Analysis of an Online Learning Platform Interaction Network

With Structural and Temporal Considerations

Emily Jensen
CSCI 5352
emily.jensen@colorado.edu

ABSTRACT

In this project, we designed and analyzed an interaction network for an online algebra learning platform (Algebra Nation). To represent interactions with course material, we defined a bipartite network where nodes representing students connect to nodes representing videos that teach specific concepts. We analyzed this network through the lens of two main research questions. First, what are the basic structural properties of this network? We answered this using the full bipartite network as well as individual 1-mode projections of the network. Second, how does this network change over time? We answered this by looking at interaction patterns month-by-month over the course of an academic semester. Based on these analyses, we learned that most students do not take full advantage of the possible resources offered through the platform. We also saw that students increase the amount of content they access over the course of the semester, peaking in the month before a high-stakes end-of-course assessment. Additionally, we reflected on some practical lessons learned from the development process for this project.

1 INTRODUCTION

We begin this report by describing the proposed problem we aimed to study as well as some background information on the data set used for this project.

1.1 Background

Students are spending an increasing amount of time using online or other digital learning environments. Traditional students in high school or college courses have access to online (and often interactive) textbooks, recorded lectures, and a plethora of multimedia supplementary material. Students outside of the traditional school system can easily learn new skills through Massive Open Online Courses (MOOCs) such as Coursera.

The rise of these online and digital learning environments provides exciting opportunities as well as challenges for both teachers and learners. In these learning environments, students have the flexibility to direct their own path through the course material and have more autonomy over their learning process than in traditional classroom settings. However, these online courses are often not designed with this self-directed flexibility in mind. This requires constant monitoring of students to assess what their current learning needs are, and adjust their study plan accordingly [7]. Many courses continue to promote a linear path from concept and concept that mimic traditional classes. In addition, these new learning environments often have thousands of students enrolled at any one time, which means that instructors cannot spend much time assisting individual students.

For this particular project, we focus our analysis on Algebra Nation (algebranation.com), an online learning platform for middle school and high school mathematics. Algebra Nation is provided to students who are enrolled in Algebra 1, Algebra 2, or Geometry in their local school. Algebra Nation is available in seven states across the United States.

Students can interact with instructional content in Algebra Nation through a variety of mechanisms. The primary form of learning in Algebra Nation is through instructional videos (Figure 1). Videos are organized according to state standards. In the state of Florida's Algebra 1 course (our focus for this project), there are 10 general sections ranging from simple expressions to two-variable statistics. Each section contains a selection of specific topic videos (e.g., using the distributive property, properties of exponents). For each topic, students can choose a video presented by one of several professional study experts, who vary by personality and level of detail. In addition to videos, students can apply their knowledge through practice quiz questions. If they are having difficulty on a specific concept or problem, students can post on a discussion wall and get responses from other students or hired instructors.

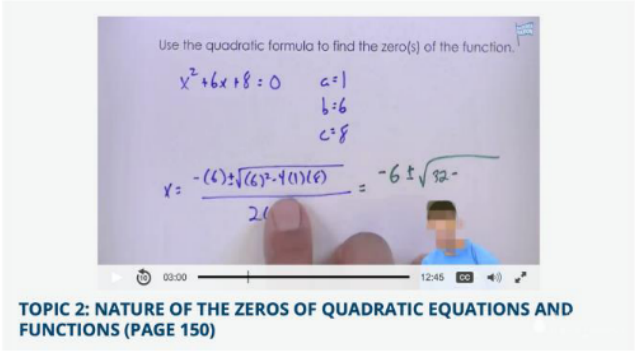

Figure 1: An example of a specific topic video in Algebra Nation. Videos are taught by professional study experts. Students can choose between several videos for each topic that differ by level of detail and instructor personality.

Algebra Nation is utilized in a variety of contexts. Some students rarely interact with the platform as a part of their normal class; any interaction is purely self-directed. In other cases, teachers integrate Algebra Nation directly into their regular lesson plan; they may have students watch videos in class and assign practice problems on the platform for students to complete for homework. This suggests that the context of use can greatly impact how students interact with Algebra Nation learning content over the course of a semester.

Student interaction with Algebra Nation is also possibly impacted by the Florida state end-of-course assessment. This assessment is administered for Algebra 1 classes at the end of the school year and is high-stakes because students must pass this exam in order to graduate from high school. We anticipate this will also impact student interaction patterns. For example, students may increase their activity on the platform as the date of the assessment approaches due to last-minute cramming.

1.2 This Project

This project aimed to repurpose and restructure an existing data set and assess the feasibility of analyzing it through the lens of network science. Specifically, analyzing an interaction network provides a different perspective on how students utilize the Algebra Nation learning platform.

