Group Evolution: CS224W Project Report

Christie Brandt

December 11, 2011

1 Introduction

Online social networks provide new opportunities to gain a better understanding of human interaction and evolution of communities. Understanding and predicting community behavior based on structure and activity patterns could be useful both for members of the social network, advertisers who wish to predict influence, and, perhaps most of all, community builders, who can attempt to alter community structure and interaction patterns to develop a successful community. Although a great deal of research has been performed to analyze communities in networks, most has focused around community detection and analysis, and work on understanding community evolution has tended to utilize explicit "friendship" or "membership links" and treat such links as permanent once created [19, 15, 16, 9, 7, 12]. However, in the domains of sociology and psychology, it is hypothesized that community size and growth are not necessarily predictors of community health.[7, 12] In addition, in networks such as social interaction networks, where link maintenance requires time or other investments, it is important to treat links as ephemeral and examine shifts in network behavior.[20]

The goal of my project was to gain a better understanding of the interplay between activity and growth, and to try to learn a method of predicting which groups will be successful in the future. I performed an empirical analysis of the social networks from the Ning platform [1] to determine what mechanisms underly group growth and evolution.

2 Related Work

Previous empirical analysis on community evolution and health in social networks differs significantly from common wisdom in sociological analysis. Since much of the analysis of communities in social networks has focused on detection, such communities are often examined in the context of static networks, and even research done on community evolution tends to treat edges as permanent once they are placed in the network[19, 15, 16, 9]. Because of this, community growth has been used in much of the empirical literature as equivalent to community health[7, 12]. However, from the perspective of sociology and psychology, community health and community growth are distinct and not necessarily related properties. From a sociological standpoint, member communication and activity are stronger indications of success.[11] Yet very little research has been done to analyze communities with transient links.

Backstrom et al.[5] analyze groups and networks as evolving structures, but treat links once formed as permanent. They also focus mainly on growth via diffusion: how agents on the "fringe" of a group, who have friends within the group, decide whether or not to join. They find that the probability of a fringe element joining the group increases sublinearly with the number of friends in the group, and the triadic closure of these friends, but the probability of growth of a community decreases with increased triadic closure. Kairam et al. [13] further analyze these properties in the Ning network and find that less than half of new memberships can be explained via diffusion growth. They characterize the remaining memberships as *non-diffusion growth*, where individuals with no previous connections to the group join, presumably because of common interests. They find that while increased in-group transitivity does increase diffusion growth, mutual ties within the group decrease the probability of non-diffusive growth, and groups which rely mainly on diffusive growth tend to be smaller overall. Kairam et al. also analyze growth rate and longevity, where longevity is defined as the amount of time before growth ceases.

Raban et al. [18] analyze transient communities in online chat channel datasets and find that high turnover is correlated with increased longevity, but note that small, homogenous groups tended to be individuals separating from a main chat to have brief private discussions. It is not clear if the correlation between high turnover and longevity observed in chat channel groups will also be seen in other social networks, where the cost of maintaining links is significantly less. The authors utilized hazard models to analyze risk factors for community failure, but did not develop an overall model for community evolution. Palla et al. [17] examine transient links in a communication and a collaboration network. They define communities using the Clique Percolation Method(CPM) and define stationarity in terms of the changes of

group membership between timesteps, $\zeta = \frac{\sum_{t=t_0}^{t_{max}-1} C(t,t+1)}{t_{max}-t_0-1}$. One potential issue with this measurement is that it does not distinguish between fluctuations and fundamental shifts in membership because an individual who contributes infrequently is treated as multiple instances of leaving and joining a group. They find that size and age of communities are positively correlated, but that while high turnover in large communities tended to indicate longevity, small communities with high turnover tended to die out quickly. Since high turnover may simply indicate a large proportion of members with varying involvement, it is difficult to determine if fundamental shifts in membership are correlated with community success. They also find stronger interaction weights within communities than outside of them; however, this may be a property of how the communities were defined.

The research of Viswanath et al. [20] relates to this research because it examines activity; however, it examines activity only over the friendship network and does not examine the affiliation network. The authors compare the structure of the friendship and activity networks in Facebook empirically. They hypothesize that the cost of actively maintaining links will result in a qualitatively different network than a membership network. They compare the social network, where an undirected edge is created between each pair of friends, and the activity network, a graph with an undirected edge between each user pair which interacted at least once during the interval the network was crawled. They find that although the activity network itself undergoes rapid evolution and fluctuations, overall macroscopic properties of the graph such as clustering coefficient, path length, and node degree remain effectively constant.

An important aspect of my proposed research is the development of a model to explain observed phenomena and make predictions about group longevity. Group evolution can be seen as a form of contagion. When links are permanent, it can be characterized as SIR contagion. Activity networks can be viewed as a form of SIS contagion. Linear and voter models are commonly used to model contagion, but could be adapted to describe group evolution.[6, 8, 14] However, since empirical evidence indicates that group formation does not follow a linear threshold model [5], graphical models like those described in [10] and [21] may be more applicable. In [10], the authors utilize a simple markv model over a weighted directed network. At each timestep, an infected agent has λ trials to transmit the disease to its neighbors with probability β per time unit. This is a Markov chain where the stationary state describes disease prevalence. [21] utilize activity to construct a latent variable model where relationship strength acts as a latent variable which affects the likelihood of activity between user pairs. However, this assumes that the latent relationship strength variable remains fixed over time, while in fact user interactions and group memberships can evolve over time.

3 Materials and Methods

As my research focused on analyzing activity in social networks, it was important to utilize a dataset with sufficiently rich information about friendship and membership links. I performed some initial analysis on the Livejournal dataset; however, I focused on the Ning social network dataset.[1] Ning provides a common web platform for organizations to build online communities with features including in-community group creation, picture sharing, messageboards, member "friending", and more. There are several million communities in the Ning dataset, ranging from only a few members to hundreds of thousands. The Ning datset is extremely rich. It consists of dozens of directories, each consisting of separate types of interactions. Each file contains the interactions from multiple communities. Each line contains various metadata, descriptions, a timestamp of the interaction, and an author. It also sometimes contains a group name or another "destination" individual.

Transforming the data into a manageable representation of the network proved unexpectedly difficult. The data is simultaneously extremely rich and incomplete. I started by breaking the activity types into three different sets: dyadic edge activity, group activity, and node ego-activity. Edge activity comprises any type of activity between two edge types–comments between two users, sharing pictures, and more. I represented these as directed edges, and used them to indicate activity (and friendship) between two nodes. I utilized the portion of the dataset that defines interactions between individuals and groups or between pairs of individuals. This data can be transformed into two networks: a friendship multigraph network, where each node represents a person and each directed edge is tagged with metadata including the time the edge was created and the event that caused it, and a bipartite group-individual network, where multiedges are again tagged with metadata describing the time and type of the edge.

In my coding and analysis, I utilized the Snap [2], Matplotlib [3], and Orange [4] libraries.

My initial goal was to perform empirical analysis and use this to develop a probabilistic graphical model; however, the empirical experiments proved unexpectedly complex and I decided to concentrate on it.

Almost no empirical evaluations have been done on community structures in activity networks. It was therefore extremely important to develop a clearer understanding of how the activity network evolved. The Ning dataset provides a unique opportunity to analyze how network size, activity, and age, as well as other features, impact group growth. However, the richness and large number of interacting parameters make it difficult to determine causation. I therefore began by examining group health and evolution in the simpler Livejournal network, as well as understanding various

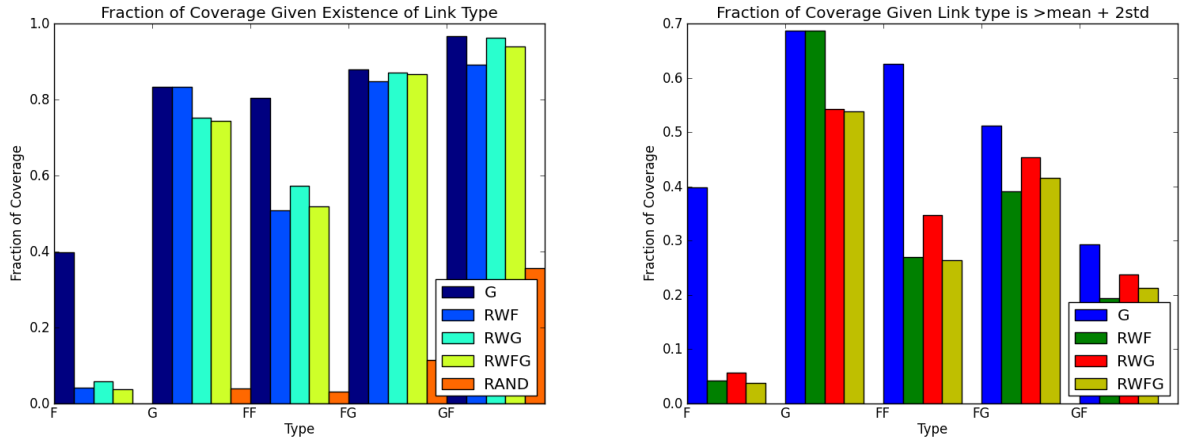
factors and parameters in the Ning networks. I hypothesized that links in the network are not formed at random, and could instead be "explained" by short paths between the two nodes of the link. I also hypothesized that differences in link and path structures were indicative of overall group health.

