Clustering and Network Analysis of Majors at Amherst College

Sara Zhu

Introduction

Over the past few years, Amherst College has become well-known for its efforts to diversify its student body. However, there is little information reported about the diversity (or lack thereof) of students within classrooms and academic majors on campus. In the first portion of this report, I explore the demographic make-up of different majors at Amherst College to examine the degree to which different demographic groups (race, sex, etc.) may be clustering in certain majors. The data set from the IPEDS database includes information on 2150 degrees conferred to Amherst College students who graduated in 2018, 2019, 2020, and 2021 (IPEDS). The clustering analysis is performed on the following six variables: male, non-resident (international), asian, hispanic/latino, black, and white. After comparing two different clustering methods, k-means clustering and agglomerative hierarchichal clustering, we plot the observations in PC space to visualize the final clustering solution. The second portion of this report uses a network model to explore the relationship between majors that have cross-listed courses in the 22-23 Amherst Course Catalog. After mining and processing the data from Amherst's public course catalog, we construct a network based on 38 majors and 369 cross-listings extracted from 754 courses. To detect clusters of majors based on cross-listings, we compare two greedy community detection algorithms - a modulation optimization algorithm and an edge betweenness algorithm. The overall goal of this report is to identify clusters of majors based on demographic and curricular information.

Preliminary Analysis

Clustering

To sample the average graduating class of Amherst College, I use IPEDS data on 2396 degrees conferred to Amherst College students who graduated in 2018, 2019, 2020, and 2021 (IPEDS). I specifically chose the last 4 years because I was worried the data would be too inconsistent if we went further back. The original IPEDS data set for each year contained information on the reported race/ethnicity, residence status (for international students), and sex of every student who graduated with a degree during the 4 sampled class years. Each major was counted once, meaning double majors were counted as two separate observations.

The data processing stage involved the following modifications:

1) Since the Native Alaskan/Indian American, Hawaiian/Pacific Islander, Unknown Race, and 2+ Races groups only made up 0, 0, 2.7, and 5 percent of the sample population, respectively, I removed these variables so they wouldn't skew the final clustering solution. The 6 remaining quantitative variables of interest are:

- male - percentage of graduates that are identified as male (from a binary man/woman selection)
- non_resident - percentage of graduates that are not U.S. citizens
- hispanic_latino - percentage of graduates that are identified as either hispanic or latino
- asian - percentage of graduates that are identified as asian american
- black - percentage of graduates that are identified as black
- white - percentage of graduates that are identified as white

After removing the 4 variables mentioned earlier, the remaining variables represent the following shares of the sample population:

```
total_pct <- degrees_data[which(degrees_data$degree == 'Total'),]
print(total_pct)
```

```
## # A tibble: 1 x 10
##   degree total  male nonresident hispanic_latino asian black white code  field
##   <chr>  <dbl> <dbl>       <dbl>           <dbl> <dbl> <dbl> <dbl> <chr> <chr>
## 1 Total   584.  48.2        11.1            14.3  16.9  10.6  47.1 <NA>  <NA>
```

In addition to these quantitative variables, my data set includes 3 categorical variables:

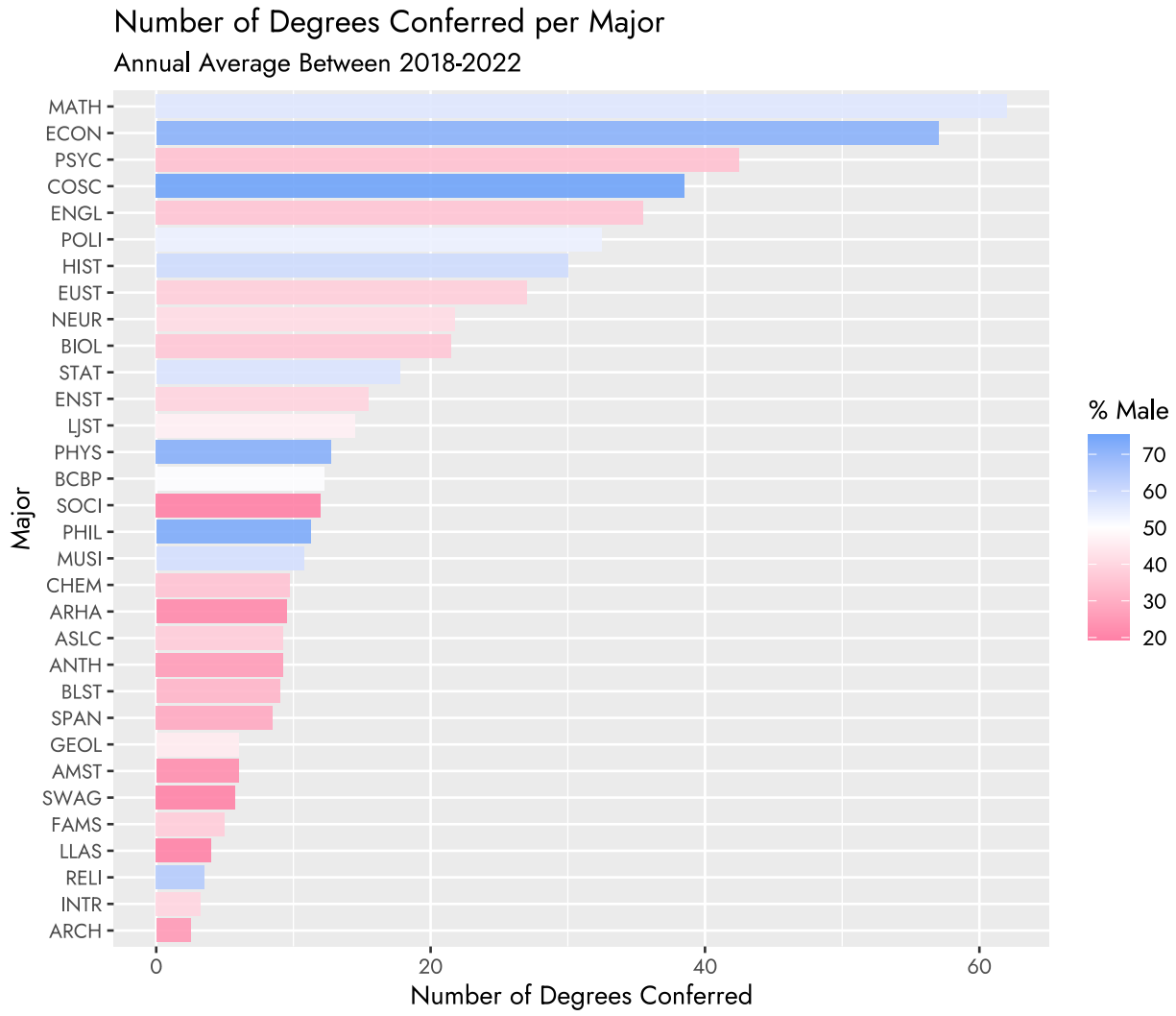
- degree - four letter major abbreviation
- total - the total number (not percentage) of degrees conferred for each major
- field - whether the major is classified as field or the Humanities (categorical/binary)

2) Because the original 38 majors differed in size, I consolidated the following majors into groups after verifying that the demographics were relatively similar within each grouping:

- Grouped Astronomy (avg of 3 majors per year) with Physics
- Grouped Theater & Dance (avg of 3.25 majors per year) with Art
- Created a new European Studies grouping that encompasses German (2.25), Classics (2.25), Russian (3.75), French (15.5)

3) After making these adjustments, I took the 4-year average for each degree category, resulting in a final data set containing 32 observations and 6 quantitative variables. The average number of degrees conferred to each major is shown below.

```r
degree_df %>%
  filter(degree != 'Total') %>%
  mutate(code = fct_reorder(code, total)) %>%
  ggplot(aes(x = total, y = reorder(code, total), fill = male)) +
  geom_bar(position = "stack", stat = "identity", alpha = 0.9) +
  scale_fill_gradient2(high = '#3992f7', mid = "white", low = '#ff80a6',
                       midpoint = 50, name = "% Male") +
    labs(title = "Number of Degrees Conferred per Major",
         subtitle = "Annual Average Between 2018-2022",
         x = 'Number of Degrees Conferred',
         y = 'Major') +
  theme(axis.text.y = element_text(size = 8.5),
        text = element_text(family = "jost"))
```



The average size of majors ranges between 2.5 - 62.

Finally, we note that the average number of degrees conferred in the final data set (566.25) is slightly lower than the number conferred in a normal year because a significant number of students chose to defer their enrollment or take a gap year/semester during the 20-21 school year. Apart from this inconsistency, there were no major differences in group proportions between years.

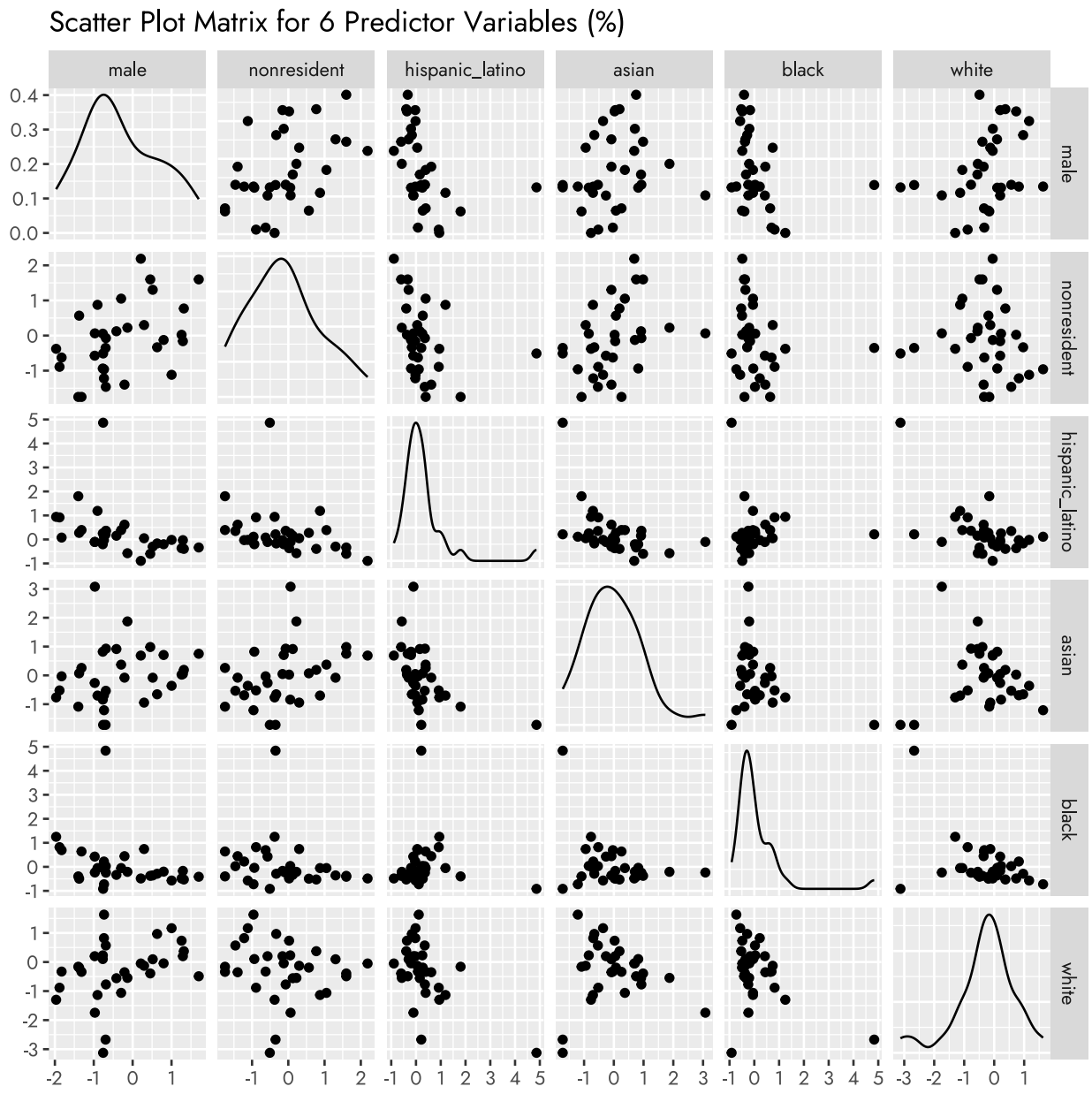
We will now proceed to our univariate analysis.

