**Introduction**

This project explores and analyzes grades and collaboration from a computer science class at Yale to better understand the internal dynamics of the class. There are two major themes will be investigated. One, this project aims to better understand collaboration, describe and catalogue how it operates in this context, and understand its effects on grades and learning. Two, this project will analyze the gender dynamics of the collaboration and grades to better understand if and how different genders are granted opportunities unequally. A clearer understanding of the impact on collaboration has the potential to make classes more data-driven and fairer.

Collaboration is an important aspect of STEM classes for several reasons. In STEM classes, assignments typically have a small range of acceptable answers that every student must reach to obtain full credit. This contrasts with humanities and social science classes where every student’s essay will likely be distinct or plagiarism will be much clearer. Additionally, problems may require a key insight which is easily communicated with other people. Furthermore, group work and brainstorming may lead more quickly to the correct answer than anyone working in isolation. Finally, collaboration’s importance is enhanced by the relatively large portion of final grades that depend on homework.

The class studied is a particularly good example of the importance of collaboration. The class has a reputation for being hard, the class has many students, and the homework makes up a significant portion of the final grade. In fact, many of the student reviews of the class mention the importance of collaboration. One student warned that “as long as you find a group to collaborate with, it’s a doable class” (Anonymous 2015). This heightened importance will make this class especially ripe for analysis.

There is a wide body of research cataloguing the differences in career choices and trajectories between men and women, especially in science and technology fields. Two overarching theories have been posited to explain these patterns: The deficit model and the difference model (Sonnert and Holton 1996). The deficit model emphasizes structural obstacles while the difference model focuses more on internal goals and behavior (Sonnert and Holton 1996). My research investigates if collaboration is one of these structural barriers preventing women from succeeding in science and technology fields. In fact, research has found that classroom climate, including group working conditions, may reduce women’s confidence, especially in male-dominated fields (Hall 1982).

Computer Science has become an overwhelmingly male field. Since the middle of the 1980s, the percentage of women working in computer-science related professions has steadily decreased. Women have earned only about 15% of the undergraduate degrees in computer science in recent years despite earning 37% in 1985 (Women in Computer Science 2017). While there is evidence that as much as 82% of the gap in STEM bachelor’s degrees is attributable to high school or before (Legewie and DiPrete Pathways 2014), the college experience is still worth studying. Many initiatives including “Girls Who Code” and “NASA Women” have attempted to combat the dearth of women in computer science by increasing women’s opportunity, participation, and interest. This essay aims to help understand and explain these trends by examining granular data from a computer science class. Focusing on the in-class dynamics that most organizations do not have access to, can expose barriers that women face inside the classroom that may be unrelated to external factors.

This project will explore these two themes to deepen the understanding of collaboration in courses.

NEED ONE MORE PARAGRAPH THAT TIES IT ALL TOGETHER

**The data**

The data consists of grades on homework and tests and self-reported collaboration on homework by students in an upper-level computer science class at Yale University called “Design and Analysis of Algorithms.” There is data for one year included in the analysis.

There were seven problem sets assigned as homework for the course. Grades on homework consisted of three problems, each of which had a max score of 10. Thus, each homework was out of 30 points.

There were also two in-class tests. These tests were entirely individual. Additionally, these tests did not pose new questions but required students to solve a problem that appeared either in class notes or on a problem set. Tests also had three questions, worth 20 points each, for a total of 60. I then scaled the test scores to be on the same range as the homework.

Information on collaboration was collected by the professor of the class after each homework assignment via an online form. On the form was a list of all the names of students in the class. The professor instructed the students to identify whom they had received help from, including the option to state that you did not receive help from any other students. Though the forms were not graded, the professor strongly encouraged providing collaboration data by stating that not filling out the data was a violation of academic honesty. Multiple emails were sent out. Students were only allowed to collaborate with up to three other students but there are rare instances of students not following this rule. This happened only seven times all semester. Admitting to this rule-breaking suggests that students were generally truthful when completing the form, even when doing so could have gotten them in trouble.

Some students may not have filled out the forms. It is also possible that these students did not collaborate with anyone and thus had no need to fill them out. This is supported by the fact that 72% of the students who list no collaborators also have no students list them as collaborators and the remaining 28% may have just had unidirectional edges. Nevertheless, this is a potential limitation and source of error.

To turn the data into network data the following transformations were made. Each student became a node in the graph. The edges represent directed collaboration ties. When student A lists student B as a collaborator, there exists an edge B🡺A. While many of the edges are reciprocal (meaning A🡺B and B🡺A both exist), it is possible for only one of these to be in the graph.

Some students also dropped the class after initially enrolling. The official drop deadline coincided with the due date of the fourth problem set. Thus, I will consider that any students who received no points this date to have dropped the class. After the drop deadline, the class had 86 men, 23 women, and 1 non-binary student meaning the class is heavily male.

To protect the privacy of the students and to comply with regulations, students were assigned random identifiers. Ensuring the privacy was made more complicated by the fact that I am a member of the network in the first year. My grades and collaborations were removed from the data set. This a source of error as removing me changes the network, but is a necessary step to ensure privacy. To further safeguard students’ identities, the grades on homework and tests were perturbed in a manner that was kept hidden from me. These grades will be taken as given throughout the rest of the paper, but this perturbation is a potential source of error.

There are other potential limitations of the data set. Students could have completed the form dishonestly to prevent the professor from knowing if they had collaborated with more than the maximum three students. Though there are instances where students listed collaboration with more than three students, students still may have underreported their actual collaboration. There was one example of students listing every student as a collaborator. This behavior indicates students did not always take the online forms seriously. These ties were removed from the data set. Furthermore, there was a single instance of student listing a student who had dropped as a collaborator. This is treated as a mistake and removed from the data set. Additionally, there are instances of students filling out the surveys multiple times. It appears that this may have been done to correct mistakes. At the advice of the professor, I will use students last response to the form. Finally, there are many other attributes that are potentially important that we do not have access to. These include overall GPA, class year, major, and whether individuals were taking the class pass/fail. These could potentially be confounding variables. Despite these potential problems, the data is rich enough and robust enough to support analysis.

