Preferential Behavior in Online Groups

Lars Backstrom[*]
Dept. of Computer Science
Cornell University
Ithaca, NY 14853.
lars@cs.cornell.edu

Ravi Kumar
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089.
ravikumar@yahoo-inc.com

Cameron Marlow
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089.
cameron@media.mit.edu

Jasmine Novak
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089.
jnovak@yahoo-inc.com

Andrew Tomkins
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089.
atomkins@yahoo-inc.com

ABSTRACT

Online communities in the form of message boards, listservs, and newsgroups continue to represent a considerable amount of the social activity on the Internet. Every year thousands of groups flourish while others decline into relative obscurity; likewise, millions of members join a new community every year, some of whom will come to manage or moderate the conversation while others simply sit by the sidelines and observe. These processes of group formation, growth, and dissolution are central in social science, and in an online venue they have ramifications for the design and development of community software.

In this paper we explore a large corpus of thriving online communities. These groups vary widely in size, moderation and privacy, and cover an equally diverse set of subject matter. We present a broad range of descriptive statistics of these groups. Using metadata from groups, members, and individual messages, we identify users who post and are replied-to frequently by multiple group members; we classify these high-engagement users based on the longevity of their engagements. We show that users who will go on to become long-lived, highly-engaged users experience significantly better treatment than other users from the moment they join the group, well before there is an opportunity for them to develop a long-standing relationship with members of the group.

We present a simple model explaining long-term heavy engagement as a combination of user-dependent and group-dependent factors. Using this model as an analytical tool, we show that properties of the user alone are sufficient to explain 95% of all memberships, but introducing a small amount of per-group information dramatically improves our ability to model users belonging to multiple groups.

Categories and Subject Descriptors

H.2.8 [**Data Management**]: Database Applications—*Data Mining*

General Terms

Experimentation, Measurements, Theory

Keywords

online community, social networks, groups, large data sets

1. INTRODUCTION

Online groups are nearly as ubiquitous as the internet itself, and social communities built around messaging were around long before the web. A Pew internet study estimated in 2001 that over 90 million Americans had participated in some form of online community related to hobbies, religious beliefs, politics, or ethnic groups [18]. Some online communities are extensions of offline organizations, others exist exclusively online, and many lie somewhere in-between. Email lists and message boards are so inextricably intertwined with the internet that it is hard to imagine a world without them.

The term "online community" is used in both the popular and academic literature to refer to many different types of entities. Consider the following three examples. First, an email list used by a local Rotary club to send announcements about upcoming events. Second, a private message board used by fathers in a Palo Alto neighborhood to share experiences in raising their children. Third, the group of people on Myspace who have elected to be part of the Aerosmith fan club, but never formally communicate with each other. Each of these three organizations would constitute an online group, but also conveys a different type of social organization, expected benefit, and social norms.

Groups. Abstractly, a group is simply a collection of people and can be divided into two high-level categories: first, some are an extension of *social identification* [31], whereby individuals affiliate with organizational memberships, religious beliefs, gender, age, or other cohorts. These types

[*] This work was done while the author was visiting Yahoo! Research.

WSDM'08, February 11–12, 2008, Palo Alto, California, USA.
Copyright 2008 ACM 978-1-59593-927-9/08/0002 ...$5.00.

of groups are most popular in social networking applications such as Orkut, Facebook, and LiveJournal, and while they are termed groups, they do not always imply group communication. The second class of online groups is more about *structured communication* [22]. These groups are built around communication, i.e., social support, political debate, civic engagement, or the discussion of specific interests.

As a group continues to communicate, the relationship between the individuals and the group change over time. Group norms develop between the members, typically made through explicit statements by core members, critical events, and behaviors from other groups [11]. Additionally, the personal relationships between members evolve, the strength of which is determined by atomic interactions [15, 21].

Key questions. In this paper we investigate the tenets of deep social engagements and evolving relationships between members within structured communication groups online. We are interested in studying the following questions.

(1) *Engagement.* We consider methodological questions around the characterization of engagement in online groups. How should we define engagement within an online group? How do activity levels and user engagement vary over time? What does the ecosystem of engaged users in a group look like?

(2) *Relationships.* What is the relationship between users, groups, and engagement? Does a user have a positive experience in a group and become increasingly engaged, or does a user destined for deep engagement behave differently upon arriving in a group than other newcomers do?

(3) *Modeling.* Is it possible to model the long-term heavy engagement of users in terms of parameters that are either user-dependent or group-dependent or both?

To answer these questions, we employ an immense corpus of online groups derived from the Yahoo! Groups product[1]. From these groups we extract a number of different group data: membership, message metadata, moderation (public, semi-private, and private), publicity (directory listed or unlisted), and more. From these aggregate statistics we derive a notion of engagement, namely the amount of activity necessary to be considered a regular contributor. This notion of "heavy" patronage is then used to observe and model user behavior, group dynamics, and intra-group user properties.

Main contributions. We present the first large-scale longitudinal study of user behavior in online groups. We formalize and employ operational definitions of lightly versus heavily-engaged users, and differentiate between short-term and long-term heavily-engaged users. We study the early, middle, and late experiences of these users within a group. Our hypothesis on beginning this work was that a new user in a group would join, issue a few tentative posts, receive some preliminary responses, and grow slowly in reputation and status into a state of fully-fledged leadership in the group.

To our surprise, we found that the reality is quite different. Users who will continue on to become heavily-engaged in the group, particularly in the long term, receive highly differentiated treatment *from the very first message they post.* At the same time, however, metrics of quality of experience such as probability that a message receives a response, and probability that a message receives a response from a central group member conditioned on receiving a response, actually decline as users remain heavily engaged in the group. Our main findings are as follows.

(1) Members who will become heavily engaged are twenty times more likely to receive a response immediately after joining.

(2) Upon becoming heavily engaged in a group, members who will remain engaged in the long term are almost twice as likely to receive responses to their messages.

(3) Consider two newly-joined members, one who will be a light user, and another who will become a heavily-engaged long-term user. If both receive a response to a message, the latter response is nine times more likely to come from a central member of the group.

(4) When a heavily-engaged long-term user joins a group, probability of receiving a response to a message post increases for a time, until the user becomes a central part of the group, and then begins to decline. Thus, likelihood to receive a response is non-monotonic over time.