Since we are trying to identify majors where demographic groups tend to cluster, we will standardize the data to ensure all variables are on the same scale.

```
df1 <- degrees_data %>%
  filter(degree != 'Total') %>%
  # standardize the variables
  mutate(male = (male - total_pct$male)/sd(male),
         nonresident = (nonresident - total_pct$nonresident)/sd(nonresident),
         hispanic_latino = (hispanic_latino - total_pct$hispanic_latino)/sd(hispanic_latino),
         asian = (asian - total_pct$asian)/sd(asian),
         black = (black  - total_pct$black)/sd(black),
         white = (white - total_pct$white)/sd(white))
```

First, we examine a scatter plot matrix with the 6 quantitative variables.

```
df1[, c(3:8)] %>%
  ggpairs(title = 'Scatter Plot Matrix for 6 Predictor Variables (%)',
          upper = list(continuous = "points", combo = "dot_no_facet"),
          lower = list(continuous = "points", combo = "dot_no_facet")) +
  theme(legend.position = "bottom",
        text = element_text(family = "jost"))
```



Black, hispanic_latino, and asian all appear to have a large outlier. All other variables appear to be normally distributed.

```
df1[which(df1$black > 4), 'code'] # black outlier
```

```
## # A tibble: 1 x 1
##   code
##   <chr>
## 1 BLST
```

```
df1[which(df1$hispanic_latino > 4), 'code'] # hispanic_latino outlier
```

```
## # A tibble: 1 x 1
```

```
##   code
##   <chr>
## 1 LLAS
```

```
df1[which(df1$asian > 2), 'code'] # asian outlier
```

```
## # A tibble: 1 x 1
##   code
##   <chr>
## 1 ASLC
```

The outliers for the black, hispanic_latino, and asian are BLST, LLAS, and ASLC, respectively. This makes sense because these majors are culturally, ethnically/racially, or regionally specific. We can also see that these groups are the least represented among whites. Since these majors clearly stand apart from the others, we will remove them from the data set. We will also remove the interdisciplinary major because it isn't a defined category and only confers around 3 degrees per year.

```
outliers <- with(df1, c(which(black > 4), which(hispanic_latino > 4), which(asian > 2), which(code == '
# remove outliers from data set
df2 <- df1[-outliers, ]
```

After removing the outliers the data set contains 28 remaining observations. We can now examine a correlation matrix for the 6 variables of interest.

```
cor(df2[, 3:8])
```

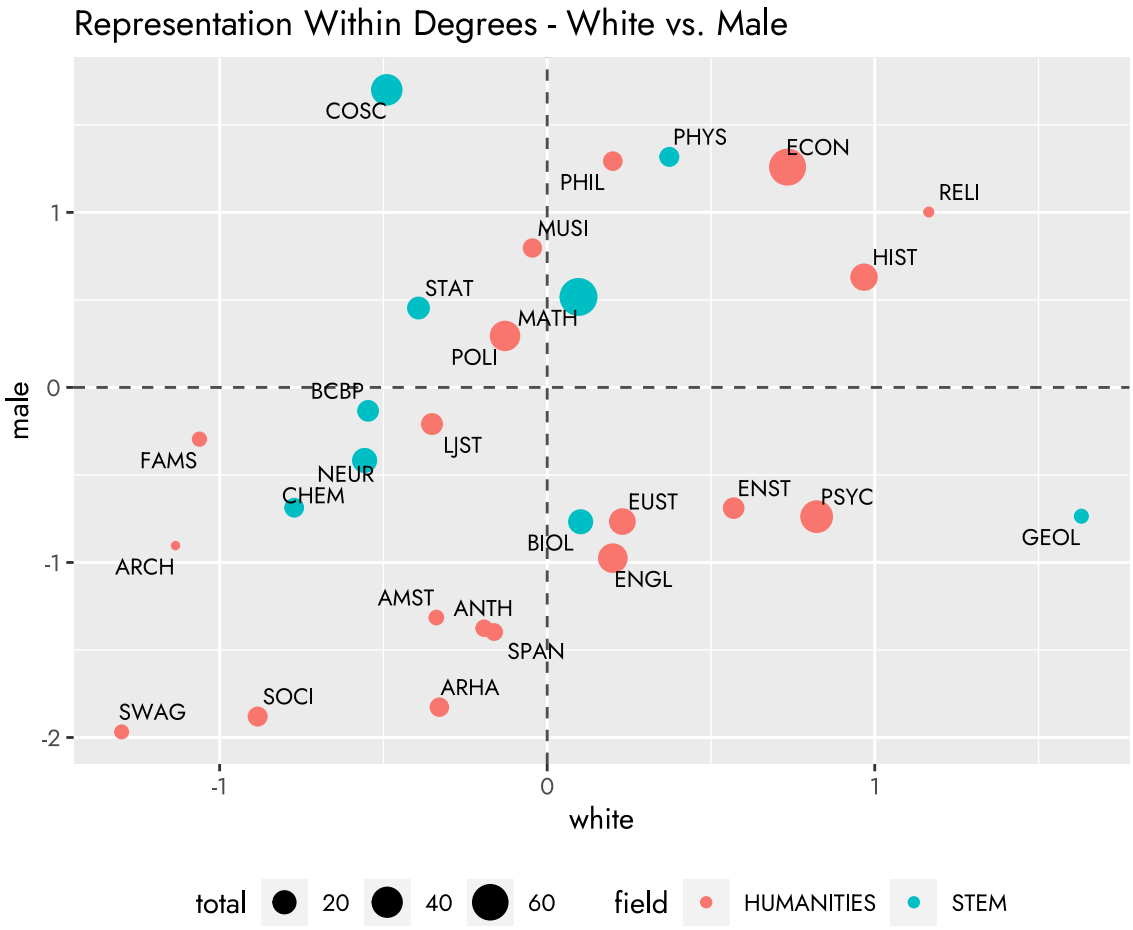
```
##                     male nonresident hispanic_latino     asian     black
## male             1.000000    0.460629       -0.639432  0.293443 -0.572299
## nonresident      0.460629    1.000000       -0.418747  0.390117 -0.323483
## hispanic_latino -0.639432   -0.418747        1.000000 -0.531057  0.347087
## asian            0.293443    0.390117       -0.531057  1.000000 -0.247826
## black           -0.572299   -0.323483        0.347087 -0.247826  1.000000
## white            0.363093   -0.327287       -0.397350 -0.345758 -0.463961
##                    white
## male            0.363093
## nonresident    -0.327287
## hispanic_latino -0.397350
## asian          -0.345758
## black          -0.463961
## white           1.000000
```

The correlation matrix shows that most of the variables are either moderately or strongly correlated with each other. Of particular interest are the strong negative correlations between hispanic_latino and male, black and male, black and white, and hispanic_latino and nonresident. There is also a strong positive correlation between male and white.

We can generate individual scatter plots to take a closer look at some of these relationships.

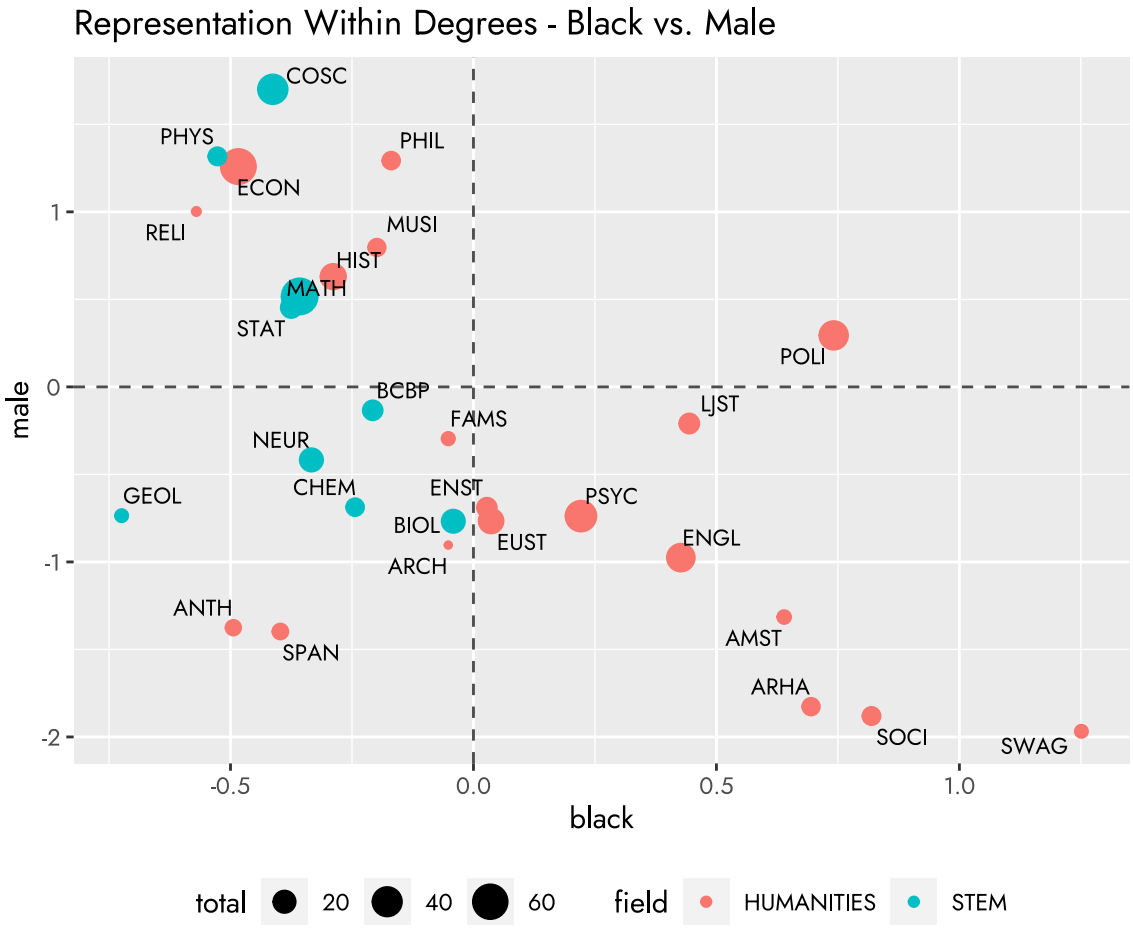
```
ggplot(df2, aes(y = male, x = white)) +
  geom_point(aes(color = field, size = total)) +
  geom_hline(yintercept = 0, color='#4d4d4d', lty = 'dashed') +
  geom_vline(xintercept = 0, color='#4d4d4d', lty = 'dashed') +
```

```
  labs(title = "Representation Within Degrees - White vs. Male") +
  ggrepel::geom_text_repel(label = df2$code, size = 3, family = "jost") +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```



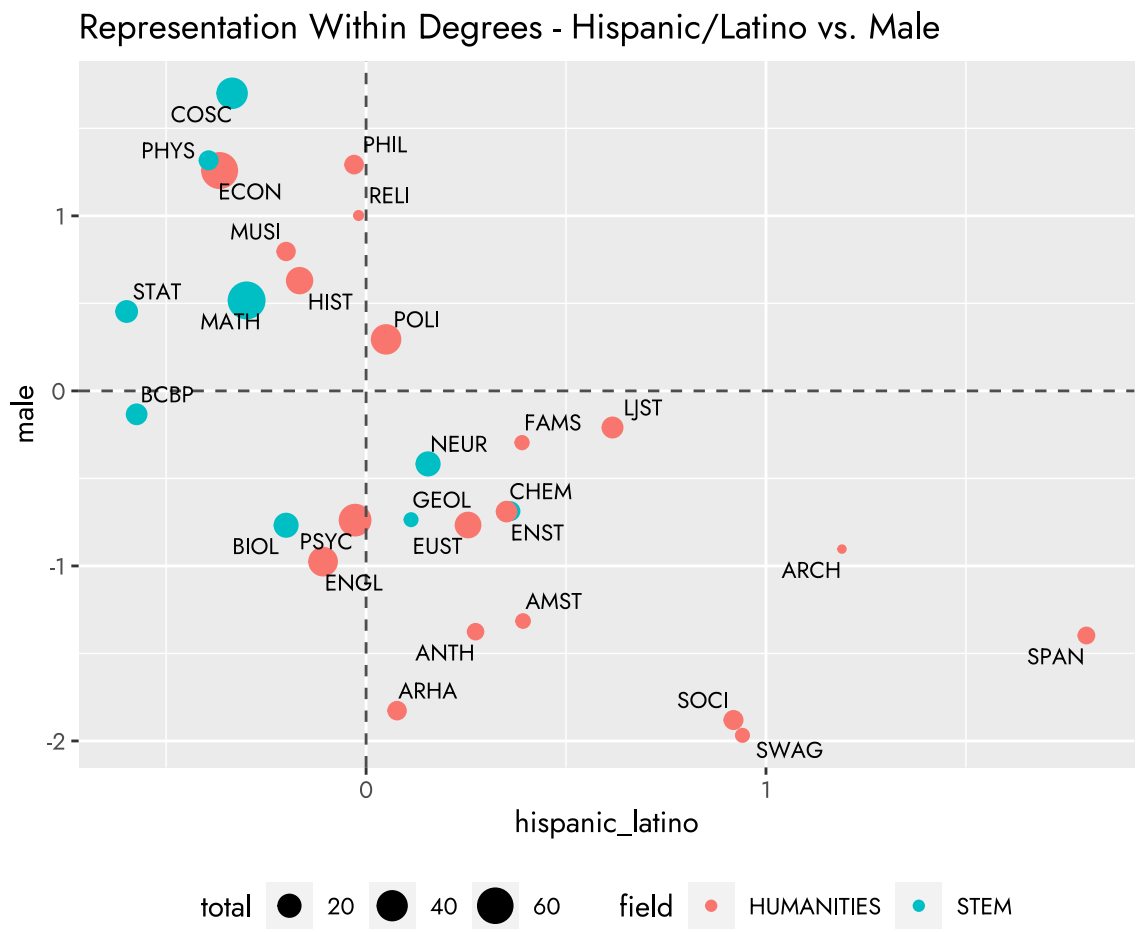
This plot shows a strong positive linear relationship between male and white with a correlation coefficient of 0.54. Both groups appear to be overrepresented in ECON, HIST, PHIL, PHYS, and RELI but underrepresented in some of the smallest majors (mainly the social sciences and arts), including SWAG, SOCI, ANTH, AMST, ARCH, ARHA, and FAMS. Males also tend to be overrepresented in COSC, MATH, STAT, and MUSI while white students are overrepresented in GEOL and PSYC.

