Honors Thesis Proposal

Title: Using Group Affinity to Predict Community Formation in Social Networks

Committee:

- Faculty Advisor - Dr. Webb: Thesis advisor with whom I currently work for as a research assistant. Dr. Webb and I have worked together for around two years now, and this thesis directly draws from that work.
- Faculty Reader - Dr. Evans: Dr. Evans is part of the Network Theory research group with Dr. Webb and has consistently provided valuable input.
- Dr. Griffin: Mathematics department honors coordinator

Timeline: Most analytical work has been completed as this project began around November 2018. Final testing will be completed by March 1st, with the proposal ready to be completed by April 1st. Publication and defense is scheduled to be completed by May 1st.

Funding/IRB Approval: No funding or other approval is necessary.

Culminating Experience: I have already presented a preliminary concept at the BYU Student Research Conference and would like to continue to present my findings at available conferences.

Project Purpose: Develop a new method to predict the future group/community formations of a network given the previous affiliations already observed in those observations.

Project Importance:

Many studies exist about link prediction and different methods of community detection, but little literature we know of suggests how to predict communities, or in other words, perform community detection in a dynamic network.[1] In the same sense that there is value in being able to predict links in a network, the capability of predicting the groups within a network structure have important implications in several settings as it tells us about that behavior of a network. This project is an approach to quantitatively measure the way communities form use such information to help predict how those communities will ultimately be established at a later point in time.

Project Overview:

Network theory has been the recipient of extensive attention in recent years thanks to the surge of many social networking websites like Facebook and Twitter, as well as big data which allows for more detailed observations on the connections between things in many fields such as biology. One item that makes the field interesting is the fact that networks found in very different settings can have remarkably similar structures, and networks in very similar settings, such as Facebook and Twitter, can be found to have very different set-ups. As a result, much information about a network can be discerned by analyzing it mathematically.

---

[1] Javed, Younis, Latif, Qadir, Baig; Community detection in networks: A multidisciplinary review.

Two of the larger questions in the field of network theory are those of link prediction[2] and community detection[3], the practice of trying to predict who are what may be connected to some other person or thing, and the attempt to detect groups within a network. A prime example of link prediction is when a social networking site such as LinkedIn suggests "people you may know" which is often based on information like where you work and who else you know. An example of community detection is when police try and see who all may belong to some criminal organization, given who has been interacting with who. In this project, we combine the two ideas, and work towards trying to predict the way groups form over time.

This "community prediction" is *a subject matter that has had very little work done* to explore it, partly because of the vastness of the problems of both link prediction and community detection, hence the contribution of this project. It makes much intuitive sense that communities could be predicted based on historical information, but actually implementing such prediction is very unintuitive. The first main question we need to tackle is what information we would use to make a prediction.

Methods[4]:

Our attempt at predicting communities which form over time is based off a measure which we label as an "affinity score" which suggests a likelihood of a node in a network belonging to a community. Using these scores, we then implement Machine Learning methods, algorithms that use computational power to reach an answer, with all the affinity scores up until the nth period to see if these scores contain predictive power. These affinity scores are derived from static community detection algorithms such as the Kernighan-Lin Bisection algorithm and Spectral partitioning. Some community detection algorithms like spectral partitioning have an inherent affinity score associated with how it partitions graphs (an eigenvector corresponding to the second largest eigenvalue that gives a hierarchal score for each node). Others, like the Kernighan-Lin algorithm can be rerun several times to extract a probability of a node being in one community or the other, and that probability is what we use as an affinity score.

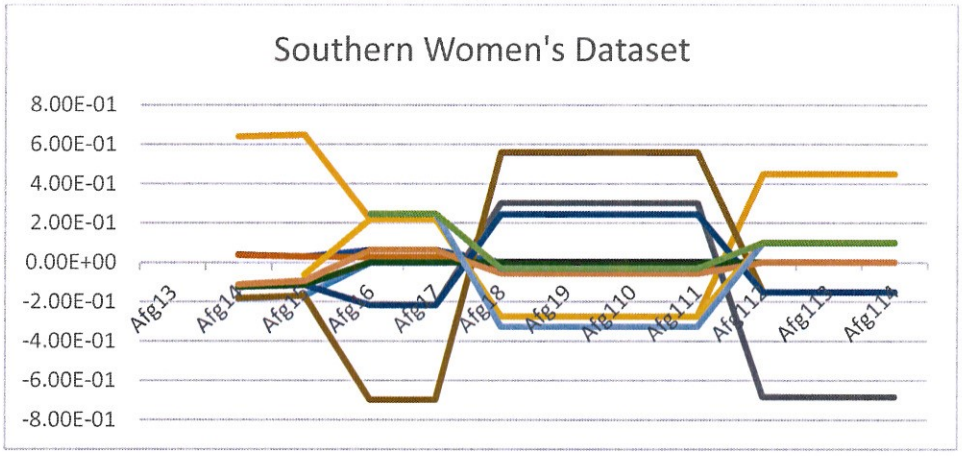
We began by using small datasets that were manageable in size, mainly one by the name of the "Southern Women's dataset." There were many limitations of using this data as oftentimes algorithms are optimized for large datasets to determine useful information. Nonetheless, using a dataset that was so small allowed us flexibility in adjusting out methodology. For example, with this dataset, we were able to see that the technique of spectral partitioning proved useful, as seen with the graph below that measures the affinity for one group or the other for each woman over time. There we can see different individuals, as labeled by the color, "switching sides" as time goes one, extremely pertinent information if we are trying to predict communities as they form.

