To Blog or Not to Blog: Characterizing and Predicting Retention in Community Blogs

Imrul Kayes, Xiang Zuo
Computer Science and Engineering
University of South Florida
Tampa, Florida, USA
{imrul,xiangzuo}@mail.usf.edu

Da Wang
Electrical and Electronic Engineering
HuBei University of Technology
Wuhan, China
wangda222@126.com

Jacob Chakareski
Electrical and Computer Engineering
University of Alabama
Tuscaloosa, Alabama, USA
jacob@ua.edu

ABSTRACT

Community blogging is a medium for publishing daily journals, expressing opinions or ideas, and sharing knowledge. Blogging has a high impact on marketing, shaping public opinions, and informing the world about major events from a grassroots point of view. However, turnover in online blogging is very high, with most people who initially join and start contributing to the community, failing to contribute in the long run.

In this paper, we ask what factors cause a blogger to continue participating in the community by contributing content (e.g., posts, comments). We crawled a sample of blogger profiles from a popular community blogging platform "Blogster". These bloggers contributed about 91% posts in the community. We derived a set of well-grounded variables related to blogger retention and built a predictive model from the variables. Our results show that the male and aged (senior) bloggers, who face fewer constraints and have more opportunities in the community are more retained than others. Other bloggers pay a high degree of attention to these retained bloggers through implicit (reading posts) and explicit (writing comments) interactions.

We have also found that a blogger has higher retention if her friends have also higher retention and a strong social tie reduces retention imbalance between two blogger friends. However, we found that a blogger's network age (e.g., how long ago she joined) has no effect on her retention. Our work has theoretical implications for the social behavior literature of bloggers, and practical implications for potential community blogging platform developers.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Social and Behavioural Sciences**]: Sociology

SocialCom '14 August 04 - 07 2014, Beijing, China

http://dx.doi.org/10.1145/2639968.2640061 ...$15.00.

General Terms

Measurement, Human Factors

Keywords

Community Blogs, Retention, Bloggers

1. INTRODUCTION

The new age of participatory web applications such as Blogs has enabled a new, grassroots form of journalism and is viewed as a way to shape democracy outside the mass media and conventional party politics [5]. A blog is a personal journal published on the Web consisting of discrete entries ("posts") typically displayed in reverse chronological order.Blogs are usually the work of a single individual, occasionally of a small group, and are often themed on a focused topic. A conventional blog may combine text, images and links to other blogs and to web pages. Blogging platforms allow the creation of online profiles in which links to other bloggers are specified. This blogger to blogger declared social ties specify the blogger's interest and endorsement of other bloggers, creating a social network through which blog updates are automatically disseminated.

Blogging has become immensely popular recently and its impact is multidimentional. For example, WordPress alone, a free and open source blogging tool, is used by over 14.7% of Alexa Internet's "top 1 million" websites and as of August 2011 manages 22% of all new websites [21]. Citizen journalism had high impact in major events such as South Asia tsunami, London terrorist bombings, and New Orleans Hurricane Katrina [5]. The *blogosphere*, the virtual universe of the blogs on the web, provides thus a conducive platform for different aspects of virtual and real life, such as viral marketing, sales prediction and counter terrorism efforts.

Despite the overwhelming impact, building and maintaining an online blogging community is difficult. First, participation is often sparse and uneven [11]. For example, Cummings et al. [2] examined a listserv-based online group and found that one-third of all listed users had no communication during a three-month observation period, and only 15% of users contributed a single message during that period. Second, churn in membership is high, with most users who initially contribute to the community never contributing again [11]. For example, Jones et al. [10] examined a sample of users from the Usenet newsgroups and found that

only 11.5% of the people who posted in one month returned to post in the second month.

The problem of participation in and contribution to a blogging community can be resolved by answering two research questions: (1) What motivates a user to join a blogging community? (2) What motivates the blogger to continue participating (e.g., posting, commenting, etc.) in the community? Existing research, based on content analysis [9] and interviews [17], provides an excellent insight into a user's motivation for joining a blogging community. These research reveals a number of reasons for joining, including the desire to publish a diary, express opinions and emotions, articulate ideas through writing, share knowledge, and form a community. However, the question of why bloggers continue participating, in other words why some bloggers have higher retention, is not well researched. Existing research [14] has used bloggers' undeclared social networks (e.g., comment networks, invitation networks) and found that social relationships and cultural elements have an effect on continued participation or higher retention.

Our goal is to examine factors that cause a blogger to continue participating on the platform. In doing so, we make the following contributions.