In addition to these findings regarding users, we also present a number of results about the structure of the central core of a group, and about the engagement behavior of users across groups. Finally, we present some modeling results to study the impact of per-user and per-group indicators of long-term heavy engagement.

2. RELATED WORK

Social scientists have long since been interested in groups: why they exist, how they start, properties that govern their growth and decline, roles members play, and other properties. For a good survey of this literature, see the book by McGrath [22].

Early work in online community analysis focused on both Usenet [13, 20, 29, 33, 6], listservs [7, 25], and email groups in the workplace [12, 17]. The relationships produced by online groups range from strong ties, as with social support groups [2] to weak ties, as with fan groups [4]. These exchanges result in both real-world relationships and latent social ties that can be activated later [16].

Research in engagement focused early on non-participant observers, who have been cast in a particular light by the pejorative term "lurker" [24, 25] commonly used in the Usenet vernacular and academic literature [10]. Recently the question of engagement has been revisited as a problem of creating incentives for users to engage in community applications [5, 27], and in some cases specifically structured communication groups [32]. Postmes et al [26] study the aspects norm conformity in online groups. The role of identity motives in shaping online behavior is discussed in [10, 23]. David and Turner [9] study the influence of social identity concerns and beliefs of an individual in determining the response behavior to messages.

Many descriptive models have been proposed for describing the activity within public groups including demographics [33], information overload [20], tenure and interactivity [13], referential information [28], structural features [29], member roles [14], and resource availability [8]. Some predictive models have also been suggested for groups on social websites such as Orkut [30] and LiveJournal [3], but these groups are largely formed for social identification purposes, and the processes governing their dynamics are assumed to

[1] These data also predate the purchase by Yahoo!, and include groups as old as 1997, including one now-defunct group called "Craig's List."

be different.

Several papers have studied the flow of information in social networks and social groups. Wu et al [34] investigate the observation that messages relevant to one person is more likely to be relevant to others in the same social circle. See also the paper by Huberman and Adamic [19].

Our primary differentiating factors from this previous work comes from the type of media; while some studies have looked at communication behavior in private and semi-public groups, none has done so at scale. The Yahoo! Groups corpus presents a unique opportunity to study the full range of interaction behavior for all types of structured communication groups.

3. DATA

Unlike Usenet and public listservs, many structured communication group software typically support varying degrees of privacy and moderation for group members. In this respect, one could create a group for their extended family that would be virtually invisible to anyone but group members. The corpus we are working with comes from Yahoo! Groups, a service that represents the entire scope of privacy and moderation settings, and at a scale that is unmatched by other services. In this section we give a detailed characterization of our data, to convey metrics that are indicative of the level of engagement of our groups and users. In the following section, we present our analysis based on this data.

3.1 Yahoo! Groups

Yahoo! Groups began as an email list service named eGroups in 1997. When it was acquired by Yahoo in 2000, it supported 18 million users who were exchanging upwards of one billion messages per month [1]. Subsequently, the product was renamed to be Yahoo! Groups, and has continued to grow. The current product contains upwards of 100 million distinct users and six million groups.

A Yahoo! Group may be created by any Yahoo! user, and this user becomes the first moderator of the group. Moderators have three families of capabilities.

(1) To control various aspects of the presentation of content within the group.

(2) To provide settings that control the privacy level of the group and other aspects of the workflow.

(3) To control the day-to-day operations of which behaviors are allowed: moderators may control group membership, ban problematic members, require pre-approval on all posts, remove objectionable posts, and so forth.

Content within Yahoo! Groups has many forms, including information pages within the group, multimedia content, and message boards. The majority of content resides in message boards, and we focus our attention on that capability. Any member may post a message on a fresh topic, or in reply to a message posted earlier. Users who belong to the group may consume message content either online, or by signing up to receive posts through email. There are roughly six million distinct groups, containing roughly six billion individual postings.

For each group, we obtained all the messages posted in the group. For each message, we have the (anonymized) user who posted this message, the time of posting, and a pointer to the original message if this message was in response to the original message.

Group sizes: small, medium, and large. When we wish to analyze groups based on size, we establish three categories — small, medium, large — based on the number of posters in a group.

– *Small groups*, with fewer than 20 unique posters.

– *Medium groups*, with 20–99 unique posters.

– *Large groups*, with 100 or more unique posters.

3.2 Privacy structure in Yahoo! Groups

We now give a brief description of the privacy structure of Yahoo! Groups, as it will be relevant to our analysis.

Listed and unlisted. Each group may be listed or unlisted, with the following characteristics.

– *Listed groups* may be discovered through either navigation of the hierarchical groups taxonomy, or through group search.

– *Unlisted groups* do not show up in the groups taxonomy, and are not visible in search results. Such groups are typically discovered either because a member sends an invitation, or because a URL to the group is posted elsewhere on the internet.

Open, restricted, and closed. Additionally, each group may be either open, restricted, or closed.

– In *open* groups, there is no access control, and nonmembers may read and post messages.

– In *restricted* groups, messages may be posted and consumed only by members, but once a user visits an online web page to request membership, the membership is automatically granted.

– In *closed* groups, messages may be posted and consumed only by members, but membership is not automatic. A group moderator must approve a request for membership before the requesting user is granted any access.

Public, semi-public, and private. We performed a study of posting characteristics in the six different types of groups given by product

$\{\text{listed, unlisted}\} \times \{\text{open, restricted, closed}\}$.

Since giving the full set of results for the six different types will be overwhelming to the reader, we abstract the six types of groups into three natural categories, where the categories were created based on user behavior. The categories are, namely, public groups, semi-public groups, and private groups.

– *Public groups* correspond to the groups that are either open and listed, or open and unlisted.

– *Semi-public groups* correspond to the groups that are restricted and listed.

– *Private groups* correspond to groups that are either closed and listed, closed and unlisted, or restricted and unlisted.

4. THRIVING GROUPS AND CORE USERS

The variation in activity across these data is vast, ranging from highly active, massive conversations to irregular conversations or abandoned groups. In this section we develop two key concepts — thriving groups and core users — to help quantify this range of activity. We then study Yahoo! Groups data using these concepts.

