res.*, 17:020136, Nov 2021.

[80] M. Sundstrom, D. Wu, C. Walsh, A. Heim, and N. Holmes. Examining the effects of lab instruction and gender composition on intergroup interaction networks in introductory physics labs. *Phys. Rev. Phys. Educ. Res.*, 18:010102, Jan 2022.

[81] D. Vargas, A. Bridgeman, D. Schmidt, P. Kohl, B. Wilcox, and L. Carr. Correlation between student collaboration network centrality and academic performance. *Phys. Rev. Phys. Educ. Res.*, 14:020112, Oct 2018.

[82] J. Bruun and E. Brewe. Talking and learning physics: Predicting future grades from network measures and Force Concept Inventory pretest scores. *Phys. Rev. ST Phys. Educ. Res.*, 9:020109, Jul 2013.

[83] J. Forsman, C. Linder, R. Moll, D. Fraser, and S. Andersson. A new approach to modelling student retention through an application of complexity thinking. *Stud. in High. Educ.*, 39(1):68, 2014.

[84] J. Zwolak, R. Dou, E. Williams, and E. Brewe. Students' network integration as a predictor of persistence in introductory physics courses. *Phys. Rev. Phys. Educ. Res.*, 13:010113, Mar 2017.

[85] R. Dou, E. Brewe, J. Zwolak, G. Potvin, E. Williams, and L. Kramer. Beyond performance metrics: Examining a decrease in students' physics self-efficacy through a social networks lens. *Phys. Rev. Phys. Educ. Res.*, 12:020124, Aug 2016.

[86] R. Dou and J. Zwolak. Practitioner's guide to social network analysis: Examining physics anxiety in an active-learning setting. *Phys. Rev. Phys. Educ. Res.*, 15:020105, Jul 2019.

[87] M. Bodin. Mapping university students' epistemic framing of computational physics using network analysis. *Phys. Rev. ST Phys. Educ. Res.*, 8:010115, Apr 2012.

[88] J. Bruun, M. Lindahl, and C. Linder. Network analysis and qualitative discourse analysis of a classroom group discussion. *Int. J. Res. Meth. Educ.*, 42(3):317–339, 2019.

[89] J. Zwolak, M. Zwolak, and E. Brewe. Educational commitment and social networking: The power of informal networks. *Phys. Rev. Phys. Educ. Res.*, 14:010131, May 2018.

[90] E. Brewe, L. Kramer, and V. Sawtelle. Investigating student communities with network analysis of interactions in a physics learning center. *Phys. Rev. ST Phys. Educ. Res.*, 8:010101, Jan 2012.

[91] K. Anderson, M. Crespi, and E. Sayre. Linking behavior in the physics education research coauthorship network. *Phys. Rev. Phys. Educ. Res.*, 13:010121, May 2017.

[92] E. Brewe. The Roles of Engagement: Network Analysis in Physics Education Research. In *Getting Started in PER*, volume 2. PER Central, College Park, MD, 4 edition, July 2018.

[93] C. Wheatley, J. Wells, R. Henderson, and J. Stewart. Applying module analysis to the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 17:010102, Jan 2021.

[94] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart. Multidimensional item response theory and the force concept inventory. *Phys. Rev. Phys. Educ. Res.*, 14:010137, Jun 2018.

[95] M. J. Lai, J. Xie, and Z. Xu. Graph sparsification by universal greedy algorithms. *ArXiv*, abs/2007.07161, 2020.

[96] M.A. Serrano, M. Boguná, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *PNAS*, 106(16):6483–6488, 2009.

[97] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[98] A. Canty and B.D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2017. R package version 1.3-20.

[99] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695:1–9, 2006.

[100] C. Hieggelke and T. O'Kuma. The impact of physics education research on the teaching of scientists and engineers at two-year colleges. In *1997 Physics Education Research Conference Proceedings*, volume 399, pages 267–288, New York, 1997. AIP, AIP Publishing.

