Using Topological Data Analysis in Social Science Research: Unpacking Decisions and Opportunities for a New Method

Dr. Allison Godwin, Purdue University, West Lafayette

Allison Godwin, Ph.D. is an Assistant Professor of Engineering Education at Purdue University. Her research focuses what factors influence diverse students to choose engineering and stay in engineering through their careers and how different experiences within the practice and culture of engineering foster or hinder belongingness and identity development. Dr. Godwin graduated from Clemson University with a B.S. in Chemical Engineering and Ph.D. in Engineering and Science Education. Her research earned her a National Science Foundation CAREER Award focused on characterizing latent diversity, which includes diverse attitudes, mindsets, and approaches to learning, to understand engineering students' identity development. She has won several awards for her research including the 2016 American Society of Engineering Education Educational Research and Methods Division Best Paper Award and the 2018 Benjamin J. Dasher Best Paper Award for the IEEE Frontiers in Education Conference. She has also been recognized for the synergy of research and teaching as an invited participant of the 2016 National Academy of Engineering Frontiers of Engineering Education Symposium and the Purdue University 2018 recipient of School of Engineering Education Award for Excellence in Undergraduate Teaching and the 2018 College of Engineering Exceptional Early Career Teaching Award.

Mr. Aaron Robert Hamilton Thielmeyer, Purdue University, West Lafayette

Aaron Thielmeyer is a mechanical engineering undergraduate student at Purdue University.

Ms. Jacqueline Ann Rohde, Purdue University, West Lafayette

Jacqueline A. Rohde is a graduate student at Purdue University as the recipient of an NSF Graduate Research Fellowship. Her research interests in engineering education include the development student identity and attitudes, with a specific focus on the pre-professional identities of engineering undergraduates who join non- industry occupations upon graduation.

Ms. Dina Verdín, Purdue University, West Lafayette

Dina Verdín is a Ph.D. Candidate in Engineering Education and M.S. student in Industrial Engineering at Purdue University. She completed her B.S. in Industrial and Systems Engineering at San José State University. Dina is a 2016 recipient of the National Science Foundation's Graduate Research Fellowship and an Honorable Mention for the Ford Foundation Fellowship Program. Her research interest focuses on changing the deficit base perspective of first-generation college students by providing asset-based approaches to understanding this population. Dina is interested in understanding how first-generation college students author their identities as engineers and negotiate their multiple identities in the current culture of engineering.

Ms. Brianna Shani Benedict, Purdue University, West Lafayette

Brianna Benedict is a Graduate Research Assistant in the School of Engineering Education at Purdue University. She completed her Bachelor's and Master's of Science in Industrial and Systems Engineering at North Carolina Agricultural & Technical State University. Her research interest focuses on interdisciplinary students' identity development, belongingness in engineering, and recognition.

Rachel Ann Baker

Dr. Jacqueline Doyle, Harvard-Smithsonian Center for Astrophysics

Jacqueline Doyle is a Postdoctoral Fellow at the Harvard-Smithsonian Center for Astrophysics. Her current research interests include professional development for K-12 science teachers; factors influencing student career interests; diversity, inclusion, and equity in STEM; and student identity development. She graduated from Florida International University with a Ph.D. in Physics.

Using Topological Data Analysis in Social Science Research: Unpacking Decisions and Opportunities for a New Method

Abstract

This research paper describes a new statistical method for engineering education, Topological Data Analysis (TDA), and considers the important decisions made during analysis and their impact on the quality of the results. We also describe why this new method may provide novel ways of understanding multidimensional data for student attitudes, beliefs, and mindsets.

In this paper, we discuss the considerations that researchers must understand in conducting TDA with engineering education data. In the data analysis, a researcher must choose a *filtering method*, *clustering method*, number of *filter slices* (n), *overlap* in data, and *cut height* (ε) for each dimension. The importance and effect on the consistency and quality of the data differ for each decision. Some have a large impact on the results of the analysis (e.g., cut height [ε]), while others have a moderate impact on the resulting map appearance but not key structural features identified (e.g., number of filter slices [n]).

We illustrate these methodological decisions as well as the results of TDA and its usefulness for engineering education using data from a project investigating first-year engineering students' underlying attitudes, beliefs, and mindsets to characterize the latent diversity of these students. A paper-and-pencil survey was administered to 3,855 students at 32 ABET accredited institutions across the U.S. in Fall 2017. After cleaning the data using attention checks within the survey, 3,711 student responses were examined for validity evidence. Exploratory factor analysis (for newly developed scales) and confirmatory factor analysis (for existing scales) were conducted. The resulting factors with strong validity evidence and high variability among engineering students were used in the TDA to map students' latent diversity. The results of this map indicate six distinct data progressions as well as a sparse group of students whose responses were not similar to the majority of the dataset. This work illustrates the opportunities for using TDA and provides a discussion of the different researcher decisions that are involved in this statistical technique.

Introduction

Recent quantitative research in social science and engineering education has begun to focus not just on a single aspect of participants' experiences or psycho-socio processes but rather a constellation of aspects that are important for particular outcomes like retention or academic success [1]-[4]. For example, studies of students' self-efficacy in engineering contexts provide valuable insights into how students' perceived abilities to accomplish particular tasks may influence important student outcomes; however, these studies do not fully account for other aspects of students experiences and identities including attitudes toward subject material, motivation, background experiences, social identities like race and gender, and other salient and interwoven student attitudes, beliefs, and mindsets. Accounting for multiple and overlapping measures can provide additional explanatory power to understand student outcomes, but this approach also brings methodological challenges in analyzing complex data with multiple correlated dimensions.

