Degrees of Arnn

Modeling the Social Interconnectedness of Hillsdale College

Hillsdale Applied Math Club

Fall 2022

Abstract

Early during the fall semester of 2022, Hillsdale student and Applied Math Club (AMC) member Daniel Brand made the assertion: "Hillsdale is such a small campus. I bet, if I don't know somebody on campus, I could go to one of my friends and they would know them." Mathematically, this can be expressed in terms of graph theory, where students are nodes and edges exist between students who know each other. Dan's assertion is equivalent to the statement: the shortest path between any two students is no greater than 2. The Hillsdale AMC performed two different surveys to test this claim. The first survey focused on how many other students one student knows (and in what capacity). The second survey focused on how well a network of students know each other. In the end, both surveys supported the claim that Hillsdale students, on average, know about 40% of campus. Furthermore, the surveys both suggested that Dan's assertion is very likely true, because it was true in the projections based off the first survey, and only 4% of student connections in the second survey required a greater path length than 2.

1 Introduction

1.1 Motivation

Early during the fall semester of 2022, Hillsdale student and Applied Math Club member Daniel Brand made the assertion: "Hillsdale is such a small campus. I bet, if I don't know somebody on campus, I could go to one of my friends and they would know them." The other club members thought this could be true, and began wondering how they could verify it. Among many ideas, one modern analogue came to mind. There is a well-known game called "Six Degrees of Kevin Bacon" that shares many similarities to the AMC's conjecture. Instead of connecting actors through movies to Kevin Bacon, the AMC decided to connect students through knowledge to each other. In the end, they named their idea after Hillsdale College President Dr. Larry Arnn, and the fall semester modeling project "Degrees of Arnn" was born. The goal: verify whether or not every student was within 2 "degrees" of every other student.

1.2 Results

The Hillsdale AMC conducted two different surveys to gather the relevant social information about students on campus to answer this question. In the first survey, we found that, on average, students recognized 39.2% of campus (although students really only "knew" 30.1% of campus). From this we generated an estimated network of students to match this statistic. In that network, we found that every student was, in fact, within 2 "degrees" of every other student. In our second survey, we found that, on average, students report knowing 40.9% of campus. Furthermore, we discovered that, on average, students knew 10.2% more people than they were known by. We examined the network of the second survey to find that, while most students were only 2 "degrees" from every other student, 4% of student connections were of "degree" 3. Therefore, we conclude that Dan's assertion is *very likely true*, and if it is not, then at least *the vast majority* of students are within 2 "degrees" of each other.

2 Interpretation of Problem

2.1 Graph Theory

We will translate this problem into graph theory. Each node will represent a student. Two nodes will have an edge between them if the students "know" each other. Since it is difficult to determine what classifies as "knowing," this will be dealt with on a case by case basis. In the first survey, we will make distinctions between different ways of students knowing each other; in the second survey, we will leave interpretation up to participants. Additionally, we will use directed graphs instead of undirected graphs because "knowing" does not necessarily go both ways. In theory, your "knowing" someone does not necessitate their knowing you (a celebrity, for example). To represent this, we will use directed graphs to allow for one-way edges. Also, the graph will often be represented as an adjacency matrix.

2.2 Degree of Arnn

In terms of graph theory, we now define what a "Degree of Arnn" is:

Definition 2.1 (Degree of Arnn)**.** The Degree of Arnn (hereafter DoA) between two students is the length of the shortest path between the two nodes that represent the two students.

Please note that the DoA, which only exists between two students, is entirely separate from the standard definition for the "degree" of a node (and, in fact, inversely proportional). To illustrate the DoA between two students in a graph, consider the following example. There are 4 students: Jack, Emily, Spencer, and Lydia. Jack knows Spencer and Emily, Emily knows Jack, Spencer knows Jack and Lydia, and Lydia knows Jack and Spencer. The directed graph of these students can be seen in Figure 1. The DoA between Jack and Spencer is

