Characteristics of Self-Selecting Social Lists

Jessica Parker, Phoebe Whitwell

Abstract

We examined the network created by living group mailing lists and their intersections. The network was compared to bipartite complex networks and found to share some characteristics but not the characteristic of having a degree distribution which follows the power law. Additionally, "power users" and interactions between halls were identified and evaluated to reveal social trends.

Introduction

We live in the dorm East Campus, which has ten halls. These halls often function as individual entities for social activities and each have social mailing lists. These mailing lists function as opt-in discussion forums that generally contain residents and often contain alumni or other friends of the hall. Some people choose to subscribe to multiple of these lists because they socialize with multiple halls. There are stereotypes within East Campus about which halls are more or less social. We thought it would be interesting to see if these stereotypes were reflected in the interconnectedness and more defined network properties of the hall social lists.

In order to analyze this system, we created a bipartite network with all of the users who subscribe to these mailing lists as one set of nodes and the ten mailing lists as the other set. We used networkx to create a weighted projection of each set of nodes to get one graph that is the connected lists and another graph that has all of the users connected through the lists. Our analysis focuses

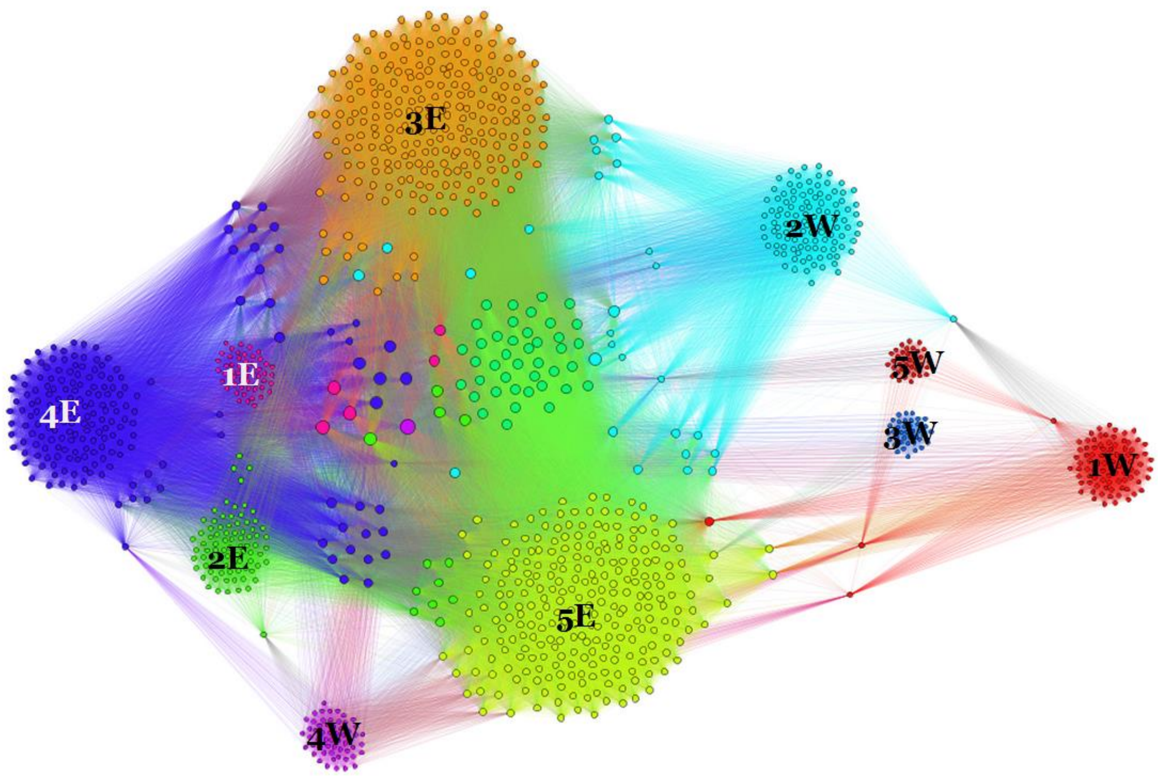


Figure 1: East Campus Hall Email Lists Network

on the graph of users which has 1249 nodes and 158800 edges. Fewer than 400 people live in East Campus and only a few subscribe to lists under multiple aliases.

