Modeling Relationships Between Student Majors and Collegiate Friendships

Aru Bhoop[1]
Future Computing Technologies Lab: Creative Inquiry

[1] *Clemson University, SC*

April 27, 2020

This paper investigates the relationship between a student's undergraduate major and the major of their friends in college. Machine learning and other statistical models are used to predict student majors from those of their friends, the results of which illustrate how starkly different this relationship is for an individual's general friend group and their closest friends, the latter almost always sharing the same major as the student. These conclusions are corroborated by the macroscale patterns, and by using social network analysis, the tendency for individuals in similar majors to "cluster" in densely-connected groups is established.

1 Introduction

One's choice of friends in college are shaped by factors such as their academic major, socioeconomic background, and interests. However, to what extent does each factor play a role in determining friendships, and which factors are the most influential in the preference of some friendships over others? In this paper, I concentrate on the role that a student's major plays in both the selection of their closest friends and their general friend group within the context of Clemson University. I first provide a broad overview of how the data was collected, highlighting some of the technical obstacles, before discussing the form, distribution, and preparation of the data. Next, I construct several statistical models used to predict majors of students from those of their friends and analyze them. Finally, I summarize my findings and provide opportunities for future research.

2 Data Collection

2.1 Data Sources

I gathered data about student majors from the Clemson Student Directory (*Student Directory*) and friendships from the mobile-payment and social media app Venmo (*Venmo - Share Payments*), which, because of its pecuniary nature, is a much more reliable indicator of friendships than other social media platforms. One's closest friends were extracted from their recent transactions, while their general friend group consisted of all their Venmo friends. Both sources are publicly accessible, though an account may be needed to access certain parts.

2.2 Data Scraping

By completely automating the data collection process, data was obtained inexpensively, quickly, and on a much larger scale than possible by manual collection. However, this approach also came with several obstacles, some which could be preempted - such as authentication issues - but most - including broken proxies, duplicate profiles, session timeouts, missing data references, network failures, and expired cookies - had to be handled during run-time. To collect as many students as possible, Venmo was scraped in a breadth-first-graph-traversal based fashion starting from an arbitrary student. Before an individual's friends were added to the queue, they were cross-referenced with the student directory. The program would continue running until the queue had been exhausted. Trial runs of the program scraped at 26 profiles/second, but by using parallel processing and running it on a sixty-four processor Palmetto Supercluster node, scraping speeds soared to 1,750 people/second (6,640% faster). Within three hours, the program had finished scraping: 320,000 profiles had been searched, out of which 26,240 student profiles had been scraped.

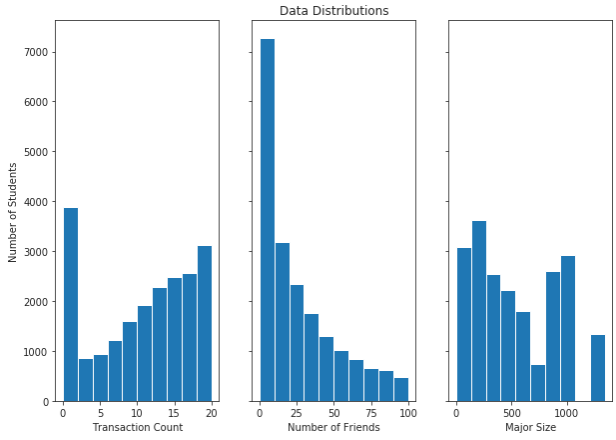

Figure 1: *Data Distribution*

3 Data Preprocessing

Quality training data ensures accurate model performance. From the 26,455 individuals collected, removed were 6,000 sans-major graduate students and 3,000 students without any transactions (likely due to hidden profiles; See Figure 1). To reduce noise, individuals with insufficient data - less than five transactions and in majors with less than twenty people - were also removed. Majors were one-hot encoded, and the data was normalized by calculating the distribution of majors for each person's friends. Finally, two pairs of training and test sets were formed based on general friend groups (x_1) and transaction-based friends (x_2).

