Analyzing the Social Networks in Stanford's CS106A

Alyssa Vann[1] and Hugo Valdivia[2]

I. INTRODUCTION

Our project is focused on understanding the social networks that exist in Stanford Computer Science courses; in particular, we aim to understand how these social networks affect student outcomes in these courses. For this paper, we are analyzing Stanford's introductory Computer Science course, CS106A, using data from the course's iteration in the Winter of 2017 when it was taught by Lecturer Chris Piech.

We would like to explore what types of connections students form, and more broadly what types of communities exist, in CS106A. The idea of homophily, that "contact between similar people occurs at a higher rate than among dissimilar people" [1], is a driving concept motivating our investigation of CS106A.

The Washington Post does an effective job explaining this phenomenon in their article "Three quarters of whites dont have any non-white friends" [2]. Figure 1 illustrates some of the findings of the article. In his book *Whistling Vivaldi* [3], Claude Steele elaborates on how this affects student performance in courses in his description of the observations of the mathematician Philip Uri Treisman at the University of Berkeley. Triesman wanted to understand why Black students entering Berkeley with the same preparation and academic performance as White and Asian students were under-performing in Calculus I (a gateway to many majors). He performed an anthropological study, and followed students from different racial groups throughout their days. He found that while Asian students (and White students to a lesser degree) worked through problem sets together, pulling their intellectual resources, Black students frequently worked in isolation. Black students were working just as hard, if not harder than, other students, with less success because their only check on their understanding was the back of the book. Treismans work highlights that students who work with their peers gain a deeper conceptual understanding of course material, which translates into better performance overall. The concept of homophily does not only apply to race, but to gender, and any other grouping of 'similar' people.

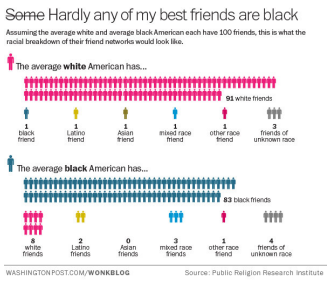


Fig. 1: Homophily Infographic from *The Washington Post*

If most of peoples friends are from a similar group, then certain students are immediately at a disadvantage when it comes to pulling 'intellectual resources when they enter Computer Science classes. These problems are likely exacerbated as students move past introductory classes, and the already small group of other students they might be acquainted with begins to take different sets of classes. As Computer Science classes move toward pair programming and group assignments, without assisting students in creating pairs, and continue grow in size (making it harder, if not impossible, to get help in office hours), students networks of friends in the field become essential to their success.

These concerns motivate our work investigating the social networks that exist in Stanford's CS106A.

II. RELATED WORK

A. Community Structures

The clustering coefficient is a key property that we can compute to begin to understand a graph; it provides a probability that for a particular node, two neighbors of that node are themselves connected. Like the clustering coefficient, community structure is yet another network property that helps us understand the underlying structure of a network.

The 2001 paper "Community structure in social and biological networks" by M. Girvan and M.E.J. Newman [4] and the 2006 follow-up paper "Modularity and community structure in networks" by M.E.J. Newman [5] discuss traditional and (at the time of writing) current approaches to the problem of revealing community structure within a network. In [5], Newman describes the need for a definition of communities by arguing that we commonly see "subsets of vertices within which vertex to vertex connections are dense, but between which connections are less dense." This view informs Newman's eventual choice to create an objective measure, modularity, in which "community structure in a network corresponds to a statistically surprising arrangement of edges" that differs from what we would expect from random chance. Newman outlines an algorithm that repeatedly splits graphs into two if his modularity metric detects that two subcommunities exist, and determines that his algorithm is suitable for networks up to a certain size. Newman's algorithm (the Spectral Newman algorithm) "clearly outperforms" the betweenness-based algorithm of Girvan and Newman [4] and fast algorithm of Clause et al., but is on par with the extremal optimization algorithm of Duch and Arenas

*This work was not supported by any organization

[1]Alyssa Vann is with the Department of Computer Science, Stanford University, Stanford, CA 94305, USA avann at stanford.edu

[2]Hugo Valdivia is with the Department of Computer Science, Stanford University, Stanford, CA 94305, USA hugov65 at stanford.edu

[pg. 8581]. Ultimately, Newman's modularity measure is one choice of many for functions that formalize the notion of "community."

