An Exploration of Topological Properites of Social and Information Networks

Wendy Shi
Stanford University

Alfred Xue
Stanford University

Abstract

Although historically, network researchers have not come up with an agreed definition of what a social network is and what an information network is, work on social networks has focused primarily on graphs whose nodes are an avatar of person, and the edges represent some form of social relationship between these avatars. For example, the Facebook graph, which is considered the epitome of a social network, has profiles as nodes and friendship as edges.

In this paper, we hope to demonstrate that the topological properties of a social graph extend beyond such a set of networks, and can include networks whose nodes are not representative of people so long as their edges are driven by some human-based relationship.

1 Introduction

Plenty of research has been done on the topological properties of social and information networks. However, the social networks studied tend to have representations of people as its nodes, and their relationships as edges. For example, the Stanford SNAP website presents Facebook, Gplus, Twitter, Epinions and other networks that follow the above definition as social networks. We hope that by demonstrating this, the number of networks that can be studied will expand greatly.

Prior research has demonstrated that while there are not strong definitions of what is a social graph and what is an information graph, there are topological features that are associated with them. For example, Meyers et. al notes defines social networks as follows: Our operational definition of a social network is simply one that exhibits characteristics we observe in other social networks. These include high degree assortativity, small shortest path lengths, large connected components, high clustering coefficients, and a high degree of reciprocity. Although we feel that such a definition is circular in nature and thus not a strong definition for what a social network is, the topological features that they mentioned have been observed in almost all social networks.

One topological feature not mentioned in the above definition but a popular metric for determining how social a network is is the spid, or dispersion of path length distribution. This is measured as a ratio of the variance of the path length distribution to the mean of the path length distribution. Per Backstorm et. al., spid values smaller than one indicate that a graph exhibits characteristics of a social network, and values greater than one exhibit characteristics of a information network. The spid of the Facebook network has been computed to be 0.09, and that of Twitter to be 0.115.

Formally, our work is largely based on the research produced by Backstorm et. al. and Meyers et. al. The papers focus on the topological features of Facebook and Twitter, respectively. The Facebook paper lists computed values for path lengths, spid, clustering coefficients, and so on, creating a ground truth for what topological features a social network should exhibit. Meyers et. al presents similar topological features of the Twitter graph, but using the values computed by Backstorm et. al. as a basis, demonstrate that Twitter exhibits a combination of information and social features. While the Twitter paper demonstrates that the distinctions between social and information networks are fluid, our work demonstrates that graphs that are not thought to be properly social can exhibit topological features of social networks.

We selected two graphs that have historically not been considered a social network to explore. First, we curated a dataset from MyAnimeList, a popular web-

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Flordia, USA. JMLR: W&CP volume 54.

site that aggregates metadata about released Anime. Specifically, for each Anime, it gives individuals the ability to recommend other animes that users would enjoy if they enjoyed the current one. This presents an interesting graph, where the nodes are animes, something that isnt an avatar of people, and the edges are recommendations by people, something that is human, and we would argue social in nature.

Second, we took a citation network of physics papers to study. In this graph, the nodes are papers, which once again are not proxies of humans, and the edges are papers that have cited each other. The edges here are less social than what we would expect from a social network, but we argue are still social since they are based off being in the same academic circles, which is analogous to same physical area.

2 Methods

To evaluate each graph, we compare it to a set of null (generated) graphs and accepted social and information networks. The null graph we used was the Erdos-Renyi graph, which generates any potential edge with probability p. The social graphs we used were the following:

- Brightkite, a location-based social networking site. It formed a friendship network with 58,000 nodes and 214,000 edges.
- IMDB movie database, where the nodes represented actors and edges existed between actors if they co-starred in at least two movies. This graph has 17,000 nodes and 280,000 edges.

The information graph we used was a crawl of the Stanford network, where nodes were webpages and edges were hyperlinks to other pages within the Stanford network. There are 281,000 nodes and over 2 million edges in this dataset.

