Comparing IMDB Network of Actors to Random Graph Models

Dan Kimball
School of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: kimbad@rpi.edu

Eric Herdzik
School of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: herdze@rpi.edu

***Abstract*—The abstract goes here.**

I. Introduction

We sought to determine whether there was/is an elite circle of actors in movies. To this end, we attempted to find a random graph model that represents the graph of actors connected by appearance in the same movie, which maintains the measure of clustering coefficient. We have compared traditional random models to the IMDb dataset, and found them lacking. We could get the degree distributions to match, but the clustering coefficient was far lower than it should be. We attempted to develop a new model that captured the clique nature of the network. Several models were theorized, and one implemented. Computational complexity of building a model based off cluster coefficient (a $O(n^3)$ operation to find) means it was unfeasible to implement all of them in the time frame.

II. Related Work

Much work has been done on social structures, but not very much on systems built out of cliques (as the IMDb set is). So, we focus our look at work done with the IMDb dataset, to ensure they are doing work relevant to this unique structure.

[1] shows that community size tends to be inversely proportional to the quality of the community. The quality of a connection in the IMDb dataset is the number of movies two actors were in together. This suggests that as an actor acts with more other actors, the quality of the community they are in degrades, and they start to blend into a larger super-community. This makes some intuitive sense with regards to actors; we would expect that lesser known actors would act repeatedly with the same small set of actors, and as they become more well known, they will branch out and act with other big name actors.

[2] attempts to translate bipartite graph data (such as the IMDb data) into a unipartite affiliation graph, and estimate an expected degree distribution on this new graph using attributes of the original graph. This work showed that the affiliation graph of social networks, inclduing the IMDb network, exhibit strong preferential attachment.

[3] attempted to develop random models for both bipartite and unipartite graphs to maintain degree distributions. They succeeded in developing a model for the IMDb dataset that matches the degree distribution of both actors and movies, using a method that builds clusters with strong preferential attachment. Although the degree distribution is maintained, the clustering coefficient is not.

[4] develops a model that accounts for both the aging of participants (unable to participate in infinite movies) and accounts for local clustering. The local clustering in particular is applicable to the IMDb dataset; each movie forms a clique, and actors are clustered by the movie they were all in.

[5] defines an attribute *Social Inertia* as a measure of the participants' desire to maintain social links. The IMDb dataset was measured, and an unexpectedly low Social Inertia was found. The author is unsure as to the reason, but this result means that preferential attachment might not be the best way to model the system.

III. The Collaboration Graph Model

The IMDb data is presented as a list of movies with the week of release and a list of the major actors in the movie. We represented the data as a collaboration graph consisting of one node per actor, with an edge between two nodes if the two actors had been in a movie together. These edges have a weight equal to the number of movies the two actors had appeared in together. So, each movie adds a clique of edges to the graph. There are [number of nodes] actors in the graph and [number of movies] movies between them.

To allow analysis of the weights, we used a thresholding model. We only added an edge to between two actors if the number of movies they acted in together exceeded a certain threshold. We decided to use a threshold of 3 movies because it provides a reasonable balance between network size and link significance. Defining links between actors through 1 or 2 movies would likely result in abnormally high clustering coefficients due to the sheer number and size of cliques building the network since movies contain multiple actors. This implies any analysis of triadic closure, etc. could likely yield meaningless results as they would be primarily influenced by the definition of the model rather than the actual relationships between actors.