4.1 Basic definitions

Our goal in this paper is to study groups that show both ongoing and high activity levels. To enable this, we develop the notion of thriving groups.

Definition 1 (Thriving groups). *We say that a group is* thriving *if it satisfies the following three criteria.*
(i) Baseline traffic. *For a one-year period,*[2] *the group must have at least two messages posted during every 30-day interval.*
(ii) Baseline users. *At least ten distinct users must post during the year.*
(iii) Dense period. *The year must contain a two-month period during which every seven-day interval has at least ten posts.*

The baseline traffic and baseline users requirements ensure that the groups are alive for a long enough period to enable longitudinal study. The dense period requirement ensures that the group has had significant activity in terms of messages and posting.

We have now developed a notion to capture high activity groups. We turn next to the behavior of users within those groups. As our goal is to study users with high levels of social engagement in the group, we must introduce actionable definitions of such users. To this end, we develop the notion of core users of a thriving group. We define the k-*core* of a group at time t as follows.

Definition 2 (k-core). *A user belongs to the* k-core *at time t if he/she satisfies the following the two requirements within the two week period centered at t:*
(i) the user has replied to $\geq k$ other distinct users, and
(ii) the user been replied-to by $\geq k$ other distinct users.

Note that we intentionally allow overlap in these two sets of k users. One should think of a person as being in the core if he/she has reached out to multiple people, and in turn has had multiple people respond to him/her, all within a brief period of time.

4.2 Analysis

Thriving groups. The initial dataset supported approximately 6.3M distinct groups. By adding the dense period requirement, this set shrinks to about 77,409 groups and 1.3M users. Adding the baseline traffic and baseline user requirements reduces the set again to 44,473 thriving groups, encompassing about 1M users.

Out of this almost 44K thriving groups, we have roughly 22K large, 13K medium, and 1K small groups in terms of size. In terms of privacy status, the decomposition is: 5K private, 13K semi-public, and 19K public groups.

Figure 1 shows the cumulative distribution of the unique poster count in a group during the period of interest. The figure shows that roughly half the groups have under 100 unique posters during the year, and very few have more than 1000 unique posters. In fact, 13% of thriving groups have fewer than 20 unique posters.

Structure of core users. To begin our study of k-cores, Figure 2 shows the cumulative distribution of k-core sizes for $k \in \{1, 2, 3\}$. This figure is computed for every group and every day of the year. Very few group/time pairs have more than twenty users in the 2- or 3-core, and between 1/4 and 1/3 of all group/time pairs have an empty core. Thus, for groups with a core, the core typically ranges from zero to ten users, with perhaps 10% of groups having from 10 to 20 users in the core. Even if the group is enormous, the core size is manageable. By definition, the users in the core will show up frequently in discussions in the group, and we may naturally view them as being highly visible users who would be well-known to consumers of the group's message content.

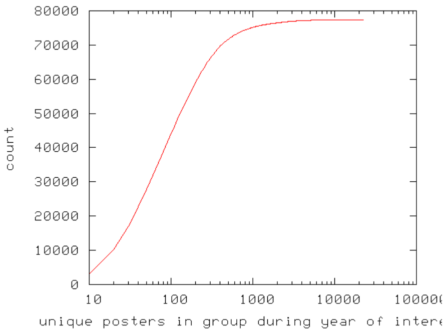

Figure 1: Cumulative distribution of number of unique posters in a group over a one-year period, over all groups.

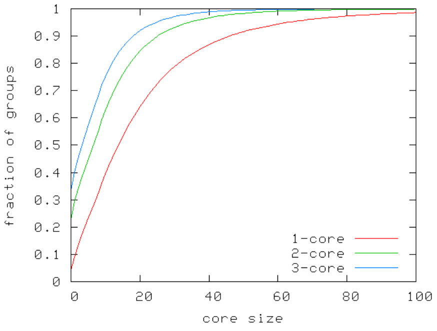

Figure 2: Cumulative probability of number of groups with a given core size, for $k = 1, 2, 3$.

Although thriving groups are filtered from all groups using only measures of activity, for most times and most groups, there is a non-trivial core of engaged users posting and responding to posts. For instance, 48% of group–time pairs have a 2-core of at least 6 people and 52% of group–time pairs have a 2-core of at least 7 people. This implies some significant response structure, and gives us some confidence that the engaged users we are studying are actually participating in a meaningful social environment.

Behavior of core users. We turn our attention now to the behavior of users in the core. We break out this behavior based on the privacy status of the group (public, semi-public, private) and the size of the group (small, medium, large). Figure 3 shows for each of our nine conditions the distribution of the number of days a user remains in the core. Small groups show very different behavior, with users remaining in the core for much longer than larger groups. Moreover, in private groups, people tend to belong to the core for longer than in public groups.

Private and semi-public groups behave quite similarly in

[2] We study the year from 5/15/2005 to 5/15/2006.

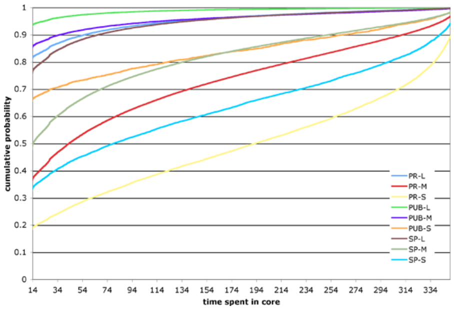

Figure 3: Cumulative distribution of number of days before a user who joins the core remains in the core, for all groups and all core users in that group.

terms of fraction of users in the core, except in the case of small groups. The results are pulled out in Table 1, presented below.

	Small	Medium	Large
Public	.40	.65	.85
Semi-public	.20	.30	.58
Private	.12	.26	.60

Table 1: Average fraction of posters not in the core.

Figure 4 shows for all groups how long users tend to stay in the core, computed as follows. For each group, we compute for each user u a value $d(u)$ representing the number of days u is in the core of the given group. Then for each time t, we compute U_t, the set of users in the core at time t,and produce a sorted list of the values $d(u)$ for each $u \in U_t$. We extract the entries at the 10th, 50th, and 90th percentiles from this sorted list. The figure shows for each percentile the distribution of values at that percentile over all groups and all times. It should be read as follows. Looking at the point halfway along the x-axis shows that for about half of all group/time pairs, 10% of people in the core are there for a very brief period, and 10% are there for almost the entire year, but the median user is there for about 100 days. We conclude that while core users in the longest-lived 10% may routinely last for several months or the better part of a year, nonetheless, the median core user is very unlikely to be around for more than 50 days.