ADD MORE LIMITATIONS AND STUFF

**Imputation of excused absences on tests**

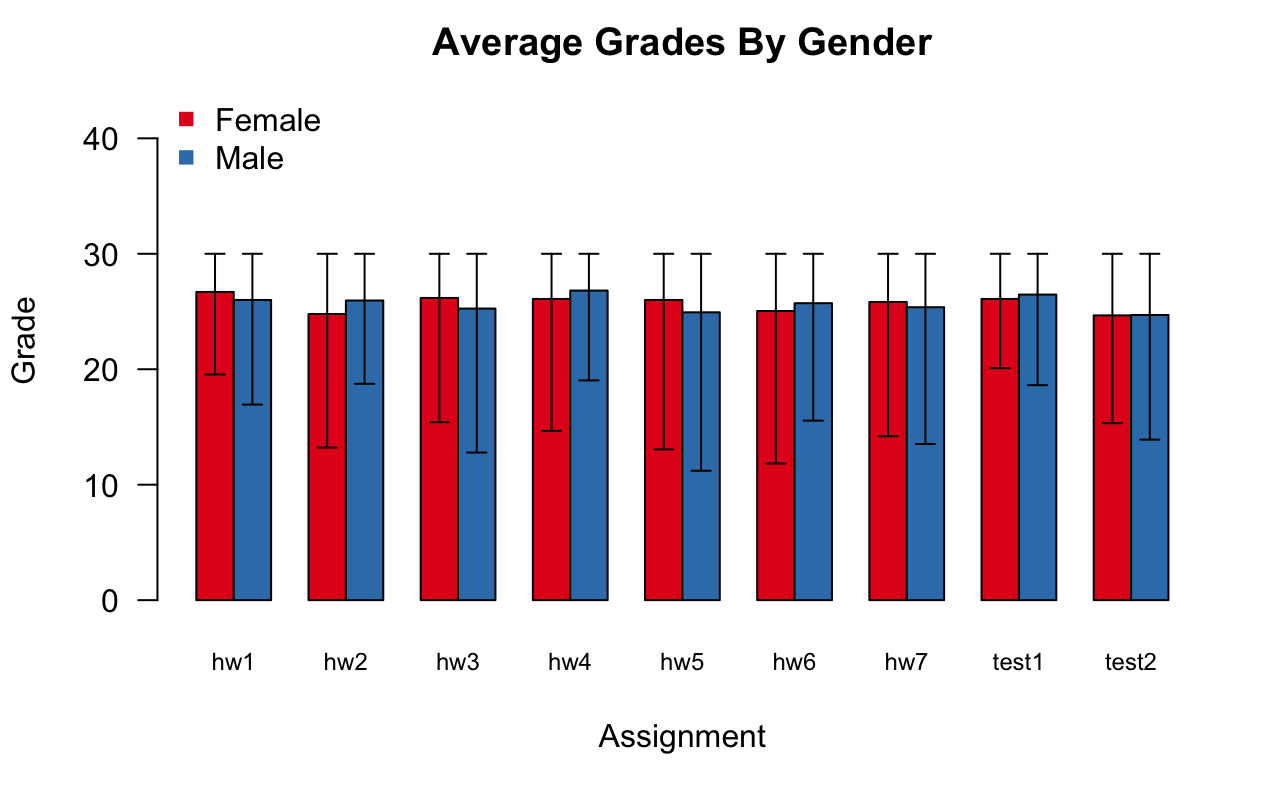
There were excused absences on tests that left 3 out of 220 test grades blank. When confronted with missing data, there are two main options. Remove the observations that have missing data in any column or impute the value. Because each student is important to the network structure, I did not want to remove any students from the data. I used the technique of computing conditioned means which leads to unbiased estimates of means but underestimates variance and overestimates covariance (Huisman 2014). Because of the relatively few instances of missing data, I am not too concerned about these issues. I tested linear models and random forests. The best model for each test was a linear model that was pruned using a combination of the Akaike information criterion (AIC) and removing predictors that were insignificant. AIC is asymptotically equivalent to leave one out cross validation (Stone 1977). Gender was ruled out as a predictor by insignificance and collaboration data was not included as a predictor. Finally, although I will proceed with the imputed data, it will be clear which grades were imputed and which were not as the imputed grades are not integers.

THERE ARE STILL SOME 0’S ON HWS, ARE THESE EXCUSED TOO?

**Descriptive Statistics of Data**

The average in-degree and out-degree for the network is 1.8. Women had higher average in-degree (1.95 v. 1.75) and out-degree (1.92 to 1.77) on average but in an insignificant manner. Although prior research has found that women are generally more attracted to collaborative working environments than men, there is no convincing evidence here to confirm that claim. (Kuhn and Villeval 2003). Over 99% of the in-degrees are less than 4 and 94% of the of the out-degrees are less than 5. The highest out-degree observed was 27. This degree distribution is different than most graphs which follow a small world structure and have a power law distribution (Liljeros et al.). While, the degree distribution did not resemble a power law for any of the problem sets individually, the out-degree distribution over all problem sets does resemble a power law distribution. This difference between the distributions of out-degree and in-degree is likely caused by the rule that you could only receive help from three other students but that there was no limit on the number of students you could help.

Scores on homework and problem sets did not differ significantly by gender. In every assignment, the grades by gender are indistinguishable and there are no clear patterns. Another insight is that grades in this class are typically high, with averages around a 26 / 30 on most assignments and medians one or two points higher because of long tails. I did not include the scores for the non-binary students in the class because there was only one who remained in the class (This student was one of the top performers in the class).

****

**Visual Analysis of the Networks**

Examining the network diagrams, a few key trends are illustrated that will guide further exploration. While network diagrams are not unique, the techniques used to draw these graphs use the properties of the graph to algorithmically determine locations of the nodes, placing nodes that are connected closer and minimizing crossing edges by assigning forces to edges. One feature of the visualizations is that genders tend to be concentrated together showing that people tended to work with other students of the same gender. The overrepresentation of male students is also apparent from these diagrams. An additional property is the highly active core of the network, where most students reside. The largest connected component was at least 75 students in every problem and grew as the course went on, reaching 94 of the 110 students in the last problem set. Finally, when the nodes are sized by grades on the assignments, the higher grades (larger nodes) tend to be focused among the highly-connected core while the worse grades are the disconnected nodes on the periphery.