Algebra Nation has been used as a research test-bed for several years through the Virtual Learning Lab group (virtuallearninglab.org) with a focus on providing a personalized learning experience for students. Previous work using the platform has focused on predicting the presence of a variety of student emotions [3] as well as analyzing these models' generalizability to new populations [4] and predicting key outcomes (under review). Additionally, some preliminary work has been done to investigate how student and teacher usage on the platform is associated with success on the end-of-course exam [8]. While all of this work has taken into account student activity and interaction with the Algebra Nation platform, none have considered approaching this data from a network science perspective. Specifically, most of the previous research on the platform has focused on either immediate actions for in-the-moment emotion prediction [3, 4] or simplified counts of general activities over the course of a school year [8].

For this project, we hoped to supplement this previous work by analyzing student activity on Algebra Nation through the perspective of a bipartite interaction network. We did this through several steps. First, we considered the practical considerations when designing and actually building such a network. Second, we analyzed some basic structural properties of this interaction network to see what interesting patterns arise. Next, we investigated how this network changes over the course of an academic semester. Finally, we discussed the main lessons learned, limitations of our approach, and some future directions for this work.

2 DEFINING THE NETWORK

2.1 Research Questions

In this project, we considered two overarching research questions.

The first main research question was: what is the structure of the bipartite interaction network? We were interested in this question because previous work simply has not considered the available Algebra Nation data set from a network science perspective. Within this overall question, we aimed to look at how the network is structured both at a bipartite level and as individual 1-mode projection networks. We could do this by considering patterns of connectivity between nodes in the network. This question could also provide insight into how actual student use of the Algebra Nation platform compares to how the creators intended for the platform to be used.

The second main research question was: how does this network change over time? To do this, we considered how students interact with content over the course of a semester. In addition to the previous question, this could give us more fine-grained insight into how students interact with the Algebra Nation platform. One hypothesis was that students increase their interaction over the course of the semester as they are motivated to prepare (cram) for the end-of-course assessment.

2.2 Design Decisions

The first decision when designing the Algebra Nation interaction network was how to define the nodes of the network. This required answering two questions: *who* is interacting with the Algebra Nation platform and *what* are they interacting with? This suggested a bipartite network composed of two types of nodes.

We considered students to be the main audience that is interacting with Algebra Nation. While teachers do have access to Algebra Nation, their set of possible interactions has a different scope than students; for example, teachers primarily perform administrative tasks such as assigning homework, rather than learning-based activities that students perform. We represented students in the platform by creating a node indexed by their unique student ID number.

Students learn on the Algebra Nation platform by interacting with lesson content. At an abstract level, we would like the content nodes in our network to represent individual learning concepts. As discussed above, these can be represented in several forms such as videos and practice problems. This makes representing learning content in the interaction network complicated. One reason is that different forms of content can address the same underlying concept. For example, a quiz on section 2 may have multiple questions assessing the same topic. In addition, quiz questions and videos cover concepts in varying levels of granularity. Including every possible type of content in our network would then be redundant and increase the complexity of our network. Therefore, we decided to define content nodes in our network based only on videos. We believed this represents an appropriate level of granularity since each video addresses a general concept (e.g., adding fractions with the same denominator) within a overall section (e.g., arithmetic with fractions). We then represented concepts in the platform by creating a node indexed by its individual video ID number.

The next decision for designing the Algebra Nation interaction network was how to define the edges between nodes. Since we were designing a bipartite network, edges should represent an interaction between a student and a piece of content. Based on the above definitions of our nodes, we decided to add an edge to the network when a student plays a video. We chose to use undirected edges since the obvious directed configuration in the bipartite structure (pointing from student to video) does not give us any new information. We also chose to maintain a simple network by not allowing repeated edges between nodes (if a student watches a video multiple times). By consequence, this created a network that represents the first instance that a student accesses each particular video.

We also made several design considerations in an attempt to improve the simplicity and overall size complexity of our model. First, we limited our generated data set to the state of Florida. Algebra

Nation is currently deployed in 7 states and is used by hundreds of thousands of students each semester. To further complicate the matter, each state uses slightly different course content in order to align with individual state standards. By limiting our data set to one state, we simplified the possible course content students can access and also found a convenient subset of the overall student population. We also chose to consider only students enrolled in Algebra 1 to limit the possible course content. Additionally, we used data limited to the Spring 2019 semester, defined as sessions that took place between January 1, 2019 and June 30, 2019. This choice of dates covered the vast majority of semester start and end times between different schools in the state.

3 DATA AND METHODS

In this section, we discuss the actual implementation of the network, keeping in mind the design considerations discussed above. We discuss collecting the data, implementing the network, and analysis of the network. The code discussed in the sections below can be found in the repository linked in Section 6.

3.1 Data Collection

The first major task of the project was to collect data in a form that can be implemented into the network structure described above. Data describing student use of Algebra Nation is maintained by Study Edge, the company which developed the platform. We have access to this data in the form of a database which can be queried. We gathered data on student when students accessed specific videos by developing and using a SQL script to query this database (see `video-data-query.sql` for details).

In order to group and analyze the data from different perspectives, we selected several different variables from the database. The key variables were the session ID number which indicates a unique log-in session, the time stamp which indicates the start of the session, the ID for the student that logged in for the session, and the ID of the video that the student played. We limited our query to only consider when a student begins a particular video; it does not matter to us if they finished it. As discussed above, we only considered students in Florida that were enrolled in Algebra 1. We also included data on the name of the video as well as the section and topic number. We did not end up using these variables in our analysis as much of this information was missing for individual videos.

