Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data

Eric Brewe
Teaching and Learning Department, STEM Transformation Institute, Physics Department, Florida International University Department, 11200 SW 8 St.; Miami, Florida 33199, USA

Jesper Bruun
Department of Science Education, Faculty of Science, University of Copenhagen, Øster Voldgade 3; DK-1350 København K, Denmark

Ian G. Bearden
Niels Bohr Institute, Faculty of Science, University of Copenhagen, Denmark and Department of Science Education, Faculty of Science, University of Copenhagen, Blegdamsvej 17, Bygning Q; DK-2100 København Ø, Denmark

(Received 13 January 2016; published 20 September 2016)

We describe *Module Analysis for Multiple Choice Responses (MAMCR)*, a new methodology for carrying out network analysis on responses to multiple choice assessments. This method is used to identify modules of non-normative responses which can then be interpreted as an alternative to factor analysis. MAMCR allows us to identify conceptual modules that are present in student responses that are more specific than the broad categorization of questions that is possible with factor analysis and to incorporate non-normative responses. Thus, this method may prove to have greater utility in helping to modify instruction. In MAMCR the responses to a multiple choice assessment are first treated as a bipartite, student X response, network which is then projected into a response X response network. We then use data reduction and community detection techniques to identify modules of non-normative responses. To illustrate the utility of the method we have analyzed one cohort of postinstruction Force Concept Inventory (FCI) responses. From this analysis, we find nine modules which we then interpret. The first three modules include the following: Impetus Force, More Force Yields More Results, and Force as Competition or Undistinguished Velocity and Acceleration. This method has a variety of potential uses particularly to help classroom instructors in using multiple choice assessments as diagnostic instruments beyond the Force Concept Inventory.

DOI: 10.1103/PhysRevPhysEducRes.12.020131

I. INTRODUCTION

Conceptual inventories have arguably played an important role in the transformation of science courses [1]. They hold the potential to help researchers and instructors identify non-normative science understandings, and they have been used to "determine if conceptual change [have] occurred during a course" [1] (p. 769). In physics the Force Concept Inventory (FCI) is the best known and most widely used conceptual inventory. It is a 30 question multiple choice assessment, with each question having a normative response (which is consistent with Newtonian mechanics) and several non-normative responses (often referred to as "distractors") based on student responses to conceptual questions [2]. This format is common among a number of conceptual inventories. The FCI was developed with the goal of helping teachers identify their students' conceptions about classical mechanics, and then modify instruction to address the conceptions [3]. Instead, the FCI, like other conceptual inventories, is often used as a measure of instructional quality, where the metric of interest is typically based on the difference of post- and preinstruction scores [4]. The simplicity and apparent objectivity of these scores may have contributed to the use of concept inventories. With this paper, we wish to add to the possible uses of conceptual inventories. We focus on the FCI as an illustrative example for two reasons. First, it can be taken as the template which other conceptual inventories follow. Thus, illustrating possible additional uses of conceptual inventories with the FCI can be seen as a proof of concept. Second, the FCI is widely used and can be seen as one of the motivators for educational change. Exemplifying how the use of conceptual inventories can be expanded by using the FCI as an exemplar may serve as a seeding point for further educational change in physics.

One of the reasons that conceptual inventories are primarily used as metrics for instructional quality is likely that interpreting student responses is unwieldy. For example, after more than twenty years, there is still discussion of both what the FCI actually measures and how FCI scores should be interpreted. This discussion includes studies of the reliability of the instrument [5], applications of Item-Response Theory [6,7], Rasch analysis [8], and two studies using factor analysis [9,10]. The factor analysis studies, in particular, have cast doubts on the FCI's ability to measure distinct constructs. This apparent lack on the side of the FCI may decrease the educational value of conceptual inventories as instruments for measuring science understandings.

This paper introduces Module Analysis for Multiple Choice Responses (MAMCR), a novel method using network analysis and techniques of community detection to identify the structure of response patterns in conceptual inventories. To do this we have chosen to use the FCI as an ideal illustrative example. MAMCR has been developed to improve the utility of any conceptual inventory as a diagnostic instrument by identifying "communities" of responses that include the non-normative responses. We then use the method to analyze post-instruction FCI data from one cohort of introductory physics students at University of Copenhagen. This analysis is meant to be illustrative of the utility of MAMCR.

A. Structural features of conceptual inventories as diagnostic instruments

One of the features of conceptual inventories is that questions include non-normative responses (alternate choices). The non-normative responses are the alternate answers that make up the other choices in a multiple choice question. For the FCI, the non-normative responses were developed based on open-ended responses during the period in which it was developed [11]. As a result, these responses are strongly related to students'experiences and their relation to the normative answer is not trivial. The presence of distinct non-normative responses means that individual test items do not (and cannot) test individual concepts. Take for example, a hypothetical question that is designed to probe student understanding of the relationship between force and acceleration. The structure of the conceptual inventory questions is to include multiple different non-normative responses as possible answers to the question. The hypothetical force and acceleration question may include non-normative responses about net force, momentum, or velocity that are taken to correspond to non-normative understandings. The presence of these responses make it so that a question is not strictly about any one concept. For example, through the non-normative responses available to question 5, the FCI has been reported to probe for the presence of "impetus as supplied by a hit," "circular impetus," "motion implies active forces," "centrifugal forces," and "obstacles exert no force" [12]. Clearly, many different concepts are possibly represented by each of the non-normative responses. This means that we cannot take any one response to represent a single concept as the student thinks about it. When students select one of these responses, we view this as partial evidence of their thinking has to be linked with other responses to understand students' conceptions. We claim that using MAMCR on the full student non-normative responses will allow us to more fully understand students' conceptions by considering the response pattern from non-normative responses.

B. Factor analyses of conceptual inventories

Factor analysis is a statistical approach to finding latent factors in a body of data. In order to find such latent factors in a conceptual inventory, the questions are correlated with each other forming a correlation matrix. Then eigenvectors are found from the correlation matrix. These eigenvectors are "factors" which identify groups of questions which make up the factor. In factor analysis the researcher uses a cut point to determine a number of factors which will then be interpreted. The interpretation of these factors involves the researcher identifying and describing commonalities among questions that load on the same factor.

Factor analyses of the FCI often have failed to produce a strong factor structure. Huffman and Heller's analysis found a two factor structure in the 1992 version of the FCI [13]. The two factors they interpreted were "kinds of force" and "Newton's third law." Huffman and Heller used this to raise questions about what the FCI actually measures, since it did not empirically identify the six conceptual dimensions of the force concept proposed by Hestenes *et al.* [2]. A subsequent analysis by Scott *et al.* [10] used a larger data corpus than Huffman and Heller. Using exploratory factor analysis Scott *et al.* found an unrotated single factor solution, which they describe as "Newtonian-ness." They also used a nonorthogonal rotation and found a five factor structure among FCI data, which they then interpreted. The five factors identified by Scott *et al.* include the following: 1. Identification of forces, 2. Newton's first law with zero force, 3. Newton' s second law and kinematics, 4. Newton's first law with canceling forces, and 5. Newton's third law. While these five factors represent categorizations of questions, they do not provide insight into the ways in which students respond. Further, the instructional utility of this factor structure would again only allow for evaluating the extent to which students answered the questions in this cluster in a way that is consistent with the normative view or not. While both of these factor analyses represent an improvement on simply understanding FCI raw percentages they are crude as a diagnostic.

One of the underlying assumptions in factor analysis is that the factor structure is evident at the question level. As Scott *et al.* point out, "…all this factor does is measure the tendency for a student to get a question right given that this student has answered another question correctly; i.e., the factor analysis looks for structure in the correlations between correct answers to questions." All factor analyses of multiple choice tests will look for correlations among normative answers to questions. In cases where questions on an instrument probe singular concepts, factor analysis is a useful data reduction technique. However, the conceptual inventories typically do not meet this criteria. Factor analysis fails to utilize the additional information about what non-normative response students chose. For this reason, we developed Module Analysis for Muliple Choice Responses, a network approach that uses community detection on the non-normative responses as a method that promotes the use of the conceptual inventories as diagnostic instruments.

C. Background on network analysis

Network analysis is a set of methods which are useful in analyzing data which are relational in nature. These methods have been used in a variety of settings and many disciplines. Network analysis is a robust methodology which has been fruitfully used in a variety of settings; disease transmission, friendship, and membership in a karate club. In educational settings, network analysis has been used to analyze participation in a student learning center [14], to understand the role that social interaction plays in future grades [15], and to map epistemological transitions during a problem solving setting [16]. Grunspan *et al.* [17] published an insightful overview of network methods in education research. One of the reasons we have chosen to employ network analysis to the FCI data is that if we expect that concepts are robust, then we should believe that student responses are related to one another. Network analysis allows us to visualize and model patterns, and test hypotheses based on these relational data. Intrinsic to network analysis is identifying nodes which interact. In this analysis we analyze a projection of a student by a response bipartite network which is described in Sec. II C. This paper is unique in that it turns the analysis to student responses on a commonly used assessment of conceptual understanding.

II. DATA AND METHODS

Because this article focuses on developing MAMCR as a method for analyzing conceptual inventory data, we elected to use data that were as clean and complete as possible. Additionally, we chose to analyze FCI data because the FCI is the most prominently used conceptual inventory in physics and has been subjected to other analyses, particularly factor analyses [9,10]. To illustrate the method we analyze only the post-instruction FCI from just one cohort. We recognize the limited scope of this data set and thus we do not expect that the modules identified from this analysis should be found in preinstruction data or across multiple universities.

A. Data sample

Data were collected from 143 first year physics majors at a Danish university. The vast majority of students were ethnic Danes while 2–3 students where of Arabic ethnicity. 78% were male. The FCI was administered pre- and postinstruction. All but six students completed the entire FCI. Two students responded to 28 questions, while four more responded to 29 questions. The statistics were $\langle \text{pretest}\% \rangle = 65 \pm 22$, and the average postscore was $\langle \text{post} - \text{test}\% \rangle = 81 \pm 18$, with an average normalized gain of $\langle g \rangle = 0.43 \pm 0.45$ and Cohen's $d = 0.42$. Students in this cohort generally have high pre- and postscores, which means that normative responses are abundant.

