Comparing conceptual understanding across institutions with module analysis

Christopher Wheatley,[1] James Wells,[2] David E. Pritchard,[3] and John Stewart[1,*]

[1]*West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA*
[2]*University of Connecticut, Department of Physics, Storrs, Connecticut 06269, USA*
[3]*Massachusetts Institute of Technology, Department of Physics, Cambridge, Massachusetts 02139, USA*

(Received 30 April 2022; accepted 17 October 2022; published 14 November 2022)

The Force Concept Inventory (FCI) is a popular multiple-choice instrument used to measure a student's conceptual understanding of Newtonian mechanics. Recently, a network analytic technique called module analysis has been used to identify responses to the FCI and other conceptual instruments that are preferentially selected together by students; these groups of responses are called communities. This study uses module analysis to explore the misconception structure of the FCI at five U.S. institutions with varying undergraduate populations (sample sizes of $N = 9606$, 4360, 1496, 466, and 213). Students from these universities had a broad range of prior knowledge in physics and of general high school academic preparation, resulting in large differences in FCI normalized gain, pretest, and post-test scores. In the current work, modified module analysis partial was applied and communities of consistently selected responses within the FCI were identified at the five institutions studied. There was substantial similarity between the communities identified postinstruction; somewhat less similarity preinstruction. This suggests that consistently applied Newtonian misconceptions exist both before and after instruction at a wide range of institutions. The most frequently applied misconceptions were "largest force determines motion," Newton's third law misconceptions, and "motion implies active forces." These misconceptions were still consistently applied even after instruction by a substantial number of students at all but the highest performing of the five institutions.

DOI: 10.1103/PhysRevPhysEducRes.18.020132

I. INTRODUCTION

The study of student misconceptions of physics concepts has long been an important area of inquiry in physics education research (PER). A misconception is a consistently and coherently applied error in students' conceptual understanding of physics. Multiple-choice instruments such as the Force and Motion Conceptual Evaluation (FMCE) [1], the Force Concept Inventory (FCI) [2], the Brief Electricity and Magnetism Assessment (BEMA) [3], the Conceptual Survey of Electricity and Magnetism (CSEM) [4], and the Quantum Mechanics Concept Assessment (QMCA) [5] have facilitated the quantitative study of student misconceptions.

The 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 [2]. The FCI has been one of the most commonly used and, accordingly, studied conceptual instrument in physics since its introduction in 1992. The FCI along with the catalog of common misconceptions it measures [6] has been transformative to PER. Hake collected FCI data from multiple institutions to show traditional teaching methods were broadly ineffective at improving student understanding [7]. 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 [8]. 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. I C.

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 [9–14]. 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. Module analysis uses a community detection algorithm (CDA) to group responses frequently selected by students into communities. A community is a set of

[*] jcstewart1@mail.wvu.edu

nodes that have stronger connection to each other than to other nodes in the network. These communities, otherwise known as modules, are then analyzed to explore student conceptual understanding of the subject. Module analysis has been particularly useful in studying the structure of misconceptions in these instruments, providing insight not available through other methods, while also suggesting that some items may not be functioning as intended.

There have been many quantitative studies of the FCI, the FMCE, and the CSEM. Many of these studies have applied factor analysis and have been largely ineffective at extracting meaningful substructure from the FCI [15–19]. The authors of the FCI argue that the instrument was not constructed to factor [2,20]. The reason for this is evident in the correlation matrices presented by Stewart *et al.* [19]; 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.

A. Research questions

The current work applied the network analytic technique called modified module analysis partial (MMA-P), detailed in Sec. II C, 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?

B. 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 [21] and can be found at PhysPort [22].

C. Prior studies of the FCI

A thorough summary of the prior research using the FCI was presented in previous module analysis studies [10,19,23]. An overview is provided below.

1. 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 [15]. This analysis identified only two out of the six factors described by Hestenes *et al.*: “kinds of forces” and Newton’s third law. When EFA was applied to 750 students at the university level, the only factor identified was *kinds of forces* [15]. Scott *et al.* performed a factor analysis of 2150 college student posttest responses and found five factors were required for the optimal model; one factor explained much of the variance [17]. 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 [18]. 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 [16]. Stewart *et al.* also performed a factor analysis using MIRT on 4716 post-test responses [19] 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.

2. *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 [2].

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 [2]; 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) [2].

