2022

Using Network Analysis to Contrast Three Models of Student Forum Discussions

Hannah N. Benston
Wright State University

USING NETWORK ANALYSIS TO CONTRAST THREE MODELS OF STUDENT FORUM DISCUSSIONS

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science

By

HANNAH N. BENSTON
B.S., Wright State University, 2020
B.S., Bowling Green State University, 2015

2022
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

February 25, 2022

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Hannah N. Benston ENTITLED Using Network Analysis to Contrast Three Models of Student Forum Discussions BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Adrienne Traxler, Ph. D.
Thesis Director

Jason Deibel, Ph.D.
Chair, Department of Physics

Committee on
Final Examination

Adrienne Traxler, Ph. D.

Jason Deibel, Ph.D.

Ivan Medvedev, Ph.D.

Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

ABSTRACT

Benston, Hannah N.. M.S. Department of Physics, Wright State University, 2022. Using Network Analysis to Contrast Three Models of Student Forum Discussions.

There is much research about how actors and events in social networks affect each other. In this research, three network models were created for discussion forums in three semesters of undergraduate general physics courses. This study seeks to understand what social network measures are most telling of a online forum classroom dynamic. That is, I wanted to understand more about things like what students are most central to the networks and whether this is consistent across different network models. I also wanted to better understand how students may or may not group together. What relationships (student to student, student to instructor, etc.) are formed, centralization, various clustering and correlation coefficients, and how participation in a forum unfolds were all things that were examined in this data set. Network model construction and measuring how these constructions may affect student interactions was another focus of this study. These attributes are analyzed among individual semesters, but also compared/contrasted across all three, to see if they maintain across different network models. It was found that in general as models increase in connectivity, a rise in network measures like centralization and average degree was observed. A drop in network measure values such as average vertex-vertex distance and diameter was also seen. Finally, it was discovered that changing a model from undirected to directed made an appreciable change in average degree outcomes. Overall, this research gave an appreciation of different network model construction and how different network measures may help describe social networks. It was discovered that centralization metrics may be more telling of social networks than what was anticipated. Average degree, average vertex to vertex distance and diameter followed trends we would expect to see. Other measures looked into were transitivity, average Barrat coefficient and degree correlation

coefficient.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Network analysis is a constantly evolving method of analysis that helps better understand the dynamics of how many types of systems work. Furthermore, social network analysis specifically, has been able to give us a closer view of how groups of people interact and form. This has become increasingly interesting in the form of online discussion forums with the rise of technology as a medium for communication for work, school, activity groups and all kinds of social formations. Unfortunately, there are currently a lot of gaps in current research, which lends more of a need for research in the area. Some of these gaps include model construction and variation of models used.

In this research, an online discussion forum which was used complementary to an introductory general physics course was looked at. By working with the data gathered across three separate, sixteen week semesters, it was hoped that a better understanding how students interact with each other in the classroom could be gained, as well as what this might say about the importance of the interactions students have with each other. The primary research question here is: what network measures are most descriptive of an online discussion forum by looking at trends between models that were constructed differently from one another. To do this, seven network measures are looked at and the calculated values are examined for any indications that trends may exist.

Three different models were created from the transcripts of the introductory physics courses. Model 1 is built as having every student in the same thread having a connection to each other. Model 2 only defines connections between the student who began a thread and the student who posts in the

thread. Finally, model 3 is built so that students have a connection to anyone who has posted above them in a particular thread. These basic model structures were designed to show a highly connected network, a loosely connected network and then have one model serve as an in-between step for connectivity level. Model 2 is the type of model that is most commonly used in current literature on these types of studies. Model 1 is designed to be a polar opposite of the model 2 design. Model 3 is our median option between the two. After the creation of each of these models, we analyzed seven network measures: average degree, centralization, average Barrat clustering coefficient, diameter, transitivity, average vertex-vertex distance and correlation coefficient. We looked at how these compared and contrasted across three semesters and within the three models. We also began to look at which students might be most central across the models.

1.1 Literature Review

A literature review was conducted to create a better understanding of current research on network analysis. One reason this was performed, was to better understand the current successes that have been made in network analyses thus far such as beginning to research more on centralization, metrics of degree and other characteristics that may tell us more about networks. On the other hand, the other benefit of reviewing existing studies, is to understand how it can be improved upon and therefore gain footing on where to begin on expanding research. The following goes into some of the papers which were reviewed in preparation for research. Before delving in, one should know the meaning of the terms ”node” and ”edge”. A node, also known as a vertex, is the subject in a network which is connected by an edge. An edge can have direction, or simply exist as a connection between nodes.

1.1.1 Online Forum Studies

Some research centers around class discussion forums that are completely carried out virtually. Online forum studies generally allow nodes in a network (such as students in a classroom) to interact via a threaded discussion format in which an original post is created and other participants can respond. These discussion forums can give ideas about different characteristics that describe networks and allow for more in depth study as to what these characteristics say about the network. In our study, one of the main characteristics we are exploring is centrality. We are looking to answer

questions about what centrality of a student in an online classroom discussion forum says about that student's performance in the class and overall experience.

Aviv et al. [1] set out to compare structured and unstructured online discussion forums using Social Network Analysis. There were structured groups which were given a schedule and certain goals to hit throughout three months. There was also an unstructured group that was there for the student usage as desired. Different factors were analyzed such as role, power, content and cohesion for their possible effects on the network. Aviv based their study off a five stage knowledge construction process which is as follows: Step 1: Sharing/ comparing knowledge; Step 2: Discover/Explore Disagreements; Step 3: Synthesis via negotiating meaning; Step 4: Testing/modifying proposed synthesis vs. schemas, theory, facts, beliefs; and step 5: Proofs of reaching agreements or meta-cognitive admitting change of knowledge. The effects in the structured asynchronus learning network were considered more robust and yielded a higher success rate for participating students, thus the null hypothesis of the paper was rejected by the findings. By use of their analysis of power, it was found that in their study, power equated to centrality.

The paper ended with a discussion of further research that could be done, which made opportunity for our own research. One point brought up was effective construction of networks. This was a point that helped lead to our creation of three different network models. It seemed as if there has been a research gap when it comes to exploring how the creation of networks might effect resulting descriptions of the networks. Furthermore, position analysis was mentioned in the paper. This led to our having a stronger curiosity in centralization and what it might say about different models of networks that we create.

In Traxler et al. [2], Traxler explores how online discussion forums build community within the classroom, and what effects this may have on students in a cohort. Data is obtained from a discussion forum which was used in accompaniment to an in class learning environment. The online discussion forum is used to better understand the dynamic of how connections may be formed within a classroom. There are three specific research questions that drive the study done in the paper: 1: How online discussion forum networks differ in multiple semesters of an introductory physics course and if the information in these networks can be taken with more simplicity from participation statistics; 2: If a student's final grade correlates with how central they are to the network; 3: If these correlations exist, do they change when backbone extraction is performed on the data.

Network analysis is done on the discussion forum data, specifically a bipartite network model is used, in which actors (students) and events (discussion threads) are both considered. An actor projection was performed on the bipartite network to provide a student - student network. Backbone extraction was used in order to simplify the dense network, but strangely did not strengthen the data set at all. PageRank and Target Entropy both proved positively correlated with grades in the first and third semesters, but the second semester showed no correlations. These findings indicate that PageRank and Target Entropy could be attributes of interest, and are worth studying more, which is why they were chosen as characteristics to analyze in our own study.

The study discussed in Poquet et al. [3] analyzes Learning Analytics using a null models approach in which graphs were randomly simulated to observe a network. Within this, they used two methods to express networks: post tree network and student to thread networks. The group created random models for the purposes of their research. Their study suggested that degree and frequency might not reflect social dynamics of the class analyzed, and also suggested that weighted clustering might be more indicative of this instead. In their study, they noticed that the number of replies to each post was relatively similar across courses. Another topic the researchers addressed had more to do with global network structures of network projections. Specifically, they wanted to know how much of student network global structures is explained by the posting behavior of students in the network. Unfortunately, the researchers were not able to characterize the centralization of these structures. The researchers were able to identify cases in which null models were more or less descriptive of various attributes. Learner degree and weighted degree were both relatively well explained by null models. However, weighted and non-weighted clustering were not represented well by null models. Since degree seemed to tell a story here, this might be something to explore more and is a reason why it was chosen to be examined in this study.

This paper also gave indication that there may be better visual methods for expressing results. Reading and comprehending the charts and graphs in this paper proved to be a challenge. This provided a better perspective for making the visuals for this paper.

The study conducted by Cho et al. [4] discussed how Computer-Supported Collaborative Learning (CSCL) and Cooperative Work (CSCW) manifest in an online student collaborative network. The experiment discussed in Cho et al. [4] follows 31 college engineering students in two distant universities who collaborated for design of aerospace systems using online tools. It follows two

semesters, in which a survey was given in the second week of the first semester and then a final survey at the end of the second semester. The surveys ask questions about who students interact with meaningfully to better understand what relationships existed at the beginning versus the end of the study.

There are three main purposes of the study. The first was to explore how distributed learners create and maintain collaborative learning networks in CSCL and CSCW settings. The second purpose was to identify what structural and personal factors influence the collaborative structures. The third purpose was to test how these social network properties influenced learning outcomes (i.e. student grades).

Two main hypotheses were investigated in this study. First, it is hypothesized that higher willingness to communicate (WTC) students will explore more social ties than lower WTC students. The hypothesis that follows this is that these higher WTC students will also be more central to the network.

To sort through the results, multiple regression analyses were done to determine how much CS and pre-existing networks affected the ways in which new social ties were created. Hypothesis 1 was supported by the experiment. Students that displayed a higher willingness to communicate were more likely to engage in more significant interactions in the network. The second hypothesis was not supported by the data. Willingness to communicate had seemingly no effect on network centrality. This paper opens up for more discussion on the topic of centrality and its effect on a social network. This applies specifically to our work, in that it takes place in an online forum, classroom setting.

1.1.2 Blended/Offline Settings

Research can be conducted on groups that do not interact through a specific medium. For example, a community of students working a study group in which individual interactions cannot be quantitatively or qualitatively examined, but the participation could be recorded would be an example of this. Although the individual interactions of students would not be recorded, if a student is known to participate in a study group external to the classroom, it could still be analyzed how the student performs in the classroom.

