Developing Network Modeling and Analysis Methods for First-Year Mathematics and STEM Student Course-Taking Sequences

Jennifer Nimtz
Western Washington U.

Kimihiro Noguchi
Western Washington U.

Jothi Ramesh
Western Washington U.

Student success is typically measured by grades in coursework and 6-year graduation rates (York, Gibson, & Rankin, 2015). The purpose of this exploratory data analysis is to develop statistical methods that will provide a deeper understanding of student success, specifically success in first-year mathematics (FYM) prerequisite courses as well as the STEM courses they serve. We consider the many possible student course-taking sequences (paths) as a complex network system. We model the paths that students take with a network of vertices (courses) and edges (links between courses) by using 10 years of registrar data to generate vertex and adjacency matrices, and associated graphical representations (e.g., heat maps and network graphs). We also will conduct relevant exploratory data analysis to determine the importance of certain mathematics courses and transitions to overall student success.

Keywords: student success, mathematics prerequisites, modeling, statistics

Introduction

Student success is typically measured by grades in coursework and 6-year graduation rates (York et al., 2015). Regarding the latter, we know that student placement and success in mathematics courses are some of the primary factors that mitigate student time to degree (Aiken, Bin, Hjorth-Jensen, & Caballero, 2020). In addition, students who have STEM interests and place into lower-level mathematics courses take longer to complete their degrees. Furthermore, they are more likely to change their majors or drop out completely (Klingbeil, 2013).

The long-term purpose of this research is to gain a deeper understanding of student success, specifically success in first-year mathematics (FYM) prerequisites as well as the STEM course they serve. To accomplish this, we consider the many possible student course-taking sequences as a complex network system. We model student course-taking sequences as paths through a network graph of vertices (courses) and edges (links between courses taken in consecutive academic terms). Next, we generate a vertex matrix (courses) and adjacency matrix (edges that link courses) from registrar data of student course taking and generate associated graphical representations (e.g., heat maps and network graphs). Lastly, we will utilize statistical analysis of networks to determine the importance of certain mathematics courses to overall student success.

In this paper, we first provide a brief background of network analysis followed by the prior use of network representations to analyze the complex system of undergraduate coursework. Second, we outline our long-term research goals and contrast our research with the prior research about using networks to analyze undergraduate coursework. Third, we present our current research results in relation to our long-term research goals. Lastly, we discuss our future research agenda in more detail.

Background

In this background section, we define what a network is, provide a brief overview of the history and uses of network analysis, and culminate with the use of networks to model undergraduate courses. Networks, represented as graphs of vertices and edges, have emerged as useful tools in the modeling and analysis of complex systems across a diverse range of applications. Vertices represent the objects of study and are illustrated by the nodes on the graph.

Edges are the links between two vertices and are illustrated by the lines or arrows connecting two vertices. Edges represent the connection or flow between the objects of study.

Networks are useful tools in a wide range of research. Kolaczyk and Csardi (2014) discussed the history and expanse of network analysis, summarized in this paragraph. Due to advances in computer science in the 1950's, the use of network analysis became prevalent in areas such as transportation, allocation of resources, and distribution of products. In this initial phase, a few sociologists also utilized networks to characterize interactions in social groups. Then, interest in network modeling and analysis increased significantly in the 1990's due to the expansion of the internet and associated social media platforms. In addition, network modeling and analysis has proved useful in research areas such as computational biology (Mason & Verwoerd, 2008), engineering (Chen, 1997), finance (Abrams, Celaya-Alcala, Baldwin, Gonda, & Chen, 2016), marketing (Webster & Morrison, 2004), neuroscience (Farahani, Karwowski, & Lighthall, 2019), political science and public health (Goodin, Moran, & Rein, 2008). Two factors have contributed to the proliferation of network modeling and analysis: the increased interest in the study of complex systems, and the ability to collect, store, and maintain large data sets. These two factors prompted the development of statistical methods to analyze large, complex, connected data sets.

