Math 168 Essay 2

Pratyusha Majumder

October 2020

I am interested in applying network analysis techniques to many types of networks, particularly their application in online social networks, and understand how this analysis can be leveraged to build intelligent recommendation systems. An ideal data set that can be used to study this is one that incorporates user-user relationships and a particular type of user behavior, so that one can investigate if patterns in behavior arise between users who have a noteworthy user-user relationship. User-user relationships should ideally be able to be quantified (i.e, by edge weights that indicate level of friendship) to further determine how acquaintance-ships and friendships differ in their influence on user decisions on a platform.

Many consumer-facing technologies leverage user social networks to build their own recommendation systems, often as a means to achieve a business goal. Companies such as Google use user data relating to videos and channel subscriptions in order to recommend new channels and videos, on their Youtube platform. Facebook analyzes pages users engage with (user-page relationships and user-ad relationships) and uses it to determine the most effective advertisements to show you. This showed me how social networks are an important commodity in today's world of technology and informed my interest further. I was able to find a Youtube social network data set that provides data on user-user friendships on the platform, provided by Mislove et al. (2007) [1]. This may provide a good basis for the project, however it does not provide any data on the similarities and differences between the content consumed by the people on the platform. Also, accessing social network data can be challenging as some of it may be classified as intellectual property for their respective companies and this may be a hurdle we face as a group. To combat this, we could use an API to access the data however I have a lack of experience in using APIs to extract data from websites so it may present as a personal challenge.

Two of the most promising data sets I found from the Stanford Network Analysis Project (SNAP) were the Twitter data set and the Twitch data set. The Twitch data set represents user-user friendships of gamers which are categorized based on the language they stream in. The nodes also have multiple features such as the games played and liked, location and streaming habits of users, provided by Rozemberczki et al. (2019)[2]. The Higgs Twitter data set from SNAP, retrieved from a study done by De Dominico et al., represents four directional networks based on retweets, replies, mentions and follower networks based on users participating in these engagements [3]. Both these data sets have the potential for extensive analysis as they provide extensive information about user-user relationships and

their behavior.

I'd like to evaluate these networks by investigating closeness centrality within the networks. Closeness centrality will give us an idea about the mean distance from a node to the other nodes, as stated in Newman's book of Networks [4]. For nodes representing users on the Twitter platform, one can hypothesize that those with higher closeness centrality to others would have their ideas spread through their community faster than those with a lower mean distance to others. The Higgs Twitter data set was created to see how fast information flowed before and after the announcement of a new particle with features similar to the Higgs Boson. Closeness centrality analysis in this case could help identify whether some Twitter users played a larger role in the spread of this information than others. This has future implications about which users to target in order to quickly and efficiently prevent the spread of false information, should it be necessary.

Using community detection algorithms to identify different communities either within the Higgs Twitter data set and the Twitch data set may also be insightful. Investigating the community structure of the networks with multiple community detection algorithms could be interesting to see help use see which algorithm performs the best and provides the most contextually meaningful results for our data set. Using multiple community detection methods may help us identify different communities that would not have arisen if we used one method alone. After this process, splitting the whole network into relevant clusters and conducting analyses on each of the clusters could provide us with some useful insights.

Newman's book mentions the analysis of cores within a network, which is another method that piqued my interest. Finding k-cores within the data set involves finding connected sets of nodes where each node is joined to at least k other nodes, and this can be done for multiple values of k. This process can be conducted recursively by starting with a network and removing any nodes that have degrees less than k, and the edges connecting to it, for k = 1, 2, . . . , n [4]. Separating the data into different cores and seeing how the nodes and edges in the highest k-cores differ from the peripheral nodes can also provide useful insights about more highly connected groups and less highly connected groups within the networks.

1 References

[1] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B., Measurement and Analysis of Online Social Networks, in *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, Association of Computing Machinery, San Diego (2007).

[2] Rozemberczki, B., Allen, C., and Sarkar, R., Multi-scale Attributed Node Embedding, (2019).

[3] De Domenico, M., Lima, A., Mougel, P., and Musolesi, M., The Anatomy of a Scientific Rumor, in *Scientific Reports 3*, (2013).

[4] Newman, M. E. J, Networks, Second Edition, Oxford University Press, New York (2018).