Stanford Memes Group: Network Construction, Community Detection, and Link Prediction

Shelby Marcus, Yardena Hirsch

Abstract—Prior to social media, constructing and analyzing social networks relied on individual survey responses, analysis of cellular phone logs, or email records. With the advent of social media, we have the opportunity to collect publicly available data and construct social networks. As students at Stanford, we have the ability to collect data from a very active Facebook group of over 28,000 members where students post memes. The group contains rich content; there are daily posts that often receive hundreds of comments, and it is common practice for friends to tag each other in comments. We extracted the comments from posts and constructed a graph that we hope is representative of the social network that exists offline at Stanford. We apply motif detection, link prediction, and community detection to this graph in an effort to learn more about this community.

Index Terms—Social Networks, Motif Detection, Community Detection, Link Prediction

1 Introduction

Since its emergence in November 2016, "Stanford Memes for Edgy Trees," commonly referred to as Stanford Memes, has created a virtual community stronger and better known than any other online community on campus. With over 28,000 users and high daily engagement, the Stanford Memes group is a rich data source for constructing and analyzing a social network. The Stanford Memes group is a very active community: there are multiple daily posts and some posts receive up to a thousand likes and hundreds of comments where friends tag their friends.

We construct a weighted directed graph where nodes are users and an edge between node A and node B signifies that user A tagged user B in a post. It is common for a user to tag their friend in a post that they think their friend relates to. We will perform several analyses on this graph, including motif detection, link prediction, and community detection. Our goal in conducting our network analyses is to quantify the types of friendships observed in the graph and analyze whether the graph we constructed accurately depicts a social network with strong communities and triadic closure.

2 Related Work

Some of our work builds off of and incorporates findings from two papers in network analysis. The first paper analyzes a network the authors created from mobile communication logs, and the second paper analyzes a social network of high school students.

2.1 Mobile Communication Network

In "Structure and Tie Strength in Mobile Communication Networks", Onnela et al. use mobile phone logs to construct and analyze the structure of social networks. They found that social networks are resilient to the removal of strong ties but that they disintegrate with the removal of weak ties, which stood in opposition to the widely held belief that strong ties were responsible for maintaining network structure. We will use their observations as guidance for the expected structure of social networks. We will determine if the Stanford Memes network remains intact when removing strong ties and collapses upon the removal of weak ties. By comparing our findings to Onnela et al.'s, we will determine if the Stanford Memes group can be used to construct a social network that is representative of the real life social network.

2.2 High School Social Network

In "Suicide and Friendships Among American Adolescents", Bearman and Moody analyze a social network of high school students to understand social risk factors for suicidal ideation and suicide. The suicide rate among adolescents has increased dramatically in recent years. For adolescents between fifteen and twenty four years old, suicide is the third leading cause of death. Bearman et al. found that, "The friendship environment affects suicidality for both boys and girls. Female adolescents suicidal thoughts are significantly increased by social isolation and friendship patterns in which friends were not friends with each other."

The first factor, isolation, decreases a person's estimate of self-worth and reduces self-confidence, while the second factor, intransitive ties, leads to competing pressures from different groups and can ultimately contribute to suicidal thoughts and/or suicide. Suicidal ideation and suicide are major issues on college campuses, where twelve percent of students experience suicidal ideation at some point in their four years of college. As Bearman and Moody point out, isolation is a key risk factor for suicidal thoughts among men and women, and intransitive friendship patterns increase the risk of suicidal thoughts for females. The Stanford Memes network is an incomplete representation of the Stanford community, so we cannot conclude that students who are not tagged in posts are isolated. However, we will analyze the transitivity of friendships by analyzing the prevalent motifs of friendships for users in the network.