The topological properties we evaluated are the following:

- Average node degree: This is a standard computation of the number of edges each node contains. For social networks, we expect this to be relatively small research has demonstrated that individuals can hold at most 150 social relationships at any point in time.
- Size of two-hop neighborhoods: This is computed as the number of nodes that can be reached within a path of two edges from each node. Two-hop neighborhoods should be small in social networks due to local clustering.
- Largest Connected Components: This property is simply a computation of the components a graph can be partitioned into. A connected component is simply a subgraph such that there exists a path between any two nodes in the subgraph.
- Shortest path distance: This is computed as the shortest path between any two edges. This operation is expensive ($O(n^2)$) to compute, so we approximate it through random sampling. The average shortest path distance should be small for social networks.
- Shortest path index of dispersion: This is a feature introduced recently that is the ratio of the variance to the mean of the shortest path between nodes. Low variance is a feature of social networks researchers have proposed that a network can be defined to be social if its SPID is less than 1.

3 Results and Analysis

3.1 My Anime List and the Null Model

The first thing we set out to do is compare various properties of the MAL recommendation graph with a null graph in this case, Erdos Renyi to verify that the characteristics we observed were not simply a product of network properties.

We generated a graph based on the Erdos Renyi model with similar number of edges and nodes as the MAL recommendation graph.

3.1.1 Degree Distribution

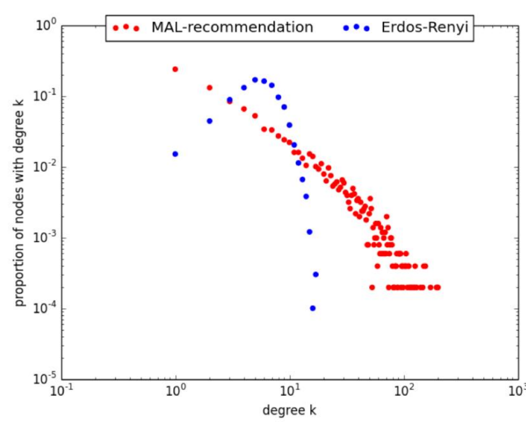

Figure 1: Degree Distribution of MyAnimeList compared to an Erdos Renyi graph

It is evident that the degree distribution of the MAL recommendation graph is nothing like what would be generated from a random graph. The MAL recommendation graph has a concentrated group of nodes with extremely high degrees in the hundreds while a randomly generated graph falters out before a hundred. This significant difference is also seen in the average node degree of the two graphs 11.53 for MAL recommendations and 6.01 for Erdos Renyi.

3.1.2 Two-hop Neighborhoods

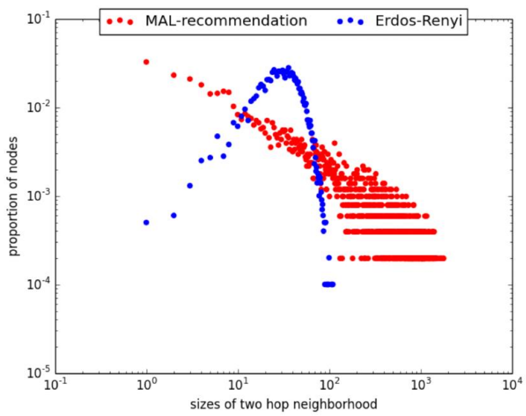

Figure 2: Two-hop neighborhoods of MyAnimeList compared to an Erdos Renyi graph

The difference in degree distribution also explains the stark difference in sizes of two hop neighborhoods. Especially when taking into consideration the ratio between the size of the two hop neighborhood and the size of the one hop neighborhood in both graphs, which is 20.7 for MAL recommendations and 5.99 for Erdos Renyi. Just by taking a single step, we are gaining access to a much broader portion of the nodes in the MAL recommendation than we wouldve with Erdos Renyi, suggesting a much more interconnected structure for MAL recommendation.

