An Analysis of Friendship Groups and Social Connectedness in Flixster Community Network

By:

Gnana Teja Peddi

ABSTRACT

With the tremendous amount of data that is created every day (about more than 2.5 quintillion bytes), collecting, storing, parsing, querying, modeling, and making decisions from data has become more challenging. The user base has grown to billions with the rapid growth of social networks in the past decade and a half. Social networks like Facebook boast about 2.38 billion active users per month. We have various social networks available for various purposes, to connect people with varied interests and personalities. Also, in an organizational context, the various participants of a network and their subsequent interactions assume a lot of significance. Such a study is possible through well-developed social network theory, handed down through several years of thorough research. The primary difference between traditional research and social network analysis is that, emphasis is on the relationships of actors in network analysis while emphasis is on actors and qualities in traditional research. In network analysis, individuals and their qualifications are not taken into consideration.

With such a powerful tool to model and analyze the data, in this project I try to derive insights from a dataset relating to a popular social-networking movie website, Flixster. It is a social movie site allowing users to share movie ratings, discover new movies and meet others with similar movie taste. If we know when users rate movies, we can find an optimal network of users that best explain the propagation of information which influences their movie watching behavior. Unveiling such a network would be useful to identify how product information is spread among users. Subsequently, we analyze several network metrics and network cohesion measures of the Flixster community. We perform various community detection algorithms to identify significant communities present in the network. Furthermore, we try to draw a parallel between what we obtain from the social network analysis as well as perceived notion.

1. INTRODUCTION

Flixster is a popular American social-networking movie website for discovering popular and new movies, learning about movies, and meeting others with similar tastes in movies [1]. The site enables users to watch movie trailers as well as learn about upcoming movies at the box office. It was started in San Francisco, California and was founded by Joe Greenstein and Saran Chari in 2007. Flixster was the parent of website Rotten Tomatoes from January 2010. In a latest development, on February 17, 2016, Flixster, including Rotten Tomatoes, got acquired by Fandango.

Flixster's growth was described in the trade press as attributable to "its aggressive viral marketing practices," including "the automated selection of your email account's entire address book in order to send a Flixster invitation to all of your contacts." Although the company claimed this procedure to be an industry standard used by other services, Flixster differed in that its system automatically selected all contacts in the user's address book and required the user to manually un-select each address to prevent email from being sent to a user. Co-founder Joe Greenstein described the advantage of Flixster as being able to spread and share among friends, as well as engage with them. As a consequence of its policy of emailing users' entire address books with advertisements for the site, the website was criticized on numerous Internet blogs. At one time, email from Flixster to Hotmail users was being filtered and deleted as spam.

Social network analysis has been a primary research area recently due to the availability of wide range of large datasets [2]. One of the main problems in social network analysis is to recover the underlying network structure given information about how nodes in the network interact with each other in the past.

In the context of a social movie website, if one knows when users rate movies, then one can easily find an optimal network of users that best explain the propagation of information which

influences their movie watching behavior. Unveiling such a network would be useful to identify how product information is spread among users. Moreover, it is useful for the recommendation of new products to users since recent studies have pointed out that using social network information could potentially improve the accuracy of recommendation. But many movie websites such as Netflix do not have a built-in social network to take this advantage.

Many researchers including Homans (1958), Blau (1964) and Emerson (1976) have contributed to the development of social network theory (Blau, 1964; Emerson, 1976; Homans, 1958). Relations with network theory and structural analysis have been associated with each other and have begun to be studied in areas such as sociology, social psychology, and anthropology. In order to visualize interpersonal connections with sociometry, people are expressed as dots and the connections are also represented as the line between the dots. This representation is often called sociogram [3].

When one looks at social network analysis from the point of view of management science, it is a tool that allows to study strategically important networks within an organization, to recognize informal groups and to work with important groups to facilitate effective collaboration. Perhaps one of the most important benefits of social network analysis is that managers focus their attention on informal networks that can be critical to organizational effectiveness [4].

2. Data Overview

The social network of Flixster is a movie rating site on which people can meet others with a similar movie taste. In this friendship network two users are linked if they are friends. The data for this analysis was obtained from the network data repository of Arizona State University [5].

The Flixster network dataset consists of two files. The first file contains information about the nodes in the network. It is the dictionary of all users who use Flixster and there are about 2,523,386 nodes in this file. The second file contains information about edge relations between the nodes. This file contains 9,197,338 edges and the edges indicate friendships among users.

3. General Information and Overall Structure of the Network

The graph we have is an unweighted and undirected graph. Since friendships are usually bi-directional it makes to not have any directed edges from one to another in the network. Some of the edges in the data have been duplicated multiple times. hence, I simplified the graph by removing the redundant edges. Now we have 2,523,386 nodes and 7,918,801 edges. In order to analyze the network, several network metrics and network cohesion measures are used which are discussed in sub sections below.

3.1 Network Statistics

By analyzing the graph, I have found that the user node with ID "46246" has the most number of connections in the network with 1474 edges connected to it. Network density represents the proportion of connected ties over the total possible connections in the network. This gives an idea of how connected the network is. The network density is 2.4873×10^{-6} edges/$(\text{vertex})^2$ with an average degree of 6.2763 edges/node. The degree refers to the number of ties a node (user) has in the network and could help to reveal the most powerful individuals in a network.

On average a user has ties to around 6 other users. The average path length of the Flixster network is 4.82 edges. It refers to the average number of steps in the shortest path to navigate the network. It gives us a sense of how efficient the flow of information is through the Flixster network. We see that generally it takes about 5 ties on average to get all the way around in the network. Further, about 1.37% of the connected triples in the network form a triadic closure.

Since the network is huge and it is computationally expensive to analyze the entire network, I have taken a random subgraph of 5000 nodes. These 5000 nodes have 24,507 edges connected overall in the network with an average degree of 9.8028 edges/node. The network density of the subgraph has now become 1.9609×10^{-3} edges/$(\text{vertex})^2$ which is much higher than that of the complete network. This means that the nodes are more densely connected to one another in the subgraph. The average path length of the subgraph is 3.49 edges. We can see that average path length is correlated with density to some extent. A larger average path distance often implies that we have a less dense network. Also, the diameter of the subgraph is 4 edges, which represents the length of the longest geodesic.

A network with high clustering has a higher proportion of closed triads to all triads. In other words, when there is mutuality there will be high transitivity. Subsequently, the subgraph has a global transitivity of 7.1% i.e. about 7.1% of the connected triples form a triadic closure which is interesting to observe.

3.2 Network Visualization

I have visualized the graph in Gephi using the Fruchterman and Reingold algorithm. Fig 1 below shows the Flixster network subgraph using the Fruchterman & Reingold Algorithm. Nodes with the highest betweenness centrality are those which are the biggest in size as evident in Fig 1 below. These nodes are the users in the Flixster network who act as bridges to forming new connections with the other users in the network. Smaller nodes in Fig. 1 have smaller betweenness.