Results

The resulting network is shown in Figure 1. This particular visualization was made in Gephi using the ForceAtlas2 layout. It sorted the network into ten distinct clusters and various sparser middle-ground communities. The clusters can each be identified as containing most of the members of one particular hall.

By running Gephi's built-in Modularity function with a focus of 0.675, we were able to find that the network has a Modularity value of 0.47 and ten modularity classes which are shown by color in Figure 1. Two of the smaller clusters, labeled 5W and 1W, were grouped in the same modularity class while the extra community consisting of users on the lists for 3E and 5E (and no other lists) was given its own modularity class. Users on multiple lists were pulled toward the largest cluster due to the gravity used by the layout algorithm. For example, the small group of teal nodes between 3E and 2W are residents of 2W, as the modularity algorithm calculated, but are pulled closer to the 3E cluster. The section of the 5E cluster closest to the 4W cluster is actually mostly 4W residents but the 5E cluster is so large that it pulled them all the way in. This is perhaps more an interesting commentary on the layout and modularity algorithms than on the network.

The macro layout of the clusters has some interesting ties to the community stereotypes mentioned in the introduction. We can see a sort of triangle formed by the three largest clusters (3E, 4E, & 5E) which encompasses a large majority of the edges in the network. These halls are known for having the largest mailing list followings and are some of the more outwardly social halls within the dorm. The 2W cluster is similarly connected but smaller and slightly offset which matches with its personality of being social but more exclusive in membership. The 1E and 2E clusters are smaller communities that are sort of encompassed by the larger triad and the 4W cluster is outside the triad but strongly connected to 5E, all of which matches the real-world stereotypes for these halls. The three clusters to the far right of the graph are known for being smaller social communities with fewer ties to the rest of the dorm social groups.

Hall	**Users on List**	**Users Unique to List**	**Exclusivity (unique/total users)**
5E	346	241	0.697
3E	326	233	0.714
4E	201	148	0.736
2W	142	111	0.782
2E	122	78	0.639
1W	93	85	0.914
4W	79	57	0.722
1E	77	52	0.675
5W	42	38	0.905
3W	41	39	0.951

Table 1: List Membership

By examining the degrees and clustering coefficients of the nodes, we can quantize these behaviors as any node with a clustering coefficient of one will be on only one list. Table 1 demonstrates the differences in size and exclusivity between the communities. It suggests that the main influence on network behavior is exclusivity, as networks of similar size can vary widely in behavior. For instance, 2E, 1W, and 4W are all within the middle range of networks in terms of size, but are towards the center of the network, far away from the main group, and at the edge of main group respectively which corresponds to their exclusivity percentages of 64%, 91%, and 72%. Figure 2 shows the relationship between clustering coefficient and degree. The large clusters of users at a clustering

