CSE 416A Final Project Written Report

Team: Mason Hall (431598), Zac Christensen (441569), Min Choi (429081)
Github repository: https://github.com/mchoi63/416aFinalProject

For our project, we wanted to look into the overall habits of Wash U, specifically looking at where students prefer to eat after class. Our overall thought leaned towards several hubs around campus: the DUC for the general population, Whisper's as a central node, Stanley's for the engineers, etc… This project gave us an opportunity to poll the campus and get a sense of how student flow moves through the dining network around Wash U.

Online Survey

Our initial plan was to gather all of our information from online surveys that would be sent out throughout the school. This survey asked the same question for several different academic areas around campus: if you have a class in a certain area, where would you go and get something to eat afterwards? Our survey also took note of the We broke campus into 12 different academic areas, and listed 14 different dining areas. Those who filled out the survey had the option to mark off several dining halls, depending on their preferences. Since we asked based on multiple different academic areas with students listing several different halls for each, we were able to flatten the data and generate around 1000 edges for the edge list.

In-Person Questionnaire

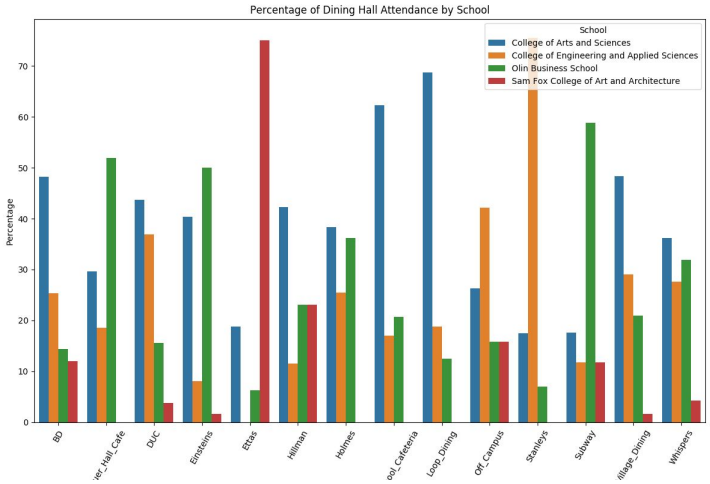
We also wanted to supplement our data with some in-person questionnaires. In a few of the dining halls we asked the same information that was asked online, such as year in school and the school that the individual was enrolled in. We then asked what building they just came from; this was slightly different from asking them about every area around campus, but was enough to give us a little bit more data.

Formatting the Data

To make things easy to work with and compatible with several different network/graphing packages, we stored all of our data in CSV files. We also worked with the data in Python, utilizing Pandas and Seaborn to the best of our ability in order to format the data in an understandable way. NetworkX, iGraph, and standard Python libraries were used to generate all of the analyses and graphs that have been seen throughout the project.

Basic Data Analysis

As an example of our data, here is a graph showing the normalized attendance at each dining hall based on school enrollment. The tall



yellow bar represents engineers at Stanley's, while the tall red bar shows the Sam Fox students at Etta's. This falls right in line with our expected outcome, and while this is not grounds for causation, we can certainly say there is a correlation.

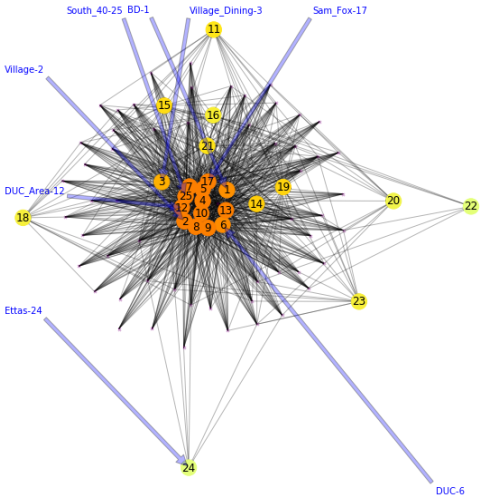
While there were similar things seen in the graphs relating to year in school, the results were not as striking, nor were they as interesting as what was found in centrality and community detection.

Centrality Measures & Graphing

Once the data was cleaned and exported to a CSV, it was loaded into our python notebook as a Pandas Dataframe before being converted into multiple NetworkX graphs:

1. Multigraph with the nodes being sources and dining halls
2. Graph with there being a node for each "pathway"
3. Graph with a node for each student response

In the multigraph and pathway-node networks we had the opacity of edges set to 0.3, so darker edges (and students for pathway-node network) means that there are multiple responses overlaid.

