Friendship Networks in College

Saurabh Khanna (SUNet ID: khanna90)

Overview

Social integration in diverse societies plays a crucial role in creating and maintaining positive political, economic, and social benefits. Education is an important tool in promoting integration, and colleges in particular are unique melting pots. However, virtually no systematic, large-scale evidence exists about whether college can contribute to social integration. I intend to fill this knowledge gap by looking at friendship networks among colleges students coming from multiple social, cultural, and economic backgrounds. I envision my analysis as consisting of three components. The first component will involve exploratory analysis of college students' friendship networks, where I will consider these networks' degree distributions, path lengths, clustering coefficients, and presence or absence of giant components. The second component will look for motifs and roles driving segregation in college communities. The third component will consider the properties of multiple clusters generated using unsupervised learning approaches.

Related work

The first paper I have considered is 'Identifying the roles of race-based choice and chance in high school friendship network formation' by Currarini et al[1]. The authors specifically consider the way friendships at the high school level play out among students belonging to diverse racial and ethnic groups. The outcome that they are testing here is homophily – which is the tendency among nodes to form connections with other nodes that share the same attributes. Currarini et al. highlight two different mechanisms that can lead to homophily in schools. The first one is the 'bias in preferences' of individuals in who they form friendships with, and the second one is the 'bias in meeting opportunities' they have with individuals from diverse social backgrounds. To this end, they use revealed preference theory to create a parametrized version of the friendship formation model, through which they bifurcate effects along the dimension of preference bias and meeting opportunity bias. Any student's preference bias is represented by a utility function [$U_i(s_i, d_i) = (s_i + \gamma_i d_i)^{\alpha}$]. The parameter γ_i captures the bias in preferences with γ_i less than 1 indicating that different friends are valued less than similar friends (in terms of race). The parameter α captures diminishing returns to friendships overall. The meeting opportunity

[1] Currarini, S., Jackson, M. O., & Pin, P. (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences, 107(*11), 4857-4861.

bias is captured by a metric that measures the 'proportion of each racial group' in the overall school population. The authors find that biases in preferences and biases in meeting rates are both highly significant and differ significantly across races. For instance, 'Asian and Black students are biased toward interacting with their own race at rates >7 times higher than Whites' (p. 4857), whereas Hispanics exhibit an intermediate bias in meeting opportunities. Meeting opportunities are also more biased in larger schools with more than 1000 students enrolled.

One of the primary missing pieces I find in Currarini et al.'s work is the presence of a sound benchmark model. For instance, for the opportunity of meeting metric, the authors simply assume that the meeting opportunities directly vary with the proportion of students from each background in the school population. A better metric would have been to use an Erdos-Renyi random graph model as a benchmark (that could at least confirm or refute the authors' logic around proportional representation). Another issue with the paper is that the authors refrain from considering school-level features (such as policy of assigning students to different classrooms) that might be leading to segregation among students, as opposed to just individual biases and population proportions.

I also considered Elena Grewal's work providing a comprehensive description of the patterns of friendships by socioeconomic status among adolescents[2]. First, she describes how many students form friendships with students from different socioeconomic backgrounds. Second, she describes how segregation into schools, neighborhoods, extra-curricular activities, and course work explains the levels of socioeconomic friendship segregation. Third, she investigates whether socioeconomic status is a significant predictor of friendships when controlling for opportunities for interaction, an approach similar to Currarini. Fourth, she tests variations in the patterns of socioeconomic status friendship segregation, looking at whether students' friendships become more or less diverse as they progress through school, whether diversity varies with the closeness of the friendship, whether there are communities within a school that are more segregated, and whether there are differences between boys and girls in the diversity of friendships formed. Her work uses Erdos-Renyi random graph models to control for a host of student covariates as well as the dependencies between friendship ties. Grewal concludes that students form diverse friendships when in diverse communities, and though there is still friendship segregation by socioeconomic status, friendship segregation by race is much higher. One of the problematic pieces in Grewal's paper is the homophily measure that has been used. Grewal uses a non-level homophily measure which calculates homophily as the ratio of same group friends to the total number of friends. This measure can be problematic as it does not consider the ethnic composition of the classroom a student might be situated in. For instance,

[2] Grewal, E. T. (2013). Exploring socioeconomic friendship segregation in schools.

having 10% friends from a dissimilar group should be given higher weightage in a classroom having very few students from that particular social group.