---

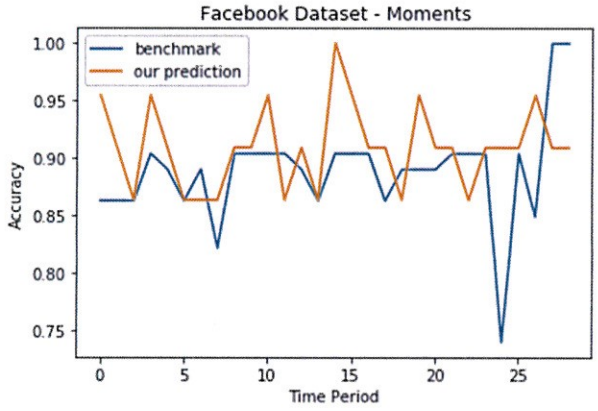
[2] Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." *Journal of the American society for information science and technology* 58.7 (2007): 1019-1031.

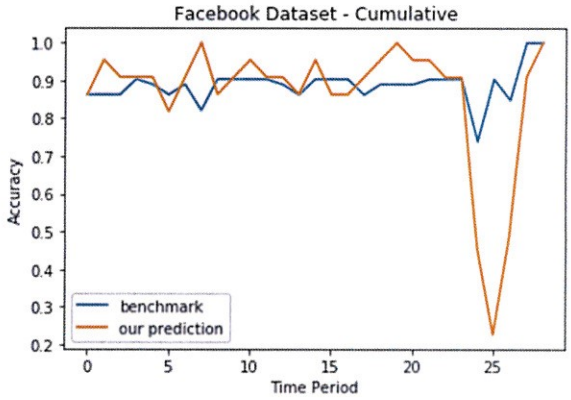
[3] Fortunato, Santo. "Community detection in graphs." *Physics reports* 486.3-5 (2010): 75-174.

[4] To truly understand much of the methodology one would need some basis of understanding in Linear Algebra. In this section I try to minimize the use of mathematical vocabulary, while still retaining the key terms, to explain the process that this project undergoes.



Once we tested with the southern women's dataset, we moved on to larger datasets[5] such as Facebook to see if there was some consistent pattern, which there was. Below we see two graphs measuring the overall accuracy that demonstrate the effectiveness of using the affinity score from spectral partitioning. We note that using cumulative information over time normally gives us a stronger result than if we compared that information at a specific moment. This supports the idea that we can predict communities by using previous information with these affinity scores.





While spectral portioning proves to be a useful technique, it does have its own limitations, mainly how it only observes two groups. Though we ultimately look at our ability to predict how two groups form, other techniques we tried can predict into many kinds of groups. We eventually generalize into a centrality-based affinity score that allows for the use of any community detection algorithm (in our tests we used Kernighan-Lin, Fluid communities, and k-cliques) by considering the centrality measure of a node in a community as discerned by a static community detection algorithm. Thus, in this case, our affinity score is a tuple that uses the community a node is determined to be in, indicated by the node with highest centrality in the

[5] Data sets come from the Konect database: http://konect.uni-koblenz.de/

new community, and the node's centrality score in that new community. We also capture the measure of the change in centrality a node may have, thus allowing our affinity score to be in the form of a triple. This technique is what most of this project centers around, and a full explanation would go beyond the scope of this paper.

Results:

While our results are generally positive, the performance does vary depending on the type of community detection used the nature of the network used. We measured our results against two kinds of benchmarks, the first being how closely the community detection algorithm's result aligned with the final time-period's results, and machine learning on the affinity score but without incorporating the time-series. In some cases, this second benchmark was shown to be on par with machine learning on the entire time series (e.g. MIT Reality Experiment). However, in most cases, this we see that using a history of affinity scores proves more effective.

To fully complete this project, most of these details regarding the results will be finalized and formalized, as this analytical work has already been done. According to how the research has gone so far, I anticipate that it will help introduce a new section into understanding network theory and bridge two pieces of that field in an intuitive way that has not yet been rigorously done.