3 Data Collection Process

The data collection process proved more difficult than we expected. It used to be possible to use Facebook's Graph API to get data from Facebook groups. After we realized we could no longer use Facebook's API, we thought to crawl and scrape the group ourselves, but the legality is unclear and certain security measures prevented us from doing so. We were able to find a way to collect the data within these constraints by downloading and parsing HTML for each post. However, we were unable to collect all the data we wanted to because of the amount of time it takes to scroll and click on each post to expand all comments. We were able to automate the process of expanding comments for each post by using the Google Chrome bookmarklet "Expand All," which saved us from having to manually click the "View previous comments" button. However, we were not able to find a way to avoid manually scrolling through the page to retrieve older posts and clicking on each post, so our data is limited to posts from May 2018 to present. We used the Python package BeautifulSoup to parse the original poster name, caption, and comments for each post, and we constructed our network using NetworkX and SNAP libraries.

In addition to gathering data from the Facebook group, we collected users' majors and degree positions from the website Stanford Who. We used this data for detecting community structure within the network.

4 Algorithms

4.1 The Louvain Algorithm for Community Detection

We compared the Clauset-Newman-Moore algorithm with the Louvain algorithm for community detection and found that the Louvain algorithm performed better on our graph. The Louvain algorithm was first described by Blondel et. al in their paper "Fast unfolding of communities in large networks," where they use the algorithm to detect language communities in a Belgian phone network of 2.6 million users. The Louvain algorithm uses a heuristic based on modularity optimization, where modularity is a measure that compares the density of edges within and between communities, and is formally defined by the equation that follows.

$$Q = \frac{1}{2m} \sum_{1 \leq i,j \leq n} \left(\left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j) \right)$$

In the first phase of the Louvain algorithm, the algorithm greedily puts a node i into the community of its neighbor j that maximizes the change in modularity. Instead of computing the change in modularity by subtracting the old modularity

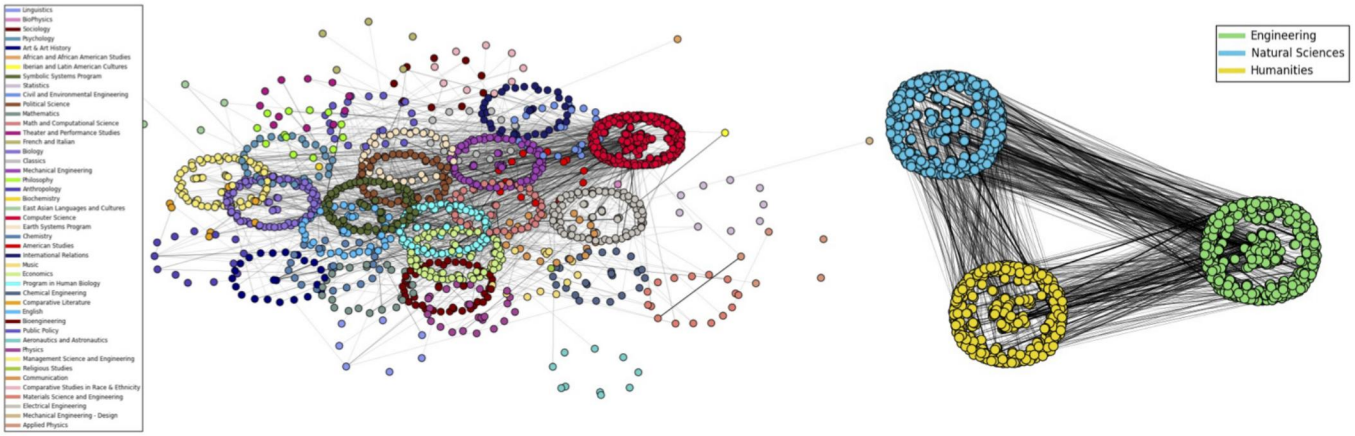

Fig. 1. Stanford Students Partitioned by Major(*left*) and Discipline(*right*)

from the new modularity, the Louvain algorithm uses the equation that follows to speed up the calculation of the change in modularity.

$$\Delta Q = \left[\frac{\Sigma_{in} + k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

4.2 The ESU Algorithm for Motif Detection

In order to both quantify and qualify the structure of relationships between users in the graph, we rely on motifs, which are defined as repetitive and significant patterns of connections between nodes. In order to count the frequency of motifs, we rely on the Exact Subgraph Enumeration (ESU) algorithm. The ESU algorithm enumerates all subgraphs of a given size and places them into non-isomorphic classes. We can then apply the ESU algorithm to a randomly rewired version of our graph, and compare the motif counts in our original graph to the randomly rewired graph to obtain a network significance profile. The equation for motif i's normalized score, Z_i, is given below, where $N(i)$ is the number of times motif i appears in the original graph and $N(i)_{sampled}$ is the mean number of times motif i occurred over the sampled null model.