The misconception framework is not the only framework applied to students’ alternative ideas in physics [24,25]. Alternate frameworks include the knowledge-in-pieces framework and the ontological categories framework. The knowledge-in-pieces framework has been investigated by many authors who have conceptualized the reasoning fragments for the framework as resources [26–29], phenomenological primitives (*p* prims) [30,30,31], or facets [32]. In this framework, students’ conceptual beliefs about physics are not understood as hypotheses about general phenomena. Instead, student thinking consists of basic building blocks that are applied in different combinations dependent on the specific context presented to them. In an ontological categories framework, entities are classified into mutually exclusive categories, which determine what characteristics an entity can have [33–35]. A soccer ball would be in the ontological category of *matter* and not the *processes* category. In this framework, incorrect student thinking results from classifying physical concepts into the wrong category.

In previous works using network-based methods to analyze the FCI, the findings have conformed most closely to the misconceptions framework [9,10,12]; as such, misconceptions are used to describe the ideas represented by the incorrect answer clusters in these works. This is hardly surprising as the FCI was built in the misconception framework. It is possible that a concept inventory designed to elicit resources or ontological categories would result in communities which conform better to those frameworks. The use of misconceptions is not an endorsement of that framework over the other frameworks, but a convenient shorthand to compare the findings with the stated intentions of the inventory creators. Module analysis is fundamentally a quantitative analysis of the answer choices students consistently select; these patterns can provide only tangential support for a correct cognitive framework to understand those answering patterns.

D. Network analysis

Network analysis is a versatile set of techniques that have been applied across many disparate research areas. These techniques have been used in a variety of studies outside of education, such as mapping electrical signals in the brain as functional networks [36], the difference between passing patterns in different teams at the World Cup [37], plants’ response to bacterial infection [38], and the probability of becoming a homicide victim when living within a disadvantaged neighborhood [39]. Network analysis has also been fruitful within educational research to study the structure of classrooms through the social interactions of students and teachers [40], undergraduate student representations of the relatedness of physics concepts through concept maps [41], and the difference between high school students’ and interdisciplinary professionals’ emotional perception and conceptual knowledge of science, technology, engineering, and mathematics [42].

1. *Social network analysis in physics education*

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 PER to characterize and test active learning environments [43–45], to predict future performance [46,47], to predict retention and persistence within a degree program [48,49], to explore physics self-efficacy and anxiety [50,51], to explore interactions between lab groups by gender [45], to study conceptual change in student responses and discussions [52,53], to determine the effect of informal learning environments and out of class relationships on class involvement and commitment [54,55], and to explore the change in co-authorship behaviors in PER over time [56]. For an overview of network analysis in PER, see the review by Brewe [57].

2. Module analysis

Module analyses are a set of network analytic techniques used to analyze multiple-choice instruments [9–13]. 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 [9]. 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 [10]. 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) [10]. The communities extracted by MMA are generally small, which simplifies the identification of the reasoning which lead to the responses being selected together. MMA was applied to 4500 responses to the FCI from an introductory calculus-based physics class [10]. 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 third law misconceptions. All of these are described in detail in Hestenes and Jackson's [6] 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 [2], the detailed description of misconceptions measured by the instrument provided by Jackson and Hestenes [6], and the detailed mapping of the granular knowledge measured by the instrument provided by Stewart *et al.* [19].

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.* [12]; 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 the current study, MMA-P, is the same algorithm as developed by Yang *et al.* [12]; this algorithm will be used to construct the networks. The current study introduces network comparison algorithms available by combining different networks into a multiplex network.

II. METHODS

A. 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 [58]. All samples contained only matched pretest and post-test responses with no missing responses.

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

TABLE I. 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	4000	33–35
5	213	10 000	33–35

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

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

Sample 5: 38% White, 18% Asian, 12% Hispanic/Latino, 12% nonresident 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 I.

These samples among others were collected by Pritchard as part of a work to improve item response theory analysis of the FCI [59]. 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.

B. 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 for two continuous random variables X and Y as

$$r_{XY} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \qquad (1)$$

where $E[X]$ is the expectation value, μ_i is the average of variable i, and σ_i is the standard deviation of the same variable.

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 as

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1-r_{XZ}^2}\sqrt{1-r_{YZ}^2}}. \qquad (2)$$

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.

C. 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 below. For a more detailed explanation see Yang *et al.* [12] where the method was introduced.

MMA-P first constructs a network out of the correct and incorrect responses to a multiple choice instrument. 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-P constructs edges between nodes from the partial correlation $r_{XY|S}$ of the response vectors correcting for total instrument score S.

MMA and MMA-P perform a number of operations which result in nodes and edges being removed from the networks; this process is called sparsification in network analysis [60]. 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. Edges with correlations that were not significant at the $p < 0.05$ level after a Bonferroni correction was applied were removed. Note, for brevity we will shorten partial correlation to correlation and use r as the partial correlation coefficient in this work. Item responses that were chosen by fewer than 5% of students were removed. Edges with negative correlations were removed. A correlation threshold was then applied to remove small correlations. In most previous studies, $r > 0.2$ was used where r is the partial correlation coefficient correcting for overall instrument score; the current study introduces a graphical method to set the threshold for each sample (Sec. II D). This threshold was selected to produce compact communities with theoretically understandable structure. The sparsification process and its relation to sample size is discussed in more detail in Sec. II E.