As we refer to the color coding in Fig. 1 we see that nodes that are red in color are those with the least number of connections and the nodes with the lightest pink shade are those with the highest degree, hence the highest number of connections in the network.

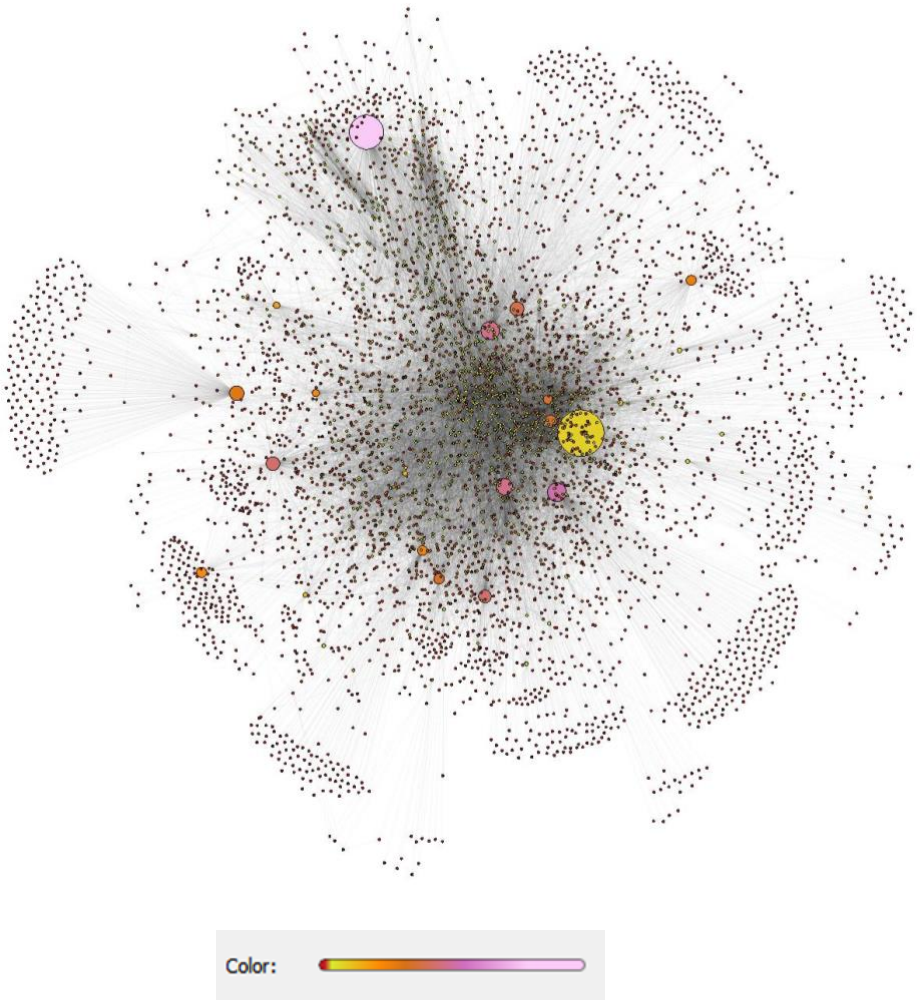


Fig. 1. Flixster Subgraph Network Layout Using the Method of Fruchterman and Reingold

3.3 Network Analysis: Degree Distribution

The users' relative connectedness in this network can be described by the degree distribution. The vertex strength distributions i.e. weighted degree distribution is shown in Fig 2. The plot shows that there are a lot of nodes of lower degree and shows skewness.

Given the nature of the decay in the distributions of Fig 2, a log-log scale could be more effective in summarizing the degree information as shown in Fig 3. We see that there is a nearly a linear decay in the log-frequency (cumulative probability) as a function of log-degree.

The log-log distribution shows a close to linear distribution and a **scale-free** network with a **power-law** degree distribution. We can say that there is a preferential attachment in play here with the "Rich Get Richer" phenomenon [6]. The nodes with a lower degree tend to connect with those of lower degree and the nodes with a higher degree tend to connect with those of higher degree.

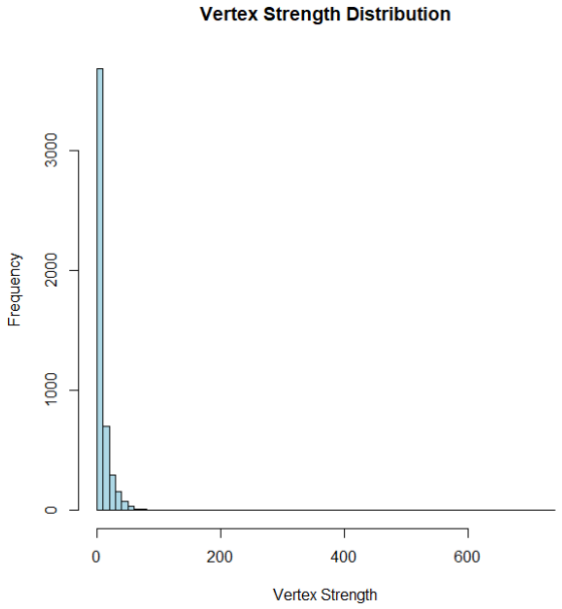


Fig. 2. Distribution of vertex strength in Flixster network

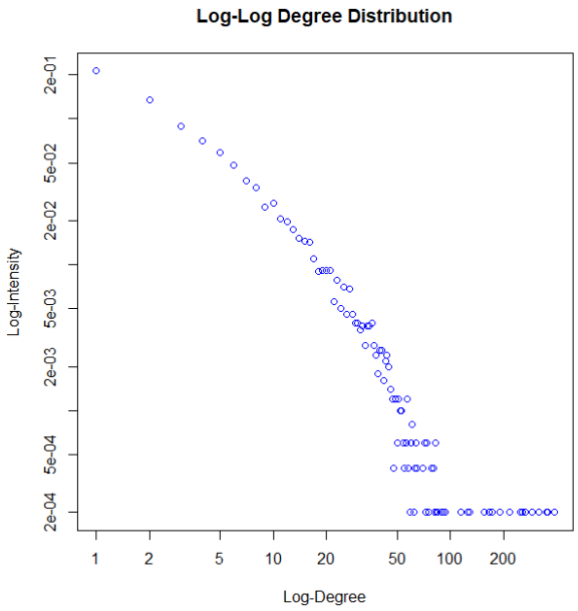


Fig. 3. Distribution of degree in log-log scale

Next, to analyze how nodes of different degrees are linked with each other, a plot of average neighbor degree versus vertex degree in the network is shown in Fig. 4. The plot suggests that there is a tendency for nodes of the lower degree to link with nodes of both lower and higher degrees. But as we go towards the nodes with higher degrees they tend to connect with only a specific range of degree of nodes. Here, we can again say that some amount of preferential attachment is in play here as users with a higher number of connections choose to connect only with other users who have a certain range of connections.