```
ggplot(df2, aes(y = male, x = black)) +
  geom_point(aes(color = field, size = total)) +
  geom_hline(yintercept = 0, color='#4d4d4d', lty = 'dashed') +
  geom_vline(xintercept = 0, color='#4d4d4d', lty = 'dashed') +
  labs(title = "Representation Within Degrees - Black vs. Male") +
  ggrepel::geom_text_repel(label = df2$code, size = 3, family = "jost") +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```



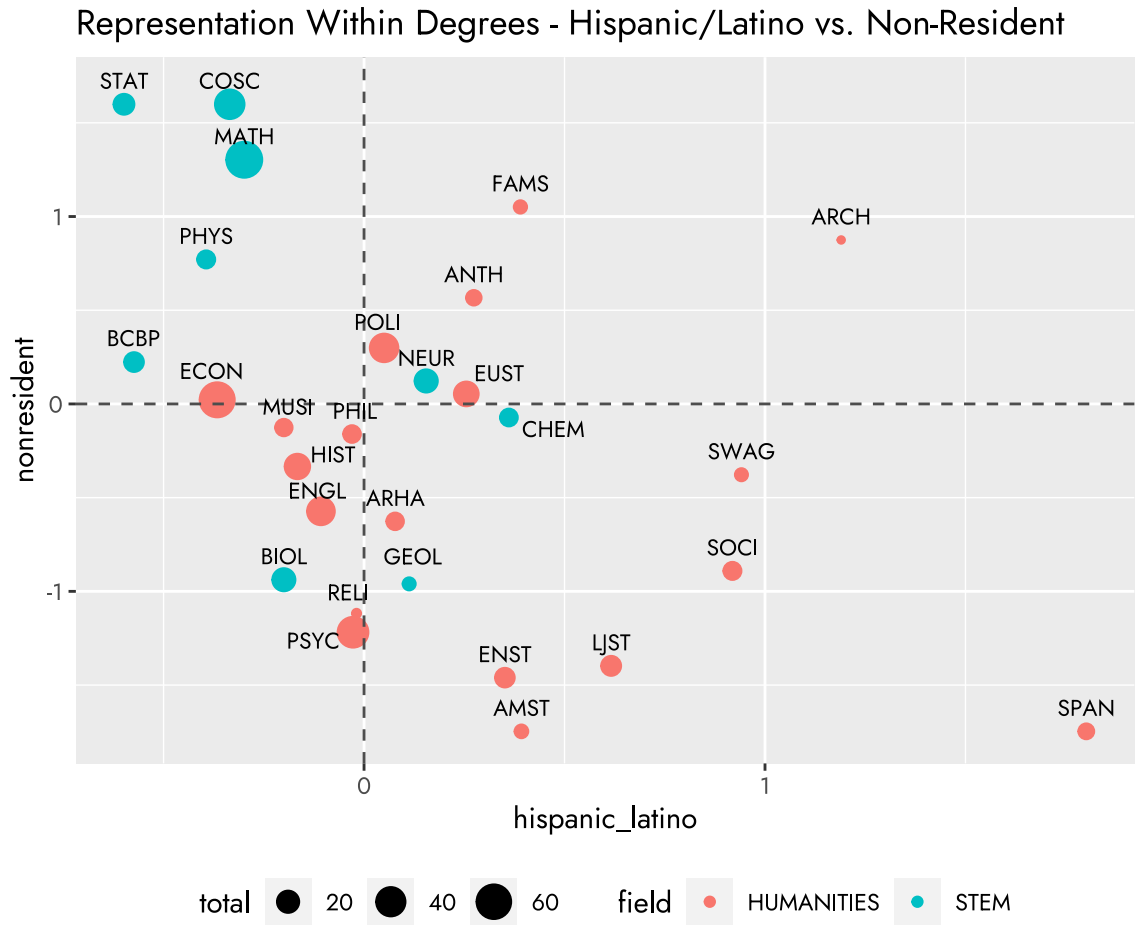
This plot shows a strong negative linear relationship between male and black with the lowest overall correlation coefficient of -0.64. Interestingly, black students seem to be underrepresented in every single STEM field but overrepresented in most of the small majors that were underrepresented by whites and males. Black students are also more likely to major in POLI, LJST, ENGL, and PSYC.

```
ggplot(df2, aes(y = male, x = hispanic_latino)) +
  geom_point(aes(color = field, size = total)) +
  geom_hline(yintercept = 0, color='#4d4d4d', lty = 'dashed') +
  geom_vline(xintercept = 0, color='#4d4d4d', lty = 'dashed') +
  labs(title = "Representation Within Degrees - Hispanic/Latino vs. Male") +
  ggrepel::geom_text_repel(label = df2$code, size = 3, family = "jost") +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```



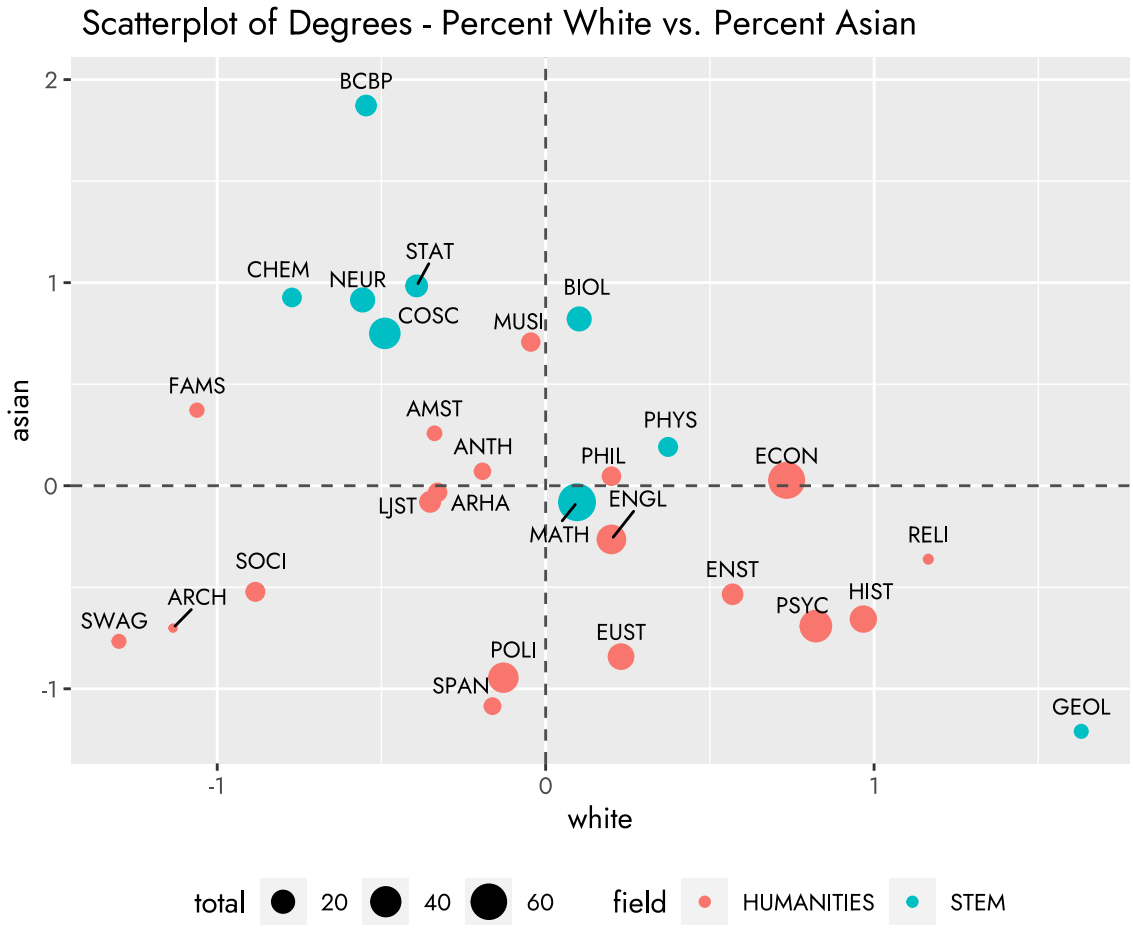
This plot shows a strong negative linear relationship between hispanic/latino and male with one of the lowest overall correlation coefficients of -0.63. The strength of this relationship can partially be attributed to the fact that some of the smallest majors at Amherst are extremely *overrepresented* by hispanic/latino students while extremely *underrepresented* by males. Because all majors are weighted equally, this coefficient should be interpreted as measuring the degree of separation that impacts the average major, not the average student.