The paper "Statistical Properties of Community Structure in Large Social and Information Networks" [6] asks the question of whether community-finding algorithms have a general structure. The authors outline the general structure of community-finding algorithms as (1) pre-processing data to model as a graph, (2) hypothesizing that the graph contains communities that interact more strongly within themselves compared to the rest of the graph, (3) choosing an objective function to formalize the notion of community, (4) developing an algorithm to calculate exactly or approximate the objective function, (5) evaluating the communities found by the algorithm. The most novel contribution of the paper is the introduction of the Network Community Profile (NCP) for the evaluation step (pg. 695-6). They apply this paradigm, using conductance as their objective function and NCP for evaluation, to real world networks to find that in large networks, the large communities begin to "blend in" more and more. Later in "Higher-order Organization of Complex Networks," [7] the objective function of conductance is further developed into function called "motif conductance," where motifs are small subgraphs. These findings are what follow in the literature; however, this work is not as relevant to our project since it focuses on methods that are particularly optimized for large networks.

These community finding algorithms are at the heart of our project. We need to understand the literature on the fastest, most accurate algorithms in order to have confidence in the conclusions we derive. At the current time, there are two community finding algorithms that have been implemented for SNAP.PY - they are the betweenness centrality-based approach from Girvan and Newman and the fast greedy hierarchical algorithm of Clause-Newman-Moore, which are mentioned in [5]. We've included these algorithms in our findings; however, we will also implement two of the more advanced algorithms (in particular, the Spectral Newman [5] and the Clique Percolation algorithm [8]) for our project.

B. Social Networks

In the paper "Affiliation Networks," [9] Lattanzi and Sivakumar build on existing models of social networks. They describe how social networks densify (with the ratio of edges to vertices growing) and shrink in diameter to a constant. Their findings, however, ultimately build on work from the social sciences describing the bipartite nature of social networks with societies and actors, which can be used to describe affiliations, and are therefore called affiliation networks. They create a social network amongst actors by folding the graph, which they do by replacing paths of length two in the bipartite graph among actors by an (undirected) edge. Using this method, one creates a network of actors in which people share an edge if they share an affiliation. Lattanzi and Sivakumar focus on the evolution of such networks using the ideas of preferential attachment, the idea that new nodes in a network are more likely to attach themselves to nodes with higher degrees, and of edge copying, the idea a new node picks a node prototype and copies its edges. These are ideas are useful for the study of networks. They could be further developed by using them to complement work done to find communities in networks.

In the paper "Social networks that matter: Twitter under the microscope," [10] Huberman, Romero, and Wu study the Twitter social network. They refute the idea that "a social network embodies the notion of all people with whom one shares a social relationship." Rather, they confront the reality that most people interact with only a small subset of those listed as part of their network, and hold that these hidden networks "that are made out of the pattern of interactions that people have with their friends or acquaintances," are the networks from which one can truly study influence. They define a friend as someone who a user has interacted with through directed posts or messages, and show that a users number of friends is a good predictor of how active a user is. Ultimately, they find that social networks are really composed of two networks, a dense network of listed followers-followees, and a sparse network of actual friends. Huberman, Romero, and Wus paper is helpful for thinking more deeply about the significance of links in social networks, and rightfully point out that a link does not imply true interaction or influence moving from one user to another. It would be helpful if they defined friends in terms that could be applied more broadly across networks, though it is certainly possible to take their idea and apply it to different contexts. They also do not provide details of how they define density or sparsity, which would strengthen their discussion of the two different networks that emerge in social networks. As readers, we are left with a very useful framework to think about social networks differently, but with the need to define their terms in ways that may applicable across different networks.

The idea of social networks being composed of a dense and sparse network applies quite directly to our work. We would like to explore what changes we see if we create a dense network based on 'affiliation' [9] in which we create links between all students within a given section or with a given section leader, versus a sparse network of links between nodes only if we are certain those nodes have interacted in a significant way (i.e. collaborating on an assignment or receiving direct help and feedback from a section leader).

III. METHOD

A. Data

Our project required us to combine data from multiple sources. Our data focuses on the networks in CS106A from the Winter of 2017. We received data from Stanford's CS198 program, which runs the section leading program for CS106A. From CS198, we received the following data: which section each student in CS106A was a part of and which section leaders they received help from in the LAIR (the CS106A evening office hours). From Lecturer Chris Piech, we received student self-reported data on their gender,

year, major, reasons for taking the class. We also have anonymized information on the grades of each student (including overall assignment score, midterm score, final score, and participation score), and the comments students wrote on each of their assignments. We used student comments to find student self-reports of who they collaborated on assignments with.

To augment our data, we used the NamePrism API [11] to estimate students' ethnicities. The API provides 13 ethnic categories (British, French, Germanic, East European, Jewish, Nordic, Italian, Japanese, East Asian, Indian Sub-Continent, Africans, Muslim, and Hispanic), which we translated into the categories found on the United States census (White, Black, Asian, Hispanic). This method certainly does not completely accurately capture the ethnicities of students, for example, Black-American names are consistently misclassified by NamePrism; however, augmenting our data with NamePrism allows us to begin answering questions about the nature of community structures in Stanford's CS106A. Because our project is primarily concerned with the experiences of underrepresented minorities in Computer Science, we further subdivided the racial and ethnic categories we estimated from NamePrism into the categories of underrepresented minority versus not. We created this category because using our diversity measures, a group composed only of White and Asian male students is labeled as diverse. Without a doubt, this is a diverse group; however, this metric does not tell us how integrated or not underrepresented students are in communities in CS106A, something we are deeply interested in finding out.

