2011 William Aberhart High School Network Analysis Project

The project:

We attempted to construct the association/friendship network for the tenth grade at William Aberhart High School. In this network, students are *nodes* and two students are connected with a *link* if they are friends. We developed a survey (see *Methods*) that would allow us to determine not only who was friends with whom, but also the *strength* of such friendships. We could include these strength measurements by assigning *weights* to the network's links; "best friends" would then have a higher weight on their connecting edge than "acquaintances".

The idea behind this project was that we could analyze some graph-theoretical properties of our network to determine, for example, which individuals have the most social links, who the most *central* people are (i.e. who are the people through which the most "friendship paths" pass), what cliques exist in the network, etc. Ultimately, we wanted to make a simulation of a disease propagating in this network and then try different network rewiring strategies to see how we could most effectively stop or slow the propagation of the disease without too much disruption of the underlying network (e.g. Which individuals or groups might we target for vaccination or quarantine?).

After developing a survey that would help us extract the social network of Aberhart's 10th grade, we ran into some difficulties getting permission to administer the questionnaire because of possible privacy issues. Instead, we were able to obtain anonymized student class schedules, and from these we constructed a network of student-class relationships (to be discussed in more detail in *Networks*). We were unable to get as far as simulating disease dynamics on our network(s), but we did look at the graph-theoretical properties of these networks (discussed in *Network Properties*) and from this analysis, we were able to draw some interesting conclusions about the structure of the student-class relationship network(s).

The Networks:

Initially, we had planned to take the results of our survey (see *Methods*) and construct a single network in which nodes would represent students, links would indicate existing social relationships, and link weights would correlate with the strength of these relationships. However, as mentioned above, due to confidentiality issues, we ended up constructing networks from anonymized student schedules. In the largest of these networks, which included all tenth grade students and all of the classes that they take, a node can be *either* a student or a course. A link *from a student to a course* indicates that a student is enrolled in a particular class. This type of network is *bipartite*, meaning that it can be separated into two groups (in this case, *Students*, and *Courses*); there are no links *within* a group (i.e. students are not connected to other students and courses are not connected to other courses), but there are links *between* the groups.

This type of bipartite network can be transformed into a *unipartite projection*, meaning that it can be "reduced" to include only one type of node. In our case, we could transform the "total network"-- i.e. the network containing students and courses-- into a network of "just students" or a different network of "just courses". Although we lose some information, we can create a network of "just students" by directly connecting individuals who share one or more courses. A link between two students would then be weighted by the *number of courses they share*. A similar game can be played to construct the "just courses" graph. In this case, two courses are connected if they share one or more people. A link on this graph would be weighted by the number of people who share the courses. All analysis reported on in the *Results* section was performed on the unipartite projection graphs.

Methods

(1) Creating a questionnaire for determining the social network of Aberhart's 10th Grade:

Although we did not end up using this questionnaire (see above for discussion), we spent a good deal of time brainstorming questions and figuring out controls that would allow us to extract-- to a rough approximation-- the friendship network of 10th grade students at Aberhart. The questionnaire we eventually devised is given below:

Survey of 10th Grade Students

All information gathered from this survey will be kept strictly confidential.

Instructions:
-For all answers involving the names of *people*, please use the name/number reference list provided by the survey administrator to find the random number associated with a name. To ensure confidentiality, please report this number and NOT the name when answering questions.
-Students included in this questionnaire must be 10th graders at Aberhart.
-All durations of interactions need only be approximate.

Your Name: _____

1. If you often attend a club or other group activity during the **lunch** hour, please list the club/activity name.

2. With whom do you interact at **lunch** on days during which you do not participate in a club? (i.e. With whom do you sit for the majority of the lunch period?)

3. Please list any **school sponsored** extra-curricular activities (before & after school) in which you participate.

4. **Non school sponsored**, organized extra-curricular activities:
Please complete the following table for all out-of-school activities in which you participate with 10th grade students from Aberhart

Activity Name	List of fellow students	Days & Times the activity meets

5. **Other outside-of-school interactions** (including friendships & transportation) – at least 30 minutes in duration.

Student's name	Length of interaction (on average)	Number of days per week of interaction

6. **Virtual interactions** (playing online games, live chatting, texting, etc.)