3.1.3 Degrees of Seperation

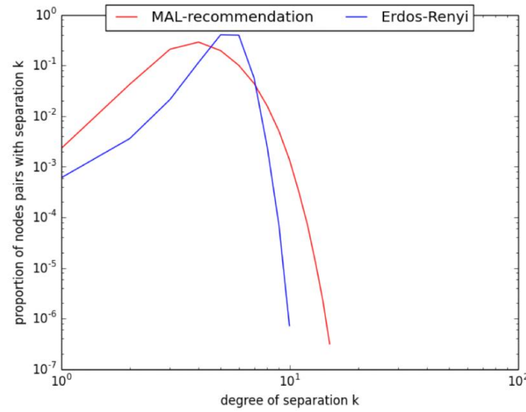

Figure 3: Degrees of Seperation of MyAnimeList compared to an Erdos Renyi graph

It should come as no surprise that the average number of hops on the shortest path between two arbitrary nodes is much lower for MAL recommendation graph. To generate the graph of degree of separation, we sampled 5000 nodes as source nodes from which we traversed to every other node, recording how many hops we took. The counts are averaged across all samples and plotted against the number of hops.

The average degree of separation of the MAL recommendations graph is 4.36, compared with Erdos Renyis 5.34, indicating that the MAL graph is by nature more closely interconnected than Erdos Renyi any arbitrary pair of nodes are on average 4 hops apart instead of 5.

Having observed differences in the above graph properties between Erdos Renyi and MAL recommendation graph, we can safely conclude that the MAL graphs characteristics are not simply a byproduct of network properties.

Our question then turns to the nature of the MAL recommendations graph itself. It is not what one would define as a conventional social graph representing societal link.

3.2 MyAnimeList and the Stanford Information Network

We first compare the MAL recommendation graph against a graph that has been classified as an information graph – the graph representing the connection between different pages that make up the Stanford website. The Stanford web graph is an information

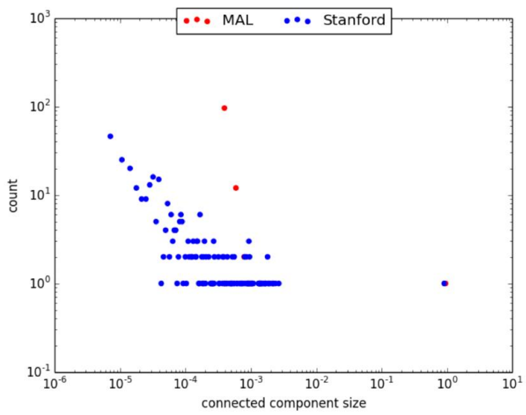

Figure 4: Connected Components of MyAnimeList and Stanford Information Graph

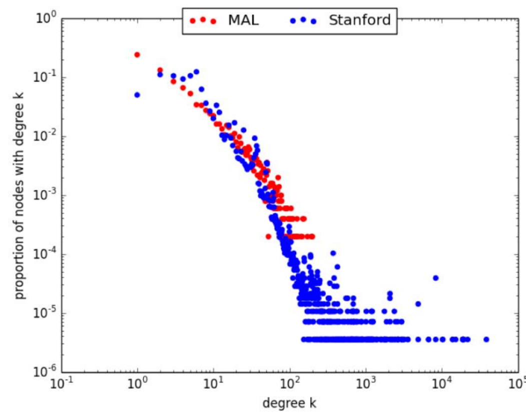

Figure 5: Degree Distribution of MyAnimeList and Stanford Information Graph

graph as its edges and nodes do not represent any sort of social interaction, but are instead channeling a flow of information.

3.2.1 Connected Components

Through Figure 4 we can already see the first signs of differences between the MAL recommendation graph and the Stanford web graph. In the graph, we are plotting the proportion of nodes for a given component size to the number of components in that size.

Though both of those graphs have one large component, the largest component of MAL recommendation graph actually accounts for 5% more nodes than the Stanford web graph. Furthermore, the pattern of the sizes of smaller connected components is very different. We see that MAL essentially only has three, maybe four component sizes aside from its biggest component, while the Stanford graph has more than a dozen tiny component sizes. Social graphs can manifest themselves in one huge connected component with many loner nodes, as we can see demonstrated by the MAL graph here, but it is less usual for a social graph to have many components of very different sizes as represented by the Stanford graph.

