Early Activity Diversity: Assessing Newcomer Retention from First-Session Activity

Raghav Pavan Karumur, Tien T. Nguyen, Joseph A. Konstan
GroupLens Research
University of Minnesota
{raghav, tien, konstan}@cs.umn.edu

ABSTRACT
Online communities suffer serious newcomer attrition. This paper explores whether and how early activity diversity – the degree to which a newcomer engages in a wide range of a site's activities in the first session – is associated with their longevity. We introduce a metric (DSCORE) to characterize early activity diversity in online sites and run our analyses on an online community 'MovieLens'. We find that DSCORE is significant both by itself and in conjunction with a measure of quantity of activity in predicting longevity. This finding is robust to different measures of longevity (aggregate number of sessions and attritions after sessions 1, 5, and 10). The immediate implication is an effective classifier for identifying users with higher (or lower) expected longevity from the first-session activity. We also find DSCORE is more useful than a traditional measure of measuring diversity such as the Gini-Simpson index. We conclude by discussing how early activity diversity may be more broadly effective in supporting design and management of online communities.

Author Keywords
Online communities; newcomer engagement; dealing with newcomers; activity; first session; activity diversity; early user experience; newcomer retention; longevity.

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: web-based interaction; H.1.2 Software psychology; H.5.2 User Interface; H.m Miscellaneous; K.4.3 [Computers and Society]: Organizational Impacts.

INTRODUCTION
In this paper, we study the relationship between an online user's diversity of early activity and their retention. Newcomers are essential to online communities not only because they replace others who leave, maintaining the critical mass [23,32], but also because they serve as sources of new energy, activity, and innovation. Retaining newcomers is hard, though, for their connection with the community is fragile. Studies of several communities show that a loss of about 60% of newcomers after their first session is not uncommon [1,11,31,45].

Predicting whether or not a newcomer returns is useful because acquiring new users is, in general, expensive compared to retaining old ones [35,36]. However, an inherent problem with predicting *new user churn* is that, not much previous activity history is available and often, demographic information is incomplete. Therefore, we look at early (first-session) activity diversity in an attempt to use information that is available early to assess newcomer retention in a way that can also generalize across different communities.

Our investigation of activity diversity is motivated by prior research that newcomers are happier and stay longer if they have a complete picture of the community while joining [2]. During their early interaction with the community, they investigate and evaluate it on a variety of dimensions to see if it fits their needs. They decide whether to invest effort in it or move on to explore other alternatives. If they find it suitable, they join and remain in it longer [6,24,25,26,33]. While some online communities provide access to their archived content without the need to join, others require that the users login to see what it has to offer. In either case (particularly, the latter), it is evident that their exploration of the community's features during their first login session can affect whether they leave for good or return for a second session.

In this paper, we introduce a metric called DSCORE to characterize this early (first-session) activity diversity. We had three specific goals that led us to develop a new metric instead of using one of the popular [9,17,19,37,38] diversity metrics. First, based on our grounding in the "complete picture," we wanted a diversity metric that focused on exposure and not quantity – the metric should ignore repetitions of an activity and consider only the breadth of activities a user tries. Second, we wanted to measure diversity in a manner that recognizes that different activities may be more or less similar, awarding higher diversity scores to sets of dissimilar activities. Third, we wanted a metric that would generalize to communities with different activity structures, and that would in turn scale to

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA
 ISBN 978-1-4503-3592-8/16/02 $15.00
DOI: http://dx.doi.org/10.1145/2818048.2820009

different numbers of activities. The metric we introduce has these desirable characteristics and is based on a distance tree analysis of the online site's activities.

Thus, using DSCORE, in this paper, we explore the utility of naturally-occurring early (first-session) activity diversity in assessing new user retention in an online recommender community 'Movielens'.

Research Questions

We organize our research around the following questions:

RQ1*: How is early activity diversity (measured using DSCORE) associated with new user longevity?*

We first establish feasibility by showing a correlation between number of distinct activity types tried and new user retention. We then build and test successive models to explore the degree to which early activity diversity is associated with new user longevity, considering a variety of model types and additional factors such as overall quantity of activity. We use the model with the best fit to illustrate the increase in average longevity associated with marginal increases in a new user's first-session activity level and diversity.

RQ2*: How can we most effectively measure early activity diversity for purposes of predicting new user longevity?*

Once we have established the value of DSCORE as a predictor of user retention, it makes sense to examine how it compares with more traditional metrics. We therefore compare the models built using DSCORE with the ones built using the Gini-Simpson index.

Movielens Dataset

We conduct this research using log data from the classic version of MovieLens (http://classic.movielens.org) from December 20, 2007 to January 1, 2014.[1] MovieLens allows users to rate and receive recommendations for movies. In addition, they can add, edit or tag movies; add buddies; or participate in other ways such as by answering questions about movies. The presence of multiple activity types and a large pool of users along with their activity logs from their very first interaction with the community (48,784 users in total) made this dataset useful for our research.

Contributions

We make the following contributions in this paper:

Early activity diversity predicts retention of new users. Based on an analysis of the usage logs of more than 48,000 users of Movielens, we find that activity diversity in the *very first session* is a significant predictor of new user retention. We also show that diversity adds significant value when combined with measures of activity level, with both measures helping predict new user retention over 1, 5, and 10 sessions.

DSCORE: A new and more effective diversity metric. We introduce DSCORE, a metric to measure early activity diversity in a general way based on a similarity tree classifying activity types. It is designed to isolate diversity from quantity of activity and can be applied to different sites that support multiple activity types. Also, we find in the context of Movielens that DSCORE is more useful than traditional measures capturing diversity such as the Gini-Simpson index.

Implications for design and research. We discuss implications both for designers and for researchers. Diversity can be used to customize experience based on predicted retention, or to assess and improve site design for engagement. Further research is proposed to isolate causal factors underlying the relationship between early activity diversity and retention.

Organization

The rest of the paper is organized as follows. In the next sections we discuss related work, followed by introduction of the novel metric (DSCORE) and its rationale, followed by our research design, assumptions made and the activity types considered for our analysis. We then present the results of our analysis followed by a discussion. We then conclude by discussing how early activity diversity may be more broadly effective in supporting design and management of online communities.

RELATED WORK

Newcomer Retention and Churn

Statistics

Online communities face heavy new user churn. 54% of newly registered developers never returned to the community after their first post in The Perl Open Source Development Project [11]. In Usenet groups, 68% of newcomers did not return after their first post [1]. Newcomers to Wikipedia have high probability of leaving within few days with only 40% of contributors continuing to use after 500 days [45]. Others found that 60% of newcomer editors never make another edit on Wikipedia after their first 24 hours. [31]. Therefore, studying churn in newcomers is valuable.

Prior work on factors that determine retention

User churn and retention have been studied in highly popular communities like Wikipedia [31], Yahoo! Answers, Naver Knowledge iN, Baidu Knows [10,46], Stack Overflow [34], and Massively Multiplayer Online Role Playing Games [3,22]. Yang et al. [46] looked at length of the first question posted by users to predict longevity. But such metrics fail to generalize to communities with other participation types without such attributes as content length.