Given my interest in student integration and segregation, a third paper I considered was Bojanowski and Corten's work on metrics measuring social network segregation[3]. The authors outline the advantages and disadvantages of multiple segregation indices, such as the E-I Index, the Assortativity Coefficient, and the Spectral Segregation Index. I like the authors' argument that two critical aspects for the choice of a particular segregation measure are whether the network ties are assumed to be static or dynamic, and how the measure treats the presence of isolated nodes in the network. At the same time, I find it limiting that the authors suggest only one measure of segregation that can be a good node-level measure, namely the Spectral Segregation Index (p. 27). This is especially limiting because its usage is considered limited to undirected networks.

Objective

Based on my review of the existing literature on assessing segregation among students in educational settings, I find three crucial gaps in knowledge:

1. All existing research in educational social networks applies to K-12 data. There is no representative information on how segregation plays out at the college level.
2. Segregation measures are defined based at institute-level or at department-level using 'umbrella metrics'. Node level network data has not been used.
3. Most of the segregation research does not detail the sub-components of a network or network roles that drive segregation.

I intend to fill this knowledge gap by looking at friendship networks among colleges students coming from multiple social, cultural, and economic backgrounds.

Data and Methods

As part of my project, I aim to proceed along three research dimensions in order to plug the three knowledge gaps mentioned below:

Firstly, I will be applying network analysis methods to social network data collected at the college level. In order to fulfil this objective, I have unique nationally representative, granular, complete social network data collected from 19,542 students at 50 colleges (42 non-elite and 8 elite)[4] in India. These 50 colleges were chosen based on a stratified random sample from the

[3] Bojanowski, M., & Corten, R. (2014). Measuring segregation in social networks. *Social Networks, 39*, 14-32.
[4] An institution is considered as 'elite' if it appears in the top 100 institutions in the National Institutional Ranking Framework – 2017 adopted by the Ministry of Human Resource Development, Government of India, to rank institutions of higher education in India.

population of non-elite and elite Indian engineering colleges. At each college, I have social network data available longitudinally for 2 cohorts of students – I) from the start of Year 1 to the end of Year 2, and ii) from the start of Year 3 to the end of Year 4. As a result, I can examine gains/losses in segregation for both student cohorts at each college. The student cohorts at the first and second time points are considered as the 'baseline' and 'endline' cohorts respectively.

Each student was asked to nominate up to ten friends in her current class (same department and year). Hence, I have a *directed friendship network* for each classroom, for a total of 200 college classrooms. For these college networks, I will be observing the trends in degree distributions, average path lengths, clustering coefficients, and the sizes of connected components.

Secondly, I will be using the Louvain Algorithm to partition each network into non-overlapping segregated communities. Since this is an unsupervised learning approach, I can further analyze each cluster to see which social, cultural, or ethnic attribute (such as race, socioeconomic status, gender, religion, etc.) ends up segregating college students the most. In line with my goal of assessing integration in college among diverse social groups, I use homophily (also called assortativity) as a metric since it measures the tendency of students to befriend students from similar backgrounds. While considering homophily across student categories, I consider two disjoint and exhaustive categories of students for a preliminary analysis – a) Reservation students (i.e. those who gain admission based on affirmative action) and Non-reservation students (i.e. those who gain admission based on merit), b) female and male students, and c) students adhering to different religions.

I calculated this homophily based on Newman's formula on assortative mixing[5]. Newman defined assortativity coefficient *r* as

$$r = \frac{\left(\sum e(i,i) - \sum a(i)b(i)\right)}{\left(1 - \sum a(i)b(i)\right)}$$

where *e(i, j)* is the fraction of edges connecting vertices of type *i* and *j*, *a(i)* is $\sum(e(i, j), j)$, and *b(j)* is $\sum(e(i, j), i)$. *r* ranges between -1 and 1, with a larger value indicating higher assortativity or homophily.

Thirdly, I will be looking for the building blocks in networks that are significantly characterizing and discriminating networks. More specifically, I will be employing the Exact Subgraph Enumeration (ESU) algorithm to first enumerate all motifs of size between 3 and 5, and then count the number of occurrences of each motif type. I can also set up a null model (using degree distribution from a randomly generated graph) to derive the significance of a particular high frequency motif. I can conduct these analyses for networks with both high and low segregation, and see what kinds of motifs drive segregation or integration (and to what extent).