That said, Figure 5 shows how long it takes on average for half the membership of the core to disappear from the core. This figure shows that within 50 days almost all cores have changed substantially.

Classes of group users. Based on these results, and a detailed analysis of Figure 3, we select 50 as the cutoff point for users who have significant ongoing interaction in a group. This gives us a natural categorization of the users in a group into three classes — LIGHT, SHORT-CORE, and LONG-CORE.

– LIGHT: A LIGHT user is one who is not part of the core.

– SHORT-CORE: A SHORT-CORE user is part of the core, but for fewer than 50 days.

– LONG-CORE: A LONG-CORE user is part of the core for at least 50 days

For thriving groups, the number of users of each category is the following.

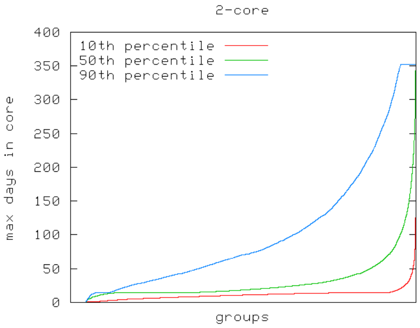

Figure 4: Cumulative count of number of users who are in the core of a group for a given number of days, over all members of all group cores.

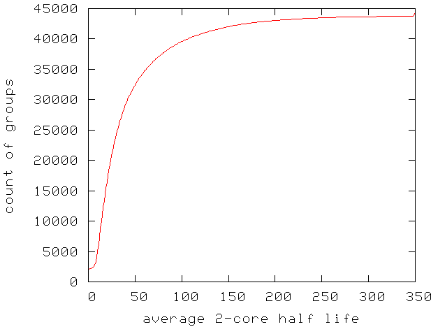

Figure 5: Cumulative distribution of number of days before half the users in a core have disappeared from the core, over all groups and all times.

LIGHT	SHORT-CORE	LONG-CORE	Total
774,493	133,507	89,966	997,966

It shows that most of the users are LIGHT, but there is a non-trivial fraction of SHORT-CORE and LONG-CORE users overall.

5. LONG-CORE USERS

We will begin our study at the level of users, trying to understand the behavioral characteristics of LONG-CORE users, either within a group or across groups. Subsequently, we'll move to a study of individual groups and explore the differentiated roles LONG-CORE users play therein, and the other members they tend to engage with.

5.1 Behavior across groups

We begin by asking whether users who are LONG-CORE in one group tend to be LONG-CORE in other groups they belong to. Figure 6 explores one view of this question. The figure shows the probability that a user's $(i + 1)$-st group will be LONG-CORE given that their first i groups are also LONG-CORE. This probability is monotonically increasing for the first 13 groups, indicating that becoming a LONG-CORE user is clearly a property of the person, rather than simply a property of the environment (see Section 7). By the 13-th group, the actual number of users is quite small, so the variation towards larger memberships should be considered less significant.

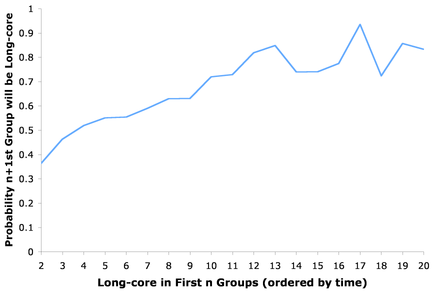

Figure 6: Conditional probabilities for additional long-core memberships.

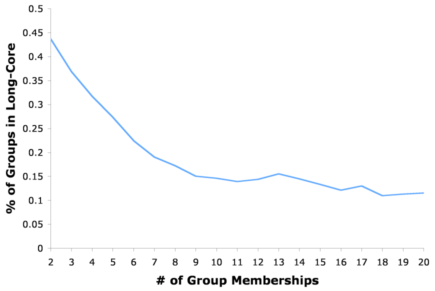

Figure 7: Ratios of long-core memberships for users with multiple groups.

Another view is given in Figure 7. The x-axis shows the number of groups to which an individual belongs. The y-axis shows for such individuals, the probability of being a LONG-CORE in one of those groups. The figure decreases almost monotonically from one group to around twenty groups, showing that users are much more likely to be LONG-CORE users if they belong to a smaller number of groups. A natural explanation is that users in fewer groups are able to focus their attention on those groups at the level necessary to become a LONG-CORE.

5.2 Size of core

Figure 8 shows the number of users in the core simultaneously. The mode is a core with a single user, with a slight local maximum at six, declining smoothly from there. The small spike at the far right of the graph is spurious. The conclusion is that even for very large groups, the core is almost always of manageable size; 90% of group/time pairs have a core of fewer than 25 people, and most are much smaller.

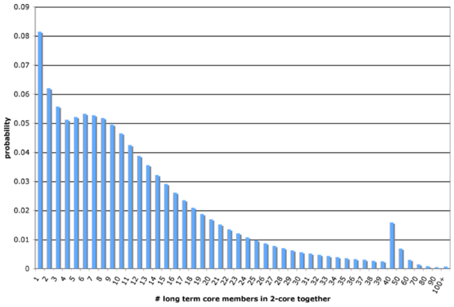

Figure 8: Distribution of number of users in the core simultaneously.

	Large	Medium	Small
LIGHT	425099	74890	2475
SHORT-CORE	84103	18934	601
LONG-CORE	102787	41895	2086

Table 2: Populations of newcomers.

5.3 Response behavior

A key moment in the evolution of thriving groups comes when an individual decides to cross the threshold and join in the conversation of an existing group. At this moment their fate is undecided: will they become an integral part of the group, or will they simply make their statement and leave? To understand this behavior, we must look at different degrees of engagement, and the experiences that lead to these states.