[1] In November 2014, a newer version of Movielens was released with a lot of redesign and change in its overall structure and features. For this reason, we did not incorporate users who joined after January 1, 2014 into our analysis.

Some of these works used metrics based on activity history for a few sessions, weeks or months, or semantic attributes that capture general mood or immersion of the user across sessions until the point of analysis to predict user longevity. Since not much previous activity history is available about a new user, these are factors we cannot assess very well at the first session. Waiting for a few weeks of activity to do the analysis would mean running the risk of losing a vast majority of users until the point of analysis.

Others used demographic information about the user, but this may be either too sparse or not always available as most communities these days have minimal registration barriers with a one step signup process using their Gmail or Facebook accounts. In Movielens, age and gender were available for only 715 out of the 48,784 users we analyzed, and so we could not use these factors for prediction.

Some of the above works used overall time spent on site as a predictor. Most users multi-task and so, the exact time a user spends on the community of interest is hard to estimate. Also, users switch to other browser tabs or windows; or close browsers or tabs without ever logging out. Therefore, metrics based on time are often inaccurate representations of amount of user activity (despite sometimes showing moderate correlations with it) and therefore cannot be relied upon.

Some of the above works also used social influence of other users to predict a particular user's survival in the community. Again, for new users who are not necessarily well-networked yet, or for new users in communities which do not have an active social component, metrics based on social influence are not suitable.

Desires to volunteer online, help others, gain reputation, pursue shared values and beliefs, voice humanitarian concerns, develop careers, develop positive attitude and protect oneself from negative feelings; having previous experience; or just enjoying what the community does have all been identified to be factors that motivate contribution (and thus retention) in online content communities [5, 12,15,21,29,40,44]. Site policy changes, personal life changes, or a sense of feeling that they can no longer fulfill their perceived role in the community on the other hand may lead to user churn [47].

Many of these works had extensive user data available about these attributes for their analyses. But we are focusing on the specific challenge of *new users* for whom we have very little to no data about any of these attributes. With de-identified log information, we cannot contact individual users who stayed or left the system either. Therefore we do not have information regarding motivations and prior experience of individual users. Hence, we focus here on identifying the relationship, if any, between early activity diversity and retention, leaving questions of association in presence of other factors, causality and manipulability for future research.

Interventions

Prior work showed also that responding to a user's first interaction, eliciting feedback from them through lightweight tools, providing assistance and recommendations early on and properly welcoming them into the community improve user retention [4,7,8,14,20]. Also, commercial practice suggests that there is an interest in interventions aimed at new user retention. A lot of sites offer additional gifts, e-coupons, membership discounts, special promotions, free premium account access for extended periods, etc., in order to retain users who do not return for long durations of time. We therefore hypothesize user retention may improve when introduced to other types of participation.

Early Activity as a Predictor of Longer-Term Behavior

Even outside questions of churn, people have found value in early activity as a predictor of longer-term behavior. It was observed in an analysis of "power users" of Wikipedia that users' activity patterns, even in the earliest days, had an ability to predict future amount, quality and frequency of activity [31]. Also, Pal et al. [30] looked at the first few weeks of activity to detect experts in a community. Burke et al. [4] found that newcomers' exposure to different features on Facebook through the newsfeeds of their friends' activities moderately affects (positively) their future usage of those features. These works strengthen our interest in studying the relationship between measures based on early activity and future retention.

Diversity

Diversity in Online communities

Zhu et al. found that greater diversity in subgroup membership was associated with greater longevity of Wikia members [49]. However, specialization in participation type is most commonly found in online communities. Categories like 'lurker', 'Questioner', 'Answer Person', 'Uploader' and 'Contributor' have been identified based on specialization [27,28,41,42]. But these works did not look at the question of whether those who chose to specialize did so after being aware of the wide range of possibilities. Our measure – DSCORE is specifically designed not to penalize people who specialize after being aware of the alternatives. We hypothesize that someone who tires of their specialized activity will be more likely to be retained if they know there are other things they can fall back on.

Diversity and Community Success

In order to accomplish goals that are important to the community, some attempts have also been made to direct users to other opportunities even if they did not match their interest in the context of Wikiprojects. Examining weekly collaborations, Zhu et al. established that explicit setting of goals and implicit social modeling can help diversify a self-identified user's participation in such a way that tasks important to the community may be accomplished [48]. So,

we understand that diversity is a characteristic that can be nurtured in users, if we find value in it.

Diversity Metrics

Diversity metrics quantify distribution of entities across various available class types and have been studied extensively in biology, ecology and in social and informational sciences. Many diversity metrics have been proposed based on the need of the community under consideration. Richness [9], Shannon Entropy [37], Simpson index [38], Gini-Simpson index [17,19,38] are the most widely used ones. Definitions of diversity have varied widely based on what the proponents of those metrics assumed diversity to be. Diversity metrics in general deal with a richness component - characterizing the number of distinct class types the set of interest contains and an abundance component characterizing number of entities per class type - sometimes using both components, and sometimes just one of them.

Richness does not account for class hierarchies or similarities between entity (activity) types. Other metrics such as Entropy, the Simpson Index or the Gini-Simpson index have quantity of activity included in them. Because we are interested in the marginal value of diversity over quantity, we introduce a diversity metric that separates diversity from quantity of activity. To validate the usefulness of our new metric over existing ones, we redo our analyses replacing DSCORE with the Gini-Simpson index (which is popular in social psychology literature).

EARLY ACTIVITY AND EARLY ACTIVITY DIVERSITY

Identifying Activities

An activity is a single interaction with any feature of a particular online community and an 'activity type' refers to one of the several types of activities that exist in the community. For instance, on Movielens, a user could rate a movie 3.5 stars, or search using the tag "Animation", or use the "My Wishlist" feature. Each of these is a different activity type but the user has performed three activities in all. We occasionally use the term 'participation' to refer to an activity and the phrase 'participation type' to refer to an activity type.

Based on consultation with the site maintainers and other site experts (more than 20 researchers – faculty, former and current students involved in development of and research on Movielens), we classified the features of Movielens into 17 distinct activity types. A brief description of each activity type is shown in Figure 1.

1) *Edit profile* – The descriptor indicates that a user edited his profile or visited the edit profile page to make changes to his profile to represent himself to the community.

2) *Create an RSS feed* – User uses this feature to create an RSS feed for herself.

3) *Invite a buddy* – User uses this feature to invite a buddy to Movielens.

4) *Use help pages* – User uses this feature to learn more about and understand different features of the system.

5) *View movie detail* – User uses this feature for viewing complete details about a specific movie such as actors, directors, genres, language, ratings, a brief description of the storyline and tags applied to the movie.

6) *Search by tag* – User uses the feature to browse for a movie using a list of displayed tags.