Dawson [5] used social network analysis methods to look at learner interactions and the spatial/temporal requirements a classroom imposes, and the possibility of computer mediated communication to break through this potential barrier. Previous work for this paper gave the authors an idea that learning is obtained by individual participation in social interactions. These allow learners to interact with each other and therefore learn more by making connections. Dawson performed a quantitative as well as qualitative analysis on 25 classes. The qualitative analysis consisted of looking at discussion forum content. Student interviews were also looked at to better understand position in social network and perceived sense of community. The quantitative analysis consisted of social network analysis and centrality. These attributes helped determine a learner's level of "sense of community" and position within the network. Face to face lectures were not used in the study, it was strictly online forum analysis.

Ordinary Least Squares regression was used as an analysis tool. The study found a relationship existed between a student's sense of community and position within the conceived social network, which supported their hypothesis. Students may engage less if they feel they do not need to, which is characterized as weaker ties to surrounding classmates because they do not share resources. Conversely, students who feel isolated may experience frustration and a lack of needed help.

Florida International University (FIU) implemented a physics learning center for students to build a sense of community, and studied the effects this had on the classroom [6]. These physics learning centers are important because they exist informally and therefore do not require any specific background for participation, so they are open and inclusive networks for students to collaborate in their learning process. The learning centers could house up to about 30 students at one time. Social network analysis was used to examine outcomes. When doing the social analysis with this study, there were four main assumptions being made: actors and interactions are interdependent, links (such as resources, shared information, etc.) aids in flow between actors, network models for individuals simultaneously restrict, but also open up opportunities for individual action and finally, network models illustrate structures of patters and relationships among actors. The idea was that identifying patterns in these collaborations would create a better understanding of how the learning of physics happens, as well as understanding what could lead to retention of students within the major. One idea investigated was whether or not there was a power distribution evident enough to trace through actor attributes such as gender or ethnicity. To study this idea, an eigenvector central-

ity measure was calculated for all nodes in the network data. The data being analyzed came from an online survey students voluntarily participated in. The survey looked at six major attributes: Major, whether the student was in a modeling instruction course, gender, ethnic background, number of days per week in physics learning center and number of hours per week in physics learning center. A matrix was created that illustrated whether or not students interacted. Correlation coefficients, as well as hierarchical multiple regression analysis were also used. The hierarchical multiple regression analysis created predictive models for use in study. A notable finding from this study was that ethnicity and gender did not contribute to any perceived disadvantages in academic performance. Days per week spent in the physics learning center was extremely predictive of success in the classroom. More days spent increased a student's centrality score. This study led us to think more about how participation outside of the classroom would affect centrality and performance and thus, the use of online discussion forums such as those used in our study.

1.1.3 Literature Reviews and Methodology Papers

Cela et al. [7] performed a systematic review in which 37 papers were studied to look into three major focuses of SNA (if the use of SNA is increasing, to identify research questions and constructs used in SNA, and identify gaps and suggest future research). Throughout the literature reviews, certain things were identified like whether a study used strictly SNA methods or paired with content management systems. Another topic looked into was what technology was used to perform SNA and if that tech was pre-existing or created by the researcher. From the literature reviews performed, it was noted that centrality and density are some of the most relied on constructs. Suggestions for future research were made at the end of the paper, like broadening the range of what is being analyzed, as well as using 2 mode e-learning networks. A lot of this directly relates to what is being done in our own study. We also use centrality and density, but we use many other characteristics seen in SNA methodologies. For our SNA analyses, we use R programming, which would be the used technology of the study.

The purpose of Freeman [8] was to dive more into what centrality actually is and to better clarify how it can be used in the study of social interaction. Centrality describes how an individual is connected to and influences others in a network. More central nodes tend to have more connections and hold more influence over others in the network.

Centralization is a measure that describes how edges are distributed. Point and graph centralization were both discussed and the problems people have had understanding them over the years. It is stated that centrality can be determined by referencing degree, betweenness and closeness. Degree describes how many connections a node has to other nodes in a network. Betweenness describes how often points fall between other points on a geodesic path. Closeness is a descriptor of the independence of a point.

Other papers explore non-human research in which the points of interest are things like papers instead of people. Looking into the patterns and characteristics these exhibit can also be telling of networks that are formed. These could provide a unique insight to network analysis that one would not gain from purely human samples.

Barrat et al. [9] looks at two different network types as well as attributes to use to analyze them. This paper was the first in which two different network types were looked at: an airport and authors collaborating on papers. This paper had a heavy emphasis on the use of clustering coefficients. Topological and Weighted clustering coefficients are used to compare the two networks. Each of these have their own merits for the study.

This paper looks at two different types of clustering coefficients. Topological clustering coefficients can be described as the average clustering coefficient whereas the weighted clustering coefficient takes into account the weighted edges of a network. Topological coefficients indicate constraints on network structure, while weighted clustering coefficients focus more on nearest neighbor degree in accordance with the normalized weight of connected edges. The weighted clustering coefficient gives a better idea of cohesiveness in areas of the network. One observed network looks at airports and flights coming in and out. This network was known as the Worldwide Airport Network. The weighted clustering coefficient in this network was extremely variant. This says that in the network, the higher degree airports tend to create interconnected groups with higher traffic links. The other network looks at author collaborations when working on scientific papers. This network suggests that authors who do not work with as many collaborators work in a well defined group, while higher degree authors work with a wider array of collaborators. These higher degree authors have a lower clustering coefficient, while the low degree authors have a high clustering coefficient. In the network dealing with authors, weighted and topological clustering coefficients were almost identical while the weighted coefficient in the airport network was slightly larger than the

topological, probably due to the constant flux in and out of the airport of flights that are generally unconnected. This study claims to offer a general, quantitative approach to understand convoluted make ups of real weighted networks.

Chapter 2

Methods

2.1 Overview

In the following sections, an explanation of how the research was carried out will be given. This will begin with a section about network measures that were calculated and used to describe our networks. This will help one better understand how our networks were characterized and what these metrics were used to describe about our models. After the network measures are stated, we will go into how they were calculated. Finally, the specific creation of each individual model created for the study will be explained. This will help set the three models used apart from each other.

2.2 Network Measures

In this section we will discuss some of the network measures used in analysis. These are tools which were highlighted in the literature review and noted as useful, specifically in [10]. Here, our nodes are the students in the network and our edges are the post interactions between students. In the following definitions, it can be assumed N_E = number of edges and N = number of nodes.

Undirected Network: A network in which edges are not directional to and from nodes. Generally, an undirected network will have less edges than a directed network.

Directed Network: A directed network is a network in which edges are directed to or away from nodes, giving opportunity for more edges to exist in the network. There can be more edges in a

directed network because there could potentially be an edge to a node as well as a returning edge. In other words, there could be two edges between two nodes as interactions are directed to and away the starting and ending node. In an undirected network, since the direction isn't taken into account, there would only be one edge per interaction.

Degree: The degree of a node is how many connections it has:

$$k_{avg} = N_E/N \tag{2.1}$$

If the network is a directed network, there are separate degrees for in and out of nodes. The average degree of a network is the number of edges divided by the number of nodes. A higher average degree value means a network is more densely connected.

Density: The density of the network is the ratio of actual edges to total possible edges. This is a very common measure used in network analysis.

$$\rho = \frac{N_E}{N(N-1)/2} \tag{2.2}$$

and for directed networks:

$$\rho = N_E/(N(N-1)) \tag{2.3}$$

Centralization: describes how the edges in a network are distributed. Degree centralization looks at this in terms of it being a proportion. The highest value possible for this metric is 1.00 which would indicate that every single edge is connected to a singular node. The lowest value possible here would be 0.00 which indicate an isolate without connection [11].

$$C = \frac{\sum C_{Dmax} - C_D(n_i)}{max \sum C_{Dmax} - C_D(n_i)} \tag{2.4}$$

Above C_{DMax} represents largest degree centrality in the data set, while C_D is the degree centrality for a particular node.

Geodesic Path: The geodesic path is the "shortest length" of edges between two points. "Shortest length" is quoted here because when talking about geodesic path and diameter we are not talking about a physical length, but a number of edges. So, a longer length means there are more edges be-

tween two nodes and the shortest path means the least number of edges. It is not a single calculation that can be made for a network, but instead is a concept that can be used for different measurements.

Diameter: the length of the longest geodesic path between two vertices. So, a higher diameter would indicate more edges between nodes, or further length to travel. This would indicate a more loosely connected network.

Average Vertex - Vertex Distance: The average distance between vertices in a network. The calculation for average vertex-vertex distance includes distances from a vertex to itself, which would equate to a zero distance. For instances where there a connecting path between vertices does not exist, infinity is used in calculation.

Transitivity (clustering coefficient): compares the number of closed triangles of nodes to the number of all connected node triplets. In other words, this is the probability of adjacent vertices in a network to be connected. This helps to display small clusters. Essentially, this is a ratio between the closed clusters seen and the maximum possible clusters. You could see this as the probabilty of two people I know, knowing each other.

$$C = (3 * N_t)/V_c \tag{2.5}$$

Where N_t is the number of connected triangles and V_c is the number of connected triples of vertices.

Averaged Barrat clustering coefficient: This is a weighted clustering coefficient, whereas the previously mentioned clustering coefficient is an average. It helps to examine local cohesiveness of triplets.

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh} \tag{2.6}$$

Above k_i is the degree of node "i", while s_i is the sum of all the edge weights which are connected to the node. Together the fraction on the left of the summation counts for the normalization factor of the equation. $w_{i,j}$ and $w_{i,h}$ are weights on endpoint nodes' degrees. Finally, $a_{i,j}$, $a_{i,h}$ and $a_{j,h}$ are all points in the network's adjacency matrix. The values of these matrices can be 0 or 1 depending on whether or not the nodes are connected by an edge. For example, if node i and node j are connected by an edge, then the value of $a_{i,j}$ would be 1. If they are not connected, the value of $a_{i,j}$ would be 0 (Barrat et al. [9]).