B. Overview of data processing and analysis

In this section, we provide a detailed description of how data are processed in preparation for interpretation. However, it is useful to have a broad view of the process as well. First, the response data for each student are stored as a bipartite network (Sec. II C). The bipartite network is then projected (Sec. II D) and normative answers are removed. The important underlying structure, or "backbone," of this response network is then extracted (Sec. II F). The InfoMap community detection algorithm then is used to identify modules of response items which cluster together (Sec. II G). These modules are then interpreted.

C. From FCI responses to response networks

In order to convert FCI responses to networks we first constructed a matrix of STUDENTS × RESPONSE ITEMS. The FCI contains 30 questions each with 5 different response items, resulting in 150 possible response options. For each student, we indicate their responses to each question with a 1 if the student chose a particular response item and 0 if the student did not. Thus, we start out with the raw responses in a matrix of 143 STUDENTS × 150 RESPONSE ITEMS, as seen in Fig. 1.

This matrix can be treated as a network with two different kinds of nodes, students and response items. This is commonly known as a bipartite network [18]. In the bipartite network a student is connected to all the response item (s)he has selected on the test. Thus, no two response items are directly connected, just as no two students are directly connected. Twenty-four items were not chosen by any student, leaving 126 response nodes in the bipartite network. The bipartite network contains 269 nodes with 4282 links. In the bipartite network 143 of the 269 nodes represent students, and 126 nodes represent response items. The number of links in the bipartite network is equal to the sum response items chosen by the entire population. Since two students chose 28 response items and four students chose 29 response items, while the remaining 137 students chose 30 response items, the

$$\left(\begin{array}{ccccc} & 1a & 1b & \dots & 30e \\ \text{Student1} & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{Student}i & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{Student}N & 0 & 0 & \dots & 1 \end{array}\right)$$

FIG. 1. The raw data are initially stored in a 143 × 150 matrix because our data set included 143 students and 150 response items.

number of links in the bipartite network is 4282. As a first exploration of the bipartite network, we use the ForceAtlas2 lay-out algorithm [19] implemented in software Gephi [20]. This algorithm treats links as attractive forces and has a built-in repulsive force for all nodes. The algorithm then finds a lay-out configuration that minimizes the energy of the system; see Fig. 2.

In the bipartite network, the normative responses form a central cluster with two non-normative responses on the periphery of that cluster. Because the layout algorithm seeks to minimize the energy, nodes with many connections are at the center of the graph, and less connected items are further from the central core as seen in Fig. 2. The bipartite network shows relationships between students and response items, and it is also possible to identify some patterns among responses, for example, the central cluster of primarily normative answers. However, to focus on the detailed relationships between response items for this cohort, we collapse the bipartite network into a response item network.

D. Bipartite network projection

Bipartite networks can be projected through matrix multiplication into two separate networks, in this case STUDENTS × STUDENTS and RESPONSES × RESPONSES. We did this using the igraph package [21] in the R environment [22]. To identify relationships between these responses, we analyzed the RESPONSES × RESPONSES projection of the bipartite network. In this response network, two response items are connected if at least one student has chosen both responses (see Fig. 3), and the weight of the connection is equal to the number of students who have chosen both responses. Figure 3 shows how two students response patterns are projected into a response network.

One problem with this network is that all the non-normative response items chosen by a student are connected. This means that the response network is extremely dense. While it is possible to discern some characteristic pattern, most are obscured by the large number of

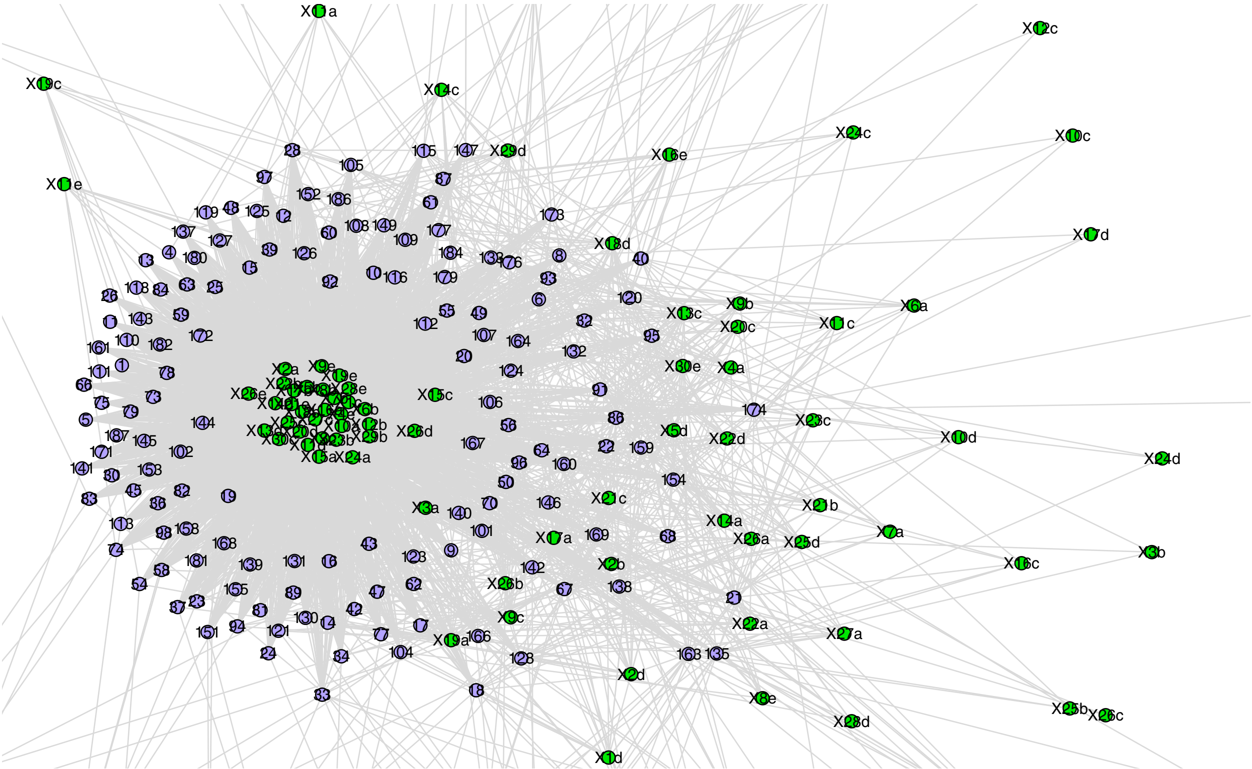

FIG. 2. Cut out of bipartite network showing students (purple) and response items (green). The normative response items form a central cluster in the middle. Two non-normative response items, 15c and 26d, are on the outskirts of this cluster. The eastern parts of the network seems to show more complex patterns.

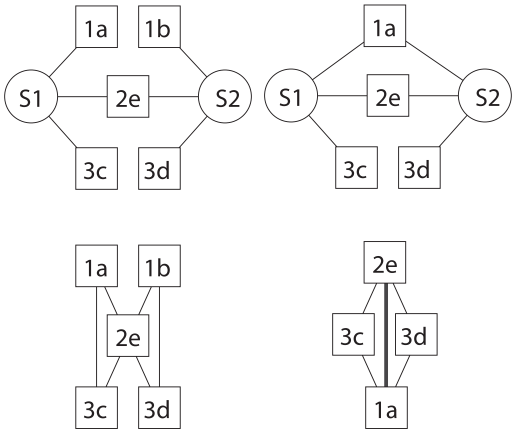

FIG. 3. Two example schematics of how a bipartite network (top row) is projected into a response network (bottom row).

connections. This is reflected in the degree distribution of the network; almost all nodes have a large amount of links—even response items that were only chosen by one student might be linked to many other items based on that student alone.

E. Removal of items

Because 24 items were not selected by any student in this cohort the bipartite network contains 126 response item nodes. We have removed the 30 response items that correspond to leaving the response network with only 96 nodes. As seen in Fig. 4, normative items were chosen much more frequently by students in this cohort than non-normative responses. When we applied the procedure described above, we found that these items formed a tight cluster which obscured the relationships between nodes representing distractor items. This meant that we could not discern any meaningful patterns and our later attempts at finding different modules (see Sec. II G) were also unsuccessful. To circumvent this problem, we removed normative responses in the following analyses. Removing the 30 normative item responses and the 24 non-normative response items which had not been chosen by any students lead to the network of 96 item response nodes and 1788 links.

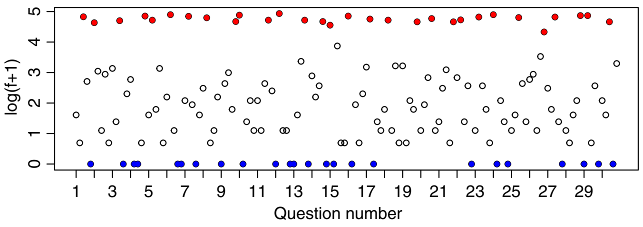

FIG. 4. Normative responses (red) have been chosen much more frequently by students than non-normative (white and blue). In a network, they will become hubs or attractors that yield little information about student thinking. Some responses (blue) have not been chosen at all, and thus do not show up in the response network. Our analysis focus on non-normative responses that have been chosen at least once (white).

We stress that this is a choice made on the basis of this cohort, while other cohorts may show different patterns. It may be that another cohort will show preferences for other items, and that another set of response items will be as frequent and connected as the normative items in this case. In such cases, they will provide some information in themselves, and we suggest that researchers analyze these separately from the rest of the network. We have not developed a systematic method to determine when this would be appropriate, but future development of this method might investigate the information lost or gained by removing nodes.

We are confident that we have gained information about residual structural patterns of non-normative responses, but the drawback is that we cannot investigate how these patterns relate to normative responses.

F. Backbone extraction of response network

A common challenge in network analysis is the reduction of data and a number of algorithms exist to deal with this issue [23]. It is akin to identifying signal and reducing noise. This process is often described as identifying the backbone of a network. In our response network the majority of connections between items were established on the basis of one or two students choices. This means that many connections between response items might be considered random. One way of dealing with this problem is define a cutoff and remove connections that are below this predefined limit. However, this strategy has been shown to potentially remove valuable information stored in weak links [23].

To address this problem of sparsifying networks while keeping important information, Foti *et al.* [23] developed the locally adaptive network sparsification (LANS) algorithm. The LANS filter works by comparing links locally for each node. A link from a node is kept if its weight is larger than or equal to a percentage of other link weights. For example, with an α level of 0.05, a link from a node needs to be larger than or equal to 95% of all other links from that particular node. This means that if all links have equal weight, all links are kept, but if one link has a higher weight, that link is kept while the others are deleted.