Once the network is constructed, a community detection algorithm (CDA) is applied to identify communities within the network. In network analysis, a set of nodes that have stronger connections between themselves than with nodes outside of the set is called a community or a module. Note, in this work very strong levels of sparsificaiton are used which 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 will continue to use the term community. This work used 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 [61]; 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. This study applied the fast-greedy CDA [62] to identify communities within the network. Wells *et al.* [10] 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 [63] in R was used for bootstrapping and the igraph package [64] in R was used for the community detection.

D. 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 [9–11,13,65]; 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.

E. 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 the Supplemental Material [66] 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. The Supplemental Material [66] presents 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 1–4. 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 1–4 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. The Supplemental Material [66] 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 1–4 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.

F. 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" [67]. 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.* [67] or Kivelä *et al.* [68].

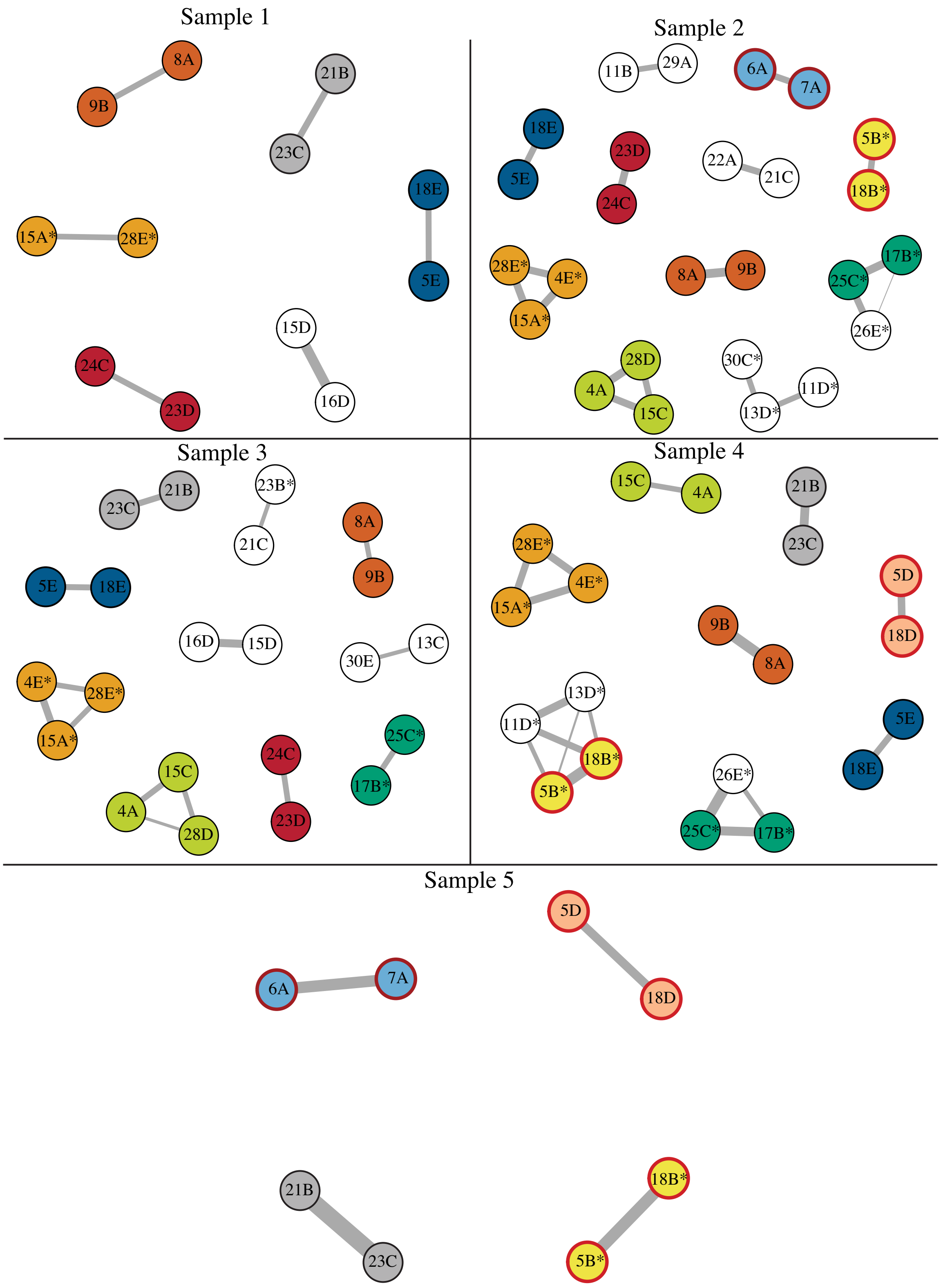
A multiplex network was formed applying MMA-P to the FCI response data from each institution studied individually, creating 5 distinct networks. These networks were then added as layers forming a multiplex network. As the networks were computed independently using the MMA-P algorithm of Yang *et al.* [12], 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 [69] was used to construct the multiplex network.