$$Z_i = \frac{N^{(i)} - \bar{N}^{(i)}_{\text{sampled}}}{\text{std}(N^{(i)}_{\text{sampled}})}$$

Positive normalized scores indicate motifs that are overrepresented in the original network, meaning they occurred at a higher frequency than in the null model. Negative normalized scores represent motifs which are underrepresented in the original network, meaning they occur less frequently than they do in the randomly rewired graph. These scores offer a metric to quantify the types of relationships that are observed in our network.

4.3 Link Prediction via Proximity

Given a graph G at time t, the objective of link prediction is to predict new links in G at time $t+1$. We use link prediction via proximity, which computes the score of all pairs of nodes using some scoring function, sorts these scores in descending order, and predicts the top n pairs as new links. We use the Adamic/Adar scoring function, which is calculated using:

$$A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log|N(u)|}$$

where $N(x)$ is the neighbors of node x.

5 Results and Analysis

5.1 Preliminary Analysis

We constructed a weighted, directed graph where nodes are a subset of the members in the Facebook group and an edge exists from user A to user B if user A tags user B in a comment. The graph contains 8,042 nodes and 10,047 edges, and the sizes of the largest weakly and strongly connected components are in Figure 3.

The graph has an average clustering coefficient

of .052. The average clustering coefficient is low, which might be due to incomplete data or the fact that a person's friends may be friends in real life but they are not active in the group. From Figure 2 we see that the majority of users have a low degree, but there are some nodes with higher degrees of around ten.

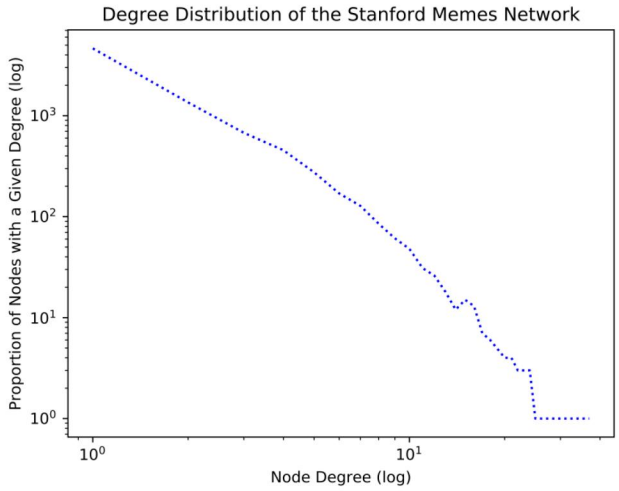

Fig. 2. Degree Distribution

Nodes	8042
Edges	10047
Nodes in Largest WCC	4480
Edges in Largest WCC	7621
Nodes in Largest SCC	801
Edges in Largest SCC	2042
Average Clustering Coefficient	0.052

Fig. 3. Graph Statistics

5.2 Community Detection

5.2.1 Partitioning on Academic Majors

We explored whether strong community structure exists within academic majors at Stanford. We used the data we scraped from the website Stanford Who to partition the users into communities based on their majors and then measured the modularity of this partition. Figure 1 shows the result of this partition, where each node in the graph is a user, colors correspond to majors, and the thickness of an edges is proportional to the tie strength. The graph in Figure 1 is a subgraph containing academic majors in the School of Engineering and School of Humanities and Sciences, but it does not contain students in the Stanford School of Law, Graduate School of Business, or the School of Medicine. The modularity of this partition is 0.16, so there is not significant community structure within majors. From the plot in Figure 1 we see that there are many edges leaving each community. If the network we constructed represents the Stanford community in real life, then Stanford students have more close friends outside of their major than within their major.