We will now look at the experiences of new members who will eventually fall into one of three states: LONG-CORE, SHORT-CORE, and LIGHT. We are interested in understanding the experience of LONG-CORE users at the moment they join a group, then later when they join the core of the group, and finally when they have spent 50 days in the core and formally become LONG-CORE users in that group. Likewise, for SHORT-CORE users, we wish to study their experience upon joining, and upon entering the core. And for LIGHT users, we study their experience upon joining the group.

To avoid conditioning, we restrict our attention to users who both (1) posted their first message in a given group during our sample window and (2) posted at least 20 messages during this time. Through these individuals we can observe what effects lead towards LONG-CORE engagement. Table 2 shows the number of users who meet both our requirements above and hence form the population of this experiment. We omnisciently define a user as a LONG-CORE user if they will eventually become a LONG-CORE user for the group.

For each of these member types, we observe the first 20 messages posted to the group, we tag the message as either initiating a new thread or responding to a message within an existing thread. We record the engagement group of the user who first responded to the message (if any), and how many days elapsed before the first response arrived. Table 3 shows the probability that a member receives a response given the category that they will eventually join.

The raw results for this experiment are show in Table 4. The table should be read as follows. RESP indicates probability of response. cRESP-{G,P,Y} indicates probability of response from a LONG-CORE, LIGHT, or SHORT-CORE user respectively, conditioned on a response being given. All results here are for messages posted by an original user in an existing thread, rather than messages than initiate a new thread. Each section category indicates the privacy status and size of the groups in that section: PR indicates private, PUB indicates public, and SP indicates semi-public; S, M, and L indicate small, medium, and large. The column headings should be read as follows: "new" indicates the first 20 messages posted by a user. "in-core" indicates the first 20 messages posted by the user after joining the core (valid only for LONG-CORE and SHORT-CORE users). And "in-core-50" indicates the first 20 messages posted by the user after being in the core for 50 days (valid only for LONG-CORE users).

There are a number of conclusions to be drawn from the data of Tables 3 and 4. We detail these next.

Privacy-Size	LONG-CORE			LIGHT	SHORT-CORE	
PR-L:	in-core-50	in-core	new	new	in-core	new
RESP	0.3121	0.3348	0.3289	0.0147	0.2135	0.2077
cRESP-G	0.8508	0.8432	0.8429	0.1266	0.7116	0.7099
cRESP-P	0.0327	0.0366	0.0378	0.6886	0.0832	0.0911
cRESP-Y	0.1165	0.1202	0.1192	0.1848	0.2052	0.199
PR-M:	in-core-50	in-core	new	new	in-core	new
RESP	0.2994	0.3181	0.3136	0.0213	0.2442	0.2385
cRESP-G	0.9478	0.9419	0.9415	0.3959	0.8764	0.882
cRESP-P	0.0082	0.0094	0.0097	0.4969	0.0204	0.0207
cRESP-Y	0.044	0.0487	0.0488	0.1072	0.1032	0.0973
PR-S:	in-core-50	in-core	new	new	in-core	new
RESP	0.3079	0.3168	0.3142	0.0581	0.2399	0.2385
cRESP-G	0.9808	0.9817	0.982	0.8372	0.9639	0.9656
cRESP-P	0.0036	0.0035	0.0035	0.0465	0.0038	0.0057
cRESP-Y	0.0156	0.0148	0.0145	0.1163	0.0323	0.0286
PUB-L:	in-core-50	in-core	new	new	in-core	new
RESP	0.301	0.3258	0.3203	0.0149	0.1969	0.1888
cRESP-G	0.8447	0.8375	0.8383	0.0939	0.6794	0.6781
cRESP-P	0.034	0.0403	0.0409	0.7384	0.0954	0.1065
cRESP-Y	0.1213	0.1222	0.1208	0.1677	0.2251	0.2154
PUB-M:	in-core-50	in-core	new	new	in-core	new
RESP	0.3093	0.328	0.3239	0.0122	0.2375	0.2306
cRESP-G	0.947	0.9409	0.9407	0.1735	0.858	0.8691
cRESP-P	0.009	0.0098	0.0098	0.7081	0.0324	0.0289
cRESP-Y	0.044	0.0493	0.0494	0.1184	0.1096	0.1019
PUB-S:	in-core-50	in-core	new	new	in-core	new
RESP	0.3043	0.322	0.3145	0.0239	0.2275	0.2227
cRESP-G	0.9818	0.9817	0.9829	0.7553	0.9549	0.9639
cRESP-P	0.0013	0.0022	0.0016	0.2021	0.0077	0.0049
cRESP-Y	0.0169	0.0161	0.0155	0.0426	0.0374	0.0312
SP-L:	in-core-50	in-core	new	new	in-core	new
RESP	0.3107	0.3333	0.3281	0.0156	0.2199	0.2146
cRESP-G	0.846	0.8409	0.8419	0.1501	0.7225	0.7214
cRESP-P	0.0333	0.0379	0.0379	0.6738	0.0785	0.085
cRESP-Y	0.1207	0.1212	0.1202	0.1762	0.199	0.1936
SP-M:	in-core-50	in-core	new	new	in-core	new
RESP	0.3088	0.3285	0.3237	0.0178	0.2486	0.244
cRESP-G	0.9487	0.9401	0.9402	0.3651	0.8785	0.8861
cRESP-P	0.0079	0.0096	0.0098	0.5002	0.0226	0.0206
cRESP-Y	0.0433	0.0503	0.05	0.1347	0.0988	0.0933
SP-S:	in-core-50	in-core	new	new	in-core	new
RESP	0.3124	0.3302	0.3266	0.0326	0.2768	0.2698
cRESP-G	0.9876	0.9808	0.9809	0.75	0.9513	0.9563
cRESP-P	0.0015	0.0019	0.0021	0.2365	0.0041	0.0021
cRESP-Y	0.0109	0.0173	0.017	0.0135	0.0447	0.0416

Table 4: Raw data for all user engagement levels and all group sizes; see text for details on naming conventions.

6. DIFFERENTIAL TREATMENT TO LONG-CORE USERS

Differential response behavior. The first is that these data show a startling differentiation between the treatment given to future LONG-CORE and SHORT-CORE members than LIGHT ones. In most cases, a new member who will eventually be engaged has 20 times the probability of getting a response early on. Further, all three types of users are treated differently immediately upon joining the group, and LONG-CORE and SHORT-CORE users are treated different when they join the core. This difference is dramatic. Probabilities of response for thread-initiating posts range from 38–43% for LONG-CORE users at various points in their life, 22–25% for SHORT-CORE users at various points, and 2% for LIGHT users.

