Social Network Analysis: Lab 3

Shreyans Kothari

4/17/2022

1) Describe the social network(s) to me, in terms of how it was collected, what it represents and so forth. Also give me basic topography of the network: the nature of the ties; direction of ties; overall density; and if attributes are with the network, the distribution of the categories and variables of those attributes.

The dataset contains undirected and unweighted Facebook networks for Columbia University students. It is one of many college-level facebook datasets collected for a study published in 2011. The dataset in total contains facebook network data from 100 different universities- I chose to stick with Columbia University mainly due to the sheer size of the complete dataset.

The data was found as a sparse matrix with the following attributes: a student/faculty status flag (i.e., one represents student, and faculty otherwise- which has more gradation in terms of job), gender, major, second major/minor (if applicable), dorm/house, year, and high school. A connection exists if the nodes are connected on facebook, i.e., if the nodes are "friends" on facebook, where the value is 1 if a tie exists and 0 if it does not. The entire Columbia dataset has a total of 11770 nodes. I picked 4 consecutive graduating classes, 2005 to 2008, to reduce the size of the network.

Initially, I removed all majors with less than 10 students and tried to run the Girvan-Newman algorithm. However, due to the size of the network (and the massive number of nodes), the community detection algorithm kept running for hours without any result. Thus, I decided to pick a single major from the network- I chose the major which was coded as 75 in the dataset as it had the most number of students/nodes. Despite a lot of searching, I could not locate a codebook which contained the code for each major. So for the context of this report, we will just call the major "major 75", which is very likely a social science major according to Google.

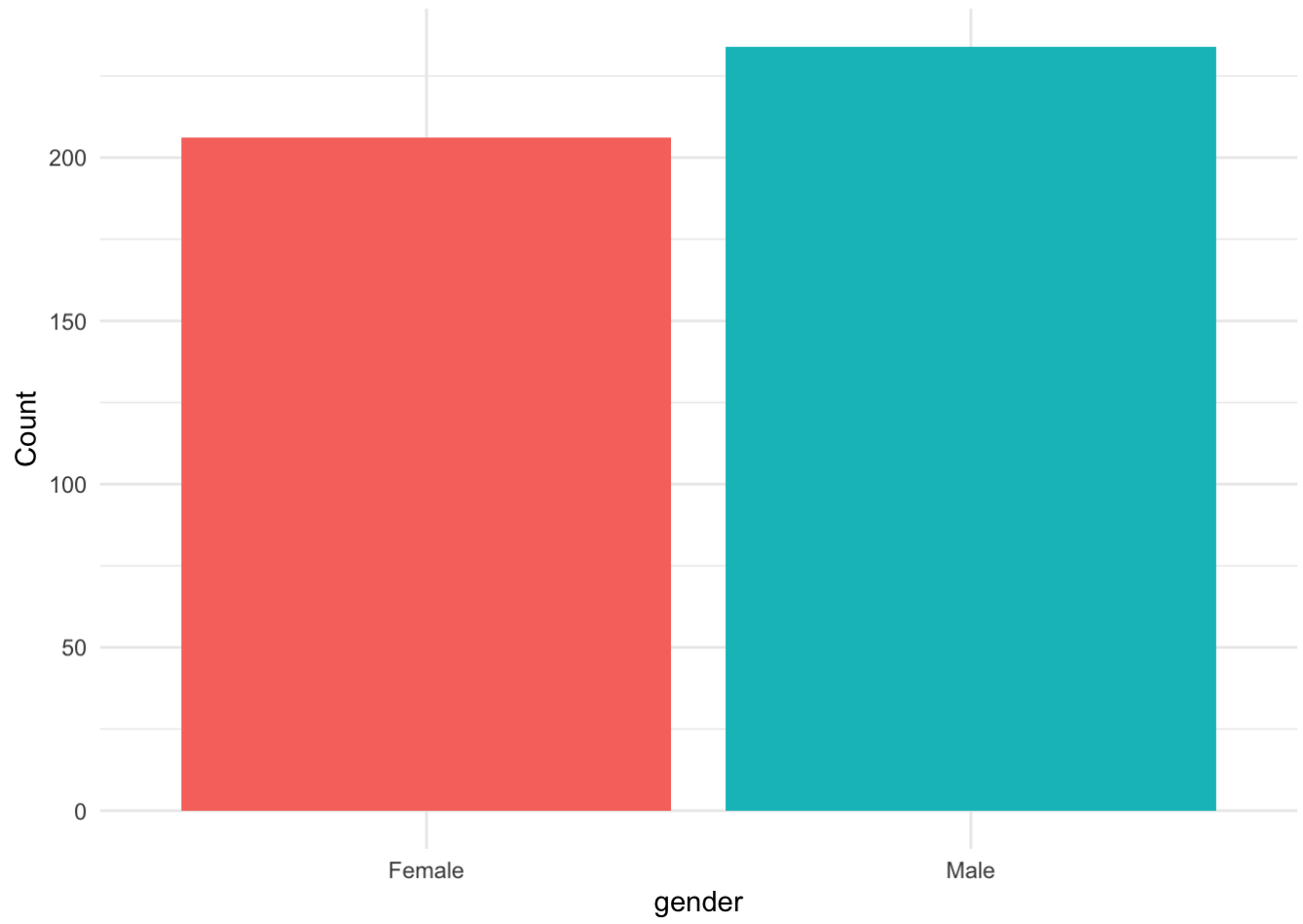
The network density of the facebook friend network for Columbia University students in major 75 is:

```
## [1] 0.02234417
```

