Retention in Online Blogging: A Case Study of the Blogster Community

Imrul Kayes and Jacob Chakareski, Senior Member, IEEE

Abstract—Community-based blogging platforms can be rich sources of information on a variety of specialized topics, from finance to parenting. The usefulness of such platforms depends heavily on user participation and contribution. However, one potential problem is lower retention: users’ fail to contribute in the long run. This paper is an investigation of retention in the popular community blogging platform “Blogster.” We use the points users earn for their activities as a proxy of retention and explore the attributes that are associated with their retention. We find that highly retained users are most central in the network and their blogger friends are mutually less connected. They get more views and comments to their posted blogs. We also examine the homophily of retention in the social network. Based on our empirical observations, we build a classifier that is able to detect top retained users with accuracy as high as 94%. Our work has theoretical implications for the social behavior literature of community bloggers and practical design implications for potential community blogging platform developers.

Index Terms—Bloggers, community blogging, retention.

I. INTRODUCTION

MILLIONS of Internet users use blogging for publishing daily journals, expressing opinions or ideas, and sharing knowledge. A blog is a personal journal published on the Web consisting of discrete entries (“posts”) typically displayed in reverse chronological order [1]. 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 web pages. Community blogging platforms have made traditional blogging more interactive by adding additional social and participatory features. They allow the creation of online profiles in which links to other bloggers are specified. This blogger to blogger declared social ties that specify the blogger’s interests and endorsement of other bloggers, creating a social network through which blog updates are automatically disseminated.

Community blogging platforms have enabled a new, grassroots form of journalism and are often viewed as a way to shape democracy outside the mass media and conventional party politics [2]. As such, they have become immensely popular and have already shown multidimensional impacts. 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 [3]. Citizen journalism had a high impact on major events such as South Asia tsunami, London terrorist bombings, and New Orleans Hurricane Katrina [2]. The blogosphere, the virtual universe of the blogs on the web, thus provides a conducive platform for different aspects of virtual and real life, such as viral marketing [4], sales prediction [5], [6], business models [7], and counter terrorism efforts [8].

Although community blogging is very popular and rapidly growing, they are facing a number of challenges. First, participation is often sparse and uneven [9]. For example, Cummings et al. [10] examined a listserv-based online group and found that one-third of all listed users had no communication during a 3-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 [9]. For example, Jones et al. [11] examined a sample of users from the Usenet newsgroups and found that only 11.5% of the people who posted in 1 month returned to post in the second month.

Existing research, based on content analysis [12] and interviews [13], provides an excellent insight into a user’s motivation for joining a blogging community. These studies reveal 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. In [15] and [16], we present some early results on bloggers’ continued participation. This work is a follow up on our previous study and we extend it by performing more in-depth analysis of user retention.

Our goal is to examine factors that are related to a blogger’s continued participation 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 bloggers contributed about 91% of posts in the community, and hence are a good candidate for a representative sample of the bloggers in Blogster (Section II). The Blogster social graph has power law degree distribution and a small world property (Section III).

Manuscript received September 12, 2014; accepted October 09, 2015.

I. Kayes is with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: imrul@mail.usf.edu).

J. Chakareski is with the Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL 35487 USA.

Digital Object Identifier 10.1109/TCSS.2015.2495135

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

We proposed five categories of variables: network-specific (e.g., centralities, clustering coefficient (CC)), 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 are related to continued participation or higher retention and to what degree (Section IV).

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 V). We also use the variables to predict top retained bloggers. We try a number of machine learning classification algorithms and found that stochastic gradient boosted trees (SGBT) can predict top retained users with a high accuracy (93.62%).

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.

II. DATA COLLECTION

Blogster is a mature and popular blogging platform, which has a combination of blogging and social networking features (see the front page of the community in Fig. 1). In the platform, bloggers can create their online profiles and blogs. Bloggers can write blog posts on a predefined set of categories (e.g., Computing and Internet) post comments and ratings on other bloggers’ blog posts. They can also add other bloggers as friends, chat with other bloggers, join groups, and upload multimedia contents. Fig. 2 shows a profile of a blogger with recent posts and various statistics on her profile.

A. Sampling Technique

Breadth-first-search (BFS) algorithm is the most widely used technique for sampling OSNs [17], [18]. Unfortunately, BFS sampling is biased toward high degree nodes [19], [20]. Another popular technique random walk (RW) sampling also leads to bias toward high-degree nodes. Hence, we modify the metropolis-hastings RW (MHRW) algorithm [21] to crawl the Blogster network. The MHRW algorithm is capable of obtaining a uniform sample (or more generally a probability sample) of OSN users [21].