Given the great disparity between getting a response, we also expect that this response will come from varying subsets of members. Figure 5 shows the pairwise probabilities for getting a response from a given member type for groups of the Public-Large variety. As one would expect, LONG-CORE members do a majority of the work, covering most of the interaction for newcomers *except* in the case that the newcomers will not ever join the core. These members largely receive their response from other LIGHT members, or from

	LIGHT	SHORT-CORE	LONG-CORE
Public Large	0.019	0.218	0.407
Public Medium	0.016	0.270	0.410
Public Small	0.033	0.291	0.420
Semi-Public Large	0.019	0.235	0.403
Semi-Public Medium	0.022	0.285	0.409
Semi-Public Small	0.028	0.340	0.455
Private Large	0.018	0.235	0.408
Private Medium	0.027	0.287	0.401
Private Small	0.045	0.318	0.440

Table 3: Probability of response.

	Newcomer Type		
Responder	LIGHT	SHORT-CORE	LONG-CORE
LIGHT	0.740	0.151	0.055
SHORT-CORE	0.169	0.266	0.110
LONG-CORE	0.091	0.583	0.836

Table 5: Probability of response by member type for Public-Large groups.

SHORT-CORE ones. This suggests that at a very early stage, members are performing some assortative mixing into subgroups: on one side we have members who will join the core at some point, and others who will not.

And conditioned on a response arriving, the probability that it comes from a LONG-CORE user is 84% for LONG-CORE users, 58% for SHORT-CORE users, and 9% for LIGHT users. So clearly our three classes of users see very different treatment, perhaps due to who they are, and perhaps due to how they behave.

At the same time, however, the response to LONG-CORE users at the moment they join the group appears almost identical to the response to such users as they enter the core, or at the moment they become LONG-CORE users (at least fifty days later). This similarity is even more striking given the great disparity in response rates among the user classes. It appears that LONG-CORE users join the group with their status already determined.

Plausible hypotheses. One hypothesis is that, when the behavior of these users was first measured, they were already longstanding members of the group. We corrected for this possibility by considering only users who had never posted to the group before the beginning of the year we studied.

A second natural hypothesis is that these LONG-CORE users might join an online group of friends whom they already know well offline. For instance, a user might join a group of old high school friends, and might immediately interact with the group as an insider. We performed an experiment to test this hypothesis. We collected 100 random instances of a LONG-CORE user joining a public group, and visited the group to determine whether a relationship existed between the new member and the existing group members, prior to the member joining. The results are shown in Table 6.

Category	Count
Friends with group members	2
Unknown to group members	13
Impossible to decide	20
Total	35

Table 6: Counts of status of relationship between group members and long-core user upon moment of joining group.

The row labeled “Impossible to decide” arises because in many cases, the first post does not contain any information to indicate that the poster knows or does not know the people in the group. For the “Friends with group members” case, there is information within the first post to indicate that the user was already friends with members of the group. Likewise, for the “Unknown to group members” case, there is clear evidence that the newcomer is a stranger introducing himself or herself to the group members.

Given that over 1/3 of users are strangers to the group, we may conclude that if the strangers were to be greeted with an experience akin to the LIGHT users, the results would be clearly visible in our response probabilities, differentiating between LONG-CORE users upon joining, and LONG-CORE users after spending significant time in the core; this differentiation does not exist, leading us to conclude that users whose interaction modalities lead them to become LONG-CORE users immediately receive differentiated treatment in the form of faster and more frequent responses, and a larger fraction of responses from LONG-CORE users, even when they are still strangers to the standing membership of the group.

To restate this conclusion, heavily-engaged users are treated as such from their earliest moments in the group, either because of their personal characteristics, or because of their fit with the group, or some combination. The standard model by which we would anticipate that a new user joins a group and rises to influence is that the new user joins, issues a few tentative posts, receives some preliminary responses, and slowly grows in reputation and status into a state of fully-fledged leadership in the group. This model does not appear to describe LONG-CORE users in our data; such users may be said to be "born" at their inception into the group, and not "made" over time.

More observations. We may observe a secondary take-away from the results of Table 3 and Table 4, and the underlying data form which it was generated. Namely, for all sizes, and all privacy levels, and whether responding to a message or initiating a thread, LONG-CORE users are more likely to receive a response at the moment they join the core than the moment they join the group, and both these probabilities are larger than the probability of receiving a response after becoming LONG-CORE users. The differences are much less dramatic than the differences between users, but are nonetheless very consistent. It is natural to postulate mechanisms by which a user would be more likely to receive responses after being in the group for a while, but it becomes slightly more awkward to suggest natural mechanisms by which the probability of receiving a response is not monotonic through time, as is the case here. We do not have a conclusive explanation for the phenomenon.

7. MODELING LONG-CORE ENGAGEMENT

To shed some light on the nature of the data, we propose the following highly simplified perspective on LONG-CORE engagement. When a user joins a group, there are three factors at work. First, the user might intrinsically have a personality which causes the user to become a LONG-CORE user of every group she joins. Second, the group might be so welcoming, or its topic so engaging, that users joining the group are likely to become LONG-CORE users of the group. And third, the particular user might happen to "click" with the particular group, causing LONG-CORE engagement even though neither user nor group have a particular propensity towards this behavior. We now define a simple model to study the first two of these reasons, both separately and in tandem, and we attribute unexplained behavior to the third reason. This model is meant to be an analytical tool to explore the data, rather than a reflection of human behavior.

The model. The modeling problem we define is the following. The input is a 4-tuple consisting of a set $\mathcal{U}$ of users, a set $\mathcal{G}$ of groups, a set $E \subseteq \mathcal{U} \times \mathcal{G}$ of memberships, and a set $H \subseteq E$ of LONG-CORE memberships. The goal is to reproduce the LONG-CORE memberships using the first two mechanisms above. Formally, we assign to each user a propensity $p(u)$ to become a LONG-CORE user when joining a group. Likewise, we assign to each group a propensity $p(g)$ for a user joining the group to have LONG-CORE engagement. When a membership (u, g) arrives, the user flips a $p(u)$-biased coin and decides to be LONG-CORE for the group if the coin lands heads. Simultaneously, the group flips a $p(g)$-biased coin and decides that the user will be LONG-CORE for the group if the coin lands heads. The membership will be LONG-CORE if either coin flip succeeds. Thus, $\Pr[(u, g) \in H] = 1 - (1 - p(u))(1 - p(g))$. Each membership is evaluated independently according to this rule. In discussion, we will consider other possible rules to assign memberships to H; for now, we adopt this "OR rule."