Correlation Coefficient: Measures how strong a relationship is between two variables. Correlation coefficients can range from -1.00 to 1.00. A data set that shows absolutely no correlation would have a correlation coefficient of 0.00. By "no correlation", this means a data set is completely scattered and shows absolutely no trend. A negative correlation coefficient describes a data set that trends downward, while a positive correlation coefficient indicates an increasing trend. The closer to 1.00 on either side of the positive or negative spectrum a correlation coefficient is, indicates that the data increases or decreases more linearly. For example, a positive correlation coefficient of 1.00 depicted on a graph, would look like a diagonal line starting at zero and increasing positively on a y - axis at a linear rate, as it goes up the x-axis. Here, the correlation coefficient is describing whether or not high degree nodes tend to link with other high degree nodes and if low degree nodes tend to link with other low degree nodes.

$$\text{correlation coefficient} = \frac{\sum(x_s - x_{avg})(y_s - y_{avg})}{\sqrt{(x_s - x_{avg})(y_s - y_{avg})}} \tag{2.7}$$

Community Detection: Groups of nodes which are densely connected to each other. These can be the "clusters" talked about in the transitivity definition, but even moreso can be much larger clusters than just triangles. There are a plethora of ways to calculate community detection, In our calculations we use the edge betweenness methods and the [12] function "infomap". There are other ways such as the "spring embedding" algorithm or cluster analysis, which are both discussed in Newman [10].

2.3 Foundations/Code

All three models were created using the "igraph" [12] package in R Studio [13]. The data begins as a data frame in which the categories from left to right are: type, student code, target, thread, thread week and creation time. Type describes if the post is the original post in the thread, or if it is a reply to the post. Student code is the identification number of the student who posted, or commented on the original post. Target gives the student identification code of the student who created the thread. Thread week describes the week in the semester that the post took place and creation time is the time stamp of the post for following chronology. The data frames containing this information are

the basis of all calculations, plots, tables, etc. done in this study. Table 2.1 below shows an example of the data frames we were working with.

2.3.1 General Model Construction

Each model was created using its own separate code file due to each model being distinctly different from one another, but the beginning formations of the models were done in a code file together. The first few steps of model creations were the same across all three, so having this separate foundational file served as a good reference for coming back to whenever needed. The three models use the same data sets which are supplementary online forum discussions between students from fall semesters of a introductory physics course for fall of 2014, 2015 and 2016. As mentioned earlier, every post is given a line in the data frame that contains its student code, target code, thread, week and creation time. These can be seen in the following table.

Table 2.1: Sample Data Frame

Type	Student.Code	Target	Thread	Week	ThreadWeek	Creation.Time
Refl	86689010	56573045	935	16	16	1418910988
Refl	16571160	56573045	935	16	16	1418915573
Refl	86978019	56573045	935	16	16	1418924031
Refl	98365241	56573045	935	16	16	1418926187
Post	97883254	97883254	934	16	16	1418829664
Refl	18783138	97883254	934	16	16	1418830073
Refl	96870056	97883254	934	16	16	1419043937
Post	18783138	18783138	933	16	16	1418828996
Post	28783140	28783140	932	16	16	1418828354
Post	97777328	97777328	931	16	16	1418819566
Refl	18783138	97777328	931	16	16	1418829870

Table 2.1 is a small subsection of threads. You can see the all pieces of the tables that we worked with when creating these data frames. It tells us about specifically week 16, (which can be seen in the fifth column from 2.1) of the fall 14 semester. The student code, target, type and thread columns are the building blocks of what differentiates each model from one another.

2.3.2 Model 1 Construction

The premise of model one is that all participants in the same thread are connected. More connections lead to weighted edges in this model. In Figure 2.1 this is illustrated at the most basic level with

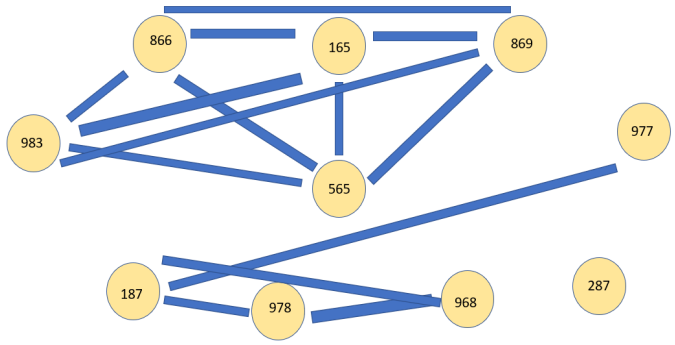

Figure 2.1: Example Model 1 network from data in Table 2.1. Nodes are labeled with the first three numbers of the StudentCode. Each student is connected to all others in the same thread.

a sample data set in Table 2.1. Model 1 utilizes a bipartite network structure. The data is divided into two ”types” that the code can decipher between (this being the indication that it’s a bipartite network). The two types of data are actors and events. In the case of our study, the actors are students and events translate into threads. In a bipartite network, nodes can only link to the other node types, so relating this to our study, students can only link to threads. So, a projection links all student nodes that link to the same thread nodes. When linking the student nodes that link to the same threads, weighted edges are created. Weighted edges in the network illustrate that there are multiple links between two students. Within this process the loops and isolates are removed from the data set as well. This means links are not shown when a student replied to themselves (loops) and there are not illustrated nodes for posts in which there are no replies (isolates). It is important to note as well that direction is not used in model 1, meaning that it is not taken into account of who made the post versus who is making the reply to the post, only that an interaction between two students happened. Since there is no direction of the threads (our edges), this means model 1 is an undirected network.

2.3.3 Model 2 Construction

The idea behind model 2 is that the only edges are those between a student who posted in a thread and the student who started the thread. The links in this network model are directed links meaning they point from one node in the direction of another (i.e. a directed network). In our study, they point FROM the student who commented TO the student that began the thread, which can be seen

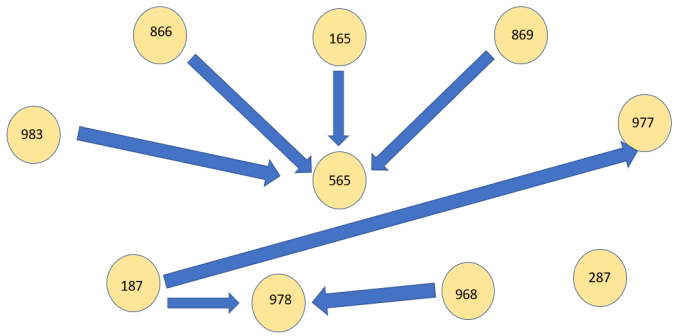

Figure 2.2: Example Model 2 network from data in Table 2.1. Nodes are labeled with the first three numbers of the StudentCode. The only edges are those between a student who posted in a thread and the student who started the thread

above in Figure 2.2.

A weight of 1 is assigned to each edge, otherwise the program will treat the edges as unweighted. Then the ”simplify” command is used to get rid of loops, as well as combine multiple edges into weighted edges. So, if a single student has interacted with another student more than once, simplify will combine this to a weight of however many interactions was exchanged.

Next, the simplified model 2 objects are created by clearing away anything with a degree of 0 from the model objects. This gets rid of isolates from the data system, which are nodes that have no connections to any other nodes in the system. In this model, isolates would be defined as students who never commented on a post, or had any replies sent to their posts.

2.3.4 Model 3 Construction

The overarching concept of model 3 is that each student replying to a thread has an edge to everyone else who has posted above them in the thread (including the student who started the thread). This can be seen in Figure 2.3 above at a foundational level. An empty object is created for later use in the model construction. Next the data is looped through. First, the R program separates all the data frames into the individual threads that make it up. From here, the code parses through all the student codes that appear within a thread of the data frame. After the program loops through, a plot can be generated that shows the outcome of the loops. Essentially, the plot shows how the student codes connect. For any node that has an edge pointing to itself, it represents a student responding

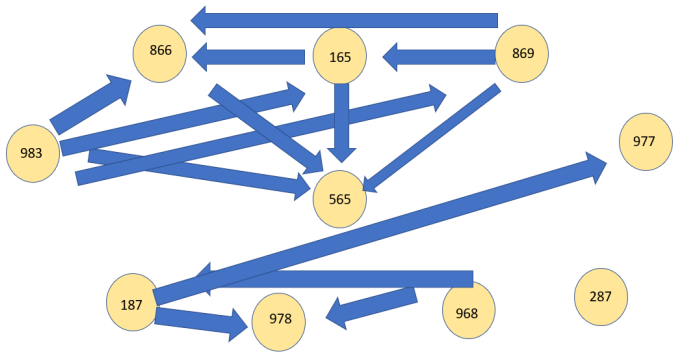

Figure 2.3: Example Model 3 network from data in Table 2.1. Nodes are labeled with the first three numbers of the StudentCode. Each student replying to a thread has an edge to everyone else who has posted above them in the thread (including the student who started the thread)

to their own post at some point of the thread. After looping through the data, an edge weight of one is assigned to the set, then the network objects are simplified to remove loops and collapse multiple edges. This process is repeated for all three semesters of data.

2.3.5 Additional Comments

Once all of the models were created various calculations were done on the models to test different network measures. The models were created so that model 2 would in theory be the network with the least amount of connectivity. Model 3 would have an intermediate level of connectivity of the three, and model 1 should have been the most highly connected model. The calculations that will be discussed in the results section are as follows: centralization, average degree, average Barratt clustering coefficient, diameter, transitivity, average vertex-vertex distance and correlation coefficient. There metrics were all calculated using the ”iGraph” package [12] in R Studio [13]. Community detection was also investigated, as well as the beginnings of a more in depth analysis of specific aspects of centrality such as Pagerank and degree and corresponding quartiles for both of these metrics.

Chapter 3

Results

There are a few different ways to view and think about the results. Each way can be helpful in its own way when considering the data and results that are gleaned from the study. One way to consider the results is by looking at them by characteristic, that is, looking at the results of measures and comparing/contrasting the models within the measure being analyzed. This allows one to look at how results may vary from one model to another. Another way to compare is just looking model by model at the results which helps when looking for similarities and differences between semesters.

With this in mind, to begin, we will look at a subset of the data from each model and show how they are mapped out visually. The example used is a subset of data from week 7 of the semester.