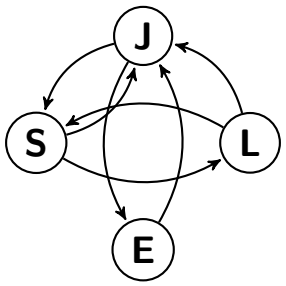

Figure 1: Directed Graph of 4 Students

1, because they know each other. The *DoA* between Jack and himself is 0, since he is himself. The *DoA* from Lydia to Emily is 2, since Lydia doesn't know Emily, but she knows Jack who knows Emily. In contrast, the *DoA* from Emily to Lydia is 3, since Emily knows Jack, who knows Spencer, who knows Lydia. The corresponding adjacency matrix looks like this:

$$\begin{array}{c} \\ J \\ E \\ S \\ L \end{array} \begin{array}{c} \begin{array}{cccc} J & E & S & L \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Note we do not mark any edges between a student and himself. To model the *DoA* between each student on campus, we first need to take a sample of students who know other students, turn it into a graph, and finally determine each *DoA*. This final step will be done using Floyd's algorithm to find the shortest path between nodes.

2.3 Floyd's Algorithm

Floyd's algorithm is a recursive algorithm for efficiently finding the shortest paths between nodes in a graph. It begins by recording paths between nodes that do not go through any intermediate nodes. That is, it records all length 1 nodes, regardless of cost (or weight). Then, it considers one intermediate node (say, node J) and updates the record of shortest paths by any of those through J that have less cost. That is, it considers all length 2 paths that go through J, updating the record of the path is less expensive. Then, it takes another intermediate node (say, node S) and updates the record of shortest paths by any of those through S that have less cost. That is, it now considers all length 3 paths that go through S (and also J by the previous step). For example, if a path that goes between two nodes passes through both J and S, then if it costs less, it would replace the previous shortest path recorded between those two nodes. This continues until all nodes are considered as intermediate nodes. The resulting path costs are stored dynamically and can be accessed later, making

Floyd's algorithm efficient for many large graphs.

We will perform Floyd's algorithm on an adjacency matrix representing the student network. This will generate a matrix displaying the DoA from each node to every other node. In order to implement Floyd's algorithm, the adjacency matrix will be slightly modified. Each 0 indicating that two students do not know each other will be changed to a sufficiently high number, such as 9999, to estimate *the infinite cost* of travelling from that node to the other (written here as ∞), while each 1 still represents the low and equal cost of travelling to another person by knowing that person. Then, Floyd's algorithm will transform the adjacency matrix into one with each entry showing the (column) student's DoA with each other (row) student. For example:

$$\begin{array}{c} \\ J \\ E \\ S \\ L \end{array} \begin{array}{c} \begin{array}{cccc} J & E & S & L \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & \infty \\ 1 & 0 & \infty & \infty \\ 1 & \infty & 0 & 1 \\ 1 & \infty & 1 & 0 \end{pmatrix} \end{array}$$

by Floyd's Algorithm becomes:

$$\begin{array}{c} \\ J \\ E \\ S \\ L \end{array} \begin{array}{c} \begin{array}{cccc} J & E & S & L \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 3 \\ 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix} \end{array}$$

Note that the matrix is not symmetric. This is a result of using an undirected graph, i.e., Lydia knows Jack but Jack does not know Lydia.

2.4 Evaluation of Dan's Conjecture

Dan's conjecture, then, can be expressed as the following falsifiable theorem:

Theorem 2.1 (Dan's Conjecture)**.** *The maximum DoA for any node in a graph corresponding to the students on campus at Hillsdale College is 2.*

The rest of this paper will evaluate how likely and to what extent this conjecture is true. It will not be formally *proven*, since that would require asking every student on campus about every other student. Instead, we will model smaller subsets of campus to determine how likely it is that Dan's conjecture is true. For instance, if the conjecture were for the previous example, it would be false since the DoA from Emily to Lydia is 3. To help with this estimation, we introduce a way of describing an individual student's interconnectedness to the rest of campus.

Definition 2.2 (Average Degree of Arnn)**.** The average DoA (denoted $\overline{DoA}$) of a graph is the arithmetic mean of the DoA for each ordered pair of students in the graph (excluding a student with himself).

