The Life and Death of Online Groups: Predicting Group Growth and Longevity

Sanjay Kairam
Stanford University
Computer Science
sanjay.kairam@gmail.com

Dan J. Wang
Stanford University
Sociology
djwang@stanford.edu

Jure Leskovec
Stanford University
Computer Science
jure@cs.stanford.edu

ABSTRACT

We pose a fundamental question in understanding how to identify and design successful communities: What factors predict whether a community will grow and survive in the long term? Social scientists have addressed this question extensively by analyzing offline groups which endeavor to attract new members, such as social movements, finding that new individuals are influenced strongly by their ties to members of the group. As a result, prior work on the growth of communities has treated growth primarily as a diffusion processes, leading to findings about group evolution which can be difficult to explain. The proliferation of online social networks and communities, however, has created new opportunities to study, at a large scale and with very fine resolution, the mechanisms which lead to the formation, growth, and demise of online groups.

In this paper, we analyze data from several thousand online social networks built on the Ning platform with the goal of understanding the factors contributing to the growth and longevity of groups within these networks. Specifically, we investigate the role that two types of growth (growth through diffusion and growth by other means) play during a group's formative stages from the perspectives of both the individual member and the group. Applying these insights to a population of groups of different ages and sizes, we build a model to classify groups which will grow rapidly over the short-term and long-term. Our model achieves over 79% accuracy in predicting group growth over the following two months and over 78% accuracy in predictions over the following two years. We utilize a similar approach to predict which groups will die within a year. The results of our combined analysis provide insight into how both early non-diffusion growth and a complex set of network constraints appear to contribute to the initial and continued growth and success of groups within social networks. Finally we discuss implications of this work for the design, maintenance, and analysis of online communities.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database applications – *Data mining*

General Terms: Measurement; Theory.

Keywords: Social Networks, Group Formation, Online Communities, Information Diffusion

1. INTRODUCTION

The formation and evolution of groups has long been a focus in social science research. While the proliferation of online communities has made it possible to study the dynamics of group membership in richer detail than ever before [2, 15, 19, 31], several important questions about the factors that influence group growth remain unresolved. Why do some groups grow faster than others? Why do some groups continue to attract new members over time while others do not? What causes groups to *stop* growing? These fundamental questions in the study of group dynamics have important implications for research on communities, both online and offline.

Motivation. Researchers have only recently begun to take advantage of large-scale data from online social networks and communities to study the dynamics of group formation and evolution. A key motivation in this paper is not only to contribute to the understanding of the relationship between structural features of a group and its future growth, but also to understand how this relationship might change over time. While past work has measured growth rates over a fixed period of time, there is no reason to expect that what makes groups grow fast in the short-term is the same as what makes them grow fast or continue growing over longer periods. In addition, past research has not examined future growth rates as a function of past growth or a group's current size or age. These gaps represent important issues in the study of online communities.

We also seek to address unresolved questions from the more recent work on social networks and group evolution. Prior research on online community growth has resulted in sometimes puzzling findings indicating that ties spanning group boundaries may induce individuals to join a group while also slowing down overall growth. In this work, we are motivated in part by the desire to explain these seemingly inconsistent results.

An Empirical Puzzle. Past research in the social sciences has advanced important, but sometimes, conflicting hypotheses about the relationship between group growth and group network structure. Citing the example of a Boston community which failed to cultivate enough support to preserve itself aganst a threat, Granovetter hypothesizes that this community failed to grow because it was too clustered, though this hypothesis is not tested empirically [12]. The intuition is that if the members of a group have a disproportionate amount of friendship or communication ties with people within the group (as opposed to with individuals outside the group) then the group is too inwardly focused to ever grow. For Granovetter, weak ties outside of the group could have facilitated the mobilization of resources from members outside the community to defend against being taken over by adjacent neighborhoods.

The strength of Granovetter's 'weak ties' traditionally lies in their ability to facilitate the spread of information. Centola and Macy, however, identify that the act of joining a group may be
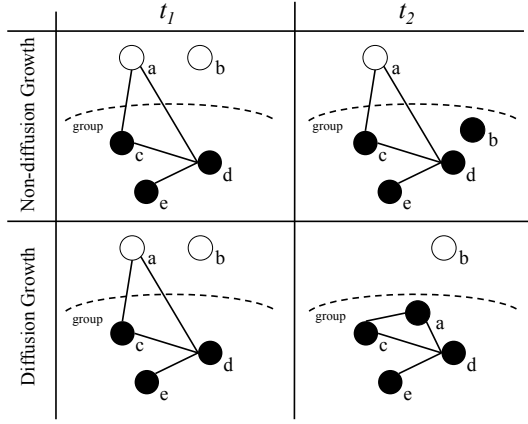
Figure 1: Illustration of diffusion vs. non-diffusion growth

categorically different from simply being informed of it [6]; using simulated networks, they provide examples of how such 'complex' contagions may not spread via weak ties in the same way [5, 6]. In contrast to Granovetter, Centola and Macy's results suggest that diffusion and growth processes in groups are facilitated by the presence of strong ties and clustering within the group.

Backstrom, et al. provide one of the first empirical investigations of group evolution using data from online social networks [2]. Examining the LiveJournal friendship network and DBLP coauthorship network, the authors identify how certain network structural features of a group influence its growth patterns over time. Their work casts group growth mainly as a diffusion process similar to Centola [6] and Granovetter [12] — that is, they assume that a group grows through the ties its members have to individuals outside the group. In other words, researchers envision membership status spreading through the network structure of a given community. Among other insights, they uncover the seemingly paradoxical finding that clustering in a group both attracts new members and decreases overall group growth, bearing out the predictions made by both Centola [6] and Granovetter [12].

Specifically, for an individual with social ties to members of a group, Backstrom, et al. find that the probability of joining the group increases with the number of ties among the individual's in-group friends. This suggests that as the clustering of a group increases, individuals with ties to the group will be more likely to join, boosting group growth. At the same time, however, the authors also find that as the ratio of open to closed triads in a group (a measure of clustering) increases, the average four-month growth rate of a group decreases. Thus, they leave a key question about their findings unresolved: If individuals with prior connections to a group are more likely to join a highly clustered group, then why do highly clustered groups experience lower growth rates overall?

1.1 Overview of Results

In this paper, we contribute to research on group evolution in social networks by making a conceptual difference between diffusion and non-diffusion growth in groups. Diffusion growth describes the process in which groups attract new members through ties to existing members; in these cases, the ties form the 'skeleton' over which membership spreads. In non-diffusion growht, individuals with no prior ties to any group members become members themselves. Here, group membership apepars to 'jump' across the community network. This is an important distinction not explicitly made in prior work, but vital to understanding why some groups grow larger, faster, and for longer preiods than others.

