An Analysis of Sexual Interactions in a Student Group

David Borowitz
Stanford University
borowitz@stanford.edu

Fred Wulff
Stanford University
frew@stanford.edu

ABSTRACT

We explore a novel dataset of sexual interactions between members of a student group. We find that the dataset exhibits organic clustering that does not coincide with official social groupings. We compare the set to the Jefferson High set in Bearman's "Chains of Affection" paper and find that dataset is a plausible extension of similar graph properties taken over a longer period of time. Our results may have applications to STD research in epidemiology, and we suggest directions for a more comprehensive work that would likely result in an evolutionary model for this type of sexual interaction network.

1. INTRODUCTION

In this paper we analyze a unique social network containing a record of sexual interactions over 20 years between students in a large voluntary student organization at Stanford University. Our goal in this project is to analyze the structure and composition of this network in the context of other networks capturing similar interactions, as well as from a more generic community-finding perspective. Previous research in this area has been well cited, both by researchers from the field of epidemiology working towards effective models of sexual transmitted diseases and by studies of other network types, suggesting wide applicability of this type of analysis[6].

2. DATASET

The network comprises 477 nodes and 732 edges, where an undirected edge between two nodes indicates that those students had a sexual encounter at some point after joining the student group. We also obtained the sex and graduation year[1] of each student. In addition to this demographic data, we also recorded two pieces of data specific to this student group for each node. First, the student group itself is subdivided into several subgroups, which the students self-select themselves into upon joining the group[2]; we will call these subgroups *sections*. Second, in order to capture individuals' level of involvement in the group, we count the number of leadership positions (*staff positions*) each student holds over the course of membership; since each leadership position involves a significant extra time commitment, we regard the students acquiring more staff positions as more involved in the group.

The sexual encounters recorded in the network span students from 20 graduation years (1985-2014), and were collected over the course of several years through oral interviews performed by one individual as a hobby, starting around 2004. Naturally, although total membership in the student group has remained roughly constant over that time period, data for older interactions is somewhat more sparse. This is primarily due to the large amount of time elapsed between the sexual encounters themselves and the time of interview, as well as the geographic and social dispersion of students in the years after graduation. In order to analyze denser portions of the whole network (and to enable computational feasibility of certain algorithms), some of our analysis focuses on the induced subgraphs from nodes having graduation years in the non-overlapping ranges 2001-2004 and 2005-2008; these ranges taken together account for approximately 30% (144/477) of the total nodes.

For a general idea of the full network's structure, the network is dominated by one giant connected component, containing 424 of the 472 nodes, which has a diameter of 15 and average shortest path length of 5. Note that due to the prevalence of heterosexual encounters (92% of all edges are cross-sex), this diameter is roughly twice as great as would be expected in a typical social network without such a heterophilic restriction.

3. NETWORK STRUCTURE

3.1 Chains of Affection

Our starting point for our analysis of the network structure was Bearman, et al.'s analysis of the sexual interaction network at "Jefferson High"[1]. Bearman's main conclusion is that the network can be closely modeled as a spanning tree

[1]Since students may graduate in more or less than four years and we are primarily interested in using the measure to identify them by chronological cohort, we define *graduation year* as the end of the fourth academic year after a student enrolled at Stanford.

[2]A small minority of students join multiple subgroups, which we represent by equal fractional membership in each subgroup, except where explicitly mentioned.