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 an initially selected node (seed node) and proceeds iteratively. The selection of a seed node is described later in this section. In each iteration, we visit a node and discover all its neighbors. If the current node is $u$, the next hop node $v$ is chosen according to the following transition probability:

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

where $P_{u,v}^{\text{MH}}$ implies the following algorithm.

Fig. 1. Front page of the Blogster blogging community.
Algorithm 1. Crawling algorithm

\[ u \leftarrow \text{seed} \]

while stopping criteria not met do
    Select node \( v \) uniformly at random from neighbors of \( u \)
    Generate uniformly at random a number \( 0 \leq p \leq 1 \)
    if \( p \leq \frac{k_u}{k_v} \) then
        \( u \leftarrow v \)
    else
        Remain at \( u \)
    end if
end while

At each iteration, at the current node \( u \), the algorithm randomly selects a neighbor \( v \) and moves there with probability \( \min(1, \frac{k_u}{k_v}) \), where \( k_u \) is the number of friends of user \( u \) and \( k_v \) is the number of friends of user \( v \). The algorithm accepts the move toward a node of smaller degree, and rejects some of the moves toward higher degree nodes. This pattern of moving eliminates the bias toward high-degree nodes.

B. Selection of the Seed Node

We wanted to select highly retained bloggers as seed nodes. Blogster shows a list of bloggers who are online at any given time on its who is online? feature. We wrote a Python crawler to collect all the listed users from April 13, 2014 to June 19, 2014 (67 days). The crawler was activated on each minute of those days and collected 91,965 observations. An observation contains a list of users who were online at that time. We calculate a metric, online presence that quantifies how many times a user was online on the observations. Fig. 3 shows the complementary cumulative degree distribution function (CCDF) of the online presence. In general, the churn was very low. Only 0.12% bloggers were logged on the system for one time.
However, 99.66% bloggers were logged on the system for more than 30 times. About 10% bloggers were logged on the system for more than 7843 times.

Fig. 4 shows the online presence (exact values) of the top 20 users (we anonymized user IDs). It appears that top users spent a significant amount of time on blogging. All the top 20 users were found on 21% of the observations. About 30% of the observations had the top three bloggers. We selected these top three bloggers as seed nodes and attempted three separate crawls from different seed nodes.

C. Implementation and Challenges

We wrote a multithreaded Python web-crawler for the sampling and crawling purposes. The crawler ran on a machine equipped with 2-GHz Intel Core i7 processor, 4-GB 1333-MHz DDR3 RAM, Mac OS X Lion 10.7.5 operating system. Blogster is a small but focused community of bloggers. As such, a distributed crawler was not required. However, there are several practical challenges we faced while crawling the Blogster. First, we found that the application program interface (API) calls of Blogster are restrictive. So, we were forced to scrape the HTML pages. Second, friends list page of the Blogster enables asynchronous loading of web content (i.e., AJAX). This demands more sophisticated customization of the crawlers (e.g., multiple http calls for a users’ friends). Finally, online social networks (OSNs) usually have defense mechanisms against large-scale crawling. So, we limit our crawler to 500 requests per hour.

D. 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 different seed node (as discussed earlier), but each time ended up discovering 17,436 nodes. The social graph formed by these nodes has 72,907 edges with 17 connected components. The largest connected components has 14,323 nodes and 64,888 edges. So, the largest component has about 82% nodes of the network. This confirms an 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 [22].

For each user, we collected all visible attributes scraping HTML pages. These attributes are: visibility of the profile, age (from Birthdate), sex, location, marital status, joined date, job, language, blog traffic, number of blog posts, number of comments the user made, number of comments other users made to the users’ posted blogs, number of photos, the list of friends, points, and the time when the user was last online. Blogster has three types of privacy settings for a profile: public, community, and private. If a blogger sets her profile to public, anyone can see her profile. However, if a profile is set to community, only community members can see it. A private profile is not visible to anyone other than that blogger. We found that 92% bloggers’ profiles are public, 3.63% are open for community members, and 4.38% are private. These figures are consistent with a study on Facebook. Gross and Acquisti’s study [23] on Facebook shows that more about 90% user profiles are public. Although expected, this confirmation warns again about the importance of correct default settings (e.g., privacy as contextual Integrity [24], [25]) in online social networks.