The density of the advice network is about 0.022. The range for a network density is between 0 and 1: the density value of 0.022 is towards the lower end of that spectrum. This makes sense because the data contains students from four different years. In any university, you would not expect students to be friends with the entire university, or even their entire year.

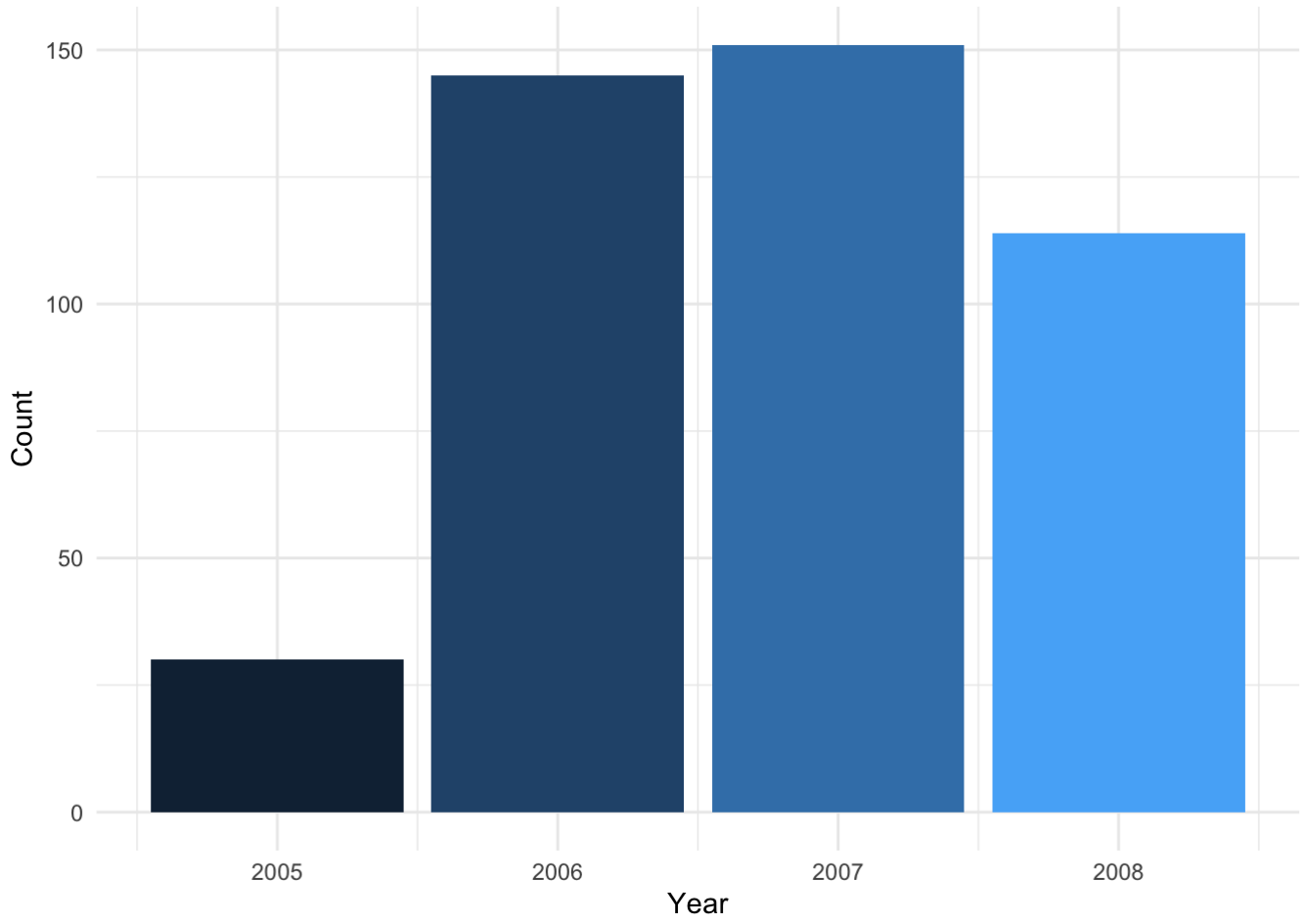
Exploring attributes

Gender



The above histogram compares the number of female Columbia students (coded as 1 in the data) to the number of male Columbia students (coded as 2 in the data). As we can see, major 75 has more males than it does females.

Graduation Year



This histogram visualizes the number of students from each year in major 75. The dataset has the least number of students from 2005. This might be because facebook did not allow students from universities other than Harvard to join until 2004- maybe it took a while for the social media platform to catch on in the other schools. After 2005, 2008 has the lowest number of nodes, followed by 2006. 2007 has the most number of nodes in the dataset.

```
The dataset can be found here: https://archive.org/details/oxford-2005-facebook-matrix
```