4 Understanding Group Formation: Livejournal Case Study

Due to the richness and complexity of the Ning dataset, I started out by performing several experiments on the Livejournal dataset. The Livejournal dataset includes several snapshots of both the friendship and group networks, but does not include activity information. Previous work showed that diffusion edges can account for 40% or more of new group memberships, but hypothesized that the remaining 60% was effectively uniformly at random because the membership edges are due to common interest rather than common relationships. However, I hypothesized that both exposure and common interests are at least partially encoded in the network structure. If this is the case, then we would expect a new member, n, of a group g to have short paths through the friendship and membership networks. In my initial investigation, I found this to be the case.

In initial investigations, I found that although the $90-$percentile effective diameter of the Livejournal network is 6.5, approximately 95% of new group membership edges are created between node-group pairs with F,$F-F$, or G paths. For the sake of comparison, I also examined $G-F$ and $F-G$ paths.

The following provides fractional coverage results over the various link types, compared to various baselines. RWF indicates rewiring of the friendship network only; RWG indicates rewiring the group network, and RWFG indicates rewiring of both networks.



(a) Plot of coverage, given existence of the various types of links compared across different rewiring types.

(b) Plot of coverage, where an element is counted as 1 (rather than 0) if it is $\geq \mu + 2\sigma$.

Figure 1: Coverage plots

These two plots demonstrate that the mere existence of a link is actually not a particularly strong signal, except in the diffusion edges. The FF links (links formed through two-friendship paths) also provided stronger signals than the group links. This may indicate that the signal is disrupted by very high-degree nodes and groups. Although nodes may be connected to them, they do not provide strong signals. Another possibility is that such edges are simply common. To test this, I plotted the fraction of coverage given that the number of paths is significant: at least two times the standard deviation above the mean. This provided a stronger signal, but was still quite noisy.

One issue brought up by the plots above is the overlap in coverage types. I examined this by plotting the overlap types. Since there are 5! potential combinations, I did plotted this as a pie chart, not including any portion which accounted for less than 1% of the types.

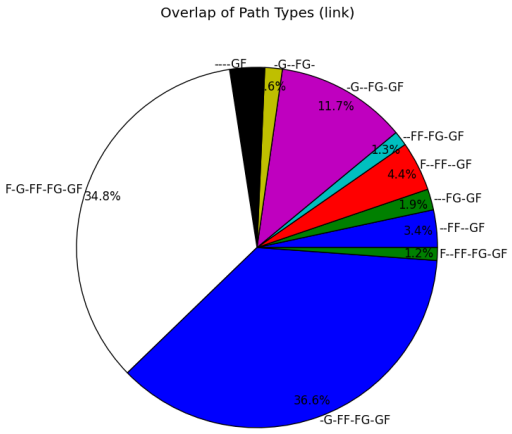


Figure 2: Overlap of path types

Interestingly, node that nonexistence of paths is less than 1%. Over 70% of new edges had all types of links except, possibly, the diffusion edges. All but 5% of the remainder had at least one of G,FF, or F edges. We can also examine the overlap from the perspective of each path type.

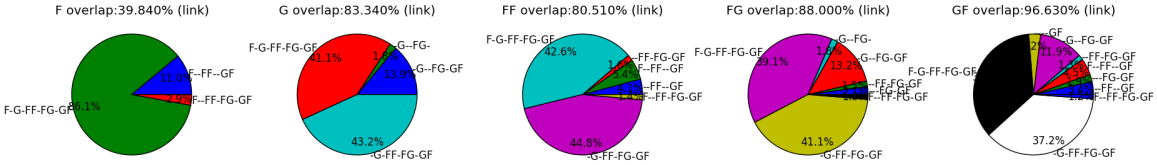


Figure 3: Overlap in G

From this, it is clear that there is very heavy overlap between G and FF paths.

I next tried to find a trend between path counts and degree. However, I found that there was no pattern between the degree and the existence of paths in the graph. After these initial experiments, I decided to determine if there were differences in network behavior in the activity graph versus the membership graph. I therefore moved on to examine the Ning network.

5 The Ning Social Networks: Macroscopic Properties

The Ning dataset provides millions of networks. It is important to determine how properties of the Ning networks change with respect to network size, what remains constant over the networks, and how the data can be used constructively.

The most obvious properties to examine are the node and group distributions of the networks. Many Ning networks are extremely small. With such small groups, path length becomes difficult to analyze, group membership is not particularly significant, and connectivity becomes very sparse. To ameliorate these effects, I constricted analysis to networks of size 100 nodes and up, where each included network also utilized the group feature to explicitly construct subcommunities. The distribution of networks by size is given below.

Min size	Max size	Number of networks
100	299	5741
300	499	1311
500	999	1076
1000	4999	841
5000		100

I first examined the relationship between the number of groups and the number of nodes in the network. I had hypothesized a linear relationship; the relationship appears to be either linear or slightly sublinear.

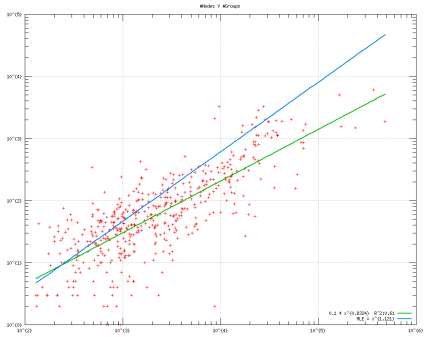


Figure 4: Plot of number of nodes (x-axis) vs. number of groups (y-axis)

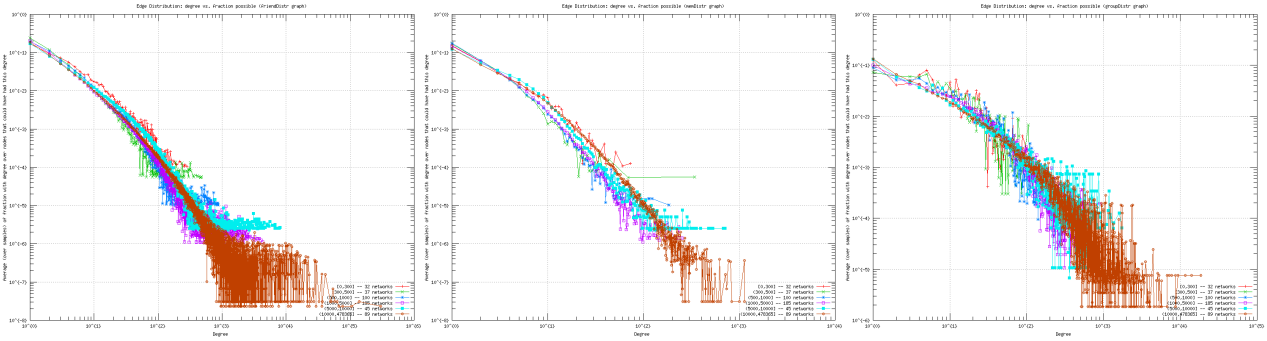
This is a useful observation, because it indicates that groups can be assumed to be approximately proportional to the number of nodes in the network.

I decided to sample the networks to get at most 100 of each type according to this breakdown. Considering this set of smaller networks as a "supernetwork" made up of separate connected components, the network contained 2088288 nodes and 64614 groups. Treating edges as directed, I found, on average, that nodes had degree 7.39646 when degree was considered in terms of unique edges. When degree was considered in terms of activity edges (an edge can occur multiple times at different times) the degree was 15.85073. In the membership graph, each node had on average 1.14875 unique membership edges and 1.64180 activity edges. Each group had on average 37.12708 members and an activity of 53.06218.

5.1 Effect of Network Size on Degree Distribution

It was important to understand how changes in network size affected the degree distribution. This could be examined from several perspectives, but most importantly, from the perspective of a member of the friendship graph, in the membership graph from the point of view of the nodes, and in the membership graph from the point of view of the groups.

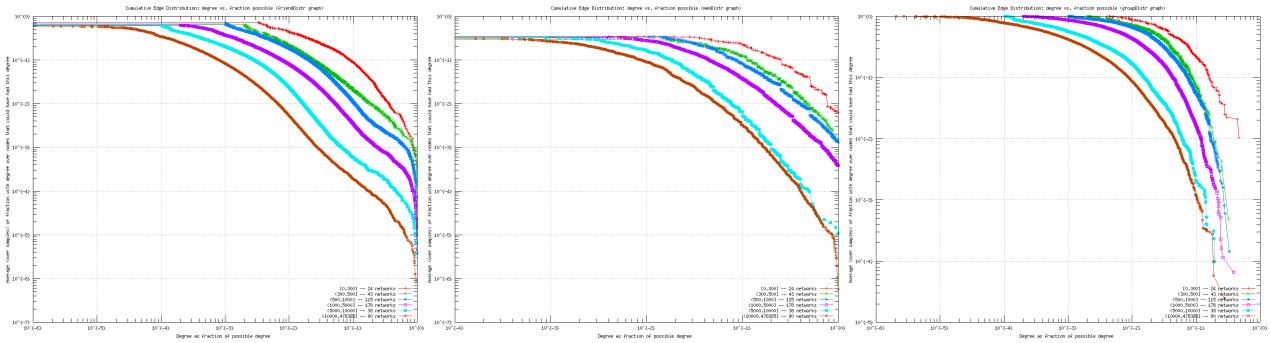
These are given below. So that comparisons can be made across the different network sizes, the x-axis represents the degree as the fraction of the maximum possible, and the y-axis represents the average, over the samples, of the fraction which could have had this degree.