4 Models

4.1 Model Creation

The first model was a single-layer neural network, which correctly classified students with the first dataset 30% of the time, but - with the second dataset - accuracy jumped to 93%. In other words, the model was able to predict a student's major from those of all their friends a third of the time but, by using their best friends, over 93% of the time. To better evaluate these results, I created a simple baseline by taking the most frequent major among a person's closest friends. Unexpectedly, the baseline performed better than the neural network with accuracies of 33% and 95%, respectively. Given these figures, I reasoned that the less flexible methods would outperform the overfitting neural network and created models using Logistic Regression, K-Nearest Neighbors, and Decision trees.

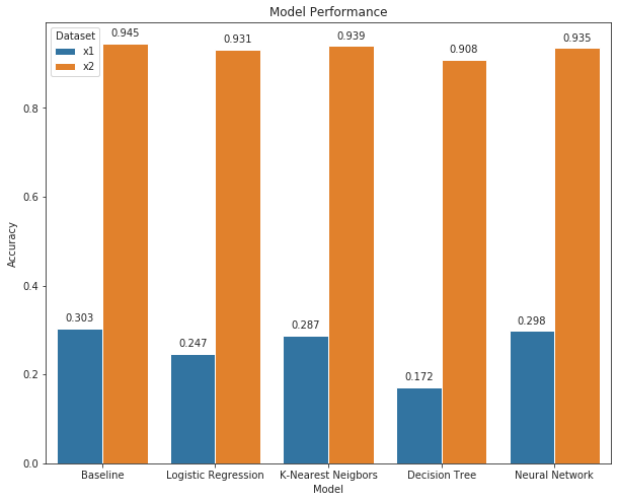

Figure 2: *Model Performances*

4.2 Model Results

Model performances were just shy of the baselines, with the Neural Network outperforming all the models for the first dataset, but coming in second to K-Nearest Neighbors for the second dataset, which fell short of the baseline by less than half

a percentage point. Surprisingly, the least flexible model, Logistic Regression, came second to last, only performing better than Decision Trees (Figure 2). The models possibly underperformed due to unobserved factors, including social ones such as membership in a sports team or organization, which may have taken precedence over major in influencing one's best friends, as well as biases in the data itself, as Venmo is increasingly being used to pay strangers. Collectively, these factors form the irreducible error, which limits how well these models can approximate the true friendships-student-major relationship.

5 Reflection

Though I cherished several components of this Creative Inquiry Course, I particularly enjoyed its self-driven nature as it encouraged exploration of topics beyond the coursework, which is something that I miss in most lecture-based courses. I was able to learn about a wide variety of topics while working on my semester project, such as concurrent computing, probability-density distributions, and force-directed graphs. This heuristic pedagogical style coupled with frequent interactions with the mentor through emails proved to be very effective, and the only drawback to this course is that more courses are not like it. By being introduced to a variety of topics beyond machine learning, such as Latex, Jupyter Notebooks, high-performance computing, and the Palmetto Supercluster, I'm much more confident of the courses I want to take later in the future. Finally, this course has helped solidify my decision to pursue machine learning in graduate school.

6 Conclusion and Future Work

The results show that one's closest friends are almost always in the same major, but their larger friend group is often more diverse. This pattern is reinforced on the macroscale by the "clustering" behavior of students in a major, most prominently among those in General Engineering, Pre-Business, and the Biological Sciences. It is interesting to note that this pattern also appears between majors within a college, such as between students in Nursing and the Biological Sciences (Figure 3). Though understanding the underlying causes of this phenomena is outside the scope of this project, it may be of interest to those wishing to extend this research. It would also be worthwhile to explore this phenomena at other universities and how factors such as college size affect major clustering.

7 Acknowledgments

I wanted to acknowledge Clemson University for its generous allotment of compute time on the Palmetto cluster, as well as our mentor Ben Shealy and the rest of the Creative Inquiry team for their continued support during the Covid-19 Crisis.

Bibliography

Student Directory. URL: https://my.clemson.edu/.