One newer statistical technique, Topological Data Analysis (TDA), may be one answer to some of the issues with studying complex data in social science and engineering education more specifically [5], [6]. TDA is a statistical method that can map structure within highly dimensional, noisy, and incomplete data. It is also insensitive to the particular distance function chosen to detect the persistent structure or typology in the data. In some ways, TDA is like a robust cluster analysis. However, unlike cluster analysis, which attempts to break datasets into distinct (or probabilistic) groups, TDA allows for data with progressions rather than clear distinctions. Rather than being focused on breaking data into defined groups, TDA maps the connections among data and provides additional details within the data structure that cannot be captured using cluster analysis. This approach uses the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of complex data sets. This hybrid method can identify subgroups in datasets that traditional methodologies fail to find. Since its development in 2009, TDA has been used in a number of different fields including medicine, business, and sports. However, few studies have used this technique with social science data. We believe that this technique can be particularly useful for engineering education researchers who deal with complex data that is often multidimensional, noisy, and incomplete.

Background

Quantitative studies allow broad investigations to establish general patterns of behavior and phenomena across different settings and contexts. Using random sampling techniques can allow researchers to generalize specific findings for a subset of participants back to a larger population without needing to measure the phenomena of interest for the entire population. Quantitative measures in the social science of cognitive and affective factors often provide a numeric assessment of some outcome or underlying trait to examine relationships among variables, often controlling for confounding variables. These measures must be carefully developed to minimize measurement error and accurately capture the underlying phenomena being measured for all students. The validity of measures is a significant consideration in quantitative research and has been written about extensively [7]. In this paper, we will not focus on validity aspects of quantitative measurements but rather the challenges of methodological considerations in data analysis.

Several methodological challenges commonly exist with quantitative analyses of students' cognitive and affective factors. One particular challenge, especially in a white male-dominated context of engineering, is that many statistical techniques fundamentally rely on examining the mean (or median for many non-parametric tests) and variation from that central measurement for different groups to make comparisons, investigate relationships, or test hypotheses. This fundamental norm can result in essentializing results of quantitative analyses to all members of a group. For example, a researcher might make claims about engineers being more or less likely than non-engineering peers to have some particular belief. This finding would be true of the "average" engineering student attitude, but not necessarily true of all engineers. The large majority can often overshadow the individual in quantitative results. For all engineers, this overrepresentation often results in the attitudes, beliefs, and mindsets of white men being emphasized and reported.

There are ways to analyze data to represent better students at the intersections of multiple underrepresented categories, but these approaches also have challenges. First, students at the

intersections of multiple underrepresented categories are small in number [30]. Small numbers of students can be viewed as "anomalies" not representative of the whole and dismissed. Additionally, statistical power to detect differences or understand students at multiple intersections is impossible to obtain in smaller datasets. Finally, these small numbers of students can be disaggregated from the larger dataset in ways that re-identify participants and make their responses non-anonymous, which have ethical implications [8].

Qualitative research often focuses on rich and thick descriptions of students' individual experiences that can be used as powerful examples [8]. This approach has strengths, especially in understanding the experiences of a small number of students. While this approach can deeply answer research questions, there are many potential pitfalls in collecting rich data including significant time for analysis, accurately representing students' narratives and words in ways that remain true to their experiences and the use of these studies in large-scale reform.

Both quantitative and qualitative research have the power to answer different types of research questions, but traditional approaches from both paradigms in answering research questions related to general trends in data for small groups are limited. Newer person-centered statistical techniques, like TDA, that provide ways to address these types of research questions that sit between traditional approaches to research questions, data collection, and data analysis in quantitative and qualitative research paradigms.

Using TDA to characterize students' cognitive and affective factors can address some of the statistical power issues in quantitative studies to understand students' diversity by focusing on how an individuals' response connects into the larger structure of the data rather than binning students into specific groups and comparing the means of these groups. This approach can also support qualitative research that captures the rich and detailed nuances of individual differences and aids in understanding a wide range of student' attitudes, beliefs, and mindsets. Using TDA to capture the variation of the individual along multiple affective and cognitive dimensions protects against potential re-identification within the data, moves away from dismissing small samples as an anomaly, and refrains from essentializing diverse groups of individuals. This methodological approach to understand diversity in engineering education is not a panacea for all methodological issues. Rather, this approach provides a new way of examining multiple affective and cognitive dimensions at once to understand how an individual experiences engineering.

How TDA Has Been Used in Prior Literature

TDA is an advanced statistical clustering technique that examines the topology, or the landscape, of the data to find common, dense areas in the dataset. TDA arose from a field of statistical theory concerned with "shapes" within data (i.e., topology); one of the most simplified forms of this approach is a traditional cluster analysis [9]. Carlsson and co-authors developed Python code to conduct TDA, which enables researchers outside of statistics to leverage the powerful analytic processes within TDA. This code has also been converted to a package in R [10] called TDAmapper that can be used for this analysis [11].