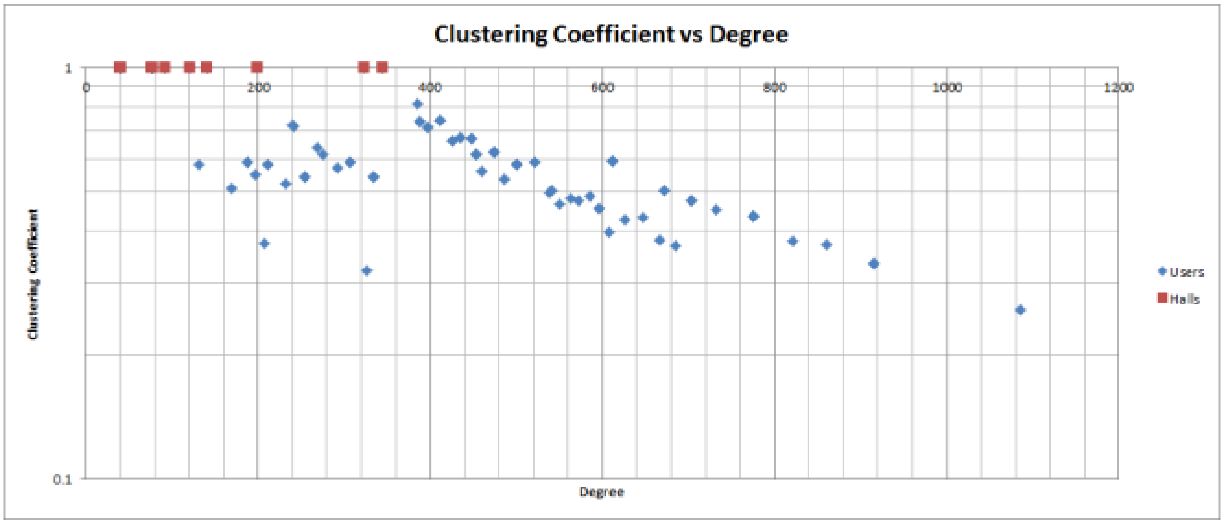


Figure 2: Clustering Coefficient vs. Degree for Halls and Individuals

coefficient of one correspond to the halls and represent the users unique to each hall list. Individual users who are on multiple hall lists trail off down and to the right, corresponding to lower clustering coefficients and higher degrees.

Average Degree	254.3
Average Weighted Degree	259.1
Density	0.204
Modularity	0.47
Average Clustering Coefficient	0.943
Average Weighted Cluster Coefficient	0.946
Average Path Length	1.892

Table 2: Network Properties

Our network has the properties found in Table 2. It has a high clustering coefficient and a small average path length. These properties correspond to those frequently found in bipartite networks, particularly those corresponding to social networks [1]. Figure 3 shows the characteristics of some of these networks. Our network appears to have properties the most similar to the "Co-occur" network, but we believe this to be a coincidence as this network shows groups of words in the Bible.

	Internet	Web	Actors	Co-auth	Co-occur	Protein
n	75885	325729	392340	16401	9297	2113
m	357317	1090108	15038083	29552	392066	2203
density	1.2e-4	2.1e-5	1.9e-4	2.2e-4	9.1e-3	9.9e-4
α	2.5	2.3	2.2	2.4	1.8	2.4
c	0.171	0.466	0.785	0.638	0.822	0.153
d	5.80	7	3.6	7.18	2.13	6.74

Table 1: The main statistics for the complex networks we use in this paper. For each network, we give its number of vertices n, its number of links m, its density, the value of the exponent α of the power law that fits best its degree distribution, its clustering c, and its average distance d.

Figure 3: Example Properties of Complex Networks [1]

Another characteristic frequently found in bipartite network projections is a degree distribution which follows the power law [1]. We graphed the degree distribution of our network and binned the data for clarity. We then tried various fits to our data (see Figure 4), including a linear fit, log fit, and analytical fits using the typical gamma values found in Figure 3. Though the fit lines themselves appeared to approximate the power law, when we checked them against our original data they did not match and therefore cannot be taken as correct fits.

Similarly, many bipartite social networks exhibit small world patterns when projected [2]. We graphed the degree distribution and binned data against a small world (Barabasi-Albert, so as to incorporate preferential attachment) distribution, shown in Figure 5. The small world distribution matches well with neither the original nor binned data. Given this evidence, we conclude that our network does not follow the power law as is typical for social bipartite networks. This may be because it is opt-in or because it does not have enough communities.

Despite the lack of a coherent degree distribution, we were still able to find useful information in our network. For instance, we used the hall mailing lists, but the dorm itself also has a mailing list, "ec-discuss@mit.edu". Notable to this list is the presence of "power users" [3]. These users, despite representing only 4% of the membership of the list, are responsible for 40% of emails sent to ec-discuss. The top twenty of these users and their interactions are depicted in Figure 6. We were curious whether the ec-discuss power users corresponded to high degree nodes in our network. By inspecting the center of our network, we were able to identify ten of the top sixteen unique dorm-resident power users (listed in Table 3). These nodes are marked in Figure 7.