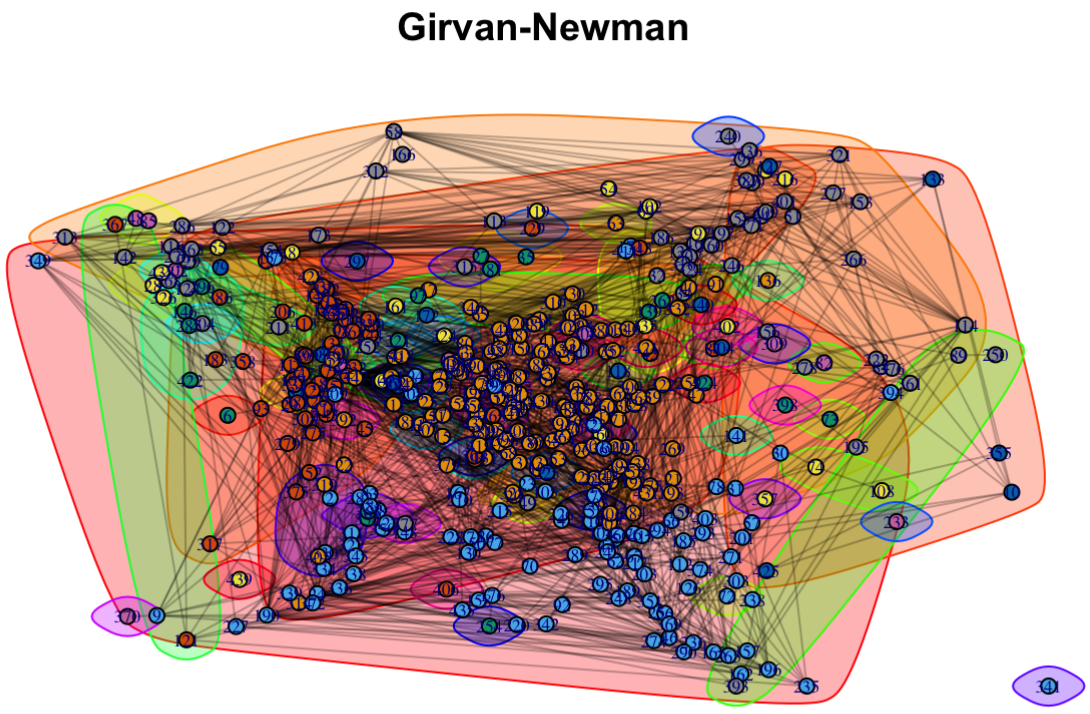
---

2) Run the Girvan-Newman community detection algorithm. Then run the random walk community detection algorithm.

Girvan-Newman community detection algorithm

```
## $`1`
##   [1]   1   3   7  11  15  17  19  21  22  33  36  38  39  40  41  43  47  52
##  [19]  53  59  60  61  65  68  71  75  80  87  90  95  96 100 103 106 107 111
##  [37] 117 128 131 132 137 138 147 149 150 151 164 165 168 170 174 175 177 178
##  [55] 180 181 187 188 192 193 194 198 199 205 212 213 214 219 223 230 231 234
##  [73] 237 239 245 255 257 258 259 264 267 269 270 272 280 281 283 284 287 288
##  [91] 291 292 298 300 301 302 305 306 310 314 318 320 322 325 328 329 331 332
## [109] 337 340 344 346 347 348 352 353 356 359 379 382 384 388 389 390 391 396
## [127] 400 407 412 416 417 418 419 420 426 428 429 430 431 432 433 437
##
## $`2`
##  [1]   2   4   5   9  24  25  26  30  31  44  46  50  56  57  62  66  67  70  76
## [20]  77  78  83  88  92 104 105 110 112 126 130 148 157 158 162 163 167 171 176
## [39] 185 189 190 196 201 203 210 215 217 220 226 227 228 235 236 238 242 248 265
## [58] 271 273 274 282 289 290 293 294 295 303 311 327 330 334 336 338 339 342 345
## [77] 349 350 351 364 366 369 372 373 374 385 394 403 409 411 415 427 435
##
## $`3`
## [1] 6
##
## $`4`
##  [1]   8  14  54  55  69 102 119 135 216 397
##
## $`5`
## [1]  10 127 133 355 425
##
## $`6`
##  [1]  12  18  29  45  86  93  98 109 115 120 134 140 143 144 145 152 169 179 182
## [20] 183 200 206 218 232 241 244 247 251 256 261 275 279 308 317 323 333 335 358
## [39] 360 365 375 380 383 386 399 413 414 434
```