G. 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 [70]. The CPM identifies communities which share k edges in m layers. Figures 1 and 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 $k = 3$ [71]. A triangle is a completely connected subnetwork with 3 nodes. As an example, consider the sample 4 pretest network in Fig. 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) [72] 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 [73] in R as shown in Fig. 4.



Sample 1

Sample 2

Sample 3

Sample 4

Sample 5

FIG. 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.

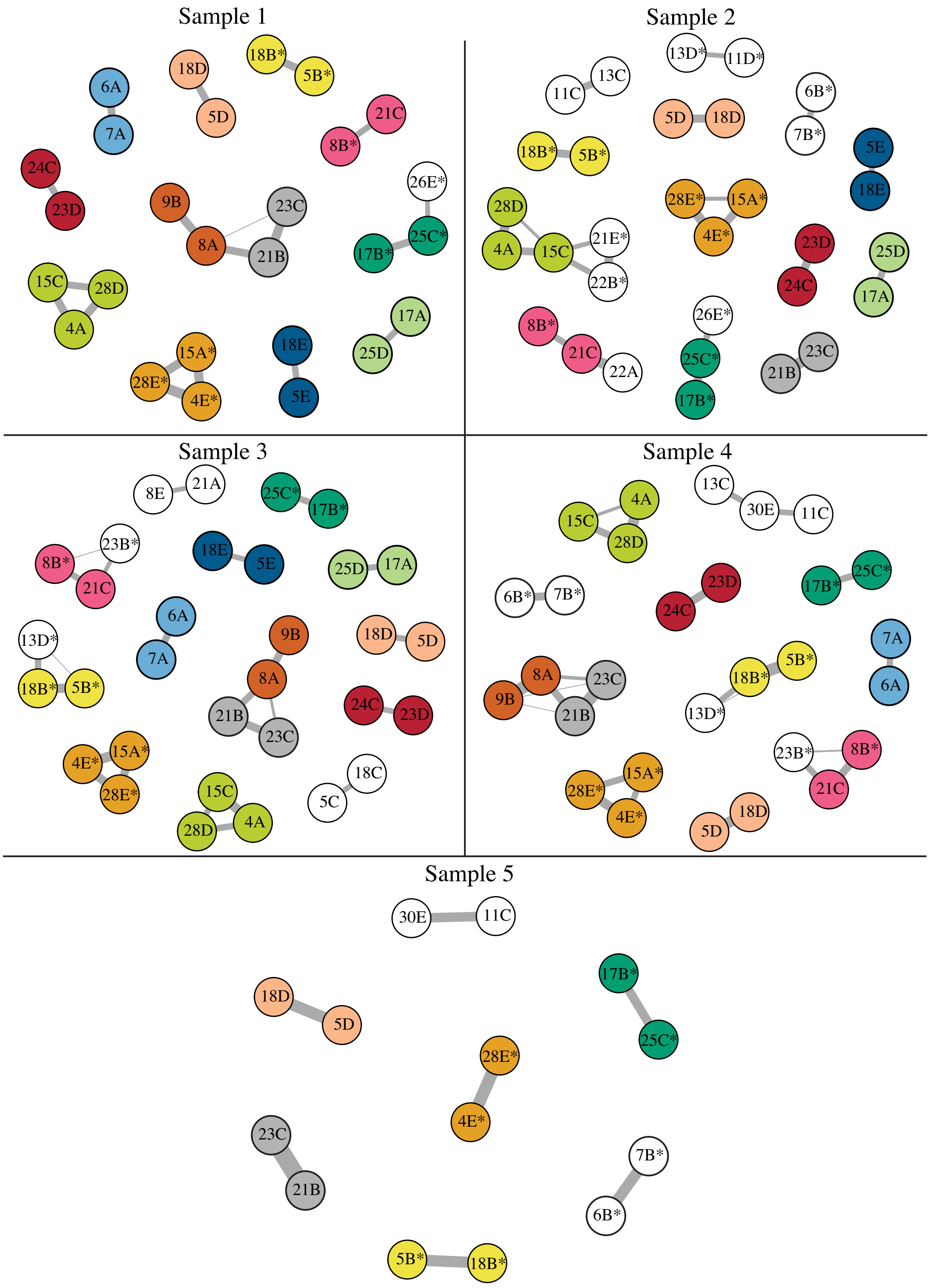


FIG. 2. Post-test networks. Communities found in three of the four largest samples are shaded with the same color. Correct responses are marked with an asterisk. The size of the partial correlation between the responses is proportional to the edge width.