Figure 3.1 shows the week 7 network for model 1, in which anyone in the same thread has a connection to each other. This one differs from the next two in that there are no arrows to indicate the direction of interaction. This is what illustrates it as an undirected network. Another important thing to note when looking at Figure 3.1 is that the lines denoting edges and the yellow circles denoting nodes are different thicknesses/sizes. The edges are sized as a function of the edge weight. This means edges with higher weights (multiple interactions between two nodes) have a thicker line than an edge where there was only one interaction between the nodes (an edge weight of one). The node size is a function of the degree of the node. This means a higher degree node, so one with more connections, will have a larger circle than a node that only has one other connection. By our way of model construction, model 1 should have the most dense looking network of the three models.

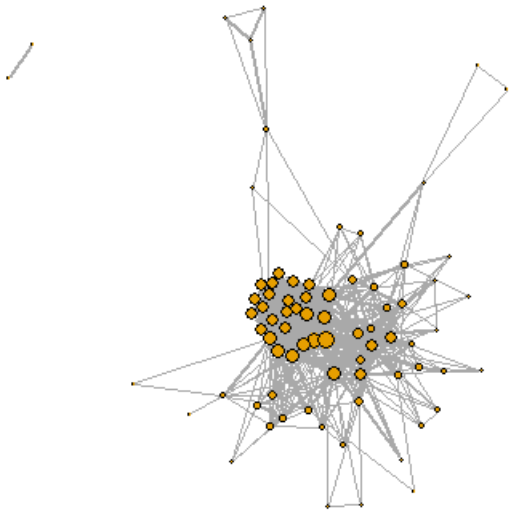

Figure 3.1: Week 7 Subset of Model 1 Network; the edge width is scaled by the edge weight and the node size is scaled by the square root of the degree of the node.

Let us remember that in model 2, the only edges are those between a student who posted in a thread and the student who started the thread. Figure 3.2 shows the arrows which indicate the direction of an interaction. These arrows delegate this figure as a directed network. Again, in the model 2 network plot, you can see different edge and node sizes to show which students (nodes) participated in more interactions with other students, and the weight of these interactions. By construction of the models, this is the model that should have the least amount of connections in its structure.

In model 3 each student replying to a thread has an edge to everyone else who has posted above them in the thread (including the student who started the thread). Figure 3.3 shows the arrows that designate this as a directed network. Again, in model 3 the edges are sized as a function of edge weight and nodes are sized as a function of their density. According to model construction, Figure 3.3 should be more of a middle ground between Figure 3.1 and Figure 3.2.

By looking at these figures, you can get a better idea of how the networks are connected to each other, and what it might look like for nodes of higher or lower connectivity. For example, in 3.1 there are some nodes to the top left of the figure that are off to the side and connected only to each

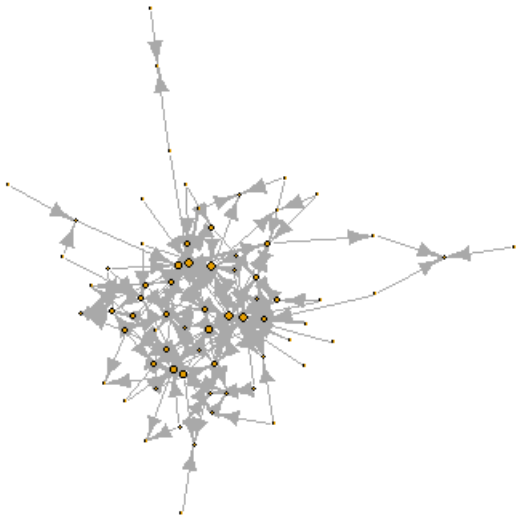

Figure 3.2: Week 7 Subset of Model 2 Network; The edge width is scaled by the edge weight and the node size is scaled by the square root of the degree of the node.

other. The nodes themselves are smaller compared to some of the others as these are examples of nodes that are less connected to the network, versus one of the nodes that is in the middle of the web-like network, where it is way more interconnected with the network infrastructure and therefore larger in size as well.

3.1 Model Overview

3.1.1 Model 1

Table 3.1: Model 1 results with columns from left to right: semester, average degree, centralization, average Barrat coefficient, diameter, transitivity, average vertex - vertex distance, correlation coefficient

semester	AvgDeg	Central	Barrat	Diameter	Transitivity	AvgVertToVert	Correlation
1	52.97	0.43	0.72	4	0.64	1.69	-0.06
2	29.40	0.48	0.70	8	0.53	1.90	-0.19
3	42.14	0.47	0.79	7	0.63	1.78	-0.14

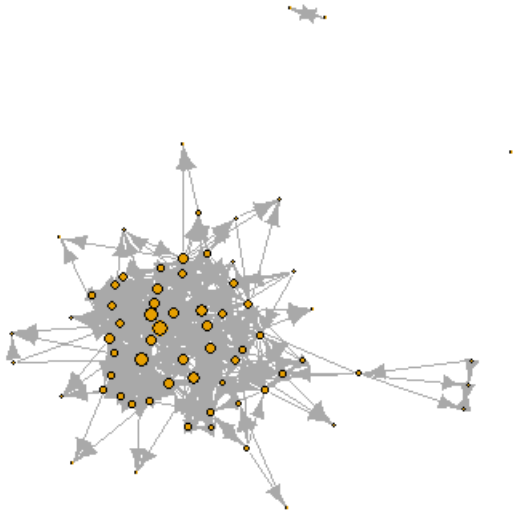

Figure 3.3: Week 7 Subset of Model 3 Network; the edge width is scaled by the edge weight and the node size is scaled by the square root of the degree of the node.

Table 3.1 shows an overview of the descriptive network measures calculated for each semester in model 1. Looking at average degree across the semesters, Fall 14, 15 and 16, it shows that the highest value for average degree was in Fall 14 and the lowest in Fall 16. All three semesters exhibit highly comparable centralization values. The values for average Barratt coefficient and transitivity are also extremely close to one another. Fall 15 had the highest value for diameter while Fall 14 was the lowest for the three, being half of what Fall 15 was. Fall 15 had the highest vertex to vertex distance of the three semesters while Fall 14 had the smallest value. Semesters 2 and 3 have comparable correlation coefficients, but there is a gap from these to semester 1. Semester 1's correlation coefficient is almost zero, showing an almost total lack of correlation here for high degree nodes connecting to other high degree nodes and low connecting with low. The correlation coefficients all being negative numbers is something to note as surprising, but this will be discussed more later.

3.1.2 Model 2

Table 3.2: Model 2 results with columns from left to right: semester, average degree, centralization, average Barrat coefficient, diameter, transitivity, average vertex - vertex distance, correlation coefficient

semester	AvgDeg	Central	Barrat	Diameter	Transitivity	AvgVertToVert	Correlation
1	23.15	0.28	0.41	7	0.31	2.39	-0.11
2	20.40	0.32	0.58	7	0.36	2.39	-0.18
3	23.58	0.34	0.55	6	0.39	2.42	-0.16

Table 3.2 shows an overview of the descriptive network measures calculated for each semester in model 2. Starting with average degree there are three values across the semesters that are all comparable. Centralization values in model 2 are also very comparable, as they were in model 1. The average Barratt values varied from 0.41 (fall 14) to 0.58 (fall 15). Diameter values are almost identical to each other, as are average vertex to vertex values. Transitivity values are extremely similar to each other as well. Fall 15 and 16 exhibited similar Barratt clustering coefficients, but Fall 14 was noticeably lower than the other two semesters for this characteristic. The correlation coefficients here were also comparable to one another, as well as in general being relatively close to zero and mostly showing a lack of any correlation.

3.1.3 Model 3

Table 3.3: Model 3 results with columns from left to right: semester, average degree, centralization, average Barrat coefficient, diameter, transitivity, average vertex - vertex distance, correlation coefficient

semester	AvgDeg	Central	Barrat	Diameter	Transitivity	AvgVertToVert	Correlation
1	70.85	0.39	0.82	5	0.64	1.84	-0.06
2	40.81	0.43	0.84	6	0.53	2.00	-0.16
3	58.55	0.46	0.99	4	0.63	1.91	-0.13

To begin, remember that model 3 is a network in which each student replying to a thread has an edge to everyone else who has posted above them in the thread. This includes the student who started the thread.

Average Degree values in this model vary from 40.81 (fall 15) to 70.85 (fall 14). The average degree numbers in this model indicate a moderately connected network. The centralization values in this network are all very comparable, ranging from 0.39 (fall 14) to 0.46 (fall 16). Average

Barrat coefficient values here were all reasonably high, especially in comparison to the other two models. All of the diameter values are one unit apart from each other showing consistency within the diameter attribute for this model. The average vertex to vertex values in this model are also very similar to each other. The lowest average vertex to vertex value is 1.84 (fall 14) and the highest being fall 15, by not too much more. The correlation coefficients in this model are very comparable as well. The most positive value appears in fall 14 and least positive being fall 15. These values are relatively close to zero, which indicates an overall lack of correlation in the model.

3.2 Compare All Models

Table 3.4: Comprehensive Results Table

Model	semester	AvgDeg	Cent	avgBarrat	Diameter	Tran	AvgVert	Corr
1	1	52.97	0.43	0.72	4	0.64	1.69	-0.06
1	2	29.40	0.48	0.70	8	0.53	1.90	-0.19
1	3	42.14	0.47	0.79	7	0.63	1.78	-0.14
2	1	23.15	0.28	0.41	7	0.31	2.39	-0.11
2	2	20.40	0.32	0.58	7	0.36	2.39	-0.18
2	3	23.58	0.34	0.55	6	0.39	2.42	-0.16
3	1	70.85	0.39	0.82	5	0.64	1.84	-0.06
3	2	40.81	0.43	0.84	6	0.53	2.00	-0.16
3	3	58.55	0.46	0.99	4	0.63	1.91	-0.13

Table 3.4 shows a comparison of all three models for the following attributes: average degree, centralization, average Barratt clustering coefficient, diameter, transitivity, average vertex-vertex distance and correlation coefficient. The top three rows show semester, Fall 14, 15, then 16 for model 1. Respectively, models 2 and 3 follow. Seeing these measurements all together in one location helps us compare and contrast results in an easier to visualize way.

The following plots show all three semesters, of all three models for each attribute. The x-axis indicates which model is being viewed and is ordered from left to right: model 2, model 3, model 1. This order was specifically chosen to express what was expected to be the least connected models, to the most connected model. Viewing them in this specific order allows us to better understand if our expectations were met for the connectivity of the models. The y-axis shows the value for whichever attribute is being examined.