Girvan-Newman Visualization



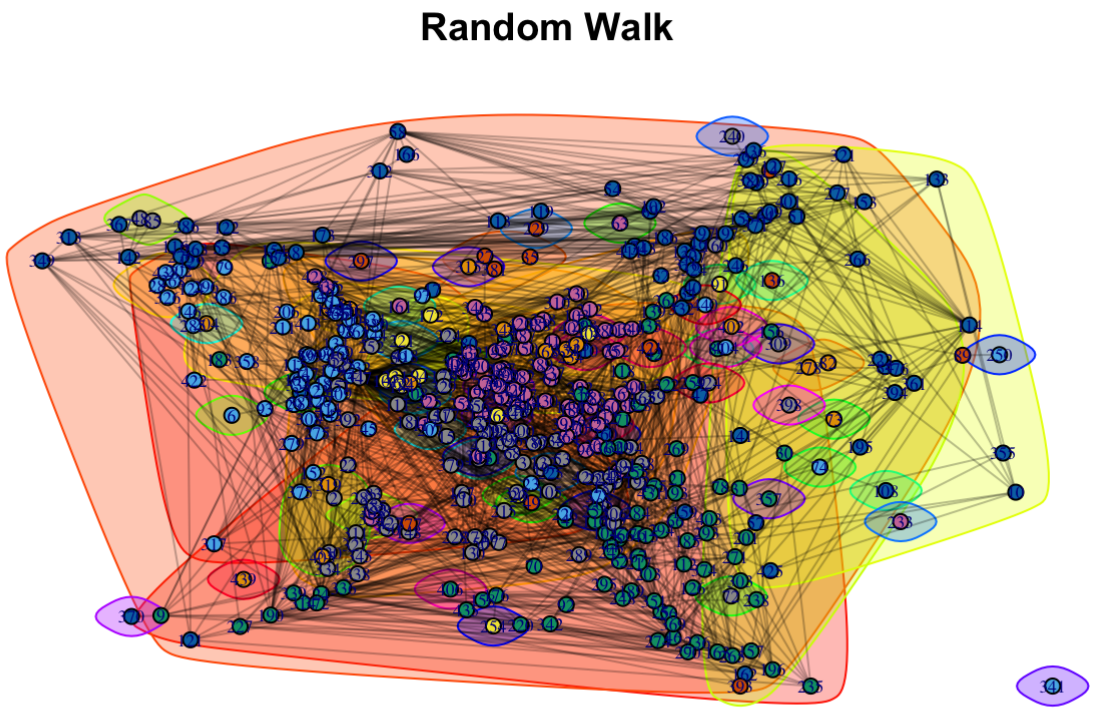
Summary of the Girvan-Newman membership

```
##  gn.membership
##  Min.   : 1.00
##  1st Qu.: 1.00
##  Median : 2.00
##  Mean   :10.47
##  3rd Qu.: 8.00
##  Max.   :77.00
```

Random Walk community detection algorithm

```
## $`1`
## [1]  82 278
##
## $`2`
##  [1]  18  23  29  45  79  86  93  97  98  99 109 115 120 124 134 140 143 146 152
## [20] 169 179 182 200 202 206 211 218 226 232 241 244 247 261 268 275 279 285 296
## [39] 308 317 323 333 358 360 365 375 378 380 383 386 399 409 413 414 418 422 434
##
## $`3`
##  [1]   4   5   9  21  24  25  30  31  42  44  46  52  56  62  66  67  70  71  76
## [20]  78  92 104 110 112 126 128 131 135 156 157 158 163 175 176 185 189 190 193
## [39] 196 201 203 217 220 227 235 236 238 242 248 255 259 265 269 271 274 284 290
## [58] 293 294 295 303 314 320 322 336 339 342 351 364 366 369 372 374 403 415 431
## [77] 435
##
## $`4`
## [1]  94 101 231 362
##
## $`5`
##  [1]   8  16  17  32  34  40  47  51  54  55  57  58  64  69  80  84 102 113 114
## [20] 116 119 121 122 125 129 141 142 153 154 155 159 162 166 171 173 186 195 197
## [39] 209 216 222 246 249 263 266 276 277 286 299 307 312 313 316 321 324 332 343
## [58] 349 361 367 371 373 376 387 394 397 401 411 436
##
## $`6`
## [1]  27  35 381
```

Random Walk Visualization



Summary of the Random Walk membership

```
##  walk.membership
##  Min.   : 1.000
##  1st Qu.: 3.000
##  Median : 7.000
##  Mean   : 9.732
##  3rd Qu.: 8.000
##  Max.   :66.000
```

---

3) Tell me how many groups each algorithm finds. Analyze how similar the two partitioning algorithms are in terms of putting nodes into groups with each other.