5.2.2 The Louvain and Clauset-Newman-Moore Algorithms

We used the Louvain and Clauset-Newman-Moore algorithms to detect communities in our network. A summary of the detection statistics can be found in Figure 6. Visualizations of the Louvain and Clauset-Newman-Moore detected communities can be found in Figures 4 and 5, respectively. These visualizations only display communities with more than 100 members for simplicity. Overall, the Louvain algorithm detected higher quality communities, as inferred from its modularity. Both algorithms detected a large number of communities with the majority of communities having few members. It is our understanding that users tend to tag their close friends in comments, so it makes sense for there to be many small communities within this network.

5.3 Link Prediction

We use link prediction via proximity with the Adamic/Adar scoring function. We split the data into two graphs: the first graph, $G_{0:t}$, has edges from posts created before time t and the second graph, G_{t+1}, has edges from posts created after time t until the present. First we created the core graph of $G_{0:t}$, where the induced subgraph contains only nodes with degree greater than two. We apply the scoring function to all pairs of nodes in the core, and we select the top n scoring pairs as new edges. This method predicted 19 correct links out of 1,000, so our link prediction model does not make accurate predictions. This may be due to the fact that people generally make new friendships as the school year progresses and their circles of friends cross, and we are making predictions over a shorter period of time that doesn't allow these relationships enough time to develop.

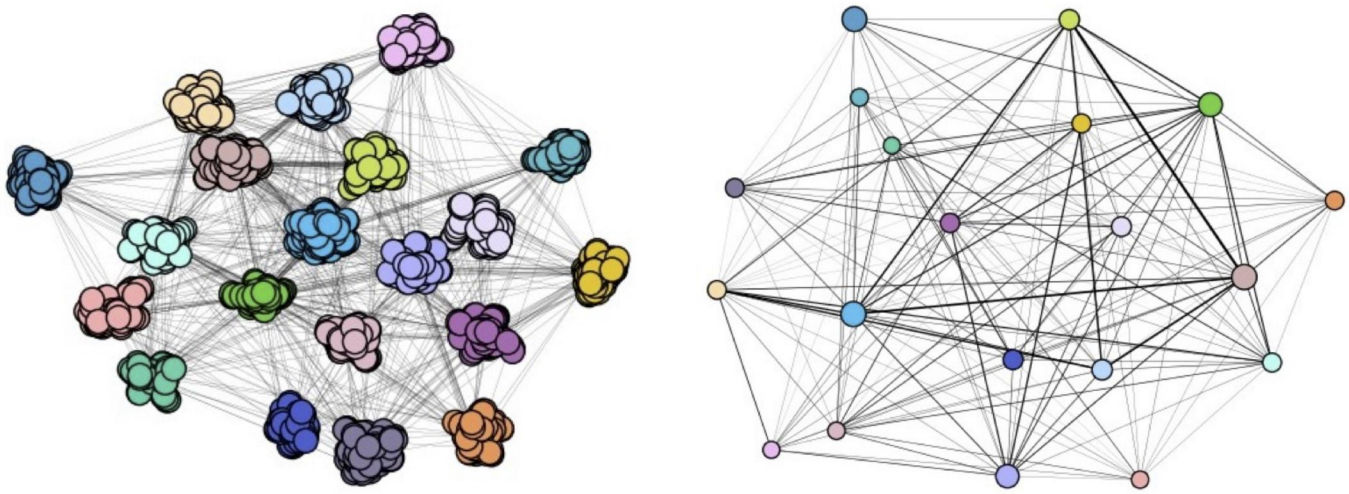

Fig. 4. Community Detection using **Louvain Algorithm**: Non-Contracted (*left*) and Contracted Supernodes (*right*)