- We crawled a sample of bloggers' profiles from an online community blogging platform Blogster. These users contributed about 91% of posts in the community, and hence are a good candidate for a representative sample of the bloggers in Blogster (Section 2).
- We proposed five categories of variables: network-specific (e.g., centralities, clustering coefficient), user activity specific (e.g., posts, comments, photos, network age), physiological (e.g., age, gender), interactional (e.g., blog traffic, other users' comments) and relational (e.g., social tie strength, friends retention) to understand which factors contribute to continued participation or higher retention and to what degree (Section 3).
- Finally, we put together the variables to predict a blogger's retention in Blogster. We found that our model fits well the data and predicts retention with *Adjusted* $R^2 = 0.84$ (Section 4).

Our work has multiple practical and theoretical implications. First, future research on social behavior of bloggers will benefit from the understanding of the variables that predict continued activity. Second, the variables revealed in this study, which contribute to continued blogging activity, may allow the developers of a new community blog to make more informed design decisions. Finally, one could imagine a "retention score" from these variables that can complement the incentive oriented scores (e.g., points) in a blogging platform.

2. DATA COLLECTION

2.1 Sampling technique

We modify the Metropolis-Hastings Random Walk (MHRW) algorithm [6] to crawl the Blogster network. The MHRW algorithm is capable of obtaining a uniform sample (or more generally a probability sample) of OSN users [6] .

We consider the social graph of Blogster as an undirected graph $G = (V, E)$, where V is a set of nodes (bloggers) and E is a set of edges (ties among bloggers). The crawling of the Blogster starts with a node and proceeds iteratively. We selected a random online blogger as a seed. In each iteration, we visited a node and discover all its neighbors. If the current node is v, the next hop node w is chosen according to the following transition probability:

$$P_{u,v}^{MH} = \begin{cases} \min(\frac{1}{k_u}, \frac{1}{k_v}) & \text{if v is a neighbor of u,} \\ 1 - \sum_{x \neq u} P_{u,x}^{MH} & \text{if v=u,} \\ 0, & \text{otherwise} \end{cases}$$

In each iteration, at the current node u, our algorithm randomly selected a neighbor v and moved there with probability $\min(1, \frac{k_u}{k_v})$. The algorithm accepted the move towards a node of smaller degree, and rejected some of the moves towards higher degree nodes. This pattern of moving eliminates the bias towards high degree nodes.

2.2 Description of the dataset

Blogster is a community blogging platform that features specific-interest blogs. Blogster features are a combination of blogging and social networking. Bloggers can create both their profiles and their blogs. They can also add other bloggers with the same interests as friends, chat with other bloggers, join groups, upload multimedia contents, and even incorporate their blog RSS feed, Twitter and Flickr accounts. Blogster is a small and focused community of bloggers where a small portion of users are active and socially engaged. We attempted multiple crawling, each time from a different seed node, but ended up discovering $17,436$ nodes. The social graph formed by these nodes has $72,907$ edges with 17 connected components. The largest connected component has $14,323$ nodes and $64,888$ edges. So, the largest component has about 82% nodes of the network. This confirms the existence of a giant component in the network. In real world networks, the giant component fills most of the network—usually more than half and not infrequently over 90%—while the rest of the network is divided into a large number of small components (e.g., 17 components in the network) disconnected from the rest [18].

For each user, we collected all visible attributes scraping HTML pages. Blogster has a public post counter, where it shows the number of posts bloggers have already contributed. Our crawled bloggers contributed $329,114$ posts out of $362,123$ posts on Blogster, as seen by the post counter. So, bloggers from our dataset contributed about 91% of total posts. Moreover, note that we could not crawl about 8% profiles of bloggers in our dataset due to their "community" and "private" profile settings. These bloggers should contribute a portion of the 9% blog posts that we were unable to trace. As such, our dataset is a good representative sample of the Blogster community.

3. RESEARCH QUESTIONS

User retention is extremely important not only for community blogs, but also for any organizations where users contribute to the profit. For example, according to Bain and Co., a 5% increase in customer retention can increase a company's profitability by 75% [13]. We are interested to examine factors that cause higher retention in Blogster. We have several hypotheses that are related to the following research questions:

1. What variables predict high retention?

2. How well do these variables predict user retention?

Measuring retention is not a trivial task. In Twitter, Java et al. [9] found that active users have higher retention. So, simply putting this forward in the context of Blogster, retention is how active a blogger is in the network, in other words the user's engagement or participation with the community. One can measure a user's participation in the network by continuously monitoring her activities. In Blogster, participation is measured by points. To encourage participation Blogster has a point system. The point a user gets depends on the specific actions she takes [1]. We take the points explicitly stated on a blogger's profile as an indication of her retention. Users with higher points have higher retention.

We categorized the predictor variables of retention into five categories:

- Network metrics specific variables (e.g., centralities, clustering coefficient)
- User activity specific variables (e.g., posts, comments, photos, network age)
- User physiology oriented variables (e.g., age, gender)
- Interactional (e.g., blog traffic, other users' comments)
- Relational (e.g., social tie strength, friends retention)

Based on these variables, we have made the following hypotheses.