The Girvan-Newman algorithm found 77 total groups whereas the Random Walk algorithm found 66. Let's look at the correlation between the membership groups of each algorithm:

```
##                 gn.membership walk.membership
## gn.membership       1.0000000       0.8270279
## walk.membership     0.8270279       1.0000000
```

The correlation coefficient between the two algorithms is positive and pretty high at 0.83; the two algorithm divide the nodes into pretty similar groups.

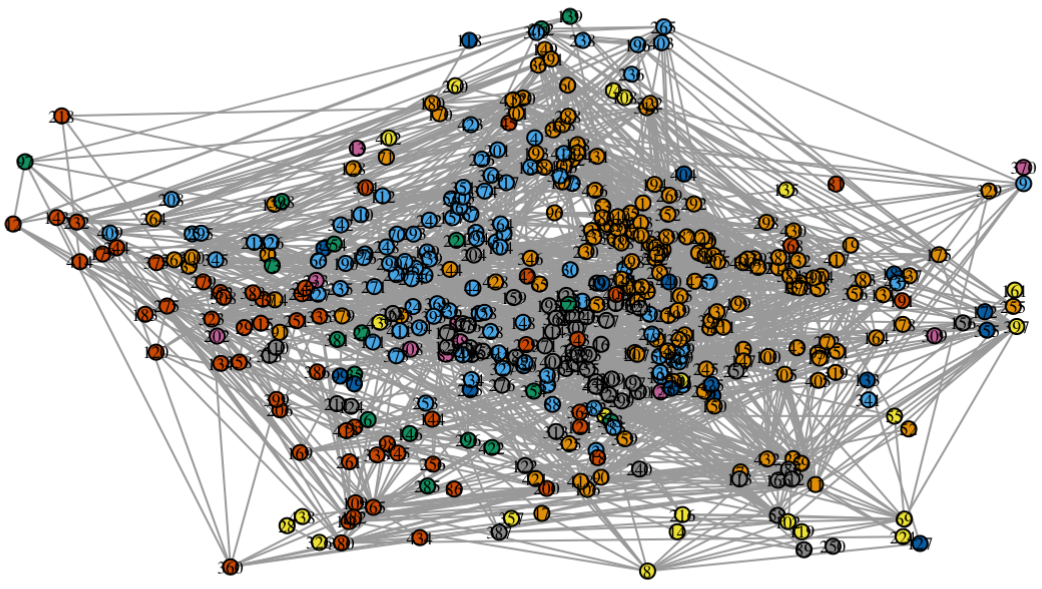
```
## [1] 0.7129782
```

It is not a perfect match but the two algorithms are dividing the nodes into pretty similar groups as the value of 0.713 is pretty high.