Because we were also missing some information on gender, we augmented our demographic data using the Genderize.io API [12]. We use this API to estimate the genders of students who did not self-report their gender during the class. The API does not capture non-binary gender information, but does allow us to create a slightly fuller picture of student demographic data in CS106A.

B. Networks

We turned this data into two undirected graphs: a dense network and a sparse network [10]. Our data fits the idea of there being dense and sparse networks very well. In our sparse network, we capture hyper-meaningful connections from students to other students and section leaders, who they directly collaborated with or received help from. In our dense network, we capture connections between acquaintances who would have seen each other in sections, but not necessarily interacted with each other in a substantial way.

Both networks have a node for each student (who at least took the midterm) and for each section leader. Both networks have edges from each student to their section leader, from each student to the students they collaborated with on assignments (parsed from self-stated collaborations written on student assignments), and from each student to the section leaders who assisted them in the LAIR. These edges make up our sparse network, in which we prioritize connecting nodes that have definitively interacted at some point during CS106A. Our dense network includes the same connections, but is also composed of edges connecting all students in a given section. These connections are weaker than the previously listed connections, but still form a part of students' experiences in CS106A.

C. Algorithms

For our project, we compare the results of four different community-finding algorithms on our networks. Two of these community-finding algorithms, the greedy Clause-Newman-Moore (CNM) algorithm and the Girvan-Newman algorithm, are available in SNAP.PY, so we use SNAP functionality as the basis for our analysis with these two algorithms.

We go beyond the methods provided in SNAP.PY, and implement two more community-finding algorithms: the spectral community-finding algorithm outlined by Newman [5] and the Clique Percolation algorithm for finding overlapping communities. Finding overlapping communities is especially important for our analysis because in the real world, people are often a part of many different communities. In our networks, students may be part of communities that are informed by their gender, their race and ethnicity, and their minority status.

From Newman [5], we know modularity is calculated using the following equation

$$\begin{aligned} Q &= \frac{1}{4m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})(s_i s_j + 1) \\ &= \frac{1}{4m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})s_i s_j \\ &= \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s} \end{aligned}$$

in which A_{ij} is the number of edges between nodes i and j (we are currently using an undirected graph that allows for only one edge), s_i and s_j take on a value of 1 or -1 depending on whether they correspond to group 1 or group 2, m is total number of edges in the network, and

$$B_{ij} = (A_{ij} - \frac{k_i k_j}{2m}).$$

Newman states that "the modularity can be either positive or negative, with positive values indicating the possible presence of community structure. Thus, one can search for community structure precisely by looking for the divisions of a network that have positive, and preferably large, values of the modularity." Newman goes on to outline a spectral algorithm for finding more than two communities in a network, which we implement on our networks:

1) Find one division of the network into two communities, if possible. We do this by choosing the appropriate $\mathbf{s}$ in the following equation:

$$\begin{aligned} Q &= \frac{1}{4m}\sum_i a_i \mathbf{u}_i^T \mathbf{B} \sum_j a_j \mathbf{u}_j \\ &= \frac{1}{4m}\sum_{i=1}^{n}(\mathbf{u}_i^T \cdot \mathbf{s})^2 \beta_i \end{aligned}$$

where β_i are the eigenvalues of $\mathbf{B}$ corresponding to the eigenvector u_i in decreasing order. We do this, in practice, by finding the eigenvector of the largest eigenvalue and dividing the nodes into two groups based on the signs of the elements of this vector.

2) Calculate the additional contribution of ΔQ to modularity upon a further division of an existing community g of size n_g, and maximize ΔQ.

$$\Delta Q = \frac{1}{2m}[\frac{1}{2}\sum_{i,j\in g} B_{ij}(s_i s_j + 1) - \sum_{i,j\in g} B_{ij}]$$

$$= \frac{1}{4m}[\sum_{i,j\in g} B_{ij} s_i s_j - \sum_{i,j\in g} B_{ij}]$$

$$= \frac{1}{4m}\sum_{i,j\in g}[B_{ij} - \delta_{ij}\sum_{k\in g} B_{ik}]s_i s_j$$

$$= \frac{1}{4m}\mathbf{s}^T\mathbf{B}^{(g)}\mathbf{s}$$

where

$$B_{ij}^{(g)} = B_{ij} - \delta_{ij}\sum_{k\in g} B_{ik}$$

3) Repeat this process until there is no division of a community that yields a positive value of ΔQ. At this point there is nothing to be gained by further dividing the graph.