**FIGURE OUT IF I AM DOING GIFS OR SOMETHING ELSE HERE?**

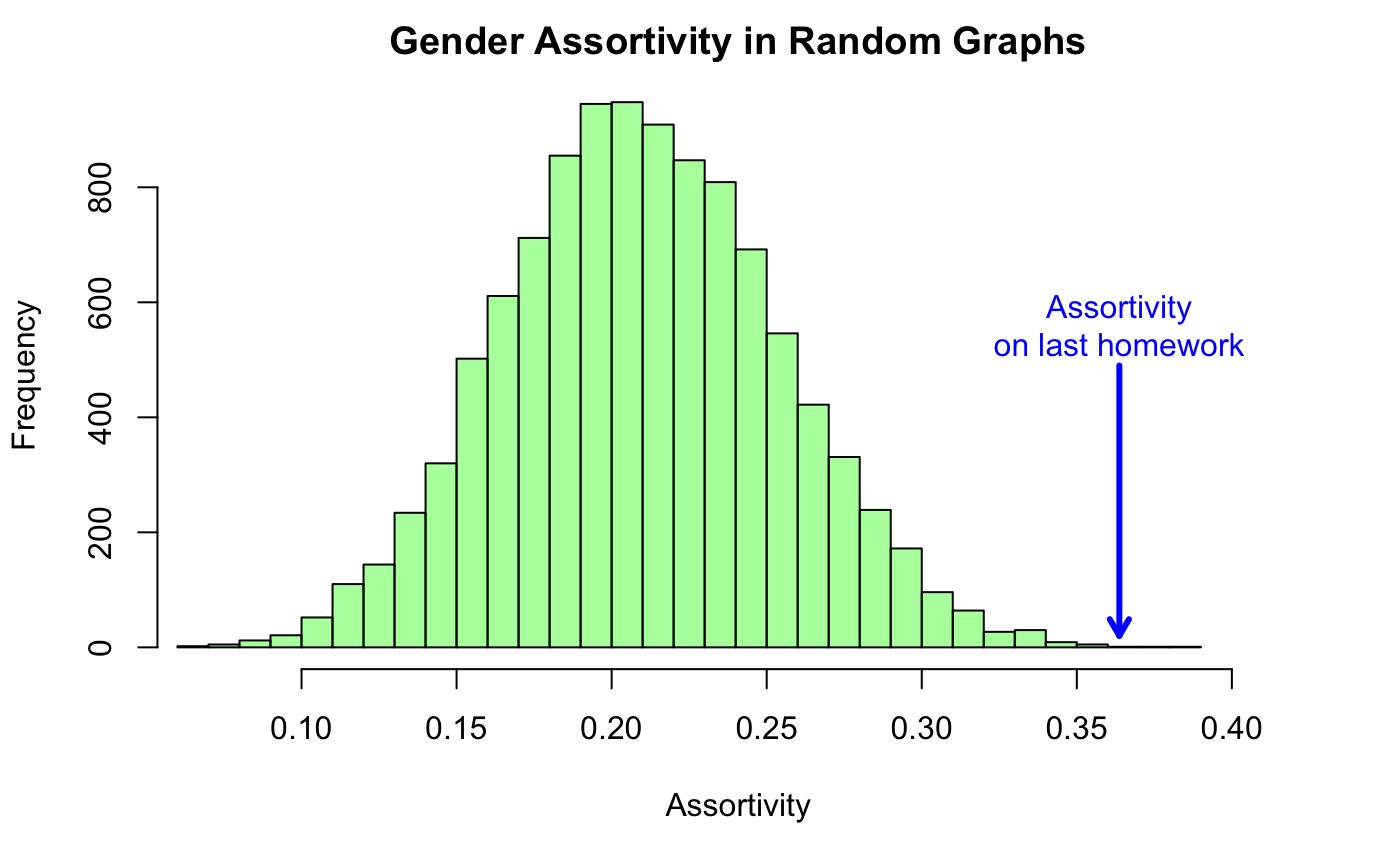
**Maybe one of the static plots from igraph as well as the dynamic evolution video?**

**Subsection: Evolution of the network overtime**

One of the striking features that evolved over time is the increasing gender homophily, measured by assortativity coefficients. Homophily is the tendency for nodes that are similar in an external property (gender in this case), tend to connect to each other. The assortativity coefficient is positive when homophily is present in a graph and ranges from 1 to -1. Assortivity is calculated as where is the mixing matrix of the graph. The assortativity of the collaboration networks increased from .26 in the first problem set to .36 in the last problem set. This increase was caused both by inter-gender ties breaking more and more intra-gender ties forming. To test whether this increase was significant or not, I took the core network, the set of edges that appeared in both the first and last problem, and randomly added the number of edges required to have the same edge count as in the last problem set. Randomly creating networks is a common technique in network studies because there are not enough experiments to determine significance otherwise (Bearman et al. 2004).

It is useful to distinguish between baseline homophily and in-breeding homophily. Baseline homophily is the homophily you would expect from random ties due to the prevalence of different groups while in-breeding homophily is the deviation above that random model (McPherson et al. 2001). In-breeding homophily is homophily that is most interesting because baseline homophily only corresponds the relative prevalence of different groups, while in-breeding homophily corresponds to the differential treatment of in-group and out-group members.

I created 10,000 simulated graphs, and in only 3 of them was the assortativity coefficient higher than the observed assortativity coefficient in the last problem set. The histogram shows the observed result's distance from the simulated results showing evidence of in-breeding homophily. This strong finding gives credence to the idea that gender is a salient feature to the students which influences which new connections are made and which connections are kept.

****

This finding may also be related other research about the way that men and women interact with groups differently. In other research, men have more negative perceptions of potential working partners, particularly women (Kuhn and Villeval 2013). These negative perceptions may also have contributed to greater lack of cross gender ties than would have been expected by random chance. The new connections are less likely to be people known before the class and thus quick judgments of other students may have a greater impact. Another research finding that may influence this pattern is the observation that men tend to talk much more in group work and are more likely to interrupt (Hall 1982). This may lead to mixed gender working groups being less effective and enjoyable, reinforcing the tendency to work with one’s own gender.

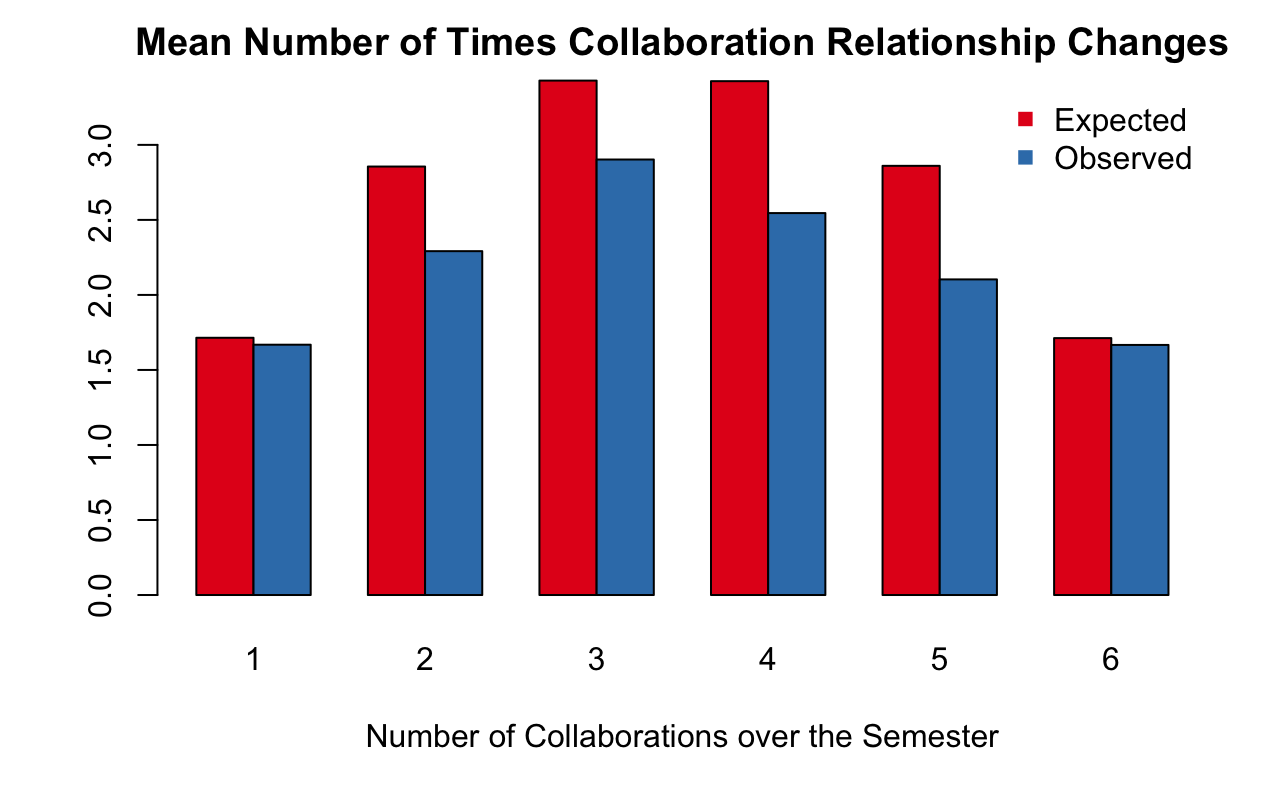
There is also evidence that more integrated social environments can lead to more women participating in STEM fields (Legewie and DiPrete 2014). This suggests that the segregated nature of the collaborative ties may be indicative of a larger trend at Yale that contributes to the imbalanced nature of STEM majors. Integrating men and women socially, and especially in classroom collaboration, could be an effective way to increase the number of women in these classes.