[101] J. Yang, C. Zabriskie, and J. Stewart. Multidimensional item response theory and the Force and Motion Conceptual Evaluation. *Phys. Rev. Phys. Educ. Res.*, 15(2):020141, 2019.

[102] R. Henderson, C. Zabriskie, and J. Stewart. Rural and first generation performance differences on the Force and Motion Conceptual Evaluation. Physics Education Research Conference Proceedings 2018 (accepted)., 2018.

[103] D.P. Maloney. Charged poles? *Phys. Educ.*, 20(6):310, nov 1985.

[104] C. Guruswamy, M.D. Somers, and R.G. Hussey. Students' understanding of the transfer of charge between conductors. *Phys. Educ.*, 32(2):91–96, mar 1997.

[105] S. Törnkvist, K.A. Pettersson, and G. Tranströmer. Confusion by representation: On student's comprehension of the electric field concept. *Am. J. Phys.*, 61(4):335–338, 1993.

[106] I. Galili. Mechanics background influences students' conceptions in electromagnetism. *Int. J. Sci. Educ.*, 17(3):371–387, 1995.

[107] L. Viennot and S. Rainson. Students' reasoning about the superposition of electric fields. *Int. J. Sci. Educ.*, 14:475–487, 10 1992.

[108] S. Rainson, G. Tranströmer, and L. Viennot. Students' understanding of superposition of electric fields. *Am. J. Phys.*, 62(11):1026–1032, 1994.

[109] M. Planinic. Assessment of difficulties of some conceptual areas from electricity and magnetism using the Conceptual Survey of Electricity and Magnetism. *Am. J. Phys.*, 74(12):1143–1148, 2006.

[110] K. Kreutzer and A. Boudreaux. Preliminary investigation of instructor effects on gender gap in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 8(1):010120, 2012.

[111] P.B. Kohl and H.V. Kuo. Introductory physics gender gaps: Pre-and post-studio transition. *AIP Conf. Proc.*, 1179:173–176, 2009.

[112] J.M. Wilson. The CUPLE physics studio. *Phys. Teach.*, 32(9):518–523, 1994.

[113] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler. Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 13:020114, Sep 2017.

[114] A. Madsen, S.B. McKagan, and E. Sayre. Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Phys. Rev. Phys. Educ. Res.*, 9:020121, Nov 2013.

[115] J. Leppävirta. The effect of naïve ideas on students' reasoning about electricity and magnetism. *Res. Sci. Educ.*, 42(4):753–767, 2012.

[116] D.E. Meltzer. Analysis of shifts in students' reasoning regarding electric field and potential concepts. In L. McCullough, L. Hsu, and P. Heron, editors, *2006 Physics Education Research Conference Proceedings*, volume 883, pages 177–180, New York, 2007. AIP, AIP Publishing.

[117] N.I. Karim, A. Maries, and C. Singh. Exploring one aspect of pedagogical content knowledge of teaching assistants using the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 14:010117, Apr 2018.

[118] US News & World Report: Education. US News and World Report, Washington, DC. `https://premium.usnews.com/best-colleges`. Accessed 4/30/2017.

[119] S. DeVore, J. Stewart, and G. Stewart. Examining the effects of testwiseness in conceptual physics evaluations. *Phys. Rev. Phys. Educ. Res.*, 12:020138, 2016.

[120] C. Wheatley, J. Wells, D. E. Pritchard, and J. Stewart. Comparing conceptual understanding across institutions with module analysis. *Phys. Rev. Phys. Educ. Res.*, 18:020132, Nov 2022.

[121] D. Huffman and P. Heller. What does the Force Concept Inventory actually measure? *Phys. Teach.*, 33:138, 1995.

[122] M.R. Semak, R.D. Dietz, R.H. Pearson, and C.W. Willis. Examining evolving performance on the Force Concept Inventory using factor analysis. *Phys. Rev. Phys. Educ. Res.*, 13:010103, Jan 2017.