7) *Search attribute / metadata* – User uses this feature to search for a movie by entering a search phrase or word. Feature 6 is different in that it does not let the user enter anything. The user can only click on existing tags to search for lists of movies.

8) *View "Most Often Rated" list* – User uses this feature to view the most often rated movies on Movielens.

9) *View "Top Picks for you" list* – User uses this feature to see a personalized list of movies recommended to him by Movielens.

10) *View "Newest Additions" list* – User uses the feature to see the list of new movies added to Movielens.

11) *View "Rate Random Movies" list* - User uses this feature to browse through a list of random movies in order to rate them.

12) *View "Your Wishlist"* – User uses this feature to see all the movies he has added to his wish list.

13) *View "Your Ratings" list* – User uses this feature to see all movies she has rated and their corresponding ratings.

14) *Rate* – User rates a movie on Movielens.

15) *Tag* – User tags a movie on Movielens.

16) *Participate in Q&A* – User uses the feature to participate in a Q&A discussion.

17) *Add / Edit movie* – User uses this feature to add a movie to Movielens or edit a movie on Movielens. The classic version of Movielens was structured such that the same control was used for both purposes.

Figure 1. A chart showing a brief description of each activity type on Movielens

Activity Metrics

To answer the question of how early activity diversity is associated with user longevity in the community, we need to separate diversity from quantity. So we introduce two metrics: DSCORE and ASCORE.

Early Activity Diversity Score (DSCORE) Early Activity Diversity Score for user *u* is a metric characterizing the number and degree of dissimilarity of distinct activity types performed by the user *u* in the *first* session based on the hierarchical ontological relatedness of these activity types. We denote it by DSCORE.

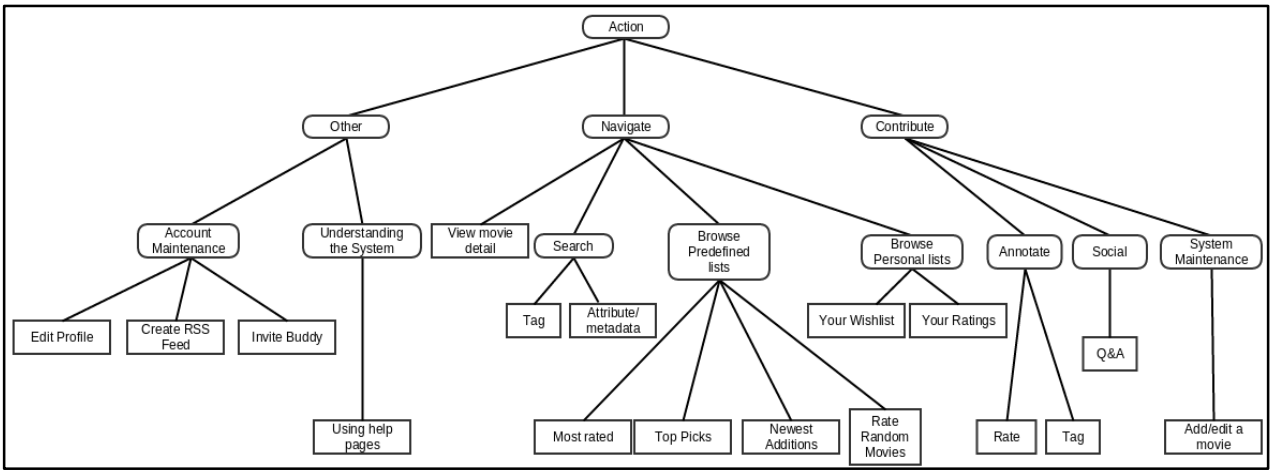


Figure 2. Classification of activity types in Movielens

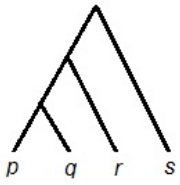
Design Challenge: The available activity types in an online system range from highly related (e.g., rate an item, tag an item) to fairly distant (e.g., invite a buddy, view a movie). While each is different, we want a measure of diversity that adequately reflects that carrying out three very different activities may have more diversity than carrying out four or five very similar ones. Intuitively, this is the same as we might find with biological diversity: a zoo with five different types of primate does not have as diverse a collection as one with a chimpanzee, a whale, and a lizard.

Our approach to this challenge is to build our diversity metric in a manner that is tied to a hierarchical taxonomy of activities – a taxonomy that is built specifically to group similar activities together and to separate dissimilar activities. Unlike many of the diversity metrics discussed in the related work section, this approach allows us to take hierarchical ontological relationship between entity types into account. We build upon simple 'richness' accommodating various degrees of dissimilarity between different activity types. So we first model the relationship between various activity types in a community.

Modeling relationship between activity types: One could interview the users of the community, analytically look at which participation types go together, or speak to experts or community moderators to understand the degrees of dissimilarity between various activity types. In our case, we engaged the experts in a card sorting activity [39], a standard usability technique used to understand the information architecture of a site. We asked them to cluster the activity types into as many natural clusters as would make sense to them and provide a brief explanation of why they believed in such architecture. The experts were also asked if they would further cluster them into smaller or larger clusters and some of them did. In the end, we had a tree that depicted the relationship between various activity types on Movielens (Figure 2).

In this tree, all distinct activity types appear as leaf nodes. Each internal node of such a tree represents a hypothetical activity type that encompasses all activity types represented as its child nodes. We call this hypothetical node an ancestor. There can be multiple levels of ancestors with multiple activity types sharing the same ancestor. Each node in the tree represented in Figure 2 is labeled, but one may choose to not label the hypothetical nodes. To make a distinction we depicted all the ancestors with rounded rectangles and the leaf nodes with rectangles. We now use an approach similar to that used in phylogenetics [43] to study evolutionary relationships between various species of organisms. Note that such a tree need not be a binary tree.

We use a distance matrix D to quantify the amount of dissimilarity between any two leaf nodes in the tree. The amount of dissimilarity between two leaf nodes is simply the number of edges in the shortest path connecting them. Let d_{ij} denote the dissimilarity between leaf nodes i and j in the tree. d_{ij} also denotes the ij-th element of the matrix D.



$$\begin{array}{c} \\ p \\ q \\ r \\ s \end{array} \begin{array}{c} \begin{array}{cccc} p & q & r & s \end{array} \\ \begin{pmatrix} 0 & 2 & 3 & 4 \\ 2 & 0 & 3 & 4 \\ 3 & 3 & 0 & 3 \\ 4 & 4 & 3 & 0 \end{pmatrix} \end{array}$$

Figure 3. Classification tree and its corresponding distance matrix *D*. The distance d_{qr} is 3 meaning that there are 3 edges in the path connecting leaf nodes *q* and *r*.

Definition: We define early activity diversity score of a set of distinct activity types as the normalized mean value of pair-wise dissimilarity (defined above) between all activity types in the set. More formally, if n is the number of distinct activity types represented in a set S of activity types, then the early activity diversity score of the set S is given by

$$\text{DSCORE} = Div(S) = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} d_{ij}}{n-1}$$

The proposed metric *Div(S)* has the following properties:

1. *Div(S)* is zero if a user performs activities of only one type.