Our query only selected distinct combinations of the variables described above. This was done in an attempt to simplify the possible data set. This limitation means that we only have records of the first time that a student played a particular video in a particular log-in session. For example, if a student logs in to Algebra Nation on Monday afternoon and plays video 1, takes a quiz, and plays video 1 again, we only have a record of the first play of video 1. However, if the student plays video 1, takes a quiz, and plays video 2, we would have a record of both videos. Similarly, if a student logs in on Monday afternoon and plays video 1, logs out, logs in on Tuesday morning and plays video 1 again, we would have a record of both plays since it is a new log-in session.

Due to the size of the available data, we pulled data from the database in 2-week intervals and exported the data to CSV files. This yielded twelve 2-week data sets which we later combined to contain all data for the entire semester.

After compiling this data, we explored the distribution of students, videos, and log-in sessions. For the Spring 2019 semester, there were 594,227 log-in sessions, 101,028 unique students, and 1,568 unique videos. While there are 93 individual topics covered by Algebra Nation, each topic contains videos from the perspective of several different study experts. In addition, there are additional videos for on-boarding and bonus content on the platform, which we did not exclude from our data collection. Overall, we gathered 1,287,278 instances of a student playing a unique video for a particular log-in session. The mean number of entries for a log-in session was 2.17 (SD 2.37, median 1) which means a student watches on average around 2 unique videos per session. This makes sense if a student logs in to Algebra Nation for focused practice on a concept. The mean number of entries for a student was 12.74 (SD 16.42, median 6) which means that watches around 13 unique videos over the course of the semester. This means that most students are not accessing the majority of the available content on Algebra Nation. Lastly, the mean number of entries for a video was 820.97 (SD 1638.63, median 22). This indicates that there is a wide spread of how many times each video was accessed. This could be due to the purpose of a particular video (e.g., the on-boarding video is watched by everyone, but not the bonus material).

3.2 Implementing the Network

As discussed in Section 2.2, we designed the network to have a bipartite structure, with one set of nodes corresponding to students and the other set of nodes corresponding to course content (videos). Edges in this network correspond to when a student accesses a particular video.

The implementation and analyses of the Algebra Nation interaction network were done using Python and related packages. First, we implemented the interaction network using the NetworkX package [1]. This provided convenient data structures and pre-built functions for structural analysis. Data wrangling and manipulations were done using the NumPy [9] and Pandas [6] packages. Finally, basic data visualizations were done using the Matplotlib [2] package. The code for generating the results can be found in the `video-processing.py` file.

To build the interaction network, we implemented a simple network with one node for each unique student ID number and one node for each unique video ID number in our collected data set. We labeled the nodes with a bipartite property of 0 for student nodes and 1 for video nodes to distinguish between the two types. We then added an edge between the appropriate nodes when a student accessed a particular video (not allowing for multiple edges). This network now represents which videos each student has watched at least once.

3.3 Structural Analyses

This section addresses the first main research question, which asked about the structural properties of our bipartite interaction graph. We considered this question in two cases. First, we looked at the original bipartite graph. Second, we considered the 1-mode projections of the original graph.

3.3.1 Bipartite Graph. We first considered the structure of the original bipartite graph consisting of both students and videos in Algebra Nation. To begin, we calculated the degree for each student node. The degree of a student node corresponds to the number of unique videos they watched over the course of the semester (not the total number of video views). Figure 2 shows the degree distribution of student nodes. The mean degree for student nodes is 11.22 (SD 14.02, median 6). This is similar to, but less than the results of the analysis in Section 3.1. This makes sense because the previous analysis includes videos that a student watched multiple times, but in different log-in sessions. However, the fact that the mean number of total videos watched is similar to the mean number of unique videos watched means that students do not generally go back and re-watch videos between log-in sessions.

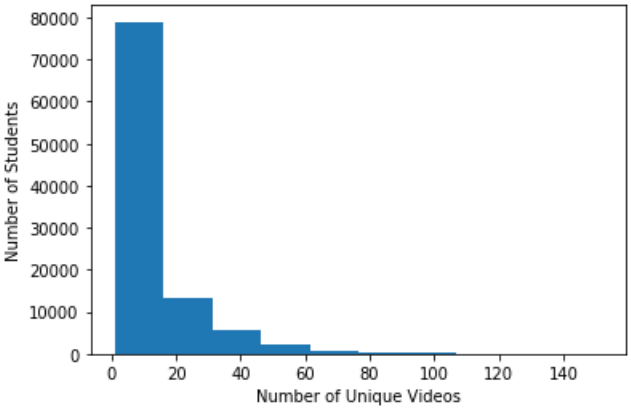

Figure 2: The distribution of degrees for the student nodes. This indicates the number of unique videos viewed over the course of the semester.