The LANS algorithm evaluates links based on significance of the link for one node. A link that is significant for one node might be insignificant for another node. For example, for one node, a link with weight 1 might survive, while for another node the same link is insignificant. We chose to keep all links that were significant for at least one node. This guarantees a connected network, since by construction, a link will always be important to at least one node.

Figure 5 shows the structure of the backbone including 96 nodes and 401 links. The backbone network is

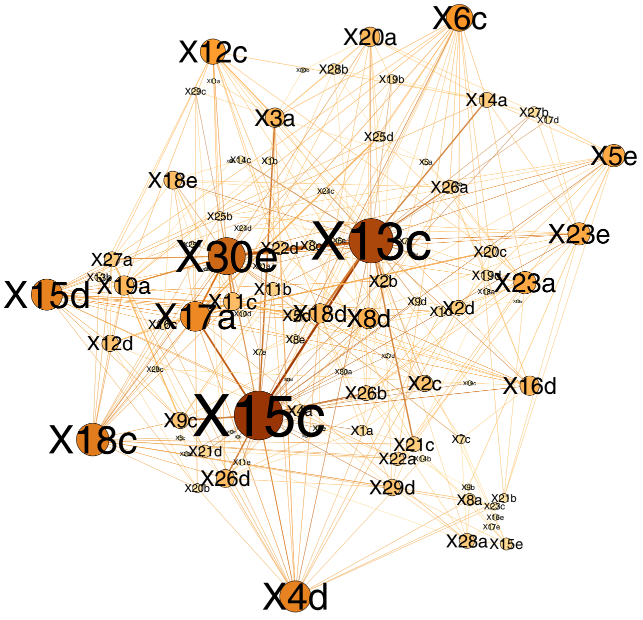

FIG. 5. The response item backbone network also contains 96 nodes but only 401 weighted connections. Node and font size represent item frequency, node color number of connections (red means more). Very infrequent nodes (small font sizes) can be viewed by zooming in on the image online.

dominated by the node that represent items 15C, while 30E is the most central. Item 30E shares a link of weight 17 with 13C. This means that 17 students picked both of these, and this is the strongest (highest weight) link in the network. Two-hundred sixty-three links have weight one (weight $= 1$, so the majority of links in the backbone are based on one student linking two nodes. It is hard to claim that two nodes are meaningfully connected just because one student chose both items. On the other hand, 15C has many connections of weight one, which may indicate that this item has a kind of random attraction. Thus, the strategy of keeping as many links as possible warrants care in the interpretation; links that are locally significant might not be globally relevant.

G. Partitioning response network using InfoMap

Network theory offers many methods for detecting communities within a given network [24]. Communities are groups of nodes that share more connections within the group than with other nodes outside the group. InfoMap [25] is one algorithm (among many) that has proven both stable [26] and useful in physics education research [16,27,28].

The way InfoMap partitions a network into communities can be understood in terms of a random walker traversing the network using links between nodes. For each node this results in a node visit frequency, which can be understood as the information flow through a node [29]. For the response network, it can be seen as a simulation of how the cohort answers the FCI. For example, many students have chosen both 30E and 13C, which means that the walker is likely to pass between these two nodes. Furthermore, students who chose 30E have chosen a diverse set of other items, which boosts the flow through node 30E.

From here, the idea is to "envision a communication process in which a sender wants to communicate to a receiver about movement on [the] network" [29]. The task is to use as little information as possible while still conveying all information about how students answered the FCI. A simple strategy is to assign a code word to each node. In order to minimize the information content in the communication process, nodes that are visited more frequently get shorter codes. The optimal process for assigning codes is called Huffman coding [30], and the theoretical lower limit for the information is given by Shannon's source coding theorem [31]: $H = \sum p_\alpha \log_2 p_\alpha$, where the p_α's are node visit frequencies. Instead of assigning a unique code to each node, InfoMap partitions the network into modules and now codes only have to be unique within a module. While the random walker is within a module, the walk can be described using short code names, but jumping between modules also carries an information cost. With many small modules, one would achieve very short codes in each module, but would suffer from many jumps between modules. On the other hand very big modules would require nodes with large names within a module, albeit few jumps between modules. InfoMap minimizes the information needed to describe the random walk by balancing the need for short code names within a module with the need for few jumps between modules. The existence and composition of a particular module is dependent on how many steps the random walker takes inside the module before it finds a connection that leads out of the module. The walker will take many steps inside tightly knit parts of the network, and that forms the basis of modules. Thus, if a random walk on the response network is seen as a simulation of how students answer the FCI, then InfoMap's partition can be seen as areas in the network with particular response patterns.

For the purposes of this paper we follow Bodin [16] and make use of the visualization tools available from Edler and Rosvall [32] to generate a map (Fig. 6) of the community structure of the backbone response network. Each of the nodes in the map represent a module with an internal structure as seen in Fig. 6. In this paper we focus on the internal structure of modules to make interpretations.

H. Stability of the solution

InfoMap relies on a certain degree of randomness when finding the partition that optimizes the minimum description length. It is possible that more than one stable minimum exists, and partition algorithms find one of these for each run. We have chosen to work with a particular

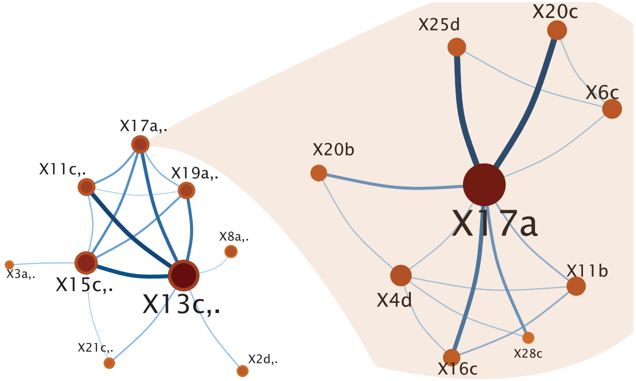

FIG. 6. Result of InfoMap partitioning of the backbone response network. (Left) InfoMap finds 9 different modules, each of which are connected to each other. (Right) Each module also has internal structure, here shown for the module labeled "X17a."

partition, M, so the significance of the results are dependent on the stability of this solution. To test the degree to which the partition is representative of all the partitions, which InfoMap is likely to find, we run InfoMap N times and compare this set of solutions, $\{m\}$, with M. We compare the partition found in two ways.

First, we mirror Bruun and Bearden and calculate the normalized mutual information [26], $I_{\rm norm}(M, m_i)$, between M and each solution $\{m_1, m_2, \ldots, m_N\}$. The normalized mutual information measures the degree to which to partitions overlap [33]. In terms of information theory, $I_{\rm norm}$ measures the information obtained about m_i given knowledge about M modified by the total information content in both M and m_i. The result for each calculation is a number between 0 and 1, where 0 would indicate no overlap and 1 indicates total overlap.

The normalized mutual information is a measure of the consistency of partition solutions on the network. However, it does not provide information about the details of any discrepancies between solutions. As physics education researchers we are interested in knowing which particular items are likely to be partitioned together. Thus, in addition to calculating $I_{\rm norm}(M, m_i)$, for each solution m_i we ask: Is response item i in the same module as response item j? Doing this with all pairs of nodes, we create a 96 × 96 module co-occurrence matrix where each element represents how many out of N times any two items where grouped together. We then reorder the matrix to show the modules of M. If InfoMap finds the same partition often the modules in M should be clearly visible in this matrix (see Fig. 8).

I. Representing response items

In the backbone network, a particular response item is represented by a node with a text label including the question number and letter. Thus, 13C refers to question 13, response C. Since our goal is to interpret the structure of modules, we found it helpful to illustrate each response item with iconic representations of the item. We have designed the icons to represent the wording in the question text and response item text as closely as possible while maintaining succinctness. When a question text includes a drawing, we have used that drawing as a template for the item response icon. Figure 7 shows three example icons representing response items 6C, 13C, and 15C. We have replaced the standard node representation of circles and text in community maps with scaled icons of the item. Thus, big icons in a module represent more important items in terms of information flow. Finally, links of weight 1 (meaning that only one student connected two items) and nodes with only links of weight 1 in a module have been made transparent in the graphical display.

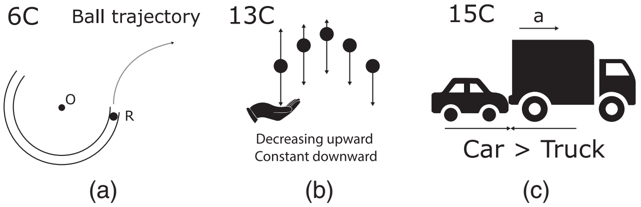

FIG. 7. Three example icons.

III. RESULTS AND ANALYSIS

Our illustrative analysis of the post FCI for this particular cohort identified nine modules. Much like the factors in factor analysis, the modules in this approach need to be interpreted. As with all interpretation, these are subjective; our goal is to provide interpretations which can serve as a starting place for this type of analysis and can be improved upon as more data are collected and as researchers find other nuance in the data. We start by analyzing the stability of InfoMap's clustering of the backbone network.

A. Variable clustering

We ran InfoMap (version 0.18.2) $N = 1000$ times, and compared each solution with the one we used. The normalized mutual information was $0.8 \pm .1$. InfoMap finds 7–9 (8.2 ± 0.7) modules with 8 being the most frequent number of modules. Modularity is a common metric used in community detection to measure how effectively the grouping algorithm identified separate communities. Modularity is defined as the fraction of edges that are grouped in the given module minus the expected fraction. The modularity for the solution we use for further interpretation has $Q = 0.39$, which is relatively high. Using the module co-occurrence matrix in Fig. 8 we elected to use the 9-module solution that seems more likely to capture subtle structural characteristics of the backbone network.