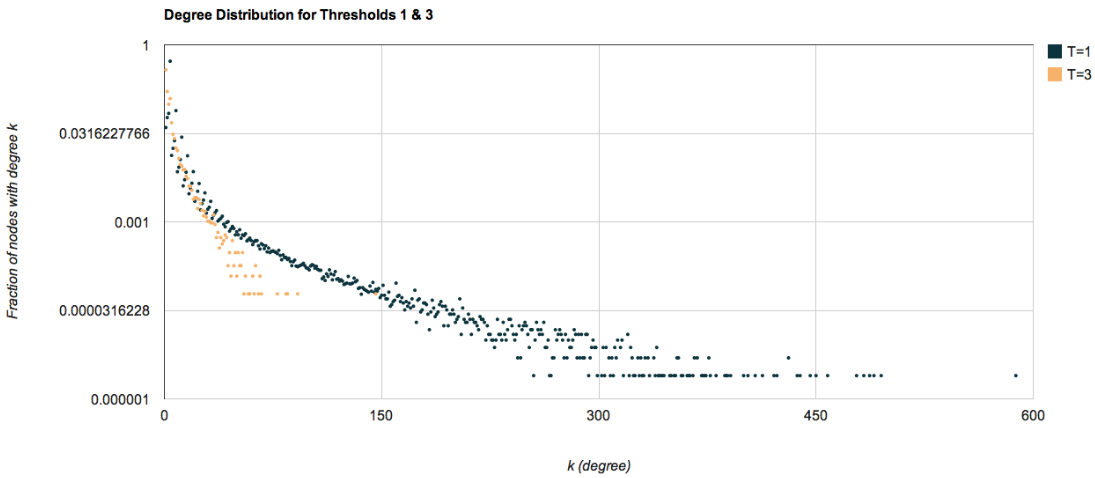

Fig. 1. Degree distribution for thresholds 1 and 3. Blue is data for a threshold of 1 and orange is data for a threshold of 3. The y-axis is plotted on a logarithmic scale, illustrating a polynomial decay in both plots.

IV. General Properties

For threshold of 1, i.e. the entire network as is, the graph contains 404006 nodes and has a clustering coefficient of approximately 0.752. An explanation for the clustering coefficient for being so large is that each movie added to the network represents a clique of actors where each triangle within that set of actors must be closed. Also consider that the average number of movies an actor acted in is only 2.82. If an actor a is connected to actors b and c, the probability that b and c close the triangle increases with decreasing number of movies that a acted in. This is true in general for the collaboration graph over any threshold because the less movies for a, the more likely the sets of movies defining links between a, b and a, c will overlap.

For threshold of 3, the graph contains 16619 nodes and has a clustering coefficient of approximately 0.289. This clustering coefficient is still fairly large compared to the random graph models as seen in the results. In general for any threshold, the degree distribution decays polynomially rather than exponentially as seen in fig. 1 indicating likely some degree of preferential attachment during the growth of the network.

V. Our Results

We first modeled the data using the various random models [numbers, graphs here]

We attempted to develop a new random model, that would be able to capture both the degree distribution and clustering coefficient. We experimented with adding cliques to the graph incrementally, rather than edges. This yielded a perfect degree distribution, and maintained the distribution of actors per movie, but very inaccurate clustering results. For a threshold of 1, the clustering coefficient was very high [number here]. This is because edges are added in clusters, every movie that had more than two actors contributed massively to the clustering coefficient. However, when the threshold was increased to 2, the clustering coefficient dropped to 0. This demonstrates that there is preference in the selection of actors toward each other, since choosing actors randomly led to no three actors all being in two movies together.

We believe that combining the principle of clique generation with preferential attachment would yield strong results. One way to do this would be to preferentially place actors in movies with actors whom they have already acted with. Another would be to preferentially place actors in movies with actors whom have acted in more movies, regardless of who those movies were with. Both models represent reality, and we believe a weighted mix of the two would yield the best results.

Another model, which does not use the clique principle, is thought to produce an accurate model of the strong triadic closure shown in the IMDb dataset. This model would be constructed as two steps. First, seed the network with no triangles using preferential attachment based on the number of movies an actor was in (edges can be created multiple times; increase the weight each time this happens). For the second step, iterate across all non-added edges that would close a triangle if added. We believe the odds of the triangle being closed is directly related to the degree of each of the three nodes, and the weight of the two existing edges. If the two nodes on the possible edge have high degree, then the odds of the triangle closing go up. If the third node had high degree, then the odds of the triangle closing go down, as the shared connection becomes less unique for the two other nodes. If the existing edges have a high weight, then the odds of the triangle closing go up.

VI. Conclusion

We have determined that traditional random graph models are inadequate to describe the social processes occurring within the IMDb dataset. While the models were able to accurately describe the degree distribution, they lacked the ability to capture the clustering properties of the data.