If a graph has a low $\overline{DoA}$, it more likely to have a smaller maximum DoA. Thus, the $\overline{DoA}$ gives an estimate for how interconnected a given graph of students is. In the previous example,

$$\overline{DoA} = \frac{1+1+2+1+2+3+1+2+1+1+2+1}{12} = 1.5, \tag{1}$$

which shows that Dan's conjecture is more likely true, or at least *mostly true*, since it is true for a very large subgraph (even though it is not completely true in this particular case).

3 The First Survey

3.1 Survey Process

For the first survey, we set up a table in the student union with a sign that said, "Do you have friends? Prove it! How much of campus do you know?" For several hours during the peak of the day, we had students (some of whom we knew and asked to participate) voluntarily agree to take a survey. The survey was a Google form, consisting of fifteen questions. Each question required two answers. One answer was a name, with the prompt "Person X." The other was a prompt asking "Relation" with the multiple choice options being "By Face," "By Name," "Both," and "Neither." This can be seen in Figure 2. We

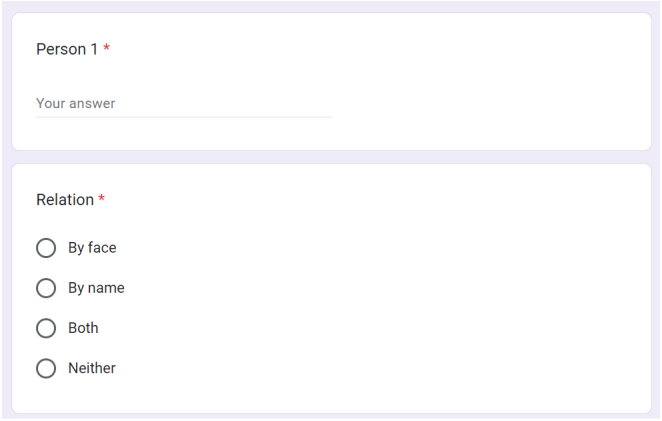

Figure 2: Participant's Prompt in the First Survey

showed each participant the name and picture of a random student currently enrolled and had the participant fill out a question. Once they completed fifteen questions, they submitted their survey. In practice, due to technical difficulties, there was variation in how students were randomly generated. This variation can be split into two main categories, as explained below.

3.2 Automatic vs Manual Generation of Students

Most of the time, random students were generated "automatically." The Hillsdale Coding Club wrote a program to generate a random student. The program opened a window that went to the Hillsdale student directory[1] and chose randomly between Freshman, Sophomore, Junior, and Senior. Then it would randomly click on one student to bring up that student's name and face. The participant was instructed to answer each question and automatically generate another student fifteen times. This method gave us a uniform distribution of freshmen, sophomores, juniors, and seniors.
When the internet was too slow, we generated a random student "manually." From the student directory, we copied a list of all undergraduates. Then we would randomly select fifteen of those names, using a Python random number generator to select random index locations, and print them to the screen. Then we would use those fifteen names and look up pictures in the student directory if the participant required them. This method did not consider the class of the student during random selection.

3.3 Statistical Results

In the end, we surveyed 32 students. According to the Hillsdale Student Directory, as of the time of this project, Hillsdale has 543 registered freshmen, 396 registered sophomores, 399 registered juniors, and 453 registered seniors for a total of 1791 registered students. Thus, we surveyed about 1.8% of campus.

We note that students are sometimes registered in a class (freshmen, sophomore, etc.) by credit, not necessarily by year. We also note that the number 1791 might include students registered online, though this number is likely small. In both cases, we believe this does not have a large impact on the data.

We found that **on average, participants knew 39.2% of randomly generated students** (i.e., knew them by name, by face, or both). Participants knew 5.6% by name only, knew 12.6% by face only, and knew 21% by both name and face. The standard deviation for percentages of students knowing others is 16.1%, with the maximum value being 73.3%, minimum being 6.7%, and the median being 46.7%. Since many people would not consider simply recognizing someone's face or name as "knowing" them, we also report a weighted average.