3.1 Network metrics with retention

Sociologists agree that power is a fundamental property of social structure. Network analysts often describe the way that a user is embedded in a social network as imposing constraints on the user, and offering the user opportunities, at the same time. Users who face fewer constraints, and have more opportunities than others are in favorable structural positions. Having a favored position means that a user may extract better bargains in exchanges, have greater influence, and that the user will be a focus for deference and attention from those in less favored positions [7]. Intuitively, such positions demand an extensive attention from the user, e.g., one has to show a significant amount of activities. So, we may expect that theses bloggers should have higher retention.

Briefly "having a favored position" can be explained by having "more opportunities" and "fewer constraints". Although, there are no uniquely acceptable measures to quantify this phenomenon, sociologists proposed different centrality metrics. For example, the larger the number of direct neighbors, the larger an audience the node has for direct communication. Alternatively, the larger the number of paths between other pairs of nodes a node is part of, the more it can control the communication between distant nodes. We hypothesize that a blogger's retention can be determined by its centrality in the blogging community. The more central positions a blogger occupies in the network, the more retention she has. We selected five representative centrality metrics as the focus of our study: degree, betweenness, closeness, pagerank, and communicability centrality.

In Blogster, relationships are dyadic. However, social scientists have shown that triadic relationships are crucial as they offer far greater insights into the connectedness of egonetworks [23]. For example, in real world, relationships between two individuals is stronger if they have a mutual friend rather than having no mutual friends. While it is common in social networks, for the neighbors of a node to be connected among themselves, lack of triadic relationships are not also common. This lack of triadic relationships is called "structural holes" in the network and has first been studied in this context by Burt [1]. Structural holes around node u's neighborhood can be a good thing for her, because lack of connections between two of her friends give u power over information flow between them. If two friends of u are not friends and their information about one another comes instead via their mutual connection with u then u can control the flow of that information [18]. The local clustering coefficient is a measure of such structural holes. The local clustering coefficient measures how influential a user is in this sense, taking lower values the more structural holes there are in the network around the user. We hypothesize that the more local clustering coefficient a blogger has (less structural holes), the less retention she has.

3.2 Activities with retention

Bloggers are involved with three types of content specific-activities. They write blog posts, upload photos and comment on other bloggers' posted blogs. It is expected that an active blogger will produce higher number of content. We hypothesize that user activities (e.g., posts, comments, photos) predict retention but to different degrees.

3.3 Physiology with retention

We hypothesize that physiology such as gender and age have effects on blogger retention. How gender is expressed in and influences online social interaction has been explored in an article written by Herring [8]. Commonly held belief is that online spaces lack physical and auditory clues, thus makes the gender of online users irrelevant or invisible, and allows men and women to participate equally, in contrast with traditional patterns of male dominance observed in face-to-face communications. However, the rise of social networks has changed this picture dramatically: for example, recent news [2] shows that women form a majority of Facebook and Twitter users, as well as they are dominating Pinterest; however, men are the majority of users on Google+ and LinkedIn. Not only social networks, collaborative online spaces like Wikipedia has shown a similar pattern. A 2010 study co-sponsored by the Wikimedia Foundation discovered that barely 15 percent of Wikipedia contributors are women, with the lion's share of the articles being written, edited and updated by men in their mid-20s[3]. Based on these studies, we believe that male and female will have different levels of activities and thus retention in Blogster. Also, taking lessons from Wikipedia editors, our intuition is that most of the users in BlogSter are in their mid-20s and their retention increase with age.

3.4 Interaction with retention

We think the extent to which bloggers interact with high retention bloggers is different than they do with low reten-

[1]www.blogster.com/help#earnpoints

[2]http://mashable.com/2012/07/04/men-women-social-media/

[3]http://news.discovery.com/tech/is-there-a-gender-gap-online.htm

tion bloggers. This interaction can be measured in two ways: explicit and implicit. Explicit interactions are those that contribute content, e.g., bloggers might post comments on other bloggers' blog posts. We expect that high retention users will receive more explicit interactions from others (as a form of comments). Implicit interactions do not produce any content, rather they show other bloggers' interest on the blogs posted by the blogger. For example, the web traffic a blogger gets could be an indicator of how many times her blogs are read by others. We expect that high retention bloggers' blogs will be read by more people, they will receive more web traffic.

3.5 Friendship with retention

Similarity fosters connection—a principle commonly known as homophily, coined by the sociologists in the 1950s. Homophily is our inexorable tendency to link up with other individuals similar to us. The result is that our personal networks are homogeneous with regard to many sociodemographic, behavioral and intrapersonal characteristics [16]. The presence of homophily has been discovered in a vast array of social network studies, including age, gender, class and organizational role [16]. Taking lessons from those studies, we hypothesize that if two users have higher tie strength between them, they will show homophily in terms of retention. In other words, a strong tie will reduce retention imbalance between two individuals in the network.