We then calculated the degree for each video node. The degree of a video node corresponds to the total number of students that watched the video at least once (not the total number of times the video was watched). Figure 3 shows the degree distribution of the video nodes. The mean degree for video nodes is 723.19 (SD 1,460.39, median 21). Like the discussion of student node degrees above, this supports the results in Section 3.1 where there is a large spread of how many students watch each video, and a few students watch a video between multiple sessions.

Besides this, we also calculated the degree centrality for each of the nodes in the interaction network. The degree centrality of a node is the fraction of nodes to which it is connected. The mean degree centrality of the bipartite interaction network was 0.0002 (SD 0.002, median 0.00006). This result highlights the sparsity of the interaction graph. Interestingly, the maximum degree centrality for any one node in the projection network was 0.09, which shows that there are some highly connected nodes within the network.

Additionally, we calculated the average bipartite clustering coefficient for the interaction network. The average bipartite clustering coefficient is defined in [5] as the overlap of neighborhoods between nodes; this is an extension of the definition for 1-mode networks counting the number of triangles, which do not occur in bipartite networks. The average clustering coefficient for the network

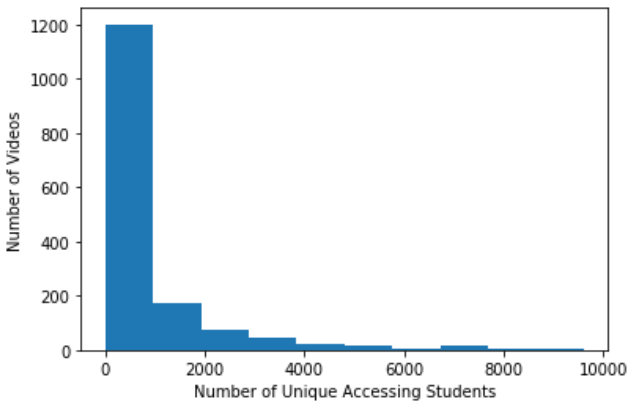

Figure 3: The distribution of degrees for the video nodes. This indicates the total number of times a video was viewed over the course of the semester.

was 0.11, which is higher than the examples given in [5] for the Actors-Movies and Authoring data sets.

Finally, we calculated the density of the bipartite network, which was 0.007. The density of a network is defined as the proportion of actual edges compared to the total number of possible edges in the graph. For example, if each student watched each possible video on Algebra Nation, then the density of this network would be 1. This definition suggests that our Algebra Nation interaction network is extremely sparse. This supports our earlier analyses that most students are not fully exploring the selection of possible content that Algebra Nation offers.

Other potential structural measures of networks include reciprocity and homophily/assortativity. As discussed in 2.2, we chose not to implement directional edges in the network because they did not add any meaningful information. Therefore, the notion of reciprocity does not make sense in the context of our interaction network. Assortativity would have been a very interesting area to investigate; for example, do students with similar demographic traits or in the same school tend to watch the same videos? However, we were not able to obtain the sort of personal metadata that would make this analysis possible. It would be interesting to pursue in future work. We also attempted to calculate other measures of centrality, but were limited by processing power.

3.3.2 1-mode Projections. After considering the entire bipartite interaction network, we considered the individual 1-mode projections. First, we considered the projection for the video nodes. In this context, two video nodes are connected if they have both been viewed by the same student (regardless of the time that they were viewed). We then calculated the video node degrees in this projection. In this case, the degree for a video is the number of other videos that have also been viewed by a student. Figure 4 shows the degree distribution of the projected video network. The mean degree of this network was 345.39 (SD 315.92, median 205). These results show that there are many videos have few neighbors (watched by a few or only one student). Compared to the analysis of the full bipartite network, we see a small increase around degree 600, which shows

that there are some videos that are especially popular and watched by many students. The density of the video projection network was 0.22, which is less sparse than the entire bipartite network. This is not surprising since most students watch more than one video over the course of the entire semester, so connections between videos in the projection should be relatively common.

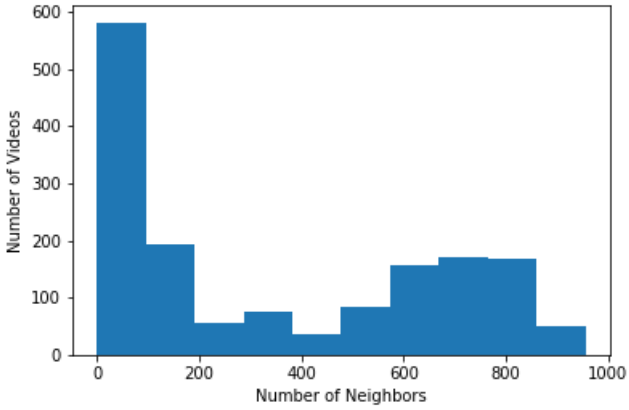

Figure 4: The distribution of degrees for the 1-mode video projection network. This indicates which videos share student viewers.

Besides this, we also calculated several centrality measures of the video projection network. We first calculated the degree centrality of the projected network. Similar to the analysis above, degree centrality indicates how videos share viewers. The mean degree centrality of the video projection network was 0.22 (SD 0.20, median 0.13). While this is still a rather sparse network, the video projection network is much more connected than the full bipartite interaction network. This is not surprising as most students watch more than one video, so many videos will be connected in the projected network. The distribution of node degree centrality follows the same shape as Figure 4, but with normalized values.