```
ggplot(df2, aes(y = nonresident, x = hispanic_latino)) +
  geom_point(aes(color = field, size = total)) +
  geom_hline(yintercept = 0, color='#4d4d4d', lty = 'dashed') +
  geom_vline(xintercept = 0, color='#4d4d4d', lty = 'dashed') +
  labs(title = "Representation Within Degrees - Hispanic/Latino vs. Non-Resident") +
  ggrepel::geom_text_repel(label = df2$code, size = 3, vjust = 0, nudge_y = 0.08, family = "jost") +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```



This plot shows a moderate negative linear relationship between nonresident and hispanic with a correlation coefficient of -0.47. Non-resident students are particularly well represented in MATH, STAT, and COSC, which may be due in part to the heightened pressure international students feel (because of visa constraints) to acquire technical skills that are 'marketable' to employers. This effect is probably enhanced by the fact that English skills are relatively less important in these fields. The plot reveals that non-resident students are also overrepresented in ANTH and FAMS, while hispanic/latino students are overrepresented in SPAN, SWAG, SOCI, LJST, and ENST.

```
ggplot(df2, aes(y = asian, x = white)) +
  geom_point(aes(color = field, size = total)) +
  geom_hline(yintercept = 0, color='#4d4d4d', lty = 'dashed') +
  geom_vline(xintercept = 0, color='#4d4d4d', lty = 'dashed') +
  labs(title = " Scatterplot of Degrees - Percent White vs. Percent Asian") +
  ggrepel::geom_text_repel(label = df2$code, size = 3, vjust = 0, nudge_y = 0.08, family = "jost") +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```



This plot shows a moderate negative linear relationship between asian and white with a correlation coefficient of -0.43, Aside from GEOL and MATH, Asian students are overrepresented in every STEM subject, particularly in BCBP. Asian students are particularly underrepresented in GEOL, RELI, SPAN, SWAG, POLI, PSYC, HIST, and EUST.

Network

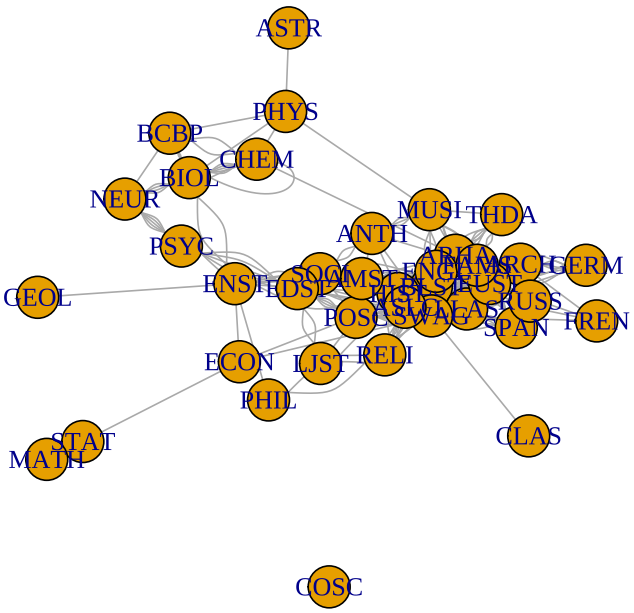
My second data set contains information on the number of course listings and cross-listings for each major in Amherst's 22-23 Course Catalog. After removing senior topics courses, labs, discussions, and other half-credit courses, the resulting data set contained 754 courses. In the pre-processing stage, I consolidated the Latin and Greek majors into the Classics major and the Chinese and Arabic majors into Asian Languages and Civilizations. I also removed the 6 Colloquium courses from the data set, resulting in a total of 38 nodes (majors) and 369 edge weights (cross-listings). In addition to the demographic variables examined in the clustering analysis, the second portion of my report includes the following variables:

- majors listed in the 22-23 course catalog (nodes)
- total number of times a major pair is cross-listed in the 22-23 Amherst College Course Catalog (edge weight)
- total number of distinct courses offered by each major during the 22-23 school year node size

Univariate Analysis

```
g <- graph_from_data_frame(d = edge_list, vertices = node_list, directed = FALSE)
```

```
set.seed(13)
plot(g)
```



The network contains one central component and one isolated component containing COSC. In the central component, MATH, STAT, GEOL, ASTR, and CLAS appear to be the least connected nodes.

```
vcount(g)
```

```
## [1] 38
```

```
ecount(g)
```

```
## [1] 369
```

```
is.weighted(g)
```

```
## [1] FALSE
```

We confirm that there are a total of 38 nodes (majors). Each edge is multiplied by its weight, there are a total of 369 edge weights (cross-listings). Next we will examine some descriptive statistics.

```
mean(degree(g))
```

```
## [1] 19.4211
```

```
transitivity(g)
```

```
## [1] 0.455556
```

```
average.path.length(g)
```

```
## [1] 2.42042
```

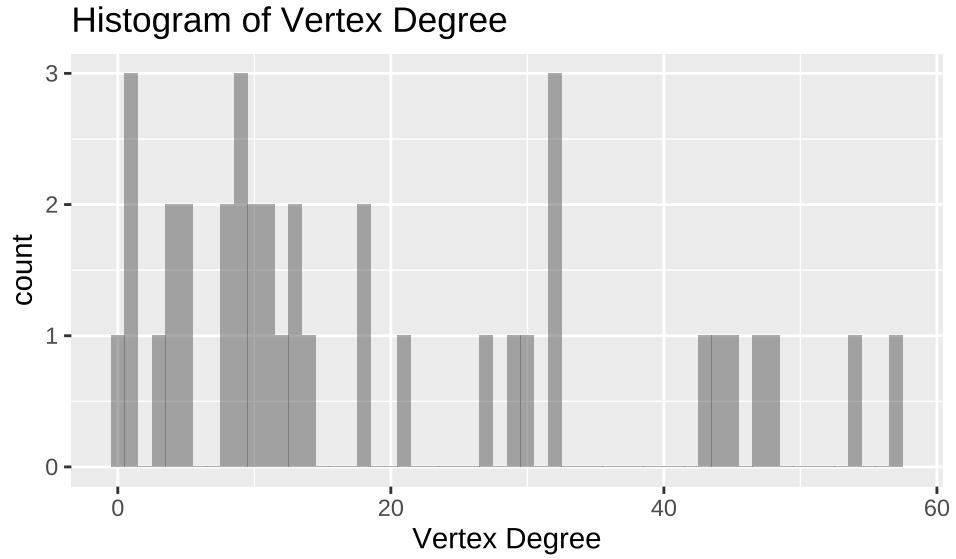
```
diameter(g)
```

```
## [1] 6
```

The graph has an average degree of 19.42, indicating that the average major is cross-listed approximately 19 times. We also observe a fairly high transitivity coefficient of 0.46. This represents the likelihood that two majors are connected if they are both connected to the same third major. For example, if ENST and STAT are both individually cross-listed with ECON, the probability that ENST and STAT are cross-listed is 46%. Finally, we note that the average path length of 2.42 is fairly small, while a diameter of 6 is fairly large.

Next, we will examine the distribution of degree, strength, and betweenness for all nodes.

```
gf_histogram(~ degree(g), binwidth = 1, xlab = 'Vertex Degree', title = 'Histogram of Vertex Degree')
```

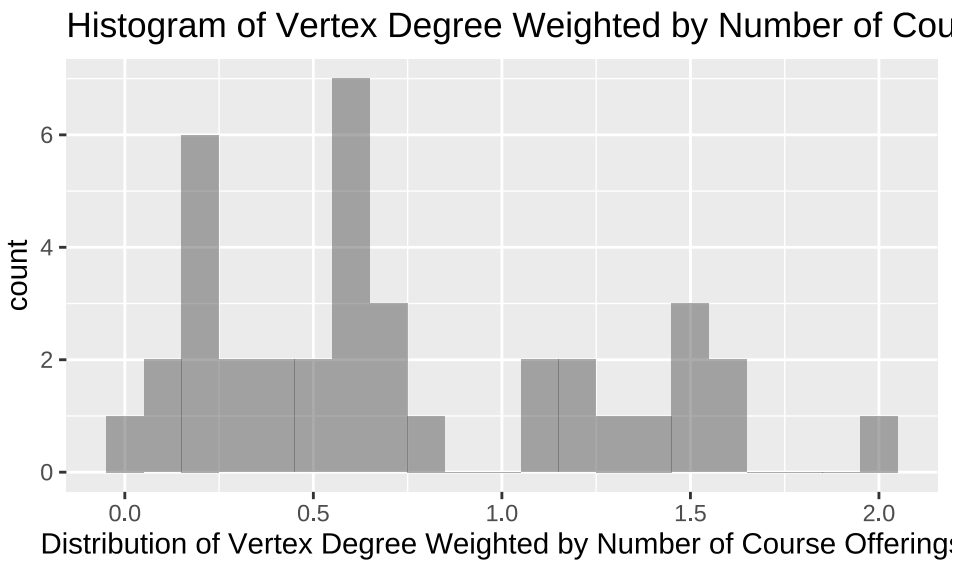



```
sort(degree(g), method = "shell", index.return = TRUE, decreasing = TRUE)$x
```

```
## HIST BLST SWAG ARHA ENGL EUST FAMS AMST ASLC EDST LLAS ARCH SOCI POSC BIOL SPAN
##   57   54   48   47   45   44   43   32   32   32   30   29   27   21   18   18
## STAT MATH MUSI ENST RUSS RELI PSYC NEUR THDA ANTH CHEM BCBP GERM LJST PHYS ECON
##   14   13   13   12   11   11   10   10    9    9    9    8    8    5    5    4
## FREN PHIL ASTR GEOL CLAS COSC
##    4    3    1    1    1    0
```

HIST, BLST, SWAG, ARHA, ENGL, FAMS are the most frequently cross-listed majors. Interestingly, COSC is cross-listed 0 times despite being one of the most popular majors at Amherst. The degree distribution is somewhat normally distributed with a strong right skew. Most majors are cross-listed between 0 and 20 times.

```
gf_histogram(~ degree(g)/node_list$num_classes, binwidth = 0.1, xlab = 'Distribution of Vertex Degree W
             title = 'Histogram of Vertex Degree Weighted by Number of Course Offerings')
```




```
sort(degree(g)/node_list$num_classes, method = "shell", index.return = TRUE, decreasing = TRUE)$x
```