Figure 1 illustrates the difference between diffusion and non-diffusion growth in a group. In both scenarios at time t_1, nodes c, d, e are members of a group, and nodes a and b are not members. a, however, is connected to two of the group's members, whereas b has no ties to the group (borrowing the language of [2], we explain in section 3 that a is part of the group's *fringe*). The group experiences diffusion growth if at time t_2, a joins the group. The group experiences non-diffusion growth if b joins the group at time t_2.

Analysis of Diffusion and Non-Diffusion Growth. Our analysis focuses on the differences in the processes which govern diffusion and non-diffusion growth. In our two main analytical exercises, we explore the tension between these two types of growth as both outcomes and inputs in models of group formation and future growth. We conduct our analysis using data on online groups drawn from communities built on a common platform, Ning. Through empirical study of these networks, we explain the puzzling finding that clustering within a group can both increase the probability of a non-member joining and decrease the overall growth rate of a group.

We gain this insight by examining the mediating role that diffusion growth plays in explaining the relationship between clustering and group growth. Our main finding is that group clustering does increase the diffusion growth of a group, but that groups which grow primarily through diffusion reach smaller sizes eventually. In this way, the apparent inverse relationship between clustering and overall group growth is actually moderated by the different effects of group clustering on diffusion and non-diffusion growth. If a group is highly clustered, it is more likely to experience diffusion growth, but if more of a group's growth comes from diffusion, then it is also less likely to grow larger overall. In addition, we find that small groups within smaller communities tend to experience more diffusion growth than similarly sized groups in larger communities. This suggests that understanding group growth requires attention not only to aspects of individuals and the group, but also to the wider network setting in which these groups are situated.

Predicting Group Growth and Longevity. The second part of our analysis uses these insights about diffusion and non-diffusion growth to generate models which predict the growth and longevity of groups mined from a diverse set of thousands of independent networks. We analyze the predictive accuracy of various group structural features for two outcomes: (1) the *rate* of group growth over a fixed period of time, and (2) the *longevity* of a group's growth (i.e. the amount of time that passes before a group stops growing). Our models incorporate features capturing the rate of growth and the proportion of that growth occurring due to diffusion processes, as well as network features such as the group's transitivity and the size of cliques within groups. We demonstrate the effects of these features on group growth and longevity. In addition, we extend past work on group growth by predicting growth for groups at different stages of evolution (i.e. groups of different ages and sizes).

Specifically, we find that the predictive accuracy of certain group features is contingent on the current size and age of a group. For predicting short-term growth, models that incorporate past growth rates have the greatest accuracy. Models that include network structural features predict long-term growth more accurately. Regarding longevity, we find that the size of the largest clique within a group predicts positive growth more accurately than all other features. Our models using all features generally achieve AUC values greater than 0.78, meaning that we consistently perform better than chance, and our most accurate models achieves AUC close to 0.87. We utilize our findings to first show how features relevant to growth may vary depending on specific prediction goals and discuss how the results of our analysis can help inform insights about the struc-

Feature	All Groups	Small Grps	Large Grps
Number of Nodes	32.9	14.9	51.6
Number of Edges	108.2	25.4	194.3
Giant Component	.89	.86	.91
Transitivity	.30	.33	.28
Fringe Size	488.2	344.2	638.1
Longevity	336.4	283.9	390.9

Table 1: Network descriptive statistics for Ning group sample (mean values).

tural properties which predispose certain groups to continue growing and others to stop.

Finally, we conclude our paper with a discussion of the implications of our findings for future work on online communities and group evolution. We generalize our results to other settings and describe how past work can be reinterpreted in light of our findings. By paying closer attention to different types of growth processes, researchers can better focus their analysis on linking individual online behavior to macro dynamics of group formation and evolution. In addition, we contextualize our findings in practical applications by explaining how understanding group growth can help architects of online communities design platforms that better promotes membership and participation. We also discuss limitations of our work, open questions, and future directions for research.

2. RELATED WORK

The growth of groups in social networks has often been viewed as a diffusion process. Consequently, many researchers treated growth as a process similar to the diffusion of innovations [26] and other behaviors. Early studies, such as Milgram's small-world experiment [18], demonstrated that information could spread along social networks quickly, and Granovetter's theory of the "strength of weak ties" [12] provided a conceptual framework for how this might occur. Granovetter's theory explained these global patterns of information transmission using local features built into the structure of the network. It was natural, then, that social scientists interested in group evolution used the language of diffusion to describe how group growth.

Research on social movements is one area which has emphasized the importance of diffusion growth in groups. A *social movement* refers to the phenomenon in which a group of individuals join to take collective action in order to press for some social change or express some grievance against an authority figure. Consider, for example, the protest activity in the Middle East in February 2011 and the role played by social networks in mobilizing protestors [27]. Past work has shown that such movements grow by attracting individuals with prior ties to group members [9, 17]. Most social movement studies share the view that participants are recruited through pre-existing ties with group members [28], with estimates that 74-100% of members of such movements joining due to such ties [16].

Diffusion processes in group growth has also been the focus social science research in other settings. For example, sociologists and economists have investigated how pressure from co-workers can make individuals more likely to join labor unions [30]. In addition, in political parties and civic associations, membership growth is linked to the influence of existing members who recruit friends and families as new members [25]. Even studies of religious conversion take into account the effect of being associated with existing members of a given religion [29].

The shift to online tools for social organization has provided a host of new opportunities to study the dynamics of communities.
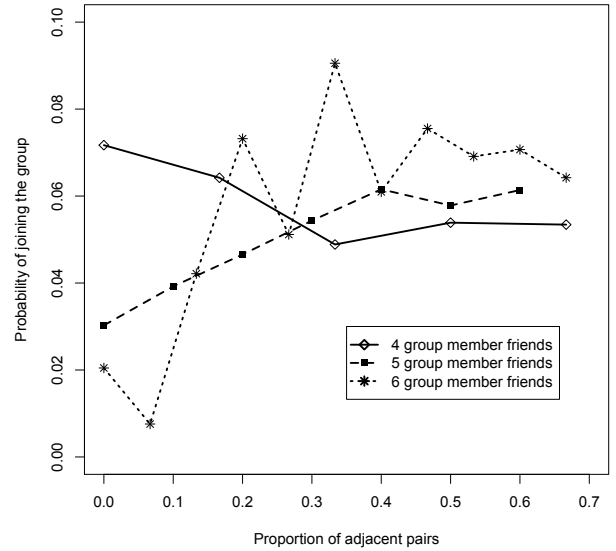

Figure 2: Proportion adjacent pairs vs. probability of fringe member joining.