TDA arose from the field of statistics but has been applied in a variety of disciplines. In the natural sciences, the technique has been used in a variety of applications, including cancer detection [12], brain activity mapping [13], and gene detection [14]. These applications share common traits of

complex, high dimensional datasets and the need to understand underlying patterns within the data. Within the social sciences, TDA techniques have been less widely used, although some applications in text-based data mining have been demonstrated [15]. As an example of the wide utility of TDA, Lum and coauthors [16] demonstrated how TDA can delineate types (or profiles) of basketball players in the NBA. Instead of the traditional five positions (i.e., point guard, shooting guard, small forward, power forward, and center), they identified up to eleven different types of positions based on their performance statistics and traditional characteristics of gameplay. This example gives some insight into the utility of detected complex structures within complex and multidimensional datasets. Within education research specifically, TDA has been discussed as a means to describe patterns in student online learning [17].

TDA has seen little application in engineering education research. The first published use of TDA described its utility in uncovering latent attitude profiles of undergraduate engineering students [18]. This strategy examined student survey responses to constructs such as identity, motivation, and belonging. Although a successful proof of application, this sample was not nationally representative and thus did not capture the full range of potential attitudes and beliefs. Recent work [19] extends the line of inquiry to a nationally representative sample of engineering students and highlights the ways in which groups of student attitude profiles emerge out of a large dataset with multiple attitudinal constructs.

Although TDA has grown in popularity in multiple fields of research, there remains a lack of a knowledge base relating to the specific steps taken during analysis. Prior literature tends to focus on mathematical processes and algorithms and is not specifically designed with educational survey data in mind. Given the complexity of the technique, it is paramount that researchers are knowledgeable about the decisions necessary to produce a high-quality analysis. This paper presents the process by which a successful TDA was constructed from survey data, including the creation of filters and cut-offs to illuminate the clearest and most meaningful shapes within the dataset.

Below, we describe the considerations that researchers must understand in conducting TDA with engineering education data through a specific engineering education example. In conducting TDA, multiple researcher decisions must be made for different modeling parameters. In the data analysis, a researcher must choose a *filtering method*, number of *nearest neighbors* (k), number of *filter slices* (n), *overlap* in data, and *cut height* (ε) for each dimension. The importance and effect on the consistency and quality of the data differ for each decision. We present an example of studying students' underlying or latent diversity in attitudes beliefs and mindsets to illustrate the effect of these different researcher decisions on the resulting data progressions. We also discuss the strengths of this new approach to understand better how multiple measures can be mapped into a single representation that considers the responses of each individual rather than an average of a group.

Topological Data Analysis in Engineering Education: An Example

Study Background

This work is part of a national study where we investigate the underlying attitudes, mindsets, and beliefs, i.e., latent diversity, of first-year engineering students to improve our understanding about

what contributes to innovation in engineering solutions. We developed a survey based on existing literature regarding epistemic beliefs, innovation self-efficacy beliefs, identity, recognition, belonging, motivation, and agency. These constructs were selected to characterize latent diversity, attitudes, beliefs, and mindsets that are not readily visible by are vastly different by students' prior experiences and backgrounds. We hypothesize that this latent diversity can contribute to students' abilities to engage in novel problem solving and may help develop more innovative engineering graduates. Concurrently, we used findings from a pilot qualitative study to identify and refine questions for the instrument. In addition to the latent diversity constructs, we included background factors such as race/ethnicity, gender identity, sexual orientation, academic performance in high school, parental educational level, ZIP code, and email addresses to aid the following phase of research in which we will follow-up with students to capture stories about their pathways in and throughout engineering. We used topological data analysis to identify the variation in mindsets across the various groups. These groups were used to identify and recruit students to participate in longitudinal interviews that were designed to capture students' identity trajectory.

Data

Data for this analysis were collected in Fall 2017 semester as part of a grant titled CAREER: Actualizing Latent Diversity: Building Innovation through Engineering Students' Identity Development. The dataset is from a nationally representative sample of ABET accredited institutions with first-year engineering programs. Schools sampled were stratified by the number of engineering students enrolled by small (7,750 or less), medium (7,751 to 23,050), and large (23,051 or above) based on data from Integrated Postsecondary Database System [20]. One-third of the sample was randomly recruited from each stratified list to ensure that the sample had representation from the numerous small schools in the U.S. and avoid overrepresentation in responses from large, public engineering institutions. The surveys were administered via paper-pencil at 32 four-year, ABET-accredited institutions in students' introductory engineering courses.

The data were cleaned of indiscriminate responses using attention checks within the survey. Following, an expectation-maximization single imputation was conducted using the Amelia II package [21].

Overall Student Demographics from the Large-Scale Dataset

The overall sample of first-year engineering students collected from 32 four-year universities was 3,711. Of the overall sample of first-year engineering students that took this survey the gender breakdown is as follows: 720 (19%) female-identified students, 2150 (58%) male-identified students, 14 (0.04%) genderqueer, 17 (0.06%) agender, 70 (2%) transgender, 75(2%) a gender not listed and 782 (21%) students did not indicate any gender from the list provided to them in the survey. The race/ethnicity breakdown is 380 (10%) Asian, 209 (6%) African American/Black, 347 (9%) Latino/a or Hispanic, 65 (2%) Middle Eastern or Native African, 34 (1%) Native Hawaiian or other Pacific Islander, 49 (1%) Native American or Alaska Native, 2089 (56%) White, 72 (2%) another race/ethnicity not listed, and 793 (21%) did not indicate any race/ethnicity listed. Often students do not complete surveys due to fatigue, lack of time or loss of interest; this may have been the case for some of the students who did not report a gender or race/ethnicity. Students were allowed to select any and all gender and race/ethnicity with which they identified. For example, out of the 2,089 (56%) students who identified as White, 291 (14%) of them also identified with another race/ethnicity. Additionally, students were asked to report their home ZIP code; these ZIP

codes were plotted on the U.S. map to provide a geographic distribution of the overall first-year engineering student sample in the dataset, Figure 1.