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 from 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 community and private profile settings. These bloggers should contribute a portion of the remaining 9% blog posts we were unable to trace. As such, we think that our dataset is a good representative sample of the Blogster community. A summary of the dataset is shown in Table I.

III. Social Graph

Before proceeding with the analysis, we want to confirm that the graph we crawled resembles properties of a “social” graph. More specifically, we want the Blogster graph to have a power-law degree distribution and a small world property, two natural outcomes of real-world social networks [26], and the Blogosphere is not an exception [27].
A. Degree Distribution

The CCDF of the network is shown in Fig. 5. The degree distribution follows a power-law distribution [28]. About 80% of bloggers have less than 12 friends, where about 1% bloggers have more than 80 friends. The exponential fitting parameter \( \alpha \) for the distribution is 2.06 when \( X_{\min} > 2 \). Most real-world networks with power-law degree distributions have values of \( \alpha \) in the range \( 2 \leq \alpha \leq 3 \) [22]. Blogster’s \( \alpha \) is similar to what Kwak et al. [29] found (\( \alpha = 2.27 \)) in microblogging platform Twitter. So, similar to Facebook and Twitter, Blogster is a scale-free network as all of them have power-law distribution. The most notable property of a scale-free network is the occurrence of hubs, which hold a much higher number of links than the average node. As hubs control the “connectedness” of the network, we expect that there are hub bloggers in the Blogster community who are essential for the platform.

B. Small World

A notable characteristic of the Blogster network is the high CC compared to a random network of the same size. Given a network \( G = (V, E) \), the CC \( C_i \) of a node \( i \in V \) is the proportion of all the possible edges between neighbors of the node that actually exist in the network [22]. The CC of the Blogster network is 0.217, which is quite high compared to a same size random network (0.0005). The small average path length (4.28), smaller than that of the corresponding erdos renyi random graph (4.59), together with the high average CC, places the Blogster social graph in the category of small-world graphs [30].

IV. 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% [31]. A Gartner Group’s statistic points out that 80% of a company’s future revenue comes from just 20% of existing customers. One final statistic provided by Lee Resource Inc. states that attracting new customers will cost a company five times more than keeping an existing customer [31]. We are interested to examine factors that are related to higher retention in Blogster. Note that blog sites can be categorized into individual blog sites and community blog sites [32]. Individual blog sites are owned and maintained by individuals, and they might want the blogs to be listed in a forum (e.g., blogcatalog\(^1\)). On the other hand, a community blog site is a single, stand-alone platform, where bloggers have their accounts and they write blogs on that site. The focus of this research is community blogging.

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?

However, measuring retention is a challenging task. On Twitter, Java et al. [12] 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 (shown in Table II). The point a user gets depends on the specific actions she takes.\(^2\) We take the points explicitly stated in 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:

1) network metrics specific variables (e.g., centralities, CC);
2) user activity specific variables (e.g., posts, comments, photos, network age);
3) user physiology-oriented variables (e.g., age, gender);
4) interactional (e.g., blog traffic, other users’ comments);
5) relational (e.g., social tie strength, friends retention).

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

A. 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 [33]. Intuitively, such positions demand an

---

\(^{1}\)[Online]. Available: http://www.blogcatalog.com/

\(^{2}\)[Online]. Available: www.blogster.com/help#earnpoints
extensive attention from the user, e.g., one has to show a significant amount of activities. So, we may expect that these 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 communications. 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 [34]. For example, in the real world, relationships between two individuals are 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 is also not uncommon. This lack of triadic relationships is called “structural holes” in the network and has first been studied in this context by Burt [35]. 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 [22]. The local CC is a measure of such structural holes. The local CC 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 CC a blogger has (less structural holes), the less retention she has.

B. 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.

C. Physiology With Retention

We hypothesize that physiology such as gender and age has effects on blogger retention. How gender is expressed in and influences online social interaction has been explored in the paper written by Herring [36]. Commonly held belief is that online spaces lack physical and auditory clues, thus making the gender of online users irrelevant or invisible, and allowing 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 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 [37]. Cunha et al. [38] found that male and female Twitter users differed in what hashtags they used for common topics. Not only social networks, collaborative online spaces like Wikipedia has shown a similar pattern. A 2010 study cosponsored by the Wikimedia Foundation discovered that barely 15% of Wikipedia contributors are women, with the lion’s share of the articles being written, edited, and updated by men in their mid-20s [39]. 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 increases with age.