3.2.2 Degree Distribution

From Figure 5, we note that the difference in pattern for degree distribution is a bit less distinct as both of them follow the power law. However, what iss definitely evident is the difference in magnitude of node degree. As an information graph, Stanford web graph is not constrained by social interactions in the degrees of its nodes, and thus its perfectly reasonable for some nodes to have degrees in the thousands.

The edges on the MAL recommendations graph, on the other hand, are driven by users who take their time to write a recommendation for a particular anime. The user input aspect then becomes a limiting factor in how many edges each node can have while its not surprising for a dozen or so people to write recommendations for a given anime, its extremely unlikely for tens of thousands of them to do so.

3.2.3 Two-Hop Neighborhoods

As seen in Figure 6, another difference in patterns emerges when we compare the two hop neighborhoods of MAL recommendations graph and Stanford web graph. The Stanford web graph has a significant number of nodes with enormous two hop neighborhood sizes with a significant number of outliers that dont follow the downwards trend established by the first half of the graph, whereas the MAL recommendations graph is much more reserved in its two hop neighborhood sizes. This is likely due to local clustering on the MAL graph if animes A and B are often recommended together, and A and C are often recommended together, then its very likely that there are some common elements between A, B and C which would cause B and C to be recommended together as well.

3.2.4 Shortest Path and SPID

Finally, the factor that most definitely labels the MAL recommendations graph as a different graph type from the Stanford web graph the distance of the shortest

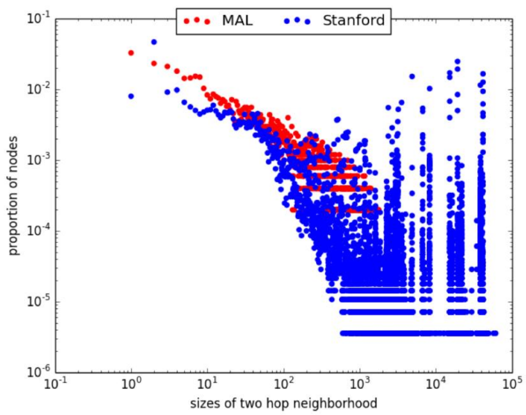

Figure 6: Two-Hop Neighborhoods of MyAnimeList and Stanford Information Graph

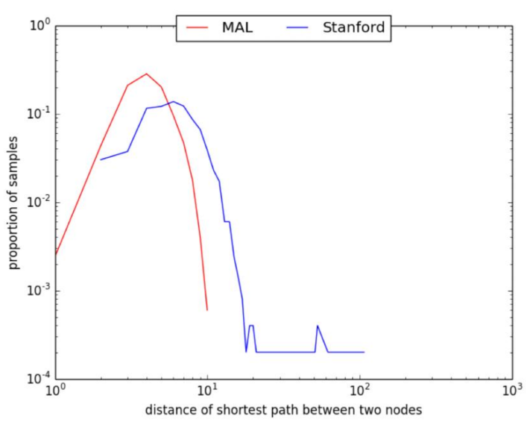

Figure 7: Shortest Path of MyAnimeList and Stanford Information Graph

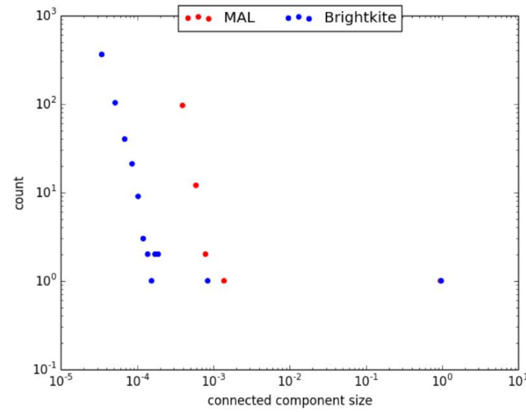

Figure 8: Connected Components of MyAnimeList and BrightKite

path between two arbitrary nodes. From Figure 7, we see that the peak of the Stanford distance graph sits to the right of the peak of the MAL graph, indicating that on average, nodes on the Stanford web graph are not as closely connected as they are on the MAL graph. The spid of the MAL graph is 0.43 while the Stanford graph has a spid of 5.31 clearly distinguishing it from both a social graph as well as the MAL graph itself.