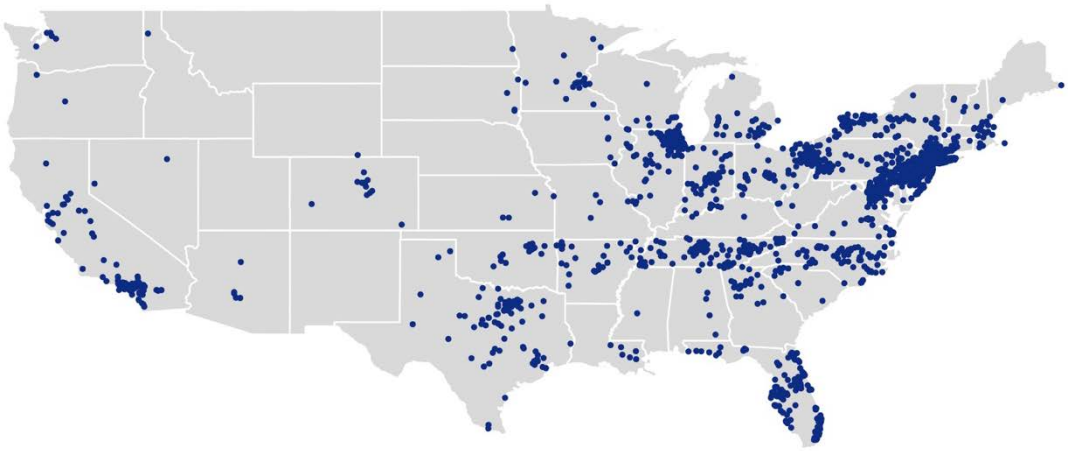

Figure 1. Home ZIP codes of all first-year engineering students who participated in the CAREER survey using ggplot2 [22].

Considerations in Conducting TDA

Pre-mapping Considerations

When conducting TDA in social science research, there are a number of important considerations before conducting the analysis. Because the mapping algorithm must be able to calculate the distance between every data point for cluster analysis, missing values are not permitted. If the data have missing values, the researcher may choose to impute missing values. Based on the current research on imputation, expectation-maximization imputation or multiple imputations with recombination to estimate the missing data are two robust choices [23]. We used expectation-maximization imputation for a single dataset for this analysis because the rules for data recombination suggested by Rubin have not been clearly defined for recombining TDA maps [24]. In addition, the measures used must be of high quality. Direct measures may have fewer issues than using latent variable measures. However, many important aspects of students' experiences in engineering education often involve latent variables. Researchers should use psychometrically valid instruments to minimize measurement error in the resulting TDA analysis.

Creating a factor space

The researcher must also create a factor space in which each data point can be placed. The factors for examination are determined from conducting factor analysis to test which items load onto particular underlying factors. The factor space is a higher dimensional space defined by an axis for each factor used in the analysis. In our example study, we created a 17-dimensional factor space using 17 factors that measure students' underlying attitudes beliefs and mindsets including measures of STEM role identities (mathematics, physics, and engineering), motivation, epistemic beliefs, belonging, and personality. A discussion of the theoretical basis for each of these dimensions and the measures used can be found in [19]. For each student in the dataset, we calculated an average score for each factor from the individual items used to measure each latent factor. These scores are used by the mapping algorithm to place the students into the factor space.

When using the Mapper function, it is important that all data points exist on the positive side of each axis; factor loadings that are negative should be made positive and the measurement scale should be inverted such that a negative factor score is not calculated for any factor for any student. For example, if a survey question has a loading of -0.78 onto a factor the item must be reverse coded to ensure that the factor loadings for each variable are positive.

Choosing Mapper parameters

After obtaining a data set with no missing values, creating a valid factor space, and computing the factor scores for each data point, the researcher must choose the model parameters for TDA. A *filtering method,* a *clustering method,* number of *filter slices (n),* amount of *overlap,* and *cut height* (ε) must be chosen in order to perform the TDA.

The filtering method is a function chosen that represents some interesting quality of each data point. The filter function can be any function that assigns to each point in the data set a single real number. For example, the filter function may be the GPA of each student, resulting in a map of students grouped by similar GPA. Typically, a filter function is usually chosen because it has a relevant geometric meaning, such as local density estimate, and these kinds of filters are known as geometric filters. These types of filter functions are typically chosen because they do not rely on representation of the data as a data matrix and therefore can account better for the underlying topology of the data within some space [16]. For our data, we used a k-nearest neighbors (knn) density estimate for the filter function, meaning the Mapper function analyzed the high dimensional factor space and assigned each point a filter value based on how close the k-th nearest neighbor was to that point.