(a) Degree distribution in friendship graph (b) Degree distriution in group membership graph, from the perspective of the nodes (c) Degree distriution in group membership graph, from the perspective of the groups

Figure 5: Degree distributions in graphs, split into bins by network size (red: 100 to 300 nodes; green: 300 to 500 nodes; blue: 500 to 1000 nodes; purple: 1000 to 5000 nodes; aqua: 5000 to 10000 nodes; brown: over 10000 nodes)

By these graphs, it is clear that the edge distributions approximately follow a power law, although it is slightly faster in the group graph. Importantly, all of the degree distributions, over the various bin sizes, appear to follow very similar distributions; there is no startling change between the distributions. We can also examine the cumulative degree distribution:



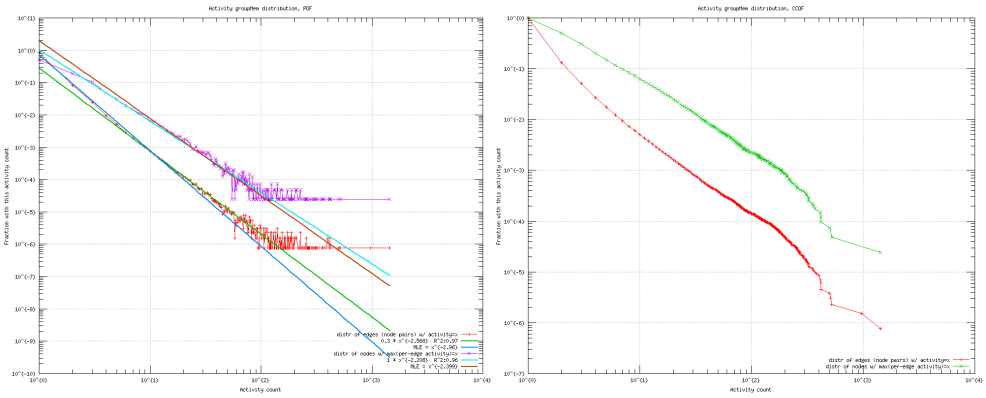
(a) friends graph (b) membership graph(node point of view) (c) membership graph(group point of view)

Figure 6: CCDF of edge distribution, loglog plot, split into bins by network size

In this graph, the difference between the group graph (from the group point of view) and the other two graphs becomes more apparent: the CCDF is not linear in the loglog plot. The tail is much shorter in the membership graph.

5.2 Understanding multiple interactions between nodes

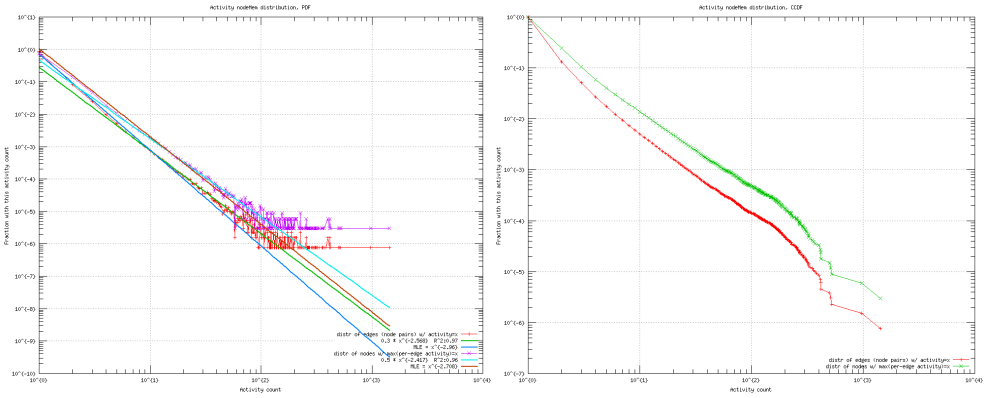
Given the fact that the difference between the average node activity and the average node degree in the group graph were extremely similar, I first more closely examined the distribution of activity edges. If, for example, activity edges were distributed with an extremely small tail, using unique edges would be sufficient in examining the graph. However, I found that this was not the case. I plotted the activity distributions of the friendship graph and membership graphs to understand how multiedges are distributed. For example, in the case of groups (the first graph shown), for each group g, I examined each of its member nodes n. Since the Ning dataset provided both group membership and group activity edges, the number of edges between n and g could be significantly higher than 1. The top line is the distribution of each such $n-g$ pair, where the y-axis is scaled by the total number of such pairs. This exhibits a long-tailed distribution. However, this information does not, on its own, indicate that many groups have multiedges rather than single interactions with their members. It could be the case that a small number of high-degree groups also exhibit high activity and contribute most of the tail, while the remainder of groups contribute to the single-edge interactions. To determine how many groups have at least one multiedge interaction, I plotted, for each group g, the maximum multiedge degree of its member set, then scaled this by the number of groups. This also exhibited a long-tailed, power law distribution. I found that the tail of the distribution for multiple interactions between groups and nodes was indeed long, and that the in general it appeared to follow a power law distribution. For example, from the CCDF plot, approximately 50% of the groups have at least one multiedge interaction, and approximately 15% of interactions are multiedge.



(a) PDF, with powerlaw fit (b) CCDF

Figure 7: Distribution of activity edges in terms of groups. The red lines indicate the distribution of the activity of node-group pairs; the pink(left) and green(right) lines indicate the distribution of the maximal node-group activity of each group.

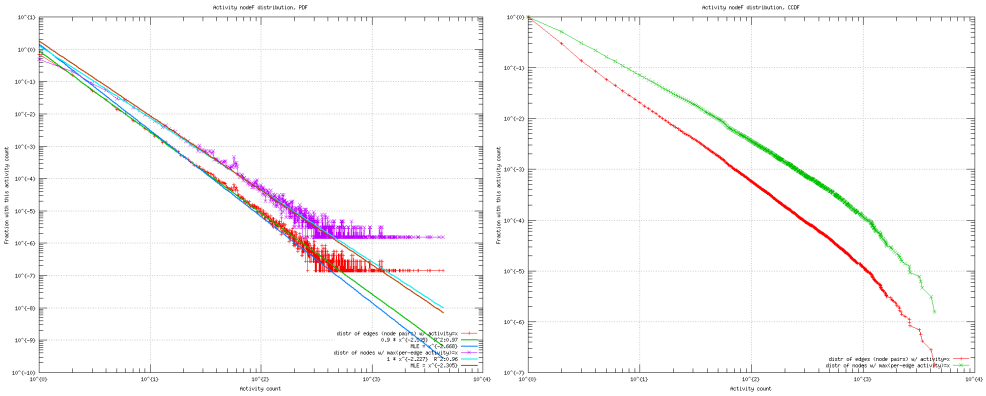
I examined the same properties for the group membership graph, but this time from the perspective of the nodes. The α for the powerlaw distribution of the pairwise activity distribution was approximately 2.3; the α for the powerlaw distribution of the maximal pairwise activity over each node was 2.9. Putting this into perspective, it shows (again) that approximately 15% of group-node interactions are multiedge interactions, and approximately 25% of nodes belonging to at least one group participate in multiple interactions with at least one of their groups.



(a) PDF, with powerlaw fit (b) CCDF

Figure 8: Distribution of activity edges in the membership network in terms of nodes. The red lines indicate the distribution of the activity of node-group pairs; the pink(left) and green (right) lines indicate the distribution of the maximal node-group activity of each node.

Last, we can examine the same distribution, but this time in terms of the friendship network. From this we see that over 50% of nodes participate in at least one multiedge interaction with a friend, and that approximately 30% of the interactions in the network are multiedge.



(a) PDF, with powerlaw fit (b) CCDF

Figure 9: Distribution of activity edges in the friendship network in terms of nodes. The red lines indicate the distribution of the activity of node-friend pairs; the pink(left) and green (right) lines indicate the distribution of the maximal node-friend activity of each node.

These results together indicate that the activity interactions are a significant part of the Ning network: they are relatively widespread, exhibit powerlaw distributions, and a large portion of nodes and groups participate in such interactions. I also examined the clustering coefficient and found it to be approximately 0.2244.

5.3 Understanding the node arrival processes

This portion of the analysis is important in undestanding how the graph evolves.