```
##     BCBP     ARCH     EUST     EDST     LLAS     AMST     FAMS     SWAG
## 2.000000 1.611111 1.571429 1.523810 1.500000 1.454545 1.433333 1.333333
##     BLST     SOCI     ASLC     HIST     NEUR     ARHA     BIOL     RUSS
## 1.227273 1.227273 1.103448 1.096154 0.833333 0.746032 0.692308 0.687500
##     SPAN     RELI     ENGL     ENST     GERM     THDA     ANTH     STAT
## 0.642857 0.611111 0.600000 0.600000 0.571429 0.562500 0.562500 0.466667
##     POSC     MUSI     CHEM     PSYC     PHYS     CLAS     LJST     ASTR
## 0.466667 0.433333 0.409091 0.294118 0.263158 0.250000 0.238095 0.200000
##     MATH     FREN     PHIL     ECON     GEOL     COSC
## 0.196970 0.190476 0.176471 0.108108 0.100000 0.000000
```

When cross-listings are weighted by the number of distinct courses offered by each major (multiple sections count as one course), BCBP, ARCH, EUST, EDST, LLAS, FAMS, SWAG, BLST, SOCI, ASLC, and HIST score the highest while COSC, GEOL, ECON, PHIL, FREN, MATH, ASTR, and LJST score the lowest.

Methods

Clustering

The first section of this report applies a clustering analysis to detect clusters of demographically majors. To select which variables to use in our data analysis, we removed all variables that comprised 5% or less of the sample population, leaving 6 remaining features. To determine which observations are most similar or different from each other in terms of the 6 features, we compare solutions for two different clustering algorithms, k-means clustering and agglomerative hierarchical clustering with Ward's method. Since neither of the methods are scale-invariant, we scaled each variable before clustering.

Broadly speaking, K-means is an iterative algorithm that partitions data into a specified number of clusters (k). Both k-means and Ward's method try to make the points in each cluster as similar as possible by minimizing some criteria in each consecutive iteration of the algorithm. In the case of k-means, we define this criteria as the Within Groups Sum of Squares (WGSS). Unlike the k-means algorithms, hierarchical clustering algorithms do not require a specified number of clusters. Our analysis will use an agglomerative hierarchical clustering procedure known as Ward's method, which starts with n clusters in the first iteration then iteratively merges the pair of clusters that minimizes the increase in within-cluster variance for the entire data set.

Both k-means and hierarchical clustering require the analyst to choose the optimal number of clusters based on solutions for multiple cluster sizes. For k-means, we make this selection by plotting the WGSS against the number of clusters and looking for the 'elbow'. For hierarchical clustering, we can examine the sizes of the changes in height in the dendrogram. Once we have obtained our clustering solutions from both methods, we can assess the validity of each solution by computing their silhouette coefficients. After we have selected the strongest of the two solutions, we will use a PCA to visualize the final cluster solution.

Network

There are many useful tools in the field of network analysis that we can use to analyze the relationships between cross-listed majors. Formally, we define a graph G = (V, E) as a mathematical structure that consists of a set V of *vertices* (also called nodes) and a set E of *edges*. Our network will model majors as nodes and the cross-listings between majors as edges. Since there is often more than just one cross-listed course between two majors, our network is *weighted*. The network is also an *undirected* graph (where edges represent unordered pairs of distinct vertices) because cross-listings assign equal weight to all participating majors. After plotting the network, we will can use descriptive statistics to assess a few basic properties, including the *diameter*, *average path length*, and *transitivity* coefficient. We will also evaluate the degree centrality distribution to determine the most central (most often cross-listed) majors. Following the descriptive analysis, we will implement two different community detection algorithms: a modularity optimization algorithm and a Girvan-Newman edge-betweenness algorithm. Both algorithms take a greedy approach to searching the space of all possible partitions by iteratively modifying successive candidate partitions. In each iteration, the least costly merge of two previously existing partition elements is executed.

Modularity is a measure of the structure of a graph. Graphs with a high modularity score have many connections within a community and relatively few that are close to other communities. In each iteration of the modularity optimization algorithm, the merge that maximizes *modularity* is executed.

Edge betweeenness measures how often an edge lies on the shortest path between all possible pairs of nodes in the graph. This means edges that connect isolated communities have high edge betweenness because all other shortest paths must go through them. In our network, edge betweenness is also inversely related to edge weight. In each iteration of the Girvan-Newman algorithm, edges with the largest *edge betweenness* are removed.

After running the two algorithms, we will describe the detected clusters and compare their respective modularity values to determine the optimal partitioning solution.

Results

Clustering

K-means Clustering First, we will apply the k-means clustering approach.

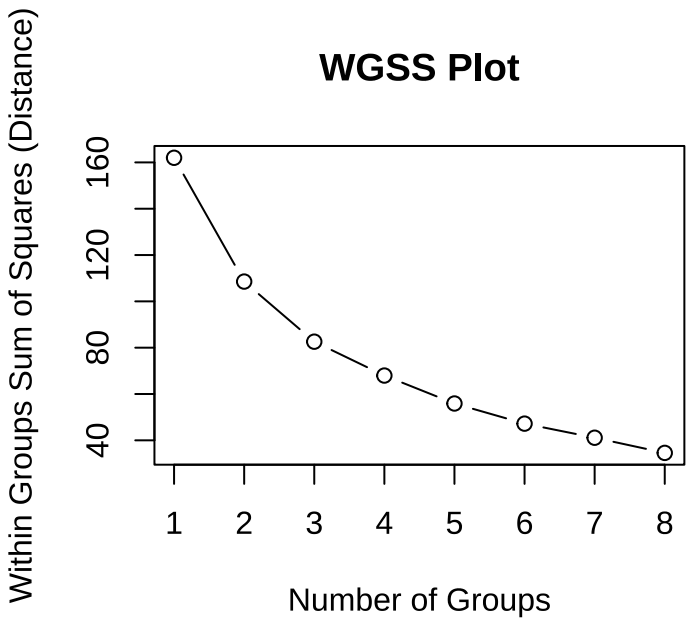
```
clust_df <- df2[, c(3:8)]

nclustmax <- 8
wss <- rep(0, nclustmax) #creates 8 copies of 0 to create an empty vector

for(i in 1:nclustmax){
  wss[i] <- sum(kmeans(scale(clust_df), centers = i, nstart = 10)$withinss)
}
```

To determine how many clusters to keep, we plot the within-groups sum of squares for one- to eight-group solutions.

```
set.seed(13)
plot(1:nclustmax, wss, type = "b", main="WGSS Plot",
     xlab = "Number of Groups", ylab = "Within Groups Sum of Squares (Distance)")
```

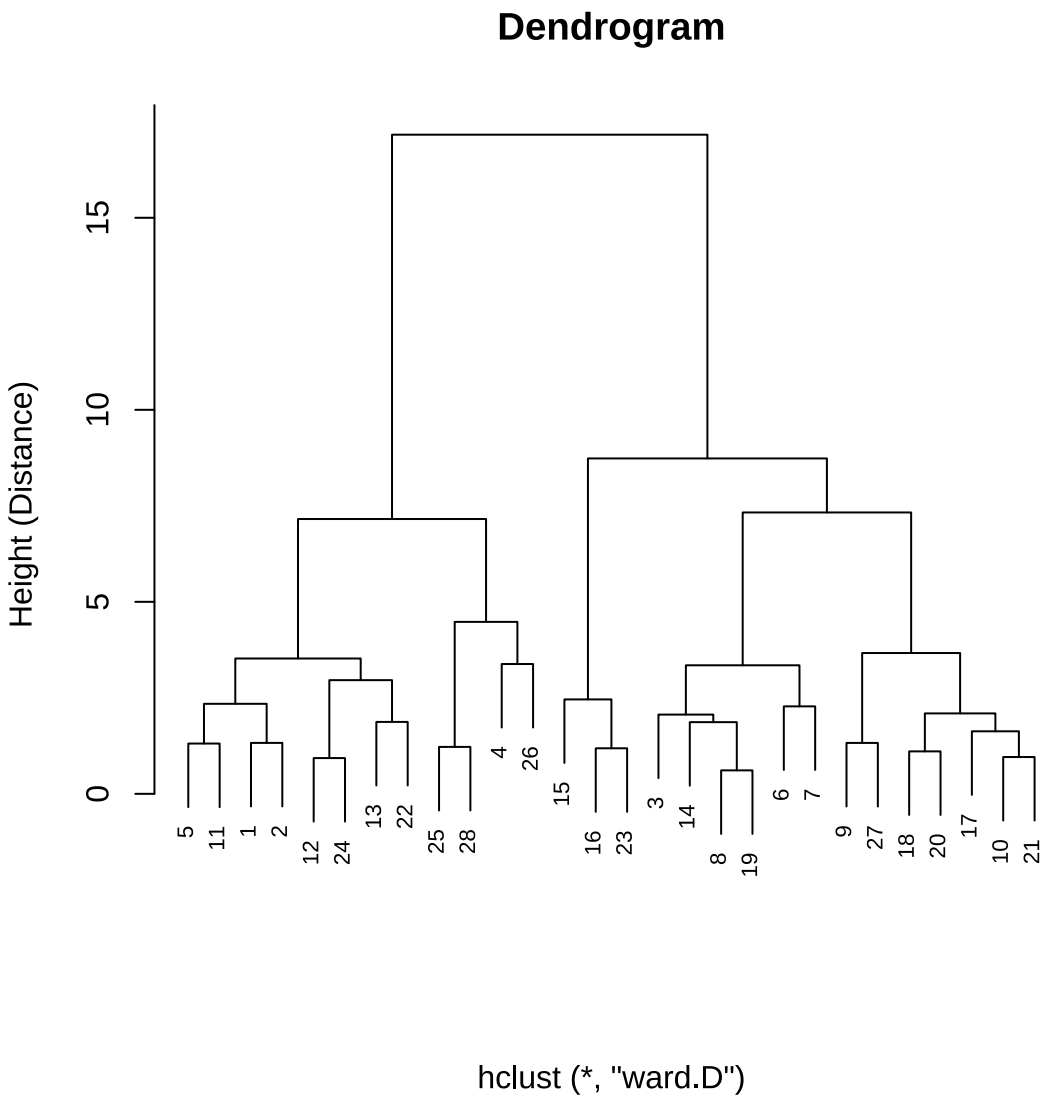


We observe an 'elbow' at six groups, so we will analyze the three-group k-means solution.

```
set.seed(13)
Ksol3 <- kmeans(scale(clust_df), centers = 3, nstart = 10)
```

Hierarchical Clustering Next, we will generate a hierarchical clustering solution using Ward's method.

```
major.dist <- dist(scale(clust_df))
hcward <- hclust(major.dist, method = "ward.D")
plot(hcward, cex = 0.7, main = "Dendrogram",
     xlab = "", ylab = "Height (Distance)")
```



We can determine the optimal cluster solution by examining the changes in height in the dendrogram. Changes in height after the third and sixth clusters form both seem like good places to cut. We will proceed with the three-group solution.

```
wardMajor <- (cutree(hcward, k = 3))
summary(as.factor(wardMajor)) #as factor to get table
```

```
##  1  2  3
## 12 13  3
```

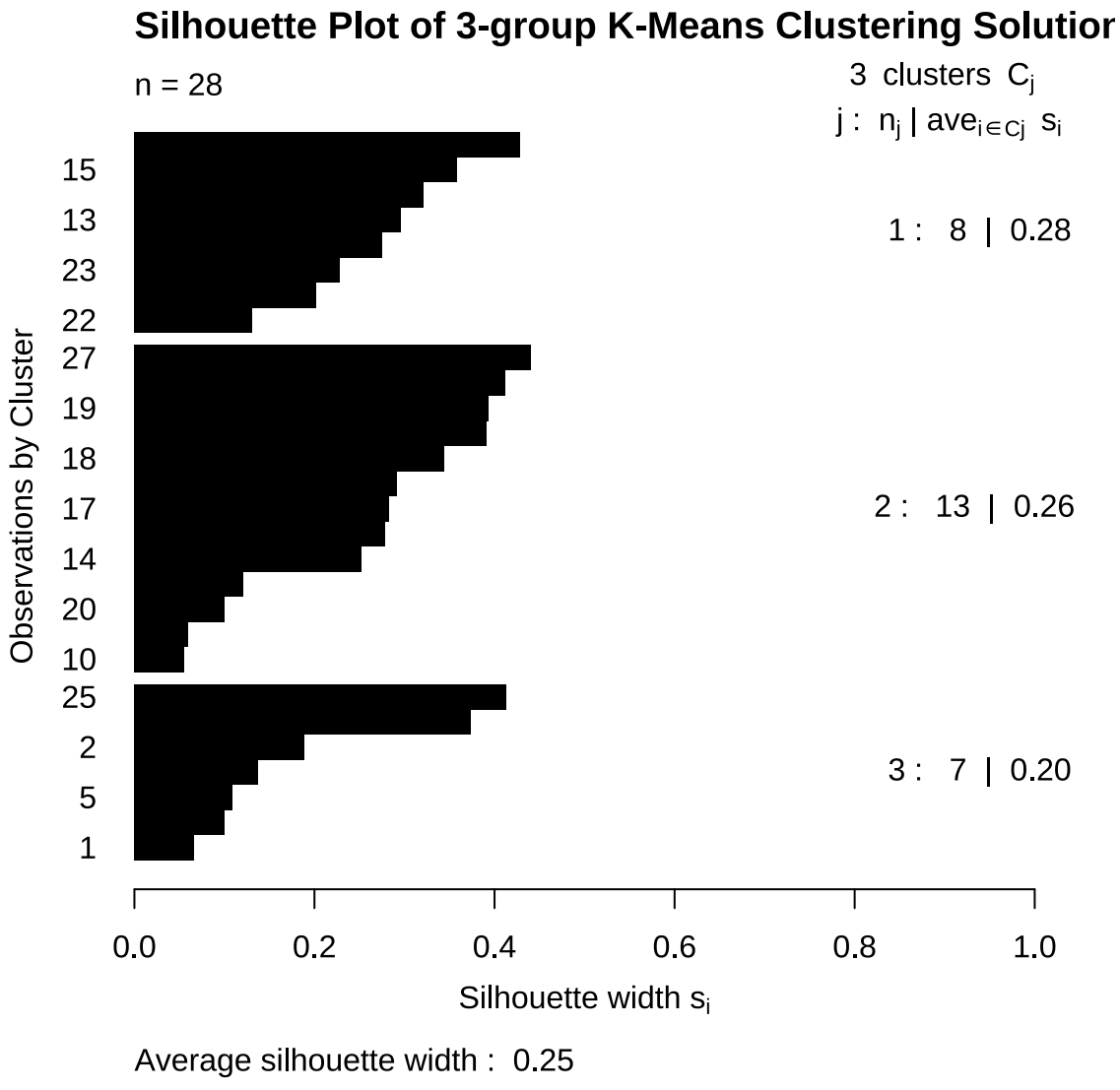
Now we will assess the validity of our two solutions.