For the filtering method chosen, the number of nearest neighbors, *k,* is selected by analyzing the distribution of sums of distances to the k-th nearest neighbor for various k-values. Different distance metrics can be used, but Euclidean distance is the simplest. In general, a higher *k* will result in more a connected map and a smaller ungrouped set of points, and a lower *k* will result in a less connected map with a larger ungrouped set of points. There is not a "best" value for *k,* and the impact of this parameter on the resulting filters should be visualized for each dataset [12], [25]. Similar to other statistical techniques, like cluster analysis, the decision of how many clusters to choose or in exploratory factor analysis, the number of factors to extract, this method is often iterative and determined through graphical means. Another method to choose *k* when little is known about the data structure is to use the formula in Eqn. 1, where *n* is the number of data points.

$$k = n^{0.5} \tag{1}$$

In our work, we examined the distribution of points in the map by differing *k* values and chose the *k* value, 59, that produced a smooth distribution of values and is an odd value to avoid ties between two class labels (points grouped into two clusters; see Figure 2 below). The distribution indicates the frequency of the distance between a point in the dataset and the nearest k neighbors or points. A discontinuous distribution would indicate that the map would fraction into subgroups because of unpopulated filter slices. Some statisticians recommend subsetting the data into a test and train sets to compute the error in the model for each *k* value. This approach works well when the inherent structure of the data is known, but may not be appropriate for some engineering education research data.

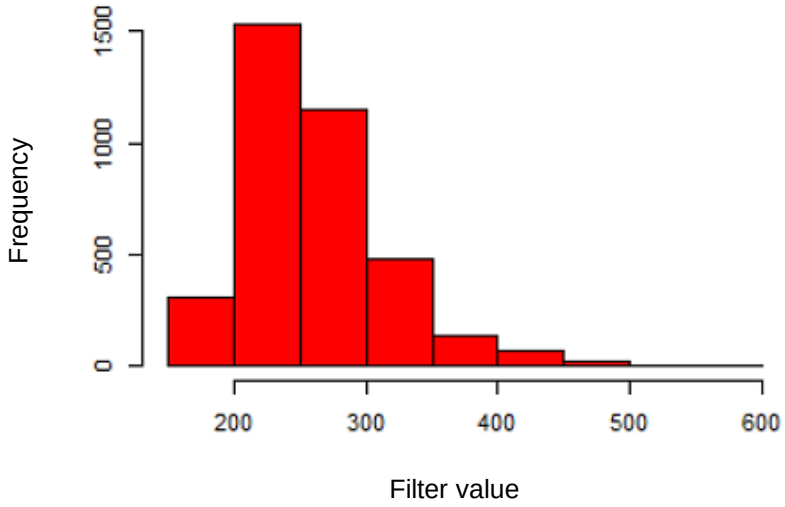

Figure 2. Distribution of *k* values for filter method using knn density with 59 k-nearest neighbors.

The result of the filter function is a set of filter values for each data point that falls within a known range. This range is then cut into *n* filter slices, that overlap based on the given *overlap* parameter. For example, if the range of the filter function is [0,100], and *n* is chosen to be 2 with an *overlap* of 50%, the filter would be sliced into two sub-ranges that overlap one another 50%, [0,66] and [33,100]. Points are grouped by which sub-ranges their filter value falls within and then clustered (see below), creating a network of nodes. If two nodes have a common point (due to the overlap in filter ranges), then they are connected in the map with an edge.

Based on our experience with engineering education data [25], the number of filter slices, *n*, should be large enough to result in enough points within each filter slice to be able to conduct future statistical analysis. Using the central limit theorem, we recommend at minimum of 30, but encourage larger slices for analyses by additional variables. We used 100 data points per filter slice as a starting point for developing our mapping, with 3,711 data points this value rounds to 35 filter slices. Due to the distribution of data within the factor space, it is likely that the final map contains a small number of nodes with more than 100 data points, and many nodes with fewer than 100 data points. A higher value for *n* results in more slices of the data which results in fewer points per slice and thus smaller nodes (often obscuring the underlying structure of the data). These smaller nodes may have less connectivity between nodes and longer sections of simple lines of nodes with few branches. The choice of *n* does change the relative the size of nodes and resolution of the resulting map but does not change the core interpretation of the map structure. This parameter has much less impact on the final mapping than other modeling parameters.

The overlap significantly affects the connectedness of the map. For the purposes of viewing the data on a continuous scale rather than grouping students into separate bins, the maximum overlap of 50% was selected. Higher overlap results in more connection between branches of the map, and a more gradual distribution between nodes and branches. A lower overlap should be used if a greater difference between nodes is desired.

The next step in TDA is to take the data that has been grouped by the filtering function into bins, in our case, k-th nearest neighbors, and apply a clustering technique to group the bins into partial

clusters (nodes within the final mapping). The clusters are determined by an agglomerative hierarchical clustering method set by the *clustering method* parameter. The Mapper function has four options: single linkage clustering, average linkage clustering, complete linkage clustering, and Ward's method. For our data, we chose single linkage clustering to identify how individual nodes within the map connect based on student membership. A single linkage connects two nodes with the minimum distance between points and can find an optimum solution [26].