2. When all leaf nodes (or activity types) have one and only one ancestor, then *Div(S)* is simply what is called "richness" in biodiversity and ecology literature.

3. *Div(S)* increases as ancestral connection increases. In the figure 2, the set $\{p, s\}$ is more diverse than the set $\{p, r\}$ which in turn is more diverse than the set $\{p, q\}$. In other words, diversity of two leaf nodes whose parent is same is less than diversity of two leaf nodes whose parent is different.

3. As additional distinct leaf nodes are added to a set of activities, *Div(S)* increases. For example, $Div(\{p, q, r\})$ will always be greater than $Div(\{p, q\})$, for $p \neq q \neq r$.

4. For sets of the same length *Div(S)* attains a maximum value when no two leaf nodes of the set have the same parent and a minimum value when all leaf nodes of the set have the same parent.

Note that we are not interested in proportional abundance of a given leaf node (or activity type), because all we care about is whether the user got an opportunity to use the feature at least once. Using this, we are interesting in predicting user churn, so we formulated the definition such that $\{p, q\}$, $\{p, p, q\}$, $\{p, p, p, q\}$ and $\{p, p, q, q\}$ are all equally diverse.

The theoretical maximum for DSCORE for our tree is 41.6.

Early Activity Score (ASCORE) Early Activity Score is defined for user *u* as the quantity of activity performed by user *u* in the *first* session. We denote it by ASCORE.

Design Challenge: In online communities, users engage in different types of activities for different periods of time. So, a simple count of all activities may turn out to be an inaccurate representation of the amount of activity performed by the user for a session as it diminishes the impact of the time- and engagement-intensive multi-step editing and adding activities. We considered two ways to adjust for this imbalance: weighing activities (i) by infrequency of use (so rare activities count more) or (ii) weighing them by time spent on the activity (so more interactive/intensive activities count more). We choose the latter as a better measure of activity that is not connected to diversity (which is related to rarity / dissimilarity / spread). Thus, we define weight of a unit of each activity type factoring in the time duration associated with that activity type. In the related work section, we have stated how time durations can often be inaccurate because users may sporadically drift to other websites for indefinite periods of time during their course of interaction with the community. Therefore, in activity-times data, one might expect to find outliers for certain user activities. In order to eliminate bias due to such time periods, we pick the median time duration of all users for each activity type as the weight for that activity type for all users.

Definition: If w_i denotes the median time duration for activity i for all users and the user u performed n_i activities of that type in the first session, then the early activity score for the user u is given by

$$\text{ASCORE} = \sum_i w_i n_i$$

METHODOLOGY

Key Steps in our Data Analysis

Dividing Activity Log into Sessions

Ideally, an activity session is defined to be the time between a user's login and logout. However, for not all users we have the information about the login and logout events. For those users for whom we do have this information, we accurately determine a login session. However, for those users whose login and / or logout events are missing for whatever reason (they stay logged in for an indefinite period or quit the browser or close the tab without logging out, etc) we use the definition of session based on log-scaled inter-activity times [16]. For such users, a session is identified to be *a set of continuous activities by a user in which any two subsequent activities are within a time difference of 1 hour*.

Representing User's Defection from the community

For analyzing the expected active period of a Movielens user, we model the user's lifetime by defining concepts of "birth" and "defection" in (from) the community as the times at which the user starts his activity and stops his activity for a considerable period of time respectively. For Movielens, we determine the threshold of inactivity to be 365 days based on activity logs which show a bi-normal distribution with the second normal at about 300 days after the first registration. So if *a user does not have an activity for 365 days since they last visited, we consider the user to have dropped out of the community.*

Ignoring activities beyond the 365 day inactive period

Based on the above threshold of inactivity, if a user is found to have an activity after 365 days, we have every reason to believe his/her movie-seeking behavior might have changed over the course of this time. So, we consider the activity thereon under a new life instance of the same user. We found a small fraction of users (2,136) with more than one life instance in our dataset. But we have carried out our analysis on 48,784 distinct users for whom only the activities of the first life instance were considered ignoring the activities beyond the 365 day inactive period.

Handling right-censored users

Note that we have ended our data collection on a certain date and therefore we do not have information about some users whether or not they return to the system after 365 days. This concept is identified in survival analyses

literature as *right-censoring* and in these analyses these users are marked as *right-censored users*. We have 8157 users who are right-censored. We model user churn using two approaches - Survival analysis and Logistic Regression. For modeling using survival analysis, we will use appropriate survival models (Cox-Regression) to handle the right-censored users. Because logistic regression is used for making prediction/binary classification and it does not handle right-censored users, we will ignore those data points in the logistic models.

Computing ASCORE weights

Recall that our ASCORE metric requires a time-measure as a weight for each activity. Based on the timestamps in the log data, we compute the weight as the median time between the start of an activity and the start of the next activity (omitting the final activity in each session). Table 1 lists the median time duration in seconds for which users of Movielens spent time on an activity before moving on to the next one.

Activity Type	Weight (median time duration in seconds)
Edit Profile	12
Create RSS Feed	13
Invite a buddy	14
Using help pages	20
View Movie detail	10
Search by tag	16
Search using attribute/ keyword	13
Visit Most Often Rated Movies list	16
Visit Top Picks list	14
Visit Newest Additions list	17
Visit Rate-Random-Movies list	10
Visit "Your Wishlist"	19
Visit "Your Ratings" list	17
Rate a movie	7
Add a tag	9
Q&A	9
Add/edit a movie	13

Table 1. Median time duration in seconds spent by Movielens users for each activity type.

Choosing predictors for the model

The data we have access to has extremely sparse age and gender information with practically no other personal information available. Nor do we have any information about the motivations or pro-social behavioral history about the users. So, we are unable to use any of these as predictors in our model. We do not use length of first session as a predictor firstly because we believe time durations are inaccurate representations of user activity due to general user drifting behavior and second because we infer activity sessions from activity log data of the user, which does not always contain login and logout. We could choose metrics specific to Movielens (that may not be generalizable to other systems) such as the number of movies rated by user in the first session (because Movielens is primarily a movie-recommender website powered by user ratings). However, we found that the number of movies rated has high correlation with amount of activity in the first session and so would not really add much to explaining the model. So, we decided to use *amount of activity* (which is only about 0.4 correlated with activity diversity) along with *activity diversity* in our model.

RESULTS AND DISCUSSION

Structure of this section

We present and discuss results in three sections. First, we explore the MovieLens log data to see the distribution of activity diversity and the prevalence of new user churn. Then we try to see how user churn is associated with early activity diversity using varying measures of longevity and different approaches to modeling user churn to establish the robustness of our results. Finally, we validate the usefulness of our DSCORE metric by comparing it with the Gini-Simpson Index in the best-fitting model.