While looking at these plots there are general trends we would expect to see with each network

attribute. Some of the attributes rely more on nodes of the network, while some of them rely more on the edges of the network. With a higher edge count, the average degree and the variations of clustering coefficients would be expected to increase with connectivity. However, with a higher edge count things like diameter and average vertex-vertex distance would be expected to decrease, so we would see these graphs go in the opposite direction. Keep in mind, a higher edge count would indicate a more highly connected network.

The number of nodes and edges for each semester of each model are shown in Table 3.5. In the directed networks, edges would be counted for interactions based on being to or from one student to another which could potentially result in two edges for an interaction. In an undirected network an edge would exist with no direction so there would only be one edge for any given interaction, which means that directed networks have the potential to have more edges than an undirected network. There is a consistent number of nodes across all three models. This makes sense, because we are working with the same number of students. The edge counts differ due to the differing structure of each model determining what we consider a connection.

Table 3.5: Nodes and Edges Comparison Table

semester	Model	Nodes	Edges
1	1	144	3814
2	1	126	1852
3	1	139	2929
1	2	144	1667
2	2	126	1285
3	2	139	1639
1	3	144	5101
2	3	126	2571
3	3	139	4069

3.2.1 Centralization

Figure 3.4 shows centralization values from the data set. Centralization showed an increase looking left to right on the plot. Semester one consistently had the lowest centralization value. Semesters 2 and 3 were higher, but varied by model for which was the highest. Centralization is a metric that there was no baseline expectation for so it is interesting that it showed an increase with connectivity.

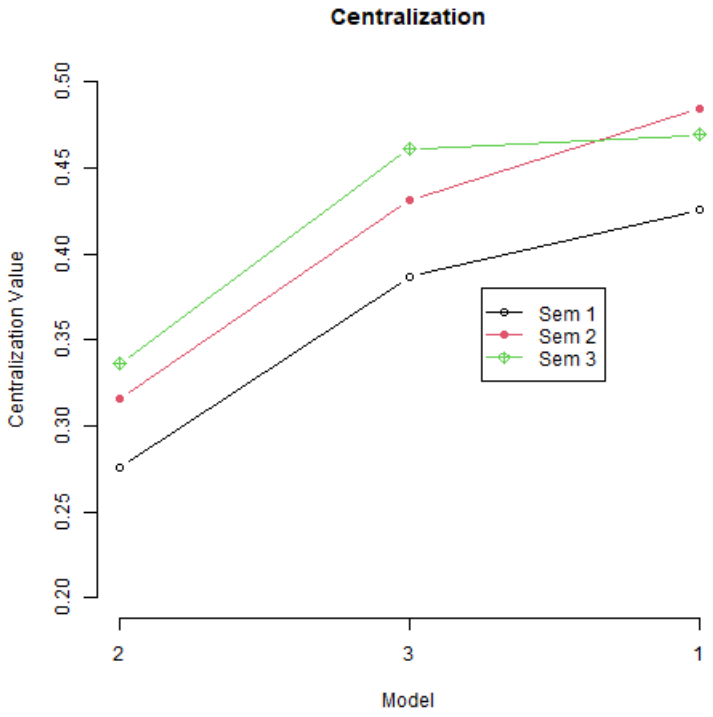

Figure 3.4: Centralization plot

3.2.2 Average Degree

Figure 3.5 showed the results for the average degree metric. On this graph, semester 1 tended to have the highest values, followed by semester 3, with semester 2 tending to be the lowest.

The Average Degree table does not exactly fit in the box of what we would expect to see. There is a spike where model 3 tends to have the highest values for all three semesters, instead of a general increase going from left to right. The reasons for this will be explained in section 4.2.

3.2.3 Average Barrat Clustering Coefficient

Figure 3.6 shows the plot for the average Barrat coefficient calculations. This data has less of a trend than some of the other graphs discussed in this section. Overall, semester 3 tends to have the higher clustering coefficient values except in model 2, where semester 2 has the higher value. Semesters 1 and 2 closely resemble each others' values in models 3 and 1. Most of the clustering coefficients in this network fall in an upper middle range showing a fairly connected network.

The Barrat coefficient follows the same pattern as the average degree values where it peaks in

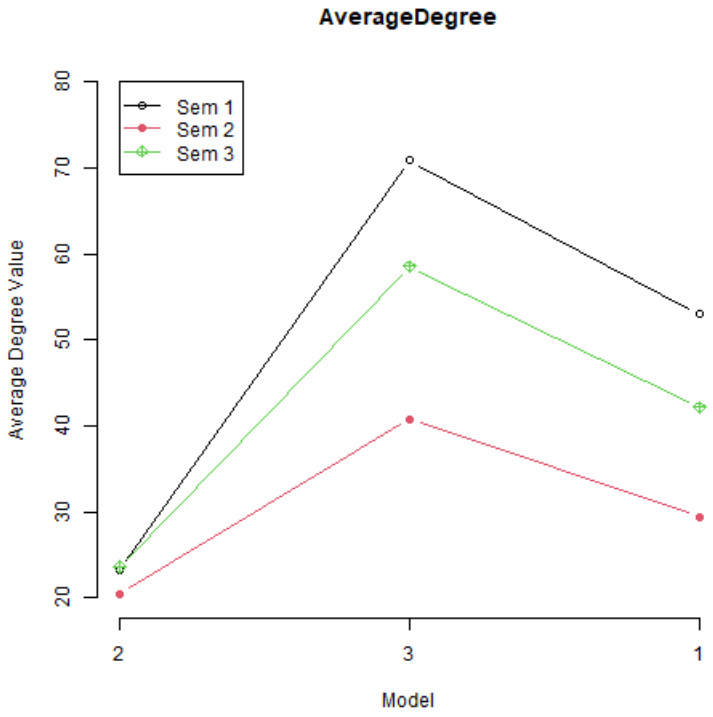

Figure 3.5: Average degree plot

values in model 3, which should be the middle values for what we would expect the graphs to look like.

3.2.4 Diameter

Figure 3.7 is distinct from the others in that it is the only figure in which one of the semesters (semester 1) completely deviates from the pattern the other two follow. Semesters 2 and 3 both decrease from model 2 to model 3 and then increase on model 1. Overall, semester two had the highest diameter values, but when making this statement it is important to note the y-axis values on the figure. The lowest value is 4, while the highest is 8, which is not a significantly large difference. The semester one line follows the expected trend for diameter with respect to the x-axis order discussed earlier and the increasing connectivity of the graphs. Diameter is the length between nodes, where length is attributed to the number of edges between the nodes. So a more connected network would typically have a lower value for diameter, as seen in the semester one values across the models. However, there is deviation from this trend in semesters 2 and 3 in the model 1 values.

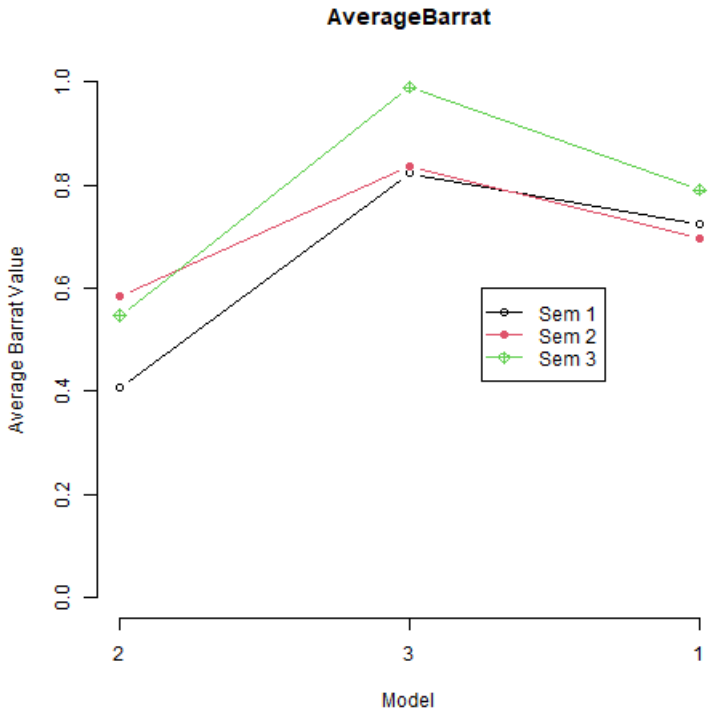

Figure 3.6: Average Barrat clustering coefficient plot

3.2.5 Transitivity

Figure 3.8 shows the results for the transitivity network measure. Transitivity is interesting in that the values from model 2 to model 3 follow what we would expect, but then from model 3 to model 1 there is no difference from point to point for each individual semester. This is due to igraph using a calculation that does not include edge direction. So, even if the weights differ, this has no effect on the calculation and explains the end plateau. Semester 1 begins as the lowest value in model 2, but for the remaining two models has the highest value. Semesters two and three are consistent with each other in that semester 3 holds higher transitivity values than model 2.

3.2.6 Average Vertex-Vertex Distance

Figure 3.9 shows the average vertex to vertex network measure results. The average vertex to vertex distances exhibit shorter distance going across the x-axis (model 2, to 3 to 1). This is exactly what would be expected for reasons similar to what is explained in the section about diameter. The general trend is that semester two shows the longest distances, followed by semester 3 and semester

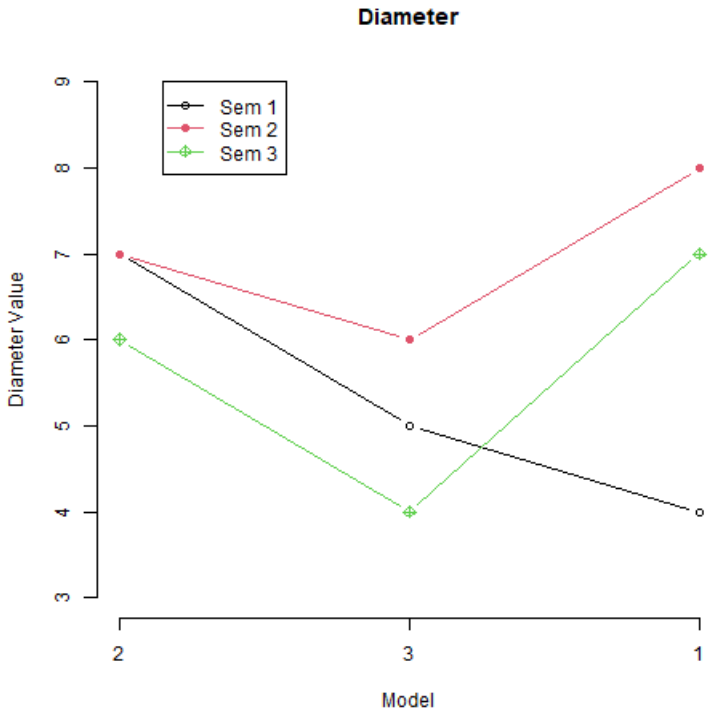

Figure 3.7: Diameter plot

1 having the shortest distances. The variant for this generality is that in model 2, semester 3 has the longest distance and semester 2 is slightly shorter and almost identical to semester 1. Another thing to note while looking at this graph, is that model 2 has the least variance in these numbers and model 1 has the most variance in its average vertex to vertex distance values.