Once the data are clustered, a constant cut height, ε, is chosen to determine the number of clusters (in other work [16], this parameter is referred to as k. Since we have already used the variable name, k, for the filter function, the symbol ε is used in this work and is consistent with the distance parameter described in prior work for Vietoris-Rips complexes [27]). The parameter, ε, dictates the point the data at which the branches are cut from the main stem of the hierarchical dendrogram to create similar groupings. The value of ε is a single value for cutting the branches for the entire mapping rather than a value for each connection point. In order to determine the constant cut height, the threshold values for each transition in the clustering can be generated and plotted as a histogram of all of the threshold values for the dendrograms of the different filter slices. The decision for ε comes from finding the last threshold before the first gap in the histogram, i.e., where points stop being clustered together and instead well-separated clusters start merging together. The resulting histogram indicated a small number of points with ε of zero (indicating the same overlapping points) and the value of 4.0 as the threshold before a gap in the data. Constant cut height is a very sensitive parameter. We have found that this number often corresponds to the square root of the number of dimensions in our engineering education attitudinal survey data. We suggest that ε values plus or minus 5-20% should also be tested. After testing multiple values of ε, the underlying persistent structure of the data can be seen, and a final ε can be selected.

TDA Results

The optimum parameters were determined for our data and the resulting map indicated that there were six interrelated groups of student attitudes beliefs and mindsets as shown in Figure 3. There was also a larger disconnected group that showed a number of students (n = 478) had very different attitudes from their peers and one another that is not shown in Figure 3. Our final parameters included a k-nearest neighbors *filtering method*, single-linkage hierarchical agglomerative *clustering method*, 35 *filter slices* (n), a 50% *overlap* in data, and a 4.0 *cut height* (ε). These parameters were determined using the heuristics above as well as the data itself. Below, we highlight some of the important features that emerge from the resulting TDA map.

We found a number of similar patterns in students' underlying attitudes, beliefs, and mindsets at the beginning of their engineering careers as measured and characterized in this study. Most students in the study (n = 3,233) had similar enough attitudes to be connected with the resulting mapping. This finding is different from previous work using TDA with four U.S. institutions [25]. We hypothesize that the larger representation of different institution types, sizes, and geographic locations may be one reason in identifying a broader common set of important attitudes, beliefs, and mindsets for students entering engineering. This study also investigated a wider set of student measures beyond students' affective attitudes and beliefs. The group that did not connect in with the resulting map had very different attitudes than the majority of students in the sample, and very different attitudes from one another. Together, these results indicate that students entering engineering are not a homogeneous group; various student attitudes exist that distinguish students

from one another and from the normed expectations in engineering. If a "one-size-fits-all" approach to engineering education is taken, it might alienate particular students or communicate through explicit or implicit messages that students do not belong in engineering [28], [29].

From TDA, the resulting data progressions allow us to see not only group membership but also interconnections between highly dimensional student attitudinal data. These progressions can be used to understand the landscape of students' attitudes, beliefs, and mindsets rather than averages of a group. In Figure 3, we have marked groups indicated by branching within the resulting map, but these groups are identified by the researchers rather than by the TDA algorithm. These groups have varying similarities across the 17 dimensions used to develop this mapping. Our future work will involve describing these characteristics and how students' profiles (general connections within the identified groups) affect their pathways into and through engineering. Ultimately, we desire to understand how students begin to see themselves as engineers and how students with diverse underlying attitudes, beliefs, and mindsets may be better supported in engineering education by acknowledging individual differences of students.

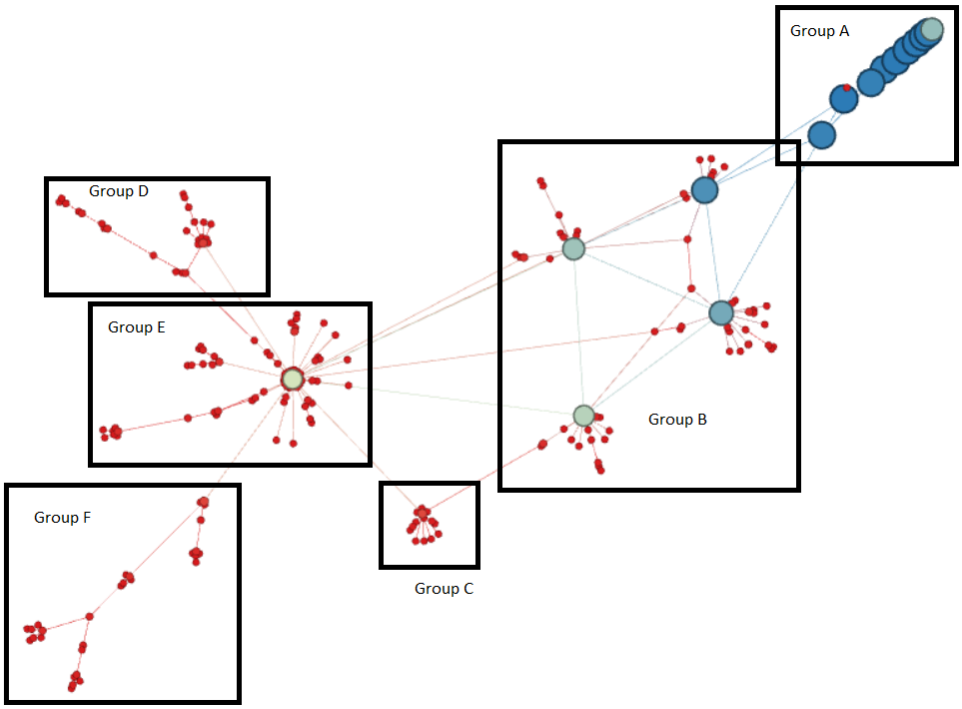

Figure 3. Resulting mapping and researcher-identified groupings from TDA. The colors in the image indicate the density of the map (number of students present in each node). Blue nodes indicate a high population size (~200 students) and red nodes indicate a smaller population size (~3 students).