[123] T.F. Scott, D. Schumayer, and A.R. Gray. Exploratory factor analysis of a Force Concept Inventory data set. *Phys. Rev. Phys. Educ. Res.*, 8(2):020105, 2012.

[124] T.F. Scott and D. Schumayer. Students' proficiency scores within multitrait item response theory. *Phys. Rev. Phys. Educ. Res.*, 11:020134, Nov 2015.

[125] D. Hestenes and I. Halloun. Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *Phys. Teach.*, 33(8):502–502, 1995.

[126] J. Wells, H. Sadaghiani, B. Schermerhorn, S. Pollock, and G. Passante. Deeper look at question categories, concepts, and context covered: Modified module analysis of quantum mechanics concept assessment. *Phys. Rev. Phys. Educ. Res.*, 17:020113, Aug 2021.

[127] I. Halloun, R.R. Hake, E.P. Mosca, and D. Hestenes. Force Concept Inventory (revised 1995), 1995. `http://modeling.asu.edu/R&E/Research.html` Accessed 7/19/2019.

[128] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell. Gender fairness within the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 14(1):010103, 2018.

[129] National Center for Education Statistics. `https://nces.ed.gov/collegenavigator`.

[130] J. Stewart, B. Drury, J. Wells, A. Adair, R. Henderson, Y. Ma, A. Pérez-Lemonche, and D. Pritchard. Examining the relation of correct knowledge and misconceptions using the nominal response model. *Phys. Rev. Phys. Educ. Res.*, 17(1):010122, 2021.

[131] M. Dickison, M. Magnani, and L. Rossi. *Multilayer Social Networks*. Cambridge University Press, New York, NY, 2016.

[132] M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, and M.A. Moreno, Y and. Porter. Multilayer Networks. *J. Complex Netw.*, 2(3):203, 07 2014.

[133] M. Magnani, D. Vega, and M. Dubik. Analysis and Mining of Multilayer Social Networks, 2020.

[134] G. Palla, I. Derényi, and T. Vicsek. Clique Percolation in Random Networks. *Phys. Rev. Lett.*, 94:160202, 05 2005.

[135] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75, 2010.

[136] Bródka P., A. Chmiel, M. Magnani, and G. Ragozini. Quantifying Layer Similarity in Multiplex Networks: a Systematic Study. *Roy. Soc. Open. Sci.*, 5(8):171747, Aug 2018.

[137] T. Wei, V. Simko, M. Levy, Y. Xie, Y. Jin, and J. Zemla. Package "corrplot". *Statistician*, 56(316):24, 2017.

[138] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York, NY, 1994.

[139] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell. Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 14(2):020103, 2018.

[140] C. Wheatley, J. Wells, and J. Stewart. Applying module analysis to the Brief Electricity and Magnetism Assessment. *Phys. Rev. Phys. Educ. Res.*, 20:010104, Feb 2024.

[141] R. Chabay and B. Sherwood. Qualitative understanding and retention. *AAPT Announcer*, 27:96, 1997.

[142] S. J. Pollock. Longitudinal study of student conceptual understanding in electricity and magnetism. *Phys. Rev. Phys. Educ. Res.*, 5(2):020110, 2009.

[143] M.A. Kohlmyer, M.D. Caballero, R. Catrambone, R.W. Chabay, L. Ding, M.P. Haugan, M.J. Marr, B.D. Sherwood, and M.F. Schatz. Tale of two curricula: The performance of 2000 students in introductory electromagnetism. *Phys. Rev. Phys. Educ. Res.*, 5(2):020105, 2009.

[144] M.W. McColgan, R.A. Finn, D.L. Broder, and G.E. Hassel. Assessing students' conceptual knowledge of electricity and magnetism. *Phys. Rev. Phys. Educ. Res.*, 13(2):020121, 2017.