Figure 8 shows the existence of a large but not highly stable module. Many of the module's constituent items are ghosted in modules 3–5, meaning that they are sometimes grouped with items from these other modules. We call this

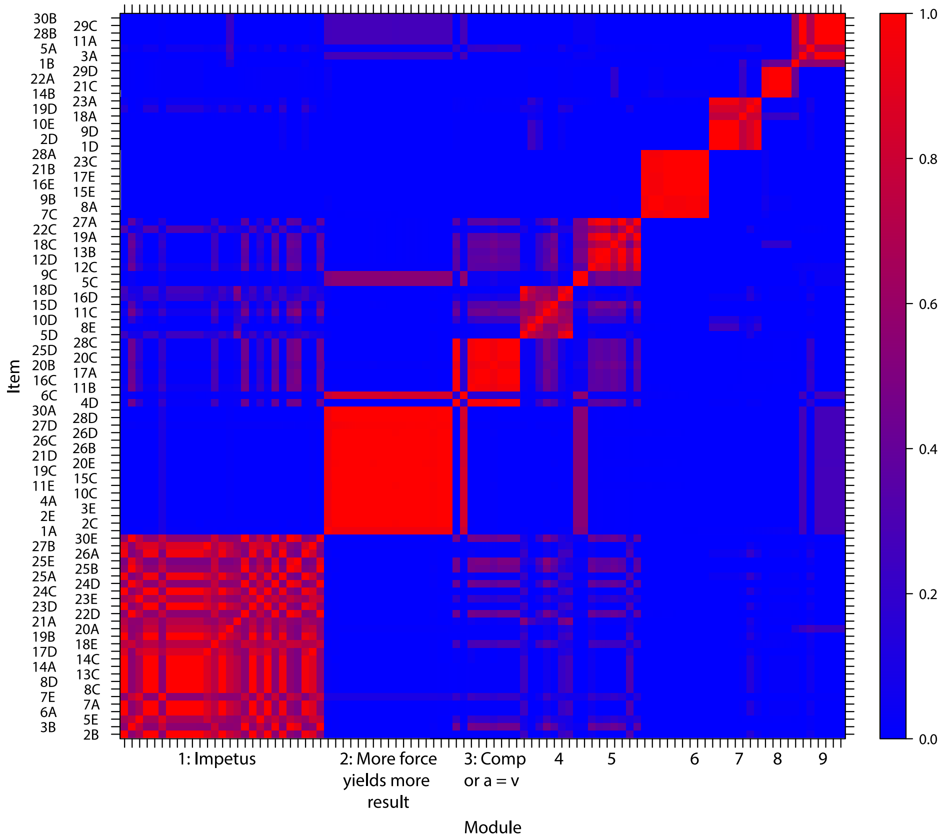

FIG. 8. Module co-occurrence matrix comparing module pairs found by 1000 iterations of InfoMap with the modules we use here. The color indicates the fraction of times two items have been grouped together, ranging from 0 (blue) to 1 (red). Note that modules 3–5 might be interpreted as one big module or as 3 separate modules.

module the impetus module for reasons that will become clearer in Sec. III C 1. In contrast, the second module—more force yields more result—is much more stable. This signifies that items in this module will almost always be grouped together by InfoMap. The exception is 6C, which is more often grouped with the elements in the third module, Comp. or $a = v$. However, 6C is a response that is not chosen frequently (only two students chose it) and thus would not contribute to our interpretation of either module. The third module also seems somewhat stable, but it also seems that it might be seen as a larger structure along with modules 4 and 5. The last four modules seem fairly stable although small.

B. Determining which modules to interpret

One of the keys to factor analysis is to interpret the factors, however, the researcher must first determine which factors to include. Researchers using factor analysis typically make a scree plot and use, as a rule of thumb, that they should interpret any factors that are to the left of the elbow. No such rules of thumb exist for this network approach. Thus our approach has been to profile modules that are prominent and interpretable—to be clear this determination is a subjective element of MAMCR. Some modules seem to arbitrarily connect response items, and upon inspection rest on a single or very few student answers. We have opted not to analyze these modules.

C. Interpreting response modules

In this section, we show how to interpret modules found using MAMCR. So that readers can follow the approach in their own research we provide a detailed discussion of the first three modules and then include diagrams of the remaining modules in the Appendix.

1. Module one: The impetus module

The first module seen in Fig. 9 has two nodes as most prominent, 30E and 13C, then seven nodes which contribute, but have a lesser overall contribution, and ~30 additional nodes which are infrequent and are represented in the background. The module has two nodes, 30E and 13C, at the core and a number of other nodes that are arranged in a starlike pattern. This suggests that the nodes at the center act as attractors and should be the core of the interpretation. Both of these items include a force of motion or an impetus force which is acting on the object (tennis ball in 30E and ball in 13C) after the object is no longer in contact with the racket or hand, respectively. The next most prominent node 22D includes a rocket accelerating to a maximum velocity

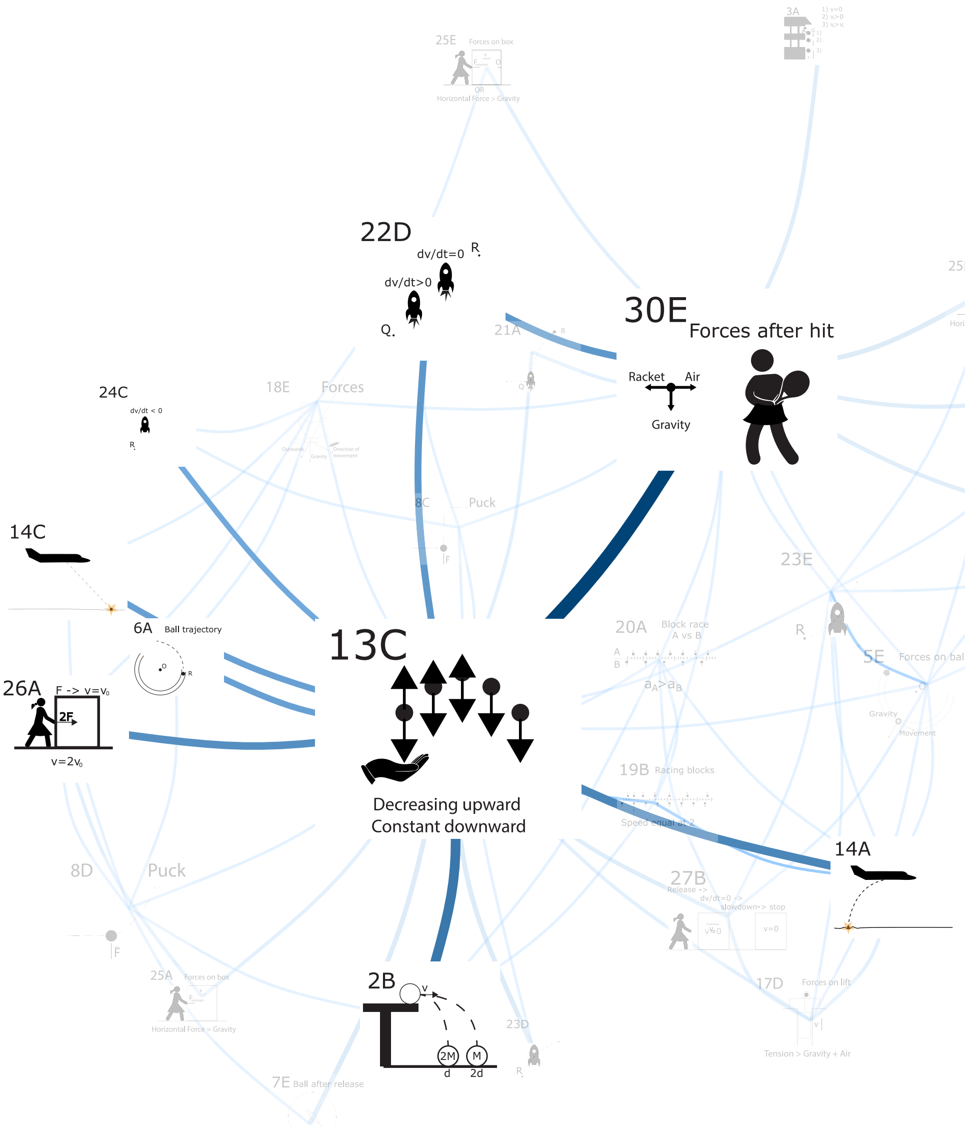

FIG. 9. Illustration of the impetus module using pictograms. Each pictrogram represents a response item.

while the rocket is on. For the rocket in 22D, while it is not itself an example of impetus, it is not inconsistent with a diminishing impetus model for force, where force diminishes over time. A number of other nodes also contribute to this module, 6A which shows a ball continuing along a curved path after leaving a curved track. Selecting response 6A is consistent with an impetus model for force. Two other nodes, 26A and 2B, both are consistent with use of an impetus model. Question 2 compares two balls with different masses rolling off a table, in response 2B the ball with twice the mass lands half as far from the base of the table. Assuming an impetus model, the reasoning in this case would be two balls with the same impetus, but different masses would have different (and proportional) outcomes. The heavier mass would then land only half as far away as is represented in 2B. Response 26A is similar in that a person who increases the force on a box would then double the velocity of the box.

Two choices, 14A and 14C, which are included in this module are mutually exclusive, meaning that a student must choose one or the other. In question 14, students are asked to predict where an object, dropped from an airplane, will land. The two choices in this module show the object far in front of the airplane and far behind the airplane. It is difficult to understand these choices from an impetus perspective. Because they come from the same question, one possibility is that a student with a strong impetus model has reason to disagree with the normative choices and these are what is left.

Finally, in this impetus module is a background of a large number of other nodes which are not prominent, meaning that they have not been selected by a large number of students. One way to interpret this is that students with an impetus model have many and varied other non-normative responses. The varied nature seems to indicate that students are not using a consistent approach to arriving at a response, but instead have many *ad hoc* approaches to arriving at an answer. This is consistent with a knowledge-in-pieces view of student answer making [34].

2. Module two: More force yields more results

The second module is a starlike module, see Fig. 10, which we interpret as "more force yields more results." At the center is response 15C. Question 15 has a car pushing a truck while accelerating, response 15C is that the car exerts a greater force on the truck than the truck on the car. Alone, this could have several interpretations. Three additional nodes help interpret the main node. Node 4A shows that in a collision the truck exerts a greater force on the car. Node 26D shows a woman pushing a box; when she doubles the force she exerts, the velocity increases for a time and then becomes constant. Based on the same question, node 26B shows a woman pushing a box; when she doubles the force she exerts, the velocity of the box doubles. Node 28D shows that a person pushing off of another person exerts a greater force than the person being pushed. In all of these nodes as a force is exerted or changed the results change. The two other nodes which are prominent, 11E and 3E, show that after a kick a puck has no forces acting on it and a ball after being dropped has a force of gravity and a force due to air. These two nodes are different than the others in this module in that the force in question is not changing. Although these last two nodes do not contribute to the more force yields more results, they are not inconsistent.