```
kmeansSil <- silhouette(Ksol3$cluster, dist(scale(clust_df)))
summary(kmeansSil)
```

Cluersting Validation

```
## Silhouette of 28 units in 3 clusters from silhouette.default(x = Ksol3$cluster, dist = dist(scale(clu
##  Cluster sizes and average silhouette widths:
##        8       13        7
## 0.280014 0.263189 0.198209
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0552  0.1283  0.2769  0.2518  0.3617  0.4404
```

```
plot(kmeansSil, col = "black", main="Silhouette Plot of 3-group K-Means Clustering Solution",
     ylab = "Observations by Cluster")
```



```r
# get silhouette coefficient
summary(kmeansSil)$avg.width
```

```
## [1] 0.251752
```

The k-means algorithm generates three clusters sized 7, 8, and 13. The largest cluster with 13 points has the highest silhouette coefficient, indicating that these points are more tightly packed toward the center than the points in the other clusters. The cluster with 7 points has the lowest silhouette coefficient of 0.23, indicating the weakest structure. A low overall silhouette coefficient of 0.28 indicates that points in one cluster cannot easily be distinguished from points in the other cluster.

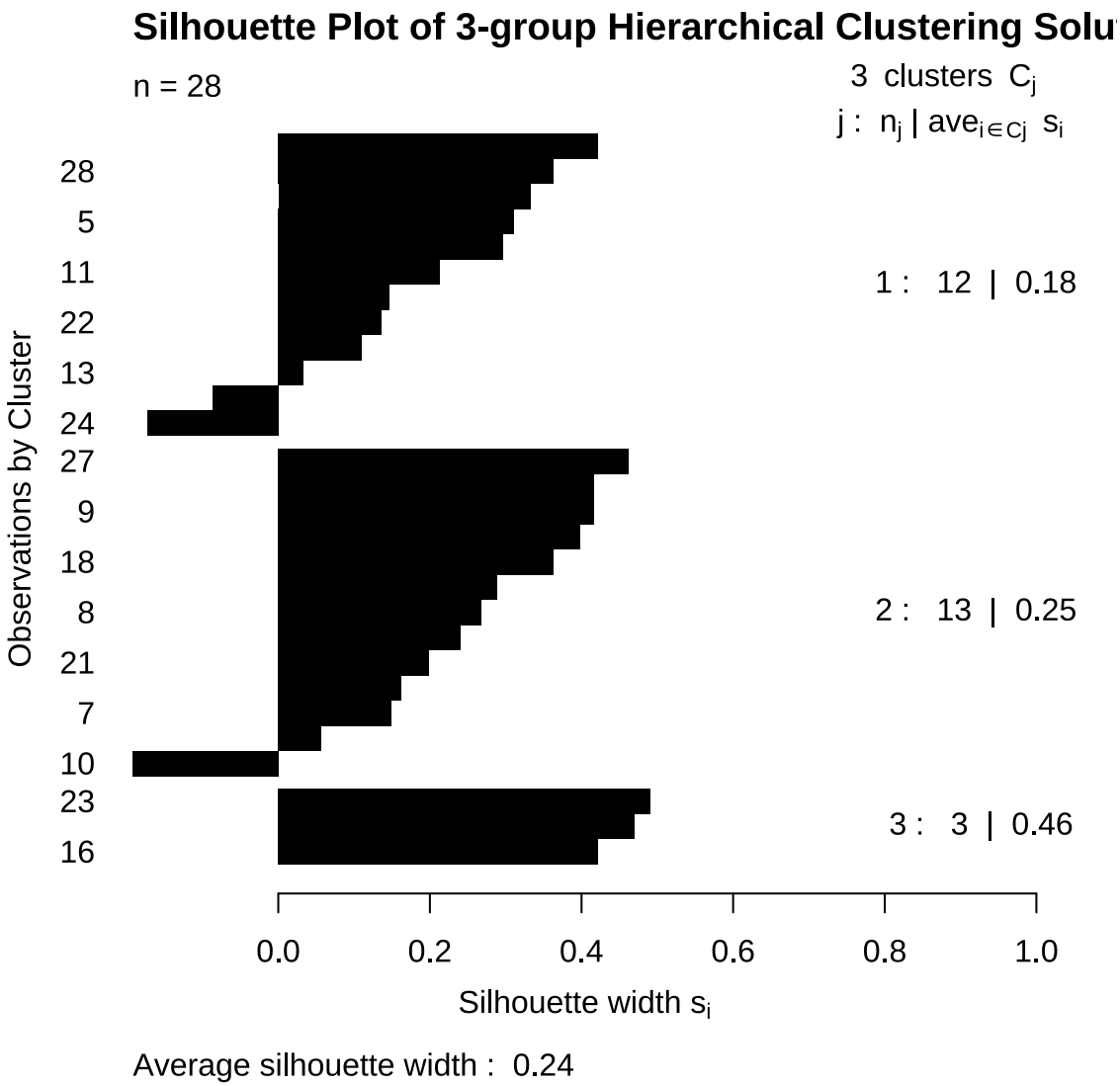
Now we turn to the hierarchical clustering solution.

```r
wardSil <- silhouette(wardMajor, major.dist)

summary(wardSil)
```

```
## Silhouette of 28 units in 3 clusters from silhouette.default(x = wardMajor, dist = major.dist) :
##  Cluster sizes and average silhouette widths:
##       12       13        3
## 0.175282 0.248366 0.461227
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -0.192   0.144   0.279   0.240   0.403   0.491
```

```r
plot(wardSil, col = "black", main="Silhouette Plot of 3-group Hierarchical Clustering Solution",
     ylab = "Observations by Cluster")
```




```
# get silhouette coefficient
summary(wardSil)$avg.width
```

```
## [1] 0.23985
```

The hierarchical clustering solution with 3 groups yields clusters of size 3, 12, and 13. The cluster with 3 points has the highest silhouette coefficient of 0.56. While this indicates moderately strong structure, the cluster is only made up of 3 points so this is not very significant (could just be due to chance). The presence of two negative silhouette coefficients indicates that two observations are more similar to another cluster than their own cluster. While the overall silhouette coefficient for hierarchical clustering (0.292) is slightly larger than the coefficient for k-means (0.284), the presence of a negative silhouette coefficient indicates that the k-means solution may be stronger overall.

Thus, we will proceed with the k-means clustering solution and run a PCA to visualize our solution.

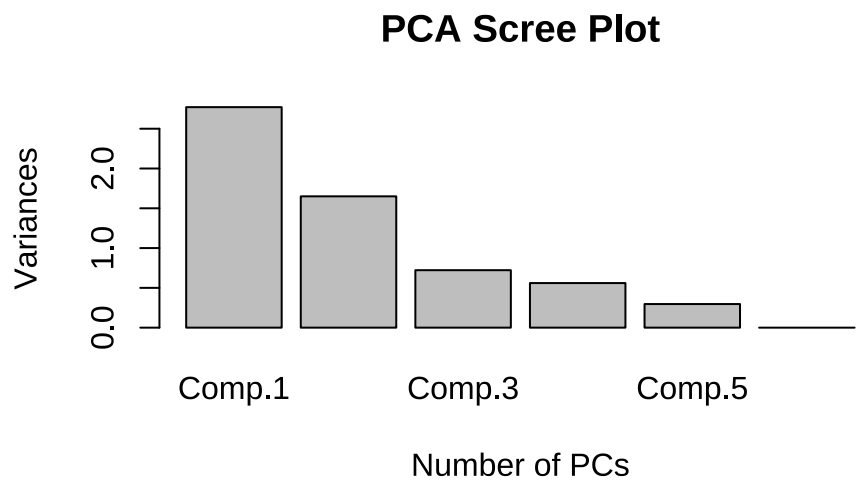
```r
myPCA <- princomp(clust_df, cor = TRUE, scores = TRUE)
summary(myPCA)
```

```
## Importance of components:
##                           Comp.1   Comp.2   Comp.3    Comp.4    Comp.5
## Standard deviation     1.664666 1.284523 0.849769 0.7485390 0.5444878
## Proportion of Variance 0.461852 0.275000 0.120351 0.0933851 0.0494112
## Cumulative Proportion  0.461852 0.736852 0.857204 0.9505886 0.9999998
##                            Comp.6
## Standard deviation     1.15198e-03
## Proportion of Variance 2.21177e-07
## Cumulative Proportion  1.00000e+00
```

```r
loadings(myPCA)
```

```
##
## Loadings:
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## male             0.515  0.120  0.144  0.260  0.795
## nonresident      0.371 -0.427  0.520  0.370 -0.391  0.346
## hispanic_latino -0.505         0.516 -0.338  0.348  0.492
## asian            0.341 -0.474 -0.478 -0.498  0.101  0.415
## black           -0.433 -0.240 -0.435  0.655  0.181  0.323
## white            0.197  0.721 -0.158        -0.227  0.602
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

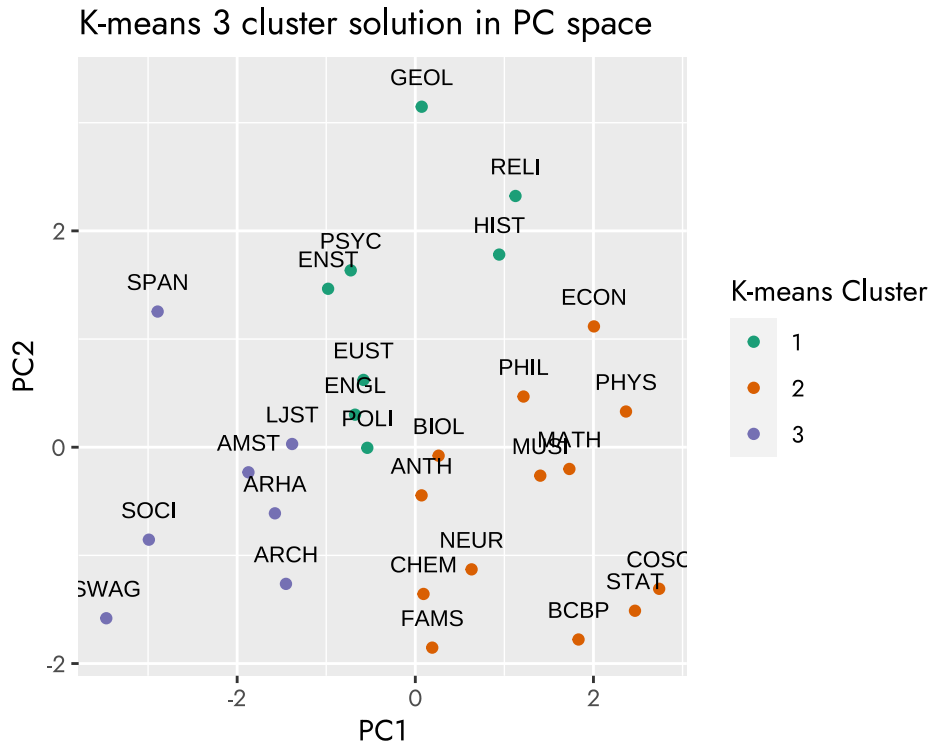
```r
plot(myPCA, main = "PCA Scree Plot",
     xlab = "Number of PCs")
```