The property assortativity of networks can also be analyzed from Fig. 4 by plotting all nodes of a network by their degree and the average degree of their neighbors. This is a measure of how preferentially attached vertices are to other vertices with identical attributes. Positive assortativity indicates a correlation between nodes of similar degree, while negative indicates a correlation between nodes of different degree. For the subgraph of Flixster network the assortativity is determined to be -0.237.

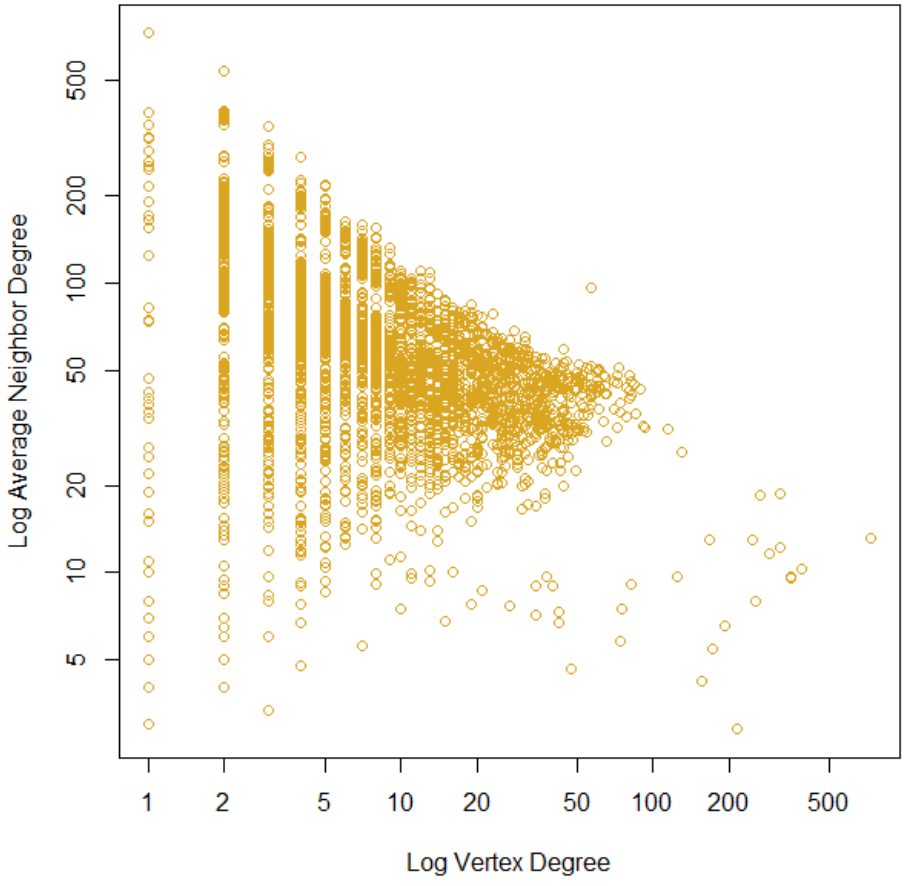


Fig. 4. Average neighbor degree versus vertex degree (log–log scale)

3.4 Network Analysis: Positional Features and Network Centralities

Various network centralities are evaluated, to be used as local measures, for the Flixster network subgraph. Some of these measures include degree, node betweenness, clustering, embeddedness, etc. These centralities are compared through correlation plots as shown below in Fig. 5. and statistically through correlation matrix in Table I.

The correlation plots are shown in Fig. 5. In the top-left plot, we see a significant negative correlation between the average betweenness of nodes and local clustering coefficients. Thus, when individuals' clustering in the Flixster network is very low, the betweenness of those individuals is quite high, and high betweenness leads to the formation of structural holes and local bridges.

Local bridges play a very important part in a community like the Flixster network. Local bridges or structural holes lead to connections among users who have no other common connection [7]. This, in fact, leads to better flow of knowledge and there is less redundancy compared to that of a strong network. Through a local bridge, critical information can be shared which may not be quite possible in a strong network. Local bridges become an important part of knowledge sharing, and information can be useful for the whole community as a whole. Because of high betweenness, there is low overlapping of information. The local bridges help the Flixster community to gain important information. Strong cohesive ties sometimes lead to data holding and redundancy and so structural holes play a very important role [8].

In the bottom-left plot, there is a negative correlation between individuals' average clustering coefficient and degree which implies that for the given network the global clustering is lower than the average clustering. In the top-right plot, we see that the average betweenness of nodes increases as their degree increases.

That means nodes with a higher degree are those that have the most connections and also the most central nodes that play an important role in connecting other nodes in the network by acting as local bridges between different parts of the network. Further, the embeddedness decreases with increase in degree in the bottom-right plot.