3. Module three: Force as competition or indistinguished velocity and acceleration

The third module, seen in Fig. 11, is interesting in that two competing interpretations are available and follow-up work is needed in order to identify which is more appropriate. This third module is again a star module with one central node, 17A, and five arms. The central node

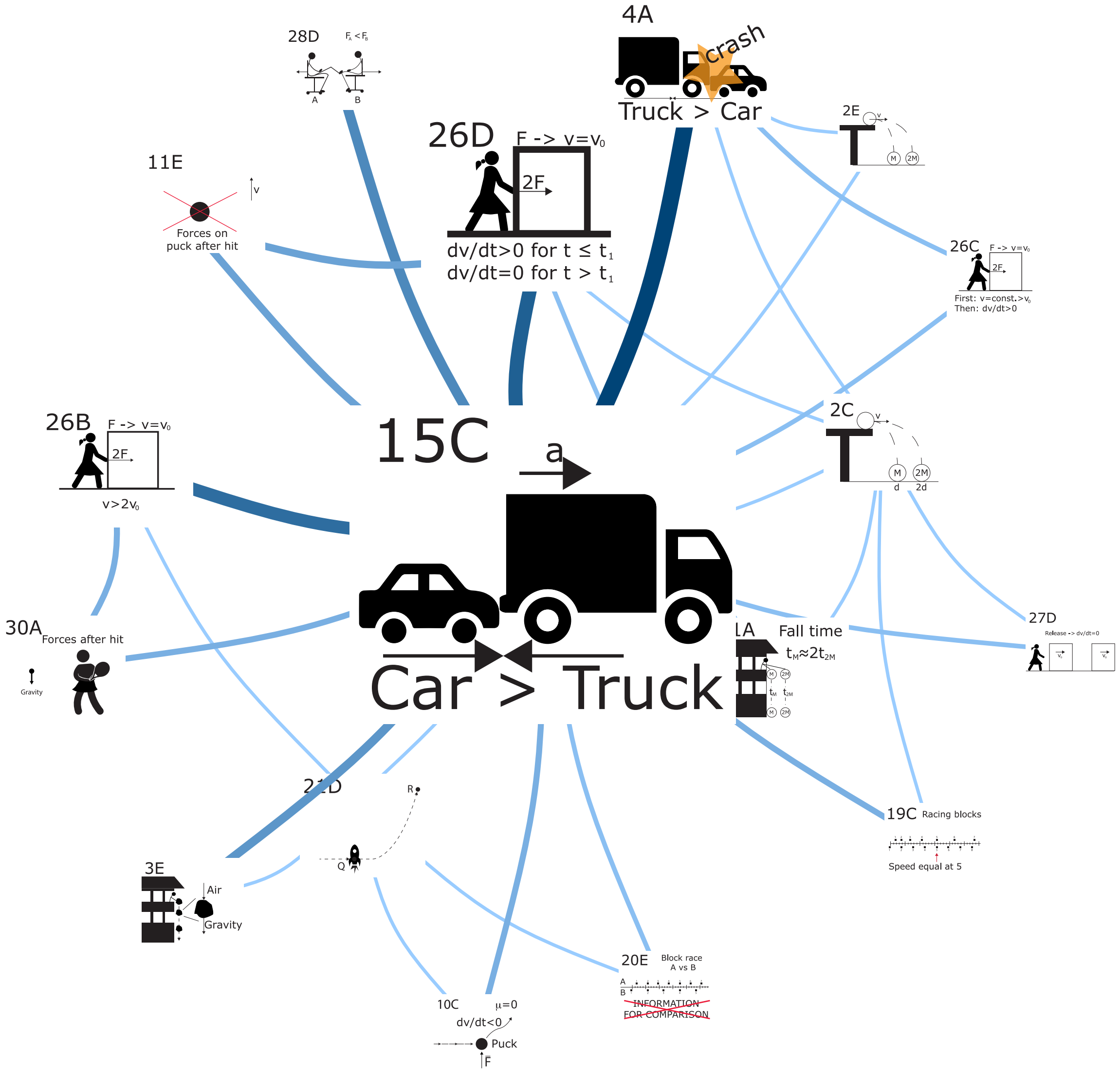

FIG. 10. Illustration of the more force yields more results module using pictograms. Each pictogram represents a response item.

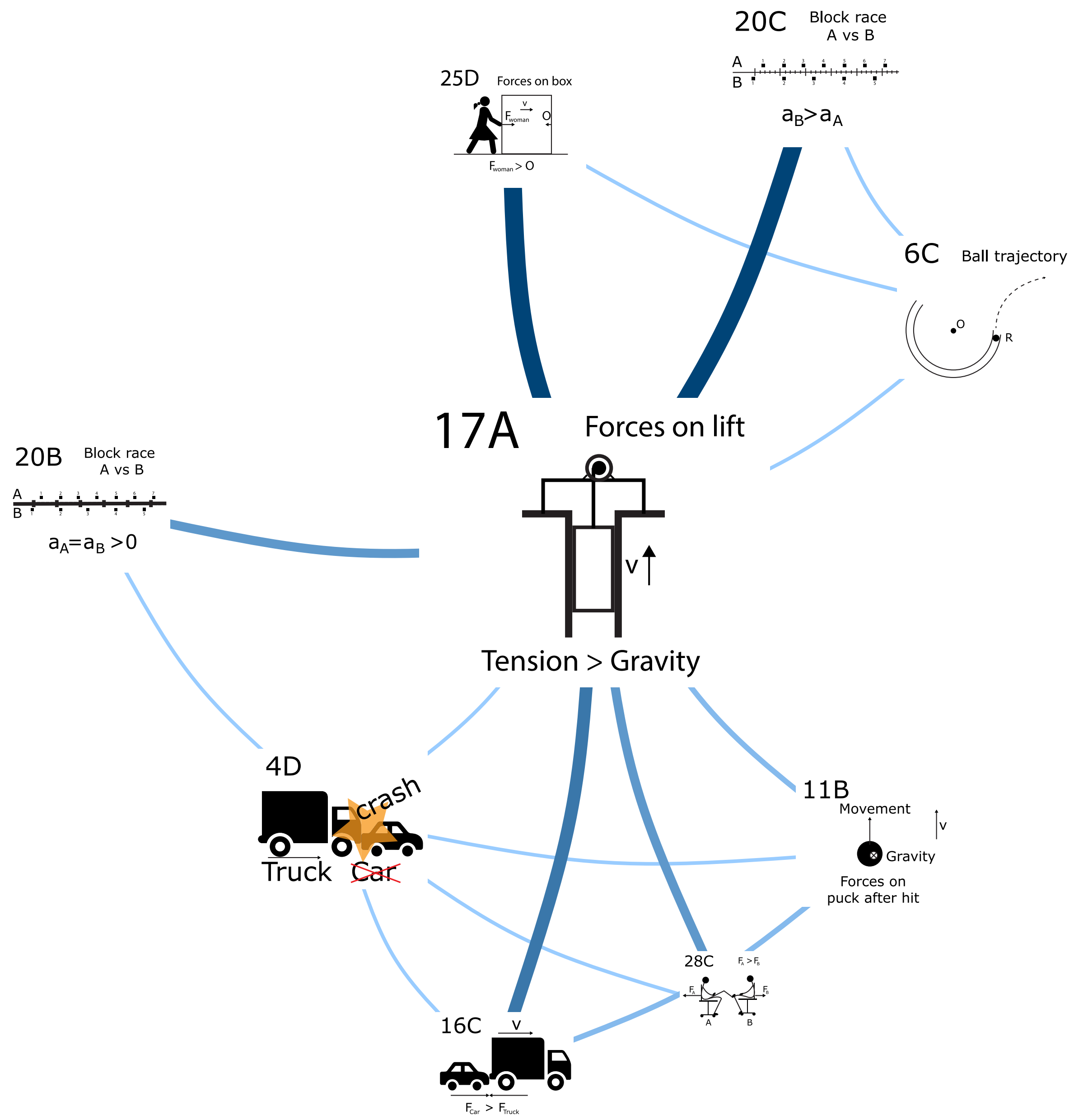

FIG. 11. Illustration of the third module, which has two main interpretations; force as competition or lack of distinction between velocity and acceleration using pictograms. Each pictogram represents a response item.

involves a box being lifted by a rope at a constant velocity; in response 17A the student ranks the force of tension as greater than the force of gravity. Three other nodes, 16C, 25D, and 28C, show an active agent exerting a force on another object. In two of these cases (16C and 25D) the object is moving at constant velocity. In all three cases the students rank the active agent as exerting a greater force than the passive agent. Alone these four suggest that students see force as a competition, with objects moving in the direction of the "winning force." Two other nodes contribute to this module, both from question 20. Question 20 shows two blocks, A and B, moving at different constant speeds, with object B having a greater speed. Response 20B has the accelerations equal but greater than zero, and response 20C has the acceleration of B greater than A. These two choices, which are mutually exclusive, suggest a different interpretation that students do not discriminate between velocity and acceleration. This second interpretation can similarly apply to the first three of the other four nodes in the module, 17A, 16C, and 25D. In these cases the students may have an appropriate understanding of forces, but not have distinguished velocity and acceleration, which would mean that the other choices are in accordance with a Newtonian view. A third possibility is that neither interpretation alone, force as competition or undistinguished velocity and acceleration is correct, but that we find that students who have a force as the competition view of force may also have difficulty distinguishing velocity and acceleration. This combined view is particularly intriguing as it suggests that treating the responses that students choose on a multiple choice instrument such as the FCI as independent is incorrect, and that a method that handles relational data such as network analysis is more appropriate for understanding student conceptions.

IV. DISCUSSION

We believe MAMCR has utility in identifying latent structure within the response patterns on conceptual inventories. Treating the FCI as an exemplar, we have applied MAMCR to post-test FCI data from a single first-year university cohort. In this section, we illustrate the utility of MAMCR by first comparing the method with factor analysis, by suggesting how MAMCR could be used in the context of a classroom, and finally by expanding the use of the methodology.