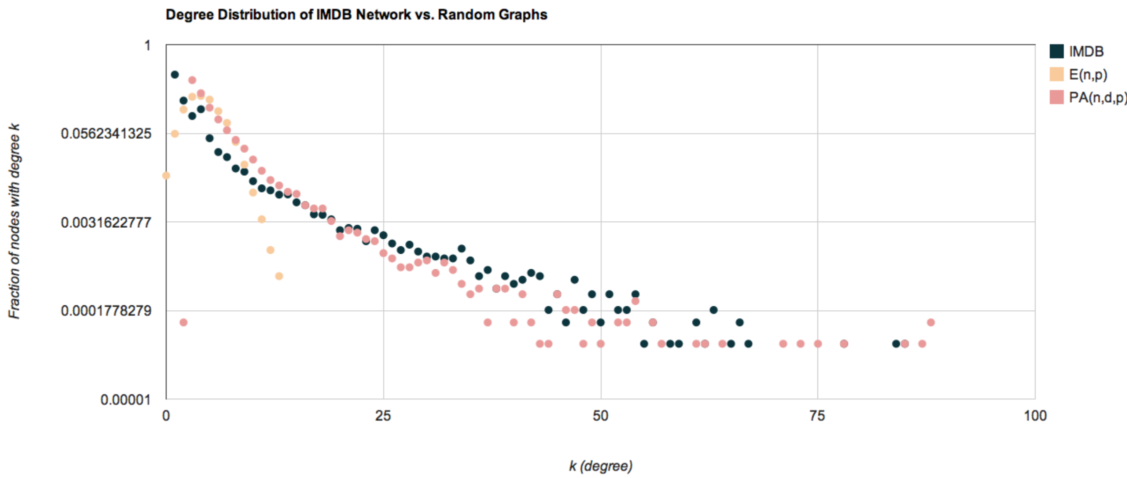

Fig. 2. Degree distribution for actual vs E-type and PA-type models. We used $PA(n, e/n, .7)$ which creates a degree distribution very similar to the actual degree distribution. The E-type graph resulted in a distribution that decayed exponentially where the actual and PA-type graphs have polynomial decays.

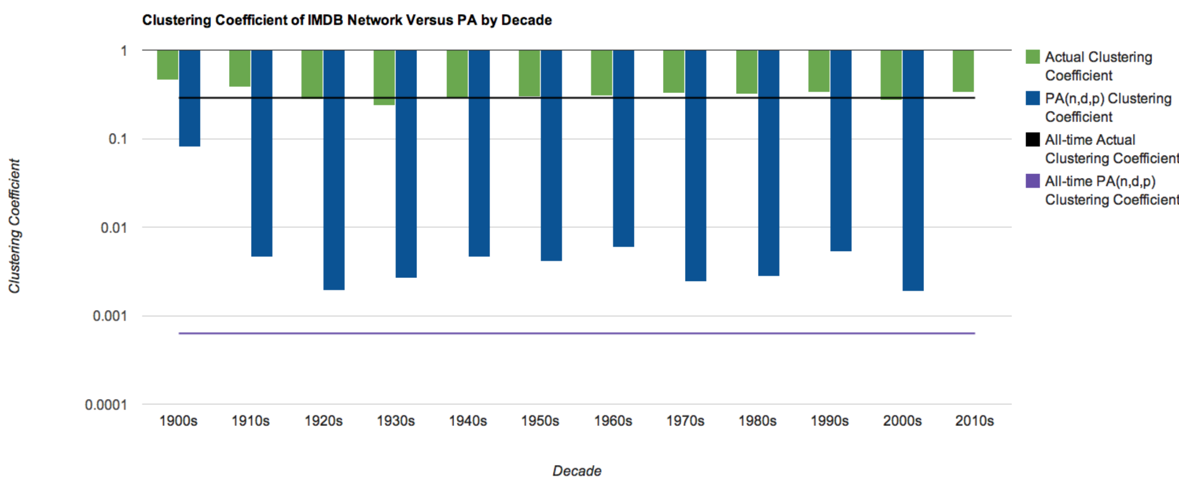

Fig. 3. Clustering coeficient of IMDb network by decade. In green is the actual network and in blue is the PA-type graph. The y-axis is on a logarithmic scale, so you can see that clustering coefficient of the PA-type graph is significantly less than the clustering coefficient of the actual network. Also, if you look at the the 2 lines representing the over-all-time clustering coefficients, it is clear that the monthly values do not differ significantly from the all-time values for their respective graphs. Because of this, time is likely not a major factor in the clustering coefficient being so high for the actual network.

References

[1] 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*, vol. abs/0810.1355, 2008.

[2] S. Chojnacki, K. Ciesielski, and M. A. Klopotek, "Node degree distribution in affiliation graphs for social network density modeling," in *SocInfo*, 2010, pp. 51–61.

[3] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. Suppl 1, pp. 2566–2572, 2002. [Online]. Available: http://www.pnas.org/content/99/suppl.1/2566.abstract

[4] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras, "Self-organization of collaboration networks," *Phys. Rev. E*, vol. 70, p. 036106, Sep 2004. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.70.036106

[5] J. J. Ramasco, "Social inertia and diversity in collaboration networks," *arxiv*, vol. physics.so, no. 1, p. 5, 2006. [Online]. Available: http://arxiv.org/abs/physics/0612006