[5] Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, *67*(2), 026126.

The results of my study will give rise to critical policy recommendations for driving integration and reducing segregation among college students having diverse social experiences.

Results

Part 1: Descriptives for college friendship networks

The summary statistics for the 'baseline' (college years 1 and 3) and 'endline' (college years 2 and 4) friendship networks can be seen below:

	Number of nodes (i.e. students)	**Number of edges (i.e. friendships)**	**Average in-degree**	**Average clustering coefficient**
Timepoint – Baseline	17699	115420	6.5213	0.276
Timepoint – Endline	18490	138251	7.4771	0.284

Even though we are looking at the same student networks across time, the number of nodes is different across baseline and endline because a few students drop in/out of college, or switch in/out of departments. Since the average clustering coefficient stays about the same across the 2 time periods, looking at clustering coefficients for each of the 200 classrooms separately might reveal more insights. Further, since the students had been asked to nominate at most 10 friends, outdegree will not be a useful metric on account of an upper bound of 10. This is the reason I am considering and reporting the in-degree distribution for this directed network. The indegree distribution of the baseline and endline graphs is seen in the figures below:

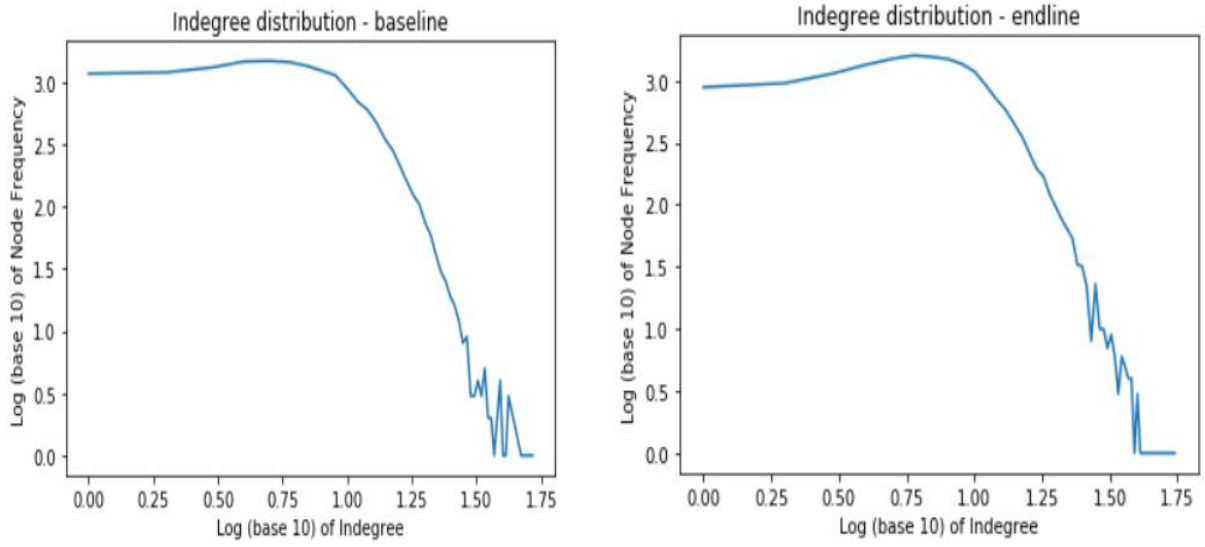

Figure 1: Indegree distribution of friendship networks – baseline (left) and endline (right)

I also considered the variation in distributions of average path length in classroom networks across the 2 time points -- i.e. baseline to endline. Since each classroom is a directed network in itself, the average path lengths were calculated considering the largest weakly connected component for each respective classroom. The path length distributions can be seen below. One clear reflection by looking at these distributions is that the standard deviation in average path length clearly reduces from the baseline to the endline. Since the average classroom path lengths are closer to the mean in the endline, it might mean that the network is evolving into a more consistent one - where the distribution of edges across nodes is more equitable - with time.