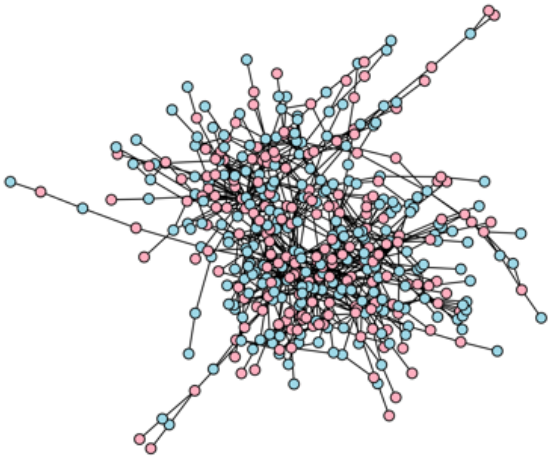

Figure 1: Graph of the gigantic component

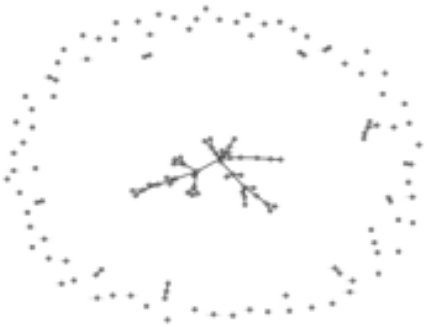

Figure 2: Simulated 18 month subset of dataset

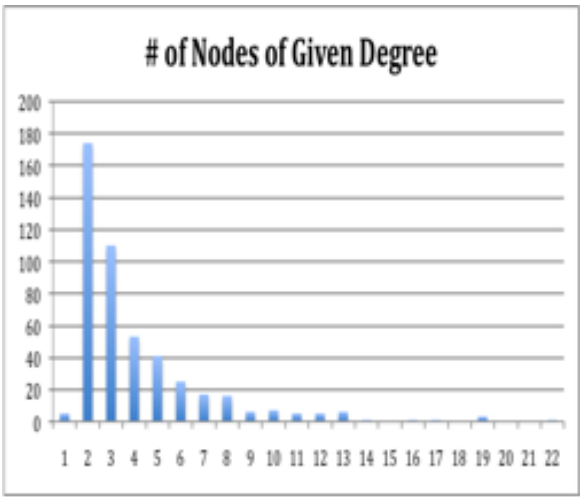

Figure 3: Histogram of node degree distribution

that compromises a gigantic component in the graph, with a number of monogamous dyads and fewer triads making up the majority of the remainder of edges. Our network seems to match the idea that most of the nodes are members of a giant component. However, at least at first glance it appears that our network's gigantic component (Fig 1) is significantly denser than Bearman's Jefferson High network. It is certainly the case that one would have to break significantly more edges to end up with a non-gigantic component than in the Jefferson High graph, where two broken edges are sufficient. However, as can be seen in Figure 3, the vast majority of nodes have relatively small degree. It appears that this follows the high clustering, low diameter small world model. It appears that many of the small world "shortcuts" are found at the high degree nodes, which have noticeably greater average graduation year deltas amongst their edges. However, due to the coarseness of year information (which is particularly confounding among high degree nodes, many of whom have been involved in the student group for longer than four years) we didn't follow this intuition with a more in-depth quantitative study.

One of the key differences between the two graphs is the length of time observed. In Bearman, the survey participants gave information about sexual encounters over the past 18 months. In comparison, our dataset lasts over 20 years. As such, it's reasonable to inquire as to whether the increased density is simply the outcome of additional connections formed over time.

In order to answer this question we simulated an 18 month period of our dataset. We did this by restricting our nodes and edges to a single four graduation year cohort. We then removed (48-18)/48 of the total edges at random to simulate the time. We repeated this process 20 times and examined the resulting graphs. Although the size of the largest connected component varied between 25% and 65% of the total all looked essentially like 2, with a spanning tree or near spanning tree component. This is very similar to the graph of Fig 2 in Bearman.

As such, our graph can fairly be considered as a chronological extension of the Bearman graph. Bearman suggests that the spanning tree may be a result of social prohibitions against having sexual relations with those already close to you in the sexual interaction graph. Therefore, it seems reasonable to suggest that the increasing connectedness is a result of these prohibitions being relaxed over time.

3.2 Homophily amongst self-selected groups

The student group encourages to students to self-select into groups in two ways: by section and by whether or not they take an officer position in the organization as part of staff. The latter affects relatively few students compared to the former, which is a selection process every member of the organization participates in. We were curious about the effect of this self-selection for two reasons. First, this is a reasonable proxy for amount of platonic social interaction. During official events, the different sections spend much of their time together, and non-official events are often organized and predominantly attended by a single section. Second, if we are to make a comparison to the Jefferson High network, to the extent that sections both have an effect on the clustering behavior and are not simply a proxy for phenomenona like level of intra-clique platonic social interaction that occur in

any social setting, we must be wary for factors that might make the dataset not directly comparable. Both of the concerns depend first on determining the significance of sections in determining cluster formation. For if there is no significance, it both illuminates a way in which level of platonic social interaction doesn't affect sexual interaction and alleviates fears of confounding factors in the comparison with Jefferson High. If there is significance, further research is required.

We were surprised at the seeming dearth of literature in the social network analysis space talking about statistical methods for determining significance of preexisting variables. Much of the literature (e.g. [4]) focused on value metrics such as conductance and modularity that work well for comparing different clustering methods but don't lend themselves to a statistical analysis of how probable the structural deviations are. Additionally, there is literature on the use of statistical correlation methods (e.g. [2] and [11], but this seems to rely on an an arbitrary cutoff on a correlation coefficient rather than a rigorous probabilistic analysis. Our eventual method was to simply compute the number of expected in-cluster edges and cross-cluster edges for a random graph with cluster of size k in a graph of size n[3] and then run a Pearson chi-square test for goodness of fit to the distribution. We readily admit that this isn't particularly novel and has some obvious shortcomings (e.g. other confounding structural elements are unaccounted for). Nonetheless, in practice it seemed to work well and we couldn't think of any reason to suspect findings for the null hypothesis.