PCs 1-2 account for 76% of the original variation in the original data set. The first 2 components capture a significant portion of the variance so we will proceed to plot our solution.

```
df5 <- clust_df %>%
  mutate(PC1 = myPCA$scores[ ,1],
         PC2 = myPCA$scores[ ,2]) %>%
  mutate(wardMajor = factor(wardMajor))

ggplot(aes(y = PC2, x = PC1), data = df5) +
  geom_point(aes(color = factor(Ksol3$cluster))) +
  scale_color_brewer(palette = 'Dark2') +
  labs(title = "K-means 3 cluster solution in PC space",
       color = "K-means Cluster") +
  geom_text(label = df2$code, size = 3, vjust = 0, nudge_y = .2) +
  theme(text = element_text(family = "jost"))
```



While the clusters do not look very distinct in terms of distance, the clusters still contain some interesting groupings. We note that cluster 2 contains every STEM major as well as ANTH, FAMS, MUSI, PHIL, and PHYS. Clusters 2 and 3 contain all humanities majors.

We can examine the centers of the k-means clusters to understand the specific feature qualities (average variable values) that characterize each cluster.

```
Ksol3$centers
```

```
##         male nonresident hispanic_latino     asian      black     white
## 1  0.0303106   -0.491575       -0.195947 -0.879370  0.0124806  1.041133
## 2  0.5327537    0.667708       -0.554117  0.820744 -0.5580793 -0.196423
## 3 -1.0240404   -0.678230        1.253013 -0.519245  1.0221694 -0.825082
```

We recall that cluster 2 (size 13) had the highest silhouette coefficient of 0.32. The points at the center of this cluster have high values of asian, nonresident, and male, and low values of hispanic_latino and black. This cluster contains all of the STEM majors.

Cluster 3 (size 8) had the second highest silhouette coefficient of 0.27. The points at the center of this cluster have high values of hispanic_latino and black, and low values of male, white, and nonresident. This cluster is essentially the opposite of cluster 2.

Cluster 1 (size 7) had the lowest silhouette coefficient of 0.23. The points at the center of this cluster have high values of white and low values for asian.

When interpreting these results, it is important to remember that all majors were weighted equally in the clustering analysis despite significant variation in major size. This means that the degree of separation exhibited by the clustering solution is based on the *average major*, not the *average student*.

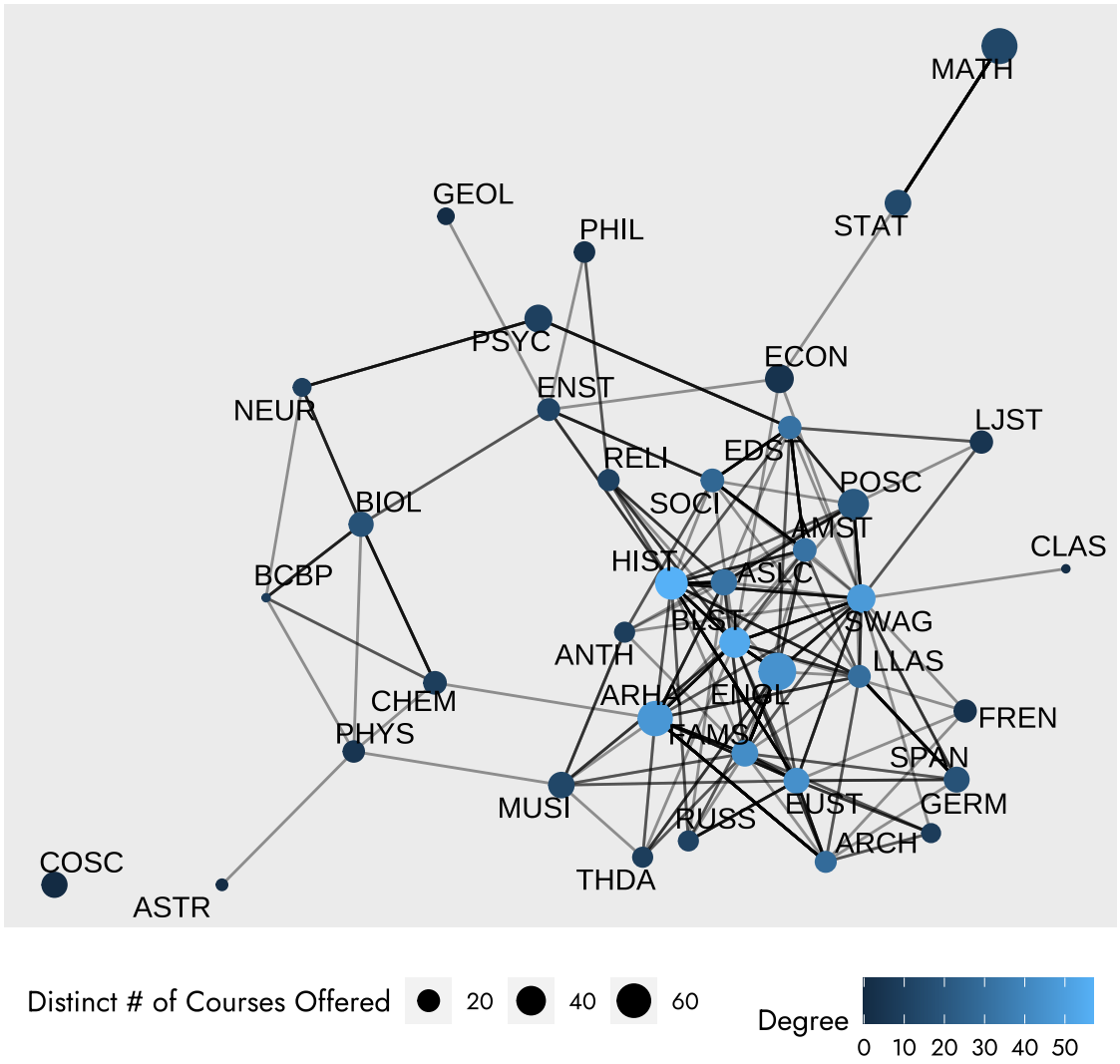
Now we will move onto the results for our network analysis.

Network

```
g <- graph_from_data_frame(d = edge_list, vertices = node_list, directed = FALSE)
```

```
ggraph(create_layout(g, layout = 'stress')) +
  geom_edge_link(alpha = .4) +
  geom_node_point(aes(size = node_list$num_classes, color = degree(g))) +
  geom_node_text(aes(label = name),  repel = TRUE) +
  labs(title = "Network of Majors Connected by Cross-Listings",
       size = 'Distinct # of Courses Offered', color = 'Degree') +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```

```
## Warning: Using the ‘size‘ aesthetic in this geom was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ in the ‘default_aes‘ field and elsewhere instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



In this network, edge weight corresponds to the number of cross-listings, node size corresponds to the number of distinct courses offered by each major (during 22-23 school year), and node color corresponds to the major's total number of cross-listings.

We can see that many of the languages, arts, cultural studies, and a few social sciences are tightly clustered together. While most of the STEM departments are disconnected from this central cluster, CHEM, BCBP, PHYS, BIOL, and NEUR appear to form their own community. We also note that MATH and STAT share a very strong connection, with STAT-ECON being the only one other edge connecting the majors.

Next we will construct the same graph but color the nodes using the data from the clustering analysis. We note that the IPEDS database did not report information about degrees conferred for EDST or EUST so these nodes will be colored grey (NA).

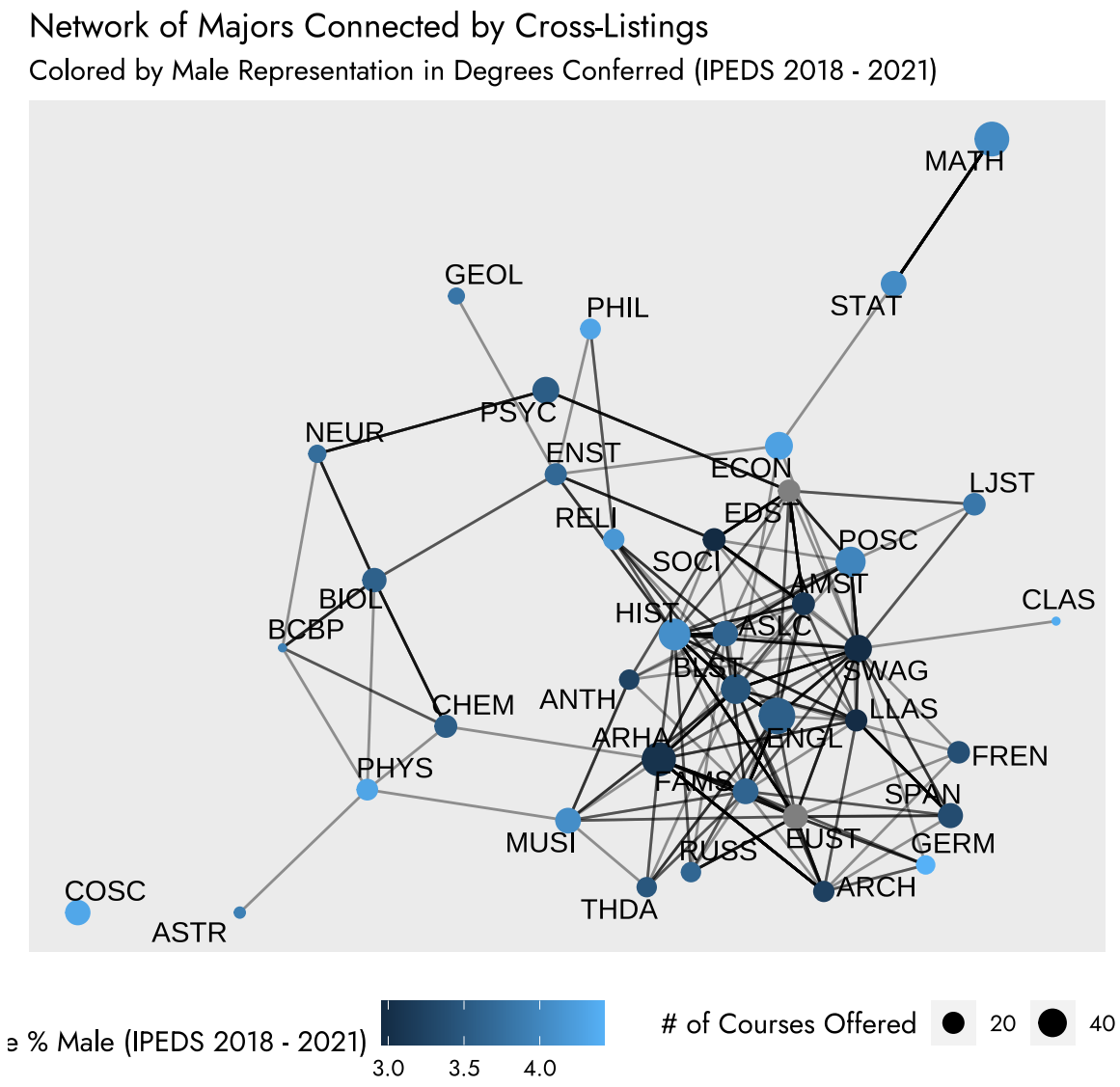
```
# data with ungrouped majors (e.g. ASTR is seperated from PHYS)
ungrouped_df <- ungrouped_df %>%
  filter(degree != 'Total') %>%
  rename(id = degree) %>%
  mutate(id = ifelse(id == 'POLI', 'POSC', id)) %>%
```

```
  filter(id != 'INTR') # filter out interdisciplinary

node_list2 <- node_list %>%
  left_join(ungrouped_df, by = 'id')
```