Frequency of activity types on Movielens

On Movielens, we find that 37.74% of activity types for all recorded sessions for all users in the data constitute 'rating movies' (most often performed) followed by 30.76% of activity types constituting 'search' using attribute/keyword. This is followed by visits to movie detail page constituting about 9.56% of total actions, followed by viewing one's own rated movies, viewing the top picks list and tagging movies accounting for another 8% of activities (See Figure 4). The remaining 11 activity types count to only about 14% of the activities on Movielens. Thus we see that most users are highly specialized in the ways they participate in Movielens although they have about 17 different activity types to engage in.

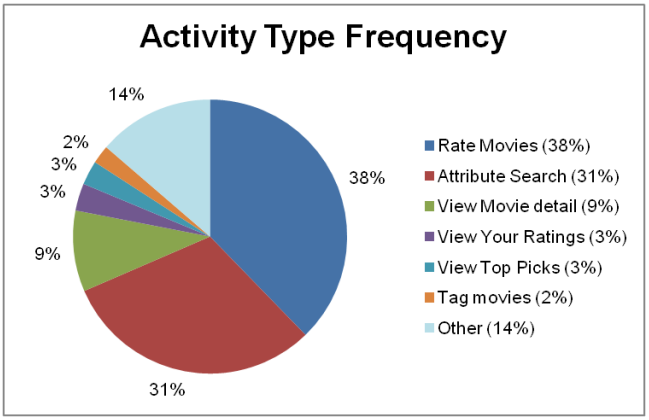


Figure 4. Frequency of activity types on Movielens

Evidence of early user churn on Movielens

Large numbers of users drop out in their first few sessions (see Figure 5, a plot based on our dataset) and particularly significant drop occurs right after the first session. So, we will use ASCORE and DSCORE of a user at the *first session*.

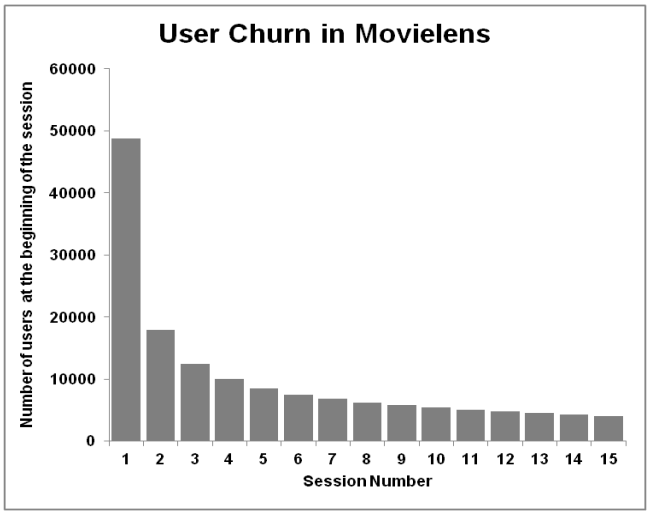


Figure 5. User Churn in Movielens

Relationship between percentage user churn and simple number of activity types tried in the first session

We define percentage user churn after the *n*-th session to be the number of users who dropped out of the community after the *n*-th session over the total number of users who tried *k* activity types in the first session where $k = 1 \ldots 15$ (although participation in all 17 activity types is theoretically possible, the users in our dataset have participated in at most 15 activity types by the end of the first session) and $n = 1,5,10$ (we report only for these sessions in Table 2.). Ignoring the users who are right-censored, we find that the lower the number of activity types tried in the first session, the greater the percentage of user churn. The results are available in Figure 6 and its corresponding Table 2.

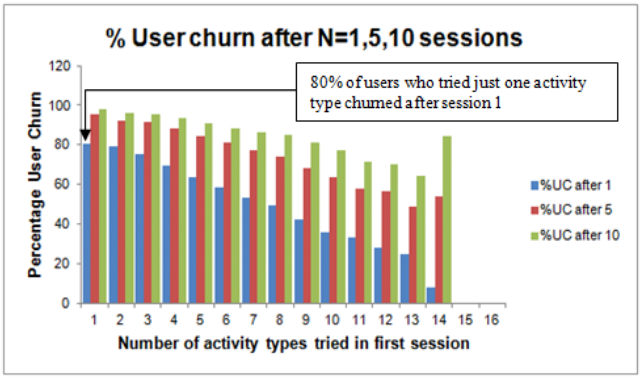


Figure 6. A graph showing percentage user churn after the first, fifth and tenth sessions against the number of activity types tried in first session

The corresponding table (Table 2) for the graph shown in Figure 6 is listed below:

#Activity Types tried in first session	#Users	%User Churn after 1st session	%User Churn after 5th session	%User Churn after 10th session
1	4519	80.7	95.42	97.83
2	4252	79.06	92.48	96.23
3	6508	75.61	91.84	95.71
4	7246	69.74	88.09	93.21
5	6836	63.77	84.54	90.96
6	5798	58.51	81.09	88.13
7	4637	53.35	77.38	86.48
8	3141	49.44	73.95	85.08
9	1842	42.23	68.33	80.97
10	998	35.52	63.73	77.33
11	421	33.44	57.67	71.47
12	219	27.88	56.37	70.3
13	72	25.09	49.06	64.15
14	17	7.69	53.85	84.62
15	1	0	0	0

Table 2. Table showing percentage user churn after the first, fifth and tenth sessions

RQ1*: How is early activity diversity (measured using DSCORE) associated with new user longevity?*

Earlier in this paper, we have identified two metrics based on activity: activity score (denoted by ASCORE) and activity diversity score (denoted by DSCORE). We will use the values of these two metrics for the first session of the new user for predicting churn or longevity in the community.

(Approach 1) Survival analysis using Cox Proportional-Hazards model

In our first approach, we use Survival Analysis using Cox Proportional Hazards because this is ideal in situations where one measures time until an event or hazard (in this case – a user leaving a community) happens with '*Number of Sessions*' (continuous measure) as the measure of longevity. Earlier in this paper, we have introduced briefly the concept of right-censoring. Because Cox Regression [13] takes care of right-censored data, we perform survival analysis for all 48,784 users.

We build two models – one consisting only of ASCORE as the predictor and the other having both ASCORE and DSCORE. We find that the difference in log likelihoods of the two models is statistically significant (p-value < 0.0001) with $\chi^2 = 410.6$. Based on likelihood ratio test, this implies that the model that includes DSCORE is better than the model that has only ASCORE.

The corresponding coefficients for the second model are shown in Table 3.

	Coef	Exp(coef)
ASCORE	- 0.00018***	0.9998
DSCORE	- 0.02304***	0.9772

Table 3. Coefficients for Cox-Proportional Hazards Model; ***indicates *p*-value < 0.001

The values in Table 3 indicate that holding the other covariates constant, a unit increase in amount of activity (ASCORE) causes a 0.02% reduction in churn hazard and a unit increase in activity diversity (DSCORE) causes a 2.28% reduction in churn hazard. Given the difference in scales, this is hard to interpret. So we address the quantitative aspect below in our logistic regression model with an illustrative example.