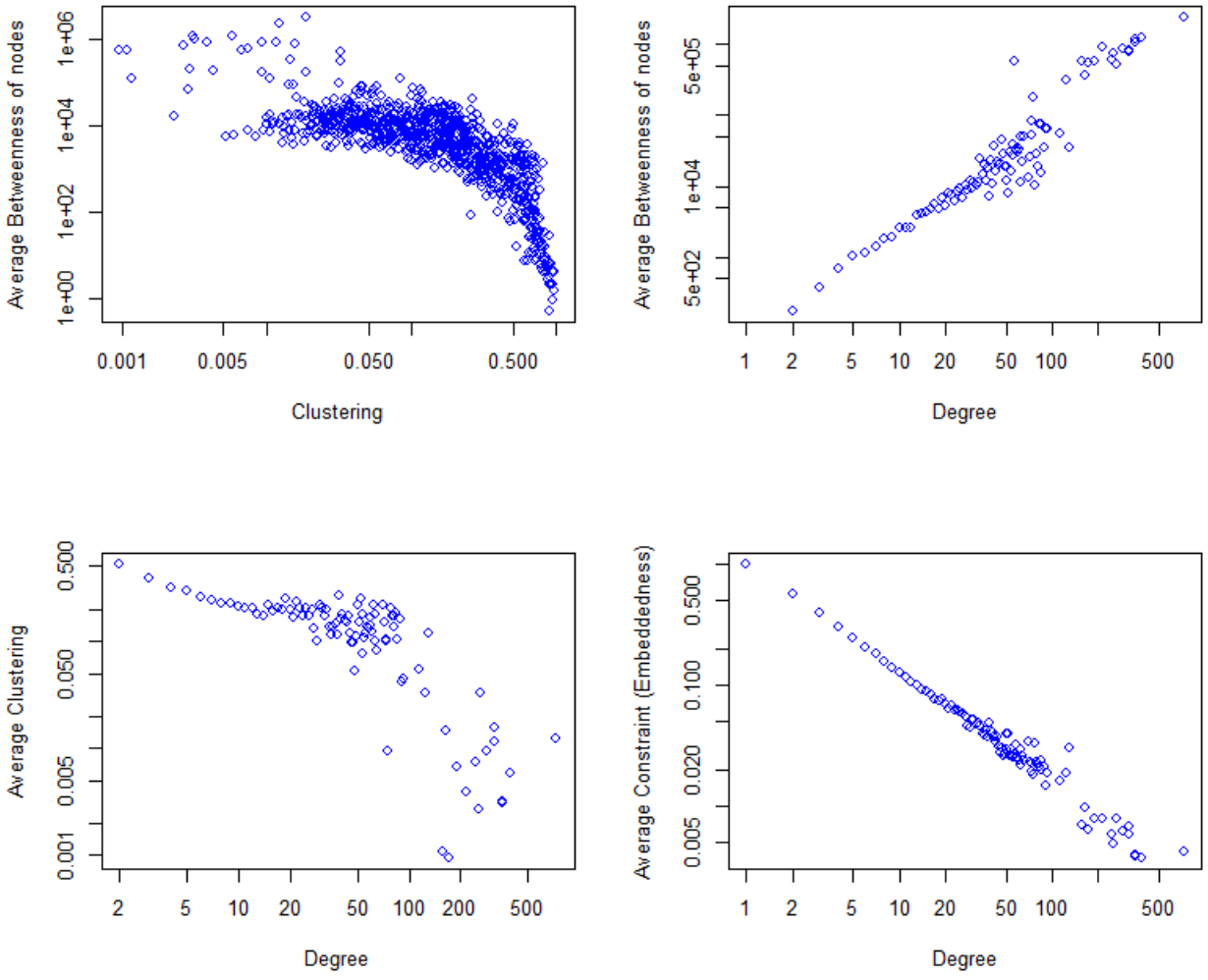


Fig. 5. Correlation plots of various network centralities

The correlation values between the network measures are shown in Table I below. Table I shows a correlation matrix with correlation values for each pair of network centralities. We can see, for example, that degree centrality is positively correlated with node betweenness centrality at a 0.58 level, while it has almost negligible correlation with edge betweenness.

Table I. Correlation matrix for various network centralities

	Degree	**Node Betweenness**	**Edge Betweenness**	**Closeness**	**Eigen**
Degree	1	0.582	-0.0024	0.2693	0.7569
Node Betweenness	0.582	1	0.0089	0.3929	0.4384
Edge Betweenness	-0.0024	0.0089	1	0.0184	-0.0057
Closeness	0.2693	0.3929	0.0184	1	0.1732
Eigen	0.7569	0.4384	-0.0057	0.1732	1

3.5 Network Analysis: Clique Census

In addition to clusters and communities that we saw before, one another approach to define network cohesion is through clique census. Cliques are complete subgraphs and hence are subsets of vertices that are fully cohesive, in the sense that all vertices within the subset are connected by edges. It is important to have weak ties between the cliques. While members of one or two cliques may be efficiently engaged, the problem is that, without weak ties, any momentum generated in this way does not spread beyond the clique. As a result, most of the population will be untouched.

Table II summarizes maximal clique census which represents cliques that are not a subset of a larger clique. We see that in the Flixster network there are 8275 edges (cliques of size 2), followed by 4478 triangles (cliques of size three), followed by 1925 cliques of size 4, and so on. The largest clique is of size 16, of which there are 12 cliques in the network. Here, we do see larger cliques because the Flixster sub network that I have considered is dense to some extent.

Table II. Maximal Clique Census

Clique Size	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Total Cliques	8275	4478	1925	1405	889	464	250	125
Clique Size	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
Total Cliques	103	130	53	77	34	30	12	

4. Detecting Communities in the Flixster Network

Typically, in a large real-life graph, such as the friendship network of Flixster, nodes can be easily clustered together into sets which are densely connected internally. Such sets can be viewed as communities within the network. It can be regarded as a collection of vertices which are densely connected, when compared to the rest of a network [10]. The underlying structure of a complex network, such as Flixster, can be understood by detecting organizational groups of the vertices (friends in the network).

Further, communities often have very different properties than the average properties of the networks, thus it becomes very important to detect these communities. Concentrating only on the average properties usually misses many important and interesting features inside the networks [11]. Presence of communities also usually affects many processes like rumor spreading or epidemic spreading happening within a network [12]. Hence to properly understand such processes, it is imperative to detect communities and also to study how they affect the spreading processes in various settings. As a result, there are several procedures of detecting communities proposed in the last decade. Some of the widely used algorithms that we apply to the Flixster network in this study are Fast Greedy, Walktrap, Spinglass and Label Propagation. They are discussed in more details in following sub sections.

4.1 Community Structure Using Fast Greedy Algorithm

Fast Greedy algorithm for community detection is a hierarchical bottom-up process [13]. It tries to optimize a quality function called modularity in a greedy manner. Initially, every vertex belongs to a separate community, and communities are merged iteratively such that each merge is locally optimal (i.e. yields the largest increase in the current value of modularity). The algorithm stops when it is not possible to increase the modularity any longer, so it gives a grouping as well as a dendrogram. This method is usually tried as a first approximation because it has no parameters to tune.

When this algorithm is applied to the Flixster network, we get a community structure as shown in Fig. 6. This algorithm produced a total of 18 communities in the network and the sizes of each of these communities are listed in Table III. We see that the largest community in the Flixster network consists of about 1131 members. This algorithm was computationally fast in assigning members to different communities.