D. Interaction With Retention

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

E. Friendship With Retention

Similarity fosters connection—a principle commonly known as homophily, coined by the sociologists in the 1950s. Even classical western philosophers observed the relationship between association and similarity. In Plato’s Phaedrus, he noted that similarity begets friendship [40]. Aristotle observed in Rhetoric and Nichomachean Ethics that people “love those who are like themselves” [41]. 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 [42]. The presence of homophily has been discovered in a vast array of social network studies, including age, gender, class, and organizational role [42]. 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.

V. RESULTS

A. 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.

Betweenness centrality, which measures the extent to which a node lies on the shortest paths between other nodes, was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network [43]. Bloggers with high betweenness centrality may have considerable influence within a network by virtue of their control over the information passing among others: they can comment, annotate, reinterpret the posts originating from a distant blogger and these altered views can be seen by other remote bloggers. The nodes with the highest betweenness are also the ones whose removal from the network will most disrupt communications between other nodes because they lie on the shortest paths between 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 [45]. If the social graph (G) is presented by adjacency matrix A, the communicability centrality of a user u can be obtained using a spectral decomposition of the adjacency matrix A

\[
SC(u) = \sum_{i=1}^{N} (v^u_i)^2 e^{\lambda_i}
\]

where \(v^u_i\) is an eigenvector of the adjacency matrix A of G corresponding to the eigenvalue \(\lambda_i\).

Table III 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. CC is negatively correlated with points, meaning that the more structural holes a blogger has, the more points she possesses.

### Table III

<table>
<thead>
<tr>
<th>Metric</th>
<th>Corr. (metric, points)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>0.60</td>
<td>0.5948374 - 0.6172682</td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.49</td>
<td>0.4649659 - 0.4922978</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.29</td>
<td>0.2775457 - 0.3046045</td>
</tr>
<tr>
<td>Pagerank</td>
<td>0.57</td>
<td>0.5565211 - 0.5805133</td>
</tr>
<tr>
<td>Communicability</td>
<td>0.37</td>
<td>0.3543655 - 0.3784447</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>-0.40</td>
<td>-0.4151026 to 0.3739151</td>
</tr>
</tbody>
</table>

All values are statistically significant with \(p < 0.05\) and 95% CI are shown.

Originally designed as an algorithm to rank web pages [46], 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. Pagerank of user \(u_i\) is computed using the following equation:

\[
PR(u_i) = \frac{1 - d}{N} + d \sum_{u_j \in M(u_i)} \frac{PR(u_j)}{L(u_j)}
\]

where \(u_1, u_2, \ldots, u_N\) are the users under consideration, \(M(u_i)\) is the set of friends of \(u_i\), \(L(u_j)\) is the number of friends of \(u_j\), and \(N\) is the total number of users.

Furthermore, we computed local CC of each blogger based on [22], which is as follows:

\[
CC = \frac{\# \text{ of pairs of neighbors of } u \text{ that are connected}}{\# \text{ of pairs on neighbors of } u}.
\]
Fig. 6. Bloggers’ posts versus points.

Fig. 7. Bloggers’ comments versus points.

Fig. 8. Bloggers’ photos versus points.

Table IV: Correlations Between Activities and Bloggers’ Points

<table>
<thead>
<tr>
<th>Activity</th>
<th>Corr. (activity, points)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posts</td>
<td>0.89</td>
<td>0.8909028-0.8983902</td>
</tr>
<tr>
<td>Comments</td>
<td>0.84</td>
<td>0.8337773-0.8461999</td>
</tr>
<tr>
<td>Photos</td>
<td>0.57</td>
<td>0.5550105-0.5867388</td>
</tr>
</tbody>
</table>

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 as follows:

\[
\text{Points}_i = \alpha + \beta_1 \cdot \text{Posts}_i + \beta_2 \cdot \text{Comments}_i + \beta_3 \cdot \text{Photos}_i + \epsilon_i. \tag{6}
\]

Table V: Results of a Multiple Linear Regression with Points as the Dependent Variable

| 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+03 | 155.88  | < 2e-16 *** |
| Photos       | 4.74e-01   | 5.30e+02 | 8.94    | < 2e-16 * |

Significance codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 ” 1.