[145] R. L. Doran. *Basic measurement and evaluation of science instruction*. National Science Teachers Association, Washington, D.C., 1980.

[146] L. Ding. Applying Rasch theory to evaluate the construct validity of brief electricity and magnetism assessment. In *2011 Physics Education Research Conference Proceedings*, volume 1413, pages 175–178, New York, 2012. AIP, AIP Publishing.

[147] R.W. Chabay and B.A. Sherwood. *Matter and Interactions*. John Wiley & Sons, Hoboken, NJ, 2015.

[148] L. Ding. Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assesment through Rasch analysis. *Phys. Rev. Phys. Educ. Res.*, 10(1):010105, 2014.

[149] Y. Xiao, J.C. Fritchman, J.Y. Bao, Y. Nie, J. Han, J. Xiong, H. Xiao, and L. Bao. Linking and comparing short and full-length concept inventories of electricity and magnetism using item response theory. *Phys. Rev. Phys. Educ. Res.*, 15(2):020149, 2019.

[150] J. Hansen and J. Stewart. Multidimensional item response theory and the brief electricity and magnetism assessment. *Phys. Rev. Phys. Educ. Res.*, 17:020139, Nov 2021.

[151] P. Eaton, B. Frank, K. Johnson, and S. Willoughby. Comparing exploratory factor models of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 15(2):020133, 2019.

[152] P. Eaton, K. Johnson, B. Frank, and S. Willoughby. Classical test theory and item response theory comparison of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and Magnetism. *Phys. Rev. Phys. Educ. Res.*, 15(1):010102, 2019.

[153] S.J. Pollock. Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA. *AIP Conf. Proc.*, 1064:171–174, 2008.

[154] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software Pact. Exper.*, 21(11):1129–1164, 1991.

[155] L. Crocker and J. Algina. *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, Mason, OH, 1986.

[156] A. F. Cunningham, W. Erisman, and S. E. Looney. Higher education in michigan: Overcoming challenges to expand access. *Institute for Higher Education Policy*, 2008.

[157] C. Howley, J. Johnson, A. Passa, and K. Uekawa. College enrollment and persistence in rural pennsylvania schools. rel 2015-053. *Regional Educational Laboratory Mid-Atlantic*, 2014.

[158] C. J. Wright. Becoming to remain: Community college students and post-secondary pursuits in central appalachia. *Journal of research in rural education*, 27(6), 2012.

[159] V. Tinto and J. Cullen. Dropout in Higher Education: A Review and Theoretical Synthesis of Recent Research. 1973.

[160] V. Tinto. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, 45(1):89–125, 1975.

[161] A. Astin. Student Involvement: A Development Theory for Higher Education. *Journal of College Student Development*, 40:518–529, 01 1984.

[162] H. McQueen. Integration and regulation matters in educational transition: A theoretical critique of retention and attrition models. *British Journal of Educational Studies*, 57(1):70–88, 2009.

[163] W. G. Tierney. An anthropological analysis of student participation in college. *The Journal of Higher Education*, 63(6):603–618, 1992.

[164] L. R. Sims and J. J. Ferrare. "since i am from where i am from": How rural and urban first-generation college students differentially use social capital to choose a college major. *Journal of Research in Rural Education*, 37(6):1—-21, 2021.

[165] Á. J. Lukács, D. Beáta, et al. Roma undergraduates' personal network in the process of college transition. a social capital approach. *Research in Higher Education*, 60(1):64–82, 2019.

[166] J.E. Eckles and E.G. Stradley. A social network analysis of student retention using archival data. *Soc Psychol Educ*, 15:165–180, 2012.

[167] D. J. Almeida, A. M. Byrne, R. M. Smith, and S. Ruiz. How relevant is grit? the importance of social capital in first-generation college students' academic success. *Journal of College Student Retention: Research, Theory & Practice*, 23(3):539–559, 2019.