First, I examined node arrival times for each network size. I plotted arrival times, where a node was considered to have "arrived" when it participates in an edge creation event. These curves appeared reasonably similar across network sizes. The general shape of the arrival curve appeared to have a very low slope initially, with a sudden change to a steeper (but still sublinear) slope. This slope decreased as it approached the maximum. The changes in slope occur at approximately the same times in all network bins. This is reassuring, becuase it indicates that in examining arrival times, we can do so over different network sizes without needing to worry about scaling factors.

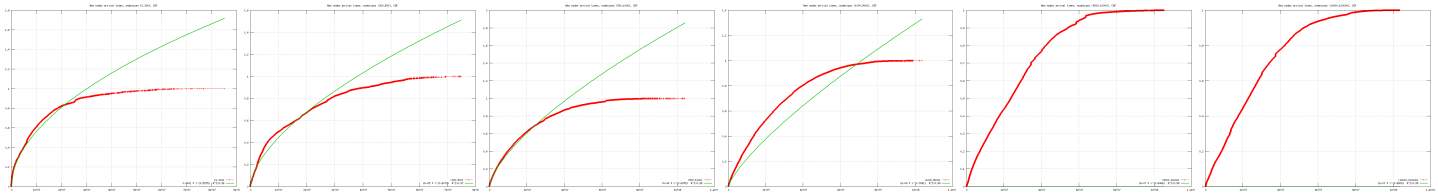


Figure 10: Node arrival distribution (CDF) over various network size bins

Note: there is an unfortunately misleading difference in scaling in these plots; without this, it is clear that node arrival times follow the same curves. We can also examine group arrival times in the same manner:

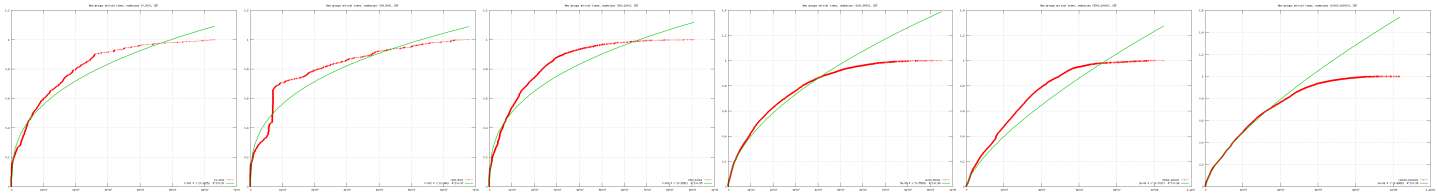


Figure 11: Group arrival distribution (CDF) over various network size bins

It is interesting to compare the arrival times of groups, not only against other groups, but also against nodes. Group arrival times are sublinear, and in small networks, there appears to be no delay in arrival times like that seen in the node arrival times (this can be seen more clearly in the loglog versions of the graphs).

In representing node arrival, typically we think of the node creating at least one edge. However, this may not be the case in the graph. There are, in fact, three ways a node can gain a first edge: it can create an edge in the friendship graph, it can create an edge in the membership graph, and it can have an edge created to it (while it still has degree 0). I had initially assumed that the nodes are activated by creating an edge, either to another node or to a group. However, upon examining the proportion of nodes activated by acting as destinations, I found that the percent of such nodes ranged from 18.6% (networks between 100 and 300) and an astounding 56.8% (in the largest networks, with size above 100000.) For networks between 300 and 10000, the fraction of "destination-initialized" nodes ranged from 26.4% and 37.2%, typically increasing with increased network size.

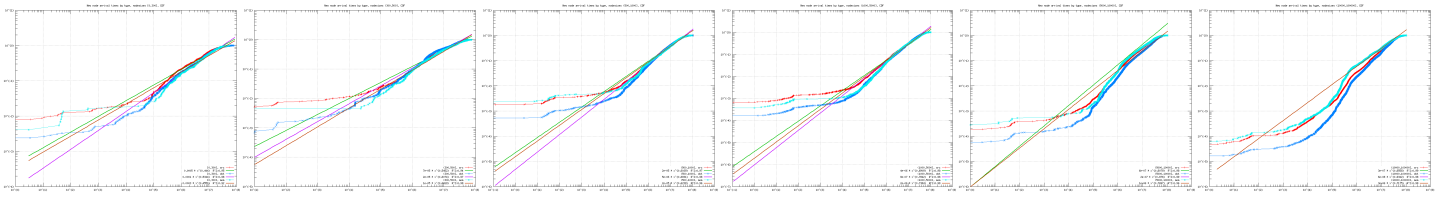


Figure 12: Group arrival distribution (CDF) over various network size bins

I also plotted node arrival times in terms of these three types: source, destination, and membership. (apologies again for the different axis scaling across graphs.) I found that the pattern of activity in terms of types was startlingly similar: each made up approximately 1/3 of the activity at each point in the plot. Interestingly, in small networks, source arrivals dominate early on; however, in larger networks, membership networks are dominant in those earlier arrival times. This again may suggest that groups play an increasingly important role in larger networks, and bears further investigation.

Node arrival time fails to capture the lifetime activity of the node, which is extremely important, especially in the activity graph. I examined the edge gap (the time between edge creation) δ_i, and plotted it for each δ_i, obtaining a separate histogram for each. I also plotted repeated edges (or "pure activity edges") and new edges ("unique edges" that actually change the graph structure) as well as their union. This was easiest to see in the form of the CDF.

Overall, the shape showed an initial sharp increase, then a "dead period", then another sharp increase which quickly tapers off. Interestingly, new nodes appear later than repeat edges in the delay times.

Although the general shape of the curves was similar, for smaller graphs, the effects of new edges decreased much more rapidly.

Since my goal was to understand activity networks rather than membership networks, I was most interested in the changes in edge relationships. However, this proved to be extremely hard to plot: averages are extremely noisy and there are simply too many edges to examine each individually. I therefore chose a selection of random edges $e = (n1, n2)$ and plotted their behavior over time. I then plotted the number of edges active for node $n1$ versus the number of edges e. I found that while there appeared to be an initial spike in mutual activity, it was then followed by periods of inactivity of edge e, which were sometimes followed by additional activity spikes.

5.4 Group connectivity and transitivity

One significant aspect of my analysis was to gain a greater understanding of community evolution. I first examined how group membership edges actually formed. In a fashion analogous to the Livejournal graph analysis, I examined the graph as each membership edge was added, and classified the edge according to whether it was a "new edge" (the first edge for either a node or a group), a "repeat edge" (an interaction between a group and node which have previously interacted), a "diffusion link", a "group superlink" or a "friend-friend superlink", where the last three define paths from the node to the group. I defined a *group superlinks* as follows: a node n and a group g have a group superlink if there exists a node n_i who is a member of g as well as another group, g_j, of which n is also a member. An *friend-friend superlink* exists between n and g if n has a friend n_i, who has another friend, n_j, who is a member of g.

I examined the same coverage statistics explored in the Livejournal experiments. From these experiments, I found that even in the large Livejournal graph, FG and GF were simply so long that they did not contain valuable information. I examined the diameter of the Ning supergraph, and found that on average, the diameter of the friendship graph was approximately 5.4 and the collapsed membership graph (constructed by considering each pair of nodes which shared membership in a group to have an edge between them) had effective diameter approximately 2.93. GF and FG paths are simply too close to this diameter to have value.

In this first plot, I also examined only newly created edges rather than multiedges seen in the activity network. I ignored all edges where the node or group was newly arrived or where the edge was already in the graph.

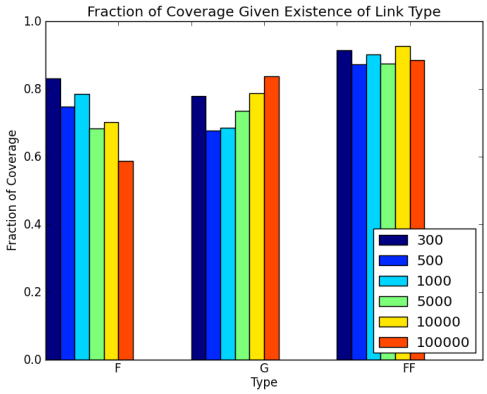


Figure 13: Fraction of coverage of newly added edges by various path types

This graph provides a potentially interesting observation: the fraction of coverage via F-paths decreases with increasing network size, but the fraction of coverage via $G-$paths increases with increasing network size. The FF-paths remain constant over varying degree sizes. There are at least two reasonable hypotheses, but they have extremely different consequences. It could indicate that G-paths play an increasingly important role in facilitating membership edges as networks become larger. Another potential explanation is that very large-degree groups and nodes act as "hubs" and that

paths become meaningless. Further experiments need to be performed to eliminate one of these explanations. Clearly, there must be rather high overlap between these types. I examined the overlap in each network size bin:

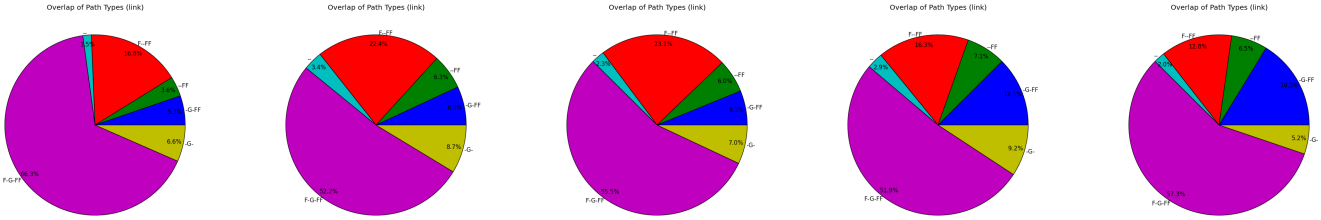


Figure 14: Overlap via link type, bin sizes (left to right)$[100, 300]$, $(300, 500]$, $(500, 1000]$, $(1000, 5000]$, $(5000, 10000]$

The colors over this plot remain constant. The aqua portion is the percent of new edges not accounted for by G,F, or FF links; it ranges from 1.5% and 2.9% of the edges. Approximately 55% of new links have all three types of paths; the fraction of $F - FF$ paths starts out around 23% and decreases to 12.8%. While G edges remain relatively constant, $G - FF$ paths increase from 5% to 16%. This again suggests that there may be a change in the role of G and F paths in different-sized networks, although more exploration is required. However, neither explanation, in itself implies that group paths–or any particular pattern of connectivity within the group–indicates future group activity.

6 Understanding Group Health and Evolution

One initial issue was that even the concept of group health was ill-defined. Group health should, obviously, be tied to a definition of group activity. As argued in the introduction, it is not clear that group growth provides a reasonable approximation of growth. However, as was shown in figure **??**, only about 15% of group edges were multiple edges. Therefore the explicit group edges were too sparse to consider group activity edges to be a true definition of activity. Group activity should capture a concept of active interaction between members. The initial intuition in using activity rather than growth was that a small group could have significant continuous interaction between members (as defined by the existence of multiedges between members) and yet not exhibit growth. This suggests another definition of group activity as the amount of interaction between two members of the group ("dyadic edges") rather than node group ("group edges"). Together, these ideas suggested four measurements of group activity over a given interval: *group growth* (the number of edges between a node and the group created over the interval when considering the network as a graph rather than a multigraph), *group activity* (the number of activity multiedges between a node and the group over the interval in the multigraph), *total dyadic activity* (the number of activity multiedges between node pairs over the interval), and *dyadic pair activity* (the number of unique node pairs which are active over the time interval). I examined all of these activity types, but with a particular preference for dyadic activity because it provided less sparse data than group growth and group activity. In addition, dyadic pair activity creates an inconsistency when considered over different time intervals. The reason for defining such a measure is to eliminate skews in the data caused by inequalities in activity between pairs. For example, total (directed) dyadic activity would consider a group g_1 of size 10 where each node pair had one activity edge to be equivalent to a group g_2 of size 10 where only two nodes were active and had 90 activity edges between them. Dyadic pair activity would consider the activity of g_1 to be 10 and the activity of g_2 to be 1. However, dyadic pair activity also exhibited an unfortunate inconsistency. If the time interval was split in half in the example above and the multiedge creation times were evenly distributed, the sum over these two intervals of the total dyadic activity would remain constant, but while the dyadic pair activity of g_1 would also remain constant, the sum of the dyadic pair activity of g_2 would become 2. As the time interval size goes to 0, the sum of the dyadic pair activity over these intervals converges to total dyadic activity. A third measurement, where only dyadic growth edges were considered, would be consistent, but would fail to capture sustained activity in small groups.

7 Analyzing activity and growth

I first examined different activity patterns over time. There were several activity types to consider: group growth, group activity (where growth edges were not counted), all group edges, dyadic activity between members, and dyadic activity where each member pair was considered only once. I examined the activity of each group for a period between one year

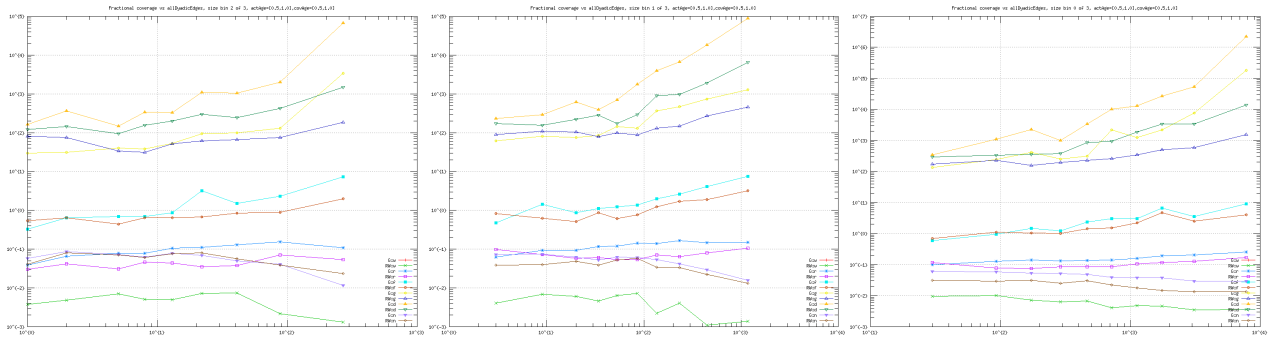
and one and two years after it was born (its first edge was created). For each pair of activity types, I used one activity type to split the groups into inactive and active (first and last quartiles, respectively). I then plotted the good and bad activities. Group activity edges were simply too sparse to be useful. Dyadic edges showed the most interesting pattern: "inactive" groups, by any metric and for any size, appeared to start out relatively high, but quickly dropped. Interestingly, when the good-bad split was done using group growth, the dyadic activity of the bad groups started out higher than that of the living groups, but very quickly dropped. This may suggest that the differences in dyadic edge activity may prove useful features for predicting growth. However, when good/bad split was done by dyadic edges, activity of "good" groups started out higher (although the slope observation remained true). Overall, dyadic activity was much more plentiful than the other measures. Group activity edges are sparse, and I believe the dyadic edges provide the best indication of activity.

7.1 Examining patterns between path values and coverage

In my first attempt to examine group health, I examined each network bin and split groups into "live" or "dead" by whether or not activity occurred in a last time interval in the graph. I then examined patterns of edge creation with respect to the different path types over time.

To examine the connections between activity and path connectivity, I binned the groups by size and plotted each group's activity versus its path value. I compared this value against a rewired null model. The use of a null model was important because I had previously found that activity (of any type, including dyadic) scaled approximately linearly with group size. Any trend I saw in increased path value could in fact be due to a confounding variable: the size. It is reasonable that both activity and path value increased with size, but it was unclear how to correct for this effect. Even with binning, the small differences in size might produce a misleading correlation between path values and activity. I therefore partitioned the groups into three bins, each with an equal number of groups inside, where size was taken at the end of the interval over which path values were sampled. The first bin contained groups (on average, since I sampled multiple times from the set of networks) of size between 10 and 19. The second contained groups between size 20 and 45, and the last contained groups of size between 45 and 1892.

In my first tests, I compared path values over an interval to activity over the same interval. I examined several different interval lengths and starting times to account for any irregular effects during the network's birth. Trends in the graphs were unclear. For example, the graphs given below give the sum of fractional path values compared to total dyadic activity. Both the rewired graph values and original graph values are shown. Note that although the increasing path values as activity increases appears to indicate a promising trend, the rewired graph follows a similar trend, throwing doubt on its value.



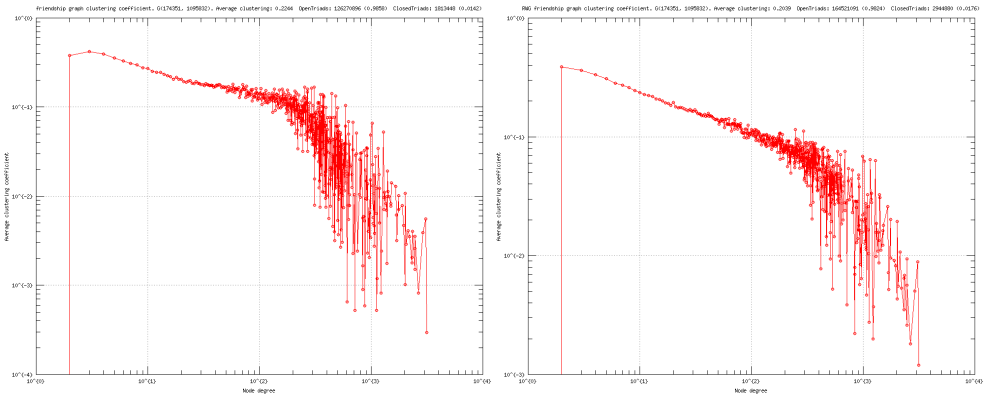
(a) Group sizes from 10 to 19 (b) Group sizes from 20 to 45 (c) Group sizes from 45 to 1892

Figure 15: Unweighted path counts plotted over total diadic activity (binned into 10 bins, median value of bin shown),measured over the interval from the groups' first to second year of life.