Implications and Limitations of this Approach

TDA has been used extensively in quantitative research, including in fields like business, sports, and medicine. There are not many examples of studies where TDA has been used to analyze social sciences, particularly in the field of engineering education. While this is somewhat unchartered territory, there are exciting possibilities for using methods such as TDA to analyze complex and

multidimensional data. Primarily, the results of TDA provide complex and rich data progressions. The resulting map, shown above in Figure 3, is descriptive rather than inferential in group determination and differences between groups. What we mean by that statement is that the results of the TDA model the structure of the data itself and emerges from the data. This approach is in contrast to other statistical methods that rely on specifying a probability at which a group is considered different (e.g., $p < 0.05$) or forcing data into deterministic groups. This approach allows for more nuanced relationships and patterns to be identified between groups and individuals.

TDA does have some inherent benefits that can help to ensure that each individual response collected is still represented, regardless of minority status. An individual's response pattern is preserved in TDA, which means that underrepresented students' contributions, both demographically and attitudinally, are holistically represented within the dataset rather than being "washed out" through comparisons of group averages or ignoring individuals as "anomalies."

As with any new statistical method, the application of TDA to the field of engineering education has some limitations and important caveats. While the usage of TDA allowed for the analysis of complex and multidimensional data, there are still drawbacks to using this approach. For example, this approach is limited in determining groups. Unlike cluster analysis, TDA does not result in deterministic or probabilistic groupings with student responses bounded to a particular cluster or grouping. The results provide a mapping of relationships within the data and the relative positioning of students. Identifying similarities and differences within the data structure can be more difficult than with other methods; this task often falls to the researcher to interpret the data and how groups should be formed from the relative positions in the map, which can result in errors or biased decisions. TDA is not a replacement for cluster analysis and its use should be consistent with the research questions posed in the study.

Another limitation is part of the data analysis itself. Datasets with missingness cannot be used in the data analysis, so imputation methods that account for missingness must be used. All of the factors used in the data analysis should be measured on the same scale or normalized before the analysis. This process is necessary because the data relies on distance parameters for analysis. In measuring latent factors, this approach does not account measurement error, like many methods outside the structural equation modeling family to data analysis techniques. All measurements should minimize measurement error because that error will be incorporated into the model and cannot be estimated separately.

Researchers also need to be clear about their role, input, and decision-making processes if they choose to utilize the TDA method in a study. TDA is emergent from the data and avoids some *a priori* assumptions about the data. However, this approach is not free of important decisions made by researchers in the making and handling of data. The process of conducting TDA involves several researcher decisions about modeling that must be empirically tested and reported. A significant amount of researcher discretion goes into the decision-making process. TDA is less sensitive to small changes in the number of filter slices, but there is a point at which large changes in this parameter would reduce the resolution of the map (too few filter slices) or result in a cloud of points with few connections or groupings (too many filter slices). TDA is especially sensitive to the choice of a constant cut value in how filter slices are grouped for the resolution of the resulting map. This analysis is not unlike other modeling approaches where the researcher has a

significant impact on the results including exploratory factor analysis and complex regression modeling. However, the norms of using those statistical techniques and the acceptable reporting statistics are much better established. The quality of future work in TDA will rely on the larger research community to provide explicit descriptions of the decisions made in these kinds of complex modeling methods.

Future work with this data will consist of investigating the different groups on the map, shown in Figure 3. This will include recruiting students from different parts of the maps in order to conduct longitudinal interviews about engineering pathways and the negotiation of identities as engineers. This corresponding qualitative work will build upon this study's existing quantitative results and will inform additional studies with the insights recorded.

Conclusion

This paper described the key model parameters that researchers must consider in using a new statistical method, Topological Data Analysis (TDA). We also presented how TDA can be useful to characterize students' latent diversity from a survey study of 3,711 first-year engineering students' incoming attitudes, beliefs, and mindsets at 32 ABET-accredited institutions across the United States. This approach identified patterns of individual student responses as well as general trends in the data. The novelty of this technique is the ability to detect the underlying structure in noisy and multidimensional data using a person-centered approach. This approach allowed the preservation of subjects' individual response patterns in the data and may address some methodological issues of statistical techniques that rely on group averages.

This study identified many similarities and differences in the experiences, mindsets, and beliefs of the engineering students included in the research. Additionally, the TDA approach also revealed that engineering students are by no means homogenous and have a wide variety of perspectives. These results indicate that the application of TDA and other statistical analysis methods to the field of engineering education can yield results that detect robust structures in complex, multidimensional data that shed light on often overlooked perspectives. This statistical method provides new opportunities to characterize engineering education data. The findings from this example study can help to identify who engineering students are, how they learn, and what problems need to be addressed within the engineering classroom and community.