(Approach 2) Logistic regression

In our second approach, we use Logistic regression for obtaining a simpler interpretation and a direct estimate of probability of survival past an arbitrary session *k*. For this we use 'presence beyond session *k*' (Binary measure) as the measure of longevity. Because logistic regression can be used for prediction, we ignore the users whose survival information is right-censored in our analysis.

Step 1: Longevity measure for logistic regression

The survival curve shown in Figure 7 indicates that the probability that the user survives is highest in session one and gradually drops as one moves towards further sessions.

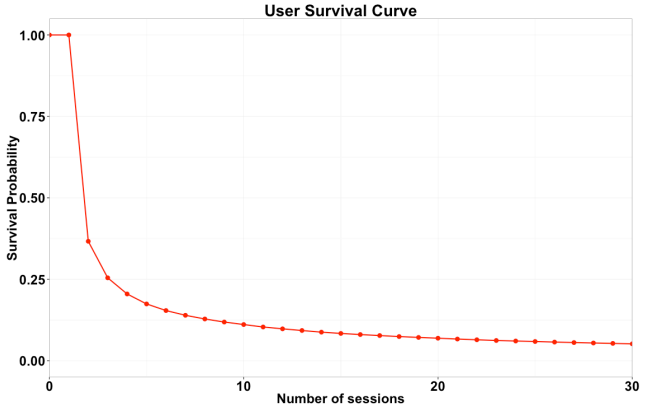


Figure 7. User Survival curve plotted to determine a suitable threshold for logistic regression analysis. The graph shows a drop in the probability of survival of users as we proceed from 0 to 30 sessions.

From Figure 7, we also see that after using MovieLens for at least 10 sessions (an average of 2 months), the probability that users continue to use MovieLens is very high. Therefore, we choose N=10 sessions as the measure of longevity to examine how well we can predict if users would stay in the community beyond 10 sessions. (We do a sensitivity check and repeat the analyses with different values of N=1 and 5 sessions, and find consistent results.) So for our logistic regression model, we use a binary response variable with values 1 or 0 indicating the two classes – Class 1 – for 'users who stayed in MovieLens for at least 10 sessions (or 2 months)' and Class 0 – for 'users who stopped using MovieLens after their 10th session'.

Step 2: Analysis

We first build three models – one having only DSCORE, one having only ASCORE and the third having both ASCORE and DSCORE. Table 4 shows the corresponding outputs:

	Model 1	Model 2	Model 3
(Intercept)	-3.542***	-2.602***	-3.352***
ASCORE		0.0003***	0.0002***
DSCORE	0.0939***		0.0604***
AIC	21302	21093	20808

Table 4. Summary of the logistic regression models; ***indicates *p*-value < 0.001

For a given dataset, AIC (Akaike Information Criterion) measures how one model performs relative to another. The models with smaller AIC have better fit. We find in Table 4 that the model having ASCORE alone is better than the one having only DSCORE. However, based on AICs we conclude that the model that includes both ASCORE and DSCORE is better than the individual models. We find also that the likelihood ratio test statistic between the models 2 and 3 has a $\chi^2 = 287.04$ (p-value ~ 0) and that between the models 1 and 3 has a $\chi^2 = 496.38$ (p-value ~ 0). So again, the model that includes both DSCORE and ASCORE is better than the individual models. The results show that both activity and diversity are important, but that retention is more sensitive to smaller changes in diversity.

Using this third model that includes both terms, we note that keeping other terms constant, a unit increase in the amount of activity (ASCORE) produces a 0.03% increase in the odds of survival beyond 10 sessions, while a unit increase in the activity diversity (DSCORE) produces a 6.23% increase in the odds of survival beyond 10 sessions.

We now use the model with the best fit (the third model) to illustrate the increase in average longevity associated with marginal increases in activity level and diversity. Consider a typical newcomer that we will call Amy with a median ASCORE (288 units) and median DSCORE (12.5 units). A typical profile for such a user would have rating 17 movies, making 5 attribute/keyword searches, using the help feature once, viewing details for 5 movies and using the "Your Ratings" feature twice. Amy's chance of surviving past the 10th session is only 7.31%.

Now let us consider a second user – Ben – who has the same activity pattern as Amy but also performed one additional and fairly different task. Ben *invites a buddy to MovieLens*. To keep Ben's ASCORE constant, we will also have Ben rate only 15 movies (two fewer than Amy). This

changes Ben's DSCORE to 15.4 units while holding his ASCORE at 288 units, but it results in 19.14% higher odds of survival – an increase to 8.6%.

Finally, let us consider a third user – Claire – who starts with Amy's level of activity but we want to increase her ASCORE to the level that would predict the same survival rate as Ben, while holding her DSCORE constant (i.e., by increasing quantity without adding new activity types). Claire would have to increase her ASCORE by 876 units which would involve (for example) 78 additional movie ratings, 10 more attribute/keyword searches, and viewing 20 more movie detail pages.

In other words, our model shows that performing one activity of a different type is associated with an increase in survival which can only be matched by performing existing activities *many more times each*. We also tested the models at N = 1 and 5 sessions and found consistent results.

Step 3: Prediction accuracy

Because of imbalance in distribution of users in both classes, we do not use precision and F-measure for gauging performance. Instead, we use sensitivity and specificity. Sensitivity, in our context is the proportion of class 1 users who are correctly classified and specificity is the proportion of class 0 users who are correctly classified.

Because logistic regression gives the probability or log odds that the output belongs to class 1, we need a suitable threshold t to compare the probability obtained using logistic regression p to say if $p > t$ then the user belongs to class 1 else the user belongs to class 0.

We use two approaches to choose an optimal threshold – the Minimized Difference Threshold (MDT) approach, which minimizes the difference between sensitivity and specificity and the Maximized Sum Threshold (MST), which maximizes the sum of sensitivity and specificity [18]. Note that while these thresholds are not biased towards positives or negatives they do not necessarily give the highest prediction in the model. To make sure our model is not data-dependent, we perform 5-fold cross validation and the average sensitivity and average specificity for the best model were found to be 0.65 using one approach and 0.66 using another.

Step 4: DSCORE: Verifying sensitivity to ontology

To verify DSCORE's sensitivity to the ontology we used, we make slight changes to the existing ontology.

In Figure 2, we move *Create RSS Feed* and *Invite Buddy* from **Account Maintenance** to **Social** and we move all four leaves of **Browser Predefined Lists** to **Search**. These changes make the ontology somewhat different but still sensible. We find that the newly computed DSCORE has a correlation of 0.9974 with the old one producing very similar results.

Step 5: DSCORE in presence of 'Number of activities'

We also include total number of activities into Model 3 and find that it is not significant in predicting survival beyond 10 sessions but is significant in predicting survival beyond 1 and 5 sessions, but in all three cases (1, 5, and 10) DSCORE is still significant and performs well.