I performed a series of additional experiments similar to the Livejournal experiments, where I compared the graph against a rewired null model in which, for each network, the friendship graph and group graph were rewired. I also examined the connections between each activity type and path type, using the null model as a basis to determine if activity induced a change in path connections. However, I found that there was no significant difference in the behavior between the rewired and original graphs. Given this surprising result, I hypothesized that the small network size, particularly the small number of groups per network, effectively eliminated the usefulness of rewiring. For example, a network of size 100 or so might be expected to have 5 or so groups, with perhaps 20 nodes participating in group membership interactions. Rewiring would most likely be unable to decouple group-node relationships. To test this hypothesis, I examined properties of the original

and rewired networks.

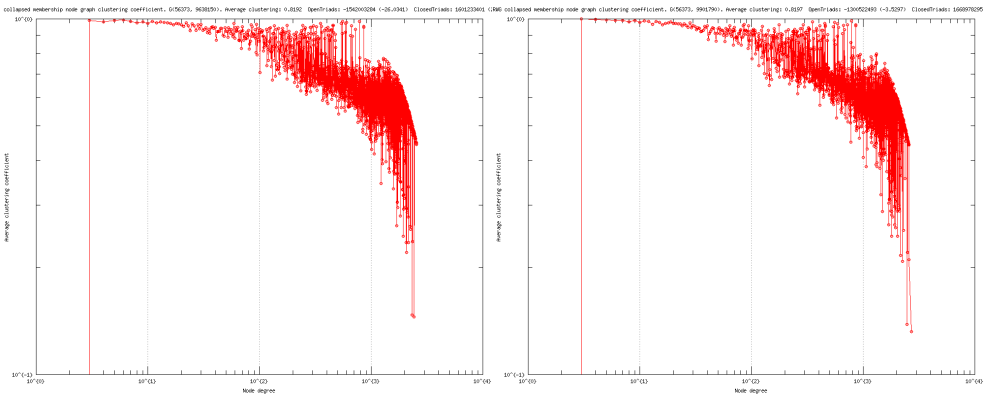
I first compared the clustering coefficients of the two graphs. There is a difference in the clustering coefficients, although it is small: 0.2244 versus 0.2039. Both are higher than a random graph of the same degree, but from this fact it appears that much of the triadic closure in the original graph remains in the rewired version.



(a) original friendship graph (b) rewired friendship graph

Figure 16: Comparison of clustering coefficients

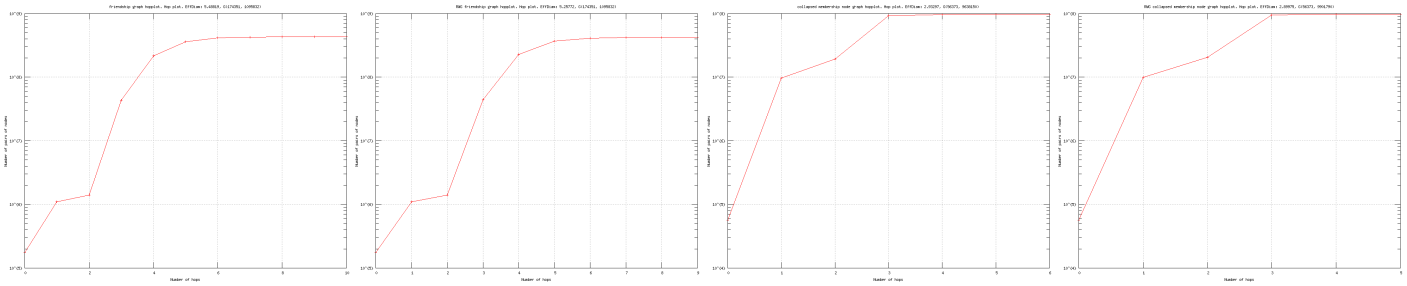
More problematic was a comparison of the clustering coefficients in the rewired membership graph. I collapsed the graph by considering each shared group membership to be an edge between nodes. Note that the clustering coefficient distributions are effectively identical, and differ by a hundredth. This meant that any pair of nodes sharing group membership in the original was likely to share membership in the new graph as well, thereby making the rewiring process ineffective in providing a null model.



(a) original (collapsed) membership graph (b) rewired (collapsed) membership graph

Figure 17: Comparison of clustering coefficients

As a last test to show that paths were effectively unaltered by rewiring, I compared the hop plots, both of the friendship networks and the collapsed membership networks.



(a) original friendship graph (b) rewired friendship graph (c) original membership graph (d) rewired membership graph

Figure 18: Comparison of hop plots

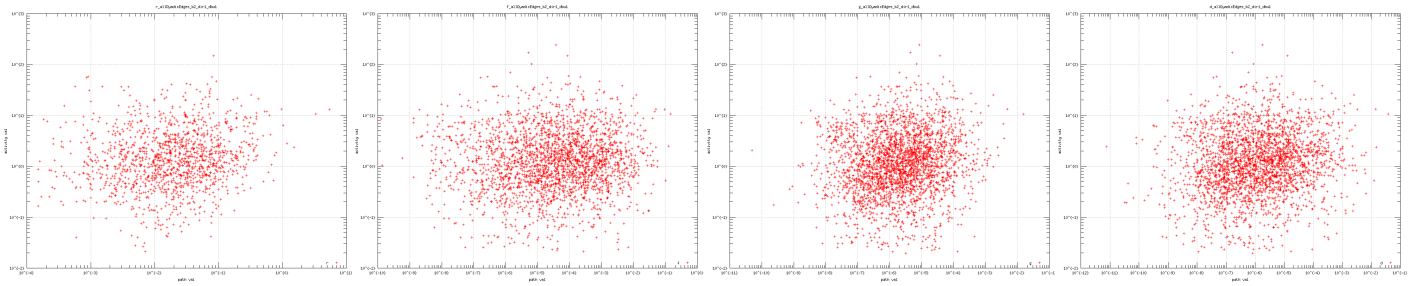
Together, the analysis from the hop plot and clustering coefficient analysis indicated that rewiring was an ineffective null model. Because it had to maintain correlations for each node in both the friend and membership graph, and the membership graph often had so few nodes that rewiring was ineffective, rewiring did not break correlations present in the original graph. The tests also provided interesting information about the original graph. As was previously noted, the effective diameter of the original graph is quite small: almost all nodes are 4 or less hops away in the friendship graph, and 3 or less hops away in the membership graph.

7.2 Examining patterns between path values and future activity

Since the eventual goal was to predict future success of groups, it seemed reasonable to compare path activity of an early interval to activity in a potentially later interval. I examined several interval pairs; for example, coverage over the first year compared to activity over the first year, coverage in the yearlong interval started after the first year of life, and coverage measurements taken over the first or second year and compared to the second or third year, respectively.

Like activity, there were multiple potential definitions for path values, such as the existence of the path, the count of such paths, and the weighted count of such paths. In addition, the paths can be constructed either by looking at edges as directed or undirected, and by looking at each edge as a single connective entity or as a multiedge. I examined boolean paths, but since such paths exist in almost every new membership edge, this did not provide useful information. I also examined unweighted paths, but found no interesting results. I hypothesized that the measure's lack of robustness most likely destroyed the signal. This measure failed to distinguish between a connection to a group "hub" and a small focused group, yet one might imagine that the second was more meaningful. I therefore weighted paths by treating each edge as though it were independently created according to a power law and weighting by this fraction.

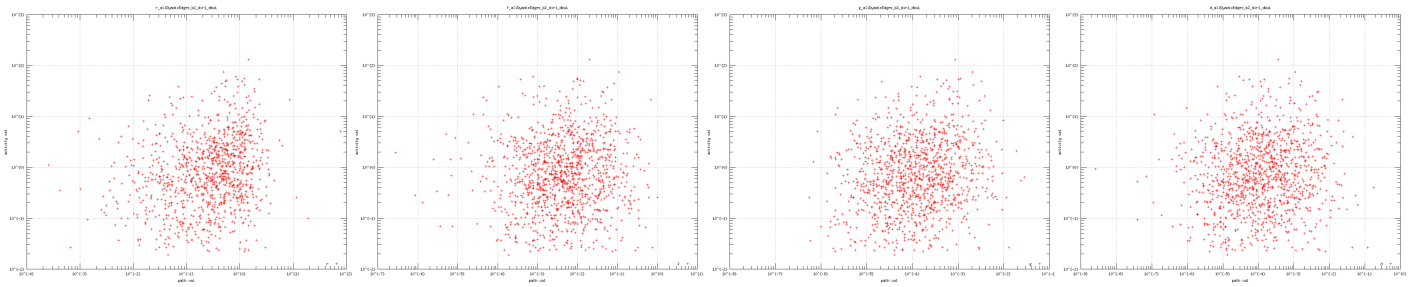
With the removal of a null graph for comparison, it became much harder to determine whether or not trends existed. Even with binning, size of the group most likely had a significant effect, both on activity and the number of paths. I examined how to represent activity so that it was relatively evenly distributed across different sizes. I hypothesized that since the dyadic activity is relative to the total number of edges in the group ($O(n^2)$), the best scaling factor would be on the order of $O(n^2)$. However, I found that scaling by n provided a better fit (although it is possible that an equation in $p(x^2)$ would provide a better fit, my data was too limited for fitting, and it was also unclear exactly how the data should be fit.) The following show weighted path values versus activity weighted by group size. I failed to find a clear trend in the results.