TABLE II. 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

III. RESULTS

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

A. The networks

The community structure identified by MMA-P is shown for the pretest in Fig. 1 and for the post-test in Fig. 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 third law for their solution. For the pretest in Fig. 1, the communities have been colored consistently with the post-test in Fig. 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.

TABLE III. 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 subcommunities 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.

	Pretest					Post-test					
Community	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5	Explanation
Completely incorrect communities											
4A-15C		×	×	⊗		×	×	×	×		Newton's third law misconceptions.
(4A, 15C, 28D)		⊗	⊗			⊗	×	⊗	⊗		Newton's third 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 third law.
15A*-28E*	⊗	×	×	×		×	×	×	×		Newton's third law.
(4E*, 15A*, 28E*)		⊗	⊗	⊗		⊗	⊗	⊗	⊗		Newton's third 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.

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 III. Only a subset of all communities are presented to highlight structures that were common across many institutions and to 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.

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 × 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 III 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 [6] while incorporating changes to this taxonomy suggested in the original MMA paper [10]. Correct answers are classified using the detailed model of the FCI constructed by Stewart *et al.* [19]. The original model proposed with the publication of the FCI [2] 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 producing 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 the Supplemental Material [66]. On the post-test, the communities identified in the networks of only one or two institutions differ from those in Table III 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 2 indicates a strong similarity between student responses postinstruction 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 postinstruction in three of the four samples include all communities identified in three of the four largest preinstruction 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 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 third law for their correct solution (items 4, 15, 16, and 28) are categorized as "Newton's third 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 [6]. The Newton's third law items do not allow these misconceptions to be disentangled. This may explain the mixing of items with different combinations of Newton's third law items shown in Table III.

Most communities identified and described in Table III have been identified previously in the FCI by either Wells *et al.* [10,11] or Yang *et al.* [12]; 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 [19]: $\{4, 15, 16, 28\}$, $\{5, 18\}$, $\{6, 7\}$, 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 [10]. 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 communities 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 [6]. 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 [6]. 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 responses 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.

B. 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 Fig. 3. Each point is calculated at a different r threshold

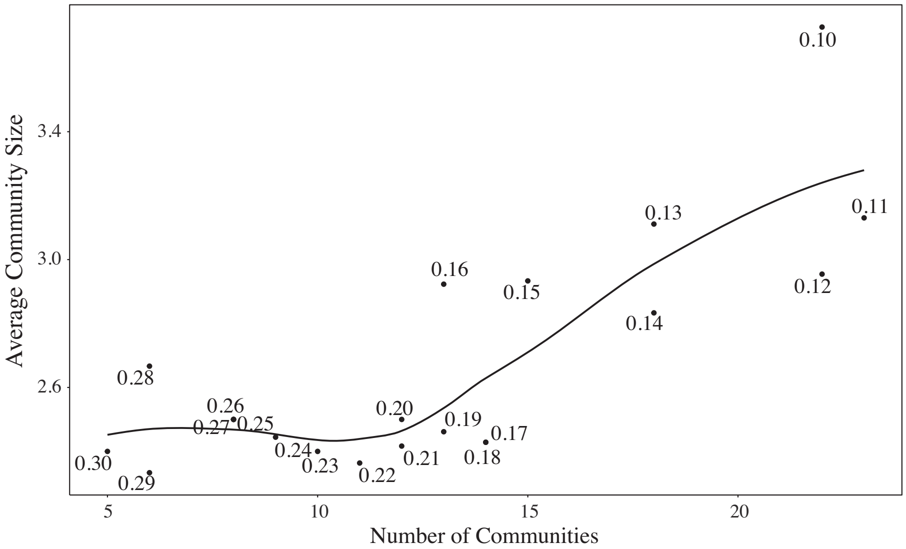


FIG. 3. The correlation threshold r for the sample 1 post-test network. Each point represents a network calculated at the labeled r value.

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 Fig. 2. Plots for other networks are included in the Supplemental Material [66].