For the first graph, we color the nodes by the variable, male. This means brighter nodes represent majors where degrees are disproportionately awarded to white students.

```
ggraph(create_layout(g, layout = 'stress')) +
  geom_edge_link(alpha = .4) +
  geom_node_point(aes(size = node_list$num_classes, color = log(node_list2$male))) +
  geom_node_text(aes(label = name),  repel = TRUE) +
  labs(title = "Network of Majors Connected by Cross-Listings",
       subtitle = "Colored by Male Representation in Degrees Conferred (IPEDS 2018 - 2021)",
       size = '# of Courses Offered', color = 'Average % Male (IPEDS 2018 - 2021)') +
  theme(legend.position = "bottom", text = element_text(family = "jost"))
```



The majors that are highly represented by men appear to be many of the least cross-listed majors. For example, COSC, PHIL, ECON, PHYS, ASTR, MATH, STAT, and CLAS all appear to be over represented by men and make up some of the most isolated nodes in the graph.

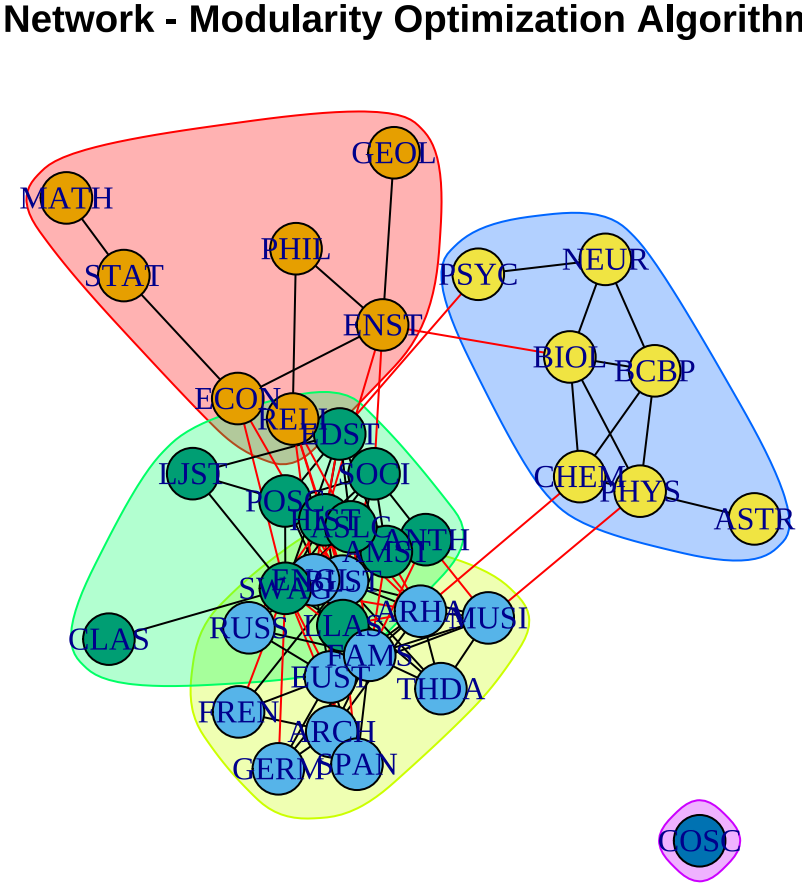
Now we will analyze the community structure of the networks.

```
network <- simplify(g) # simplify graph to remove multi-edges and loops
E(g)$weight <- count.multiple(g) # extract edge weights

fg <- cluster_fast_greedy(network, weights = E(g)$width)
sizes(fg)
```

```
## Community sizes
##  1  2  3  4  5
##  7 12 11  7  1
```

```
plot(fg, network, main = "Network - Modularity Optimization Algorithm")
```



```
#plot_dendrogram(fg, main = 'Major Network Cluster Dendrogram', ylab = 'Height')
```

The solution contains 7 communities. The first three are relatively large, containing 11, 13, and 7 majors, respectively. Communities 4-6 contain pairs of similar majors, including SPAN + LLAS, STAT + MATH, and ARCH + ARHA. It makes sense that these pairs would be highly cross-listed. The last community contains COSC, which is not cross-listed at all.

```
modularity(network, fg$membership)
```

```
## [1] 0.301671
```

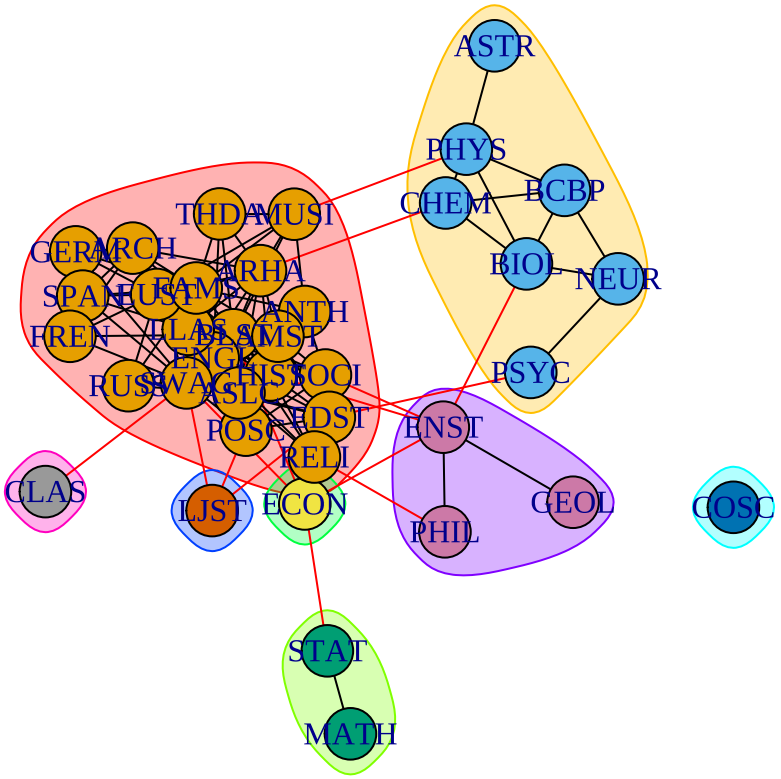
This solution has a modularity of 0.155, indicating that the communities detected from cross-listings exhibit a fairly weak structure.

Now we will apply the Girvan-Newman edge betweenness algorithm.

```
eb <- cluster_edge_betweenness(network, weights = E(g)$width)
sizes(eb)
```

```
## Community sizes
##  1  2  3  4  5  6  7  8
## 22  7  2  1  1  1  3  1
```

```
plot(eb, network, main = "Network - Edge Betweenness Algorithm")
```




```
#plot_dendrogram(eb, main = 'Major Network Cluster Dendrogram', ylab = 'Height')
```

This solution is made up of only 4 communities. The first community is by far the largest with 24 nodes. This cluster seems to contain almost all of the humanities majors. Communities 2 and 3 are smaller and similarly sized with 6 and 7 nodes, respectively. Community 2 contains mostly STEM majors that include the biological and physical sciences: BIOL, BCBP, CHEM, NEUR, PSYC, ASTR, PHYS. Students who are interested in medicine or engineering often major in these fields. The third community contains mostly math/logic related majors (STAT, MATH, ECON, PHYS) and earth science related majors (ENST, GEOL). The last community contains the lone major, COSC.

```
modularity(network, eb$membership)
```

```
## [1] 0.191142
```

This solution has a modularity of 0.210, indicating that the communities detected from cross-listings exhibit a moderately strong structure.

The Girvan-Newman edge betweenness solution yields a slightly larger modularity score than the modularity optimization solution so we conclude that the Girvan-Newman edge betweenness solution is the stronger solution.

Conclusion

The first part of this report used a clustering method to identify groups of majors at Amherst College that are demographically similar. The clustering analysis was run on the 6 quantitative variables, male, non-resident (international student), asian, hispanic/latino, black, and white. After comparing k-means and hierarchical clustering techniques, the 3-group k-means solution produced the strongest solution. Although the structure of the clusters were relatively weak overall (silhouette coefficient = 0.28), the solution still grouped all STEM majors in a single cluster (silhouette coefficient = 0.32). The majors at the center of the STEM-dominated cluster were characterized by high asian, non-resident (international) and male representation, as well as low hispanic/latino and black representation. When interpreting this result, it is important to consider that all majors were weighted equally in the clustering analysis despite significant variation in major size. This means that the degree of separation exhibited by the clustering solution is based on the *average major*, not the *average student*.

The second portion of this report used a network model to explore the relationship between majors that had cross-listed courses in the 22-23 Amherst Course Catalog. HIST, BLST, SWAG, ARHA, ENGL, FAMS were the most frequently cross-listed majors and COSC was cross-listed zero times, despite being one of the most popular majors at Amherst. After plotting the network, we identified a negative correlation between majors that were highly cross-listed and majors that were highly represented by men. Finally, a community detection algorithm based on edge betweenness detected four clusters with moderately strong structure overall. The first cluster contained 24 humanities majors, the second cluster contained 7 science related majors (BIOL, BCBP, CHEM, NEUR, PSYC, ASTR, PHYS), the third cluster contained mathematical reasoning and earth science related majors (STAT, MATH, ECON, PHIL, GEOL, ENST), and the final cluster singled out COSC.

Though we successfully identified all STEM majors in one cluster based on the demographics of students at Amherst College, the structure exhibited in the clustering solution was fairly weak. The sample size of the data is another important limitation to consider. Although we did not identify any glaring inconsistencies across the 4 years, the relatively small sample constrains the accuracy of our data. The variance in major size (total number of degrees conferred) is also a major limitation. As we demonstrated in the preliminary analysis, majors ranged from an average of 3 degrees to 62 degrees conferred per year. This led us to consolidate several of the smallest majors which exposes the data to bias. It is also important to note that over 10% of observations in the original data set were excluded from the analysis because the observations belonged to demographic groups that each made up less than 5% of the sample population. The data for the network analysis, which only contained the 754 courses offered during Amherst's 22-23 school year, was also limited by its sample size. We also note several major limitations to using demographic data from the clustering analysis to decorate the network: 1) the data sets did not represent the same class years 2) EDST and EUST were not included in the IPEDS data base.

Given that the IPEDS database and the Amherst Course Catalog both have at least two decades worth of data, it would be interesting to extend this analysis to a wider time frame in the future. The IPEDS database also has the same information for most other colleges and universities, so future analyses could also examine how Amherst compares to its peers (especially other liberal arts colleges). It would also be worthwhile to perform a clustering analysis on course-level data (demographic make up of students in each course).

Citations

U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), 2017-2020, Amherst College: Completions. Retrieved from https://nces.ed.gov/ipeds/datacenter/FacsimileView.aspx?surveyNumber=3&unitId=164465&year=2020 on November 28, 2022.

Amherst College Course Catalog: 2022-2023. Retrieved from https://www.amherst.edu/academiclife/college-catalog/2223 on December 7, 2022.