**Decrease in Number of Components**

The number of components and the number of isolated students is another of the striking features from the evolution of the networks. The number of connected components dropped from 34 to 9 while the number of unconnected students fell from 26 to 4. Meanwhile the largest component grew from 75 to 94. Following a similar randomization procedure to the method for determining the significance of the increase in assortivity, the decrease in components and unconnected students is actually in line with what would be expected to happen given the formation of random edges. In fact, the increase in the largest connected component is smaller than would be expected. This suggests that edges were not formed independently of the network structure but that the decrease in components is largely due to the increased number of edges in the network and not some pattern specific to this class.

**Subsection: Ties are more consistent that you expect just based on the number of ties that exist. Explaining this trend. THIS COULD POSSIBLY BE CUT**

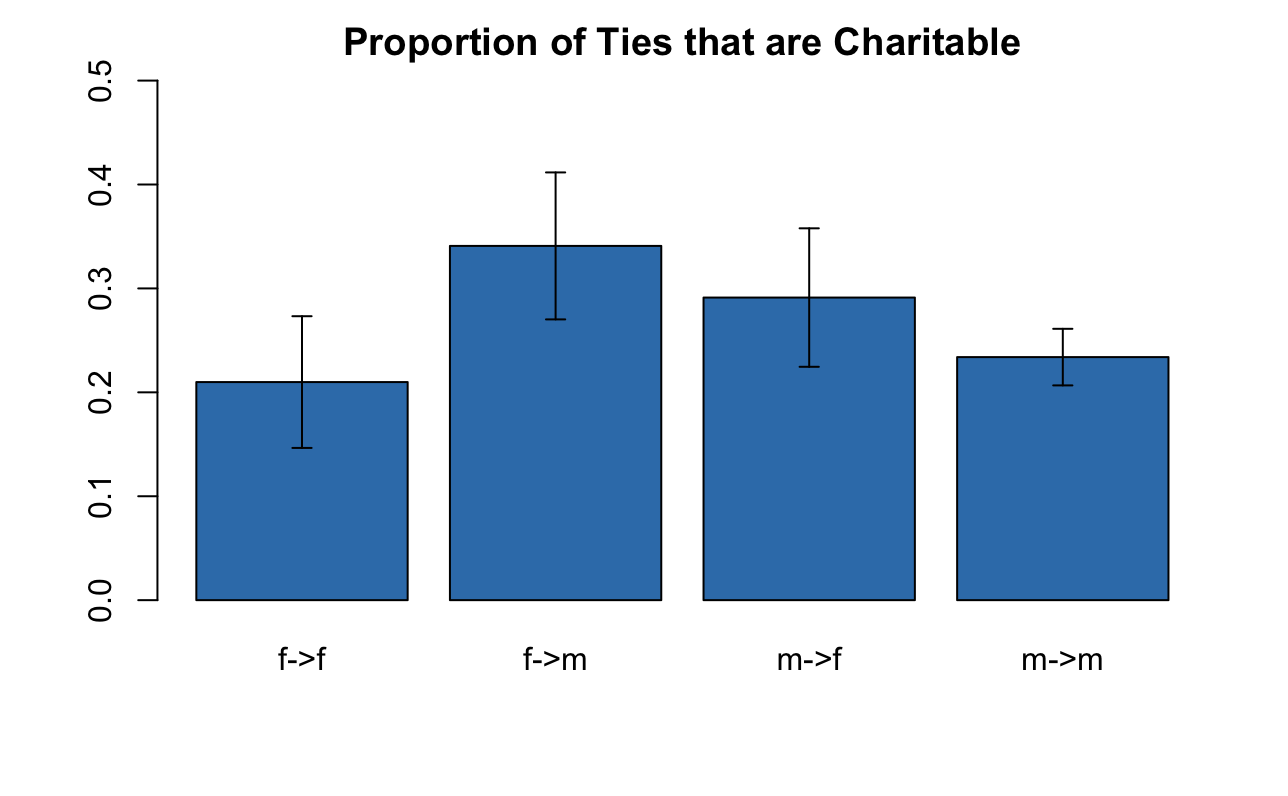
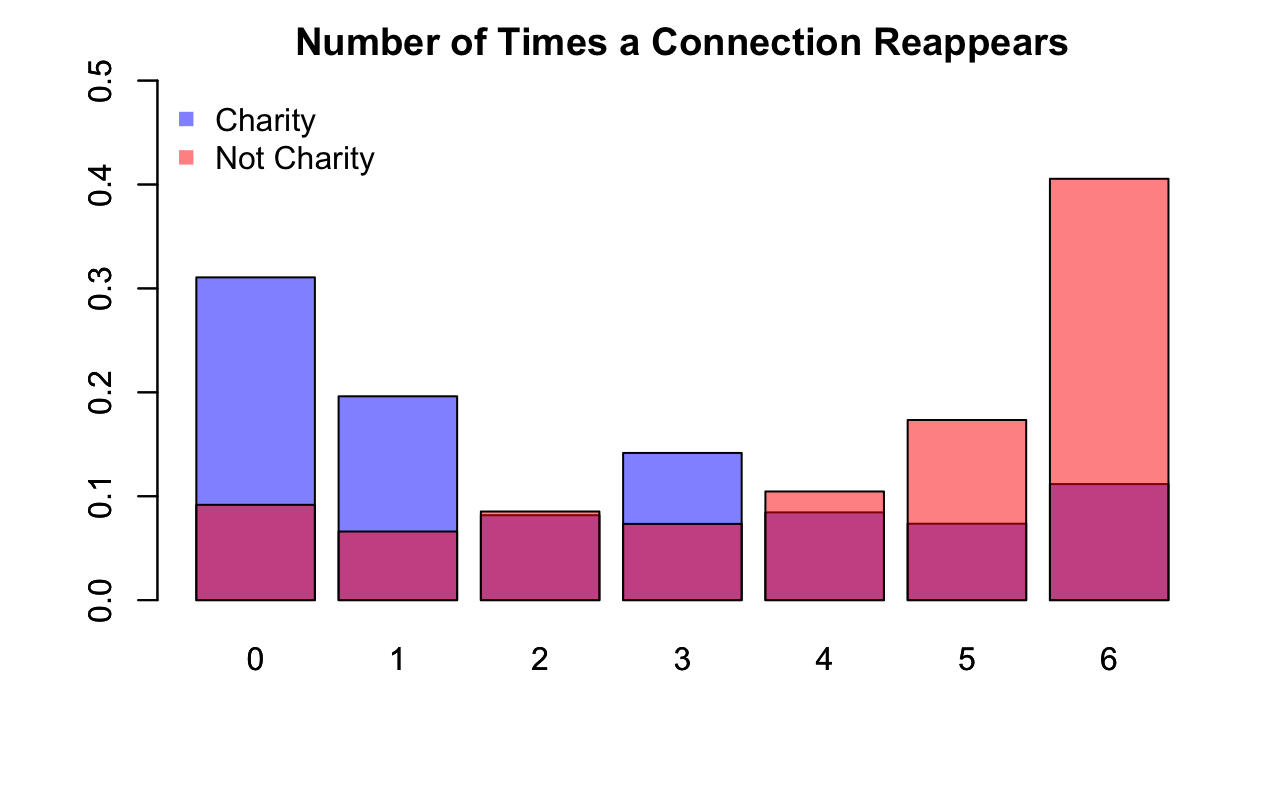
Despite significant evolution of the networks, collaboration is largely consistent over time. 48% of edges are in part of edges that appear either in 6 or 7 of the networks. Additionally, no two networks have less than 51% overlap in edges and the average overlap is 60% between networks. Furthermore, collaboration edges exhibit more consistency (fewer times when the relationship changes in consecutive problem sets) than one would expect just given the number of times that they are present in the networks. This trend suggests these relationships mostly switch on and off rather than popping in and out of existence.