The correlation thresholds selected using this method for the pretest and post-test are shown in Table IV. The sample 5 network was independent of r and, therefore, no threshold was required for this sample. As shown in the Supplemental Material [66], this behavior is the result of the sample size making the resolution of correlations below 0.35 unlikely.

C. Layer comparison results

A wealth of network comparison metrics have been developed for multiplex networks. For this work, we use the coverage index of the actors (nodes), edges, and triangles (completely connected 3 node subnetworks) to quantitatively characterize network similarity. Plots of these quantities are shown in Fig. 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.

TABLE IV. Partial correlation threshold coefficients used for each sample.

Pre or Post	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Pretest	0.20	0.20	0.16	0.21	not applicable
Post-test	0.21	0.20	0.17	0.22	not applicable

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 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, $\langle \mathrm{CI} \rangle$, of the four largest samples shows an increase in similarity from pretest to post-test (pretest $\langle \mathrm{CI} \rangle = 0.63$ actors, 0.58 edges, and 0.28 triangles; post-test $\langle \mathrm{CI} \rangle = 0.79$ 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 ($\langle \mathrm{CI} \rangle_{\mathrm{pre}} = 0.56$, $\langle \mathrm{CI} \rangle_{\mathrm{post}} = 0.60$). This stability is partially the result of

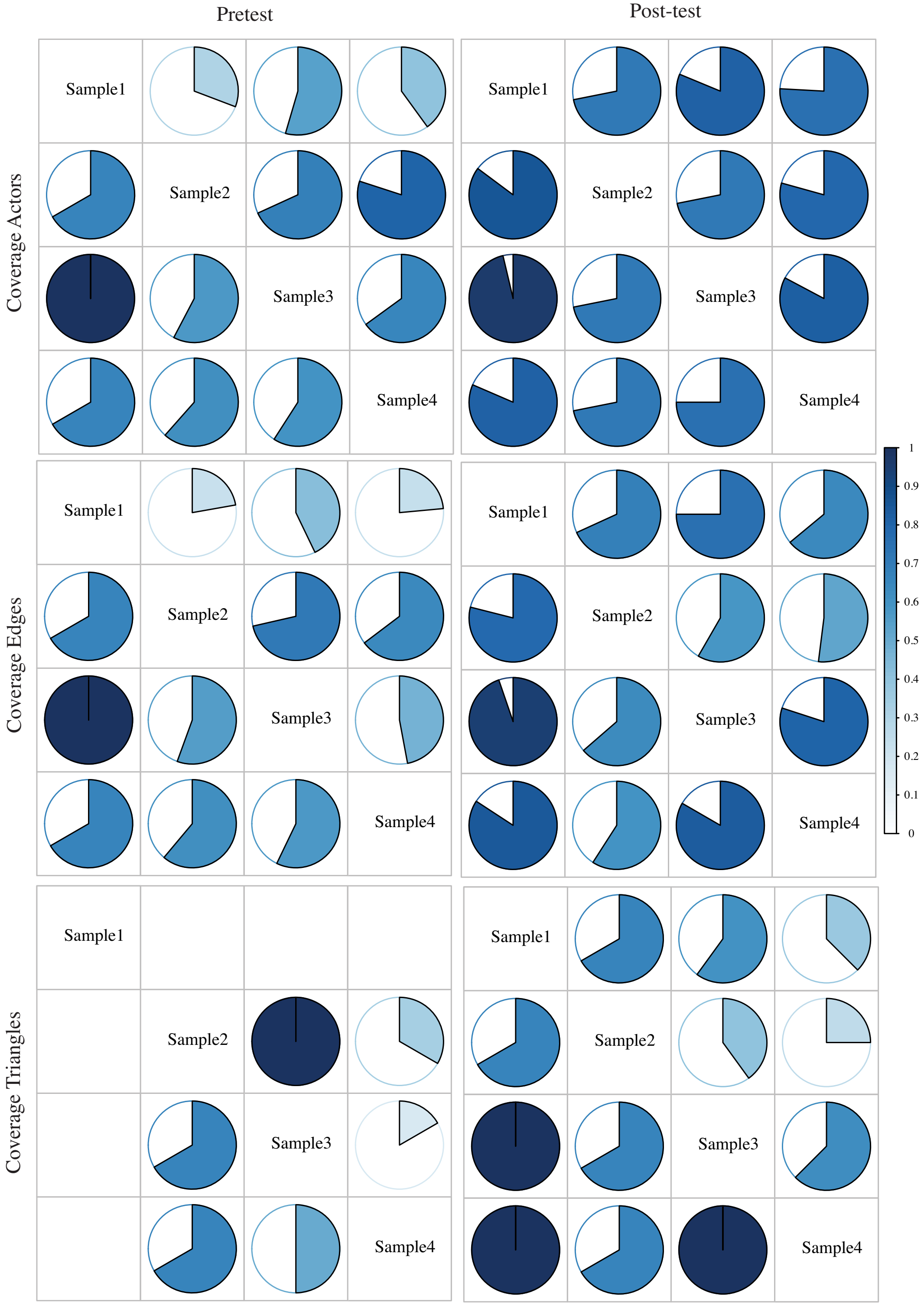


FIG. 4. Coverage actors, edges, and triangles between samples.

correct and incorrect responses to Newton’s third law items forming the majority of the triangles.