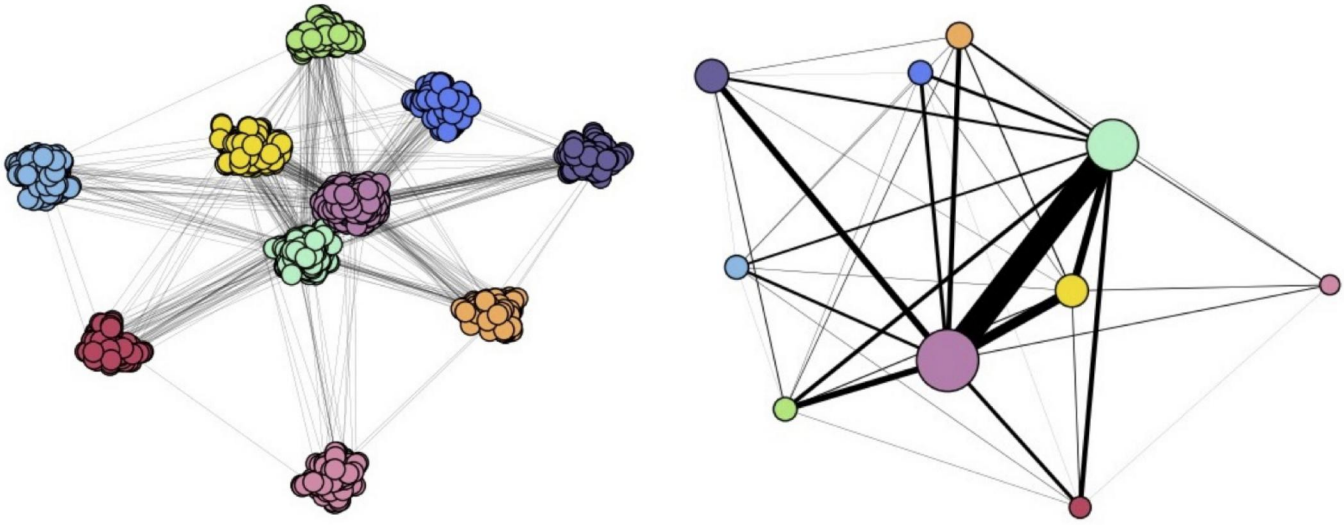

Fig. 5. Community Detection using **Clauset-Newman-Moore Algorithm**: Non-Contracted (*left*) and Contracted Supernodes (*right*)

Partition	Number of Communities	Size of Largest Community	Modularity
Academic Majors	45	464	0.16
Louvain Algorithm	1375	183	0.88
Clauset-Newman-Moore Algorithm	1397	807	0.78

Fig. 6. Community Detection Statistics

5.4 Motif Detection

We used the ESU algorithm to count all non-isomorphic directed 3-subgraphs in the original graph and compared the counts to those found in the randomly-rewired model. There are thirteen non-isomorphic directed 3-subgraphs, as seen in Figure 7. From the network significance profile in Figure 8, we see that motifs 3, 5, 6, 7, 8, 10, 11 are overrepresented in the graph compared to the null model. Motif eight is the most overrepresented subgraph, and it corresponds to a subgraph where there are bidirectional edges between node A and node B and between node A and node C, but no edge between node B and node C. This subgraph corresponds to a person being close friends with two people who are not friends with each other.

This motif is especially interesting because, as we discussed earlier, Bearman and Moody found in their analysis of a high school social network that suicidal thoughts among female adolescents are significantly increased by friendship patterns in which friends are not friends with each other. These intransitive friendship patterns are a risk

factor for suicidal ideation, which is a major issue on college campuses. Our graph is an incomplete representation of the real Stanford social network, so we cannot conclude that these friendships are truly intransitive, but this over represented motif is an interesting aspect of our network.