While extensive research has focused on identifying implicit communities within networks [10, 15, 21], this is not our present goal. Other work has focused on capturing properties of explicitly defined groups within networks. Mislove, et al. analyze groups occuring in a variety of networks (Flickr, YouTube, Orkut, and LiveJournal) finding that they tend to overlap with highly clustered subgraphs of the network and that smaller groups exhibit higher clustering than larger ones [19]. Zheleva, et al. develop a theoretical model of how users in an evolving social network might affiliate which closely matches the observed properties of groups in Flickr [31].

Ducheneaut, et al. examine the impact of network features on the longevity of groups, finding that guilds in World of Warcraft with smaller subgroups and higher density survive longer; given the intensely goal-directed nature of these groups, they hypothesize that such structures reduce coordination problems [7]. As mentioned, our work is informed by Backstrom, et al.'s investigation of the formation and growth of groups in large social networks [2]. Our analysis starts by answering the open questions left by this prior work and goes on to develop a more comprehensive view of what predisposes some groups to rapid and continued growth. We also analyze groups within thousands of independently-formed social networks, which we believe improves the robustness and generalizability of our findings.

3. NING COMMUNITY DATA

The data for our analysis comes from the membership lists of communities built using Ning (http://www.ning.com), a web platform that allow users to create their own online social networks similar to Facebook. These online communities, which we call *Ning communities* serve a variety of interests ranging from small private networks for friends and families to public networks for larger organizations. Some of the largest networks belong to entertainment properties such as the Dallas Mavericks NBA Team (http://friends.mavs.com) and the rapper 50 Cent (http://thisis50.com).

Ning communities include a variety of social features including user profiles, forums, blogs, and groups. For each user in a Ning community, we observe their group affiliations and when they joined these groups. Only some communities in our data al-
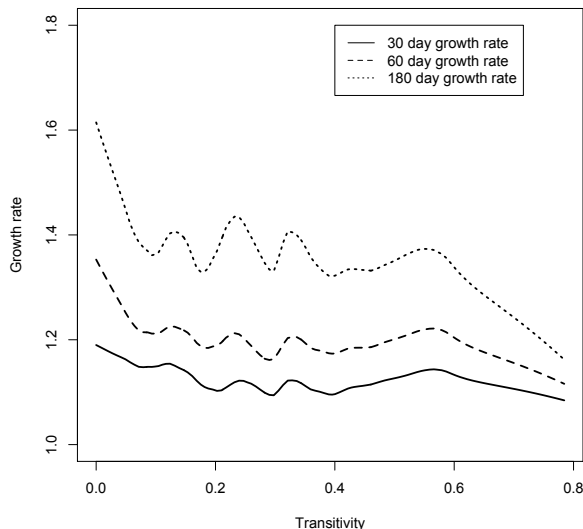
Figure 3: Future growth rates vs. transitivity for Ning groups with 50 members.

low users to explicitly 'friend' other users. For all communities, however, social interactions on a page, such as replying to a user in a forum, are stored as user-to-user comments. In our analysis, we consider an edge to have been created between two users when they have exchanged at least one comment in each direction (reciprocal communication). We found that this formulation reduced issues due to spam messages which are often unreciprocated. While these implicit edges are different from the explicit edges based on 'friend' relationships, we find that they serve as suitable network structures for our analysis. Using the group affiliation data and the comment networks, it is possible to reconstruct the day-by-day evolution of groups within a Ning community.

4. GROUP GROWTH AND DIFFUSION

In the following subsections, we provide definitions for diffusion and non-diffusion growth. We then discuss how data was collected from Ning communities and groups to analyze how these interact with network clustering and overall growth.

4.1 Diffusion and Non-Diffusion Growth

We define diffusion growth as the addition to a group of new members with existing social ties to one or more group members (see Figure 1). Non-diffusion growth comes from new members with no prior ties to group members. This conceptual distinction leads to two hypotheses about why users join a group. Under diffusion growth, users may be influenced to join due to the behavior of their friends. In non-diffusion growth, users may join because there is a feature of the group itself (i.e. a common interest) which appeals to them. In these cases, a Ning community user might be exposed to the group through some means other than an invitation from a friend, such as banner ads on the Ning community website or a search for similar groups. It is, of course, possible that some users join groups based both on shared interests and influence from a friendship tie — which can be difficult to separate — but for our purposes, we designate users who join groups as part of either diffusion or non-diffusion growth to simplify our analysis [1].

To measure diffusion growth in a Ning group, we first consider a snapshot of the friendship network containing the members of a Ning group, G, at a given moment in its evolution, t_1. Second, we identify all users in the Ning community who are not members of the Ning group but who have *at least one tie* to the members of the
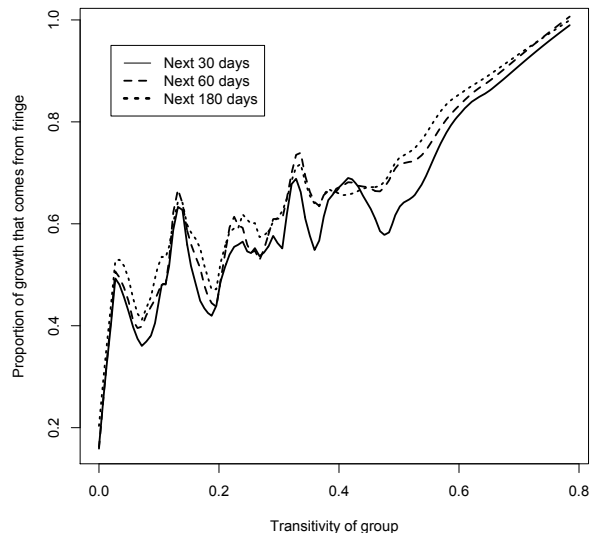

Figure 4: Group transitivity vs. proportion of future growth from the fringe.

Ning group. This set of users is what [2] call the group's fringe, f_G. Finally, we then obtain a snapshot of the Ning group network at t_2, where t_2 represents some fixed time after t_1 — call this new snapshot of the Ning group, G'. We then identify the members of G' that were not part of G as m, representing the overall growth of G from t_1 to t_2. The proportion of Ning users in m that were also in f_G signals the amount of growth in G attributable to diffusion growth whereas $(|m|/|m \cap f_G|)$ gives the total amount of non-diffusion growth in G from t_1 to t_2.