Definition 3.1 (Weighted Knowledge)**.** Let F be the number of students recognized by face, N the number of students recognized by name, B the number of students recognized by both, and T the total number of randomly generated students. Then a participant **knows (weighted)** $X\%$ of randomly generated students, where X is given by

$$X = \frac{\frac{1}{2}F + \frac{1}{2}N + B}{T} \tag{2}$$

Weighted knowledge basically says that recognizing someone's name or face is really only half knowing them. **On average, participants knew (weighted)**

[1] `https://apps.hillsdale.edu/custom/campus-directory`

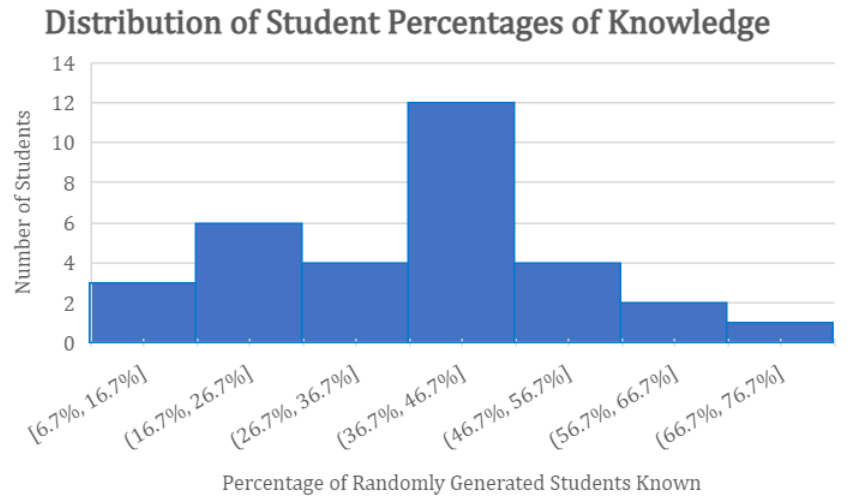

Figure 3: Results from the First Survey

30.1% of randomly generated students. The standard deviation of weighted knowledge is 12.3%, with the maximum value being 53.3%, minimum being 6.7%, and the median being 33.3%.

3.4 Graph Theoretic Results

To find the survey's $\overline{DoA}$, we must generate a directed graph to model the entire campus using the previous statistical results. We wrote a program in Python to generate a graph of 1791 nodes. Then, for each ordered pair of nodes, it generates a random number between 0 and 1. If that number is above the percent one student knows another (i.e., 39.2% in the unweighted case and 30.1% in the weighted case), it creates an edge from the first node to the second. Then, we perform Floyd's algorithm to transform the matrix. Finally, we calculate the maximum and average DoA. In the unweighted case, we found that the maximum $DoA = 2$ and $\overline{DoA} = 1.608$. In the weighted case, we found that the maximum $DoA = 2$ as well, and $\overline{DoA} = 1.699$. As expected, the $\overline{DoA}$ in the weighted case is higher than that of the unweighted case. This is because the weighted case has fewer connections in the graph, so paths from one student to another are longer on average.

3.5 Evaluation

These results suggest that Dan's conjecture is true, but we should evaluate the survey as a whole to determine its accuracy.

Roughly half of the participants experienced automatic random student generation, and the rest experienced manual random student generation. Some-

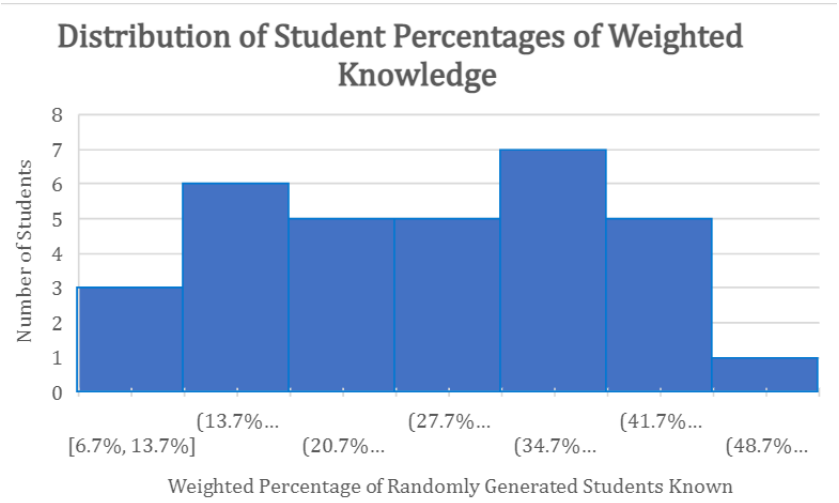

Figure 4: Results from the First Survey, Weighted

times, due to a high concentration of participants at once, multiple participants answered a question using the same randomly generated student. While not ideal, the AMC believes this variation has little impact on the resulting data. This is because the randomly generated students were generated independently of the identity of the participants, and the method of scoring the participants only depended on *how many people they knew and to what extent*, not on *which people they knew in comparison to other participants*.