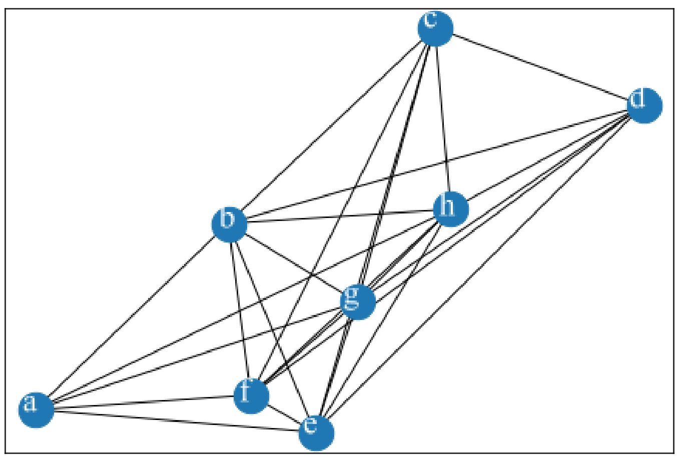

Figure 1. Network graph, G=(V,E) with vertices, V={a,b,c,d,e,f,g,h} and edges, E, that link any two vertices.

Recently, network analysis has been utilized as a method to examine the complex system of undergraduate coursework. For example, Aldrich (2015) and Ren et al. (2021) represented the intended sequences of courses using a directed network of vertices and edges in which each vertex represents a course and each edge represents a directional link between courses. To generate the course networks, university course catalogues (Aldrich, 2015) and program guides (Ren et al., 2021) can be used. In the context of undergraduate coursework network, the number of vertices represented the number of courses, and the number of edges represented the number of links between pairs of courses. They also discussed the general topology of the network by outlining the number of vertices and edges.

Long-Term Research Goals

Here, we present our long-term research goals to provide a context for our current research and results. In our research, we will utilize similar statistical methods as Aldrich (2015) and Ren et al. (2021), but our generation of the coursework network differs significantly. We define the *undergraduate course networks* presented in the referenced curriculum research as *intended* *course sequences* because those networks were generated by web-scraping data from a university course catalogue (Aldrich, 2015) and from a dental school program guide (Ren et al., 2021). In contrast, we define the *undergraduate course-taking network* we generate as the *enacted course sequences* because the network is produced from student course-taking data from a university registrar.

We currently have written a Python program to analyze registrar data of FYM course-taking, and we will expand this program to include STEM courses and General University Requirement Quantitative Reasoning courses. In our current research, we have generated a vertex matrix and adjacency matrix. Our long-term goal is to use such a vertex matrix and adjacency matrix to generate the *undergraduate course-taking network,* which will be a directed weighted network graph. In this *undergraduate course-taking network*, each vertex will represent a course and each edge will represent a directed link between a pair of courses. Each vertex will be weighted by the number of students taking that course and will contain a pie graph of the grade distribution for each course. Each edge will be weighted by the number of students who take linked pairs of courses in a sequence of terms. After generating the *undergraduate course-taking network*, we will conduct statistical analysis to determine different centrality measures of a network (e.g., degree centrality, out-degree centrality, betweenness centrality), which is discussed further in the Future Research section.

Research

To produce the *undergraduate course-taking network*, we utilized the Python programming language to generate a vertex matrix of the number of students who took each course (see Table 1) from student FYM course-taking data from a university registrar. Note that these courses are not necessarily taken in numeric order (see Figure 2). In addition, we utilized the Python programming language to generate an adjacency matrix in which each cell represents the number of students who took the sequence of the pair of courses referenced in the row (course going from) and column (course going to) (see Figures 3 & 4).

Research Questions

1. How can we represent the total number of students who take each course in a useful and flexible representation?
2. How can we represent the sequence of student course-taking in a useful and flexible representation?

Analysis

The analysis of the network of mathematics course-taking sequences is presented as follows:

1. Student course-taking is represented by a vertex matrix in which the total number of student enrollments for each course is tallied. Additional student summary data may be included in this matrix (e.g., grade distributions).
2. We generated an adjacency matrix from 10 years of registrar data. In the adjacency matrix, each cell consists of the number of students who took the sequence of the pair of courses referenced in the row (course going from) and column (course going to).

This adjacency matrix can be used in combination with the associated vertex matrix to generate a directed weighted network graph of student course-taking sequences.

Results

Our results thus far are represented in the Vertex Matrix (Table 1), Course Adjacency Matrix with Heat Map (Figure 3), and the Log Transform Adjacency Matrix with Heat Map (Figure 4). The log transform here is given by $\log_{10}(n+1)$, where n is the original number. The vertex matrix consists of a sample of FYM courses of interest and the student enrollment in those courses over a 10-year period. Note that course numbers do not necessarily indicate the prerequisite sequence. For example, a student who passes Math-90 can proceed to take Math-110 (Figure 2).

Table 1: Vertex Matrix.