RQ2*: How can we most effectively measure early activity diversity for purposes of predicting new user longevity?*

Because it takes a lot more work to compute DSCORE using a distance tree analysis, we wondered how useful this measure is over a traditional diversity measure such as the Gini-Simpson Index. We therefore replaced DSCORE in our model with the Gini-Simpson and report the results at N = 10 sessions in Table 5.

	Model 3	**Model 4**
(Intercept)	-3.352***	-2.724***
ASCORE	0.0002***	0.0004***
DSCORE	0.0604***	
Gini-Simpson		0.197
AIC	20808	21093

Table 5. A comparison of models using DSCORE and Gini-Simpson index; * indicates *p*-value < 0.001**

We find that introducing it does not add significant value over ASCORE in predicting survival beyond N= 10 sessions. We performed a sensitivity analysis by redoing it at N = 1 and 5 sessions as well and we found again that Gini-Simpson is not significant in presence of ASCORE. Also, because some activities are closely related to each other while others may be very different and Gini-Simpson does not account for this as well, we find DSCORE to be more useful in characterizing early activity diversity.

CONCLUSION

This paper is a preliminary investigation of new user activity engagement in an online community and is largely intended to describe its effects on new user survival within the community. This work stands out in comparison to previous works addressing the challenge of user retention in that it introduces a novel way of assessing retention using limited amount of information available about new users. We make use of metrics based on activity in the very *first session*: diversity and amount of activity. We also introduce a way of computing diversity in online communities. The hypotheses were tested on Movielens, an online community that gives its users an opportunity to participate in a variety of ways – from finding movies they like to rating to tagging movies to answering questions about movies, to inviting buddies and so on.

Our findings indicate that 1) the lower the number of activity types tried in the first session, the greater the

percentage of users in that category who drop out of the community; 2) early activity diversity measured using DSCORE is a significant predictor of user longevity, and that it remains a significant predictor even in the presence of amount of early activity (ASCORE); and 3) a metric that considers possible similarity between activity types based on a distance-tree is a more useful way of measuring early diversity than traditional metrics. Our results are invariant of measures of longevity and the approaches used to model user churn. We also find that the positive effect of higher early diversity being associated with greater longevity is consistent with prior research [6,24,25].

Limitations and Generalizability

The Movielens data log limited our ability to assess the relationship of other features to user retention, and of our diversity metric in presence of other features. We are however interested in seeing how this works in other contexts where such data may be available. Nonetheless, we were able to assess user retention and longevity from first-session activity data, which is readily available for all online communities.

We chose Movielens dataset for our analysis because of its richness in activities offered to its users combined with a long-term longitudinal log of user activity – a log that dates from the user's very first interactions. This work is relevant to community designers, moderators and administrators who wish to understand new user longevity in a variety of contexts: travel sites where diversity of activities (reviews/ ratings) may be with respect to categories: hotels, restaurants and places; Q&A sites and peer-production communities where diversity may be with respect to types of content posted or moderation activities, social networks where diversity may be in types of content shared, on their own profile or others'; product review and retail sites where users may buy/use a variety of products; and so on.

Applying DSCORE to other contexts requires creation of an activity taxonomy. We have not validated its effectiveness in systems with different taxonomies or different types of activity structure, and leave that to future work.

Another interesting scenario is of online communities that unlock features with increase in user reputation such as StackOverflow. New users in such spaces have limited activity choices to explore, and one may have to investigate other approaches for assessing user retention.

Future work

In the next stage of this work, it seems natural to look at questions of causality including direction – whether users who are longer surviving tend to be more diverse or vice-versa considering even the possibilities of joint causality with other factors of site design.

There are two ways in which this might be useful. A community administrator might want to:

(a) identify users who are more/less likely to return to invest effort (e.g., relevant offers, mentors, greetings) in those users who are likely to return and attempt to "recapture" their interest in those who are not.

(b) use activity diversity as a metric to assess overall site engagement. Apart from simply using for predictions about longevity, an analysis of the activity types usage distribution may lead to further opportunities to engage users. Also, it tells the site administrators what activity types users engage in and what activity types need more visibility.

Based on our observations of commercial site interactions, we expect that some sites may already be employing some of these methods and we look forward to public research results that establish or refute causality.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grants IIS 0808692, 1017697, 1319382, and Fellowship support from the University of Minnesota Informatics Institute. We also acknowledge the thoughtful discussions and helpful feedback from several members of the Grouplens Research lab along with their assistance in development of the classification tree. We thank the consultants at the Statistical Consulting Center and the Biostatistics Consulting Unit at the University of Minnesota for reviewing our statistical analyses. We also thank the anonymous reviewers for their valuable comments.

REFERENCES

1. Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '06), 959-968. http://doi.acm.org/10.1145/1124772.1124916
2. Talya N. Bauer, Todd E Bodner, Berrin Erdogan, Donald M. Truxillo, and Jennifer S. Tucker. 2007. Newcomer Adjustment During Organizational Socialization: A Meta-Analytic Review of Antecedents, Outcomes, and Methods. *Journal of Applied Psychology* 92, 3: 707–721.
3. Zoheb H. Borbora and Jaideep Srivastava. 2012. User Behavior Modelling Approach for Churn Prediction in Online Games. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust* (SOCIALCOM-PASSAT '12), 51-60. http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.84