D. Misconception scores

Wells *et al.* [10] 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. Table V presents the misconception scores for completely incorrect communities found in most

TABLE V. 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 third 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%

post-test networks in Figs. 1 and 2. Misconception scores represent the number of responses chosen that are associated with a misconception out of the total number of FCI responses that are associated with the same misconception. Misconception scores are computed independently for each institution. For example, responses 6A and 7A are related to the circular impetus 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 12.9% score shown for sample 1 for the circular impetus misconception is the average of each student's misconception score for that sample. The misconceptions are classified using the modifications proposed in Wells *et al.* [10] to the Hestenes and Jackson taxonomy [6].

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 third law misconception group contains 3 responses, the 31.9% misconception score for sample 1 indicates that on average students in this sample select $3 \times 0.319 = 0.96$ of these responses for each application of the FCI. That is, even postinstruction, students are on average answering one Newton's third law item incorrectly using a common misconception.

The largest force determines motion, Newton's third law, and motion implies active forces misconceptions consistently have the highest score across all five institutions. The average misconception scores for these three misconceptions 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 $2 \times 0.44 = 0.88$, $3 \times 0.39 = 1.2$, and $5 \times 0.31 = 1.6$ times each time the FCI is given. These misconceptions are likely some of the most widely held and consistently applied in mechanics and remain postinstruction at institutions with a broad spectrum of student populations. The rest of the misconception scores 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.

IV. 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 III shows, misconceptions related to Newton's third 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 preinstruction and postinstruction. The misconception scores presented in Table V suggest the largest force determines motion, Newton's third law, and motion implies active forces misconceptions are the most prevalent postinstruction.

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 related to the motion implies active forces misconception. This consistency suggests the misconception groups are also present in international students.

Community 5D-18D, corresponding to the motion implies active forces misconception, community (4A, 15C, 28D), corresponding the Newton's third 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 postinstruction and had the three highest misconception scores indicating the misconceptions were still frequently applied postinstruction across a broad spectrum of 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 postinstruction. Many of the incorrect communities identified postinstruction with high misconception scores were identified preinstruction much less consistently. This may be because many students answer incorrectly preinstruction because they have little knowledge of the correct physics and thus are not consistent, but students with consistently applied misconceptions retain these postinstruction.

The similarity of the community structure across the institutions studied suggests that the sets of consistently applied misconceptions present preinstruction and remaining postinstruction may be very consistent across many institutions. Misconception scores suggest many of these misconceptions are still selected by many students postinstruction 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 [15–19]. 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 preinstruction and postinstruction. 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 [10–13]. In some works, $r > 0.2$ yielded a network far too sparsified and $r > 0.15$ was chosen instead [13]. 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 [74], 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 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 $\langle \mathrm{CI} \rangle = 0.63$ for actors and 0.58 for edges on the pretest which increased to $\langle \mathrm{CI} \rangle = 0.79$ 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 third law, or motion implies active forces misconception postinstruction each time they take the FCI. The rate of consistently applying these misconceptions was much lower at the highest performing institution.

V. 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 third 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 [6] 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 subdimensions of these Newtonian force concepts. This would provide instructors with a more fine-grained measurement of student knowledge.

VI. 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 noncognitive constructs such as self-efficacy, belonging, and identity.

The current work and prior MMA and MMA-P analysis of the FCI, FMCE, and CSEM have used very restrictive correlation and community fraction thresholds so as to identify compact communities with clear theoretical explanations. With these communities identified, the correlation and community fraction thresholds can be relaxed to allow more complex structure to emerge which should show how these communities are connected producing a more complex picture of student thinking about conceptual physics.

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 [15], items which would be in the range of problematic item functioning in classical test theory (CTT) [75], and the effects of the practice of blocking or chaining items [19]. Beyond these, substantial issues of item fairness for some demographic groups have been identified in some of the instruments [75]. 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.

VII. 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 third 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.

ACKNOWLEDGMENTS

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, PHY-0108787. Data collection for this work was supported by National Science Foundation Grants No. EPS-1003907, No. ECR-1561517, and No. HRD-1834569.