Additionally, we calculated the eigenvector centrality of the video projection network, which considers the number of paths that go through a particular node. Generally, a high eigenvector score corresponds to that node being connected to other high-scoring nodes. The mean eigenvector centrality of this network was 0.02 (SD 0.02, median 0.01). Since the eigenvector centrality scores for this network are low, we anticipate that we would not see large clusters of densely connected nodes. The distribution of eigenvector centrality scores is found in Figure 5. This distribution is interesting because the centrality measures tend to cluster around the low and high end of the range. This suggests there may be groups of distinct popular and unpopular videos.

We then calculated the closeness centrality of the projected network, which considers the average path length from a node to the other nodes in the network. The mean closeness centrality of this network was 0.52 (SD 0.12, median 0.51). Compared to other centrality measures, nodes in this network have less spread and have a generally higher centrality value. The distribution of closeness centrality scores is found in Figure 6. This distribution is notable

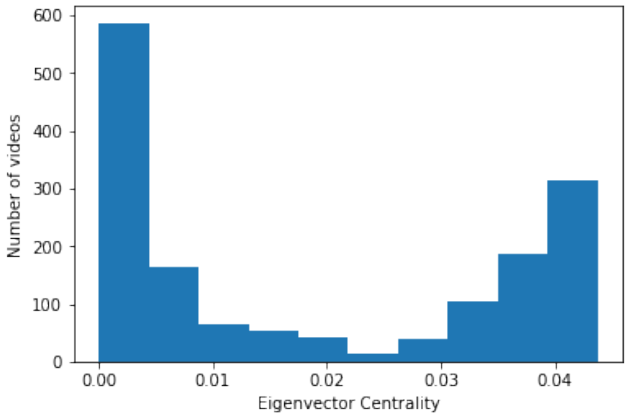

Figure 5: The distribution of eigenvector centrality values for the 1-mode video projection network.

because there are a few nodes with a centrality score of 0 while the rest of the nodes have relatively high closeness values. Clearly, the nodes with a score of 0 are the nodes that are disconnected from the rest of the graph. However, the rest of the nodes seem relatively close together. This suggests that there is not a pronounced core-periphery structure to the graph, excluding the disconnected nodes. One potential cause of this could be because it is relatively easy to form an edge between two nodes in the projection network; two videos need to share only one student viewer in order to be connected.

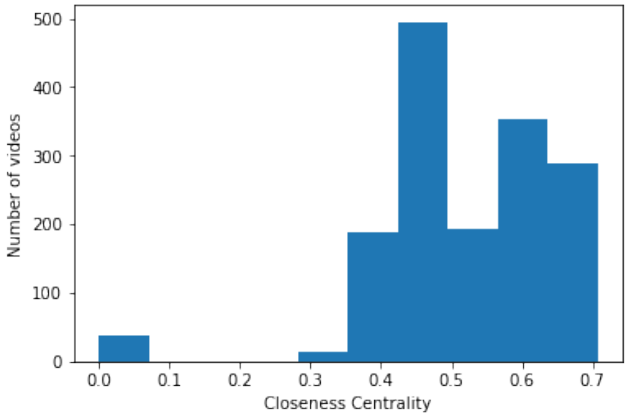

Figure 6: The distribution of closeness centrality values for the 1-mode video projection network.

Next, we calculated the betweenness centrality of the projected network, which finds the number of geodesic paths that go through each node in the network. The mean betweenness centrality for this network was 0.0005 (SD 0.001, median 0.00008). The distribution of betweenness centrality scores is found in Figure 7. We can see that this distribution is severely skewed in that most centrality scores are 0. Considering the results of closeness centrality discussed above,

this is not surprising. If most of the nodes are in fact highly and evenly connected with each other and therefore "close" together, it is unlikely that there will be connective "bridges" between clusters of nodes that would be measured by the betweenness centrality metric.

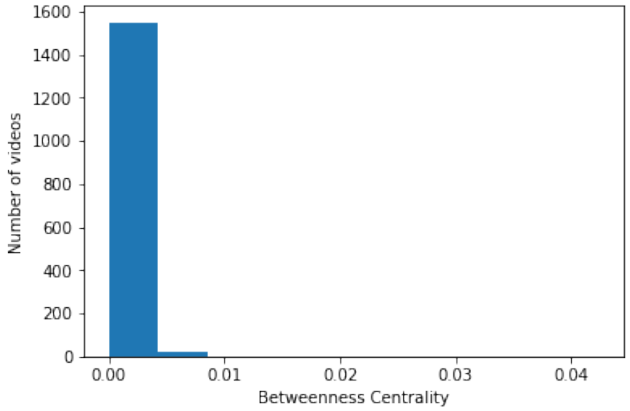

Figure 7: The distribution of betweenness centrality values for the 1-mode video projection network.