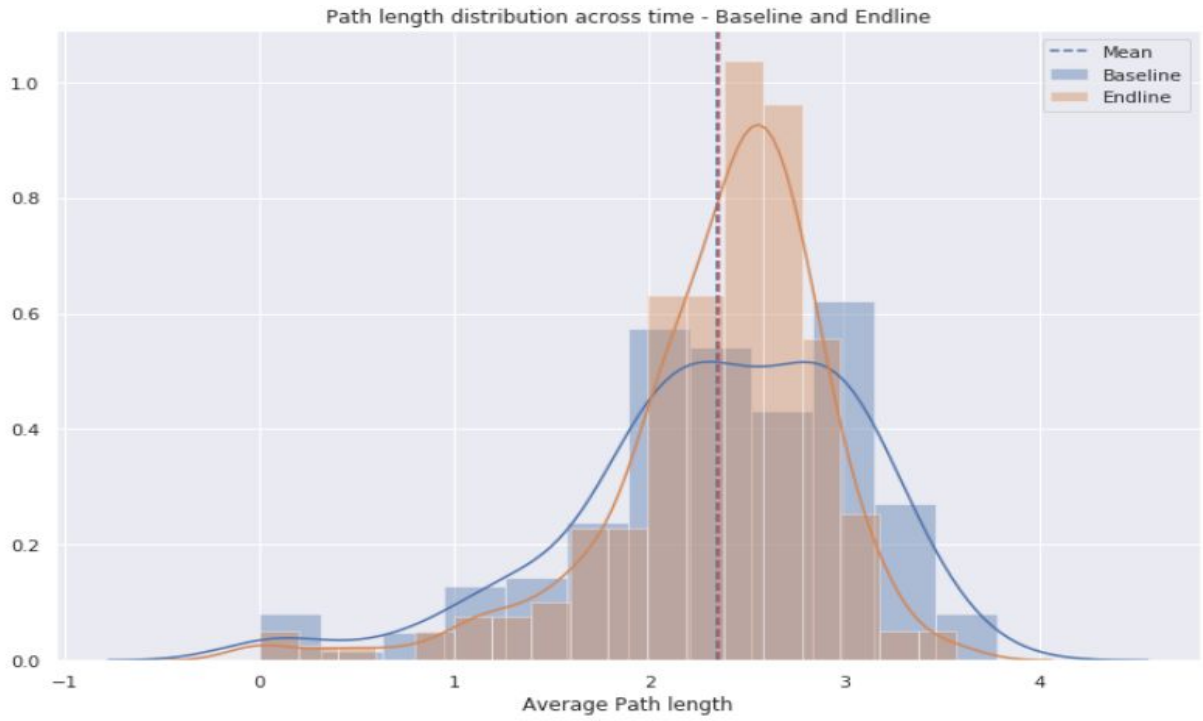

Figure 2: Distribution of average path length of 200 classroom networks - baseline and endline

Part 2a: Homophily across time

Based on Newman's algorithm for assortative mixing, I calculated the homophily (i.e. the assortativity coefficient) for all 200 classrooms. I then identified 2 classrooms – one with the highest homophily, and one with the lowest homophily - say based on a social category like affirmative action received or not. I then drew out networks for both classes across time with the aim of spotlighting any visible trends:

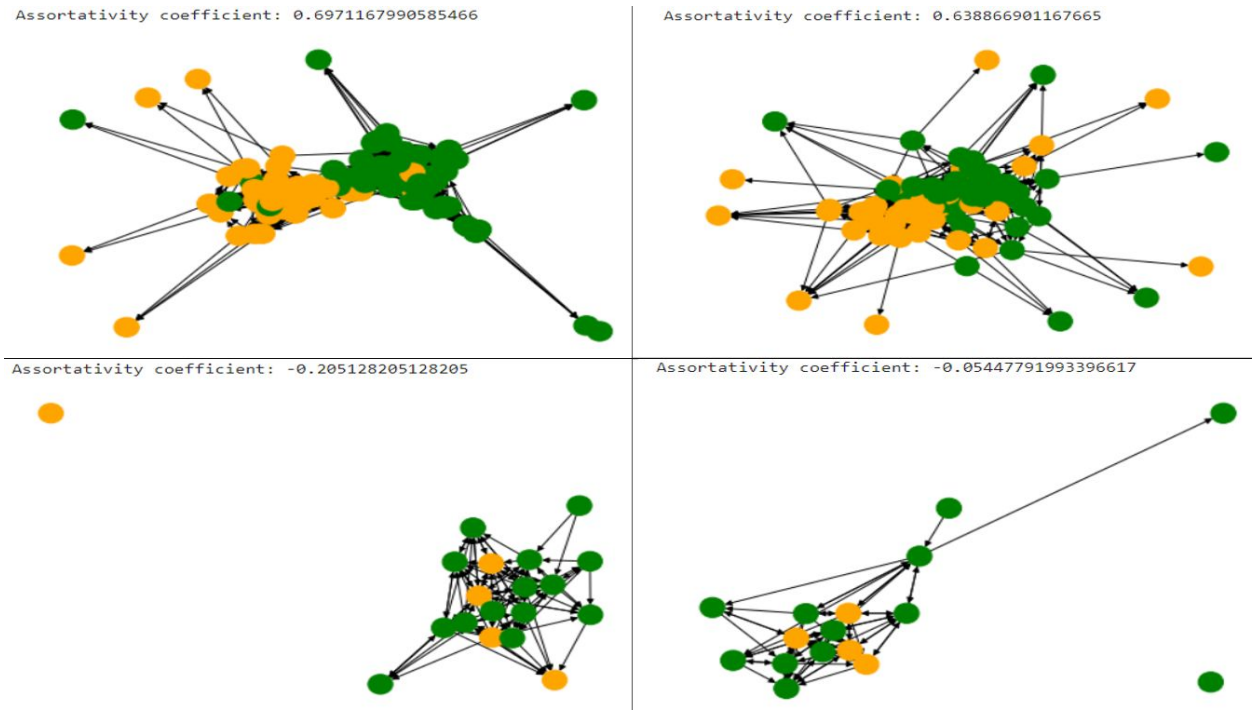

Figure 3: Homophily in classrooms across time – Green is for Reservation students; Orange is for Non-reservation students. Top 2 plots show the baseline (top left) classroom with highest homophily, and its state in endline (top right). Bottom 2 plots show the baseline (bottom left) classroom with lowest homophily, and its state in endline (bottom right). Assortativity coefficients are also reported with each plot.

We see in the figure that the homophily in both classrooms depicted above reduces (albeit slightly) with time, i.e. students start forming friendships across groups as they progress through college. This phenomenon is more evident in the top 2 figures – in the baseline plot (top left), affirmative action (green nodes) and non-affirmative action students (orange nodes) are almost entirely segregated. But in the endline plot, we see that some of the orange nodes have moved across into the green bunch, hence reducing homophily of the overall network.

Part 2b: Homophily based on student characteristics

I also performed clustered homophily calculations based on 3 different student characteristics - affirmative action, gender, and religion. I then mapped the homophily distribution for each characteristics across 200 classrooms at both baseline and endline time points. The interesting result was that segregation in classrooms is strongly driven by gender differences (assortativity coefficient of ~0.6). Religion and affirmative action have a relatively weaker effect (both with assortativity coefficients ~0.1). The patterns were strikingly similar across baseline and endline as well.