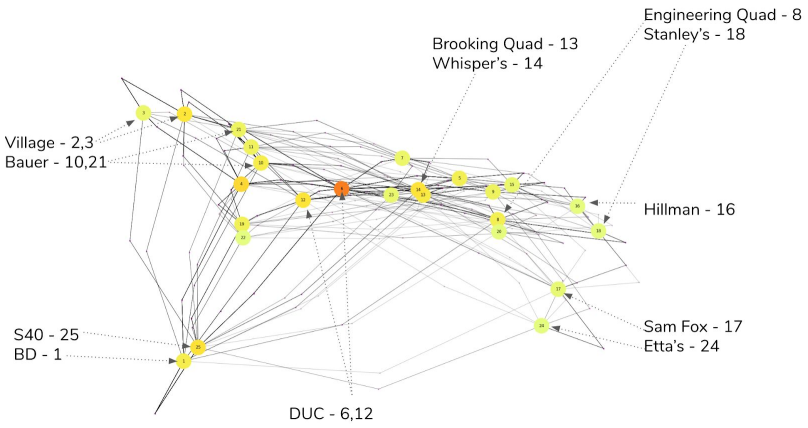


For the student nodes version, roughly all of the sources had the same centrality for each centrality measure tested (the highest). The closest dining hall was always the DUC, but it still had less centrality than the sources. This is likely due to sampling bias in our survey, since students were required to answer for *every* source but not every dining hall. Below is a partially labelled eigenvector centrality diagram:
The pathway node graph had some of the most interesting results, providing different results for the different centrality measures (relative to the source-saturated results of the first graph).

Eigenvector	Betweenness	Closeness
DUC: 0.635	DUC: 0.390	DUC: 0.408
AC & Frat Row: 0.162	AC & Frat Row: 0.135	Whispers: 0.357
DUC Area: 0.151	South 40: 0.114	Brookings: 0.354

It's interesting to note how the DUC is not that important to closeness of the graph, despite being extremely significant in all other centrality measures. It makes sense that Whispers has significance with closeness, since it is very central for the entire campus (though

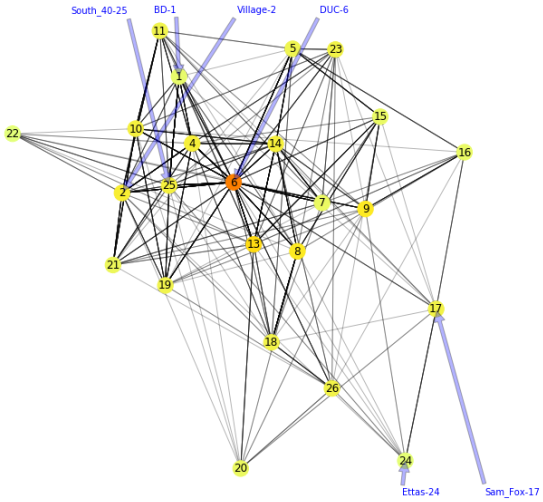
less than the DUC). Another **very** interesting observation came from the drawing of this graph. While different networkx [clustering and] layout methods weren't that notable for any of the graphs, using the spectral layout on the pathway graph created a structure similar to that of campus' actual layout:



The multigraph was similar in construction to the pathway network, but there were no "edge-nodes", just multiple edges from sources to dining halls (one for each student pathway). This graph also had a large variety of centrality across the different measurement methods and it might be worth further analysis. We also contracted the Village and DUC nodes (instead of having both a source and target for each of them).

Eigenvector	Betweenness	Closeness
DUC: 0.650	DUC: 0.115	DUC: 0.92
AC & Frat Row: 0.337	Brookings: 0.042	Village: 0.719
Village: 0.302	Engineering: 0.032	Brookings: 0.719

Below is a picture of the multigraph using the Fruchterman-Reingold method (spring layout):



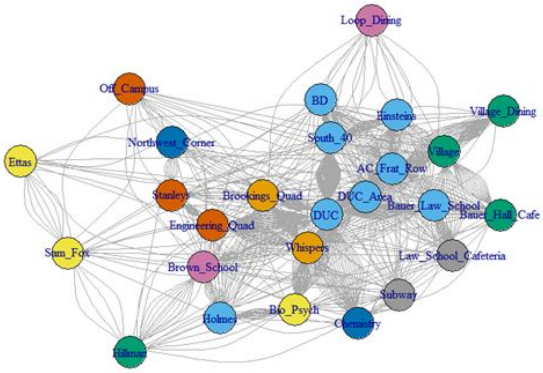
The initial conclusion from these graphs is that there is a distinct student flow through dining networks, as we expected before beginning our research. The differences in centrality measures both within and across each of the graph types holds promise for more information within the dataset. We believe further development and testing of different graph structures will provide more information.