Given the significant differences we found in our comparison between MAL recommendations graph and an information graph, we can safely conclude that MAL graph definitely does not fall into the category of an information graph, despite the fact that its nodes are pieces of information.

Does MAL recommendations graph exhibit characteristics closer to a social graph, then? We turn to comparing MAL recommendation graph against social graphs to confirm our suspicions.

3.3 MyAnimeList and BrightKite

We use the Brightkite graph as an example of a social graph. Brighkite is a location-based social service whose graph is represented in users (nodes) and friendships (edges). In that sense, Brightkite fits the image of a conventional social graph.

3.3.1 Connected Components

Once again, we start by comparing the size of the connected components between the two graphs. From Figure 8, we see that Brighkites largest connected component accounts for 98% of its nodes, while MALs largest connected components accounts for 95%. Fur-

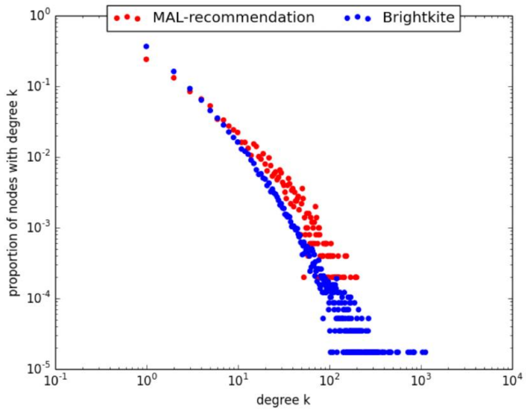

Figure 9: Connected Components of MyAnimeList and BrightKite

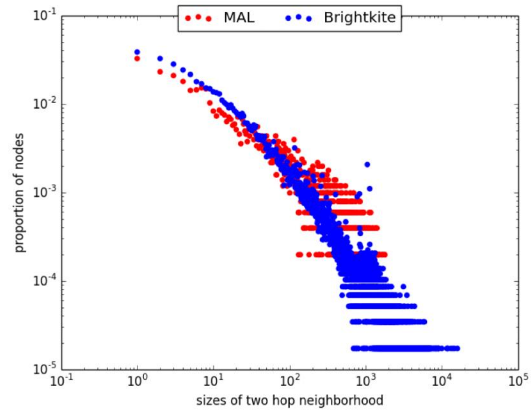

Figure 10: Connected Components of MyAnimeList and BrightKite

thermore, the distribution of the rest of the components though definitely not quite the same is not as starkly different as it was when we compared MAL to the Stanford web graph. We see a lot of loner nodes in both of these graphs, with a few mini clusters scattered about.

3.3.2 Degree Distirbution

Overall, we see from Figure 9 that the Brightkite graph seems to have an edge over the MAL graph in terms of maximum degree for a node, coming in at a little bit over a thousand. While thats still much higher than the MAL graphs maximum, it is still much more constrained when compared to some of Stanford web graphs nodes that have degrees in the tens of thousands. Its not a 1:1 similarity between the two, but MAL graph definitely leans closer to Brighkite than Stanford web graph in degree distribution.

3.3.3 Two-Hop Neighborhoods

From Figure 10, we see a very similar pattern emerging. Though the maximum sizes of Brightkite graphs two hop neighborhoods exceeds that of MALs due to its greater magnitude in degree distribution, the overall pattern of the MAL and Brighkite two hop graph follows a similar downwards curving trend. In terms of raw numbers, Brightkites average ratio of its two hop neighborhood to its one hop neighborhood exceeds that of MAL - 45 compared to 20 though that can be explained by Brighkite having a more sizable largest connected component on top of a few high degree nodes.