In terms of statistical results, some members of the AMC were surprised at how high the percentage of students known on campus was. If a participant knows 39.2% of a campus of 1791 students, then the participant knows $39.2\% \times 1791 = 702$ students. To some, this seemed like a lot. Even the weighted estimate, which is smaller, indicates that a participant knows (weighted) on average $30.1\% \times 1791 = 539$ students. To explain this unexpected result, it might be noted that the sample for each participant was only 15 students. This is only 0.8% of campus. This could cause individual participant results to be skewed. Namely, the maximum value of 73.3% knowledge of campus implies the participant recognizes 1313 students, which many of us would suspect is not true. This survey could be improved by taking a larger sample of randomly selected students for each participant.

In terms of graph theoretic results, we remark that an overly high statistical result would decrease our estimation of the DoA. That is, if it turns out students only know 20% of campus (as an arbitrary example), it is much more likely that the maximum DoA becomes 3 or more. Furthermore, we note that if one node has an edge pointed to another node, it is much more likely the second node also has an edge pointed to the first. This is not reflected in our program, which contains an independent check for each direction, both with the same set

probability. Because of these concerns, the AMC decided to perform a second survey with a different fundamental structure.

4 The Second Survey

4.1 Survey Process

For the second survey, we manually generated 300 random students from the student directory. Then, we emailed these 300 students, asking them to participate in our survey. 45 students agreed and participated. Each participant was sent an identical Google form of 45 questions. Each prompt had the name and picture of one of the 45 students taking the survey and asked if the participant "knew" them. Participants could answer "Yes" or "No." Participants were instructed to interpret "know" as instinctively as they could. The advantage of this second survey, aside from having more participants and more responses, is that each participant is asked about *the other participants*. This ensures that the correlation between one student knowing another and that other student knowing the first is captured, even in a directed graph. In the end, results for statistics were compared to the first survey. However, instead of generating a random graph to match the statistics, a graph was generated *directly from the data*. Since students were all surveyed about each other, the set of surveyed students functions as a subgraph of campus.

4.2 Statistical Results

We randomly selected 300 students, which is 16.8% of campus. 45 students participated, which is 2.5% of campus. **On average, students knew 40.9% of other students**. The standard deviation of percentage of students known is 16.9%, with the maximum value being 75.6%, the minimum being 4.4%, and the median being 42.2%. Similarly, on average, students were known by 40.9% of other students, although for many of participants, the number of students they knew was *different* from the number of students who knew them. The standard deviation for the percent of students one student was known by is 18.5%, although the maximum, minimum, and median values are all exactly the same as those for a student knowing others. Paradoxically enough, **on average, students knew 10.2% more people than they were known by.** More specifically, each student has a knowledge ratio, defined below.

Definition 4.1 (Knowledge Ratio)**.** Let p be the number of students a participant knows, and let q be the number of students who know the participant. Then the participant's **knowledge ratio (k)** is $k = \frac{p}{q}$.

The average knowledge ratio for students is 1.102, implying that, on average, students knew 10.2% more people than they were known by. Specifically, the data shows that 24 students had $k > 1$, 19 students had $k < 1$, and 2 students had $k = 1$. Note that more students had $k > 1$ than had $k \leq 1$. This shows that