4.2 Sampling Ning Communities and Groups

For the analysis in this section, we sampled Ning groups and communities using the following rules. First, based on an examination of several groups, we set a lower threshold of 10 members for the final group size. Groups which did not grow larger than 10 members may represent 'failed attempts' or 'tire-kicking' by creators and thus would not be sensible to include in our sample. We also excluded all groups which grew to be larger than 50% of the surrounding community, as these may represent groups which are not distinct from the overall community (and thus would grow artificially large). In order to eliminate 'right-censoring' issues in assessing group growth over time, we considered only groups which had stopped adding new members 2 months prior to the end of the data collection period; while this may form an artificial constraint, it allows us to analyze groups according to their final size. Table 1 reports descriptive statistics for the sample of 4,051 Ning groups from 418 Ning communities used in the analysis below. In Table 1, 'small' groups are those smaller than the median group size in our sample (22 eventual members) 'small', and 'large' are those larger than the median.

4.3 Resolving the Clustering-Growth Paradox

Using our sample of Ning groups, we replicated experiments by Backstrom, et al. [2] concerning the effects of group clustering on future group growth and the probability of an individual in a group's fringe joining the group [2]. First, for each group, G, with 50 or more members, we obtain a snapshot of the group on the date, time t, on which it gained its 50th member. We then calculate the proportion of members in G's fringe, f_G, who joined the group after 180 days — we treat this as the probability of a fringe member joining the group. Then, for each group, we calculate the average
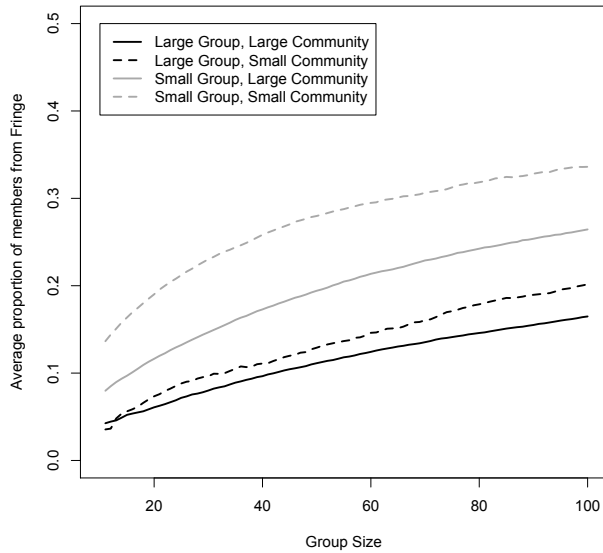
Figure 5: Proportion of group membership from fringe vs. current group size (by eventual group size).

Figure 6: Proportion of current membership from fringe vs. future overall growth.

proportion of adjacent pairs among a fringe member's group member friends. The proportion of adjacent pairs refers to the number of ties between a fringe member's group member friends divided the by the possible number of ties between them.

Figure 2 plots the average probability of a fringe member joining a group against the average proportion of adjacent pairs among this individual's in-group friends for all of the groups in our sample (with separate lines for fringe members with 4, 5, and 6 friends in the group). We observe a generally positive relationship between the proportion of adjacent pairs and the probability of a fringe member joining a group which strengthens with the number of friends that the fringe member has in the group. The explanation for this is intuitive. If non-member has six friends in the group who are highly clustered, the social pressure to join the group is stronger than that for a non-member who has only three friends in the group who are highly clustered.

In Figure 3, we also plotted the 30-, 60-, and 180-day future growth rates of Ning groups at size 50 against their clustering as measured by transitivity. The transitivity of a graph, G, is equal to the number of closed triads in G divided by the possible number of closed triads in G. According to Figure 3, the average future growth rate of a Ning group (of size 50) decreases as the current transitivity of a group increases. For example, for Ning groups that have transitivity = 0, 180-day growth rates exceed 1.6, whereas a highly clustered group, with transitivity = .8, experiences a 180-day growth rate of only 1.2.

The findings shown in Figures 2 and 3 show that group clustering appears to both increase and decrease future group growth across the wide range of groups and diverse communities available in the Ning data. Thus, the relationships that Backstrom, et al discover in their work are not particular to the LiveJournal or DBLP communities [2]. Having shown the general nature of this pattern, we now move on to explaining how it arises.

4.4 Network Clustering and Growth

Comparing diffusion and non-diffusion growth for Ning groups sheds light on these contradictory findings. In Figure 4, we plot the transitivity of a group at t, where t is equal to the time where a group reaches 50 members, against the proportion of its growth in the next 30, 60, and 180 days that comes from the fringe (i.e. diffu-
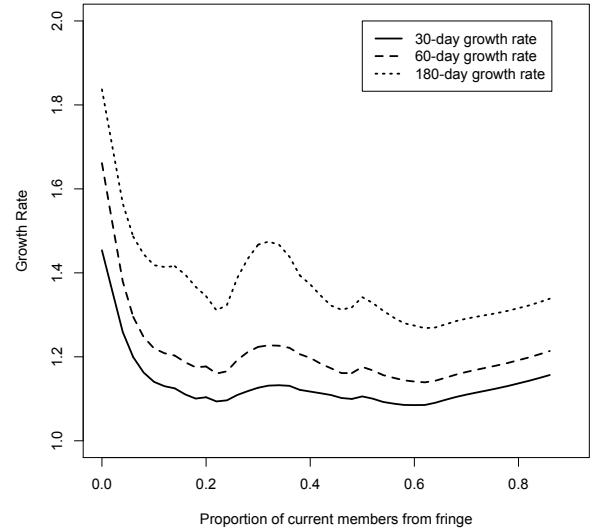
sion growth). We calculate the proportion of growth attributable to diffusion growth as described above. Figure 4 contains smoothed versions of line plots to facilitate interpretation.

The results from Figure 4 are clear. As the clustering of a group increases, the proportion of its future growth from its fringe also increases. However, even though group clustering promotes diffusion growth, this is not enough evidence to claim that higher clustering leads to greater overall growth. In fact, most of the confusion associated with the clustering-growth paradox comes from assuming that group growth is primarily achieved through diffusion growth.

Higher group clustering therefore leads to a greater proportion of its members joining diffusion growth. This, in turn, influences future rates of overall growth. Figure 5 illustrates the differences between small and large groups (embedded within small and large communities) in terms of their fringe and non-fringe membership composition over their first 100 members. To generate this graph, we divided groups into four categories based on 'final' group sizes and 'final' community sizes. We classify any group which never gains more than 113 members as 'small', and those which surpass 630 as 'large'. Communities with more than 2000 users are large, while those with less are small.