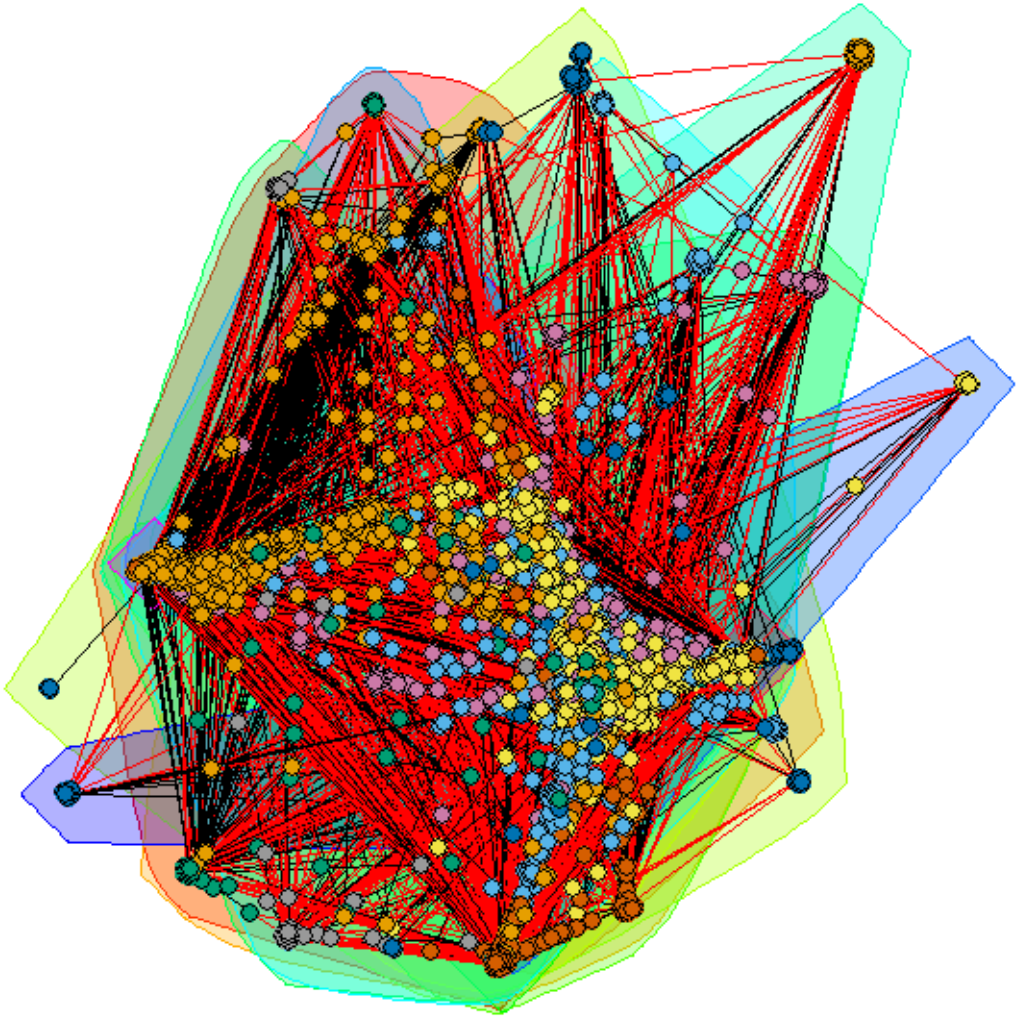


Fig. 6. Network Graph Plot Showing Community Structure Using Fast Greedy Algorithm: 18 Communities

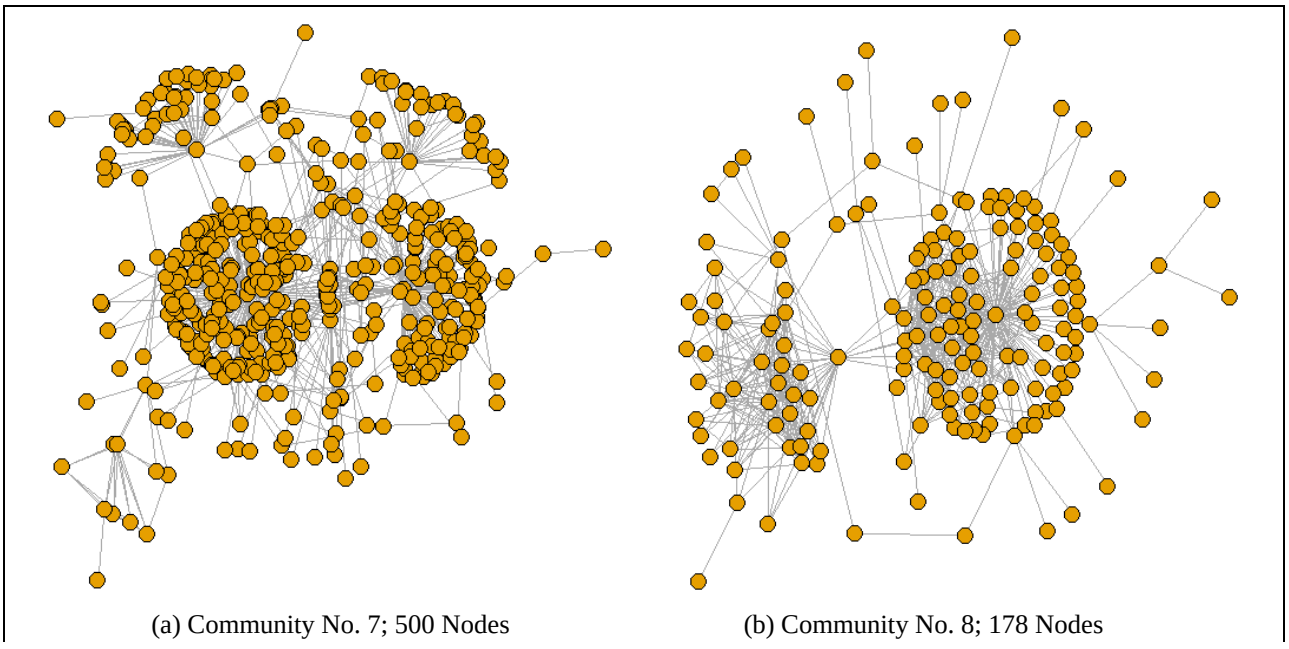
Table III. Community Significance for Communities Detected by Fast Greedy Algorithm

Community Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Community Size	1131	315	162	1252	183	632	500	178	205
Community Number	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Community Size	260	58	37	43	9	8	11	6	10

I also performed a community significance test for the communities detected by the Fast Greedy algorithm. I applied Wilcoxon rank-sum test on the internal and external degrees of a community to quantify its significance. Internal edges are the edges present within a community, while external edges are the edges connecting vertices of a community with the rest of the graph.

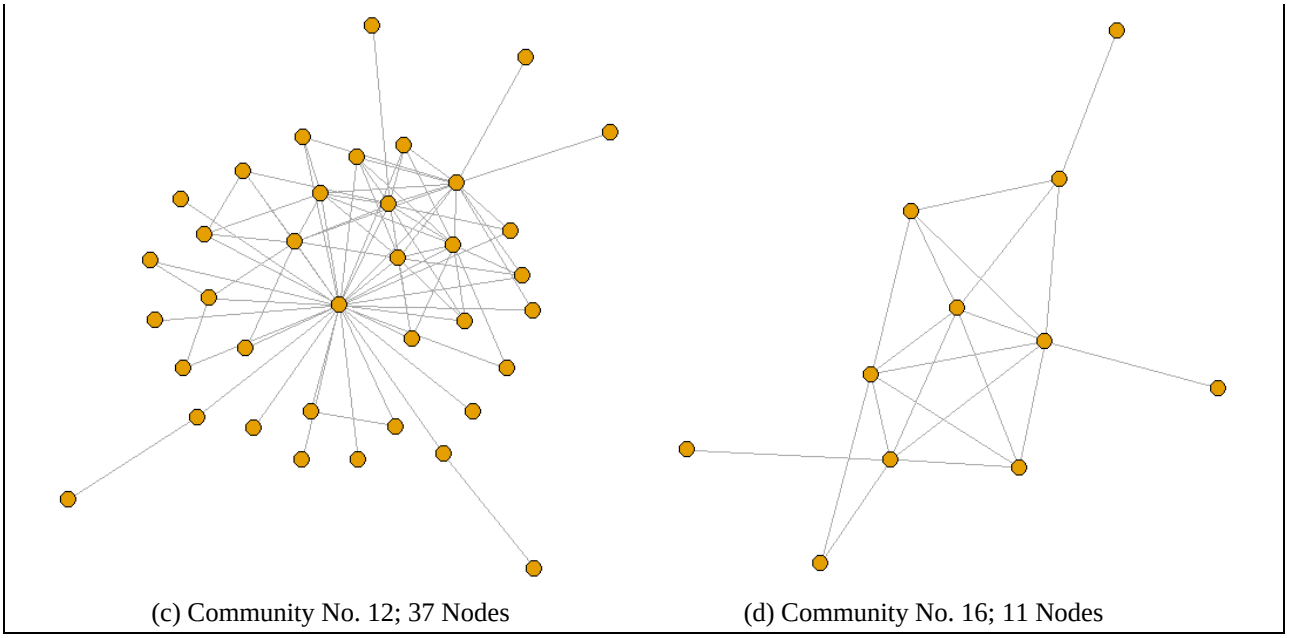
The null hypothesis of this test is that there is no difference between the number of internal and external edges incident to a vertex of the community [14]. More internal than external edges show that the community is significant; less internal than external edges show that the community is in fact an "anti-community". The *p*-value of the test performed by this function will be close to zero in both cases; the value of the test statistic tells us whether we have a community or an anti-community.