Acknowledgments

This work was supported through funding by the National Science Foundation CAREER Grant No. 1554057. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors wish to thank the STRIDE team and survey participants for their engagement with this study.

References

[1] M. Credé and N. R. Kuncel, "Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance," *Perspectives on Psychological Science*, vol. 3, no. 6, pp. 425-453, 2008.

[2] A. Godwin, "Unpacking Latent Diversity," in *American Society for Engineering Education (ASEE) Annual Conference and Exposition*, Columbus, OH, 2017.

[3] J. J. Lin, P. K. Imbrie, K. J. Reid, and J. Wang, "Work in progress—Modeling academic success of female and minority engineering students using the student attitudinal success instrument and pre-college factors," in *Frontiers in Education Conference (FIE)*, 2011.

[4] Authors, 2018

[5] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255-308, 2009.

[6] C. Epstein, G. Carlsson, and H. Edelsbrunner, "Topological data analysis," *Inverse Problems*, vol. 27, no. 12, pp. 120201, 2011

[7] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Washington DC, 2014.

[8] A. L. Pawley and A. E. Slaton, "The Power and Politics of STEM Research Design: Saving the 'Small N,'" in *American Society for Engineering Education (ASEE) Annual Conference and Exposition*, 2015.

[9] L. Wasserman, "Topological Data Analysis," *Annual Review of Statistics and Its Application*, vol. 5, no. 1, pp. 501-352, 2018.

[10] R Core Team, "R: A language and environment for statistical computing." *R Foundation for Statistical Computing*, Vienna, Austria, 2017.

[11] P. Pearson, D. Muellner, and G. Singh, "TDAmapper: Analyze High-Dimensional Data Using Discrete Morse Theory," R package version 1.0, 2015.

[12] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 7265-7270, 2011.

[13] M. Saggar, O. Sporns, J. Gonzalez-Castillo, P. A. Bandettini, G. Carlsson, G. Glover, and A. L. Reiss, "Towards a new approach to reveal dynamical organization of the brain using topological data analysis," *Nature Communications*, vol. 9, no. 1399, 2018.

[14] M. L. Dequeant et al., "Comparison of pattern detection methods in microarray time series of the segmentation clock," *PLoS One*, vol. 3, no. 8, pp. e2856, 2011.

[15] S. Gholizadeh, A. Seyeditabari, and W. Zadrozny, "Topological Signature of 19th Century Novelists: Persistent Homology in Text Mining," *Big Data and Cognitive Computing*, vol. 2, no. 4, pp. 33, 2018.

[16] P. Y. Lum et al., "Extracting insights from the shape of complex data using topology," *Scientific Reports*, vol. 3, no. 1, 2013.

[17] C. M. Ganley and S. A. Hart, "Shape of Educational Data: Interdisciplinary Perspectives," *Journal of Learning Analytics*, vol. 4, no. 2, pp. 6-11, 2017.

[18] Kirn, A., Godwin, A., Benson, L., Potvin, G., Doyle, J., Boone, H., & Verdin, D. (2016). "Intersectionality of Non-normative Identities in the Cultures of Engineering," in *American Society for Engineering Education (ASEE) Annual Conference and Exposition*, New Orleans, LA, 2016.

[19] A. Godwin, D. Verdín, B. S. Benedict, R. A. Baker, T. J. Milton, and J. T. Yeggy, "Board 51: CAREER: Actualizing Latent Diversity: Building Innovation through Engineering Students' Identity Development," in American Society for Engineering Education (ASEE) Annual Conference and Exposition, Salt Lake City, UT, 2018.
[20] National Center for Education Statistics, *Integrated Postsecondary Education Data System [Data file and code book]*, 2017. Retrieved from https://nces.ed.gov/ipeds/
[21] J. Honaker, G. King, and M. Blackwell, "Amelia II: A Program for Missing Data," *Journal of Statistical Software*, vol. 45, no. 7, 2011
[22] H. Wickham, *ggplots2: Elegant Graphics for Data Analysis*, New York: Springer, 2009
[23] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, Hoboken, NJ: John Wiley & Sons, 2014.
[24] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Hoboken, NJ: John Wiley and Sons, 1987
[25] J. Doyle. "Describing and Mapping the Interactions between Student Affective Factors Related to Persistence in Science, Physics, and Engineering," Ph.D. Dissertation, Dept. of Phys., Florida International University, Miami, FL, 2017. (Order No. 10747700). Available from ProQuest Dissertations & Theses Global. (2092654188).
[26] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, pp. 30-34, 1973.
[27] V. De Silva and R. Ghrist, "Coverage in sensor networks via persistent homology," *Algebraic & Geometric Topology*, vol. 7, no. 1, pp. 339-358, 2007.
[28] B. A. Danielak, A. Gupta, and A. Elby, "Marginalized identities of sense-makers: Reframing engineering student retention," *Journal of Engineering Education*, vol. 103, no. 1, pp. 8-44, 2014.
[29] I. Villanueva et al., "What Does Hidden Curriculum in Engineering Look Like and How Can It Be Explored?" in *American Society of Engineering Education (ASEE) Annual Conference and Exposition*, Salt Lake City, Utah, 2018.
[30] B. L. Yoder, "Engineering by the Numbers," Washington DC, 2017.