To find overlapping communities, we implement the Clique Percolation algorithm. To implement this algorithm, we first recursively found maximal cliques starting with each node in a network. We then find the amount of overlap between each maximal clique pair (i, j), and threshold these values using some k, making the values 0 if they were less than $k-1$ and 1 if they were greater than or equal to $k-1$. Finally, we combine all cliques with a value of 1 after the threshold (who therefore have an overlap greater than or equal to $k-1$) into communities, producing a set of overlapping communities for us to evaluate.

D. Measuring Diversity

Finally, to understand something about the communities that exist in CS106A, we need to measure how homogeneous the communities we discover are. To measure the diversity of communities in CS106A, we use two metrics commonly used to measure the diversity of communities, known as the Shannon Entropy metric and the Sullivan Composite Index metric. These metrics are described in the paper "Using composite metrics to measure student diversity in higher education" [13]. For both metrics, when a community has no diversity it has a diversity value of 0, and as diversity increases its value moves further and further from 0.

The Shannon Entropy metric is calculated as

$$D = \sum_{i=1}^{n} p_i \times ln(p_i)$$

where p_i is the proportion of individuals in some category i (like male or female if one is considering gender diversity).

Instead of producing different metrics for different types of categories (like race and gender), Sullivan's Composite Index comes up with one diversity score across variables V (like race and gender).

$$C = 1 - (\sum_{k=1,p} (Y_k)^2/V)$$

There are V variables, p categories, and Y_k proportions in each category.

We use these metrics to quantify how diverse the communities we find through our algorithms are, and attempt to describe how diverse, or not, communities in the CS106A network are.

IV. Results and Findings

A. Student degrees

For the sparse model with LAIR edges, we first calculated some initial statistics on our network. We found that the effective diameter of our network is 7, where we define the effective diameter (as the GetBfsEffDiam SNAP.py function documentation does) as the 90-th percentile of the distribution of shortest path lengths. We found that the average degree of nodes in the whole graph is 7.28. The average degree of section leaders is 19.72, and the average degree of students in the course is 0.43. The average degree of female students is 6.08, and the average degree of male students is 3.89, nearly half that of female students.

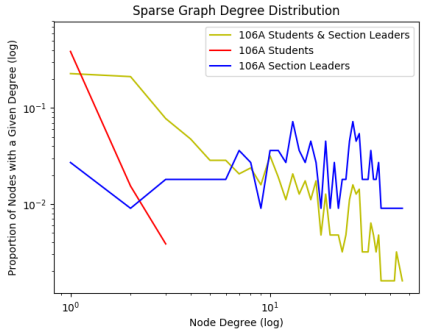


Fig. 2: Sparse network degree distributions of students and section leaders.

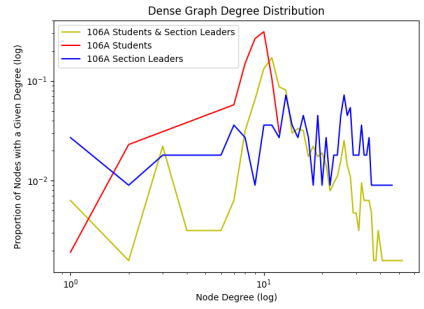


Fig. 3: Dense network degree distributions of students and section leaders.

In Figures 2 and 3, we plot the degree distributions of several portions of our network to better understand the degrees of different types of nodes. In our sparse network, we see that all nodes have a general downward trend, which aligns with the power-law distribution and is like the distribution of a preferential attachment graph, which can be seen in Figure 4 for the purposes of comparison. When we look only at student-to-student connections and only at section leaders, the distributions vary greatly. On average, section leaders have higher degrees than students because they interact with many students from different sections during LAIR. We also that see that the majority of students have very few connections to other students. As expected, the degree distribution of students in the dense network varies greatly from that of students in the sparse network. In the dense network we see that students still do not have the high degree counts of some section leaders, but do have higher degrees than in the sparse network because they have edges to all the other students in their section.

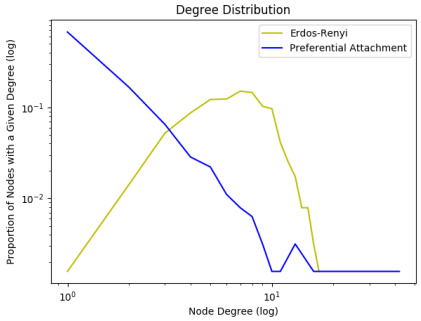
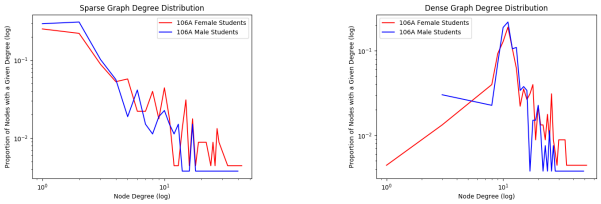


Fig. 4: Random graph degree distributions.