A. Comparing with factor analysis

Factor analysis has previously been employed in an effort to understand the conceptual structure of the FCI in a different way than we propose in this paper. For researchers it is interesting to know how the knowledge MAMCR offers is different from the knowledge offered by factor analysis when analyzing conceptual inventories. The fact that factor analysis of FCI results have been published twice [9,10] adds to the value of the FCI as an exemplary conceptual inventory.

When comparing the results of an MAMCR analysis with Scott *et al.* [10] we see both overlap and distinction. Because factor analysis uses correlations among questions answered in concordance with Newtonian mechanics and MAMCR uses selection of non-normative responses, we cannot compare results exactly. In both our impetus module and their first factor (identification of forces) two questions (13 and 30) are present. However, these two differ. Scott *et al.* describe the identification of forces factor as grouping questions where students are able to determine the forces on an object according to the normative view. What is different between our impetus module and their factor is that we identify the forces that are incorrectly included in the interaction as typically being impetus forces. The differences between MAMCR and factor analysis may be related to how each method is built on normative or non-normative responses. These differences between our impetus module and their identification of forces factor can have important instructional impacts. Instructional recommendations would be to focus on either having students identify forces or to address an impetus model.

A second notable similarity is that their fourth factor (Newton's first law with canceling forces) has a substantial overlap with our third module (force as competition or indistinguished velocity and acceleration). We both agree that items 17, 25, and 16 all include forces in opposite directions; our interpretation is that force is seen as a competition, while theirs is that students see these as Newton's first law questions. However, the additional items in our third cluster make a lack of distinction between velocity and acceleration a distinct, yet plausible interpretation.

One of the distinguishing features of our analysis is that our second module (more force yields more results) includes all the items that are identified as difficult and thus do not show up in any factor (26) structure, drop out in the nonorthogonal rotated factor solution (3), or are not present in the single factor solution (15). This might be a by-product of our method given that our method removes normative answers. A second interpretation of this difference with the factor analytic method is that Ohm's p-prim, more force yields more results, is a strong conception that students hold.

In conceptual inventories it is difficult (impossible?) to write questions that probe singular concepts, such as asking about force without asking about motion. Thus, using a factor analysis approach may not offer clear interpretation as it does when questions probe singular concepts and responses are Likert-scale such as the C-LASS [35]. In the analysis of conceptual inventories, we expect that using our method, which accounts for the variety of non-normative responses in the clusters, we will be better able to diagnose student thinking. One of the challenges to our approach is to find ways to account for normative answers, currently we have excluded these due to the overwhelming clustering that happens when normative answers are included.

B. Classroom utility of network analysis of conceptual inventories

We see three ways in which network analysis of conceptual inventories can be useful at the level of the classroom. First, MAMCR may with further developments be usable by classroom teachers, if all of the steps leading to visualization of modules for a given conceptual inventory can be automated. It is certainly not our intention to expect classroom teachers to utilize this level of analysis. At the same time, we believe this approach provides useful and novel insight into student understanding, far superior to simply knowing how many responses are in concordance with the normative view a class or even a student achieved. Further, as we described in Sec. IV A, the level of analysis is superior to that of factor analysis. Thus, one challenge is to develop the methodology sufficiently to be able to automate all steps, packaging the analysis such that classroom teachers can use these methods to analyze their own classes. In principle this is not unattainable.

Second, extending the method beyond the FCI as an exemplar is needed. We expect that analyzing additional conceptual inventories and varied classes of students (within any one conceptual inventory) will provide additional insight. In particular, understanding the relative frequency and robustness of modules has the potential to inform instruction by allowing enhanced diagnostic capacity.

Third, making conceptual inventories more useful diagnostic instruments will allow researchers to connect instruction with student conceptions on a more fine-grained level and to help develop interventions that target the ideas in classes. Specifically, it may help teachers to find activities that address whole modules. For example, an activity related to impetus might be employed taking the impetus module identified by MAMCR as a starting point. In relation to this point, MAMCR provides a new suite of tools for researchers to probe student understanding which can be incorporated into other research methodologies.

C. Limitations of this study

The analysis presented here is limited in three ways: (i) the cohort, (ii) the instrument, and (iii) the choices made in the implementation of MAMCR. In terms of the cohort, we have analyzed a limited data set in terms of student numbers, student diversity, and learning context. Thus, we advise that the results of this analysis should not be generalized extensively. In terms of the instrument, we have only shown an illustrative example of one conceptual instrument, and the FCI might be special in terms of form and content. However, for this section we elaborate on the limitations in terms of the choices one has to make during analysis. One limiting feature of the analysis is the choice of the sparsification method. Choosing to remove correct answers was based on the prevalence of these answers in the network, but we have not provided a systematically informed threshold for when to remove items. Choosing LANS and choosing to keep links that are relevant if only to a single node represents a choice that aims at maximizing the connectedness in the network. Other choices would aim at finding only very strong connections. Finally, choosing the particular partition to analyze in depth also represents a choice on our part. All these choices carry a subjective element, and they affect the subsequent interpretation. However, the transparency of the method ensures that it is possible to compare MAMCR results across various choices. This fact may be incorporated in future research using MAMCR.

D. Further development of methodology

With this paper, we have identified and demonstrated the potential of a MAMCR network analytic approach to analyzing conceptual inventories, particularly the FCI. This is an initial effort; here we outline three different ways that this methodology can and should be developed.

First, we used MAMCR on FCI as an illustration of the method. We believe that the modules we have identified can serve as a starting point for a rigorous study of modules latent within the FCI. A rigorous study would include various populations as well as pre- and post-tests. Qualitative methods, such as a microgenetic interview approach [36], would serve to identify how modules are manifest in students. Additionally, we welcome other researchers' perspectives on the modules. Interpreting these modules is subjective, and while we have been deliberate in our approach to interpretations, certainly other interpretations are possible, and the utility of the methodology will be enhanced by further input and interpretation. In so doing, it might be interesting to compare the resulting modules with the previously proposed taxonomy [2,12] to investigate the degree to which modules and taxonomy overlap or can inform each other. Here, the potential lies in the fact that the modules produced by MAMCR are driven by the data (although the reasoning behind analyzing the one partitioning we analyzed here has subjective elements), while the taxonomy was informed by experts' design of the FCI.

The second line of research is to expand the use of this methodology to various other settings. Our initial analysis was only with post-instruction data. Thus, analyzing predata could provide insight into the dynamics of the formation of these modules. Further, the data we analyzed are from a cohort of students in a Danish University; analyzing further data may identify different modules, and potentially highlight cultural differences. In each of these approaches, we will attempt to stratify our data by institution type and by teaching methodologies, necessitating a large data corpus.

The third line of research, following on the first two, would be to expand the methodology beyond the FCI to other instruments. Because we expect that analysis of modules of answers is a more robust approach to analyzing conceptual inventory data, a logical next step would be to move to other domains. Simultaneously, further development of this methodology to incorporate normative answers without overwhelming other network structure is necessary but not straightforward.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant No. PHY 1344247.

APPENDIX: ADDITIONAL MODULES

We provide an overview of the six additional modules identified in the appendix.

Module four (see Fig. 12) seems to be another variant of the impetus force module. Three choices are prominent: 11C, 18D, and 5D. In all three of these choices, the object moving has a force of movement on it, but are otherwise correct.

Module five (see Fig. 13) has three responses, 9C, 19A, and 27A, that may indicate a lack of understanding of velocity. In 27A an object comes to an immediate stop after

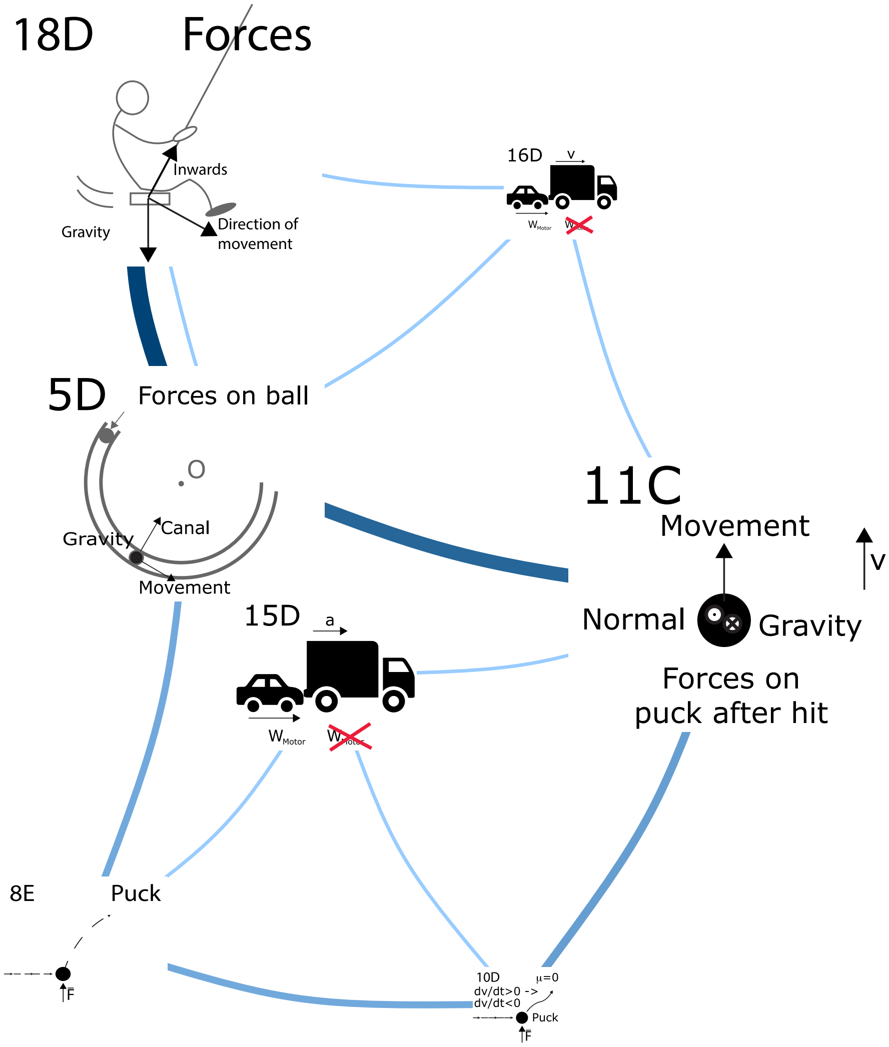

FIG. 12. Module 4.

a woman stops pushing the object. In 19A two blocks are never seen to have equal velocity. And in 9C, after a puck receives a kick the velocity is seen as the sum of the two velocities (v_0 and v_k).