Table 3: Active Users

Power Users
lroyden
pranjal
robj
mnawrot
cjtenny
wliverno
mcole
bsena
cirrus
yonadav
eddiecc
flipdog
leonidg
mspatz
oliviame
ycoroneo

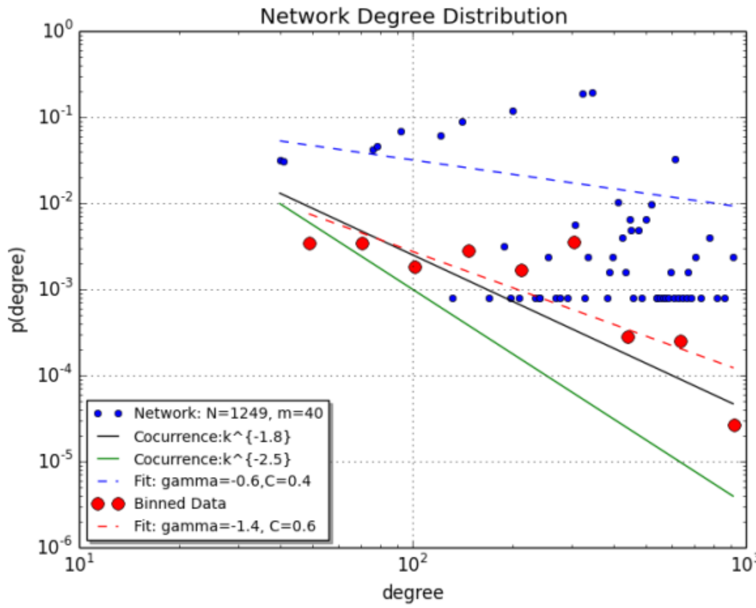


Figure 4: Degree Distribution and Fits

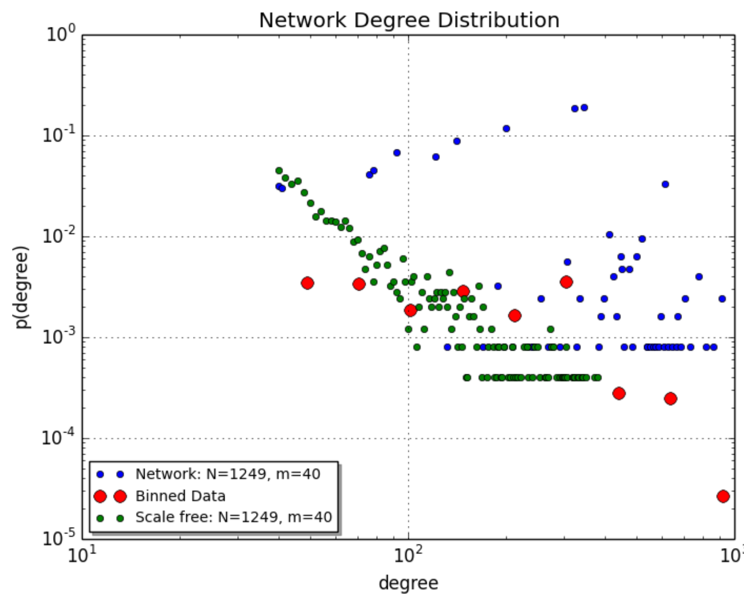


Figure 5: Degree Distribution and Small World Distribution

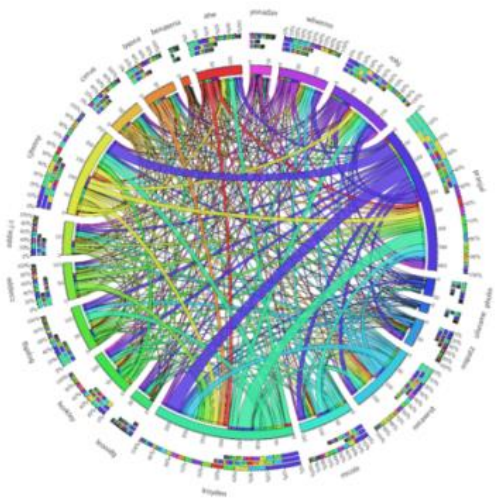


Figure 6: Power User Interactions [3]