Student's name	*Type of Interaction*	Length of interaction (on average)	Number of days per week

Thank you for your participation ☺

Had we been able to administer this questionnaire, we would have created a social network of student-student interactions that were weighted by the amount of time individuals spent together. The data could have been merged with student class schedules to give a more complete description of individuals' daily social interactions.

For reasons mentioned in *The Networks*, we were unable to administer this survey, and so our final interaction networks *only* included student schedule data.

(2) How we analyzed our networks

We chose to look at four dominant graph-theoretical measures when analyzing our networks. These were:

1) The *degree distribution*-- i.e. the probability distribution for the number of nodes in a network with degree greater than or equal to *k*;

2) The *clustering coefficient-- i.e.* the extent to which nodes tend to group together. Officially, the local clustering coefficient is the number of connections *among* neighbors of a node, divided by the total number of connections that *could exist among* these neighbors. A clustering coefficient of 1 means that all nodes in a local neighborhood are connected to all other nodes in that neighborhood;

3) The *betweenness centrality*-- i.e. the number of shortest *ij*-paths that pass through node *k*, divided by the total number of these shortest paths;

4) The *community structure*--community structure has many different possible definitions and there are also many algorithms for finding communities, or "groups" of nodes, in graphs. We chose to define a community as a set of nodes for which the interlinking was denser than linking between these nodes and other nodes in the graph. This definition is essentially the idea of *modularity*, put forth by Newman, *et al. [1].*

We wrote our own Python code and implemented subroutines from NetworkX (a Python-based code suite for analysis of complex networks) for the basic graph-theoretical analysis of our networks. For the community finding algorithm, we relied on out-of-the-box code for the *Louvain Algorithm* [2] for finding modular structure in networks.

(3) Results

After the full graph, containing all courses and all tenth grade students was collapsed into projections-- either a graph of people (PeopleGraph), connected if they shared courses, or a graph of courses (ClassGraph), connected if they shared students-- the unipartite projections were analyzed from a graph theoretical standpoint. The results are reported, below.

i) General characteristics of the unipartite graphs

The number of nodes in the PeopleGraph and ClassGraph is equal to the number of tenth grade students and courses, respectively. These numbers, and the total number of edges in each graph, are reported in Table 1. Note that as an alternative to edge-weighting, we allowed for parallel edges in our graphs-- i.e. two nodes could share *more than one* edge, meaning, in the case of the PeopleGraph, that two students could have more than one class in common, or that, in the case of the ClassGraph, two classes could share more than one student. The relative density (edges per node) in the ClassGraph is lower (12.1) than it is in the PeopleGraph (44.1). This reflects the fact that on average, if classes contain ~20-30 students and if each student takes (typically) a maximum of 4 classes, the upper bound for edges adjacent to a node in the ClassGraph is essentially ~120, but is unlikely to occur (students can, to some extent, select the courses they take, so, presumably, they might select sections in which they have friends); in the PeopleGraph, however, it is fairly likely that there will be little overlap in the students comprising each of the 4 classes an individual takes, and therefore the number of edges adjacent to an individual may well be ~120.

Table 1: Sizes of unipartite projection graphs, ClassGraph and PeopleGraph.

Network	Number of Nodes	Number of Edges
ClassGraph	131	1,587
PeopleGraph	520	22,957