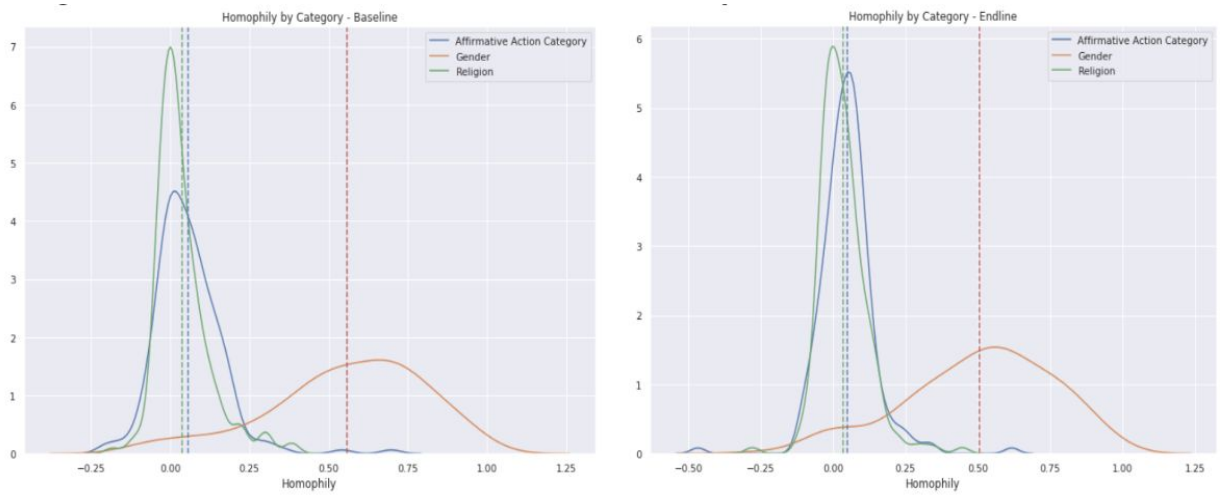

Figure 4: Baseline and endline segregation through homophily based on affirmative action, gender, and religion

Part 3: Motif Detection

After identifying gender as the segregating characteristics across networks, I used Exact Subgraph Enumeration (ESU) to identify motif significance in classroom networks with a) high homophily based on gender, and b) low homophily based on gender. Based on ESU results, it was revealed that a divergent subgraph with 3 nodes and 2 edges has a very low z-score for networks with high homophilic segregation. As a result, this motif signifies reduced gender segregation across both baseline and endline networks, effectively acting as an 'anti-motif'.

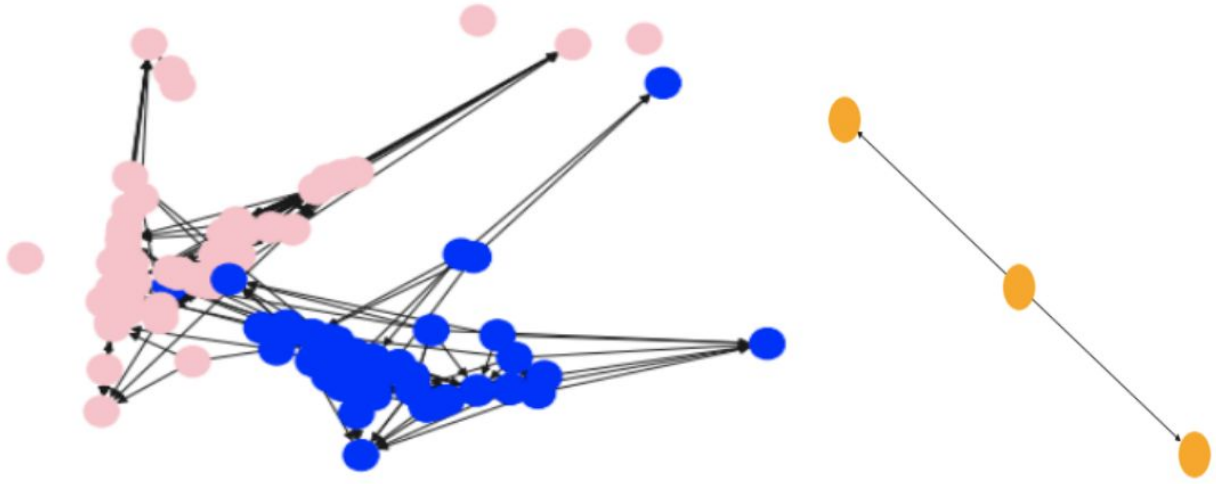

Figure 5: A network with high segregation by gender (male nodes: blue; female nodes: pink) and its divergent anti-motif

Implications

There are three clear implications from my analysis. Firstly, gender is revealed as the primary social category responsible for segregating students. College level efforts should be focused on promoting integration through activities that cut across gender boundaries and biases. Secondly, encouraging classroom groups of three students with one assigned task leader might help reduce segregation (as is visible from the motif analysis). Finally, future research can outline the role of larger motifs and anti-motifs towards reducing classroom segregation.

References

Bojanowski, M., & Corten, R. (2014). Measuring segregation in social networks. *Social Networks, 39*, 14-32.

Currarini, S., Jackson, M. O., & Pin, P. (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences, 107*(11), 4857-4861.

Grewal, E. T. (2013). *Exploring socioeconomic friendship segregation in schools*.

Grochow, J. A., & Kellis, M. (2007, April). Network motif discovery using subgraph enumeration and symmetry-breaking. *In Annual International Conference on Research in Computational Molecular Biology* (pp. 92-106). Springer, Berlin, Heidelberg.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E, 67*(2), 026126.