4. RESULTS

Network metrics and retention As discussed previously, we selected five representative centrality metrics to show their relations with blogger retention: degree, betweenness, closeness, pagerank, and communicability centrality. Degree centrality is defined as the number of links that a node has. Although simple, degree centrality intuitively captures an important aspect of blogger's potential retention: bloggers who have connections to many others are read by more people, have access to more information, and certainly have more prestige than those who have fewer connections. High degree centrality bloggers can reach many bloggers directly.

Bloggers with high betweenness centrality may have considerable influence within a network by virtue of their control over information passing between others: they can comment, annotate, re-interprete the posts originating from a distant blogger and these altered views can be seen by other remote bloggers. The nodes with highest betweenness are also the ones whose removal from the network will most disrupt communications between other nodes because they lie on the largest number of paths taken by messages [18]. Formally, the betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through :

$$C(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{1}$$

where v is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t) paths, and $(s,t|v)$ is the number of those paths passing through some nodes v other than s,t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t, \sigma(s,t|v) = 0$.

Closeness centrality measures the mean distance from a node to other nodes, assuming that information travels along the shortest paths. Formally, the closeness centrality $(C(x))$ of a node x is defined as follows:

$$C(x) = \frac{n-1}{\sum_{y \in U, y \neq x} d(x,y)} \tag{2}$$

where $d(x,y)$ is the distance between node x and node y; U is the set of all nodes; d is the average distance between x and the other nodes. In our blogging network this centrality measure estimates the amount of information a blogger may have access to compared to other bloggers. Specifically, a blogger with lower mean distance to others can reach others faster.

To account for the fact that not all communications take place along the shortest path, we also considered communicability centrality. This centrality measure is defined as the sum of closed walks of all lengths starting and ending at the node [4].

Originally designed as an algorithm to rank web pages [20], PageRank computes a ranking of the nodes in a graph based on the structure of the incoming links. The algorithm assigns a numerical weighting to each node of a network with the purpose of "measuring" its relative importance within the network.

Furthermore, we computed local clustering coefficient (CC) of each blogger based on [18], which is the following:

$$CC = \frac{\text{\# of pairs of neighbors of u that are connected}}{\text{\# of pairs on neighbors of u}} \tag{3}$$

Net. metrics	Corr. (Metric, Points)	95% CI
Degree	0.60	0.5948374-0.6172682
Betweenness	0.49	0.4649659-0.4922978
Closeness	0.29	0.2715457-0.3040645
Pagerank	0.57	0.5565211-0.5805153
Communicability	0.57	0.5543655-0.5784447
Clustering coefficient	-0.40	-0.4151026 -0.3739151

Table 1: Correlations between social network metrics and bloggers' points. All values are statistically significant with $p < 0.05$ and 95% CI are shown.

Table 1 shows Pearson correlations between social network metrics and bloggers' points. As we speculated, all centralities are positively correlated with points. However, surprisingly, although degree centrality is the most simplest centrality measure in terms of computational complexity, it has the highest correlation with points. Closeness centrality is the least predictor of points among all centralities. Clustering coefficient is negatively correlated with points, meaning that the more structural holes a blogger has, the more points she possesses.

Activities and retention We considered the activities a user involved in the community and investigated to what degree they contribute to her retention. In BlogSter, users are involved in three activities: writing blogs, commenting on other blogs and uploading multimedia contents (e.g., photos). Figures 1 (a), (b) and (c) show plots between the number of posts, comments and photos versus points respectively. These plots indicate that higher number of posts, comments or photos mean higher blogger points. We computed Pearson correlations between activities and bloggers' points, which is shown in Table 2. It seems that the number of posts and comments are very good predictors of points, correlation coefficient $r = 0.89$ and $r = 0.84$ respectively. The number of photos a blogger upload also has a good correlation with her points ($r = 0.57$).

Activity	Corr. (Activity, Points)	95% CI
Posts	0.89	0.8909028-0.8983902
Comments	0.84	0.8337773-0.8461999
Photos	0.57	0.5550105-0.5867388

Table 2: Correlations between activities and bloggers' points. All values are statistically significant with $p < 0.05$ and 95% CI are shown.

We built a multiple linear regression model to see how well only these three predictors can predict bloggers' points. Note that, bloggers' points are dependent on a number of other activities (e.g., a blogger is penalized for deleting a comment). Our model is the following:

$$Points_i = \alpha + \beta_1.Posts + \beta_2.Comments_i + \beta_3.Photos_i + \epsilon_i \quad (4)$$

The results are shown in Table 3. The model is statistically significant with p-value: $< 2e-16$ and Adjusted R-squared: 0.954. Although we knew that the predictors used in the regression are key elements of bloggers' points, it's interesting to observe that how accurate they can predict user points. Only the number of posts, comments and photos can predict 95.4% variations about points.