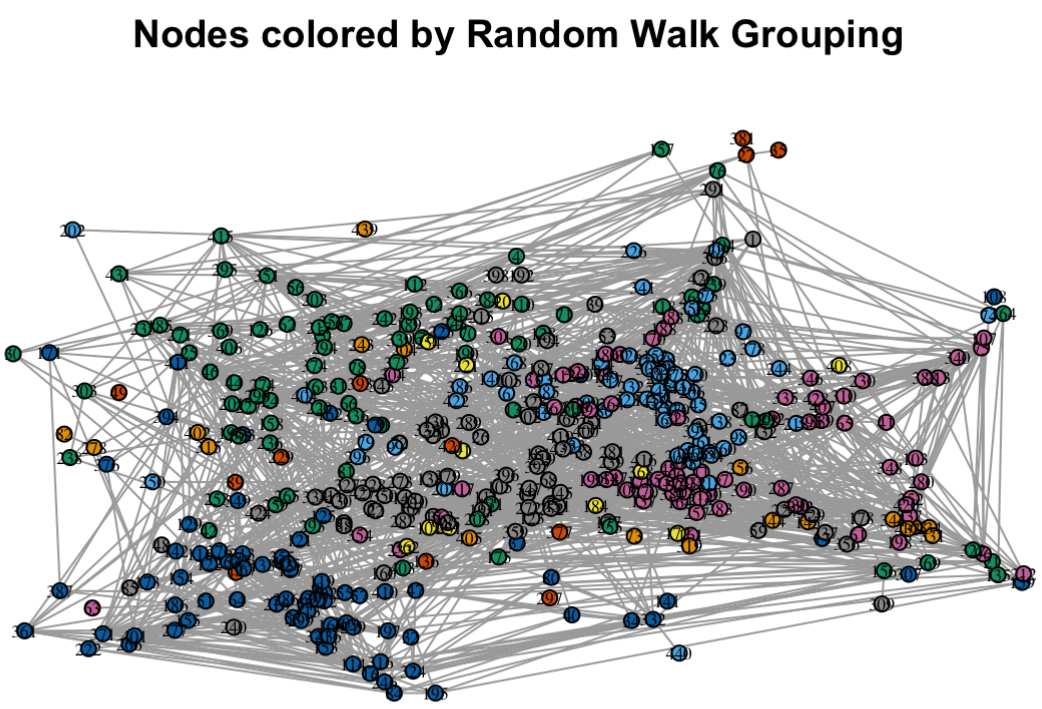
---

4) Visualize the network (either in R or Gephi), coloring the nodes by either Girvan-Newman grouping or the random walk grouping.

Girvan-Newman Colored Nodes:



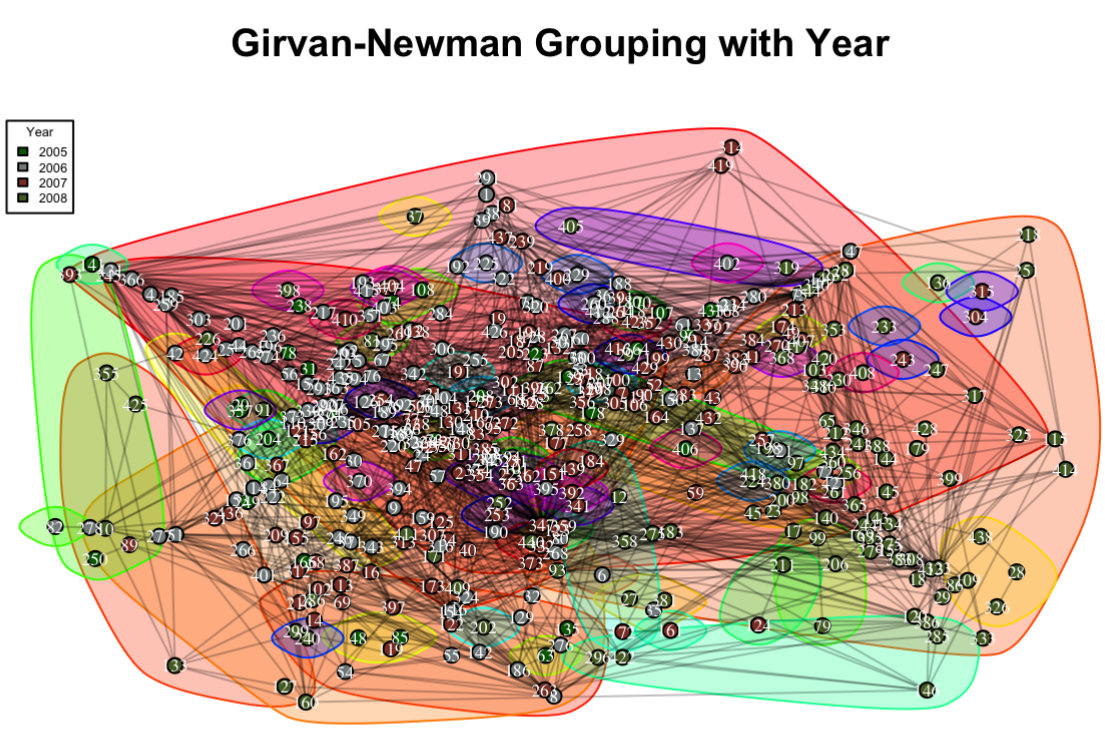
Rnadom Walk Colored Nodes:



5) Tell me anything else about whether the partitioning makes sense, based on attributes or who the nodes are, and so on.