The goal of the modeling task is therefore to select propensities $p : \mathcal{U} \cup \mathcal{G} \to [0, 1]$ so as to reproduce H as exactly as possible. Formally, the goal is to provide a function $p : \mathcal{U} \cup \mathcal{G} \to [0, 1]$. We evaluate the quality of a function $p(\cdot)$ by its likelihood of producing the correct assignment of memberships to H and $E \setminus H$. The quality of a function $p(\cdot)$ is therefore given by

$$\prod_{(u,v) \in E \setminus H} (1 - p(u))(1 - p(g)) \prod_{(u,v) \in H} 1 - (1 - p(u))(1 - p(g)) .$$

Model variants. We consider three variants of the model. In the first, we set $p(g) = 0$ for all groups, allowing the model to employ only properties of users; we consider how effectively this model captures the observed data. Next, we allow $p(g)$ to be arbitrary, but fix $p(u) = 0$, allowing only groups to influence when a membership becomes LONG-CORE. In both these cases, the optimal likelihood is trivially solvable. Finally, we allow both $p(u)$ and $p(g)$ to be arbitrary; in this case, we iteratively solve optimally for $p(u)$ given a fixed set of $p(g)$'s, then likewise resolve for $p(g)$ given a fixed set of $p(u)$'s, and so forth. This procedure converges to a solution with the same likelihood for many different iterations of the parameters, and is guaranteed to show monotonically non-decreasing likelihood.

We apply this model to the thriving groups described above, which contain around 44K groups and 7M users. In this data, about 17.6% of memberships are LONG-CORE. The following table shows the percent of time each edge is correctly assigned:

Model	% of correct edges
User-only ($p(g) = 0$)	94.9
Group-only ($p(u) = 0$)	85.6
Combined	95.1

As the table shows, the ability to set individual variables for each of 7M users allows a very accurate model, correctly explaining roughly 95% of the memberships. Using only the 44K per-group variables results in a much weaker fraction of edges correctly modeled: 85.6%. However, much of the difference comes from the 6M users who belong to a single group, who are modeled perfect in the user-only case, and imperfectly in the group-only case. A more detailed analysis of the values, and the behavior of the model on users who belong to multiple groups, shows that both per-user and per-group information are important for accurate modeling, as we now describe.

Figure 9 shows the distribution of values of $p(u)$. The spiky red line marked with plus tickmarks indicates the values when $p(g)$ is constrained to be zero; this represents a baseline. The regular spiky pattern in the graph shows an increase in counts at values of $p(u)$ corresponding to multiples of small reciprocals: 1/2, 1/3, 2/3, etc. These occur because the optimal solution assigns $p(u)$ to be the fraction of edges incident to u that are LONG-CORE, and thus low-degree nodes with d neighbors fall into one of the small number of buckets corresponding to a multiple of $1/d$. The graph shows that no single $p(u)$ value predominates; the

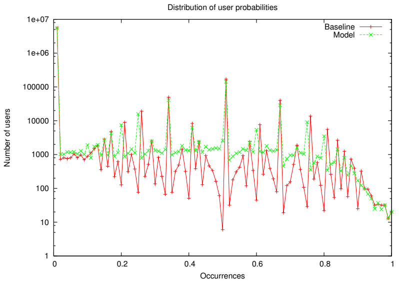

Figure 9: Distribution of number of users in the core simultaneously.

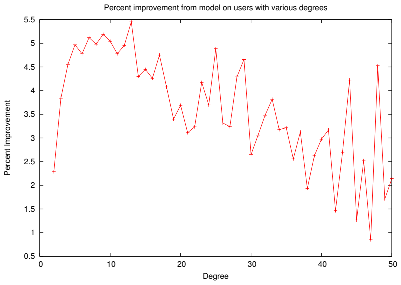

Figure 10: Improvement possible over user baseline by incorporating per-group information.

optimal assignment of values places similar mass within all deciles of probability.

The green curve marked with cross tickmarks shows the same results when groups are allowed to take on non-zero $p(g)$ values. In this case, the addition of some 44K additional degrees of freedom on top of the existing 7M completely changes the picture. The presence of non-zero values for $p(g)$ causes significant smoothing to occur. Figure 10 shows the improvement in the average log-likelihood of an edge, as a function of the degree of the left (user) endpoint of the edge. Users who belong to a reasonable number of groups, say 5 or more, show a significant increase in modeling quality.

8. CONCLUSIONS

In this paper we have investigated the social dynamics of one of the world's largest collections of online communities. Due to the massive scale and breadth of behavior, we have proposed a partitioning on the data that selects for active communities of engaged individuals. These *thriving* groups were then further examined to identify different levels of engagement: LONG-CORE, SHORT-CORE, and LIGHT. We have found that varying types of groups produce varying degrees of engagement: the average member of a small, private group will be much more engaged than a member of a large, public one.

Looking more closely at individual LONG-CORE members across different groups they belong to, we observe a diminishing return on group involvement: the more groups a person belongs to, the less likely they can be heavily engaged in all of them. We finally explore the experience of newcomers to groups, and find an environment of assortative mixing: users who will eventually become LONG-CORE are receiving preferential treatment from other LONG-CORE users, while LIGHT users receive little to no attention from this group.

The findings of this paper suggest a number of key insights that might inform the design of future community systems. First, the varying behavior of groups based on privacy and size suggest entirely different types of communication. While it is beneficial to researchers to have these all in one corpus, it might be beneficial to users to provide different interfaces based on the type of communication they wish to engage in. Second, we have observed the importance of long-term members of communities. For nearly all types of groups, these individuals hold the continuity that allow the group to maintain integrity. Tenure in a group core may be seen as an important variable to expose to new users, or in ranking systems that take advantage of individual user attributes. Finally, the preferential treatment of LONG-CORE users to newcomers who will be LONG-CORE users could be an important tool for filtering information in large groups, and focusing engaged members into smaller, more manageable interfaces.