	Estimate (β)	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1.11e+02	5.85e+00	18.95	$< 2e-16$ ***
Posts	1.26e+01	5.85e+00	228.15	$< 2e-16$ ***
Comments	8.17e-01	5.24e-033	155.88	$< 2e-16$ ***
Photos	4.74e-01	5.30e-02	8.94	$< 2e-16$ *

Table 3: The results of a multiple linear regression with points as the dependent variable. Significance codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

We found a negligible correlation between bloggers' points and their ages in the network ($r = 0.07, p < 0.05$). This result is not surprising, as one might join the network, but remain inactive afterwards.

Physiology and retention In Blogster, we found that the number of female bloggers is higher than the number of male bloggers: the ratio is 1.61:1.0. Interestingly, male bloggers have higher retention than female bloggers, as seen by their means from the descriptive statistics in Table 4. However, it's difficult to distinguish a difference between their respective cumulative distributions of points from Figure 2 (a). So, we performed two statistical tests, namely, two-sample Kolmogorov-Smirnov tests and permutation tests. Following two non-parametric test results confirm that male and female samples are not the same and the mean difference is statistically significant. Two-sample Kolmogorov-Smirnov test results: D = 0.0451, p-value = 5.791e-05. Permutation tests results: Z = 4.3974, p-value = 1.095e-05, mean difference=212.6047.

Gen.	Min.	1st Qu.	Med.	Mean	3rd Qu.	Max.	Std.
Male	0.0	101.0	178.0	705.3	369.0	97140.0	2982.639
Female	0.0	102.5	165.0	492.7	330.0	60220.0	2049.942

Table 4: Descriptive statistics of points for male bloggers and female bloggers.

A descriptive statistics on bloggers' age are shown in Table 5. The mean age of bloggers is about 32 and most of the bloggers are between 20 to 30 years of old (see a histogram

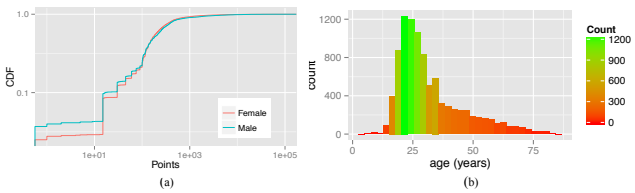

Figure 2: (a) Cumulative distribution functions of points for male and female; (b) histogram of bloggers' age.

in Figure 2 (b)). We computed the Pearson correlation between bloggers' points and their ages. We found a positive correlation between them with $r = 0.21, p < 0.05$.

Min.	1st Qu.	Med.	Mean	3rd Qu.	Max.
5.427	22.490	27.680	32.030	37.850	82.110

Table 5: Descriptive statistics of age.

Friends retention and social tie strength with retention We first show that bloggers show homophily in terms of retention. For each blogger, we computed the average retention of her blogger friends by taking the geometric mean of their points . We got a positive Pearson correlation coefficient ($r = 0.33, p < 0.05$) between blogger retention and average retention of her friends.

Next we show that a strong tie between two bloggers can even reduce their retention difference. We used social proximity as a metric to characterize the strength of the relationships between two bloggers and understand whether social proximity reduces retention imbalance between them. The social proximity metric is based on a previous study [19] that suggests that the overlap between the social neighborhood of two individuals is a good indicator of the strength of their relationship. We assessed the strength of the relationship between two connected bloggers by the overlap between their sets of friends, computed as follows:

$$Overlap_{uv} = \frac{m_{uv}}{((k_u - 1) + (k_v - 1) - m_{uv})} \quad (5)$$

where m_{uv} is the number of common neighbors between user u and v, k_u is the number of neighbors of user u and k_v is the number of neighbors of user v. Figure 3 (a) shows a CDF of the overlap in friendship networks between two connected bloggers. From the distribution, we observe that 95% of blogger pairs have less than 25% network overlap. From the distribution of point differences in Figure 3 (b), we see that 63% point differences are less than 1000.

The Pearson correlation coefficient between social proximity (represented by network overlap) and the point difference between two friends is -0.38 ($p < 0.05$).This confirms our hypothesis that if two bloggers have higher social proximity, their retention difference will be lower.

Interaction with retention We asked whether bloggers explicit interactions (e.g., comments) and implicit interactions (e.g., blog traffic) have a relation to retention. Figure 4 (b) shows a plot between traffic on blogs and bloggers' points. The plot clearly shows a positive correlation and that is confirmed by the Pearson correlation coefficient ($r = 0.74$ with $p < 0.05$). We finally run a linear regression between

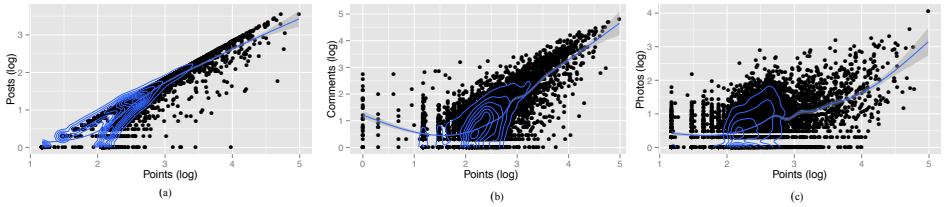

Figure 1: Bloggers' (a) posts vs. points; (b) comments vs. points; (c) photos vs. points.