Besides centrality measures, we also calculated the clustering coefficient for each node in the video projection network. For a particular node, the clustering coefficient is the fraction of possible triangles that go through a node that actually exist. There were a total of 43,083,048 triangles in the projected network. The mean clustering coefficient was 0.82 (SD 0.19, median 0.86). The distribution of clustering coefficients is found in Figure 8. We see an interesting distribution similar to the closeness centrality metric discussed above. Here, we again see a few nodes with a coefficient of 0; these are the nodes that are not connected to the rest of the projection network. For the rest of the nodes, we see generally high clustering coefficients. In fact, about one third of the videos have a clustering coefficient of 1. This again speaks to the highly connected nature of the projected graph.

Finally, we tried to calculate the diameter of the video projection network, which indicates the longest geodesic path in the network. In other words, the diameter measures how far apart the two most distant nodes in the network are. We were ultimately not able to calculate the diameter of this network because the graph is not connected; there were several nodes that were not connected to any others. This is interesting because it indicates there were some videos that were viewed by exactly one person.

Our original intention was to repeat these analyses for the student projection network as well. However, we encountered computer memory and processing limitations when attempting to construct the projection. This is likely due to the fact that our data set contains 101,028 students compared to 1,568 videos. We attempted to remedy this problem by randomly sampling a subset of students and constructing a projection using those students. However, in order to reach a comparable number of nodes as the video network, we would be sampling roughly 10% of the original network. We felt

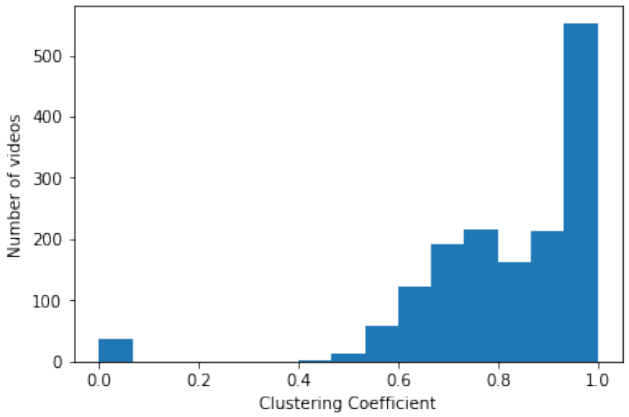

Figure 8: The distribution of clustering coefficient values for the 1-mode video projection network.

that sampling at such a low rate would not be informative of the structure of the full projection network.

3.4 Network Changes Over Time

This section addresses the second main research question, which asked how the interaction network changes over time. We considered these changes on the time scale of months over the course of the semester, which yields six time slices of the interaction network (January through June).

First, we considered how the overall number of video views changed per month. To calculate this, we counted the number of entries in our data set that occurred in each month. Unlike the simple interaction network we built in the previous section, this count includes when a student watches a video in more than one session. Figure 10 shows how the number of video views changes over the semester. We see that there is a slight increase in video views between January and March and a sharp increase in the month of April. After April, the number of video views sharply declines.

One possible explanation for this pattern is that the end-of-course assessment is administered in May. Since students tend to wait and study right before the test, it makes sense that the highest video activity happens in April, the month before the end-of-course assessment. This change might also be explained by a difference in the context that Algebra Nation is used. For example, it is possible that the primary use of Algebra Nation in the beginning of the semester (January through March) is in regular classrooms where teachers require students to use the platform more frequently; this could cause the generally stable amount of video views between these months. In April, as the end-of-course assessment approaches, we may be seeing an increase in use of Algebra Nation as other students are motivated to use the platform outside of regular class time. After the assessment, use of the platform drops dramatically as students feel there is no more need to study.

On a related note, we also considered the average number of videos each student viewed per month. To do this, we counted the number of entries in our data set that occurred per month for each

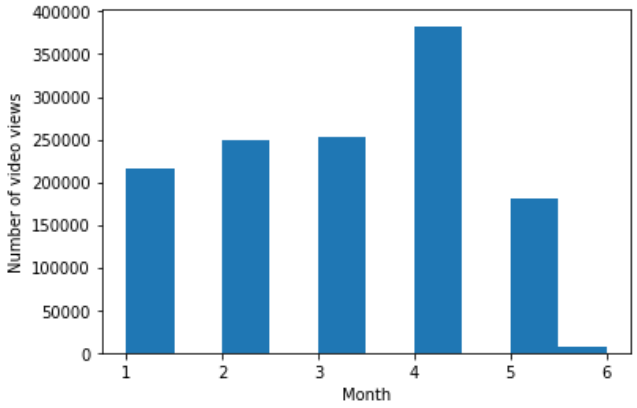

Figure 9: The number of total video views per month over the course of the semester. This can indicate the change in overall activity on the platform for each month.

student. This indicates the total number of videos a student watches per month (ignoring repeated videos within a single session). We then averaged across students to get the average number of views per student per month. Figure 10 displays these results. The pattern of activity is similar to those discussed above. We see a slight increase in video views per student between January and March. After that, there is a large increase for the month of April (right before the end-of-course assessment) followed by a decrease in activity through the end of the semester.