Course Number	Course Name	Enrollment
090	Introductory Algebra	679
100	Quantitative Reasoning	3203
110	Functions and Algebraic Methods	5660
115	Precalculus 1	5750
116	Precalculus 2	2096
120	Calculus 1	3473
140	Business Precalculus	2334
141	Business Calculus	3270
200	Statistics	2008

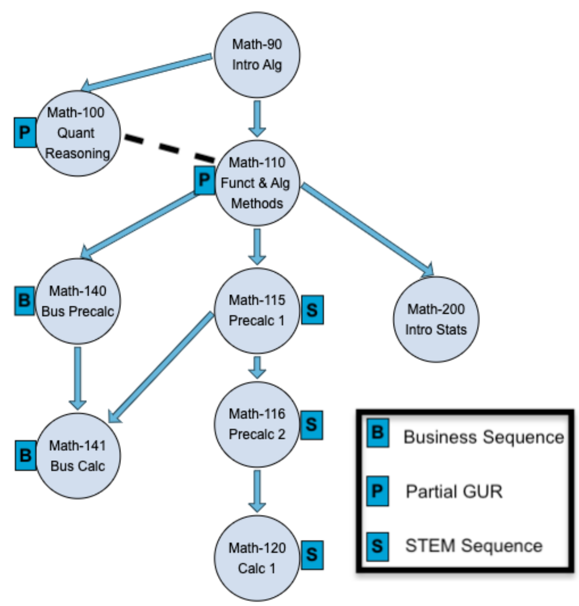

Figure 2: Intended Course Sequence.

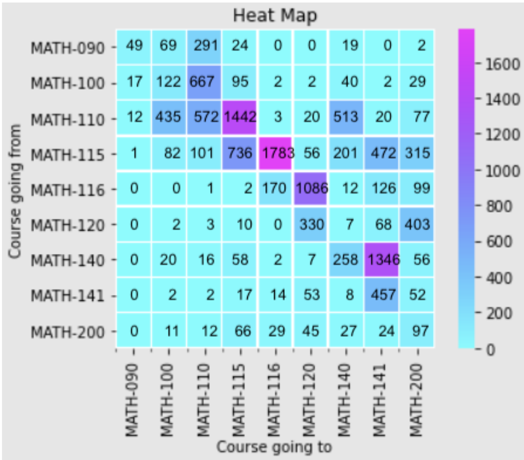

Figure 3: Course Adjacency Matrix with Heat Map

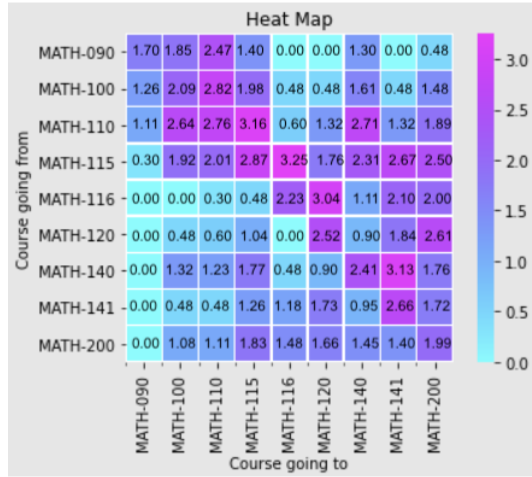

Figure 4: Log Transform of Adjacency Matrix with Heat Map

In the adjacency matrix (Figure 3), each cell consists of the number of students who took the sequence of the pair of courses referenced in the row (course going from) and column (course going to). For example, from prior data, we know that Math-110 has a higher DFW rate, so we may want to examine common course-taking sequences of those students. If we examine the row of cells indicating pairs of courses beginning with Math-110, we can determine the highest enrollment sequence pairs are [Math-110, Math-115] followed by [Math-110, Math-110]. In cell [Math-110, Math-115], we can see that 1442 students followed the STEM sequence, taking Math-110 as a prerequisite to Math-115. In cell [Math-110, Math-110], we can see that 571 students repeated Math-110, resulting in a loop in the network. Due to space limitations, we leave further interpretation to the reader and for the conference presentation.