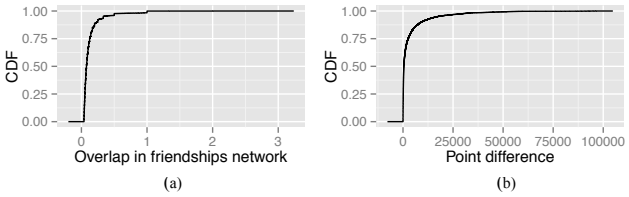

Figure 3: (a) CDF of overlap in friendship networks; (b) CDF of point difference between two bloggers who are friends.

points and blog traffic. The regression coefficient is statistically significant ($p < 0.001$), positive, and corresponding R^2 is as high as 0.56. These results suggest that the more retention a user has, the higher traffic she gets on her blogs.

The plot between the number of comments bloggers get from other bloggers and their points (in Figure 4 (a)) shows a positive relation between them and that is also confirmed by the Pearson correlation coefficient ($r = 0.83$ with $p < 0.05$). We also run a linear regression between points and the number of other bloggers' comments. The regression coefficient is statistically significant ($p < 0.001$)' positive, and corresponding R^2 is as high as 0.69. These results suggest that the more retention a user has, the higher number of comments she gets on her blogs.

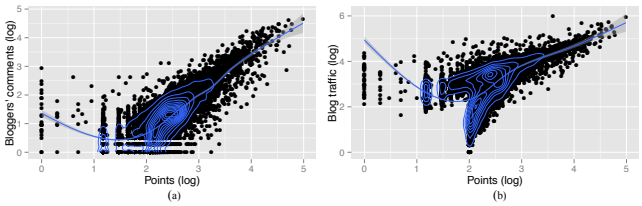

Figure 4: (a) Bloggers' comments vs. points; (b) blog traffic vs. points.

5. PREDICTING RETENTION

We now turn to the core of our results: how well do these variables predict user retention? Note that we have five different types of predictor variables—network metrics (degree, betweenness, closeness, pagerank, communicability and clustering coefficient), activity metrics (posts, comments, photos), physiological (age, gender), interactional (blog traffic, other users' comments) and relational (e.g., social tie strength, friends retention).

Our previous results show that activities can highly predict (95.4%) retention. In fact, in Blogster activities are most contributing factors of points. The number of points users get depends on the specific action they take. For example, writing a blog post contributes 15 points, writing a comment per single blog post contributes 2, adding a photo as avatar gives 15 points, deleting a blog post reduces points by 15 and deleting a comment reduces points by 2. So, we plan to exclude all activity specific variables from our prediction. Our goal is to investigate to what extent variables that are not related to point calculation can predict user retention.

As an input processing step, first, we rank users based on each centrality metric. Specifically, each centrality metric assigns each node a score that can be used to order users in decreasing order of importance (according to that centrality). This allows each blogger to receive a rank according to each centrality metric: the first ranked blogger will be the most central one, the last ranked will be the one with the lowest centrality score. Bloggers having the same centrality score are given the same rank.

	Estimate (β)	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	3.66e+03	1.86e+02	19.62	$< 2e-16$ ***
DegreeRank	-1.92e+01	1.07e+00	-17.98	$< 2e-16$ ***
Comm.Rank	-1.57e-02	4.25e-03	-3.68	0.000233 ***
CC	-8.71e+01	4.06e+01	-2.14	0.032081 *
BlogTraffic	3.62e-02	8.21e-04	44.06	$< 2e-16$ ***
UserComments	1.02e+00	1.59e-02	64.61	$< 2e-16$ ***
Age	5.03e+00	7.29e-01	6.90	$5.45e-12$ ***
Gender	-9.38e+00	1.86e+01	-4.50	0.001257 **
AvgFriendsRet	1.53e-04	2.04e-03	3.08	0.021513*

Table 6: The results of a multiple linear regression with points as the dependent variable. We include only a selection (that were not used in points calculation) of predictor variables. Estimates are not standardized; they remain on their original scales. Significance codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