Venmo - Share Payments. URL: https://venmo.com/.

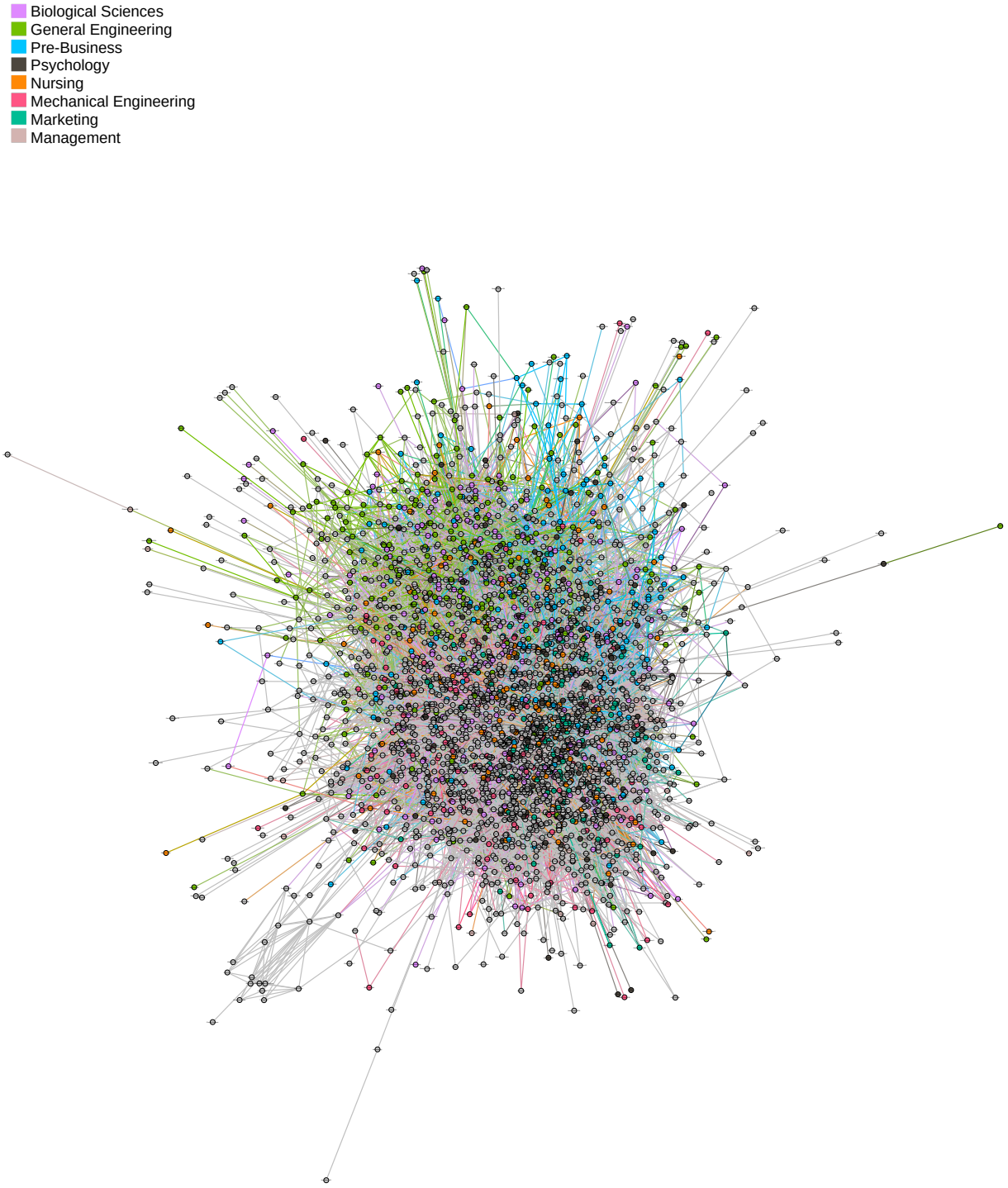

Figure 3: *Partial Network Representation of Friendships at Clemson University*