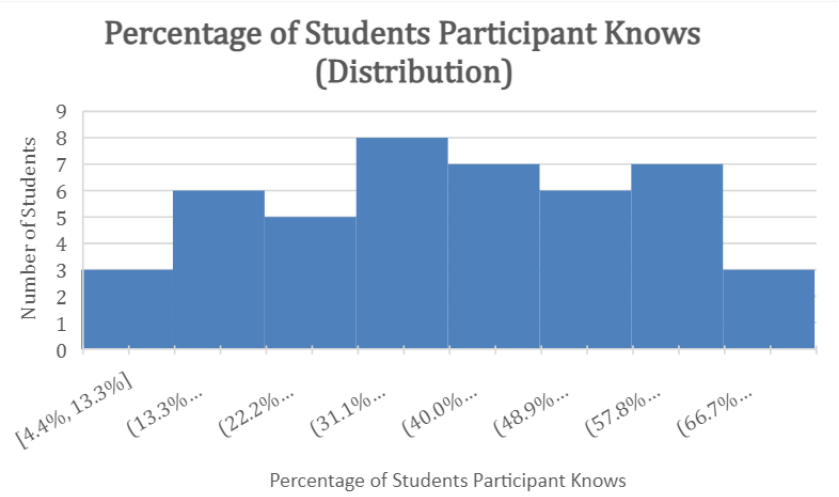

Figure 5: Second Survey Results, Actively Knowing

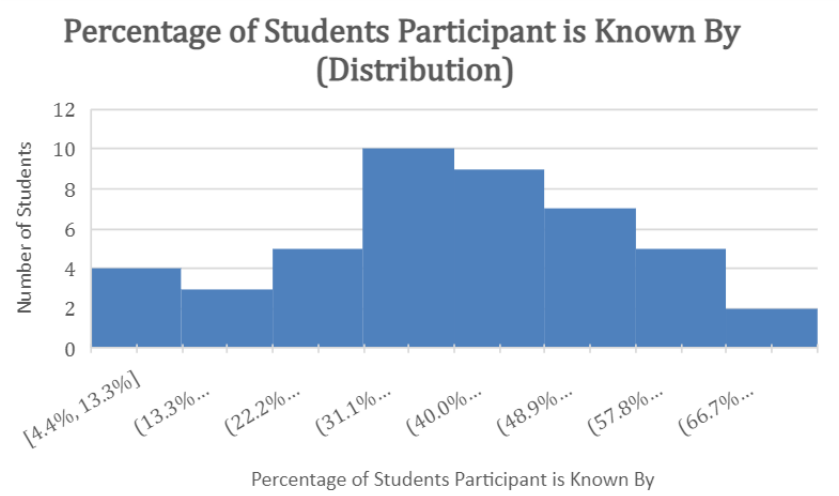

Figure 6: Second Survey Results, Being Known

a small majority of participants knew more students than they were known by, even though the average number of students a participant knows is equal to the average number of students that know a participant.

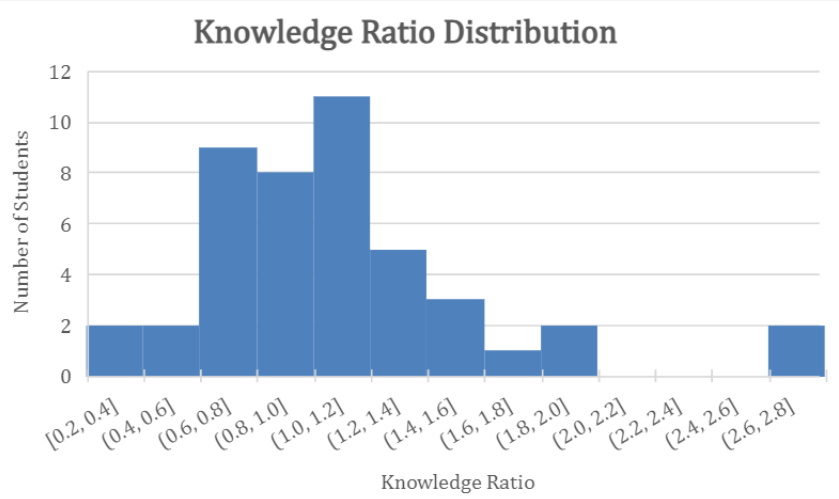

Figure 7: Knowledge Ratio Distribution

4.3 Graph Theoretic Results

Since participants were asked if they knew one another, the data can be translated directly into a directed graph. The very graph is reproduced here, in Figure 8. Having run Floyd's algorithm, we find that the maximum $DoA = 3$, and $\overline{DoA} = 1.589$. In fact, there are 83 ordered pairs of students with a DoA of 3. That is a relatively small number, considering there are 1980 ordered pairs of students in the graph (excluding the participant with themselves).