3.2.7 Correlation Coefficient

Figure 3.10 shows the values for the degree correlation coefficients. Note that the y-axis begins at a negative number (-0.020) and works its way up to zero. These numbers are all negative values close to zero, which shows more of a lacking in correlation within the data set. By doing the correlation coefficients here, we were trying to determine if high degree nodes connect with other high degree nodes, which is typically the case seen in social networks. That is, we wondered if students who tended to link to many other students would tend to link to each other. These results proved it to be more random, with no particular trend as to whether students who were ”popular” in the class tended to link to each other.

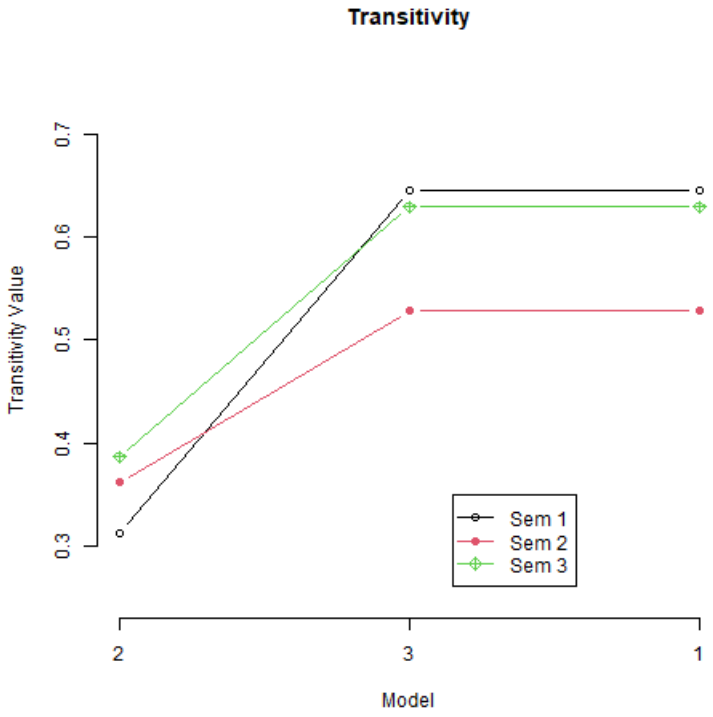

Figure 3.8: Transitivity plot

3.2.8 Community Detection

Community detection should be addressed as well in this section, although the figures are not included. This is because the results were not indicative of anything. For our calculations we used the methods of edge betweenness and informap in the igraph package [12]. When using the edge betweenness, essentially every node was defined as its own community. There were too many communities within this method to be of any significance to our research. When using infomap, all of the nodes were essentially bundled into one large community, which was also not useful to our research.

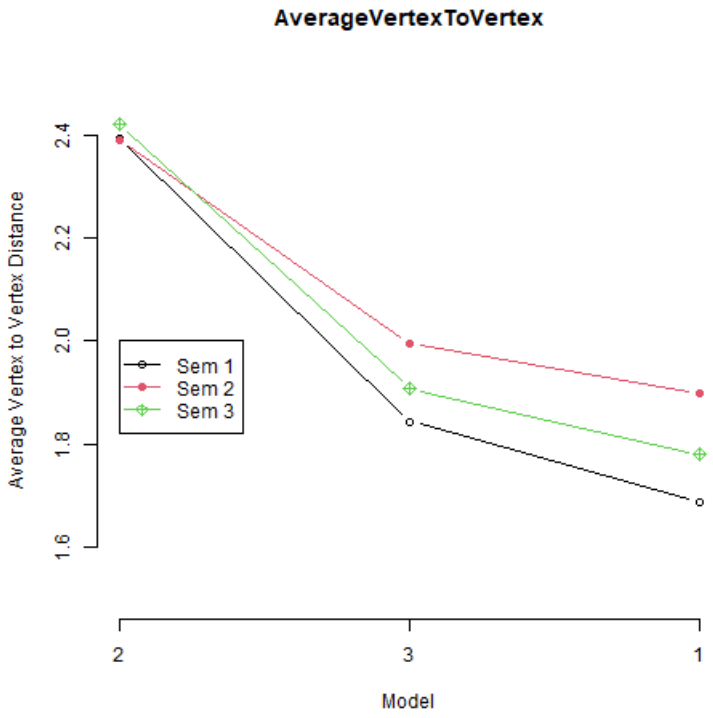

Figure 3.9: Average vertex-vertex distance plot

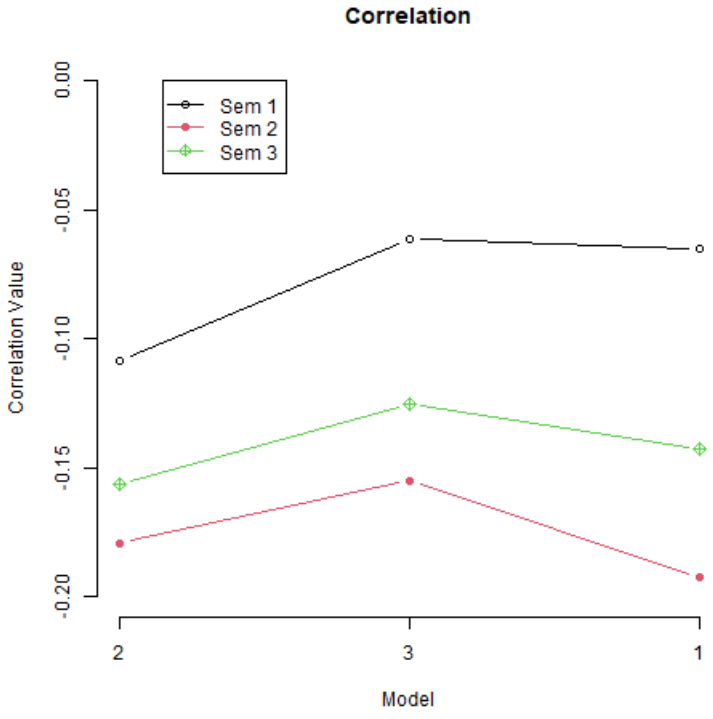

Figure 3.10: Degree correlation coefficient plot

Chapter 4

Discussion

4.1 Comparison With Example Networks

The numbers in Table 3.4 from the results section were compared with Table 3.1 from Newman [10]. The Newman paper was instrumental in our research because it covered a lot of important network concepts and measures which factored into the network measures chosen to be researched in this study. Table 3.1 from the Newman paper is a mostly comprehensive summary of the paper as a whole, which made it a good point of comparison for our results tables. Their table lists sample networks that range from four categories: social, information, technological and biological sciences. For each of these, the networks are either directed or undirected. There are eight listed attributes: vertices, edges, average degree, vertex to vertex distance, degree distribution exponent, transitivity clustering coefficient, average clustering coefficient and correlation coefficient. Their average clustering coefficient is parallel to our Barratt clustering coefficient. For our study, their category of social networks is the most relevant and will therefore be the main point of reference. Density is another commonly explored network attribute, but for the purposes of our study, did not hold a strong bearing and therefore was not examined.

In terms of nodes and edges, the Newman networks are generally much larger than our networks. Newman also saw an overwhelming amount of negative values for correlation coefficients except when it came to their social networks. Since their social network section is what is most

comparable to our work, it is an interesting thing to note that where their correlation coefficients were positive, ours ended up being negative. The ranges of the correlation coefficient values they saw were comparable to ours.

4.2 Unexpected Network Measure Trends

The Results section touched upon how a few of the network measure plots did not follow the trend one would expect to see in these models. The big question that would obviously result from this is: why not?

The commonality of the network measure plots that did not follow the expected trend is as follows. The network measures that had more to do with the edge number of the graphs followed the expected trend, until getting to model 1. Model 1 was the only undirected network, which would lead to less edges. This is a probable explanation for the drop off in edge related network values for what should have been our most connected network and therefore highest values. With this in mind, model 1 was re-examined, but as a directed network to see what might have changed. These new directed values can be seen in the table below. For the purposes of comparison to original findings, the undirected model 1 network renames named ”1” in the ”Model” column, while the newly calculated directed model 1 network is listed as ”4”.

Table 4.1: Model 1 Directed Network Node/Edge Comparison

semester	Model	Nodes	Edges
1	1	144	3814
2	1	126	1852
3	1	139	2929
1	4	144	7628
2	4	126	3704
3	4	139	5858

As seen in Table 4.1, the number of nodes remains the same from undirected to directed networks, which was seen in the original calculations of the undirected model 1 network in comparison to the directed model 2 and 3 networks. However, from the original undirected model 1 network to the newly calculated directed model 1 network, the edge numbers have doubled. This makes sense because in model one anyone in the same thread is connected, so there would be two counted edges for every interaction.

Table 4.2: Model 1 directed network results compared with original calculations for model 1 undirected network.

Model	semester	AvgDeg	Central	Barrat	Diameter	Tran	AvgVert	Corr
1	1	52.97	0.43	0.72	4	0.64	1.69	-0.06
1	2	29.40	0.48	0.70	8	0.53	1.90	-0.19
1	3	42.14	0.47	0.79	7	0.63	1.78	-0.14
4	1	105.94	0.43	0.68	4	0.64	1.69	-0.06
4	2	58.79	0.48	0.61	8	0.53	1.90	-0.19
4	3	84.29	0.47	0.73	7	0.63	1.78	-0.14

Table 4.2 shows the network measures for the undirected and directed model 1 networks. There is no change to the following network measures: centralization, diameter, transitivity and average vertex-vertex distance. However, there is a noticeable change in average degree and a small change in the average Barrat coefficients.