We ran the chi-square test on each of the sections. Two sections had statistically significant deviations from the null hypothesis: one was a single sex group, which unsurprisingly had more cross-cluster connections than expected (P=1.2e-05). The other was a section that has an informal reputation for being relatively isolated (P=1.4e-06). Two other sections had more internal connections at the 90% confidence level but not at the 95% confidence level and the remaining five sections have no statistically significant deviation. This, combined with the structural explanations, suggest that deviation is only loosely tied to the official section, and it seems likely that what deviation there is a function of natural social group selection.

4. COMMUNITY DETECTION

Since homophily on single attributes alone does not completely account for the presence of sexual choices between groups and individuals, we turn ourselves to the broader question of detecting significant communities in the network, whatever their composition. We find that distinct communities exist, and that they show significant differences between and among them in terms of the variables studied so far, but that the variables themselves are not adequate to describe the communities. That is, the communities that do exist defy easy description in terms of section, graduation year, or staff membership.

[3]In particular, we would expect $\frac{\frac{k\cdot(k-1)}{2}}{\frac{n\cdot(n-1)}{2}}$ of the edges to be contained within the cluster, $\frac{\frac{(n-k)\cdot((n-k)-1)}{2}}{\frac{n\cdot(n-1)}{2}}$ to be outside the cluster, and the remainder to connect interior and exterior nodes

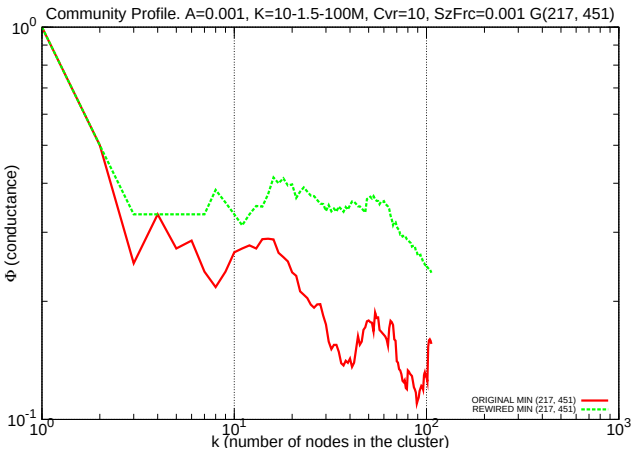

Figure 4: NCP plot of the giant connected component of the whole graph. The original graph shows substantially more community structure than the rewired graph. Note the spikes at community sizes of roughly 8 and 40.

In order to establish the existence of communities, we first present a network community profile plot, which determines the likely sizes of communities that we can find in the data. Next, we describe three separate algorithms that partition the network into two or more partitions, and evaluate their performance on our dataset, both in terms of absolute measures and relative agreement between community detection algorithm.

4.1 Network Community Profile

The *network community profile plot* (NCP plot) is a tool that "measures the quality of network communities at different scale sizes"[10]. In particular, we want to use the NCP plot to determine whether communities exist in a network. The NCP plot plots the minimum conductance $\Phi(k)$ over all communities of size k in the graph, as a function of k. The *conductance* of a set S of nodes is defined as the ratio of edges leaving S to the minimum of the total degree of S or $\bar{S}$:

$$\phi(S) = \frac{\sum_{i\in S, j\notin S} A_{ij}}{\min\left\{A(S), A(\bar{S})\right\}} \quad (1)$$

where A is the adjacency matrix of the graph G, and $A(S)$ is the total degree of a set S. In this formulation, NCP plots should have downward spikes corresponding to good community sizes, and for reasonably small social networks (i.e. those having some social community structure), the overall trend should be downward as a function of k. In particular, a network that does not exhibit any community structure should have a roughly flat graph of $\Phi(k)$ for $2 \leq k \leq \frac{|G|}{2}$. We also consider a null model of a rewired network, preserving degree sequence.

According to the NCP plot, we should expect to find some community structure as is typical in social networks. As we will see, the community detection algorithms we chose to evaluate will indeed find some communities having the optimal sizes. Although the NCP plot reflects community structure for the entire network, we only used it as a start-

ing point for determining whether communities exist. The remainder of the community analysis was performed on the 2001-2004 and 2005-2008 datasets.