We then compare the degree distributions across genders (female and male only because we did not have enough data from students with other gender identities). From these degree distributions, shown in Figure 5, we see that female students seem to have higher degrees compared to male students in both the sparse and dense network.



(a) Sparse network degree distributions by gender

(b) Dense network degree distributions by gender.

Fig. 5: Comparing degrees by gender.

To explain why female students might have more outgoing edges than male students, we compared the average number of edges from female and male students to section leaders in the LAIR in Figure 6. We see that on average, female students have more edges to LAIR helpers, which boosts their overall degree values.

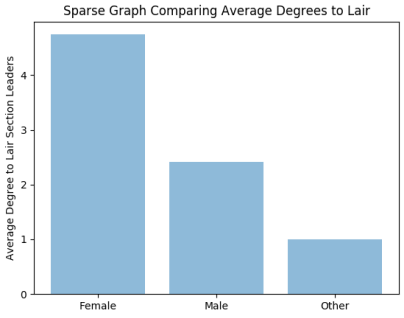
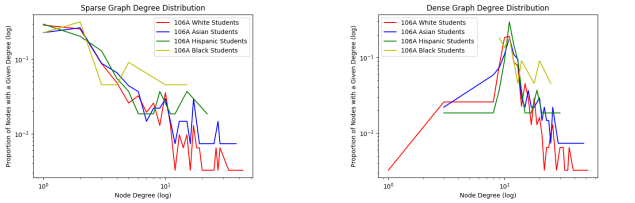


Fig. 6: Average number of connections to section leaders in the LAIR.

We then look at the degree distributions of students by race in Figure 7. From these degree distributions, we see that the plots for underrepresented minority students in Computer Science (Hispanic and Black students) cut off after a certain point; that is, they do not have the higher degree counts that some White and Asian students do.



(a) Sparse network degree distributions by race

(b) Dense network degree distributions by race

Fig. 7: Comparing degrees by race.

Further, in Figure 8, we see that on average Black and Hispanic students have lower degrees. Because our current race and ethnic categorizations are estimated from an API, which consistently labels Black-American names as White, the lower degrees of some Black students may be masked by the higher degrees of some White students in this chart. We believe that with accurate race data we would see greater disparities between Black students, and White and Asian students, than this graph presents.

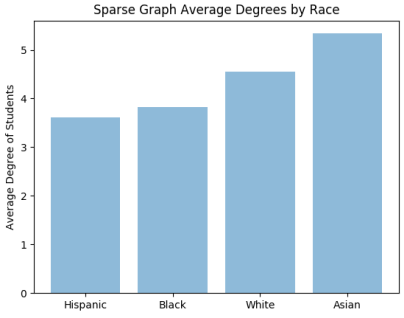
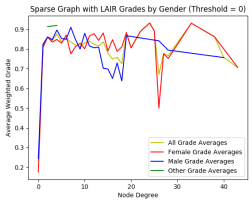
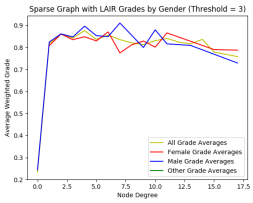


Fig. 8: Average degree of students by race.

To begin answering one of the questions guiding our project, about how student connections affect performance, in Figures 9 and 10 we visualize the relationship between students' degrees and outcomes in the course (as weighted grade averages). For each degree, we average the grades of students with that degree. We looked at results on both the sparse and dense networks, but the plots look quite similar since on average, students have about the same number degrees connecting them to other students in their sections, so these additional connections to do not reveal anything more about their performance.

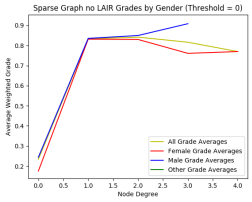


(a) With LAIR, Threshold 0

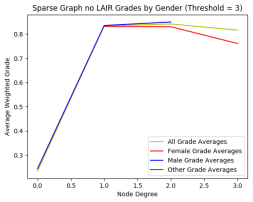


(b) With LAIR, Threshold 3

Fig. 9: Student degrees (with LAIR) and average grade.



(a) No LAIR, Threshold 0



(b) No LAIR, Threshold 3

Fig. 10: Student degrees (no LAIR) and average grade.