Figure 5 shows that in groups which eventually grow large, a smaller proportion of membership comes from the fringe than in groups which eventually grow small; moreover, this is true at any point in a group's growth from 10 to 100 members. In small Ning communities, groups which eventually grow larger than 630 members will gain only 10% of their first 40 members from the fringe, on average (black dotted line in Figure 5). In these same communities, groups which never grow larger than 113 members will gain 25% of their first 40 members from the fringe (gray dotted line in Figure 5).

The resolution to the growth-clustering paradox is evident from the patterns shown in Figures 4 and 6. Figure 4 shows that higher group clustering increases the proportion of growth which happens as a result of diffusion processes. Figure 6 shows that the group's overall growth rate decreases as the proportion of group membership coming from the fringe increases. Thus, we infer that higher clustering increases the amount of diffusion growth happening in a group, but decreases a group's growth rate overall.

We attribute this pattern to the notion that some groups grow by appealing to common interests and identities (non-diffusion growth) while other groups grow by virtue of its extra-group connections

Category	Feature	Description
Growth	Monthly Growth Rate	Fraction of users who joined in the prior month
	Fringe Growth Rate	Fraction of users who joined in the prior month who joined from the fringe
Connectivity	Group Transitivity	Transitivity of network formed by group members
	Transitivity Ratio	Ratio of group transitivity to transitivity of entire community
	Group Density	Density of network formed by group members
	Density Ratio	Ratio of group density to density of entire community
Structural	Clique Ratio	Largest fraction of group members whose edges form a clique
	Disconnected Ratio	Fraction of group members who are not a part of the group's largest connected component

Table 2: Features used in all growth and longevity models.

(diffusion growth). If a group relies on diffusion growth, it can only grow as much as the number of ties its members have to non-members. At some point, a group relying on diffusion growth, might run out of such ties, constraining its eventual growth. Thus, while clustering in a group can increase joining activity from the group's fringe, it can also diminish the group's overall growth.

Group growth and Ning community size. While Figure 5 offers insight into the relationship between the amount of growth a group experienced through diffusion and its future growth, it also shows how this dynamic operates for groups in different settings. While groups that reach eventually smaller sizes tend to experience more diffusion growth, we can also compare growth rates for eventually small groups in large communities and small communities.

According to Figure 5, among eventually small groups, those that were established in Ning communities that are eventually small tend to experience a greater proportion of diffusion growth than those in Ning communities that are eventually large. For example, in an eventually small group with 40 current members, an average of 28% of its current members will have come from the fringe if the group is in an eventually small community. By contrast, if the same group were in an eventually large community, only 16% of its current members will have come from the fringe on average.

These observations lead to the hypothesis that smaller communities may foster greater familiarity among individuals, and thus, stronger social pressures to adopt behaviors such as group membership. This is consistent with social science research finding that participation rates in civic groups or clubs, such as voter registration organizations, in small towns and rural areas is far higher than in big cities [22]. Part of the explanation is that individuals experience more peer pressure in smaller, more intimate communities than in big cities where they may be more isolated.

Figure 5 also shows the same contrast between small and large communities for eventually large groups, but the difference is much smaller. This is likely because eventually large groups generally attract many members through wider appeal. That is, individuals join not because they were invited by a 'friend' in the group; instead, they join because they have some broad interest shared by the members of the group. For example, consider the differences between a Ning group that serves as an online fan club for a professional sports team and a group used by close friends to exchange private group messages. Regardless of where the online fan club group is established–i.e. a small or large Ning community–its growth primarily comes from its appeal to other fans rather than through the influence of its members' external friendship ties.

Diffusion Growth and Future Overall Growth. Figure 6 further clarifies the relationship between the proportion of a group's current members who came from the fringe and the group's future growth rate. Specifically, in Figure 6, if no members of a 50 member group joined from the fringe, then on average, we expect the group to grow by a factor of 1.6 over the next 180 days. However, if 90% of the group joined from the fringe, then we expect the group to only grow by a factor of 1.2 in the same amount of time. We observe the same pattern when examining growth rates over periods of 30 or 60 days. Again, this suggests that fringe growth early in a group's existence can diminish the eventual size of the group. One hypothesis about why groups which grow primarily through diffusion would be eventually smaller concerns the purposes of these groups. Prior work has shown that group dynamics can differ greatly for groups based on a *common bond* versus those based on a *common identity* [23, 24]. We discuss the connection between these attachment types in Section 6.

5. PREDICTING GROWTH & LONGEVITY

Building on these empirical observations, we engage in two machine learning tasks aimed at predicting which groups will be successful in attracting new members. The goal of our first task is to understand what group and network characteristics predict whether a group will grow faster or slower than others. Specifically, we explore how the predictive value of these features vary along three dimensions: *group age*, *group size*, and *prediction interval*.

We test the effects of age by taking snapshots of groups at 60 and 180 days after their creation. For each age group, we separate these snapshots into two categories according to size, resulting in a total of 4 age-size 'buckets' of data; for each of these buckets, we generate two models, one aimed at predicting short-term growth and the other at predicting long-term growth. Building these 8 models in this manner provides results which are cleaner and more interpretable as they allow us to clearly observe how coefficients change across groups of different ages and sizes and for the two prediction intervals without having to explain complex interaction effects. Comparing these models provides a rich picture of how diffusion processes, group transitivity, and other features predict short-term and long-term growth of groups. In our second modeling task, we utilize these same features to predict whether groups will 'die' or cease to grow within a given period of time, shedding additional light on how these group and network characteristics affect group dynamics.

5.1 Features for Learning

We begin by defining a set of features to be used in our predictive models, described in Table 2. We divide these features into three rough categories which describe the data at different levels. The first category, labeled *growth* features, capture how quickly and by what means the group was growing when the snapshot was taken. These features serve as a baseline, capturing both the current 'velocity' of growth, as well as insights from the prior section about the role of diffusion processes in subsequent growth.

In addition, we consider a set of *connectivity* features, capturing the probability of edges among members of the group absolutely and relative to the community as a whole, and *structural* features, chosen because they succinctly summarize aspects of higher-level structure within the group. As most features approximated a log-

Size	Statistics	60 Days	180 Days
	Number of Groups	5871	5312
Small	2-Month Growth	1.333	1.165
	2-Year Growth	2.613	1.934
	Number of Groups	1602	2884
Large	2-Month Growth	1.147	1.076
	2-Year Growth	1.494	1.267

Table 3: Number of groups, short-term (2-month) and long-term (2-year) median growth rates for each age*size bucket.

normal distribution, values were log-transformed as part of our analysis. *Monthly growth rate* and *group transitivity* more closely approximated normality after being power-transformed.