Subsequently, performing Wilcoxon rank-sum test to the communities detected by the Fast Greedy algorithm, I found that there were a total of 16 significant communities in the network. Communities numbered 14 and 17 in Table III above resulted in a *p*-value of the test to be greater than 0.05 and thus were insignificant. Network graph of some of the significant communities is shown in Fig. 7.



(a) Community No. 7; 500 Nodes

(b) Community No. 8; 178 Nodes



(c) Community No. 12; 37 Nodes (d) Community No. 16; 11 Nodes

Fig. 7. Some Fast Greedy Significant Communities

4.2 Community Structure Using Walktrap Algorithm

Next, we apply Walktrap algorithm to detect communities in the Flixster network. It is an approach based on random walks [13]. The general idea is that if random walks are performed on the graph, then the walks are more likely to stay within the same community because there are only a few edges that lead outside a given community. Walktrap runs short random walks of 3-4-5 steps and uses the results of these random walks to merge separate communities in a bottom-up manner like Fast Greedy algorithm. Again, the modularity score can be used to select where the dendrogram is cut.

When this algorithm is applied to the Flixster network, we get a community structure as shown in Fig. 8. This algorithm produced a total of 141 communities in the network, which is a lot higher than that obtained by the Fast Greedy algorithm. The sizes of first 20 communities are listed in Table IV. We see that the largest community consists of about 1563 members. This algorithm was computationally very slow in assigning members to different communities compared to the Fast Greedy approach.

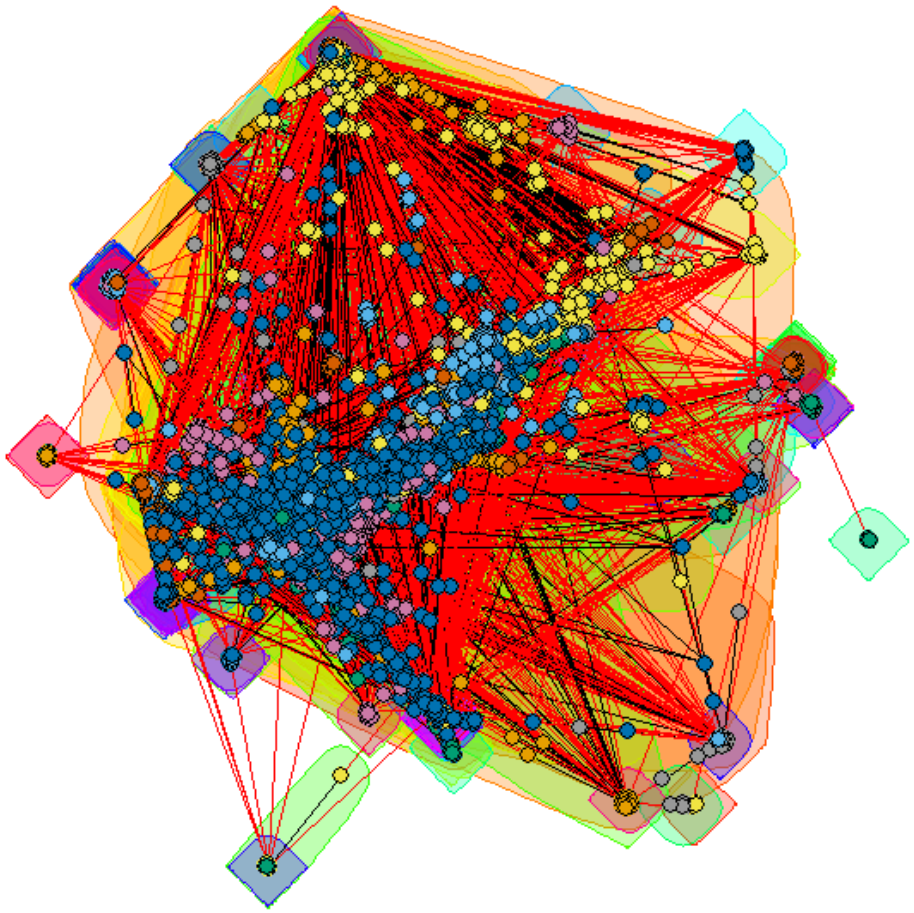


Fig. 8. Network Graph Plot Showing Community Structure Using Walktrap Algorithm: 141 Communities

Table IV. Communities Detected by Walktrap Algorithm

Community Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Community Size	26	7	18	6	216	137	19	108	15	116
Community Number	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Community Size	7	377	1563	5	243	98	33	185	55	171

4.3 Community Structure Using Spinglass Algorithm

Further, we perform Spinglass algorithm to detect communities in the Flixster network. It is an approach from statistical physics, based on the so-called Potts model [13]. In this model, each particle (i.e. vertex) can be in one of *c* spin states, and the interactions between the particles (i.e. the edges of the graph) specify which pairs of vertices would prefer to stay in the same spin state

and which ones prefer to have different spin states. The model is then simulated for a given number of steps, and the spin states of the particles in the end define the communities.

When this algorithm is applied to the Flixster network, it defines a community structure as shown in Fig. 9. This algorithm produced a total of 19 communities in the network and the sizes of each of these communities are listed in Table V. We see that the largest community consists of about 806 members. This algorithm was computationally similar to the Fast Greedy algorithm in assigning members to different communities.