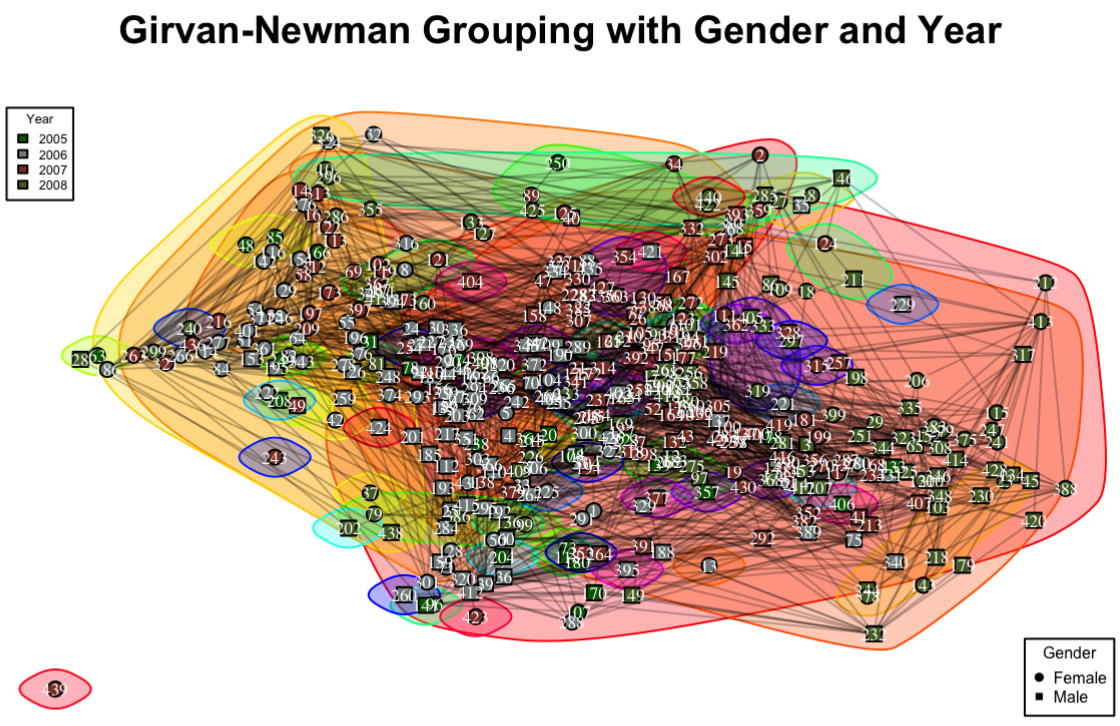
I will explore the attributes and see how the Grivan-Newman community detection algorithm relates to the gender and year of the students/nodes.

Adding color for year:



This visualization adds different colors to the nodes for the year they are graduating. There seem to be some grouping that seems congruent with the year the student is in. For example, most of the students graduating in 2008 seem to be more or less in the same groups. Same with the students graduating in 2006. There definitely seem to be a lot of connections between students across the graduating years. We have the least number of nodes for students graduating in 2005- these nodes seem to not be grouped together based on their year of graduation.

Adding shape for gender:



Based on this visualization, it seems that there seems to be some clustering of the nodes based on their gender, especially in the middle with the most connected nodes. Just by looking at the visualization, it is hard to say whether the gender influences the grouping too much. We see a mix of cross-gender friendship groups as well as same-gender friendships in both small and big groups. There definitely are some bigger groupings that have mostly one gender.