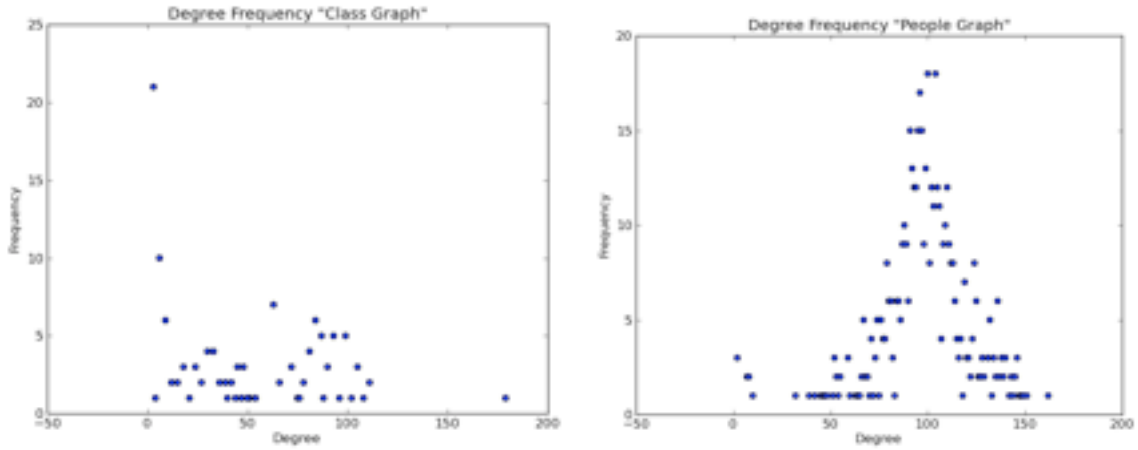
ii) Degree Distributions

For reasons explained above, the average degree of the PeopleGraph is substantially higher than the average degree of the ClassGraph (*<k>=88.3* vs. *<k>=24.2*). The argument presented in *(i)* for the relative difference in edge densities between the ClassGraph and the PeopleGraph can be seen clearly in Figs. 1(a,b). The peak in the ClassGraph degree distribution occurs at very low degree, indicating that many classes

share few students; on the other hand, the degrees of the PeopleGraph are roughly Gaussian distributed and peak around *k=100*.

Figure 1: **(a)** Degree distribution of the ClassGraph. Many classes have degree *k<10*, indicating they share few students. **(b)** Degree distribution of the PeopleGraph. Most people have degree 80<*k120*.

(a) **(b)**



The degree distributions presented in Fig. 1 give rise to interesting cumulative degree distributions (i.e. the probability that a node will have degree greater than *k*). It is often the case that complex networks exhibit so-called *scale-free* degree distributions (see, for example, [3]), where the probability that a node has degree *k* is a power-law of the degree, $P(k) \sim k^{-\alpha}$. When plotted on a log-log scale, this type of distribution is a straight line, whose slope is α. Interestingly, neither the ClassGraph nor the PeopleGraph exhibits this type of distribution. As can be seen in Fig. 2, the ClassGraph cumulative degree distribution is linear, while that of the PeopleGraph is S-shaped (Fig. 3). Again, the *S-shaped* curve of the PeopleGraph cumulative degree distribution corroborates the notion that most people probably take ~4 courses, each of which has ~20 students, and all of which are largely disjoint in terms of the sets of students that take them. This is evidenced by the fact that more than 60% of the nodes in the network have degree *k>80*.

Figure 2: ClassGraph cumulative degree distribution. Distribution is approximately linear.

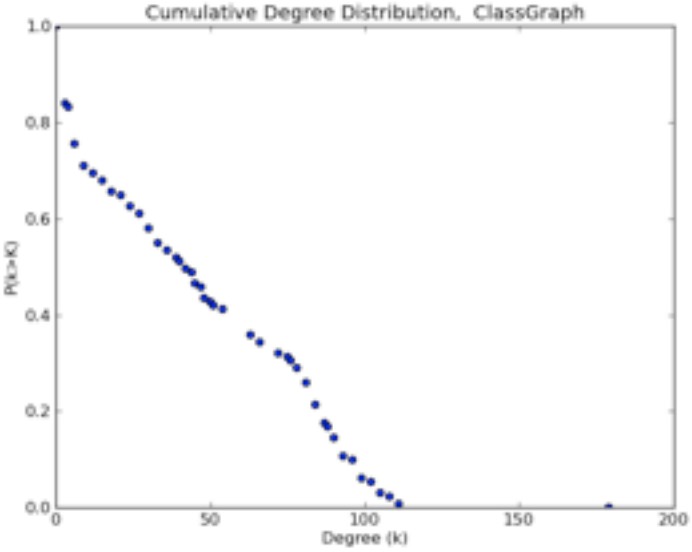
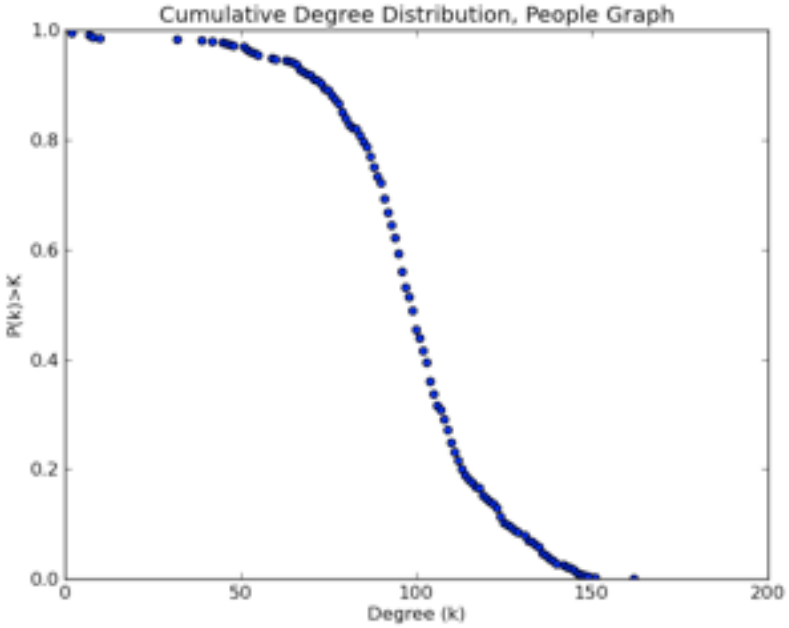