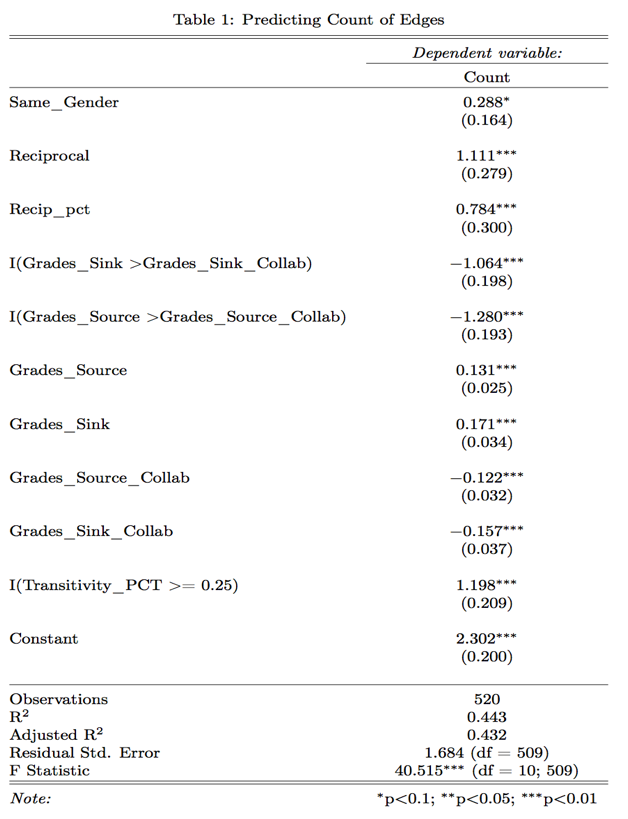
**Charity**

Another one of Kuhn and Villeval’s findings is that women seemed to be motivated by inequity more than men, related to similar findings that women are more generous in the dictator game (2003). This conclusion was drawn from the fact that women were more likely than men to choose team pay instead of individual pay when they were abler than his or her partner, raising the partner’s income at the expense of his or her income (Kuhn and Villeval 2003).

To test if a similar phenomenon was present in the collaboration data, I designated ties as being charitable if student A helped student B but did receive help from student B and student A received at least as good a grade as student B. I then took the proportions of ties that were charity in several different groups and tested for significant differences in means. After applying the Holm correction for multiple comparisons, the comparisons that showed significant differences were female to male ties had a higher proportion of charity than both male to male ties and female to female ties and mixed gender ties had a higher proportion of charity than same gender ties. While the conclusion that women were in general more likely to give charity than men that was found in the prior study was not supported, there are interesting results. The fact that mixed gender ties were more likely to be charitable hints at the intriguing gender dynamics of the class. I hypothesize this trend is because while people mostly worked with the same gender, when groups were stuck they may have asked for help solving the problems from someone that they didn’t typically work with. In fact, this intuition is correct and charitable ties appeared again in only 2.15 other problem sets on average while non-charitable ties appeared in 4.17 other problem sets on average. This difference is highly statistically significant.

****

**General Trends about Collaboration:**

The trends illustrated from the charitable ties are the important determiners of how collaboration relationships either remained or broke up. I categorized each edge that appeared any network by the number of times the edge appeared as well as structural and grade information associated with the endpoints. Again, cross-gender edges were less likely to remain despite there being no difference based on the genders involved or direction, only that edges spanned different genders. When transitivity, measured by the Jaccard index of the set of individuals that each of the collaborators collaborated with during the semester, was greater than or equal to , collaborations were far more likely to be consistent suggesting that groups are more stable than one-off partnerships. The cutoff was chosen empirically. Additionally, reciprocity, both the fact that the reciprocal tie exists at all and the percentage of times the tie is reciprocal, were significant predictors of more collaborations perhaps illustrating an aversion to free-riding by the students giving the help without receiving any.

Another important determiner factor was the success of the collaboration. When grades within the collaboration were less than students average grade in general, there was a strong aversion to continuing to collaborate. Furthermore, the extent to the difference was also important. Furthermore, in general, students who received better grades tended to have more stable collaboration. Whether this was because more stable collaboration led to better grades, because the students who received better grades preferred more stable collaboration, or the effect of another factor is impossible to determine.