[168] O. Poldin, D. Valeeva, and M. Yudkevich. Which peers matter: How social ties affect peer-group effects. *Research in Higher Education*, 57(4):448–468, 2016.

[169] M. Berthelon, E. Bettinger, D. I. Kruger, and A. Montecinos-Pearce. The structure of peers: The impact of peer networks on academic achievement. *Research in Higher Education*, 60(7):931–959, 2019.

[170] M. S. González Canché. *Geographical, Statistical, and Qualitative Network Analysis: A Multifaceted Method-Bridging Tool to Reveal and Model Meaningful Structures in Education Research*, volume 34, pages 535–634. Springer International Publishing, Cham, Switzerland, 2019.

[171] M. S. González Canché and C. Rios-Aguilar. Critical social network analysis in community colleges: Peer effects and credit attainment. *New Directions for Institutional Research*, 2014(163):75–91, 2015.

[172] T. Hogue, D. Perkins, R. Clark, A. Bergstrum, M. Slinski, and Associates. Collaboration framework: Addressing community capacity., 1995.

[173] L. M. Borden and D. F. Perkins. Assessing your collaboration: A self-evaluation tool. *Journal of Extension*, 37(2):67–72, 1999.

[174] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[175] T. Wei and V. Simko. *R package 'corrplot': Visualization of a Correlation Matrix*, 2021. (Version 0.92).

[176] C. Bron and J. Kebosch. Algorithm 457: Finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):48–50, 1973.

[177] C. Aaron, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), dec 2004.

[178] A. Nemeth, C. Wheatley, and J. Stewart. Comparing introductory undergraduate physics learning and behavior before and after the COVID-19 pandemic. *Phys. Rev. Phys. Educ. Res.*, 19:013103, May 2023.

[179] P. Klein, L. Ivanjek, M. N. Dahlkemper, K. Jeličić, M.-A. Geyer, S. Küchemann, and A. Susac. Studying physics during the COVID-19 pandemic: Student assessments of learning achievement, perceived effectiveness of online recitations, and online laboratories. *Phys. Rev. Phys. Educ. Res.*, 17:010117, Mar 2021.

[180] M.F.J. Fox, J.R. Hoehn, A. Werth, and H.J. Lewandowski. Lab instruction during the COVID-19 pandemic: Effects on student views about experimental physics in comparison with previous years. *Phys. Rev. Phys. Educ. Res.*, 17:010148, Jun 2021.

[181] I. Marzoli, A. Colantonio, C. Fazio, M. Giliberti, U. Scotti di Uccio, and I. Testa. Effects of emergency remote instruction during the COVID-19 pandemic on university physics students in Italy. *Phys. Rev. Phys. Educ. Res.*, 17:020130, Oct 2021.

[182] E.W. Burkholder and C.E. Wieman. Absence of a COVID-induced academic drop in high-school physics learning. *Phys. Rev. Phys. Educ. Res.*, 18:023102, Jul 2022.

[183] G.W. Corder and D.I. Foreman. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. John Wiley & Sons, Inc., Hoboken, NJ, 2009.

[184] G.J. Privitera. *Statistics for the Behavioral Sciences*. SAGE Publications, Thousand Oaks, CA, 3 edition, 2017.

[185] Roy C. Milton. An Extended Table of Critical Values for the Mann-Whitney (Wilcoxon) Two-Sample Statistic. *Journal of the American Statistical Association*, 59(307):925–934, 1964.

[186] D. George and P. Mallery. *IBM SPSS Statistics 27 Step by Step: A Simple Guide and Reference*. Routledge, New York, NY, 17 edition, 2022.

[187] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian J. Stat.*, pages 65–70, 1979.

[188] R. P. Dalka, D. Sachmpazidi, C. Henderson, and J. P. Zwolak. Network analysis approach to likert-style surveys. *Phys. Rev. Phys. Educ. Res.*, 18:020113, Sep 2022.