Figure 3: PeopleGraph cumulative degree distribution. Distribution has a characteristic *S-shape* and demonstrates switching behavior between *80<k<120*.



iii) Clustering Coefficients

In an unweighted graph, the local clustering coefficient is simply

$$C_i = \frac{2t_i}{k_i(k_i - 1)} \tag{1}$$

where t_i is the number of triangles in which node *i* participates, normalized by the maximum possible number of such triangles, given that node *i* has degree k_i. We can also extend the idea of a clustering coefficient to weighted networks [4], such that the definition becomes

$$C_i = \frac{1}{s_i(k_i - 1)} \sum_{i,k} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{jk} a_{ik} \tag{2}$$

where w_{ij} are the edge weights, where $a_{ij} = 1$ if there is an edge between node *i* and node *j*, and 0 otherwise, and where $s_i = k_i < w_i >$. Basically, the form of Eq. 2 causes each triangle's contribution to be weighted by a factor that is equal to the ratio of the average weight of two adjoining edges in the triangle to the average weight of node *i*.

It is interesting to compare the weighted and unweighted forms of the clustering coefficients for the ClassGraph and the PeopleGraph. While the ClassGraph shows greater clustering than the PeopleGraph in the unweighted case, when weighting is taken into account, the PeopleGraph is more highly clustered than the ClassGraph (see Table 2).

Table 2: ClassGraph and PeopleGraph (global average) clustering coefficients.

Network	Unweighted Clustering Coefficient	Weighted Clustering Coefficient
ClassGraph	0.5062	0.0734
PeopleGraph	0.4072	0.1179

Given that individuals within a specific class are completely connected (i.e. everybody has a link to everybody else), which would, *a priori*, suggest a clustering coefficient of 1, it might be a bit surprising that the unweighted global average clustering coefficient in the PeopleGraph is relatively low. We suspect that while individuals are tightly clustered within a given class, there is little overlap *between* classes, and therefore many of the neighbors of a specific node will remain unconnected, ultimately resulting in a low clustering coefficient.

iv) Betweenness Centralities

As was the case with the clustering coefficient, it is possible to calculate both an unweighted and a weighted edge betweenness centrality. In addition to calculating the global averages of these quantities for the two graphs (see Table 3), we also identified the nodes in each network with the top 10 betweenness centralities. For the ClassGraph, only 3 of the 10 classes with highest betweenness centrality were core subjects (i.e. English, Math, Science, Social Studies, Phys. Ed., etc.); the other classes were electives. In fact, this hints at the possible composition of the underlying community structure of the ClassGraph: namely, groups of core subjects are likely densely connected into small communities, which are, in turn, sparsely interconnected through electives.