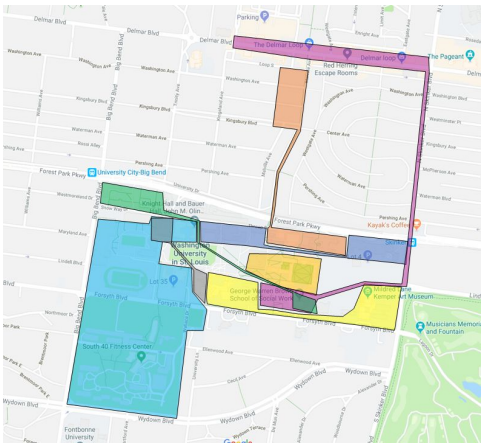
These networks also place the DUC as an integral part of dining/study-based student flow on campus. It is a popular spot no matter where students are going to or coming from.

Source *and* ***Dining Halls*** *labels are inside the notebook*

Community Detection

We used R's igraph library to perform community detection within the graph. We adopted two graph partitioning algorithms we learned in class, namely betweenness-based clustering and modularity maximization. The following graph is the graph of the buildings on campus with different colors grouping buildings into different communities.

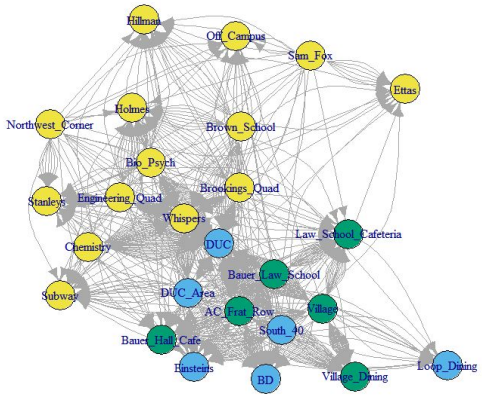


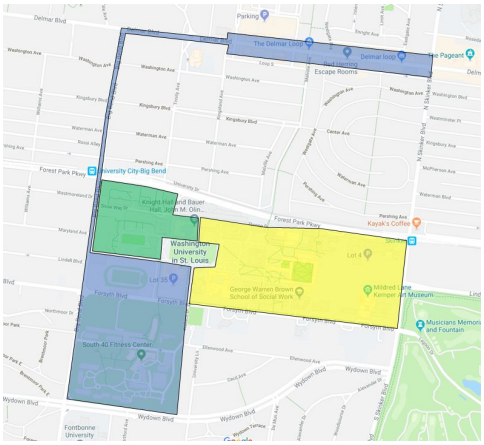


We can observe a few interesting points based on this graph:

1. There are 8 communities in the network, which can be mostly explained by the classes that are held in each of the buildings in the graph based on the schools that students are in.
2. There are two anomaly communities, which are: (Law School Cafeteria, Subway), (Hillman, Village Dining, Village, Bauer Hall Cafe). We believe that this might be because of the sampling bias that our data set has - our sample is based on undergraduates since we have asked different student groups that we are part of to fill out the survey.

The next graph is the graph with communities based on modularity maximization.





This graph makes more sense in terms of how close the buildings are, and what the areas are used for. Basically, residential areas (S40 and Village areas) seem to form their own communities with dining areas on campus that are close to them. Then, school buildings including those on the east end are all in one community - which can be explained that there are a lot of student movements among the buildings in that community, a lot denser than the movements between residential areas and the school area. It makes sense if we come to think of it as students staying on school areas and move among buildings during the day, and commuting from and to the residential areas explain the sparser connection between different communities.
One thing that we want to point out is that if we were to have more data points to be utilized for the analysis, we could have gotten better resolution in terms of the communities - having more communities (even within the school area and the residential areas).

Conclusion

After all of the analysis, we were able to come to a few conclusions. While we didn't have the largest amount of data, we were still able to see that the DUC was quite a central hub throughout campus. Likewise, there were a few dining areas that were clearly popular with different sets of students: engineers favored Stanley's, business school students prefer Bauer Cafe, and Sam Fox students congregate around Etta's.

Even when modelled abstractly as nodes and edges, the network structure mimics that of the true WashU building layout. The DUC really is a significant hub on campus though, dwarfing the importance of other dining areas.

When looking at communities, we can also see (through Modularity Maximization) that there are several defined areas throughout campus: South 40 and the surrounding buildings, the AC, Frat Row and the Village, and finally the rest of the eastern side of academic campus. Communities are definitely present when it comes to schools, but there isn't a huge difference between groups based on their year in school.