4.4 Evaluation

The AMC is much more confident in the results of this second survey. The surveying environment was far more controlled and equal for all participants, and the structure lent itself more directly to a graph theoretic solution. In comparison to the first survey, we note that the statistical results support each other. That is, 40.9% and 39.2% are very close, so that both surveys support the result that **as a general rule, on average, students know about 40% of campus**.

Since the maximum DoA is 3, Dan's Conjecture for the second survey graph is false. The members of the AMC, however, have debated over what this implies of Hillsdale as a whole. Of all the ordered pairs of students in the second survey, only 4.2% had $DoA = 3$. In contrast, 54% of pairs had $DoA = 2$, and

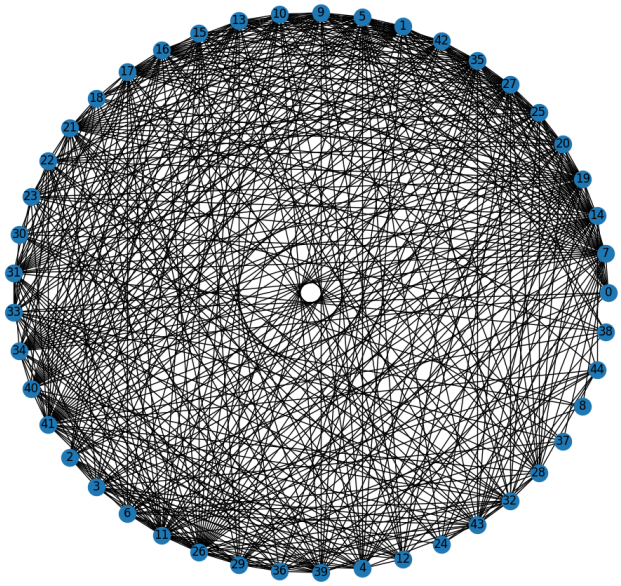

Figure 8: Graph of Student Connections in Second Survey

the remaining 41.8% of pairs had $DoA = 1$. With such a small number of degree 3 connections, it must be asked: as the number of students in a graph increases, how does the DoA change?

5 Final Remarks

As the number of students in a graph increases, how does the DoA change? Members of the AMC have argued both ways. One argument is that, as more students are added to a graph, more possible connections arise. Since there are more possible connections, there are more nodes through which a path of length 2 could be created. As a result, this argument concludes that the DoA would stay the same or decrease. On the other hand, it has been argued that the more students are added to a graph, the more likely a student who knows very few people is added. If that student is added, many more connections would be required to ensure they have a DoA of 2 with the previous students. Furthermore, for both surveys, participation was self-selected. It is possible that students who are more likely to participate in a survey have a greater likelihood of knowing and being known by others (perhaps because of their willingness to participate in social engagements). As a result, this argument concludes that the DoA would stay the same or increase. Regardless of how increasing the number of students affects the max DoA, our results indicate at least that

Dan's Conjecture is *generally* true. That is, even if the maximum DoA for any node in the graph of Hillsdale's student body is not 2, *the vast majority of* DoA *in this graph are* ≤ 2. Therefore, in the end, it would be very accurate to say, "If I don't know someone on campus, someone else that I know would know them."

6 Acknowledgements

This study was performed by the Hillsdale Applied Math Club (AMC). Our faculty advisor is Dr. Paulina Volosov. Our board for the current year is Jack Graham (President), Spencer DenBleyker (Vice President), Emily Balsbaugh (Secretary), and Lydia Hilton (Treasurer). Our members include Daniel Brand, Brianna Willhite, Eleanor Balsbaugh, Lional Armstrong, Daniel Ladzinski, Nicholas Treloar, Stephen Pearson, and many more. We also would like to thank the Hillsdale Coding Club, Ethan Cobb in particular, for help with randomly generating students during the first survey. Finally, we appreciate the cooperation of all of the survey participants, without whom this study would have been impossible.