The betweenness centrality can often help in anticipating how information (or disease) will propagate through a network, as nodes with high betweenness centrality will be "central" to the flow on the network-- i.e. a great number of paths will have to pass through them. For certain network structures, this can also be true for the nodes of high degree. To this end, we were interested in determining whether or not the top 10 betweenness centrality nodes of each network were also the top 10 nodes of highest degree. While there was some overlap between the classes of high betweenness centrality and the classes with high degree (45% overlap for the ClassGraph and 50% overlap for the PeopleGraph), the *character* of the top 10 classes of high degree is distinctly different from that of the high betweenness centrality classes: 8 of the top 10 high degree classes were core subjects.

Table 3: Global average and maximum edge betweenness centralities for the ClassGraph and the PeopleGraph.

Network	Unweighted Betweenness Centrality	Weighted Betweenness Centrality	Maximum Unweighted	Maximum Weighted
ClassGraph	0.006135	0.007627	0.041637	0.051887
PeopleGraph	0.001545	0.00158	0.003696	0.003830

We were also interested in determining how frequently the top 10 high betweenness centrality *people* appeared in the top 10 high betweenness centrality *classes* (people-class overlap), and, conversely, what fraction of a top 10 high betweenness centrality person's class schedule was constituted by top 10 high betweenness centrality classes (class-people overlap). In both graphs, while the maximum centrality is at least twice the average measure, it is still not particularly high. If the maximal centrality measures were close to one, we could probably expect that people taking classes with high

betweenness centrality in the ClassGraph would, themselves, have high betweenness centrality in the PeopleGraph (and conversely). However, this is not the case with our networks, so it is therefore not surprising that the overlap previously discussed is rather low: class-people overlap ranged from 0% to 50%, with an average of 20.45% (roughly one class), while people-class overlap ran between 0% and only 9%, with an average of 2.6% (roughly one person).

v) Community Structure

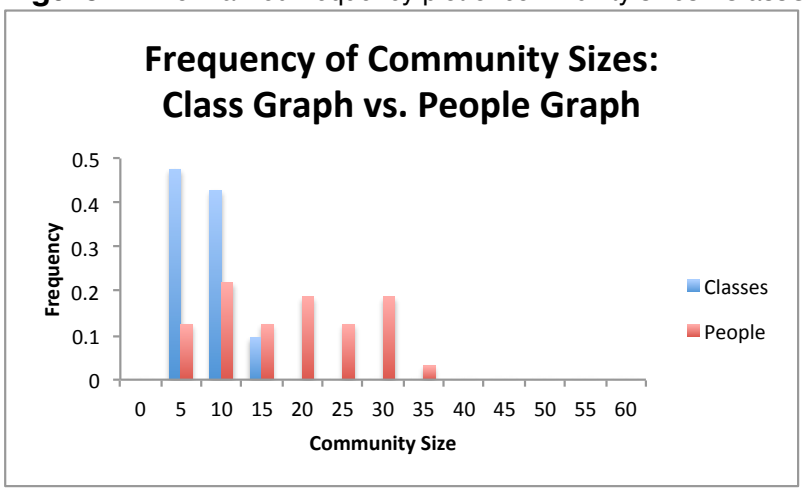
The final graph-theoretical feature that we examined was the community structure of each of our unipartite projections. When looking for communities, we were primarily interested in identifying groups of nodes-- a community-- for whom the density of interlinks was greater than the density of links to other communities. We used a community-finding algorithm to first identify large communities within the global structure of each network, and we then further divided these communities into a second stratum of subcommunities.

The community finding algorithm we employed identified 10 main communities within the PeopleGraph and 6 main communities within the ClassGraph. The algorithm was immediately successful in classifying, as one community within the ClassGraph, a group of nodes constituting Aberhart's *Assisted Learning Program*.

When we further subdivided the ClassGraph communities into subcommunities, we found that 20 of the 21 subcommunities were composed of at least 40% core courses. The average and median core course percentages were 57% and 56%, respectively. This result fits nicely with our initial hypothesis of the ClassGraph structure from studying betweenness centrality: namely, the network is composed of small groups of core courses that are then interconnected through electives.

Perhaps not surprisingly, there is a large difference between the average size of ClassGraph communities and the average size of PeopleGraph communities (6 vs. 16). The distribution of community sizes for the PeopleGraph is much broader than for the ClassGraph (Fig. 4).