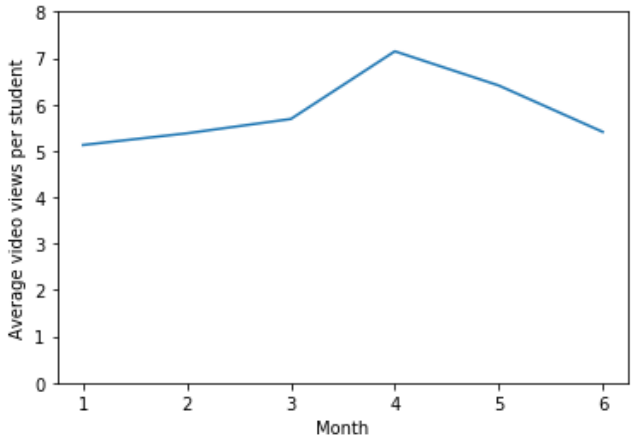

Figure 10: A timeline of the average number of video views per student for each month in the semester.

Additionally, we calculated the average month in which each video was viewed. To do this, we averaged over the month number for each time a video appears in the data set. Figure 11 shows this distribution. First, it is interesting to note that there are quite a few videos that fall on the extreme ends of the time range. For example, there are around 70 videos that are viewed exclusively in January and about 80 videos that are viewed exclusively in April. This shows that there are some topics that are almost always taught at the beginning and the end of the semester. Besides this, there is a large group of videos that fall in the middle of the semester, around March. This is likely caused by the varying teaching timelines between different schools. For example, one teacher may have students watch a video on topic 7 closer to the beginning of the semester while another teacher may teach topic 7 near the end of the semester once students have learned other concepts. This would result in an average month towards the middle of the semester.

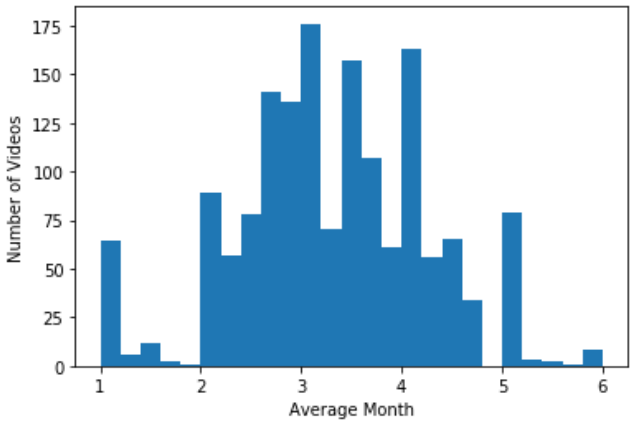

Figure 11: The average month that each video was viewed.

Finally, we looked at the time when videos were first viewed. To do this, we found the month corresponding to the first entry of each video in the data set. Figure 12 shows this distribution. It is interesting to note that the vast majority of videos are first viewed in January. This gives us a more nuanced understanding of the distribution found in Figure 11. If students were watching videos directly in time with when they were learning a particular topic in class, we would expect a much more uniform distribution of first-views over the semester. That is, more videos should receive their first view later in the semester because they are not needed until other concepts are first mastered. This peak of first views in January could also be caused by students exploring the platform at the beginning of the semester when they first get access to the Algebra Nation platform.

4 DISCUSSION

This project provided a first attempt at analyzing interactions on an online learning platform through a network science perspective. Previous work using this data has focused on student actions in a specific time window; we have expanded this scope to include an entire semester. We approached this problem with two main research questions. First, what are the structural properties of a bipartite network defined with nodes representing students and learning content (instructional videos)? We found that such an interaction network is actually sparse in the fact that students generally do not access most of the available course content. We also analyzed the projection of the network onto the video nodes. From this, we learned that there is a wide disparity in how much individual videos are accessed compared to others.

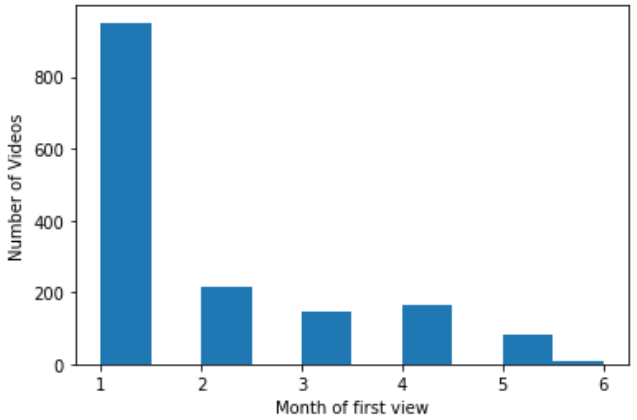

Figure 12: The month that each video was first viewed.

Our second research question asked how does this network change over time? We found a general trend that access to videos increases over the semester and peaks right before the end-of-course assessment. This is consistent between the total number of viewed videos per month as well as the average number of video views per student each month. Lastly, we found some evidence that videos are accessed in time with when students are learning content in class; while some videos were exclusively viewed at the beginning or the end of the semester, the majority of videos were first accessed at the beginning of the semester, perhaps during an exploratory period.