(a) repeat paths (b) F paths (c) G paths (d) FF paths

Figure 19: Scaled activity versus weighted path values, shown for bin with elements of sizes between 23 and 52 (mean and median both 34). Both activity and path values are computed over the yearlong interval starting after the groups' first half-year of life. Other plots, over various other intervals, appear similar.

I also attempted to find a pattern in the activity over an interval (1/3 of the total time) in terms of various factors, such as the number of nodes at the start of the interval and the number of edges at the start of the interval. I found that there did not appear to be any clear pattern; the distribution of edge activity and group activity appeared essentially random.



(a) repeat paths (b) F paths (c) G paths (d) FF paths

Figure 20: Scaled activity versus weighted path values, shown for bin with elements of sizes between 23 and 52 (mean and median both 34). Path values were computed over the groups' first years of life; activity was computed over the groups' second year of life. Other plots, over various other intervals, appear similar.

In my attempt to correct for the confounding size variable, I may have destroyed correlations that actually existed. I decided to move on to the prediction task to determine whether or not these network features encoded valuable information.

8 Experiment: Predicting Future Total Dyadic Activity

In this experiment, I treated the network activity problem as a prediction task. The goal of the task was to predict, given a set of features describing a group's initial interval of activity, its future success. Although accuracy of prediction is desirable, even more interesting is feature selection: the most significant features for this task. The maximum recorded time over the networks is 2.8 years. I therefore defined the evaluation task as follows: for each group over the sample of networks which lives at least 1.75 years, predict, using statistics taken from the group's first 0.75 years of life, the group's activity in the group's lifetime interval from 1.25 to 1.75 years. The group age was chosen so that a sufficient number of groups would be included, but still allow evaluation of activity and features over a sufficient amount of time to prevent extremely noisy data. The reason for the gap was to emphasize prediction of relatively long-term characteristics. I wanted to have at least a half-year gap between the provided information and the predicted information so that the features would be evaluated on whether they captured long-term trends in group growth rather than current group characteristics. For example, if evaluation of activity were done over the same time interval, or close to the time interval over which the features were taken, then the feature describing activity or growth would clearly be very significant. It is much more interesting to determine if activity early on indicates later activity. It is clearer to describe these times in terms of quarter-year intervals. Then the features were taken over the group's first three quarters, and activity was measured over the 5th and 6th quarters.

Initially, I took a sample of groups of various sizes and attempted to scale both features and activity. However, the scaling function for activity is itself not clear. Initially, I scaled activity by group size; however, it was unclear as to whether activity should be scaled by group size when the features were taken, or group size at the beginning or end of time over which the activity measurements were taken. For example a group that had significant growth over the activity period might end up being penalized for this. The "correct" way to handle this was nonobvious.

Instead, therefore, I changed the sampling method so that only groups of approximately the same size at the end of the interval over which features were taken (the first 0.75 years of the group's life) were sampled.

The distribution, over all the networks sampled, of groups at this age is given below.

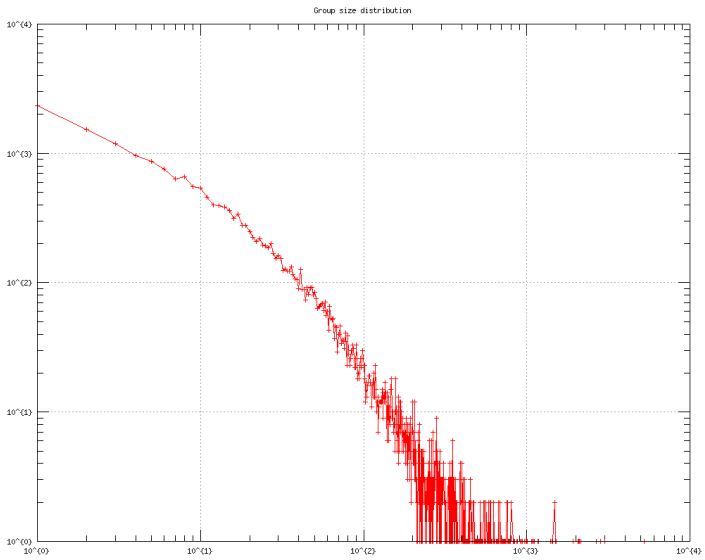


Figure 21: Group size distribution at age 0.75 years over network sample

To allow for a sufficient number of groups while still obtaining nontrivial features, I chose all groups which, in their 3rd quarter of life, of size between 10 and 12.

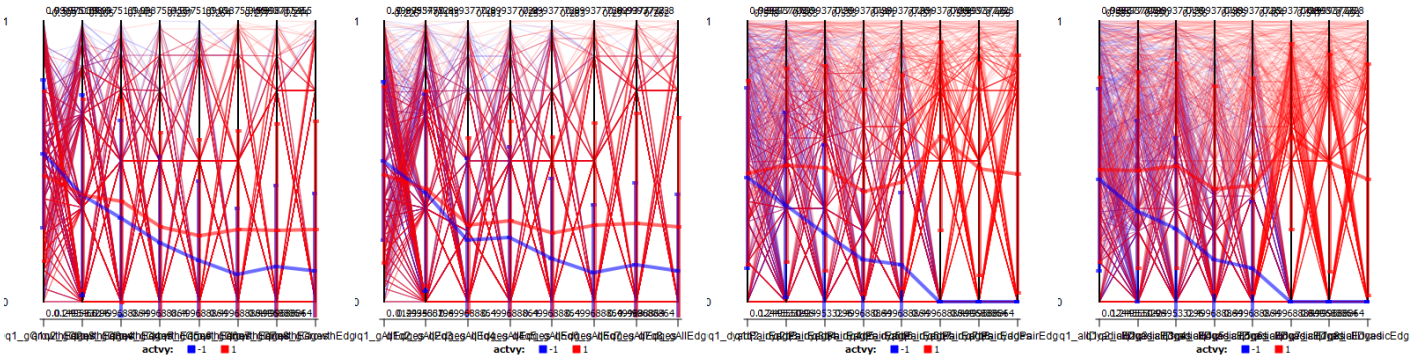
To minimize the design decisions involved in evaluating a continuous activity value, I instead decided to predict a dicrete binary-valued "activity" value. I binned the activities of the groups into three bins and considered the first bin (with the smallest activity values) to be "inactive" and the last to be "active". I used three bins rather than two to more strongly differentiate between active and inactive groups.

I considered the following features, taken at each quarter interval. Due to initial scaling issues, each feature was represented as its percentile value over the sample.

- Activity features
  - Group growth/size
  - All edge activity over the interval
  - Total dyadic activity over the interval
  - Dyad pair activity over the interval
- Network features
  - Network size
  - The number of edges in the friendship network between group members
  - The number of edges in the friendship network between a group member and an outsider
  - The modularity in the friendship network of the group
  - Average degree centrality of members
  - Average closeness centralities of members
  - Average clustering coefficients of members

- Path value features: I examined, at various points, all combinations of weighted/unweighted, directed/undirected, multiedge/unique connection paths, and including/excluding repeated edges in evaluation. I found that directed, unique connection, excluding repeated edges, tended to work best. Unweighted path counts also appeared to contain more information than weighted, although I included both as features.
- Differences (deltas) between current and previous interval values of the above features.

I was most interested in the behavior of individual features. Of the various growth measures, the total dyadic activity proved most indicative, which seemed reasonable, since activity in the later period was based on the same metric. One interesting trend seen in all activity types was that increased activity in the first interval actually had a slight correlation with later death. The graphs exhibiting this trend are shown below. The graphs provide the values for each quarter-year interval in the first two years of the groups' lives.



(a) Growth activity (b) All group activity (c) Dyad pair activity (d) Total dyadic activity

Figure 22: Parallel coordinates plot of various growth types. Each group, colored red if it is classified as active, blue if dead, is plotted as a line plot of its values over the the full set of intervals. The bars are provide means and deviations.

This trend has an obvious rationale in the case of growth. Since groups were chosen by their size after the third quarter of life, for a group to have significant growth in this quarter, the group would need to begin by growing slowly, and vice-versa. In other words, as an artifact of the group choice procedure, groups that are already "dying" by their third quarter of life must have exhibited more early growth.