5.2 Predicting Group Growth

We now describe in more detail the procedure used in our first task of generating predictions about short-term and long-term group growth. We started with a sample of 11,944 groups mined from 1,713 distinct Ning communities. These groups were selected using the same criteria for group size, community size, and expiration date as in our prior analysis. To aid in generalizability, we limited data collection to no more than 50 groups within any single community in order to avoid over-representing a single, large community in our analysis.

As mentioned above, we generated snapshots for groups at two ages (60 and 180 days) and then separated groups of the same age into 'small" (10-100 members) and 'large' (150-1000) members. Rather than including group size as a dependent variable, we bin groups in this manner for two reasons: (1) results from prior work indicate a natural threshold for group size around 150 members [8, 15], (2) our dependent variable, growth rate, does not exhibit equal variance for small and large groups (a small group can feasibly grow 100x in size, while a large group can not).

For each group, we define the 'growth rate' to be the ratio of the group size at prediction time to the group size at the time of the first snapshot. For each group at a given age, we calculate the short-term (2 month) and long-term (2 year) growth rates. Table 3 shows median growth rates over these two prediction intervals. For each combination of age, size, and prediction interval, we structure our problem as a binary classification task, with class 1 representing groups growing more quickly than the median rate, a formulation which provides a balanced sample without excluding groups from our analysis. We generate predictions using a classifier based on logistic regression.

5.2.1 Growth: Results

Below, we summarize and compare results from the 8 models (for each combination of group age, size, and prediction interval) described above. In evaluating our models, we consider two different evaluation measures: the classification accuracy and the area under the ROC curve (AUC). In each task, we generate models for each category of features and a fourth combining all features.

Short-term growth. In Figure 7, we show the accuracy of model predictions for short-term growth. Rows show results for groups of the same age, and columns show groups of the same size. In each cell, the four bars correspond to the four models: (G)rowth, (C)onnectivity, (S)tructural, and (ALL) features. For short-term growth, we can clearly see that the growth features contribute the most to the accuracy of the models and that predictions are more accurate for larger, younger groups, perhaps where the growth rate from the prior month provides more signal. For large groups at 60 days, for instance, the combined model achieves 79.2% accu-
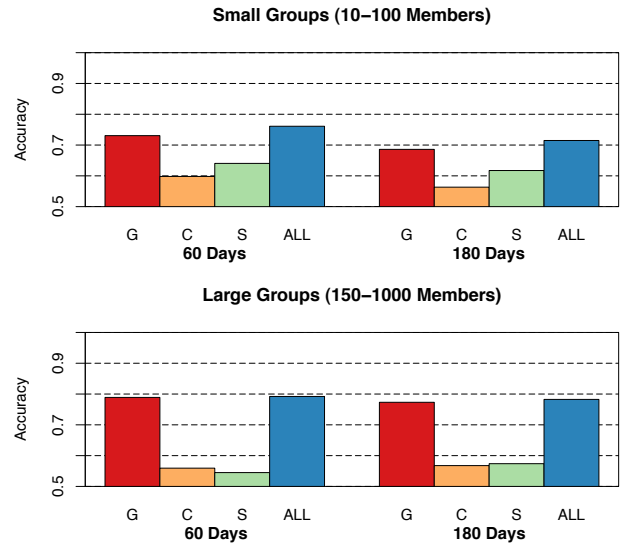

Figure 7: Short-term (2-month) prediction accuracy of models. Rows show groups of the same size, columns show groups of the same age. Note that the y-axis starts at 0.5, the expected accuracy of a random model.

racy (AUC = 0.868), with predictions driven almost entirely by the growth metrics. As shown in Figure 7, models based on growth metrics alone consistently achieved prediction accuracy over 68.6% (consistently over 77.3% for large groups).

Table 4 shows regression coefficents for the models combining all features. Within each age-size bucket, distributions for each feature were standardized to have a mean of 0 and standard deviation of 1, allowing us to directly compare regression coefficients. Here, we see clearly that short-term predictions are heavily dependent on monthly growth rate, especially for large groups. For small groups, we find that increased group transitivity and growth from the fringe signal a decreased likelihood that a group will grow rapdily, matching earlier observations about diffusion processes, transitivity and growth. An intriguing finding seemingly at odds with this observation about transitivity is that a larger clique seems to predict an increased likelihood that a group will grow rapidly, with this effect stronger for small groups. Furthermore, groups with more individuals outside the largest connected component seem to grow more quickly as well. Together, these findings hint at a more nuanced picture of how network properties predispose groups to growth, which we will discuss later on in our analysis.

Long-term growth. Figure 8 shows prediction accuracy achieved by our models for long-term growth. Again, rows correspond to group size, columns to group age, and the bars in each cell represent the four models. In general, our models still achieve relatively high accuracy, up to 78.6% (AUC = 0.868) for small groups at 60 days and consistently over 71.9% (AUC = 0.782), in predicting whether groups will grow rapidly over the subsequent two years. We see an interesting pattern where model fit for small groups is better at 60 days and for large groups at 180 days; looking more closely, we see that this is because different features appear to be contributing to predictions for small and large groups. For small groups, we see that the models based on structural features are more accurate than those based on growth metrics.

In Table 5, we show coefficients for the combined models (as with the short-term models, distributions for feature values are standardized). Again, we see a pattern where increased growth from the fringe predicts decreased growth rates for small groups and increased transitivity predicts decreased growth for all groups. In

Feature	Small Groups (10–100 Members)		Large	
	60 Days	180 Days	60 Days	180 Days
Monthly Growth	1.19***	0.88***	1.94***	2.00***
Fringe Growth	-0.20***	-0.10*	—	—
Group Trans.	-0.68***	57***	—	—
Trans. Ratio	0.22***	12*		
Group Dens.				
Dens. Ratio		.11		
Clique Ratio		0***		
Disconnected		8***		

Table 4: Reg_____ r com____ ing short-term growth (for this and following tables. *** $p <$ 0.01, ** $p < 0.005$, *** $p < 0.001$). Coefficients with $p > 0.05$ are not repor____

addition, the p_____ e of larger ___es and ___ __ memb____ e the giant com_____ incre_____ s of all sizes an_____ t least ____ e structural feat_____ y a v_____ g the outcome o_

Predictive Va_____ expect_____ e growing quickly at the time of the first snapshot were likely to continue to do so over the short-term, though the extent to which this feature predicted long-term growth for large groups may have been surprising. Of greater interest to us, however, were the findings pertaining to fringe growth, which confirm the results of our prior empirical investigation concerning the role that diffusion processes play in the growth of groups. We find that the negative effect of diffusion processes on subsequent growth is significant only for small groups, suggesting some support for our initial hypothesis about differences between groups smaller and larger than 150 members.