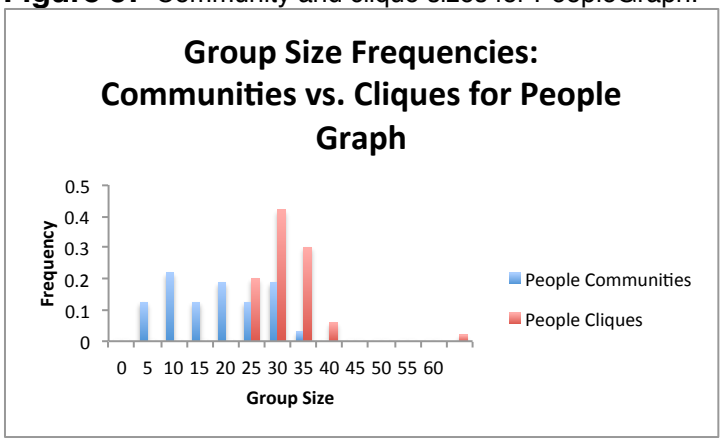
Figure 4: Normalized frequency plot of community sizes: ClassGraph vs. PeopleGraph.



We were not surprised that the community sizes for the PeopleGraph were larger than for the ClassGraph, and assumed that this result arose because a class would manifest itself in the PeopleGraph as a completely connected set of nodes, and therefore, as a-- we guessed-- maximal community. Since the average class size was ~20-30, we expected that *most* communities in the PeopleGraph would have size ~20-30. We were surprised to find that this is not the case.

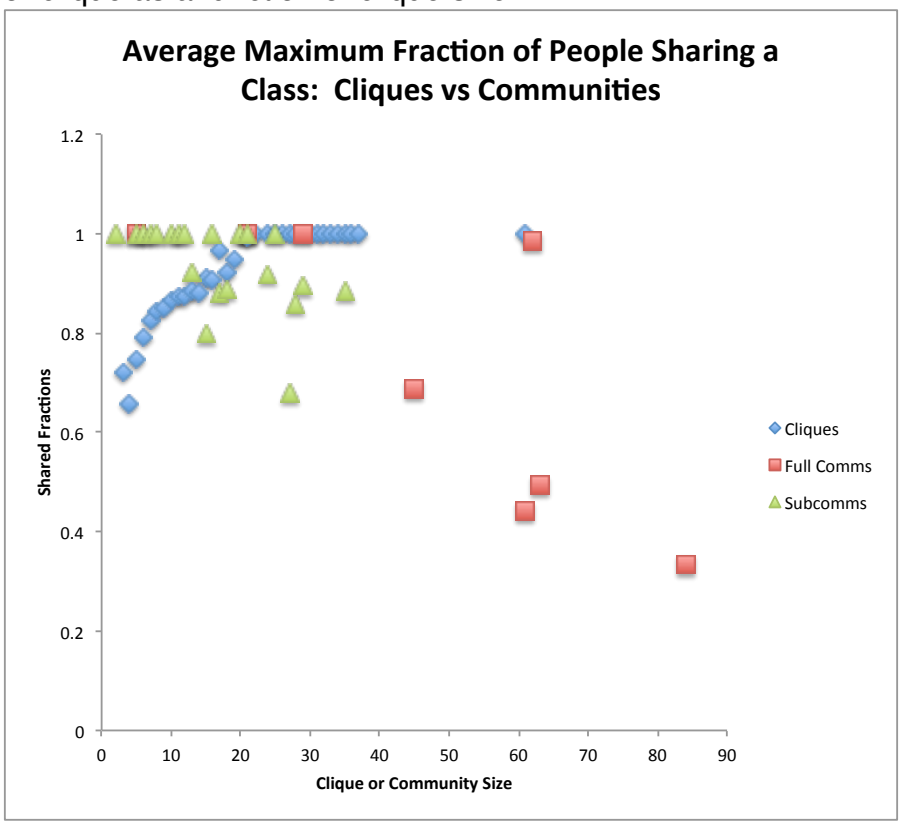
To gain a better understanding of why our hypothesis failed, we compared the sizes and compositions of the 50 largest cliques (completely connected subgraphs) in each network to the sizes and compositions of the communities identified with the community-finding algorithm. For the ClassGraph, all 50 of the largest cliques had size 5, which was comparable to the average and median of the sizes found for the communities. In the PeopleGraph, however, a marked difference was evident (Fig. 5): while the community size distribution is broad, most of the 50 largest cliques had size between 25 and 35, indicating that they are, truly, individual classes.