Table V. Communities Detected by Spinglass Algorithm

Community Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Community Size	205	238	367	418	253	23	296	663	46	549
Community Number	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	
Community Size	52	389	806	183	112	45	41	263	51	

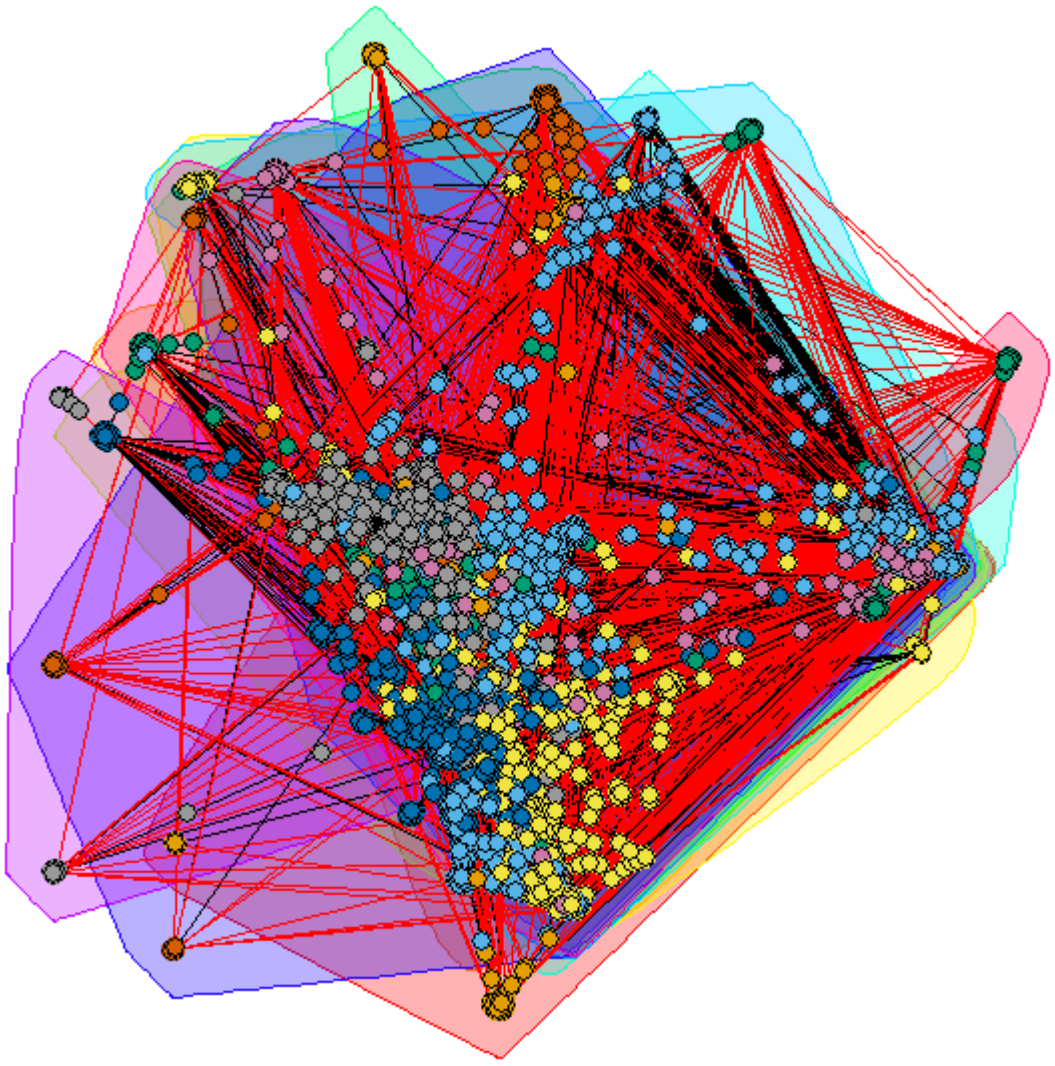


Fig. 9. Network Graph Plot Showing Community Structure Using Spinglass Algorithm: 19 Communities

4.4 Community Structure Using Label Propagation Algorithm

Finally, we apply Label Propagation algorithm to detect communities in the Flixster network. It is a simple approach in which every node is assigned one of *k* labels [13]. The method then proceeds iteratively and re-assigns labels to nodes in a way that each node takes the most frequent label of its neighbors in a synchronous manner. The method stops when the label of each node is one of the most frequent labels in its neighborhood. It is very fast but yields different results based on the initial configuration.

When this algorithm is applied to the Flixster network, we get a community structure as shown in Fig. 10. This algorithm produced a total of 28 communities in the network and the sizes of each of these communities are listed in Table VI. We see that the largest community in the Flixster network consists of about 3751 members. This algorithm was computationally the fastest amongst all the algorithms in assigning members to different communities.

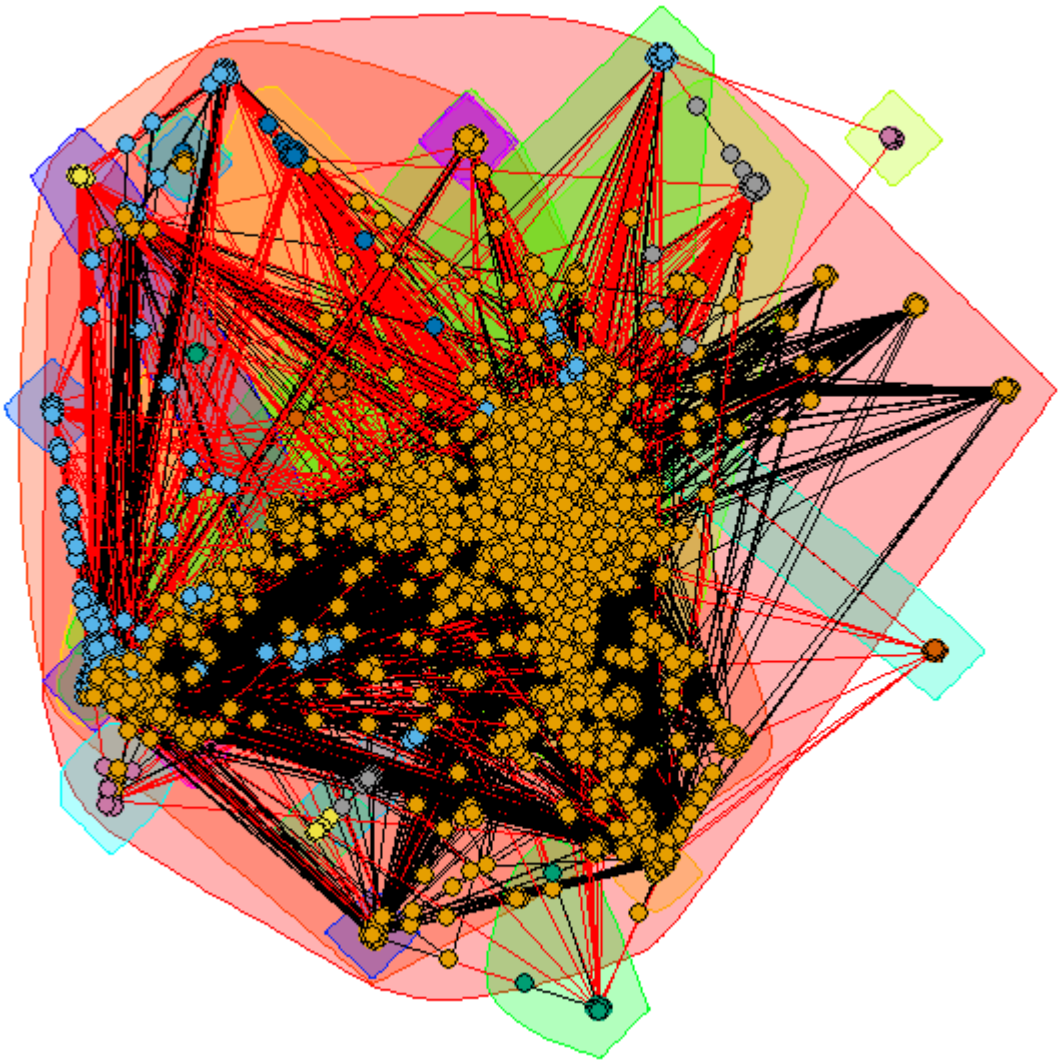


Fig. 10. Network Graph Plot Showing Community Structure Using Label Propagation Algorithm: 28 Communities