This project was largely limited by sheer size of the data set. Even when only considering students in one course in one state over one semester, the resulting data set was often unwieldy and did not lend itself well to even basic algorithms such as projections or calculating centrality measures. Therefore, we were not able to do the same analyses for the student data as we did for the video data. There are several possible options for alleviating this problem in the future. First, we could limit the scope of our data set even further than it currently is. For example, we could limit location by only selecting students from a single school district. Alternatively, we could limit the time duration by focusing on a single month's worth of data. By using a smaller data set, we could also possibly do more detailed analysis of the data by using more interaction features besides watching a video. Rather than limiting the scope of our data set, it would also be interesting to experiment with different sampling methods. This approach would be preferable because we could then maintain the richness and dynamics of our data without being burdened by an overly large data set.

Relatedly, our project was limited by how we operationalized the problem and designed the data set. In an effort to simplify the available data, we defined student interaction with learning materials only in the context of watching videos. This does not account for learning that happens through practice (taking quizzes) or sharing knowledge in social situations (posting on the discussion wall). Future work may choose to include these types of learning content and build an interaction network with more than two types of nodes; one could perhaps even imagine a hierarchical scheme for defining course content.

There are several directions that would be exciting to pursue for future work. First, it would be interesting to augment the above analyses with added metadata about the students. This could include demographic information, achievement scores, and school information. If we had access to this data, we could investigate some community-detection algorithms. It would be interesting to see if we could detect the context in which students use the Algebra Nation platform; for example, are they only accessing content because it is assigned by their teacher, or are they intrinsically motivated to do work outside of school? This would also warrant a closer analysis of the time stamps associated with each viewed video. For this project, or level of granularity when considering changes over time was one month. We could also consider more fine-grained analysis such as considering the time of day (hour) or day of the week that a student is accessing a video. With more robust data cleaning, we could match section and topic numbers to each video and investigate patterns of activity related to different levels of content grouping.

In addition, we could supplement the analyses above by including more metrics that are specific to bipartite networks. In this project, we focused on measures of centrality as well as measures such as density and clustering that were discussed in class. We could extend this analysis to new measures that have not been discussed in class. Several of these are introduced in [5]. For example, in [5], the authors discuss the concept of redundancy, which considers how the projection of a network might change if one of the nodes were removed. It would be interesting to apply these new concepts to our bipartite Algebra Nation interaction network to learn about the unique structure given by framing this data set in terms of such a network.

5 CONCLUSION

In this project, we investigated the interaction network for Algebra Nation, an online mathematics learning platform. We defined this interaction network with a bipartite structure, with one set of nodes corresponding to students and the other set of nodes corresponding to learning content (instructional videos). After collecting and wrangling the data set, we analyzed basic structural properties of the network as well as some temporal changes over the course of a semester. Along the way, we learned some practical lessons on designing such a network, such as the fact that more data is not always better; in fact, it can make your computing tasks extremely time and resource intensive! Future work can incorporate more sophisticated analyses as well as incorporate concepts of community detection. Overall, we are confident that there are many more insights to be gained from analyzing this data from a network science perspective.

6 LINK TO PROJECT REPOSITORY

The code for this project can be found on GitHub. The repository also includes the figures added to this paper and a copy of this report. Due to the sensitive nature of the data and the fact that students are minors, we are not able to publicly share our data. (https://github.com/emilykjensen/CSCI5352_Project)

REFERENCES

[1] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). Pasadena, CA USA, 11 – 15.

[2] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[3] Stephen Hutt, Joseph F. Grafsgaard, and Sidney K. D'Mello. 2019. Time to Scale : Generalizable Affect Detection for Tens of Thousands of Students across an Entire School year. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*. ACM Press. https://doi.org/10.1145/3290607.3300726

[4] Emily Jensen, Stephen Hutt, and Sidney K. D'Mello. 2019. Generalizability of Sensor-Free Affect Detection Models in a Longitudinal Dataset of Tens of Thousands of Students. In *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, and Roger Nkambou (Eds.). 324–329.

[5] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. 2008. Basic notions for the analysis of large two-mode networks. *Social Networks* 30, 1 (2008), 31–48. https://doi.org/10.1016/j.socnet.2007.04.006

[6] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 51 – 56.

[7] Daniel C. Moos. 2014. Setting the stage for the metacognition during hypermedia learning: What motivation constructs matter? *Computers and Education* 70 (2014), 128–137. https://doi.org/10.1016/j.compedu.2013.08.014

[8] Sahba Akhavan Niaki, Clint P. George, George Michailidis, and Carole R. Beal. 2019. Investigating the Usage Patterns of Algebra Nation Tutoring Platform. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*. ACM Press, New York, New York, USA, 481–490. https://doi.org/10.1145/3303772.3303788

[9] Travis Oliphant. 2006–. NumPy: A guide to NumPy. USA: Trelgol Publishing. http://www.numpy.org/