Similarly, for small groups, our models confirmed prior observations concerning group transitivity and growth, with increased transitivity predicting decreased growth over the short-term for small groups and over the long-term for all groups. An interesting result seemingly at odds with these findings about transitivity was that, over both prediction intervals, the presence of a large clique appeared to predict a greater likelihood of subsequent growth. The estimated effect of clique size on subsequent growth was especially high for small groups, hinting that these structural features may be important at the early stages of group formation. In the discussion, we provide hypotheses about types of network structures which could lead to these combinations of features. Our finding that having more members outside of the largest connected component appears to be a good predictor of growth coincides with our notion of non-diffusion growth (membership 'jumps' across the network rather than following existing connections).

We chose this classification approach to match the real-world analysis task of identifying groups which will be 'successful' in the future; an alternative approach would be to utilize linear regression to predict final group size. Though space does not permit a full exposition of these linear models, our experiments showed that group size could be predicted with some accuracy (for long-term growth predictions, we achieved adjusted R^2 values around 0.5 for models combining all features). While these final size predictions were driven heavily by growth features, an analysis of deviance revealed that adding features pertaining to network connectivity and structure provided a better model fit despite the added complexity.

5.3 Predicting Group Longevity

In this section, we focus on the closely related problem of predicting when groups will 'die', or cease to attract new members.
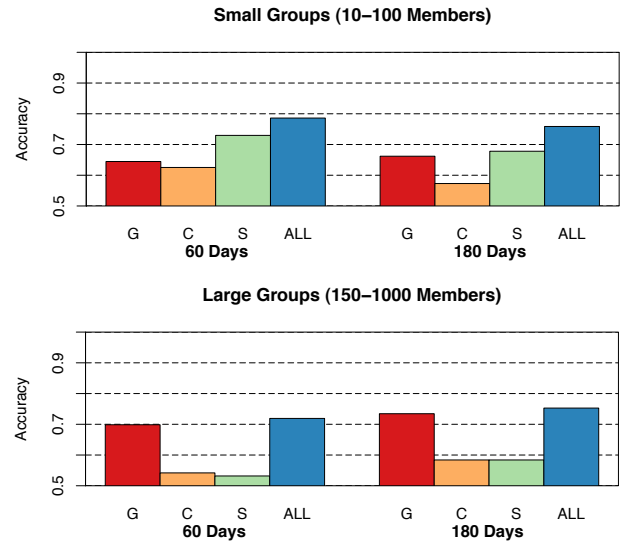
Figure 8: Long-term (2-year) prediction accuracy of models.

Instead of predicting when groups will cease to grow, we again avoid right-censoring problems by focusing on the binary classification task of predicting whether they will continue to grow after 1 year (360 days). In order to simplify our analysis in this section, we focus on groups at a single age (90 days), selecting the 1818 groups which stopped growing within a year (Class 0) and the 1818 groups which grew the fastest after 1 year (Class 1) to obtain a balanced sample. We utilize the same features and logistic regression approach used before. As our goal is now to predict which groups will 'die', we expect that the feature coefficeints learned by the model will have signs opposite those in the growth analysis (i.e. features which positively predict growth should negatively predict death).

Results. Coefficients estimated by our five models, as well as the accuracy and AUC for each, are summarized succinctly in Table 6. For the task of predicting whether groups would stop growing within a year, the group size at the time of the snapshot again provided the most information, followed by the structural metrics. The model using only structural metrics achieved a favorable 70.4% accuracy (AUC = 0.759) compared to random chance (50.0%) and the models utilizing only growth (62.0% accuracy) or connectivity (61.9% accuracy) features. Our combined model achieves 77.4% accuracy (AUC = 0.834) in predicting whether groups will 'die' within a year. Observing the magnitude and sign of learned coefficients for the longevity model, we see that they match closely the findings from the growth analysis. We summarize these observations in the list below and then in more detail in our discussion.

- A higher proportion of growth from diffusion corresponds to a higher likelihood that a group will die.

- Clique Ratio is the strongest feature: a large clique makes a group significantly less likely to die.

- We see similar patterns as with growth with respect to the Group Transitivity and Disconnected Ratio features – low transitivity and fewer members in the large connected component make a group significantly less likely to die.

- One finding which differs from the growth analysis is that groups which are more dense than their surrounding community are more likely to die.

- Groups with a high monthly growth rate are more likely to continue growing a year later.

	Small		Large	
Feature	60 Days	180 Days	60 Days	180 Days
Monthly Growth	0.49***	0.62***	1.24***	1.55***
Fringe Growth	-0.27***	-0.31***	—	—
Group Trans.	-1.34***	-1.37***	-0.31**	-0.25*
Trans. Ratio	0.37***	0.33***	—	0.21**
Group Dens.	—	—	-0.41	—
Dens. Ratio	-0.32***	-0.49***	—	-0.17
Clique Ratio	2.64***	2.33***	0.67***	0.48***
Disconnected	0.75***	0.78***	0.20	0.47***

Table 5: Regression coefficients for combined models predicting long-term growth.

Feature	Model G	Model C	Model S	Model ALL
Prior Month	-0.46***			-0.39***
Fringe Growth	0.26***			0.27***
Group Transitivity		0.23***		0.30***
Transitivity Ratio		-0.18***		-0.12
Group Density		-0.67***		0.27
Density Ratio		0.41***		0.65***
Clique Ratio			-1.05***	-1.59***
Disconnected Ratio			-0.62***	-0.47***
Accuracy	0.622	0.619	0.701	0.744
AUC	0.656	0.647	0.758	0.809

Table 6: Coefficients and prediction outcomes for modles predicting at 90 days whether groups would cease to grow within a year. Note that coefficient signs differ from the prior analysis.

6. DISCUSSION

Core-Periphery Structures. Building off of these findings, we could hypothesize that groups with one or more cores of tightly connected members and a periphery of members loosely connected or entirely disconnected from this core should experience increased and prolonged growth (low transitivity, small connected components, and large cliques). Such core-periphery structures have been observed empirically in many large, real-world networks such as communication and transportation systems [13], online media sites [11], and social networks [15]. The densely connected core allows for the swift transmission of resources [13] and the loose periphery allows for the presence of structural holes [3], or ties which bridge clusters, allowing members on the periphery to bring new information or members to the core. When all members are too tightly bound to the core, as they are in groups with high fringe growth and high transitivity, groups might become too inwardly focused, precluding the possibility of gaining new information and members outside the group.