Module six (see Fig. 14) is interesting in that the similar features of the items chosen seems to be that instantaneous change happens. In choices 21B, 23C, and 8A an object that receives an impulse immediately begins traveling in the direction of the impulsive force.

Module seven (see Fig. 15) has two responses, 1D and 2D, that indicate that mass is an important (though not linearly related) factor in time of fall and distance traveled. This module may be an iteration of the more force yields more results module, however, interestingly, in both 1D and 2D the time and distance are not a factor of 2 in proportion to the mass.

Module eight (see Fig. 16) includes only two items which are part of a paired set of questions (21c and 22A). In the scenario, a rocket is turned on, the questions ask about the path of the rocket (21) and the acceleration of the rocket (22). These two choices are consistent with each other. In 21C the response is that the rocket travels along a straight path, and in 22A the response is that the acceleration is zero. This may indicate a misunderstanding of the situation. If the understanding is that the rocket provides only an impulsive force, like the kick on the puck in question 8, then this pair would be correct.

Module nine (see Fig. 17) has response 3A at the center where an object dropped reaches a maximum velocity. Other responses in this module include 11A and 5A where the only force on the object is gravity, and 30B where there is only gravity and a force of a “hit.”

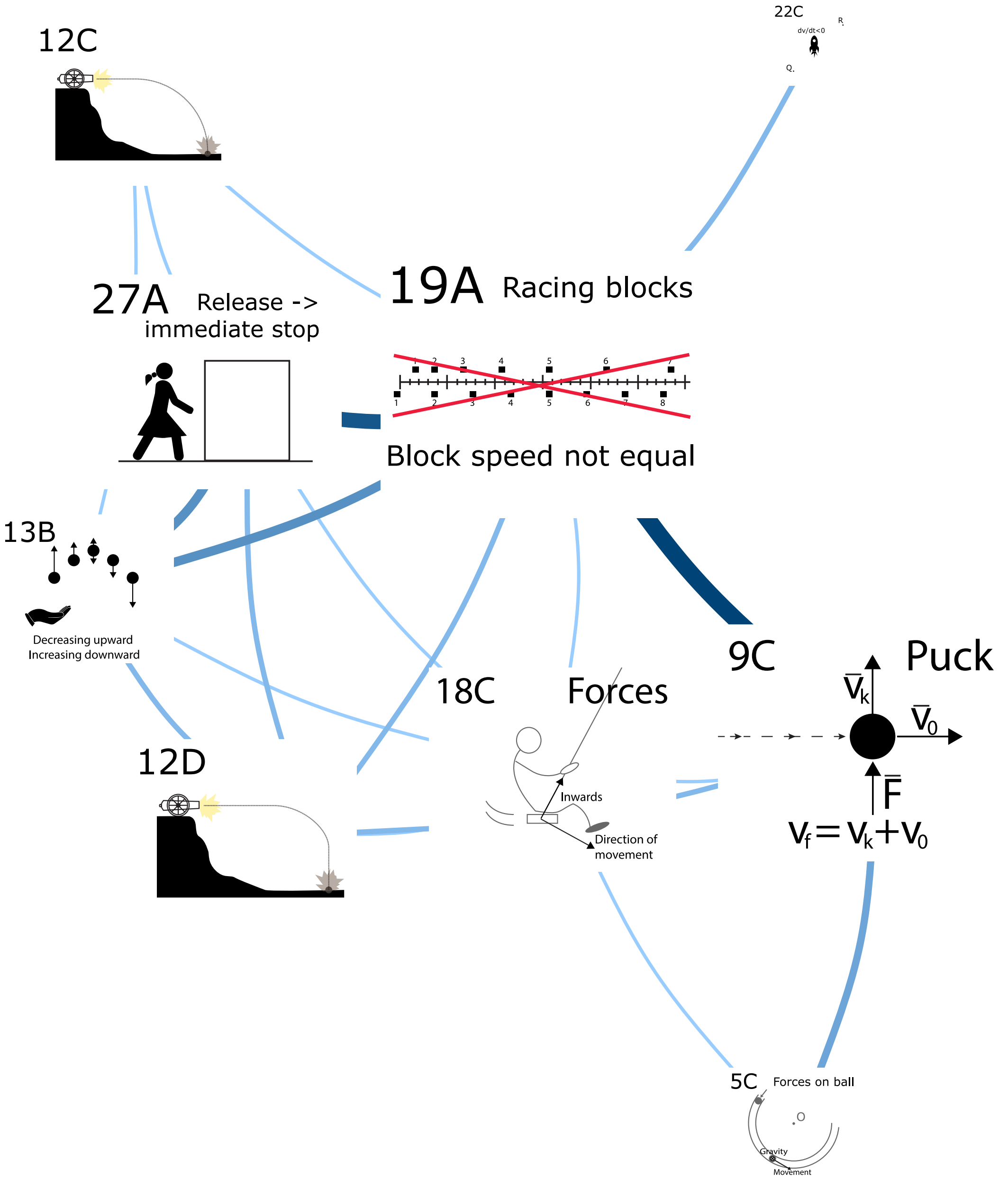

FIG. 13. Module 5.

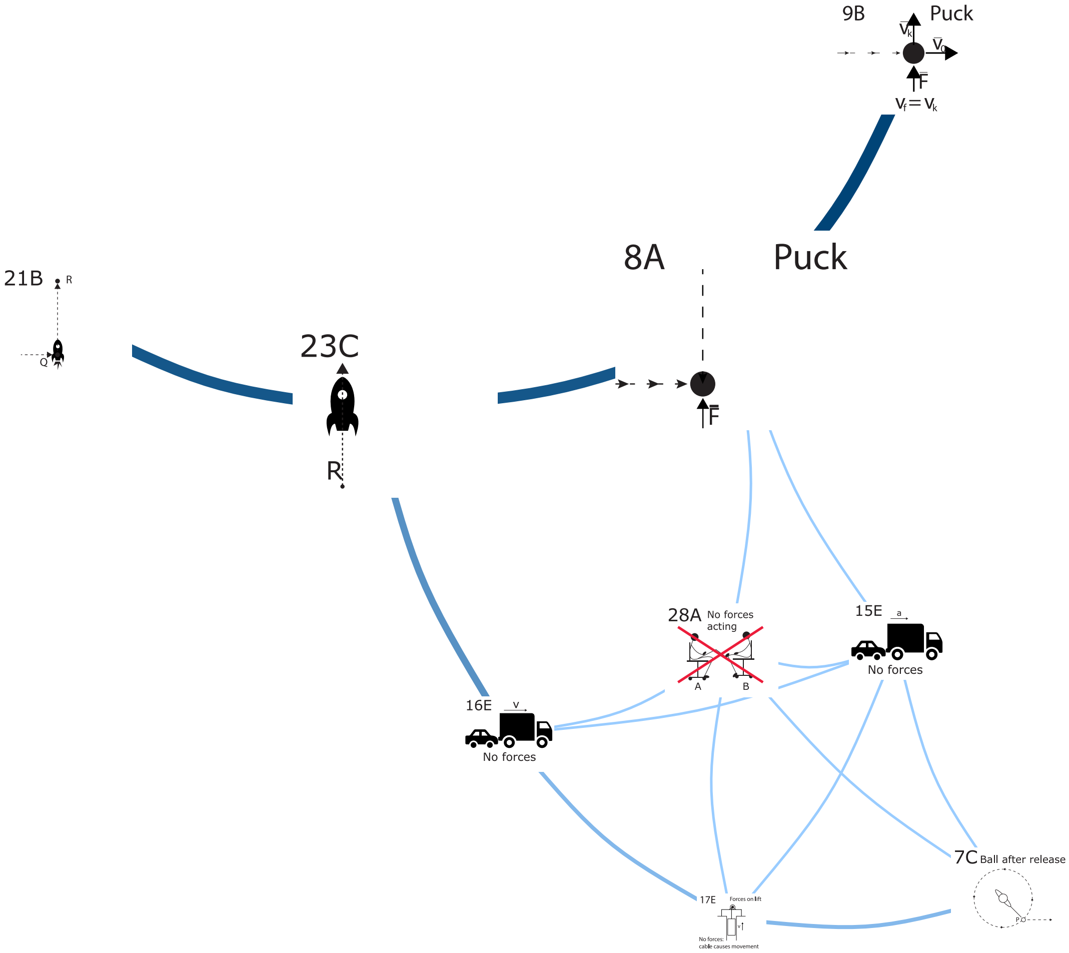

FIG. 14. Module 6.