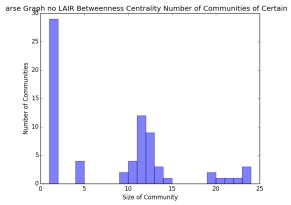
In all plots in Figures 9 and 10, we see a sharp increase in grade average from degree 0 to degree 1, explained by students in the course who may have worked remotely and may not have had access to a section or other course assistance. Further, in all plots we see a general decrease in student performance as degree increases. To highlight this trend, we applied a threshold to the data we plot, only including points with 3 or more students of a given degree. With the threshold applied, the downward trend becomes clearer.

We initially presumed this trend could be explained by students receiving help in LAIR, so students with more connections to the LAIR, the students in need of more help, are also the students who perform worse on average in the course. To test this hypothesis, we created the same plots, excluding edges to the LAIR in Figure 10. Excluding LAIR edges, we see the same downward trend as students collaborate with more than 2 other students. From this, we can conclude that in CS106A working with 1 or 2 others (section leaders or students) aids performance, but that performance is hindered as one's number of collaborators increases beyond this.

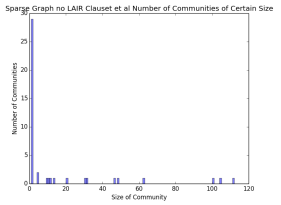
B. Community-finding algorithms

Now, we turn to the four community-finding algorithms we used on our networks. When running all of our community-finding algorithms, we excluded connections to the LAIR. We chose to exclude connections to the LAIR because students do not typically interact with a section leader from LAIR more than once, and students do not pick which section leader will help them in LAIR, so these connections are not community-like.

In Figures 11 and 12 we plot the sizes of communities found by each of these algorithms. We look at the sizes of the communities found on the sparse network, when we do not consider LAIR edges, and in other words look at the very most meaningful edges between a student and their section leader and their direct collaborators.

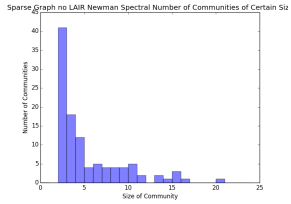


(a) Communities from the Girvan-Newman algorithm

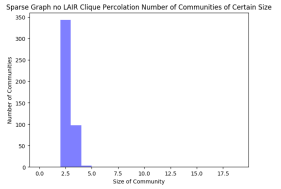


(b) Communities from the Clauset-Newman-Moore algorithm

Fig. 11: Communities found using built-in functions



(a) Communities from Newman Spectral algorithm



(b) Communities from the Clique Percolation algorithm

Fig. 12: Communities found using implemented community-finding algorithms

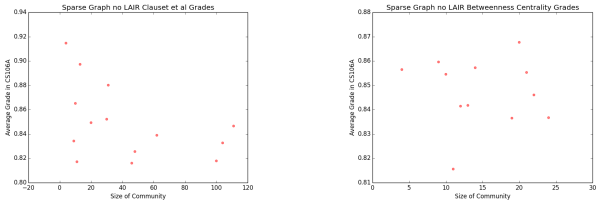
We see that the Newman Spectral algorithm produces communities varying the most in size, while the Clique Percolation algorithm has the most uniformly sized communities. We set the parameter $k = 3$ for the Clique Percolation algorithm because this produced communities with the most variety in size albeit not very much. This was to be expected, however, as we were looking at the sparse network and were relying on the smaller number of student-student collaboration links.

With these community-finding algorithms in place, we then asked several questions about the communities we found using each of these algorithms:

1) How does community size relate to how student's perform in that community?
2) How does a community's diversity relate to student performance?

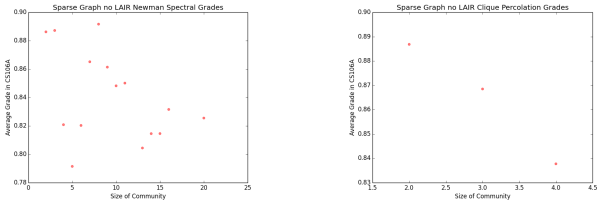
3) For Clique Percolation, how does belonging to multiple communities relate to a student's performance?
4) How does community diversity relate to community size?
5) Do sections make communities in CS106A more diverse?

To consider question 1, we plot community size and average student grades in Figures 13 and 14.