My major interest was in how much the structural features (e.g. path values, clustering, modularity, etc.) impacted prediction accuracy. I therefore performed the prediction task. I used linear svms for prediction. The training set, constructed of groups of sizes between 10 and 12 in their third quarter of life, contained 1027 examples. Due to the way the data was constructed, the bin sizes were approximately equal: 514 positive and 513 negative examples. I utilized 20-fold cross validation for evaluation. I trained several classifiers: the first, C_{ALL}, contained all of the data types given above. The second, $C_{NET+ACT}$, utilized all features except for the path values (including network features). C_{PATH} contains only path features, with no activity or network features. C_{NET} uses only network features. C_{ACT}, only contained the activity features. Note that since the classes were approximately equal in probability, the majority baseline (random guessing) would produce classification accuracy of 0.5. A better baseline (C_{LAST} simply predicts using the last dyadic activity value given (that in the third quarter).) To evaluate the results, I examined classification accuracy (the fraction of correctly classified examples), precision (the fraction of examples which are actually positive out of those classified as positive), recall (the fraction of positive examples which are correctly classified as positive), AUC (area under the receiver-operator curve; equivalent to Wilcoxon rank test), and F1-measure (the weighted harmonic mean, equivalent to $\frac{2 \cdot precision \cdot recall}{precision + recall}$). The results were as follows:

Classifier	Accuracy	Precision	Recall	AUC	F-measure
C_{LAST}	0.6884	0.6824	0.7037	0.7557	0.6929
C_{ACT}	0.7204	0.7140	0.7349	0.7939	0.7243
C_{NET}	0.6806	0.6869	0.6628	0.7394	0.6746
C_{PATH}	0.6923	0.6958	0.6823	0.7496	0.6858
$C_{NET+ACT}$	0.7303	0.7296	0.7310	0.8065	0.7303
C_{ALL}	0.7615	0.7597	0.7626	0.8324	0.7612

The results from this were interesting. It was clear that simple activity information of the groups' early stages provided a significant amount of information about its later activity. The baseline of simply predicting using the last recorded dyadic

activity provided accuracy of 0.6884 and an $F - measure$ of 0.7243. However, the results also indicated that the network and path features did indeed provide useful information. Taking each set of features on its own, the classifiers were able to achieve accuracies of close to 70%.

In C_{ACT}, the most important features were, unsurprisingly, dyadic edge activity in the last quarter, but growth activity features also were highly weighted. In C_{NET}, most of the weight was placed on the number of edges leaving the group in the second and third quarters. Other additional weighted features were the edges entering the network, the size of the network, and the degree centrality. Interestingly, the clustering, modularity, and other more structural features did not have weight placed on them. In C_{PATH}, the heighest weights were placed on the weighted path values for G and FF, followed by the weighted path values for F-paths and the unweighted path values for FF. The next few features, and the only others to have considerable weight, were the delta values–the differences in path values from one interval to the next–of the G, F, and D paths. In $C_{NET+ACT}$, weight was placed on a mixture of the two feature sets described above. In C_{ALL}, the top two features, given approximately equal weight, were the number of edges leaving the group and the dyadic activity in the third quarter. The first 8 features were all NET and ACT features: edges out, the number of nodes in the network, and they dyadic edge counts. However, 9 out of the next 10 features were path features: weighted path counts for F, G, and FF, as well as the delta values for these paths, mainly for the third quarter. All other features with significant weight on them were a mixture of earlier quarter path values and other types of activity, such as growth. Modularity, the clustering coefficient, and similar features had essentially no weight placed on them. These results seemed to indicate that the activity by the third quarter provided a good indication of later activity, although structural information provided additional accuracy. It would be interesting to see if, stopping earlier or using short intervals, the weights of the features changed to put less weight on the growth features.

9 Discussion and Conclusions

Overall, my research sought to gain a better understanding of how groups evolve and grow. The goal was to understand the basic characteristics of activity networks and compare them to growth networks, and determine if network-based characteristics could be used to predict eventual group health.

I also began by analyzing a simpler dataset: the Livejournal social network. Previous work separated community growth into diffusive and non-diffusive growth, where diffusive growth was defined as a new link created between a node and a group where the node had friendship links with one or more members of the group. From my exploration, I found that non-diffusive growth was not a correct description: in fact, almost all (over 95%) of group memberships are created by individuals who are friends with a member of the group (diffusive growth), are in a group with a member of the group (G-paths), or are friends with a friend who is in the group. In other words, new links are created with short paths rather than at random.

I then moved on to empirically analyze the Ning dataset, a massive dataset with hundreds of thousands of individual social networks, each with internal community structures. My first task was to understand the structure of the networks and communities, and to understand the degree distributions, both in terms of activity and growth. I found that the realtionship between groups and nodes in a network was approximately linear, that the friend graph and membership graphs followed powerlaw distributions, and that activity, the degree of the multiedges between pairs of nodes, also followed powerlaw distributions. In addition, I found that nodes do not obey a standard "arrival model", where each node is initialized by creating an edge. Instead, up to 50% of nodes were initialized by contact from another person–effectively, initialized as a target rather than a source. I also examined the same path structures analyzed in the livejournal graph, and found that approximately 99% of membership links were formed via short links.

I then tried to understand and analyze group activity. I found that explicit group activity links were too sparse in my data to truly characterize group activity, and instead used dyadic activity–friendship activity between members of a group–as the measure of group activity. I examined how the link formation structure correlated with activity, and found very little correlation, despite an exploration of many different characterizations of these paths. To get a more statistical understanding of these results, I performed a classification task to use features about the first three quarters of a group's life to predict its later success I found that although the path features improved the prediction slightly, just past activity values provided an extremely strong signal.

My results indicate that groups do not form at random: individuals have short paths to the groups that they join. This, in itself, may provide an interesting way of improving link prediction methods; it is not sufficient, because G and FF paths are more common than F (diffusion) paths, and although the probability of an individual link forming given the existence of such links does increase (for example, in the Livejournal graph, a potential link where a G path existed had probability on the order of 10^{-5} as compared to 10^{-8}), it is not as strong of a signal as diffusion paths (which, in Livejournal, indicate link formation with probability on the order of 10^{-3}.) However, at the same time, the actual type of paths used in formation of a group do not appear to provide a very strong signal as to whether or not the group will,

in the long term, succeed.

The weights in the SVM also indicate an interesting direction. Path values appear to be useful, but the fact that network size is so significant indicates that my decision to ignore this parameter in choosing the groups was not justified. One goal would be to determine how to reason about network size as well as group size.

This opens up a wide area for future research. In my experiments, I did explore various characterizations of the path values as well as different group sizes. I found similar results for groups of sizes between 10 and 50 in their third quarter of life. It would be interesting to explore a wider variety of starting and ending times as well as group sizes.

Characterizing activity proved to be one of the unexpectedly difficult aspects of the research, and it is not clear that the metric that I used perfectly characterized group activity. For example, the metric simply counts that absolute dyadic activity count without scaling by size; however, intuitively, a 5-person group with the same amount of activity as a 5000 person group seems to be much healthier, even though the 5000-person group has obviously grown more.

In my prediction attempts, I decided to focus on controlling for starting group size, and therefore ignored confounding factors caused by network size. In the future, more work needs to be done to understand the role of network size in activity and growth. Network size might also need to be taken into account in estimating activity. For example, a future experiment might bin network activity by the most and least active groups per network rather than over all networks. It might also prove useful to try to divide the networks used according to type. Perhaps in different types of networks, communities function in different ways and require different characterizations of growth. Overall, my research provided more questions than answers about the role of network structure on group success, but indicates many interesting areas to explore in the future.

References

[1] http://www.ning.com/.

[2] snap.stanford.edu/data/.

[3] http://matplotlib.sourceforge.net/.

[4] http://orange.biolab.si/.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 4454, 2006.

[6] S. A. Delre, W. Jager, T. H. A. Bijmolt, and M. A. Janssen. Will it spread or not? the effects of social influences and network topology on innovation diffusion. *Journal of Product Innovation Management*, 27(2):267–282, Mar. 2010.

[7] N. Ducheneaut, N. Yee, E. Nickell, and R. J. Moore. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, page 839848, New York, NY, USA, 2007. ACM.

[8] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks*, page 33, 2010.

[9] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[10] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. Discrete-time markov chain approach to contact-based disease spreading in complex networks. *EPL (Europhysics Letters)*, 89:38009, 2010.

[11] A. Iriberri and G. Leroy. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)*, 41(2):129, 2009.

[12] E. Jin, M. Girvan, and M. Newman. Structure of growing social networks. *Physical Review E*, 64(4):046132, 2001.

[13] Kairam, S., Wang, D., and Leskovec, J. The Life and Death of Online Groups: Predicting Group Growth and Longevity. In submission.

[14] M. Kimura, K. Saito, K. Ohara, and H. Motoda. Learning to predict opinion share in social networks. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI 2010)*, page 13641370, 2010.

[15] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web*, page 695704, 2008.

[16] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, page 2942, 2007.

[17] G. Palla, A. Barabsi, and T. Vicsek. Quantifying social group evolution. *Arxiv preprint arXiv:0704.0744*, 2007.

[18] D. Raban, M. Moldovan, and Q. Jones. An empirical study of critical mass and online community survival. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, page 7180, 2010.

[19] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 717726, 2007.

[20] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[21] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, page 981990, New York, NY, USA, 2010. ACM.