3.3.4 Shortest Path and SPID

Upon first glance on Figure 11, this graph seems to suggest that the MAL graph is more closely knit than the Brightkite graph, with a somewhat higher proportion of sampled pairs having short distance. However, the average of the two graphs ends up being a similar value 4.34 for MAL vs 4.92 for Brightkite.

Both of these graphs have spid values of below 1 a trait of social graphs 0.43 for MAL and 0.26 for Brightkite. All of these similarities between the MAL graph and the Brightkite graph suggest a close relationship between the MAL graph and a social graph than an information graph.

3.4 MyAnimeList and IMDB Actors

Finally, we will compare the MAL graph against another graph with humans as nodes the actors graph. The difference between the actors graph and the Brightkite graph lies in the fact that the edges in the actors graph do not represent something as clear-cut as friendship but rather simply the piece of information that two actors collaborated together in a movie. However, there is a social aspect to collaboration as well, as actors who have often worked together are likelier to share social connections. Thus, the actors graph is at its core a social graph.

How does the MAL recommendations graph compare against this slightly different social graph, then? The answer is, quite similar to how it compared against the Brightkite social graph. We will not examine each property of the graph in as great of a detail as before, but note that MAL recommendations graph and

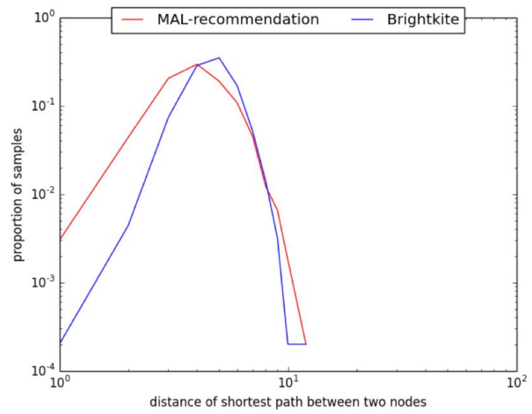

Figure 11: Shortest Path and SPID of MyAnimeList and BrightKite

actors graph exhibit similarity in sizes of connected components as well as average shortest distance between nodes. The average degree of separation is 4.35 for MAL recommendations and 4.89 for actors, and their spid values are practically identical 0.43 and 0.44, both very typical values for social graphs. The degree distribution of the actors graph has a slightly different starting curve compared to MAL recommendations graph perhaps something to do with the fact that its rare for an actor to just star in a single movie but the overall pattern is still very similar.