**Grade Regressions (Including Gender differences)**

**This is a good intro**

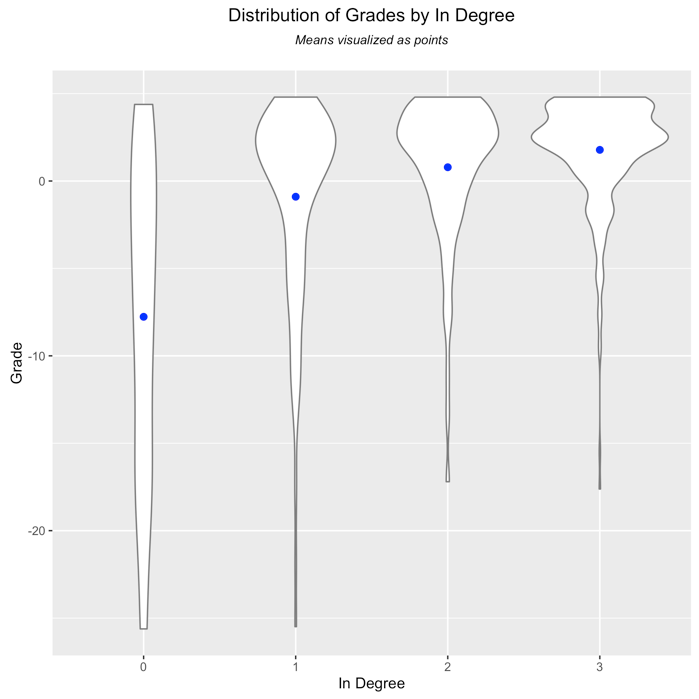
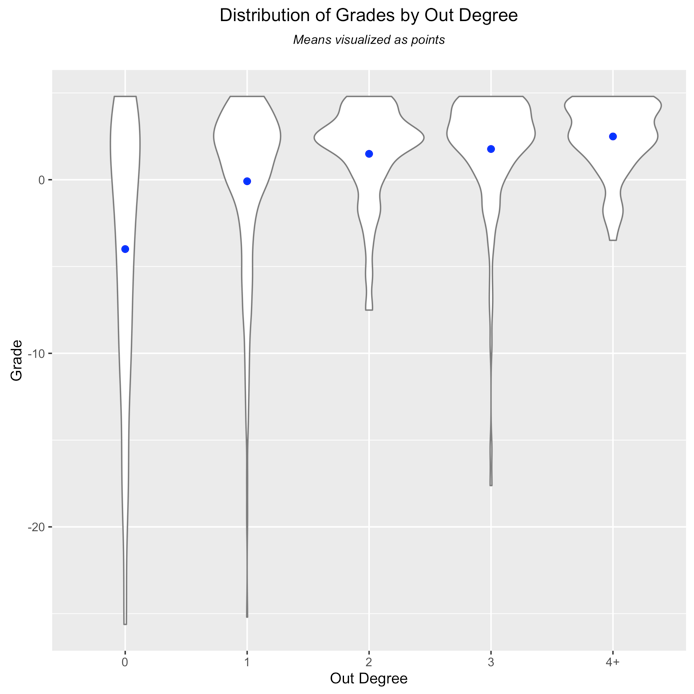
Collaboration happens in many courses but the effects are unclear. When collaboration occurs in a course it can be difficult to determine what is true collaboration where students are working together to solve the problems and what is more like one student copying the answer from another. Additionally, there is the worry that students who collaborate are at a big advantage over those who don’t. One course reviewer complained and cautioned future students that “if you don't have friends in the class, that just means that your grades will be lower than everyone else who is able to collaborate - beware of that (Anonymous 2016)!” I will analyze and explore these two phenomena, both demonstrating the effects of collaboration and analyzing the potential abuses.

Measuring the influence of collaboration is a difficult exercise. It is hard to separate out the effects that come from individual students’ abilities and efforts when the only individual assignments are tests that may be influenced through increased or decreased learning when collaborating with others. Additionally, because collaboration is quite consistent throughout the semester, grades on other problem sets reflect both the students’ abilities and their level of collaboration on that other problem set. One difficulty is that for most problem sets there are a few students whose performance is very disparate from their performance on other problem sets. This is might be due to random events that were happening at the time in students’ lives.

**Talk about methods here.**

Because grade means differed by assignment, I analyzed grades after subtracting out the means. Additionally, while there were some time dependent trends such as neighboring problem sets grades and collaboration being slightly more correlated, I found better results by removing information about which problem set or test grades came from. However, to avoid assuming independence when it was not there, cross validation on problem sets occurred by holding out each problem set, training on the rest of the data, and then predicting on the held-out data. Grades within each problem set are not independent since grades between collaborators are correlated. All model selection and accuracy over baselines are reported using the held-out data while model coefficients and significance scores are reporting using the full data set to get the most accurate numbers possible.

To get a first approximation of the relationship between grades and collaboration, I created violin plots which are like boxplots but feature rotated kernel density plots to better illustrate the distribution of data across categories. There are several important observations from these graphs. Grouping by both in-degree and out-degree, there is a big increase in grades as collaboration increases but there are diminishing marginal returns to more collaboration, especially after going from zero to one collaborators. Furthermore, the skew of the data is apparent. While most grades are around zero (the mean on each assignment), there are long tails in the negative direction. While this analysis does not illustrate any causal information, it suggests that collaboration exposes important trends for grades.

****

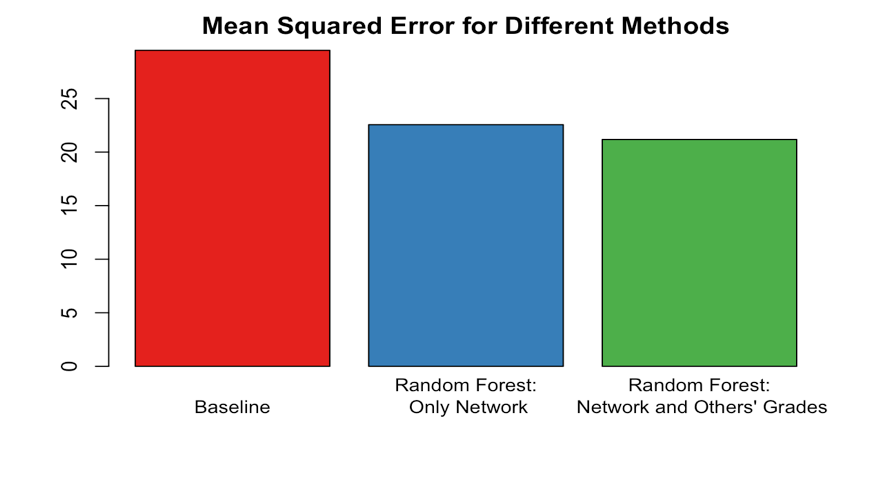
**PARAGRAPHs ABOUT RESULTS**

I then fit a variety of models with access to different information. Comparing the fits of these models will give some insight into which of these features are most important for predicting grades, and perhaps also success in the class.