(a) Clauset et al (b) Girvan-Newman

Fig. 13: Community size against average grade (built-in methods)

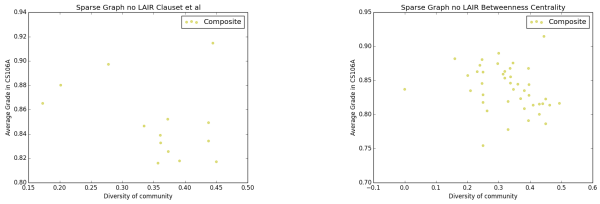


(a) Newman Spectral (b) Clique Percolation

Fig. 14: Community size against average grade (implemented methods)

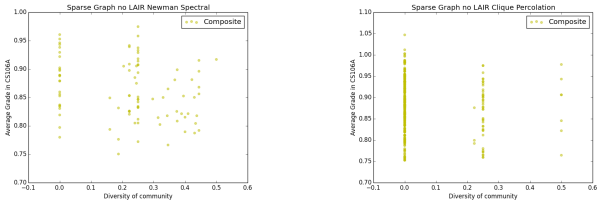
The plots in Figures 13 and 14 of community size against average grade aren't conclusive, but they do show an inverse correlation between community size and average grade. We could make sense of this as follows: Given our data, the people who will have a lot of links within themselves and not many outside of them (and thus form the small communities) will be the groups that have decided to work as partners on assignments from the very beginning and thus go to the same people for help. When the algorithms can't distinguish that a student has a surprising number of links with just a few people, it places that student in a bigger community, and these are the people that end up performing more poorly.

Question 2 is particularly interesting because we wanted to test the idea that more diverse groups perform better. Our experiment for this question involved the following plots:



(a) Clauset et al (b) Girvan-Newman

Fig. 15: Diversity of community against average grade (built-in methods)

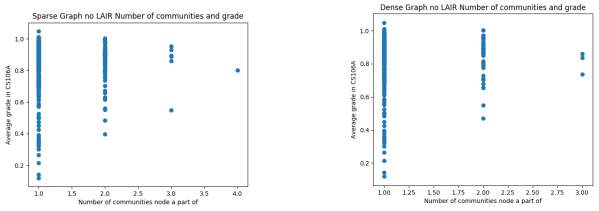


(a) Newman Spectral (b) Clique Percolation

Fig. 16: Diversity of community against average grade)

Given our data and from these plots, we unfortunately cannot conclude that these community finding algorithms support the above idea. As for future work, we need to formulate more refined methods of testing this question,

For our third question, we wanted to see how being in multiple communities might relate to a student's performance in the course. This question clearly only pertains to the Clique Percolation algorithm. In Figure 17, look at the number of communities a student is part of versus their performance in the course. We do not see any definitive relationship between these two variables, as there is a wide spread of student performance with students who are part of just one community, and very few students who are part of more than two communities. The students who do belong to three or four communities do seem to perform worse on average compared to their peers, but there are so few data points for these students, that is difficult to extrapolate from their course performance.



(a) Sparse Network Overlaps (b) Dense Network Overlaps

Fig. 18: Sparse network community size against diversity (built-in methods)

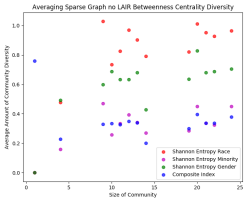
To answer question 4, of how community size relates to diversity, we plot (in Fig. 19-22) the relationship between the size of community and its diversity using Shannon's Entropy

Comparing Community Diversity				
	Girvan-Newman	Clauset et al	Newman Spectral	Clique Percolation
Shannon Entropy Sparse (Race)	0.790716097138	0.837800110151	0.71972222945	0.120561037502
Shannon Entropy Dense (Race)	0.838018033368	0.819748348936	0.853223619038	0.80628247452
Shannon Entropy Sparse (Minority)	0.314594436776	0.337307883948	0.292190309854	0.0593437596547
Shannon Entropy Dense (Minority)	0.337309714387	0.327021419662	0.347000209312	0.321727158797
Shannon Entropy Sparse (Gender)	0.6043196414	0.617909190572	0.536625877121	0.0855754702467
Shannon Entropy Dense (Gender)	0.618365834477	0.61382044685	0.627957757813	0.633692919777
Sullivan Composite Index Sparse	0.372760374466	0.357170081339	0.280307408001	0.064290529154
Sullivan Composite Index Dense	0.356913701339	0.354364904262	0.323322622882	0.320748808973

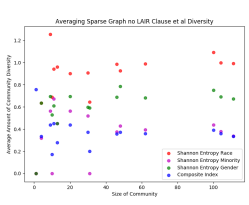
Fig. 17: Diversity found in communities.

to measure diversity in terms of race, underrepresented-minority-status, and gender and Sullivan's Composite Index, in which we look at how diverse communities are in terms of minority-status and gender. We focused the Composite Index on these two variables because these are the areas in which most Computer Science programs and the technology industry at large lacks diversity. For each of these metrics, we looked only the diversity of students within a given community, because these are the communities that are self-made to a great extent. In contrast, students do not pick their section leaders or who will help them in the LAIR, so we did not consider the race, minority-status, or gender of any section leaders.

We created these plots for both our sparse and dense networks. We see in both sets of plots, that as the size of students' communities increases the diversity of their communities does as well. This is promising because it means that as students have more contact with a broader array of other students in CS106A, they are more likely to form diverse communities.