Figure 6 shows cumulative distributions of various ranks. One of the objectives of plotting these distributions is to show how granular the ranks are, more specifically, how successful these centrality metrics are in assigning distinct scores to different nodes in the network. To this end, analyzing the distributions we get these facts: 5% of the bloggers cover the top 86% of the ranks in degree centrality scores, 10% of the bloggers corresponds to top 17.7% ranks in closeness centrality, 10% of the bloggers corresponds to top 14.9% ranks in betweenness centrality, 10% of the bloggers rank within 13.5% rank on pagerank distribution and 10% bloggers within top 15.5% ranks on communicability rank dis-

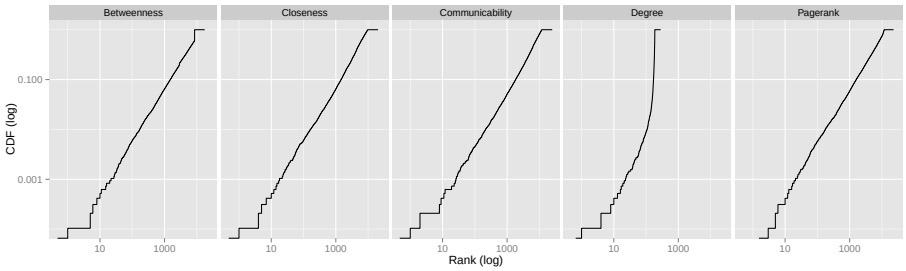

Figure 6: Cumulative distribution functions of ranks.

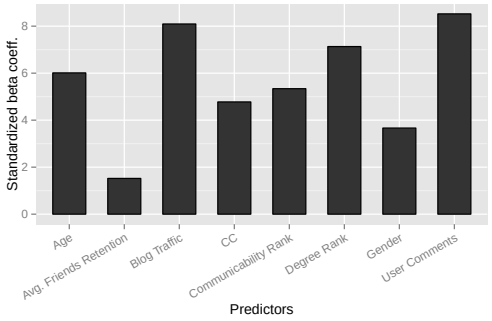

Figure 5: Standardized beta coefficients (β) show the relative effect sizes that each input variable has on points. Absolute values are taken for standardized beta coefficients.

tribution. So, we observe that all centrality measurements except degree centrality show granular scale of ranking, that is, they are typically capable of assigning a distinct score to each blogger (e.g., 10% bloggers within top 13.5% ranks).

We considered all explanatory variables other than activity specific variables to build a multiple regression model to predict retention (as reflected by points). However, not all predictors are significant in explaining the variability of points. So, our regression went through a variable selection process. We used an "all possible regression" approach to select mode predictors. The all possible regression approach considered all possible subsets of the pool of explanatory variables and found the model that best fits the data according to adjusted R^2. Finally, we considered the model that yielded maximum adjusted R^2. The model is the following:

$$\begin{aligned} Points_i = \alpha + \beta_1.DegreeRank_i + \beta_2.CC_i + \\ \beta_3.CommunicabilityRank_i + \beta_4.BlogTraffic_i \\ + \beta_5.UserComments_i + \beta_6.Age_i + \\ \beta_7 Gender_i + \beta_8 AvgFriendsRet_i + \epsilon_i \end{aligned} \quad (6)$$

The results of the regression are shown in Table 6. The adjusted R^2 of the model is 0.837, which implies the model can explain 83.7% of variation around points. The unstandardized β coefficients in Table 6 are useful in that they can be directly interpreted according to the native units of each predictor: for each one unit change in the predictor variable, the count of the response variable (points) is expected to change by the respective *beta* coefficients (all else being equal). As expected, higher degree rank, communicability rank and clustering coefficient mean lower points with corresponding β are $-1.92e+01$, $-1.57e-02$ and $-8.71e+01$ respectively. However, the more web traffic and other users' comments a blogger gets, the more points she can expect with $\beta = 3.62e-02$ and $\beta = 1.02e+00$ respectively. Also, being female suggests less points $\beta = -9.38$ and aged user might expect higher points with $\beta = 5.03$. Furthermore, if a blogger's friends has higher retention, we might expect her retention higher with $\beta = 1.53e-04$.

While β coefficients are valuable for a broad range of prediction and forecasting purposes, we are also interested in comparing the relative impact of each predictor. We report the standardized beta β coefficients in Figure 5. From the figure we observe the number of other users' comments, web traffic and the degree rank are the most influential or significant predictors. The rest of the predictors can be serialized from the most significant to the least significant ones as age, communicability rank, clustering coefficient, gender and average friends retetion.

6. RELATED WORK

Researchers have explored the factors that contribute to increased continued participation or higher retention of users in online settings. In online discussion groups, Joyece et al. [11] tested whether the responses that new users receive to their first posts influence the extent to which they continue to participate in the community. They found that those new users who receive a reply to their initial post, are 12% more likely to post to the community again. In Facebook, Burke et al. found that new users contribute more contents if they see their friends are also contributing and if they receive feedback and a wide audience. Our study complements these studies by discovering some novel factors of retention. However, outside of blogs, literature suggests that social relationships are important for continued participation. Pre-existing social relationships have been linked with recruitment on political and social movements [3, 15]. Analysis of online special interest groups (e.g., newsgroup) showed that group members with a strong sense of attachment to a group are more likely to participate [12] and closely connected groups are more supportive of members [24]. Stiggelbout et al. [22] investigated factors to continued exercise participation among older adults. They found female sex, younger age, being married, being a non-smoker and being in paid employment are good factors for continued exercise. In our study, we have also found that age and gender are also good factors of continued participation.