Acknowledgments. We are grateful to the Groups team at Yahoo! for their support in data gathering, data analysis, and direction. In particular, we would like to thank Di-Fa Chang, Lee Clancy, David Kopp, Bobby Lee, Maria Saltz, and Gordon Strausse. We also thank the anonymous reviewers for helpful comments.

9. REFERENCES

[1] Wikipedia entry on egroups. `http://en.wikipedia.org/w/index.php?title=Special:Cite&page=EGroups&id=%109507115`.

[2] S. C. Alexander, J. L. Peterson, and A. B. Hollingshead. Help is at your keyboard: Support groups and the internet. In L. R. Frey, editor, *Group Communication in Context: Study of Bona Fide Groups*, 2nd ed, pages 309–334. Lawrence Erlbaum Associates, 2003.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. 12th ACM Conference on Knowledge Discovery and Data Mining*, pages 44–54, 2006.

[4] N. K. Baym. *Tune in, Log on: Soaps, Fandom, and Online Community.* Sage, Thousand Oaks, CA, 2000.

[5] G. Beenen, K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. E. Kraut. Using social psychology to motivate contributions to online communities. In *Proc. ACM Conference on Computer Supported Cooperative Work*, pages 212–221, 2004.

[6] C. Borgs, J. Chayes, M. Mahdian, and A. Saberi. Exploring the community structure of newsgroups. In *Proc. 10th ACM Conference on Knowledge Discovery and Data Mining*, pages 783–787, 2004.

[7] B. Butler. *When a Group is not a Group: An Empirical Examination of Metaphors for Online Social*

Structure. PhD thesis, Carnegie Mellon University, 1999.

[8] B. Butler. Membership size, communication activity and sustainability: A resource-based model of online social structures. *Information Systems Research*, 12(4):346–362, 2001.

[9] B. David and J. C. Turner. Studies in self-categorization and minority conversion: Is being a member of the out-group an advantage? *British Journal of Social Psychology*, 35:179–199, 1996.

[10] J. Donath. Identity and deception in the virtual community. In P. Kollock and M. Smith, editors, *Communities in Cyberspace*. Routledge, London, 1999.

[11] D. C. Feldman. The development and enforcement of group norms. *The Academy of Management Review*, 9(1):47–53, 1984.

[12] T. Finholt and L. Sproull. Electronic groups at work. *Organizational Science*, 1(1):41–64, 1990.

[13] A. T. Fiore, S. L. Tiernan, and M. A. Smith. Observed behavior and perceived value of authors in usenet newsgroups: Bridging the gap. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 323–330, 2002.

[14] D. Fisher, M. Smith, and H. T. Welser. You are who you talk to: Detecting roles in Usenet newsgroups. In *Proc. 39th Hawaii International Conference on System Sciences*, 2006.

[15] L. C. Freeman. The sociological concept of "group": An empirical test of two models. *American Journal of Sociology*, 98(1):152–166, 1992.

[16] C. Haythornthwaite. Strong, weak and latent ties and the impact of new media. *The Information Society*, 18(5):385–401, 2002.

[17] C. Haythornthwaite and B. Wellman. Work, friendship, and media use for information exchange in a networked organization. *Journal of the American Society for Information Science*, 49:1101–1114, 1998.

[18] J. Horrigan. Online communities: Networks that nurture long-distance relationships and local ties. Technical report, Pew Internet and American Life Project, 2001.

[19] B. A. Huberman and L. A. Adamic. Information dynamics in the networked world. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, editors, *Complex Networks, Lecture Notes in Physics*, pages 371–398. Springer, 2004.

[20] Q. Jones, G. Ravid, and S. Rafaeli. Empirical evidence for information overload in mass interaction. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 177–178, 2001.

[21] P. Marsden. Measuring tie strength. *Social Forces*, 63:482–501, 1984.

[22] J. E. McGrath and D. A. Kravitz. Group research. *Annual Review of Psychology*, 33(1):195–230, 1982.

[23] K. Y. McKenna and J. A. Bargh. Causes and consequences of social interaction on the internet. *Media Psychology*, pages 249–270, 1999.

[24] B. Nonnecke and J. Preece. Lurker demographics: Counting the silent. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 73–80, 2000.

[25] B. Nonnecke and J. Preece. Persistence and lurkers in discussion lists: A pilot study. In *Proc. 33rd Hawaii International Conference on System Sciences*, page 3031, 2000.

[26] T. Postmes, R. Spears, and M. Lea. The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3):341–371, 2000.

[27] A. M. Rashid, K. Ling, R. D. Tassone, P. Resnick, R. Kraut, and J. Riedl. Motivating participation by displaying the value of contribution. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 955–958, 2006.

[28] T. Schoberth, J. Preece, and A. Heinzl. Online communities: A longitudinal analysis of communication activities. In *Proc. 36th Hawaii International Conference on System Sciences*, 2003.

[29] M. Smith. Invisible crowds in cyberspace: Mapping the social structure of the usenet. In M. A. Smith and P. Kollock, editors, *Communities in Cyberspace*, pages 195–219. Routledge, 1999.

[30] E. Spertus, M. Sahami, and O. Buyukkokten. Evaluating similarity measures: A large-scale study in the orkut social network. In *Proc. 11th ACM Conference on Knowledge Discovery and Data Mining*, pages 678–684, 2005.

[31] H. Tajfel. *Human Groups and Social Categories: Studies in Social Psychology.* Cambridge University Press, 1978.

[32] S. J. J. Tedjamulia, D. L. Dean, D. R. Olsen, and C. C. Albrecht. Motivating content contributions to online communities: Toward a more comprehensive theory. In *Proc. 38th Hawaii International Conference on System Sciences*, page 193b, 2005.

[33] S. Whittaker, L. Terveen, W. Hill, and L. Cherny. The dynamics of mass interaction. In *Proc. ACM Conference on Computer Supported Cooperative Work*, pages 257–264, 1998.

[34] F. Wu, B. Huberman, L. A. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A*, 337:327–335, 2004.