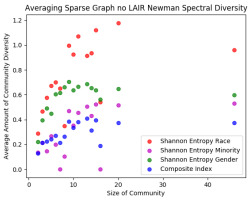


(a) Girvan-Newman

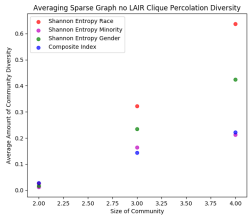


(b) Clauset et al

Fig. 19: Sparse network community size against diversity (built-in methods)

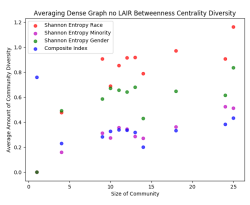


(a) Newman Spectral

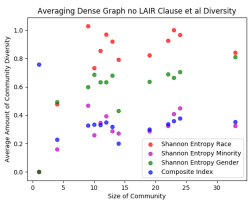


(b) Clique Percolation

Fig. 20: Sparse network community size against diversity (implemented methods)

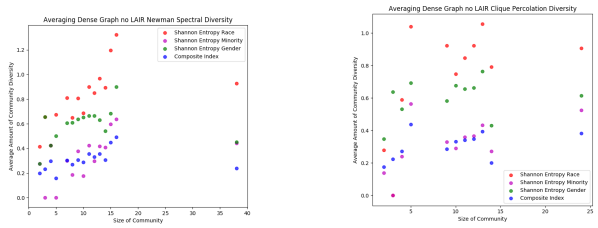


(a) Girvan-Newman



(b) Clauset et al

Fig. 21: Dense network community size against diversity (built-in methods)



(a) Newman Spectral (b) Clique Percolation

Fig. 22: Dense network community size against diversity (implemented methods)

Finally, to answer question 5, of whether sections perform the function of bringing diverse groups of students together, across the categories of race, underrepresented-minority-status, and gender, we looked at the diversity of the communities found by each of the four community algorithms on both the sparse network and the dense network. In the sparse network considered by our community algorithms, students only have connections to their section leader and direct student collaborators, and in the dense network, students have connections to their section leader, direct student collaborators, and their all other students in their section. By comparing these two networks, we are measuring how much having a section increases the diversity of communities formed in CS106A.

We summarize the results we found from running each of these algorithms in the table in Figure 17. Diversity values closer to zero indicate that there is less diversity, while diversity values that are further from zero indicate there is greater diversity. We see that the communities found in the dense network by each of the four algorithms are just about as diverse or more diverse than the communities found in the sparse network, and this is always the case for the communities found by the Newman Spectral and Clique Percolation algorithms. This means that sections do indeed perform the function of diversifying the communities students interact with in CS106A, and work to integrate students in the course.

V. FUTURE WORK

To continue the work of analyzing the social networks of Computer Science courses at Stanford would require acquiring self-reported data from students in the form of surveys. This would allow us to get more accurate information on the race and gender of students, and would allow us to get a larger picture of collaborations beyond those listed on assignments (i.e. to partners on assignments they did not list, to outside tutors, or to other friends or dorm-mates in the course who may they have worked with or studied for exams with, but are not listed in our data). Furthermore, we know that diversity means a lot more than gender and race (factors such as first-generation, socioeconomic status, just to name a few), and these added dimensions of diversity would only enrich our studies.

ACKNOWLEDGMENTS

This work would not have been possible without the support of Chris Piech, Mehran Sahami, and the CS198 program.

REFERENCES

[1] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[2] C. Ingraham, "Three quarters of whites dont have any non-white friends," *The Washington Post*, 2014.

[3] C. M. Steele, *Whistling Vivaldi: How stereotypes affect us and what we can do.* New York, NY, US: WW Norton & Co, 2010.

[4] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. cond-mat/0112110, pp. 8271–8276, 2001.

[5] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[6] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *Proceedings of the 17th international conference on World Wide Web*, pp. 695–704, ACM, 2008.

[7] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.

[8] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[9] S. Lattanzi and D. Sivakumar, "Affiliation networks," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 427–434, ACM, 2009.

[10] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," 2008.

[11] J. Ye, S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, and S. Skiena, "Nationality classification using name embeddings," *arXiv preprint arXiv:1708.07903*, 2017.

[12] F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier, "Inferring gender from names on the web: A comparative evaluation of gender detection methods," in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 53–54, International World Wide Web Conferences Steering Committee, 2016.

[13] J. E. McLaughlin, G. W. McLaughlin, and J. McLaughlin, "Using composite metrics to measure student diversity in higher education," *Journal of Higher Education Policy and Management*, vol. 37, no. 2, pp. 222–240, 2015.