[1] 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**, 338 (1998).

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

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

[4] 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**, S12 (2001).

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

[6] Table II for the Force Concept Inventory (revised from 081695r), http://modeling.asu.edu/R&E/FCI-RevisedTable-II_2010.pdf.

[7] 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 (1998).

[8] 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. U.S.A. **111**, 8410 (2014).

[9] 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 (2016).

[10] 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 (2019).

[11] 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**, 010121 (2020).

[12] 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 (2020).

[13] 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 (2021).

[14] 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 (2021).

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

[16] 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 (2017).

[17] 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**, 020105 (2012).

[18] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[19] 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 (2018).

[20] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[21] 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.

[22] Physport, https://www.physport.org.

[23] 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**, 010103 (2018).

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

[25] R. Duit, D. F. Treagust, and A. Widodo, Teaching Science for Conceptual Change, in *International Handbook of Research on Conceptual Change* (Routledge, Abingdon, UK, 2013).

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

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

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

[29] A. Elby, Helping physics students learn how to learn, Am. J. Phys. **69**, S54 (2001).

[30] A. A. diSessa, Knowledge in pieces, in *Constructivism in the Computer Age*, The Jean Piaget Symposium Series, edited by George Forman and Peter B. Pufall (Lawrence Erlbaum, Hillsdale, NJ, 1988), p. 49.

[31] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[32] J. Minstrell, Facets of students' knowledge and relevant instruction, in *Research in Physics Learning: Theoretical Issues and Empirical Studies*, edited by R. Duit, F. Goldberg, and H. Niedderer (IPN, Kiel, Germany, 1992), p. 110.

[33] M. T. H. Chi and J. D. Slotta, The ontological coherence of intuitive physics, Cognit. Instr. **10**, 249 (1993).

[34] 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**, 27 (1994).

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

[36] F. De Vico, J. Richiardi, M. Chavez, and S. Achard, Graph analysis of functional brain networks: Practical issues in translational neuroscience, Phil. Trans. R. Soc. B **369**, 20130521 (2014).

[37] J. Lopéz Peña and H. Touchette, A network theory analysis of football strategies, in *Sports Physics: Proc. 2012 Euromech Physics of Sports Conference*, edited by C. Clanet (Éditions de l'École Polytechnique, 2012), pp. 517–528.

[38] 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**, 27 (2013).

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

[40] 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**, 167 (2014).

[41] 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 (2014).

[42] 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**, 1 (2019).

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

[44] 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 (2021).

[45] 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 (2022).

[46] 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 (2018).

[47] 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 (2013).

[48] 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. Higher Educ. **39**, 68 (2014).

[49] 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 (2017).

[50] 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 (2016).

[51] 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 (2019).

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

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

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

[55] 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 (2012).

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

[57] E. Brewe, The roles of engagement: Network analysis in physics education research, *Getting Started in PER*, 4th ed. (PER Central, College Park, MD, 2018), Vol. 2.

[58] National Center for Education Statistics, https://nces.ed.gov/collegenavigator.

[59] 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**, 010122 (2021).

[60] M. J. Lai, J. Xie, and Z. Xu, Graph sparsification by universal greedy algorithms, arXiv:2007.07161.

[61] M. A. Serrano, M. Boguná, and A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proc. Natl. Acad. Sci. U.S.A. **106**, 6483 (2009).

[62] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E **69**, 026113 (2004).

[63] A. Canty and B. D. Ripley, boot: Bootstrap R (S-Plus) Functions (2021), R package version 1.3-28.

[64] G. Csardi and T. Nepusz, The igraph software package for complex network research, InterJournal, Complex Systems **1695**, 1 (2006).

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

[66] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.020132 for the ACS vs NC graphs for each community, a full table of the communities identified, a summary of the sparsification process, and an analysis of low occupation responses.

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

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

[69] M. Magnani, L. Rossi, and D. Vega, Analysis of multiplex social networks with R, J. Stat. Softw. **98**, 1 (2021).

[70] G. Palla, I. Derényi, and T. Vicsek, Clique Percolation in Random Networks, Phys. Rev. Lett. **94**, 160202 (2005).

[71] S. Fortunato, Community detection in graphs, Phys. Rep. **486**, 75 (2010).

[72] P. Bródka, A. Chmiel, M. Magnani, and G. Ragozini, Quantifying layer similarity in multiplex networks: A systematic study, R. Soc. Open. Sci. **5**, 171747 (2018).

[73] T. Wei and V. Simko, R package 'corrplot': Visualization of a Correlation Matrix, 2021, https://github.com/taiyun/corrplot.

[74] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, New York, NY, 1994), p. 243.

[75] 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**, 020103 (2018).