Referencing our original definition for degree, it is defined as ”how many connections a node has.” It makes sense that when the number of edges doubled, so did the values for average degree, because we have now doubled the connections. This explains why the average degree plot from 3.5 did not follow the expected trend. When using the model 1 calculations where it is a directed network it now follows the expected trend.

Next, we re-examined the network measure plots by recreating the original plots used in the results section, but adding a ”Model 4” on the x-axis. This ”Model 4” represents the newly calculated model 1 directed network.

Figure 4.1 shows the centralization with no difference from the original plot to this one. Making model 1 a directed network had no effect on the centralization values.

Average degree shown in Figure 4.2 showed a significant improvement by making model 1 a directed network, which was the expected outcome of this exercise. By adding more edges, this shows more connections entering and exiting the nodes, which should raise the degree values as shown. This supports our theory that directing the model 1 network would make the data follow the trend we would expect to see.

Looking at the Barrat coefficient graph, Figure 4.3, there was not a significant change from what we originally saw. Semesters 1 and 3 were similar to their original values and semester 2 slightly decreased here.

The diameter graph seen in Figure 4.4 showed no changes between the directed and undirected

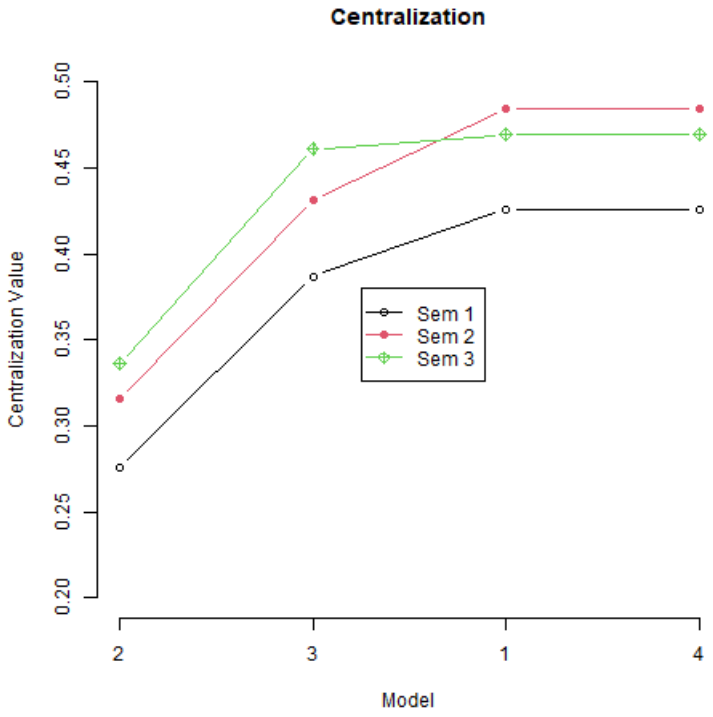

Figure 4.1: Centralization plot where values were identical between undirected model 1 and directed model 4.

model 1 networks. Our semester 1 line still shows what we would expect to see, while the semester 2 and 3 lines deviate from the expected trend.

The transitivity graph in Figure 4.5 and the average vertex-to-vertex distance in Figure 4.6 showed no change between the directed and undirected model 1 points.

Finally, the correlation coefficients in Figure 4.7 were identical to the undirected network calculations.

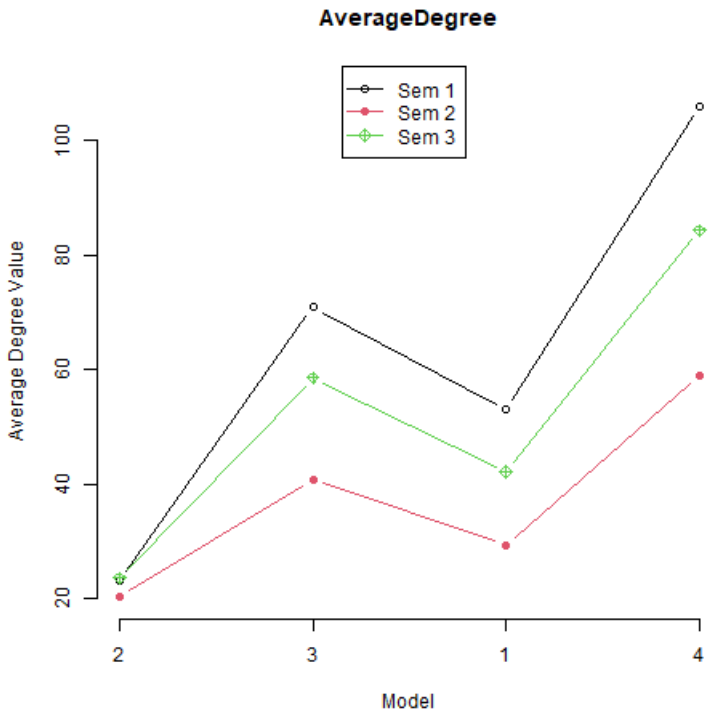

Figure 4.2: Average degree plot which shows an increase from undirected model 1 to directed model 4.

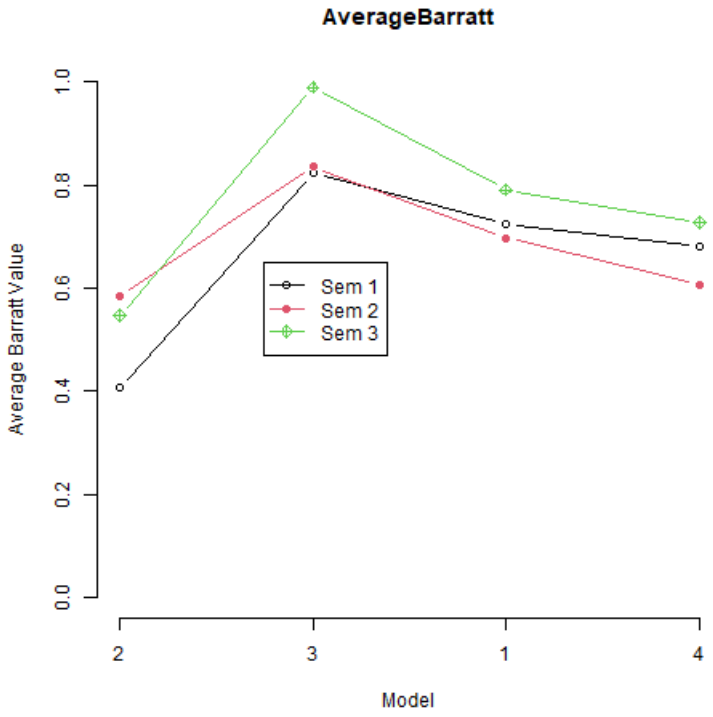

Figure 4.3: Average Barrat clustering coefficient plot which shows decrease from undirected model 1 to directed model 4.

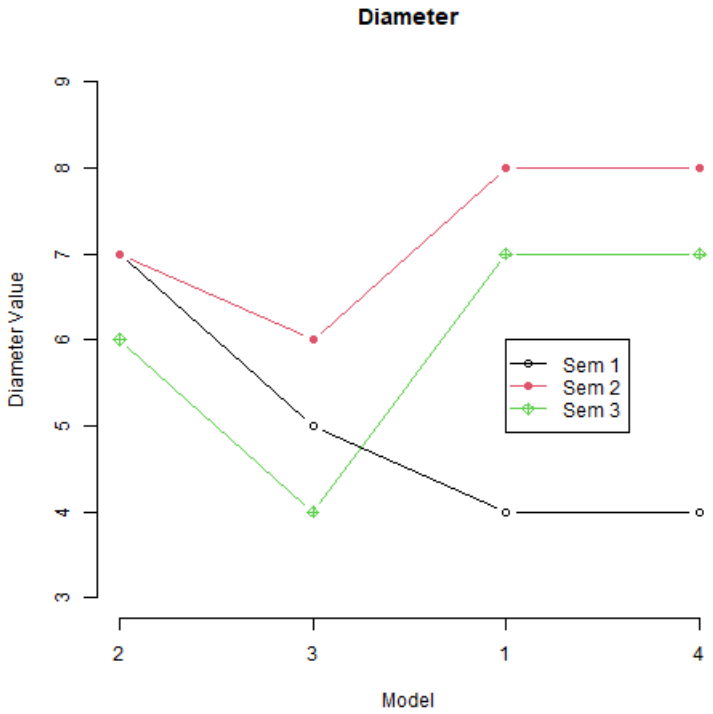

Figure 4.4: Diameter plot which shows no change from undirected model 1 to directed model 4.

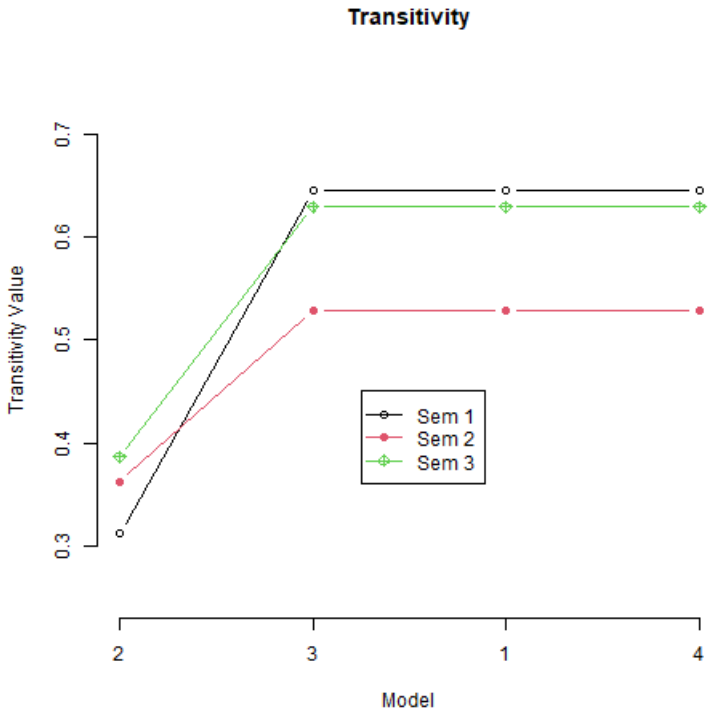

Figure 4.5: Transitivity plot which shows no change from undirected model 1 to directed model 4.