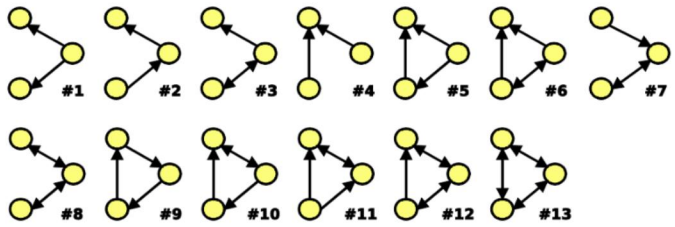

Fig. 7. All Possible Non-Isomorphic Directed 3-Subgraphs

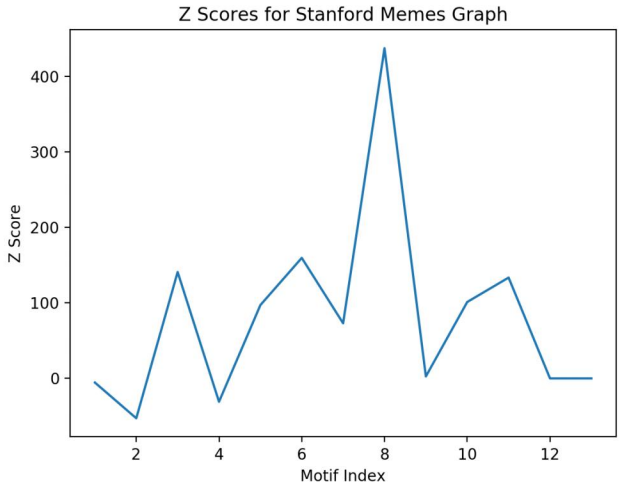

Fig. 8. Network Significance Profile on 3-Node Motifs

6 The Effects of Strong and Weak Ties on Network Robustness

Onnela et al. found in their analysis of mobile communication logs that social networks are resilient to the removal of strong ties but that they quickly collapse upon the removal of weak ties that connect communities. In Figure 9, we see that removing weak ties disconnects the network sooner than when we remove strong ties.

7 Conclusion

We were able to build a relatively large graph given the fact that the data collection process

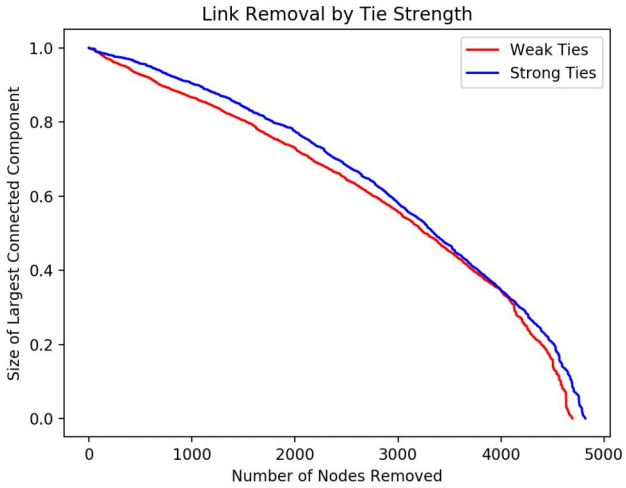

Fig. 9. The Effects of Removing Strong and Weak Ties on Network Robustness

involved a significant amount of manual work. Although the graph that we constructed is large, it does not have significantly more edges than it does nodes. This could be due to the fact that users only tag a small subset of there real life friends, or it could be due to the limited data.

We did not find significant community structure within majors, which may be due to limited data or the fact that people often have friends from diverse backgrounds and different majors. The community detection algorithms were able to partition the network into a set of communities with high modularity. Further work could be done to detect if these communities are meaningful by collecting data about student group involvement, dorm life, and academic year in combination with our academic major data.

The precision of our link prediction model was very low. We suspect that this could be due to the time frame from which we collected data. It is often the case that students live in new residents, join new clubs, and try new things at the beginning of the academic year. We hypothesize that it is less likely for friendship to develop from spring quarter through summer and into the fall. We believe that using a graph derived from data from autumn and winter quarter would better be able to predict links made in the spring.

We found that motif eight in Figure 7 was significantly overrepresented in our network. This finding is interesting because Bearman and Moody found that intransitive ties are a risk factor for suicidal ideation, which is an issue on all college campuses.

8 Future Work

Unfortunately we were not able to obtain the complete two year history of the Stanford Memes Group because the manual process of getting the data was time intensive. Future work includes getting more data and going over our analyses again. We expect that more data would lead to larger and stronger communities and more accurate link prediction.

Code Repository:
https://github.com/yardenahirsch/cs224w_final_project

References

[1] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi, *Structure and tie strengths in mobile communication networks*, PNAS, 2007.

[2] Peter S. Bearman and James Moody, *Suicide and Friendships Among American Adolescents*, American Journal of Public Health, 2004.

[3] Leskovec, J. (2018). CS224W: Analysis of Networks, Homework 2.

[4] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). *Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008.