6.3732 PSet 4 Part 2

Stephen Andrews

April 19, 2025

4.1B:

j:

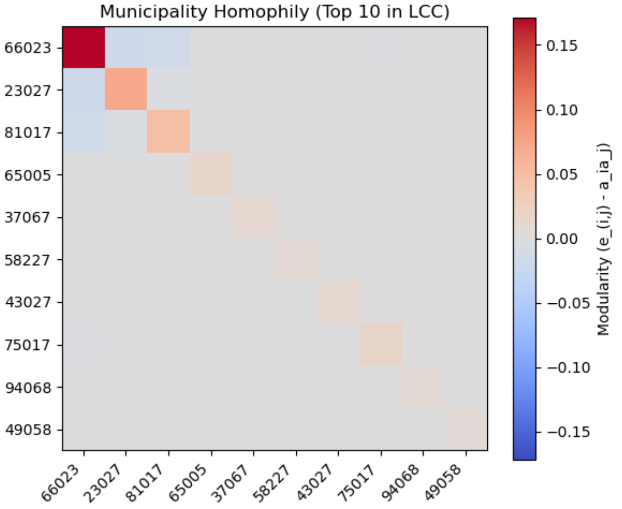
We get the following metrics on the LCC:

1. Edge density: 0.0004
2. Average clustering: 0.4784
3. Global clustering: 0.9460

The extremely low edge density of 0.0004 indicates that barely 0.04% of all possible connections among individuals in the LCC are realized, revealing that most pairs of offenders never collaborate directly. In contrast, the average clustering coefficient of about 0.48 shows that nearly half of each person's neighbors are themselves interconnected, while the global clustering (transitivity) of 0.946 means roughly 95% of all connected triples close into triangles. Together, these show the network is composed of a handful of very tightly knit co-offending clusters that are only sparsely linked to one another, yielding a structure that is locally dense but globally sparse.

k:

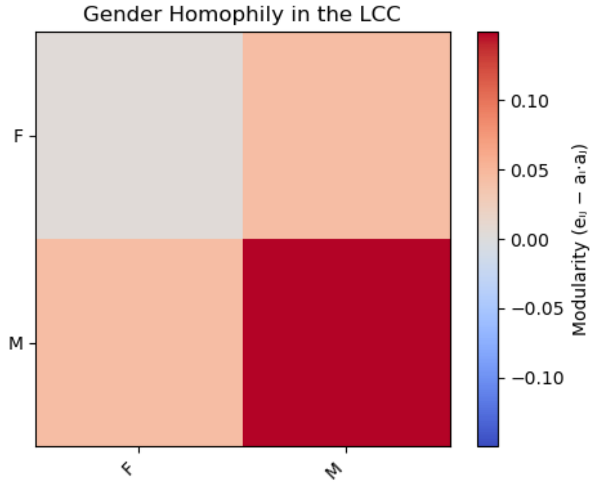
The following diagram shows the homophily matrix for the top 10 municipalities by node count. Note that we focus on the top 10 municipalities by node count in order to keep the matrix concise and readable, highlighting the most prominent groups in the LCC.



The heatmap shows that most municipalities tend to co-offend within themselves rather than with others, as evidenced by the strong positive values along the diagonal—particularly for municipality 66023, where self-ties exceed random expectation by over 15%. A few off-diagonal cells stand out. For example, the light orange between 23027 and 81017 suggests modest cross-municipality collaboration there, and similarly between 65005 and 37067. However, nearly all other off-diagonal entries hover near zero (or slightly negative), indicating that, aside from those few pairs, co-offending relationships between different municipalities occur at roughly the rate expected by chance. In sum, the LCC is composed of tight intra-municipality clusters with only limited cross-municipality edges.

l:

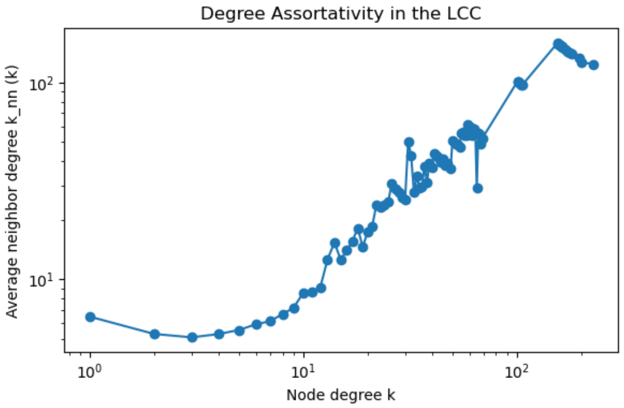
We create the gender homophily matrix as shown below:



Each cell measures how much more (or less) often offenders of gender i co-offend with those of gender j than would be expected at random. Male–male ties are the strongest, with roughly 13% more edges than chance, while female–female ties cluster almost exactly at random expectation (near zero). Interestingly, cross-gender ties also exceed random by about 6%, indicating that male offenders in this component co-offend with females more often than would occur by chance. In short, the network shows strong male homophily alongside a significant level of cross-gender collaboration.

m:

We choose to explore the degree assortativity of the LCC in this part. The following graph shows this:



Degree assortativity measures whether well-connected nodes tend to link to other well-connected nodes (positive assortativity) or to poorly connected ones (negative assortativity). In the plot above, we bin each node by its degree k and compute the mean degree of its neighbors, then display $\langle k_{nn}(k) \rangle$ versus k on log–log axes so that both small and very large degrees can be seen on the same scale. The clear upward trend—where nodes with higher k also have neighbors of higher average degree indicates high co-offenders disproportionately team up with other high co-offenders. The log scale is essential here because node degrees in the LCC span more than two orders of magnitude.

4.2:

a:

Beyond the measures calculated by Mota et al, we could compute the average clustering coefficient, and the centrality of each node. The clustering coefficient can be calculated as $\overline{C} = \frac{1}{N}\sum_i \frac{2T_i}{k_i(k_i-1)}$ where T_i is the number of triangles at node i with degree k_i. This can quantify how "cliquey" each patient's speech is beyond loop counts. We can calculate centrality as $b_i = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$ where σ_{jk} is the number of shortest paths from j to k and $\sigma_{jk}(i)$ is how many of those pass through i. This can be used to find "bridge" words that govern speech transitions.

b:

Repetition of the same word twice in succession corresponds in the speech graph to a self-loop, i.e. an edge from a node back to itself. In the adjacency matrix A, the repetitions are manifested as non-zero diagonal entries. Therefore, one quick test to see if a patient has this behavior is to look at the trace of A. A higher trace would indicate more repeats.

c:

No—you cannot recover each node's in- and out-degree from unigram counts alone. Knowing only how many times each word occurs tells you its total degree but not how many times it was spoken before versus after another word. To reconstruct directed degrees, you need at least the bigram frequencies (counts of consecutive word pairs) or information about which words begin or end the transcript, so you can distinguish incoming from outgoing edges.