Our work is conceptually closer to the work done by Lento

et al. [14] . They examined the relationship between social relationships and continued participation (as expressed through various features of the system) in the Wallop system. Wallop was a personal publishing and social networking system designed by Microsoft Research, where an individual gains an access to the system when she receives an invitation from an existing Wallop user. The study found that pre-existing networks (e.g., users who maintain a connection with the person who invited them) and the number of social ties have an effect on retention. Another interesting finding was that cultural elements play a role in user retention: chinese language users have higher retention than english language speakers. However, this study has several limitations. First, the social network is not a declared social network. It's a small and implicit network that the authors has built from users' activity traces (e.g., comments, who invites whom to join). So, the network is highly sparse ($3,119$ nodes, $4,323$ edges), an unlikely case for a social network. Second, the retention metric is poorly defined. They defined a user as active if she posted comments during a 5 week period of a month. The study acknowledged this weakness and pointed out that "content uploads would be a better measure, but unfortunately this data was not available at the time of this work" [14]. Our study overcomes the limitations of this work by using a relatively bigger ($17,436$ nodes, $72,907$ edges), declared social network and by using content upload behavior as a metric of retention.

7. SUMMARY

In this paper, we asked why some bloggers have higher retention in a community blogging platform, while others fail to contribute in the long run? We crawled a sample of blogger profiles from Blogster, who contributed about 91% posts in the community. From sociological and psychological studies, we derived five categories of variables that are related to bloggers retention. We showed to what extent these variables relate to retention and built a predictive model for retention prediction. We found that male and aged (senior) bloggers, who face fewer constraints and have more opportunities in the community and have friends with higher retention are more retained in the community than others. These bloggers also get higher attention from others as reflected by higher explicit and implicit interactions from others. Our work can be used as a foundation for further study of community bloggers retention. System developers of community blogs could also leverage results of this paper and build retention aware community blogs.

8. REFERENCES

[1] R. Burt. Structural holes: The social structure of competition. 1995.

[2] J. N. Cummings, B. Butler, and R. Kraut. The quality of online social relationships. *Commun. ACM*, 45(7):103–108, July 2002.

[3] M. Diani and D. McAdam. *Social movements and networks: Relational approaches to collective action: Relational approaches to collective action*. Oxford University Press, 2003.

[4] E. Estrada and J. A. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E 71*, 056103, 2005.

[5] D. Gillmor. *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly, 2006.

[6] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, 2011.

[7] R. Hanneman and M. Riddle. Introduction to social network methods. 2005.

[8] S. C. Herring. 9 gender and power in on-line communication. *The handbook of language and gender*, 25:202, 2008.

[9] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[10] Q. Jones, G. Ravid, and S. Rafaeli. Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Information systems research*, 15(2):194–210, 2004.

[11] E. Joyce and R. E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747, 2006.

[12] P. Kollock and M. Smith. Managing the virtual commons. *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, pages 109–128, 1996.

[13] A. Lawrence. Five customer retention tips for entrepreneurs. `http://www.forbes.com/sites/alexlawrence/2012/11/01/five-customer-retention-tips-for-entrepreneurs/`, November 2012.

[14] T. Lento, H. T. Welser, L. Gu, and M. Smith. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *3rd Annual Workshop on the Weblogging Ecosystem*, volume 12. Citeseer, 2006.

[15] D. McAdam. *Freedom summer*. Oxford University Press, 1990.

[16] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[17] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why we blog. *Communications of the ACM*, 47(12):41–46, 2004.

[18] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[19] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. *Stanford InfoLab*, 1999.

[21] L. Rao. Wordpress now powers 22 percent of new active websites in the u.s. `http://techcrunch.com/2011/08/19/wordpress-now-powers-22-percent-of-new-active-websites-in-the-us/`, August 2011.

[22] M. Stiggelbout, M. Hopman-Rock, M. Crone, L. Lechner, and W. Van Mechelen. Predicting older adults' maintenance in exercise participation using an integrated social psychological model. *Health Education Research*, 21(1):1–14, 2006.

[23] M. Tortoriello and D. Krackhardt. Activating cross-boundary knowledge: the role of simmelian ties in the generation of innovations. *Academy of Management Journal*, 53(1):167–181, 2010.

[24] B. Wellman. An electronic group is virtually a social network. *Culture of the Internet*, 4:179–205, 1997.