4.2 Girvan-Newman

The first algorithm we use for finding communities in the network is the Girvan-Newman algorithm[5]. This is a heuristic algorithm that uses the notion of *edge betweenness*, defined simply as the number of shortest paths running through an edge. Edges are deleted in order of decreasing betweenness, where the betweenness is recalculated after every removal; the goal is to eliminate edges that span communities, leaving separate connected components. Since every edge is eventually removed, the result of the Girvan-Newman algorithm is a dendrogram of hierarchically-arranged partitions. For our analysis, we somewhat arbitrarily stopped removing edges after creating 20 communities. (This facilitated comparison with other community detection methods, and passed the point of optimized modularity; see below.) The algorithm as described in [5] can be run in $O(|E|^2|V|)$ time.

4.3 Louvain

The Louvain method of community detection[3] is based on heuristically optimizing the *modularity* of a partition[12]. The modularity of a partition captures the difference between the number of intra-community edges across all communities compared to a rewired network. More precisely, Fortunato and Bartélemy define modularity as

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{|E|} - \left(\frac{d_s}{2|E|} \right)^2 \right] \tag{2}$$

where there are m communities, l_s is the number of edges inside community s, and d_s is the total degree of community s[4]. Thus partitions with high modularity (Q close to 1) exhibit much more intra-community edges than would be expected in a random network, and can be said to capture strong community behavior. It is worth noting that there is some evidence[4] that modularity optimization is unlikely to find communities smaller than some minimum resolution, but considering the relatively large optimal community size according to our NCP plot, we are not concerned about such limits.

There are two phases to the Louvain method: first, starting with one node per community, neighbors are added to a node's community if the change increases overall modularity. Next, the process is repeated with communities in place of nodes, and so on, until modularity can no longer be increased.

4.4 Latent Position Cluster Model

In the latent position cluster model (LPCM)[7][9], each node is presumed to occupy some location in a multidimensional Euclidean "social space," where the probability of a connection to another node depends on the Euclidean distance between node. Communities are estimated using Bayesian estimation of mixture models, where a parameter $K_i = s$ if node i belongs to community s, which is estimated using Markov chain Monte Carlo sampling.

The advantage of this model is that it does not assume that homophily exists on any particular set of attributes; nodes that are more likely to connect to one another should simply be estimated closer together in the social space. In this sense Conversely, however, it is difficult to succinctly explain the social meaning of each Euclidean dimension. As we shall see, this mirrors the problem with other community detection methods that it is difficult to describe the communities in terms of the node data we have collected.

Graph	Algorithm	# Comms.	Q
'01-'04	G-N	8	0.6968
'01-'04	Louvain	7	0.6975
'01-'04	LPCM	4	0.6248
'05-'08	G-N	11	0.6438
'05-'08	Louvain	7	0.6496
'05-'08	LPCM	4	0.5345

Table 1: Summary of community detection results. The best partition for each algorithm and dataset is shown.

Graph	Algorithms	Agreement
'01-'04	G-N/Louvain	0.7872
'01-'04	G-N/LPCM	0.5745
'01-'04	Louvain/LPCM	0.5957
'05-'08	G-N/Louvain	*
'05-'08	G-N/LPCM	*
'05-'08	Louvain/LPCM	0.5185

Table 2: Agreement results between best-modularity partitions. Comparisons marked with * were not performed due to computational infeasibility.

4.5 Community Analysis

Table 1 summarizes the results of the various community detection algorithms on the 2001-2004 and 2005-2008 subgraphs. Clearly, the relatively high modularities of all the best community partitions indicate that there must be some sort of community structure. However, several questions remain about the quality of the community detection results. First, we would like to know how well the communities identified by our algorithms correspond to the data collected about each node. From our homophily analysis, it should already be somewhat clear that communities do not break down easily along the lines of section or graduation year.

This argument is supported by an examination of the distributions of section membership and graduation year within the communities, as exemplified by Figure 5. Clearly from the histograms we can see that the section and year distributions differ significantly between communities, but the distributions defy easy explanation. The quality of succinct generalization we can make between two communities is along the lines of claiming community 3 is younger than community 4, owing to the oppositely skewed graduation years. It is clear, then, that although these communities undoubtedly exist, any variables sufficient to succinctly describe–if they exist–them are not captured in our current dataset.