We can also inspect the other projection of this network by using the lists as nodes (Figure 8). Social properties are conserved between projections: the halls at the center of the network and with high degrees remain so, while smaller halls with few users on multiple lists move to the edge of the graph. We can further isolate communities by removing edges (users), starting with the edges with the highest weights. After deleting the connections between the three nodes previously identified as the most connected (3E, 4E, 5E), we see in Figure 9 that a secondary group of connected nodes emerges (2W, 2E, 1E, 4W). The first node to become completely disconnected is 3E, only after removing over half of the initial set of edges (see Figure 10).

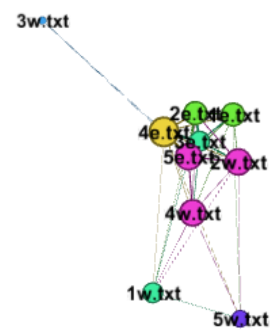


Figure 8: Initial Hall Network

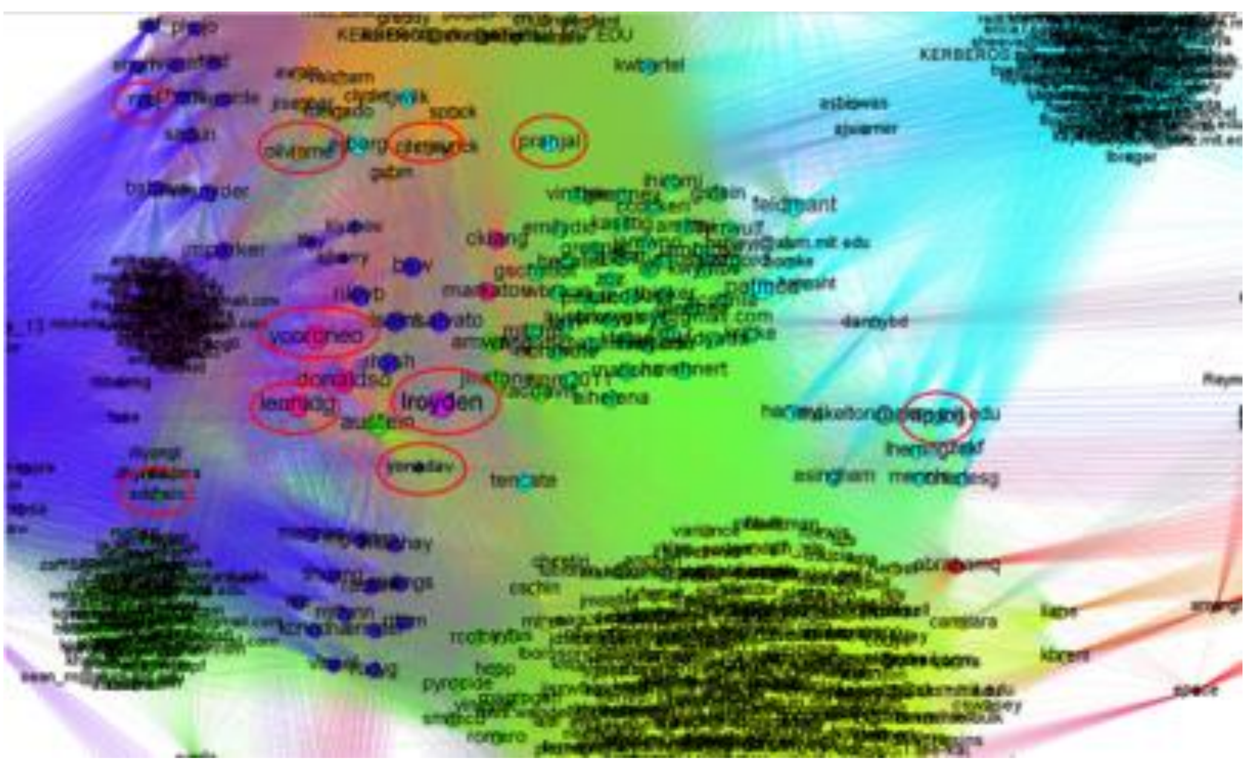
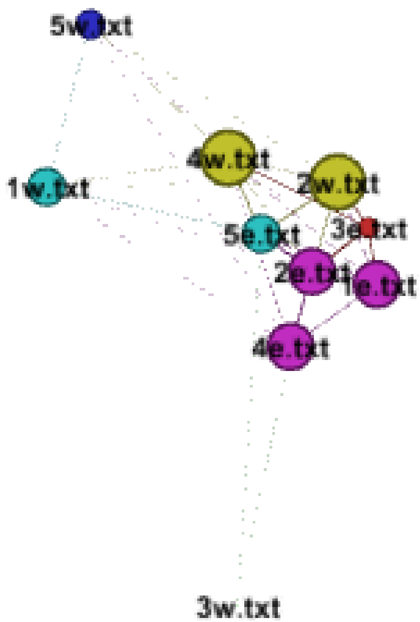