First, I tried to predict grades using information only obtainable from looking at the network structure. I then added information about the grades of collaborators but held out information about individuals’ grades on other assignments. For both models, the baseline is a naïve model that predicts the mean grade, which is 0, for every person. Both methods drastically outperformed the baseline, but performed similarly to each other despite the increase in information available. I suspect this is because collaboration and good grades are linked in a way that allowed the simpler model without grade information to get an approximation of this information. Evidence of this is that membership in the largest component was included in the model for the simplest model but was removed for the model that allowed grade information to be considered. Membership in the largest component is related to higher grades but in a weaker way than collaborators’ grades.

Additionally, the measure of centrality that was most helpful for these predictions and the was included in both models is constraint. Constraint measures access to disparate sources of information which might explain why it is so predictive here. Constraint is defined as

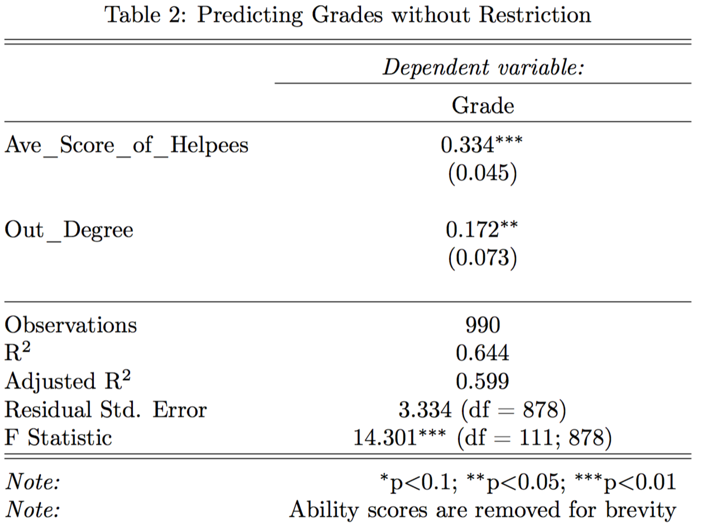
where is a vertex’s ego network, and represents elements in the row-normalized adjacency matrix. Constraint is lower for students with less redundant ties and ranges from 0 to 1. Because constraint is undefined for isolated students, I set constraint to 1.5 for these students to allow the model to be fit with all the data. Since random forests were used for these models, the important thing is that there is a cut point that disambiguates this arbitrary value from the real values of constraint. Nevertheless, the value of 1.5 also fits well with the linear trend of constraint. Ties to different groups of students might give you exposure to more ideas and a better chance of getting the correct answer to a problem set. It also may be the case that better students have lower constraint because they are asked by many different people to collaborate. Lower constraint has been linked in work settings to the formulation of better ideas, higher salary, and promotions (Burt 2004). This corroborates the finding that constraint can be an important determiner of success, especially for complex tasks.



**PARAGRAPH ABOUT USING PEOPLE’S Grade INFO**

I then predicted grades based on all available information. Importantly, the model had access to individuals’ grades on other assignments and used student identities as a factor variable. This method creates “ability scores” for each student. The baseline I compared this model against was an individual’s grades on other assignments, which turned out to be a great predictor of grades that was hard to improve over. The errors for both the baseline and the model in this section are about half as big as in the previous section.

Nevertheless, there were significant improvements to be made by including information from the network. Cross-validated mean squared error decreased from 13.5 to 12.6. This is a 6.5% decrease that is important but much smaller in magnitude than the decrease achieved above because the baseline is so much better in this model. This small improvement is reflected in the small magnitude of the coefficients trained for this model. Interestingly, Ave\_Score\_of\_Helpees, which measures the score of those whom a student helped on an assignment, is not about the information students received but about the information volunteered. Additionally, the fact that Out-Degree was a good predictor highlights the importance of the help given. I have two possible theories for why help given is so important. One is that by helping another student, a student learns the material better and this extra learning is reflected in higher grades. The second possibility that I think is more compelling is that helping other students is an indicator of a student who is confident in the fact that he/she has the right answer and is thus linked with higher grades because students are good at knowing if they are right or not. However, these conclusions are dampened because this linear model, chosen for the best performance on the held-out data and Akaike information criterion, is one of many similar models that can achieve roughly equivalent performance.



I also attempted to use expectation maximization to simultaneously fit coefficients for each individual’s own ability, , as well as coefficients and that measured the effect of the ability of edges coming in and out respectively according to the following formula.

However, this method did not produce any improvements from the baseline model of individuals’ average grades. Furthermore, the estimates of ability changed very little from this baseline. I suspect the failing of this method is the reliance more on the source of the collaboration than the fact or whether collaboration existed at all. In the other models that are more successful, the information that is most helpful is whether collaboration exists. Additionally, the reliance on ability scores neglects fine-grained information such as the score of students you helped on a specific problem set. This doesn’t allow for learning when a group of students may be preforming worse than their average on a given problem set.

**Paragraph about whether or not the results would seem to indicate that men and women are at differential access to the important things**

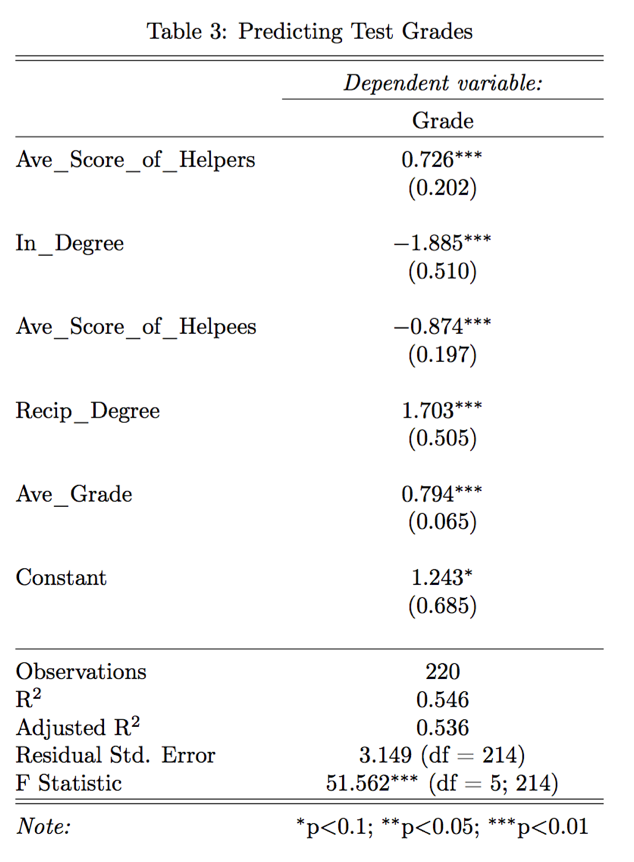
To compare the impact of collaboration on men and women, I compared the percentiles of grades to the percentiles of “ability scores.” Those with higher grades than ability scores were helped by collaboration while those with higher ability scores than grades may have benefitted from a different system. While women were more likely to receive help (61% vs. 49% for men), this difference is insignificant. This is parallel to the finding that women collaborate more than men in a statistically insignificant manner. The fact that women and men receive roughly similar amounts of help from collaboration is surprising given the large gender imbalance in the class and the notion of women being disadvantaged in STEM classes. One possible explanation for this phenomenon is that men are more confident in their abilities (Kuhn and Villeval 2003; Hall 1982) and believe that they don’t need to collaborate, even when collaboration could help them. Another explanation is that women’s status as the underrepresented group energized them to collaborate more than otherwise might happen. While there was a Facebook group for women in the class to work together, members of the group do not report it being particularly active.