The final analysis we perform is a comparison of the different community detection methods, summarized in Table 2 For each pair of community partitions of the same subgraph, we

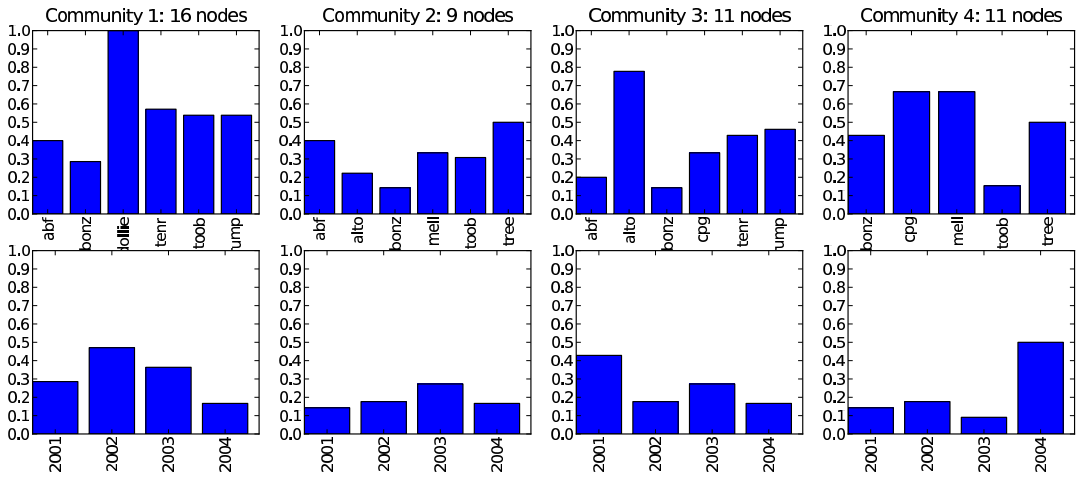

Figure 5: Distribution of year and section membership for the best LPCM community for the 2001-2005 subgraph. Bar heights are fraction of the total population of that category across all histograms.

define the *agreement* between partitions as the fraction of nodes assigned to the same community.[4] The relatively high degree of agreement between the various partitions further supports the hypothesis that the communities we have found are in fact meaningful. Moreover, there is greater agreement between some pairs of non-optimal-modularity partitions, and quick inspection of the non-agreeing nodes shows that they are mostly low-degree outliers.

5. CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

We believe our two major findings are that our dataset does appear to be directly comparable to the Bearman's[1] Jefferson High dataset and that despite the density, meaningful clusters due exist, although we haven't succeeded in explaining the major evolutionary factors behind cluster formation.

There are two major avenues of research that we would recommend pursuing but which we couldn't due to time or resource limitations. The first is to look at the embedding of the social group within the larger Stanford community: students who participate in staff, which we use as a proxy for involvement in the group, have an average degree of 3.8, compared to 2.4 for non-staff members. Since we don't believe staff are noticeably more promiscuous than their non-staff counterparts, this suggests that there are many edges to nodes outside the student group, that could have significant effects on graph measures. Kossinets[8] suggests a number of effects only examining a subgraph could have. However, he only suggests effects on a particular class of graphs and it's not at all clear that our graph exhibits the same characteristics. Additionally, we were unable to find data on the larger graph, although it seems likely that campus health organizations maintain statistics from surveys that would at least answer questions such as average degree.

The second avenue both pertain to correcting the sparsity of attributes in our dataset. For instance, we didn't have even a rough ordering of the sexual interactions. This significantly limited our ability to predict the evolution of the graph, which would answer many more questions about the connection between the Jefferson High graph and ours. An intense survey would provide data to provide analysis comparable to Bearman's chronology-based work later in the paper.

[4]More precisely, we take the maximum fraction over all community labellings. Unfortunately, this has the side effect of making the comparison run in $O(nm!)$ for n nodes and m clusters.

6. REFERENCES

[1] P. Bearman, J. Moody, and K. Stovel. Chains of affection: the structure of adolescent romantic and sexual networks 1. *American Journal of Sociology*, 110(1):44–91, 2004.

[2] P. Blau. *Inequality and heterogeneity: A primitive theory of social structure*. Free Press New York, 1977.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J.Stat.Mech.*, page P10008, 2008.

[4] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, January 2007.

[5] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, June 2002.

[6] Google Scholar, http://scholar.google.com/scholar?cites=2324137115064840048. *Bearman: Chains of affection:... - Google Scholar*, 2009. Accessed 10 Dec, 2009.

[7] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, March 2007.

[8] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.

[9] P. N. Krivitsky and M. S. Handcock. Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software*, 24(5):1–23, 12 2007.

[10] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *CoRR*, abs/0810.1355, 2008.

[11] J. McPherson and L. Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review*, 52(3):370–379, 1987.

[12] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.