Table VI. Communities Detected by Label Propagation Algorithm

Community Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Community Size	3751	543	3	6	86	8	11	93	159	99
Community Number	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Community Size	30	7	5	17	36	10	5	19	5	48
Community Number	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**		
Community Size	10	4	10	3	13	4	7	8		

5. Common Neighbors in the Flixster Network

The viola plot in Fig. 4 illustrates the role of triadic closure in the Flixster network. It shows the density distribution as well as the box plots of the number of common neighbors for linked and unlinked pairs of nodes. Also, it shows the minimum and maximum values for number of common neighbors. The plot shows that vertex pairs that are incident to each other (i.e., edge) have higher neighbors compared to the vertex pairs that are not incident to each other (i.e., no edge). Further, it shows a slightly higher variance in the distribution of common neighbors for pairs of nodes connected by an edge compared to unconnected pairs of nodes.

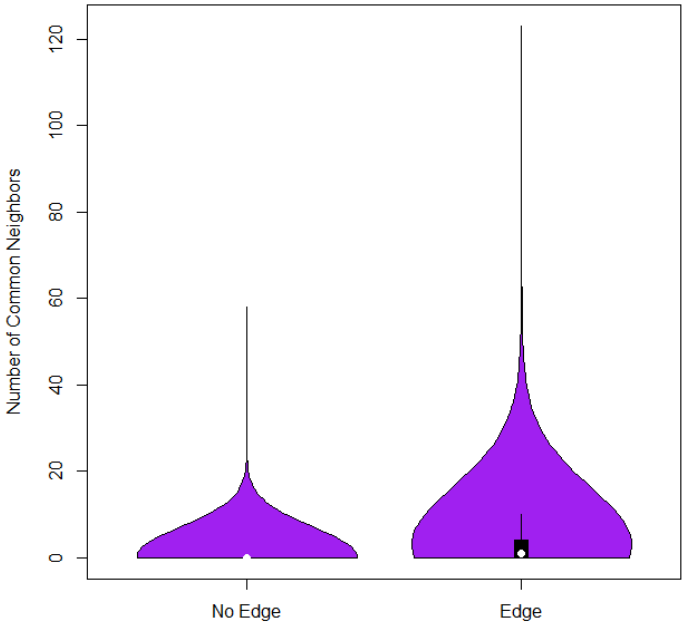


Fig. 11. Comparison of the Number of Common Neighbors Score Statistic in the Flixster Network, Grouped According to Whether an Edge is Actually Present Between a Vertex Pair, for all Vertex Pairs

6. Future Work: Epidemic Modeling

It would be interesting to see how the epidemic modeling, such as the standard Susceptible-Infected-Recovered (SIR) model, performs in the Flixster network to model the behavior of epidemics. Social contact networks and the way people interact with each other are the key factors that impact on epidemics spreading [15]. Studying epidemic modeling can help people to identify the dynamics of epidemics on networks and aid the decision makers to devise a suitable marketing/advertising campaign. In a community such as Flixster this study could be useful to target the right customers. Identifying ones who are most influential and targeting them will lead to the awareness about the movie propagating to the whole community. It can also help determine advertising campaign duration. This could potentially improve the accuracy of recommendation.

7. CONCLUSION

As an outcome of the analysis performed, I was able to discover the nature of participants in the popular movie website, as well as understand the effect of their interaction. It has a special consequence for the marketers of movie industry, as the consumers who have high level of engagement, and are actively evaluating different movies, are available. Also, the study of their interactions, helps in identifying the strong influencers in each community, as well as potential incorporation of popular ideas and themes in future movies to build more on already established interest. In fact, one of the potential opportunities, for popular movie production companies, or marketing agencies is to partner with such platforms, to leverage the data and intelligence about the consumer base, which would be more effective and efficient as well.

REFERENCES

[1] "Flixster." *WikiVividly*, wikivividly.com/wiki/Flixster.

[2] White Paper Reference: Inferring Social Networks Based on Movie Rating Data Chaofei Fan Department of Computer Science Stanford University, CA 94305 stfan@stanoford.edu Le Yu Department of Computer Science Stanford University, CA 94305 billyue@stanford.edu

[3] Cropanzano, R., & Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. *Journal of management*, *31*(6), 874-900.

[4] Cross, R., Borgatti, S. P., & Parker, A. (2002). Making invisible work visible: Using social network analysis to support strategic collaboration. *California management review*, *44*(2), 25-46.

[5] "Dataset: Flixster." *Social Computing Data Repository at ASU* - Flixster Dataset, socialcomputing.asu.edu/datasets/Flixster.

[6] Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008, April). Knowledge sharing and yahoo answers: everyone knows something. In Proceedings of the 17th international conference on World Wide Web (pp. 665-674). ACM.

[7] Aral, S., & Van Alstyne, M. (2011). The diversity-bandwidth trade-off. American Journal of Sociology, 117(1), 90-171.

[8] Granovetter, M. S. (1977). The strength of weak ties. In *Social networks* (pp. 347-367). Academic Press.

[9] Kolaczyk, E. D., & Csárdi, G. (2014). Statistical analysis of network data with R (Vol. 65). New York: Springer.

[10] Vo, N., Lee, K., & Tran, T. (2017, December). MRAttractor: Detecting communities from large-scale graphs. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 797-806). IEEE.

[11] Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, *74*(3), 036104.

[12] He, K., Li, Y., Soundarajan, S., & Hopcroft, J. E. (2018). Hidden community detection in social networks. *Information Sciences*, *425*, 92-106.

[13] Bishop, Michael BishopMichael. "What Are the Differences between Community Detection Algorithms in Igraph?" *Stack Overflow*, stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/.

[14] *Igraph*, igraph.wikidot.com/community-detection-in-r.

[15] Zhang, Z., Wang, H., Wang, C., & Fang, H. (2015). Modeling epidemics spreading on social contact networks. *IEEE transactions on emerging topics in computing*, *3*(3), 410-419.