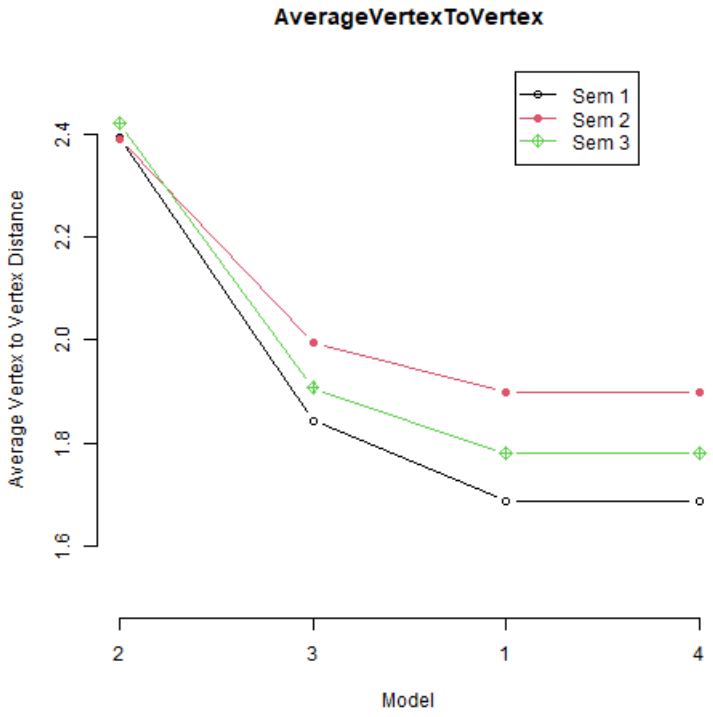

Figure 4.6: Average vertex-vertex distance plot which shows no change from undirected model 1 to directed model 4.

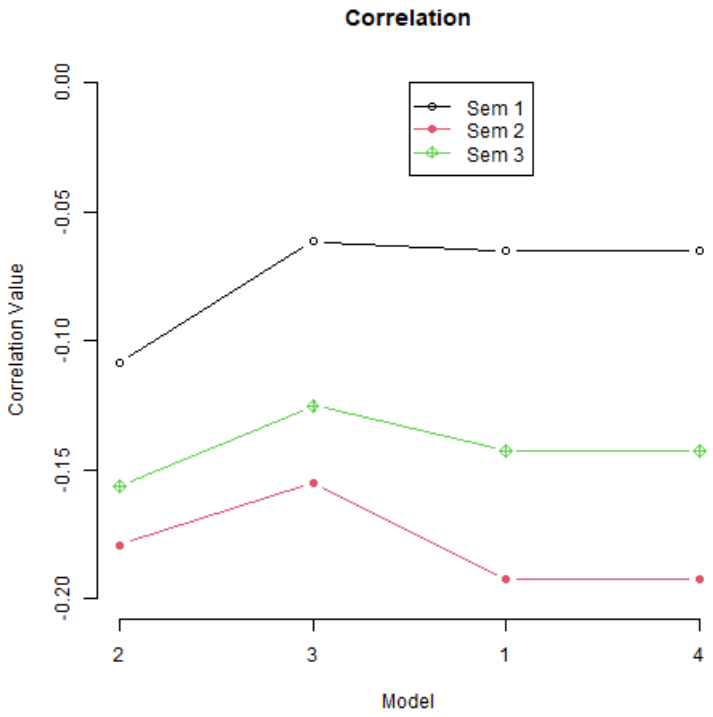

Figure 4.7: Degree correlation coefficients plot which shows very minimal change from undirected model 1 to directed model 4.

Chapter 5

Conclusion

5.1 Future Work

As research is never ending, there are an infinite number of network attributes that could have been examined and provided more information about the networks being discussed in this paper. One example of this would be a heavier emphasis on centrality specific measures.

Table 5.1: Sample Centrality Attribute Table

ID	degree	DegreeQuartile	PageRank	PageQuart
37865234	63	Q3	0.007	Q3
76771103	45	Q2	0.005	Q2
96870056	56	Q2	0.008	Q3
56573045	101	Q4	0.021	Q4
56687107	43	Q2	0.005	Q2
86689010	65	Q3	0.008	Q3
16571160	113	Q4	0.022	Q4
86978019	38	Q2	0.004	Q2
98365241	82	Q4	0.010	Q4
97883254	65	Q3	0.007	Q3

Table 5.1 shows a few extra network measurements that were beginning to be calculated in our research. Due to external constraints, these were not explored as heavily as desired, leaving room for more potential work to be done. The table 5.1 shows five columns of data: ID number, degree, Degree Quartile, PageRank, and Page Quart (short for Pagerank quartile). These attributes are associated with different centrality measures, to see if changing models also changes the centrality

quartiles that students fall into. This would give us more insight into whether or not the same students are central to the network regardless of the assumptions being made about the interactions going on within the network. It may also give more insight into how they reached a more central position via their interactions with other students.

Since community detection in this particular experiment was inconclusive, it would be interesting to further analyze this by using some different methods. As previously mentioned, our study used edge betweennness and the igraph package command, infomap. There are numerous more methods that can be used to detect and analyze communities in networks, so it would be interesting to see what some of the other types of community detection might produce. This is a very dense network which includes weighted edges. There are existing community detection methods that would be suited to doing calculations for this network and take these two attributes into account.

5.2 Final Notes

This study was done to better understand what social network measures are most telling of a online forum classroom dynamic and furthermore, understand how measures compare and contrast when the construction of the network models is altered. There was an initial hypothesis about how the connectivity of the networks would compare with each other and how this would pan out with the network measures calculated about the data. In some cases, the hypotheses were completely correct, for example, average vertex-to-vertex distance followed the exact trend predicted. As the network models became increasingly more connected, the average vertex-to-vertex distance values decreased. The results showed that centralization may be more descriptive of how social networks are affected by model construction and connectivity than what was initially thought. Average degree followed the expected trend. There were higher average degree values as connectivity increased. The average Barrat coefficient and transitivity did not seem to follow any particular trend pertaining to connectivity. Finally, degree correlation coefficients also did not seem to be extremely telling of any relationship for social networks, model construction or connectivity.

Although the numbers in this study may not be statistically significant because of the relatively small sample size, it gave good indications about measures that may be important in network analysis. For example, this research helps to indicate that centralization is something worth more

investigation. Centralization is a measure that had no preconceived expectation, but showed a general increase across connectivity, thus may be a measure of interest in the future. On the other hand, community detection methods and correlation coefficients show little to no usefulness for analysis. The small steps here are a good start in the direction of filling the holes of existing literature and optimizing network analysis to best understand the interaction of groups and the individual influence within them.

Looking at the models used, it is arguable that the model 2 construction may be the most preferred model in this study. Having every node in a thread connect to one another as in model 3, may have too many connections to be too telling of the network, where as model 2 may skip a lot of important connections with how it is set up. Model 3 is a solid medium ground for this to capture important ties, without creating an indistinguishable mess.

Bibliography

[1] Reuven Aviv, Zippy Erlich, Gilad Ravid, and Aviva Geva. Network Analysis Of Knowledge Construction In Asynchronous Learning Networks. *Journal of Asynchronous Learning Networks*, 7, July 2010. doi: 10.24059/olj.v7i3.1842.

[2] Adrienne L. Traxler, A. Gavrin, and Rebecca Lindell. Networks identify productive forum discussions. *Phys. Rev. Phys. Educ. Res.*, 14(2):020107, September 2018. URL https://www.compadre.org/per/items/detail.cfm?ID=15264. 10.1103/PhysRevPhysEducRes.14.020107.

[3] Oleksandra Poquet, Liubov Tupikina, and Marc Santolini. Are Forum Networks Social Networks? A Methodological Perspective. In *10th International Conference on Learning Analytics and Knowledge(LAK '20)*. ACM, December 2019. doi: 10.1145/3375462.3375531.

[4] Hichang Cho, Geri Gay, Barry Davidson, and Anthony Ingraffea. Social networks, communication styles, and learning performance in a CSCL community. *Computers & Education*, 49 (2):309–329, September 2007. ISSN 0360-1315. doi: 10.1016/j.compedu.2005.07.003. URL https://www.sciencedirect.com/science/article/pii/S0360131505001272.

[5] Shane Dawson. A study of the relationship between student social networks and sense of community. *Educational Technology and Society*, 11:224–238, July 2008.

[6] Eric Brewe, Laird Kramer, and Vashti Sawtelle. Investigating student communities with network analysis of interactions in a physics learning center. *Physical Review Special Topics - Physics Education Research*, 8(1):010101, January 2012. doi: 10.1103/PhysRevSTPER.8.010101. URL https://link.aps.org/doi/10.1103/PhysRevSTPER.8.010101. Publisher: American Physical Society.

[7] Karina L. Cela, Miguel Ángel Sicilia, and Salvador Sánchez. Social Network Analysis in E-Learning Environments: A Preliminary Systematic Review. *Educational Psychology Review*, 27(1):219–246, March 2015. ISSN 1573-336X. doi: 10.1007/s10648-014-9276-0. URL https://doi.org/10.1007/s10648-014-9276-0.

[8] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, January 1978. ISSN 0378-8733. doi: 10.1016/0378-8733(78)90021-7. URL https://www.sciencedirect.com/science/article/pii/0378873378900217.

[9] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, March 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0400087101. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0400087101.

[10] M. E. J. Newman. The structure and function of complex networks. *Society for Industrial and Applied Mathematics*, 45(2):167–256, March 2003. doi: 10.1137/S003614450342480. URL https://arxiv.org/abs/cond-mat/0303516v1.

[11] Christina Prell. *Social Network Analysis: History, Theory& Methodology*. Thousand Oaks, Calif. ; London : SAGE, 2012. ISBN 978-1-4129-4714-5.

[12] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL https://igraph.org.

[13] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2021. URL http://www.rstudio.com/.