Common-Bond vs. Common-Identity Groups. Qualitative research on online and offline groups has detailed differences between groups based on *common bond*, where group members form attachments to one another, and those based on *common identity*, in which group members form attachments to the group itself. Our results imply that groups with high transitivity and high diffusion growth represent common-bond groups due to the strong relationship between group members. Prior research on offline communities [23] has argued that common-identity groups may adapt better to changes in membership, as attachment is not so dependent on who is or is not in the group. This resilience may play a greater role in online communities, where membership can change much more dramatically. Research on online communities has shown that groups which accrue more traffic witness greater membership turnover [4, 14]. Ren, et al. [24] hypothesize that change of this type may be unsettling for members of bond-based communities, leading these communities to falter.

7. CONCLUSION

Summary. We investigated the relationship between a group's network features and its future growth using online community data from Ning.com social networks. We considered two types of growth: 1) diffusion growth, wherein a group attracts new members through the friendship ties of its current members to outsiders, and 2) non-diffusion growth, wherein individuals without pre-existing ties to any group members join a group. First, we showed that while group clustering increases diffusion growth, groups that grow more from diffusion tend to reach smaller eventual sizes. This explains the empirical puzzle that group clustering can increase the proba-

bility of an individual joining a group (given that the individual has ties to the group) while decreasing a group's overall growth rate.

Second, we generated a set models which use a group's structural features and past growth experience to predict its eventual size and longevity. We found that past growth features predict short-term growth more accurately while, for small groups at least, network structural features better predict long-term growth. In terms of growth rates, we find that while a group's higher transitivity leads to slower long-term growth, the larger a group's largest clique, the more likely the group will experience fast growth over a 2-year period. In addition, structural features predict a group's longevity better than a group's past growth and connectivity features. We note, finally, that we were able to achieve robust results using networks based on implicit (comment) reciprocal edges rather than explicitly defined (friend) edges.

Areas for Future Research. Our findings have useful practical implications and serve as a springboard for future research. For architects and administrators of online communities, understanding the dynamics of group growth is important for sustaining and promoting interaction [20]. Knowing how groups form within networks helps community managers assess the future growth prospects of the community. In addition, it helps administrators decide whether to implement certain features, like membership invitation requests, in online groups to regulate group growth. Because our results were based on data from a wide array of online social network settings, our insights are also generalizable to many online environments.

Our findings raise several important open research questions. While we explored one measure of group longevity, studies of group growth would benefit from research that better specifies the key moments of a group's life-cycle. In addition, we welcome further investigation into the relationship between group growth and the group content. Finally, we encourage analysis on how groups within the same community network setting interact with one another — what role do members that have multiple group affiliations play in promoting diffusion growth? Under what conditions do groups 'compete' for members? We invite future research on these questions to enhance our knowledge about the relationship between networks, group growth, and online community evolution.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, Dec. 2009.

[2] 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*, KDD '06, pages 44–54, New York, NY, USA, 2006. ACM.

[3] R. Burt. Structural holes versus network closure as social capital. *Social capital: Theory and research*, pages 31–56, 2001.

[4] B. Butler. Membership size, communication activity, and sustainability: The internal dynamics of networked social structures. *Information Systems Research*, 12(4):346–362, 2001.

[5] D. Centola. The spread of behavior in an online social network experiment. *Science (New York, N.Y.)*, 329(5996):1194–1197, Sept. 2010.

[6] D. Centola and M. Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3):702–734, Nov. 2007.

[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, pages 839–848, New York, NY, USA, 2007. ACM.

[8] R. Dunbar. Coevolution of neocortex size, group size, and language in human. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.

[9] R. M. Fernandez and D. McAdam. Social networks and social movements: Multiorganizational fields and recruitment to mississippi freedom summer. *Sociological Forum*, 3:357–382, 1988. 10.1007/BF01116431.

[10] M. Givan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, June 2002.

[11] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1019–1028, New York, NY, USA, 2010. ACM.

[12] M. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[13] P. Holme. Core-periphery organization of complex networks. *Physical Review E*, 72(046111), 2005.

[14] Q. Jones, G. Ravid, and S. Rafaeli. Information overload and the message dynamics of online interaction spaces. *Information Systems Research*, 15(2):194–210, 2004.

[15] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 695–704, New York, NY, USA, 2008. ACM.

[16] G. Marwell and P. Oliver. *The Critical Mass in Collective Action: A Micro-Social Theory*. Cambridge University Press, 1993.

[17] D. McAdam and R. Paulsen. Specifying the relationship between social ties and activism. In D. McAdam and D. A. Snow, editors, *Social Movements: Readings on Their Emergence, Mobilization, and Dynamics*, pages 145–157. Roxbury Publishing Co., 2007.

[18] S. Milgram. The small-world problem. *Psychology Today*, pages 60–67, 1967.

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[20] E. D. Mynatt, A. Adler, M. Ito, and V. L. O'Day. Design for network communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '97, pages 210–217, New York, NY, USA, 1997. ACM.

[21] M. E. J. Newman. Detecting community structure in networks. *Eur. Phys. J. B*, 38:321–330, 2004.

[22] J. E. Oliver. The Strength of Weak Ties. *The American Political Science Review*, 94(2):361–373, 2000.

[23] D. Prentice, D. Miller, and J. Lightdale. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Personality and Social Psychology Bulletin*, 20(5):484–493, 1994.

[24] Y. Ren, R. Kraut, and S. Kiesler. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*, 28(3):377–408, Mar. 2007.

[25] C. T. Reviews. *Making Democracy Work: Civic Traditions in Modern Italy*. Cram 101. Academic Internet Publishers Incorporated, 2007.

[26] E. M. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 5th edition, 2003.

[27] F. Salem and R. Mourtada. Civil movements: The impact of facebook and twitter. *The Arab Social Media Report*, 1(2), 2011.

[28] D. A. Snow, L. A. Zurcher, and S. Ekland-Olson. Social networks and social movements: A microstructural approach to differential recruitment. In D. McAdam and D. A. Snow, editors, *Social Movements: Readings on Their Emergence, Mobilization, and Dynamics*, pages 122–131. Roxbury Publishing Co., 2007.

[29] R. Stark and W. S. Bainbridge. Networks of faith: Interpersonal bonds and recruitment to cults and sects. *American Journal of Sociology*, 85:1376–1395, 1980.

[30] J. Waddington and C. Whitston. Why do people join unions in a period of membership decline? *British Journal of Industrial Relations*, 35(4):515–546, 1997.

[31] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1007–1016, New York, NY, USA, 2009. ACM.