4. Moira Burke, Cameron Marlow, and Thomas Lento. 2009. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 945-954. http://doi.acm.org/10.1145/1518701.1518847
5. Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social network activity and social well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 1909-1912. http://doi.acm.org/10.1145/1753326.1753613
6. Derek S. Chapman, Krista L. Uggerslev, Sarah A. Carroll, Kelly A. Piasentin, and David A. Jones. 2005. Applicant Attraction to Organizations and Job Choice: A Meta-Analytic Review of the Correlates of Recruiting Outcomes. *Journal of Applied Psychology* 90, 5: 928-944.
7. Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. 2010. Socialization tactics in Wikipedia and their effects. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (CSCW '10), 107-116. http://doi.acm.org/10.1145/1718918.1718940
8. Giovanni Luca Ciampaglia and Dario Taraborelli. 2015. MoodBar: Increasing New User Retention in Wikipedia through Lightweight Socialization. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 734-742. http://doi.acm.org/10.1145/2675133.2675181
9. Robert K. Colwell. 2009. Biodiversity: concepts, patterns and measurement. In *The Princeton guide to ecology* (1st. ed.), Simon A. Levin (eds.). Princeton University Press, Princeton, NJ, 257–263.
10. Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. 2012. Churn prediction in new users of Yahoo! answers. In *Proceedings of the 21st international conference companion on World Wide Web* (WWW '12 Companion), 829-834. http://doi.acm.org/10.1145/2187980.2188207
11. Nicolas Ducheneaut. 2005. Socialization in an Open Source Software Community: A Socio-Technical Analysis. *Comput. Supported Coop. Work* 14, 4: 323-368.
12. Kate Ehrlich, Michael Muller, Tara Matthews, Ido Guy, and Inbal Ronen. 2014. What motivates members to contribute to enterprise online communities? In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing* (CSCW Companion '14), 149-152. http://doi.acm.org/10.1145/2556420.2556477
13. John Fox. 2002. Cox Proportional-Hazards Regression for Survival Data. Retrieved July 24, 2015 from https://socserv.socsci.mcmaster.ca/jfox/Books/Companion-1E/appendix-cox-regression.pdf
14. Jill Freyne, Michal Jacovi, Ido Guy, and Werner Geyer. 2009. Increasing engagement through early recommender intervention. In *Proceedings of the third ACM conference on Recommender systems* (RecSys '09), 85-92. http://doi.acm.org/10.1145/1639714.1639730
15. Paul Fugelstad, Patrick Dwyer, Jennifer Filson Moses, John Kim, Cleila Anna Mannino, Loren Terveen, and Mark Snyder. 2012. What makes users rate (share, tag, edit...)?: predicting patterns of participation in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (CSCW '12), 969-978. http://doi.acm.org/10.1145/2145204.2145349
16. R. Stuart Geiger and Aaron Halfaker. 2013. Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13). ACM, New York, NY, USA, 861-870. http://doi.acm.org/10.1145/2441776.2441873
17. Corrado Gini. 1912. *Variabilità e mutabilità; contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.]* Tipogr. di P. Cuppini.
18. Alberto Jiménez-Valverde, and Jorge M. Lobo. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica* 31, 3: 361-369.
19. Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2: 363-375.
20. Elisabeth Joyce and Robert Kraut. 2006. Predicting continued participation in newsgroups. *Computer-Mediated Communication* 11, 3: 723-747.
21. Sanjay Ram Kairam, Dan J. Wang, and Jure Leskovec. 2012. The life and death of online groups: predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining* (WSDM '12), 673-682. http://doi.acm.org/10.1145/2124295.2124374
22. Jaya Kawale, Aditya Pal, and Jaideep Srivastava. 2009. Churn Prediction in MMORPGs: A Social Influence Based Approach. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04* (CSE '09), 423-428. http://dx.doi.org/10.1109/CSE.2009.80
23. Amy Jo Kim. 2000. *Community Building on the Web: Secret Strategies for Successful Online Communities* (1st ed.). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
24. Amy L. Kristof. 1996. Person-organization fit: An integrative review of its conceptualizations,

measurement, and implications. *Personnel Psychology* 49, 1: 1-49.

25. Amy L. Kristof-Brown, Ryan D. Zimmerman, and Eric C. Johnson. 2005. Consequences of Individuals' Fit at Work: A Meta-Analysis of Person-Job, Person-Organization, Person-Group, and Person-Supervisor Fit. *Personnel Psychology* 58, 2: 281-320.
26. John M. Levine and Richard L. Moreland. 1994. Group socialization: Theory and research. In *European Review of Social Psychology* (Vol. 5), Wolfgang. Stroebe & Miles Hewstone (eds.). John Wiley & Sons, NY, 305-336.
27. Michael Muller, N. Sadat Shami, David R. Millen, and Jonathan Feinberg. 2010. We are all lurkers: consuming behaviors among authors and readers in an enterprise file-sharing service. In *Proceedings of the 16th ACM international conference on Supporting group work* (GROUP '10), 201-210. http://doi.acm.org/10.1145/1880071.1880106
28. Blair Nonnecke and Jenny Preece. 2000. Lurker demographics: counting the silent. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '00), 73-80. http://doi.acm.org/10.1145/332040.332409
29. Oded Nov. 2007. What motivates Wikipedians? *Commun. ACM* 50, 11: 60-64. http://doi.acm.org/10.1145/1297797.1297798
30. Aditya Pal, Shuo Chang, and Joseph A. Konstan. 2012. Evolution of experts in Question Answering Communities. In *Proceedings of the Sixth International AAAI Conference on Web and Social Media (*ICWSM'12), 274-281.
31. Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work* (GROUP '09), 51-60. http://doi.acm.org/10.1145/1531674.1531682
32. Derek M. Powazek. 2002. *Design for Community: The Art of Connecting Real People in Virtual Places*. New Riders.
33. Jenny Preece, Blair Nonnecke, Dorine Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior* 20, 1: 201–223.
34. Jagat Sastry Pudipeddi, Leman Akoglu, and Hanghang Tong. 2014. User churn in focused question answering sites: characterizations and prediction. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion* (WWW Companion '14), 469-474. http://dx.doi.org/10.1145/2567948.2576965
35. Frederick F. Reichheld, W. Earl Sasser. 1990. Zero Defections: Quality Comes to Services. *Harvard Business Review* 68, 5: 105-111
36. Larry J. Rosenberg and John A. Czepiel. 1984. A Marketing Approach For Customer Retention. *Journal of Consumer Marketing* 1, 2: 45-51.
37. Claude Elwood Shannon. 2001. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun.* 5, 1: 3-55.
38. Edward H. Simpson. 1949. Measurement of diversity. *Nature* 163: 688-688.
39. Donna Spencer. 2009. *Card Sorting: Designing Usable Categories*. Rosenfeld Media.
40. Katherine J. Stewart and Sanjay Gosain. 2006. The impact of ideology on effectiveness in open source software development teams. *MIS Quarterly* 30, 2: 291-314
41. Tammara Combs Turner, Marc A. Smith, Danyel Fisher, and Howard T. Welser. 2005. Picturing Usenet: Mapping computer mediated collective action. In *Computer-Mediated Commun.* 10, 4: 00-00.
42. Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure* 8, 2: 1-32.
43. Jason Tsong-Li Wang, Huiyuan Shan, Dennis Shasha, and William H. Piel. 2003. TreeRank: a similarity measure for nearest neighbor searching in phylogenetic databases. In *Proc. of the 15th International Conference on Scientific and Statistical Database Management* (ICSSDM'03)*,* 171-180.

    http://dx.doi.org/10.1109/SSDM.2003.1214978
44. M. McLure Wasko, and Sameer Faraj. 2000. "It is what one does": why people participate and help others in electronic communities of practice. *Journal of Strategic Information Systems* 9, 2-3: 155-173.
45. The Wikimedia Foundation. Attracting and retaining participants. Retrieved July 21, 2015 from https://strategy.wikimedia.org/wiki/Attracting_and_retaining_participants
46. Jiang Yang, Wei Xiao, Mark S. Ackerman, and Lada A. Adamic. 2010.Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media (*ICWSM'10), 186-193.
47. Alcides Velasquez, Rick Wash, Cliffe Lampe, and Tor Bjornrud. 2013. Latent users in an online user-generated content community. *Comput. Supported Coop. Work* 14, 23: 21-50.

48. Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (CSCW '12), 935-944. http://doi.acm.org/10.1145/2145204.2145344

49. Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 281-290. http://doi.acm.org/10.1145/2556288.2557213