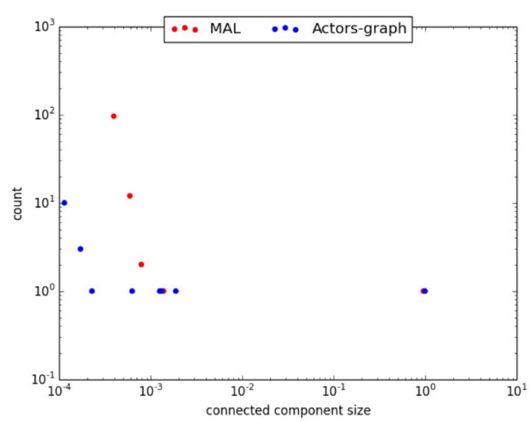

Figure 12: Connected Components of MyAnimeList and Actors network

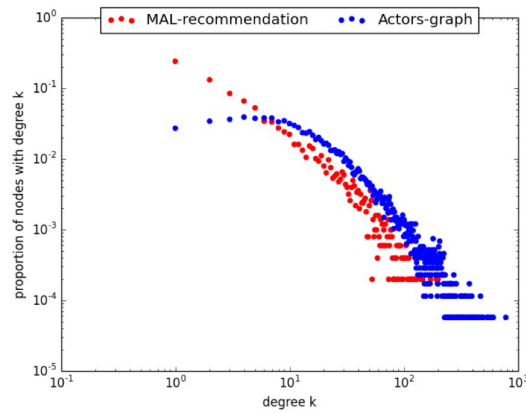

Figure 13: Degree Distribution of MyAnimeList and Actors network

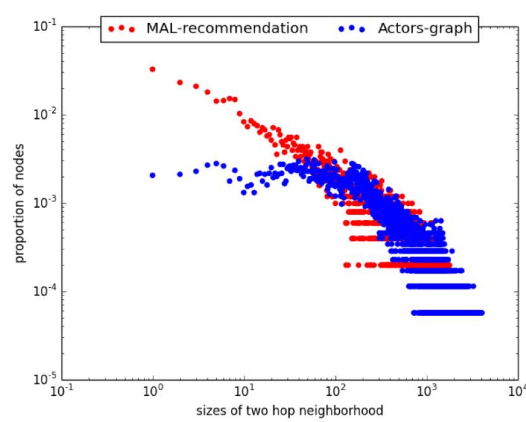

Figure 14: Two Hop Neighborhoods of MyAnimeList and Actors network

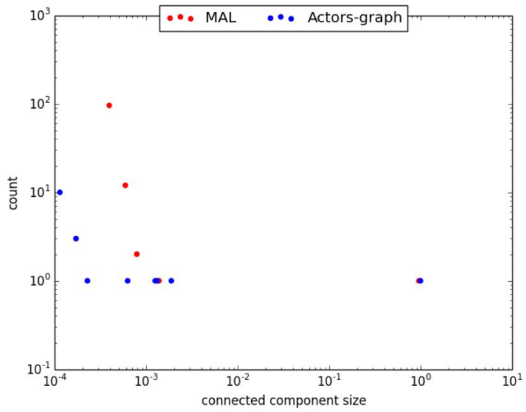

Figure 15: Shortest Paths of MyAnimeList and Actors network

3.5 MyAnimeList and a Citation Graph

We have now demonstrated that MyAnimeList does not resemble a information network nor the Erdos-Renyi null model. It is clear that MyAnimeList is toplogically similar to typical social graphs, and we want to show that MyAnimeList is not unique in that it is a graph whose nodes are not humans but whose graph structure nevertheless is that of a social graph. To do that, we compare a citation network to MyAnimeList.

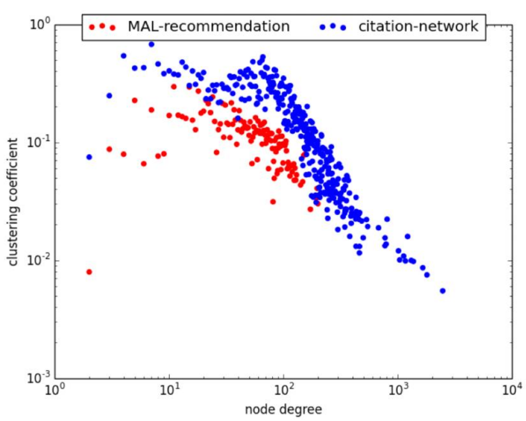

Figure 16: Clustering Coefficient of MyAnimeList compared to a a Citation Network

We can easily see from Figure 4 that the average clustering coefficent partitioned by node degree is extremely similar for MyAnimeList and the Citation graph.

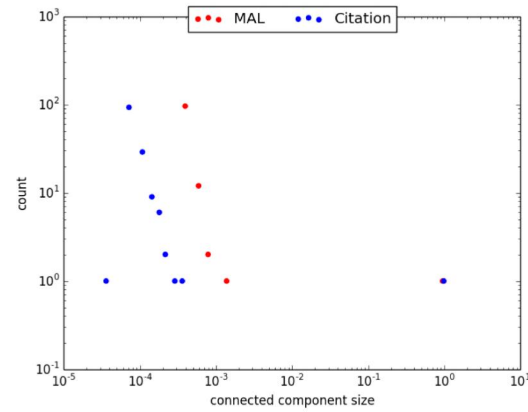

Figure 17: Largest Connected Components of MyAnimeList compared to a Citation Network

As shown in Figure 5, MyAnimeList and the Citation Network have very similar connected component distributions. The largest connected component of MyAnimeList contains 95% of the nodes, while the largest connected of the citation graph contains 98% of the nodes.