Figure 5: Community and clique sizes for PeopleGraph.



To further track this difference between cliques and communities, we asked, for each person in a PeopleGraph community, what fraction of *classes* are shared with *all* other people in the community? If communities in the PeopleGraph were formed solely from classes, and therefore constituted cliques, we would expect this fraction to be exactly 1. If we plot this fraction as a function of community or clique size, we can see that for cliques, as the size of the clique passes 21, the fraction goes to 1, corroborating our hypothesis that large *cliques* are simply classes (Fig. 6). However, we can also see that the opposite trend is true for communities and subcommunities; as the community size grows, the fraction of classes that an individual shares with all other members of his community falls off (Fig. 6), though even for very large communities, it never dips below ~30% (roughly one course). What this suggests is that the community detection algorithm tends to pick out class participation overlap *between* courses.

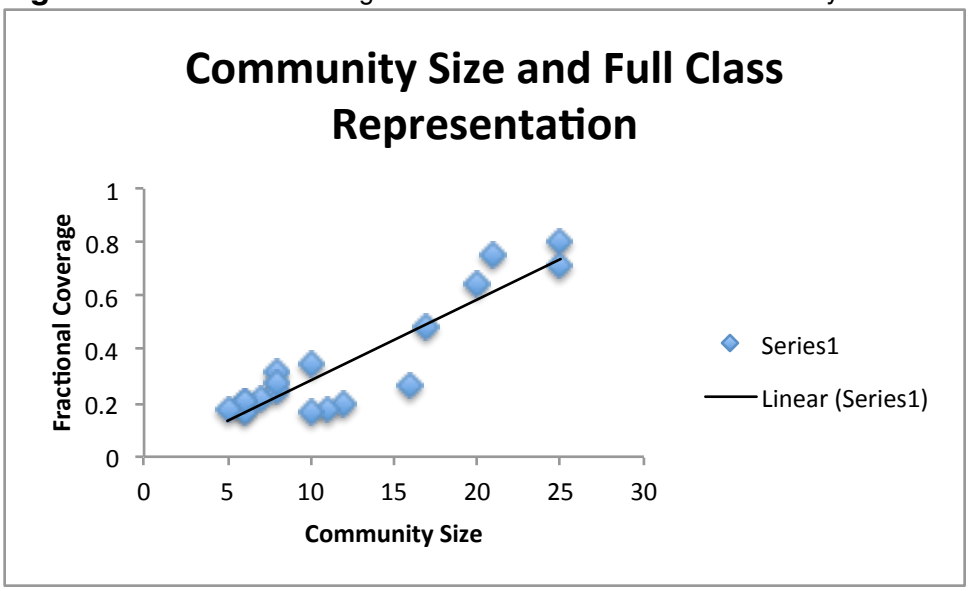
Figure 6: (Maximal) fraction of classes shared with all other members of a community or clique as a function of clique size.



This phenomenon can be seen with a different approach. If, for a community in the PeopleGraph, we identify courses in which every person in the community participates and then ask what fraction of the full course the community size constitutes, we can see (Fig. 7) that the fraction increases with community size, but never actually reaches 1.

This implies that the most dense abundance of links is likely achieved *not* by taking the full completely-connected subgraph that represents an individual course, but by taking a section of this subgraph whose nodes, in general, share *more than one* pairwise connection (i.e. students who are in more than one class together).

Figure 7: Fractional coverage of classes as a function of community size in the PeopleGraph.