**Tests:**

To test the hypothesis that collaboration may lead students to essentially copy other students work without learning the material, I predicted grades on tests using grades on homework and collaboration information. If those who collaborate are not learning as much, then I would expect that collaboration information would improve the accuracy of models and would cause predictions for those with more collaboration to have worse test scores. It is also possible that collaboration improves learning, holding grades constant, and that models would predict higher grades for those with more collaboration. I hypothesize that the first is more likely because students are probably doing the minimum amount of work required to get the grade they received.

To serve as a baseline for this section, I predict individual’s test grade by predicting their average grade on the problem sets. To prevent overfitting, I used 10-fold cross-validation that chooses which students to include completely randomly as test scores for different people are independent from each other since they are individual assignments.

A linear model with average grade on problem sets, average in-degree, average number of reciprocal ties, average score of helpers, and average score of helpees outperformed the baseline by reducing mean squared error by 15% from 12 to 10.2. Average grades are a very good predictor of test scores mostly likely because the nature of the class’s tests. The tests did not require students to solve new problems but instead required students to solve problems that appeared in class or on a homework assignment. The remaining coefficients suggest an interesting, intuitive story. After conditioning on grades, collaboration led to predictions of lower grades, particularly of those free-riding. While collaboration led to lower grades, the extent to which one’s in-degree was higher than their number of reciprocal ties the more extreme the expected penalty on one’s grade was. This effect was somewhat mitigated by having helpers have higher grades than those you are helping but this does not reverse the effect. This illustrates that the worry that students who collaborate do not learn as much is backed by the evidence. However, since students who collaborate do better on homework assignments, this method overrates the free-riding effects of collaboration. Still, the fact that collaboration, and particularly non-reciprocal collaboration, is associated with lower test scores is an important trend to recognize.



**Simulating under different hypotheses**

To measure the effectiveness of these models under different assumptions, I created 1000 simulated data sets under the null hypothesis that grades are equal to people’s average grades plus a random error term.

This reduced the performance of the random forest predicting grades solely with network information by the least, but still significant, amount. This illustrates that the features included in that model (Constraint, Reciprocal Degree, In-Degree, Out-Degree, and membership in the biggest component) are the most consistent features that mark higher grades but still are dependent on the exact dynamics of the network. The random forest that predicted grades based on the network information and the grades of collaborators declined much more in performance but still outperformed the baseline. Since this model relied heavily on the performance of collaborators and now preforms worse, this illustrates grades of collaborators, and specifically the deviations from their average grades, are more correlated than what would happen by chance alone.

In the both linear models, the one predicting grades on homework and tests conditioned on all available information and the one predicting test scores based on average level of collaboration, not a single coefficient had a 95% confidence interval that excluded 0. This demonstrates that when the deviations of grades from student’s average grades happen at random, network features are not helpful in finding patterns. This finding gives more confidence to the results that the two linear models examined are finding important features instead of simply following noise in the data and helps to reject the null hypothesis that the network structure is not connected to deviations in grades.

**Non-null hypothesis, probably get rid of**

I also tested a non-null hypothesis with another 1000 simulated sets of grades to test how well it fit the observed trends. The hypothesis I tested takes multiple steps. First, I simulate grades under the null-hypothesis that grades are equal to people’s average grades plus a random error term. The random error term has mean equal to the how much more help than average that student had. Then, I calculate the average grade of the students that each student helped and received help from. Then, I set students’ grade equal to the average of these three numbers (grades of helpers, grades of helpees, and own grade). I penalize students half a point for each of these that is missing. For tests, I change the error term so that it has mean for student and the same standard deviation as before.

This method recreates the trends on the tests well. The 95% confidence intervals for all of the coefficients that were statistically significant in the original regression do not include 0 and all have the same sign as they do in the original. While the values of the coefficients are different, this suggests that the trends in test performance can at least be partially recreated by penalizing those who have received more help than they gave and rewarding those who gave more help than they received.

This method did nothing to change the performance of the two random forests compared to the null hypothesis and did not lead to any significant coefficients for the linear model predicting homework scores. This illustrates that this average scheme is not the way that collaboration is correlated with peoples’ grades.

**Section: Conclusion of these results:**

There are several main takeaways from these different models. While more collaboration is predictive of and correlated with higher grades, it is not possible to determine causal impacts. The data does not contain sufficient natural experiments. Nevertheless, the network structure is importantly linked with grades. Diversity of collaborators appears to be advantageous because lower constraint consistently appears in a variety of models as a good predictor of better grades. There are instances of free-riding on the network and while these students can get higher grades on problem sets, the tests serve as a check against this behavior. Furthermore, an important unobserved impact of collaboration may be time savings, something that may be equally as important to students as better grades. Additionally, gender was never a useful or significant predictor of grades in any model showing while there may be differences in collaboration by gender, there are no patterns in ability by gender.

**Conclusion and Recommendations**

Collaboration in this class is characterized by a diffuse network where most activity is concentrated in the largest connected component. Gender is an important feature in the networks but neither side has an advantage. The average grades received by men and women were indistinguishable. Despite women making up less than a quarter of the class, the women in the class collaborate more and were helped more by collaboration than men in statistically insignificant ways. Intra-gender edges were more stable and became more prevalent as the course continued. While inter-gender edges were more likely to be charitable, these ties were short-lived. The other important norms enforced by students in the network were reciprocity and transitivity. Edges that were reciprocal and were between members who had other collaborators in common were much more likely to continue working together.

Collaborating with classmates is strongly linked with higher grades. One network feature that was consistently linked to higher grades is lower constraint, a measure of the redundancy of a student’s ties, which highlights the advantage to accessing multiple information sources. However, causal effects are elusive. There were no natural experiments and the consistency of collaboration made untangling the impact difficult. Although the best predictor of grades was grades on other assignments, this could be improved with network information. There is some evidence of free-riding on problem sets without fully learning the material because test scores are lower after conditioning on homework grades when a student collaborates more, especially if they received more help than they gave.

Overall, the collaboration system seems to be working well, but there are ways for it to be improved. While most students collaborate, an effort to get those students currently not working with anyone to the network could help those students as the first collaboration is very important and the large connected component offered some advantages. Encouraging more inter-gender collaboration could reduce the gender segregation in the network and lead to a more open and welcoming class. Although all genders succeeded equally in the class, the overwhelming majority is still male so attracting more non-male students to the class is another potential improvement. Putting a larger emphasis on the individual tests could reduce the free-riding and reward the students who are learning the most.