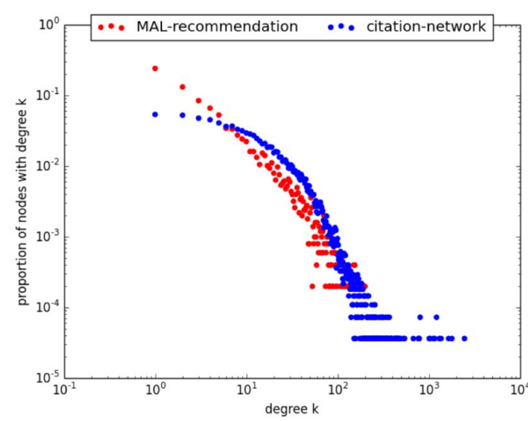

Figure 18: Degree Distribution of MyAnimeList compared to a Citation Network

Although there are a greater proportion of nodes in MyAnimeList that have a very small degree, and a greater proportion of nodes in the citation network that have very very high degree, we can see that the overall structure of the degree distirbution of the two graphs are roughly similar. However, the citation network deviates from standard social networks here, as there are too many nodes with very high degree.

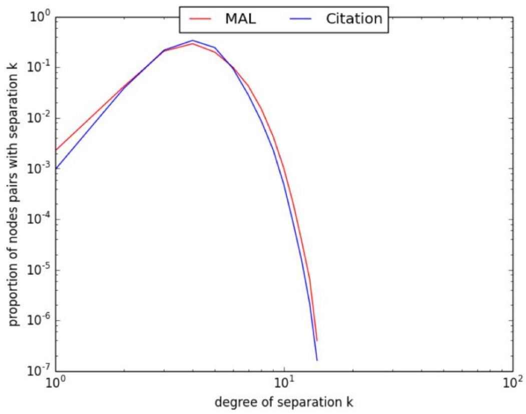

Figure 19: Shortest Paths of MyAnimeList compared to a Citation Network

Once again, the shortest path distribution of MyAnimeList looks very similar to that of the Citaiton Network. Through these graph comparisons, it becomes very apparent that the Citation network and MyAnimeList have similar topological properties.

4 Conclusion

Through the toplogical features analyzed, we see that although there are slight deviations between MyAnimeList and standard social graphs, MyAnimeList approximates social graphs far better than it does null graphs or informational ones. This indicates that as far as graph strucutre is concerned, MyAnimeList is a social graph. We demonstrate similar results for the citation network, which suggests that MyAnimeList is not a unique case where a graph with non-human nodes exhibits social graph characteristics. We do note that MyAnimeList and the citation network are closer to each other than they are to strictly social graphs, but also that the deviations are not that large.

Although our sample size is too small to make a confident conclusion, we argue that the interesting features of a social graph are derived from two properties of the graph being studied, and that these properties are not unique to human nodes and social-based edges. The first is that the number of edges per node is relatively small and does not scale significantly with the size of the graph. This is something that is observed through the degree distribution graphs above. The second, and more interesting one, is that the nodes can be graphed in such a way that the edge probabilties are approximately inversely proportional to the distances between nodes. For traiditional social networks, the nodes can be placed geographically. For MyAnimeList, node placement is derived through genre. This property allows small two-hop neighborhoods and low degrees of seperation between edges. In conclusion, the topological properties of a "social graph" are not unqiue to what are traditionally thought to be social graphs, and there exist a class of graphs that are not properly social but exhibit the same features. Both parsing and analysis code is available at https://github.com/XueAlfred/MALAnalysis/.

References

[1] Kleinfield, J. 2002. Could It Be A Big World?. Forthcoming Society 2002.

[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S.Vigna. Four degrees of separation. WebSci 2012.

[3] S. A. Myers, A. Sharma, P. Gupta, and J. Lin. Information network or social network? The structure of the Twitter follow graph. WWW Companion, 2014.