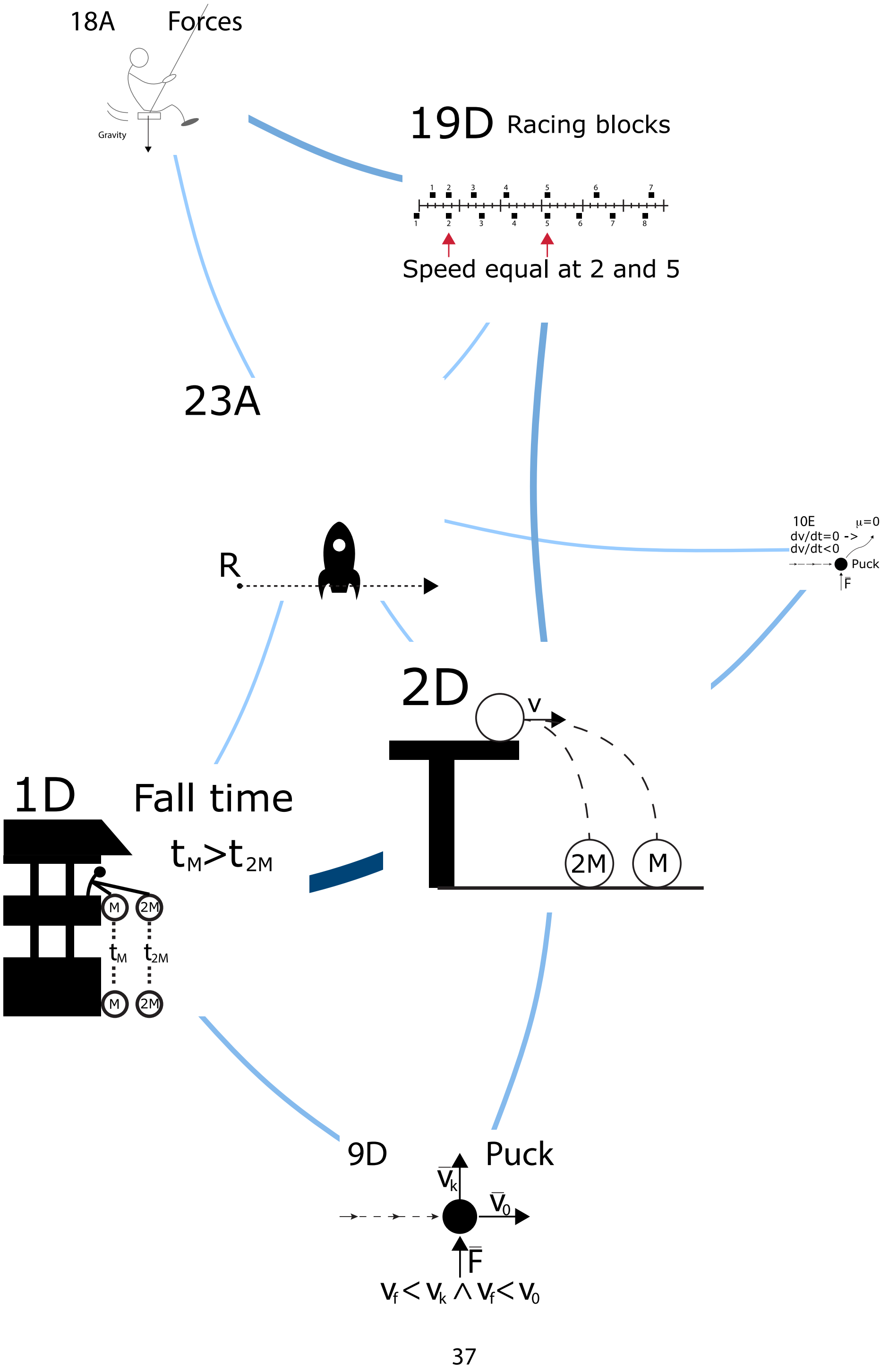

FIG. 15. Module 7.

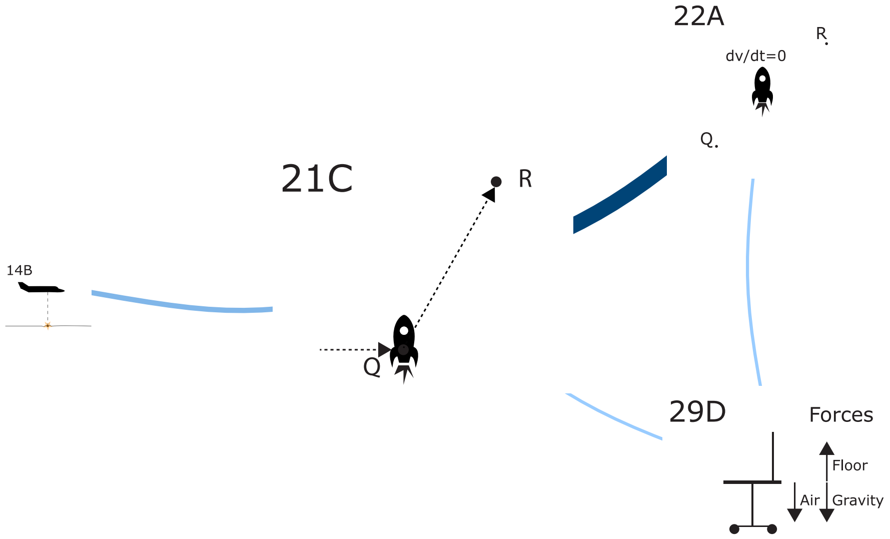

FIG. 16. Module 8.

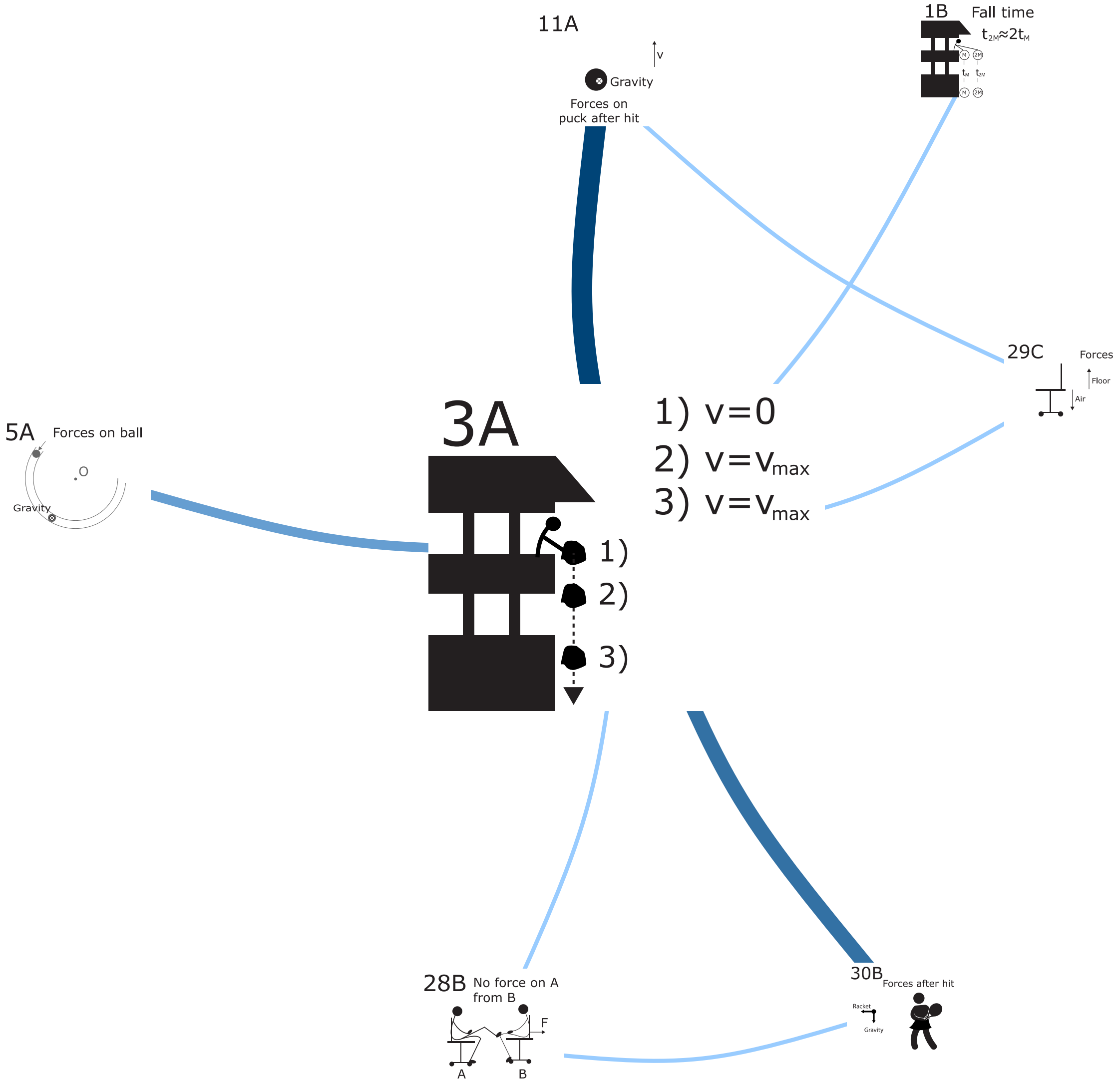

FIG. 17. Module 9.

[1] S. R. Singer, Advancing research on undergraduate science learning, J. Res. Sci. Teach. **50**, 768 (2013).

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

[3] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory, Phys. Teach. **33**, 502 (1995).

[4] L. Bradshaw and J. Templin, Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions, Psychometrika **79**, 403 (2014).

[5] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).

[6] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80**, 825 (2012).

[7] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[8] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

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

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

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

[12] D. Hestenes and J. Jackson, Revised Table II for the Force Concept Inventory (Unpublished), http://modeling.asu.edu/R&E/FCI-RevisedTable-II_2010.pdf.

[13] Huffman and Heller factor analyzed the 1992, 29 question, version of the Force Concept Inventory; however, it is improbable that their analysis would change substantially by using the 1995 version.

[14] 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).

[15] 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).

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

[17] D. Z. Grunspan, B. L. Wiggins, and S. M. Goodreau, Understanding classrooms through social network analysis: A primer for social network analysis in education research, Cell Biol. Educ. **13**, 167 (2014).

[18] M. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).

[19] Y. Hu, Efficient, high-quality force-directed graph drawing, Math. J. **10**, 37 (2005).

[20] M. Bastian, S. Heymann, M. Jacomy *et al.*, Gephi: An open source software for exploring and manipulating networks. International Conference on Web and Social Media **8**, 361 (2009).

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

[22] R. C. Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2014).

[23] N. J. Foti, J. M. Hughes, and D. N. Rockmore, Nonparametric sparsification of complex multiscale networks., PLoS One **6**, e16431 (2011).

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

[25] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).

[26] A. Lancichinetti and S. Fortunato, Community detection algorithms: A comparative analysis, Phys. Rev. E **80**, 056117 (2009).

[27] J. Bruun, Ph.D. thesis, University of Copenhagen, Copenhagen, 2012.

[28] J. Bruun and I. G. Bearden, Time development in the early history of social networks: Link stabilization, group dynamics, and segregation, PLoS One **9**, e112775 (2014).

[29] M. Rosvall, D. Axelsson, and C. T. Bergstrom, The map equation, Eur. Phys. J. Spec. Top. **178**, 13 (2009).

[30] D. A. Huffman *et al.*, A method for the construction of minimum redundancy codes, Proc. IRE **40**, 1098 (1952).

[31] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[32] D. Edler and M. Rosvall, The MapEquation software package, available online at http://www.mapequation.org.

[33] Bruun and Bearden [28] make use of the related construct Variation of Information.

[34] A. diSessa, *Phenomenology and The Evolution of Intuition* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1983), Chap. 2, pp. 15–33.

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

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