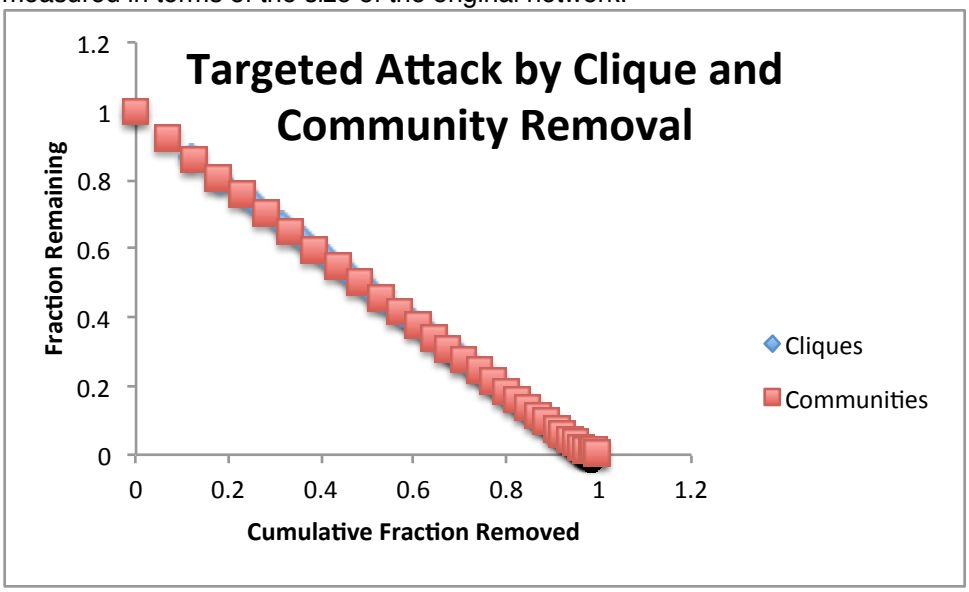


When we "attacked" the PeopleGraph network by removing cliques (respectively, communities) in descending order of size and looked at the size of the largest remaining connected component (a set of path-connected nodes), we found a linear relationship between the size of the removed clique (community) and the size of the remaining connected component. Moreover, there was almost no difference between clique removal and community removal (Fig. 8). It is interesting to note that at face value, what this tells us is that we do the same "damage" to the graph whether we remove a *clique* of size *S* or a *community* of size *S*. However, a community of size *S* presumably has a greater density of paths than an equivalently sized clique, and, as stated previously, many of these paths-- at the level of individual edges-- may be redundant. Though removing a community of a given size has virtually the same effect on network breakdown as removing a clique of the same size when edge redundancy is ignored, the scenario may change if edge multiplicity is considered important.

For example, if we think about disease propagating through a portion of a graph, spreading from infected nodes to susceptible nodes along connecting edges, the probability for a susceptible individual to become infected may in some way be tied to the edge density between the susceptible node and the infected node (for example, if each edge represents a certain amount of time spent together, more edges would correlate with a longer exposure period and a higher probability of infection). Thus, if we were looking to remove or isolate groups of students to prevent the further spread of

a disease, there could be an intrinsic advantage in removing communities, as opposed to cliques, since a disease may spread more quickly within a community.

Figure 8: Targeted attack of the PeopleGraph by selective removal of cliques and communities in descending order of size. After each removal the size of the largest remaining connected component is measured in terms of the size of the original network.



Discussion and Summary

Though our project ultimately underwent significant redesign, we were, in the end, able to capture a rough representation of the social/academic network of the tenth grade at William Aberhart High School, eventually decomposing the full network into two unipartite projections. Graph theoretical analyses of these projections yielded interesting information about the underlying social/academic "scaffolding" of Aberhart's tenth grade.

In future work, we hope to explore disease and/or information transmission on the complex networks we have developed. In particular, we would like to test our hypothesis that removal of communities might be more advantageous for preventing the spread of disease (information) than would removal of cliques.

References

[1] M. E. J. Newman (2006). "Modularity and community structure in networks". *Proc. Natl. Acad. Sci. USA* **103** (23): 8577–8582

[2] Vincent D. Blondel *et al.* (2008). "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment* **(10)**: P10008

[3] R. Albert and A.-L. Barabasi (2002). "Statistical mechanics of complex networks". *Reviews of Modern Physics* **(74)**: 47-97

[4] A. Barrat *et al*. (2004) *Proc. Natl. Acad. Sci. U.S.A.* (**101**): 3747