Figure 7: Power Users Identified at Center of Network



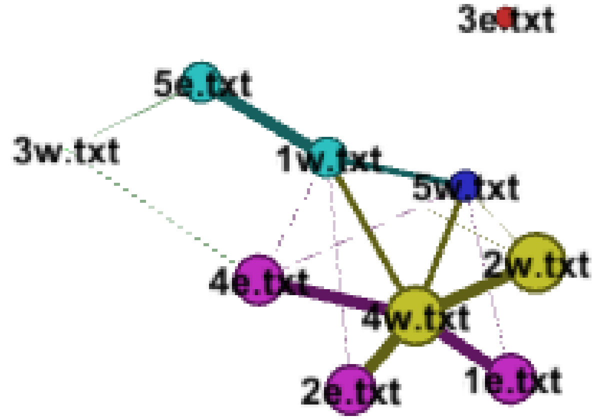


Figure 10: Hall Network after Removing Further Links

< Figure 9: Hall Network after Removing Largest Weighted Links

Conclusions and Future Work

Though this network proved difficult to relate to seemingly similar networks, it did prove useful for demonstrating broad social trends within communities. The self-selecting nature of this network proved to be very important, as it was responsible for the high centrality of some very involved users and also showed which lists in general "select" for connections with other lists and varying degrees of exclusivity. In future studies, we would hope to correct for the trend of joining lists despite not being connected socially to that hall, which we have observed in our daily interactions with dorm residents. We would do so by using nonsymmetrical weights [5]. This technique would emphasize connections that a user feels are important to their social landscape and de-emphasize links that are not socially relevant to that user. This would generate a more accurate picture of community connections.

Works Cited

[1] Guillaume, Jean-Loup. "Bipartite Graphs as Models of Complex Networks." (n.d.): n. pag. Web. 18 May 2014. <http://graal.ens-lyon.fr/~abenoit/reso06/papier/biparti.pdf>.

[2] Davis, Gerald F., Mina Yoo, and Wayne E. Baker. "The Small World of the American Corporate Elite, 1982-2001." SAGE Journals. Strategic Organization, 2003. Web. 18 May 2014. <http://soq.sagepub.com/content/1/3/301.full.pdf%2Bhtml>.

[3] Ben-David, Danny, Matthew Cole, and Stephen Suen. The Weak Shall Be Eaten: Uses and Gratifications of MIT Mailing Lists. Rep. N.p.: CMS.400, n.d. Print.

[4] Gleiser, Pablo M., and Leon Danon. "Community Structure in Jazz." (n.d.): n. pag. 25 July 2003. Web. 18 May 2014. <http://arxiv.org/pdf/cond-mat/0307434v2.pdf>.