Discussion

Our study builds on prior research that analyzed *undergraduate course networks* of *intended course sequences* generated from university course catalogues and programs of study. In contrast, we are developing methods to analyze the *undergraduate course-taking network* of *enacted course sequences* because the matrices that will be used to generate the network are produced from student course-taking data from a university registrar. The methods for generating and analyzing a directed weighted *undergraduate course-taking network* of *enacted course sequences* from actual student course-taking are the major finding of this research. Thus far, this study illustrates FYM course-taking as a vertex (course) matrix containing the number of students enrolled in the course and is useful in identifying courses with the highest enrollment. In addition, this study illustrates FYM course-taking as an adjacency matrix (transitions between pairs of courses). The adjacency matrix is useful in determining the most common course pairs taken in a sequence, which can be used to determine the emphasis of course content. The vertex matrix and adjacency matrix are generated as a first step in creating the *undergraduate course-taking network.*

Future Research

In our future research, we will conduct a network analysis on the *undergraduate course-taking network.* We will employ statistical techniques utilized by Ren et al. (2021) and Aldrich (2015) to produce the following measures of centrality: degree centrality, out-degree

degree centrality, and betweenness centrality. In the context of an undergraduate course network, these three measures of centrality signify the following: (i) degree centrality indicates the importance of a course in relation to all courses to which it is connected, (ii) out-degree centrality indicates the importance of courses in relation to those courses that follow, and (iii) betweenness centrality indicates the importance of the course in relation to the flow of information between other pairs of courses. For all three centrality measures, degree centrality, out-degree centrality, and betweenness centrality, the higher the measure, the more important the course is to an undergraduate coursework network. The degree centrality of vertices is calculated by the number of both incoming and outgoing edges to each vertex (Ren et al., 2021). On the other hand, the out-degree centrality of vertices is computed by the number of outgoing edges from each vertex (Aldrich, 2015). Lastly, the betweenness centrality is given by the extent that a vertex is situated between other pairs of vertices (Aldrich, 2015; Ren et al., 2021).

This ongoing network analysis will allow us to determine important courses in course sequences, to analyze student success in courses, and to prioritize courses in need of reform. These research techniques have the potential to apply across different departments, programs, and colleges of the university, as well as the university as a whole.

References

Abrams, J. R., Celaya-Alcala, J., Baldwin, D., Gonda, R., & Chen, Z. (2016). Analysis of equity markets: A graph theory approach. *Society for Industrial and Applied Mathematics, 10*. doi:10

Aiken, J. M., Bin, R. D., Hjorth-Jensen, M., & Caballero, M. D. (2020). Predicting time to graduation at a large enrollment American university. *PLoS ONE, 15*(11). Retrieved from https://doi.org/10.1371/journal.pone.0242334

Aldrich, P. R. (2015). The Curriculum Prerequisite Network: Modeling the Curriculum as a Complex System. *Biochemistry and Molecular Biology Education*, 168-180.

Chen, W.-K. (1997). *Graph Theory and Its Engineering Applications* (Vol. 5). Singapore: World Scientific.

Farahani, F. V., Karwowski, W., & Lighthall, N. R. (2019). Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Frontiers in Neuroscience, 13*, 1-27. doi:10.3389/fnins.2019.00585

Goodin, R., Moran, M., & Rein, M. (2008). Policy network analysis. In R. A. W. Rhodes (Ed.), *The Oxford Handbook of Public Policy*: Oxford University Press.

Hagberg, A., Schult, D., Swart, P. (2022) *NetworkX Reference: Release 2.8.5*. Downloaded from https://networkx.org/documentation/stable/_downloads/networkx_reference.pdf August 01, 2022.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Klingbeil, N. W. (2013). *A national model for engineering mathematics education: Longitudinal impact at Wright State University*. Paper presented at the 120th ASEE Annual Conference & Exposition, Atlanta, GA.

Kolaczyk, E. D., & Csardi, B. (2014). *Statistical Analysis of Network Data with R*. New York: Springer.

McKinney, W., & others. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Mason, O., & Verwoerd, M. (2008). Graph theory and networks in biology. *IET Systems Biology, 1*(2), 89-119. Retrieved from https://arxiv.org/pdf/q-bio/0604006.pdf

BLINDED REFERENCE

Python-Software-Foundation. (2022). Python Language Reference, Version 2.7. Retrieved from http://www.python.org

R-Core-Team. (2022). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

Ren, Q., Pang, X., Liu, X., Zheng, Q., He, T., Hu, D., & Jiang, X. (2021). Network Modelling and Visualization Analysis of the Undergraduate Dental Curriculum System in China. *Journal of Computer and Communications, 9*.

Webster, C. M., & Morrison, P. D. (2004). Network analysis in marketing. *Australian Marketing Journal, 12*(2), 8-18. doi:10.1016/S1441-3582(04)70094-4

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation, 20*(5), 20.