Table VI: Descriptive Statistics of Points for Male Bloggers and Female Bloggers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.0</td>
<td>101.0</td>
<td>178.0</td>
<td>705.5</td>
<td>369.0</td>
<td>97140.0</td>
<td>2982.639</td>
</tr>
<tr>
<td>Female</td>
<td>0.0</td>
<td>102.5</td>
<td>165.0</td>
<td>492.7</td>
<td>330.0</td>
<td>60220.0</td>
<td>2049.942</td>
</tr>
</tbody>
</table>

The results are shown in Table V. 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 is interesting to observe how accurate they can predict user points. Only the number of posts, comments, and photos can predict 95.4% variations about points.

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 afterward.

C. 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 VI. However, it is difficult to distinguish a difference between their respective cumulative distributions of points from Fig. 9. So, we performed two statistical tests, namely two-sample Kolmogorov–Smirnov tests and permutation tests. Following two nonparametric 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.
A descriptive statistics on bloggers’ age are shown in Table VII. The mean age of bloggers is about 32 and most of the bloggers are between 20 and 30 years old (see a histogram in Fig. 10). We computed the Pearson correlation between bloggers’ points and their ages. We found a positive correlation between them with \( r = 0.21, p < 0.05 \).

**TABLE VII**

DESCRIPTIVE STATISTICS OF AGE

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>First Qu.</th>
<th>Med.</th>
<th>Mean</th>
<th>Third Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>5.427</td>
<td>22.490</td>
<td>27.680</td>
<td>32.030</td>
<td>37.850</td>
<td>82.110</td>
</tr>
</tbody>
</table>

**D. 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 [47] 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:

\[
\text{Overlap}_{uv} = \frac{m_{uv}}{(k_u - 1) + (k_v - 1) - m_{uv}}
\]

where \( m_{uv} \) is the number of common neighbors between user \( u \) and \( v \), \( k_u \) is the number of neighbors (friends) of user \( u \), and \( k_v \) is the number of neighbors (friends) of user \( v \). Fig. 11 shows a cumulative distribution function (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 Fig. 12, 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.

**E. Interaction With Retention**

We asked whether bloggers explicit interactions (e.g., comments) and implicit interactions (e.g., blog traffic) have a relation to retention. Fig. 13 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 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 Fig. 14) 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.

VI. 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, pagerank, communicability, and CC), 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 (posts, comments, photos) from our prediction. Note that blog traffic and other users’ comments (labeled as UserComments later) a user receives from other bloggers are not activity specific variables. Our goal is to investigate to what extent the 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.

Fig. 15 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 distribution. So, we observe that all centrality measurements except degree centrality show granular scale of ranking, i.e., they are typically capable of assigning a distinct score to each blogger (e.g., 10% bloggers within top 13.5% ranks).

As discussed, we consider 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 goes through a variable selection process. We use an “all possible regression” approach to select the model predictors. The all possible regression approach considers 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 consider the model that yielded maximum adjusted \( R^2 \). The model is as follows:

\[
\text{Points}_i = \alpha + \beta_1 \cdot \text{Degree Rank}_i + \beta_2 \cdot \text{CC}_i + \beta_3 \cdot \text{Communicability Rank}_i + \beta_4 \cdot \text{Blog Traffic}_i + \beta_5 \cdot \text{User Comments}_i + \beta_6 \cdot \text{Age}_i + \beta_7 \cdot \text{Gender}_i + \beta_8 \cdot \text{Avg Friends Ret}_i + \epsilon_i.
\]

(8)

The results of the regression are shown in Table VIII. The adjusted \( R^2 \) of the model is 0.837, which implies that the model can explain 83.7% of variation around points. The unstandardized \( \beta \) coefficients in Table VIII 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 CC mean lower points with corresponding \( \beta \) 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 \( \beta \) 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 \( \beta \) coefficients in Fig. 16. From the figure, we observe that 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, CC, gender, and average friends retention.

We also consider the prediction task as a classification problem. Our goal is to build a classification model that can determine whether a given user is highly retained or not.
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.

Unfortunately, our dataset does not have class labels (e.g., retained or not retained) for the users. The dataset only has points for the users that can be used to infer to what extent a user is retained. As such, we divide the users into two classes: top retained and regular. Similar to [48], we label the top 10% of users based on their points as top retained, and regular otherwise. We formulate the prediction task as a binary classification problem with two classes.

However, generation of class labels leads to an unbalanced dataset, where the positive class (e.g., regular) outnumbers the negative (top retained) class. Various approaches have been proposed in the machine learning literature to fix the unbalanced dataset. We use ROSE [49] algorithm to create a balanced dataset from the unbalanced one for training the classifiers. ROSE creates balanced samples by random over-sampling minority examples, under-sampling majority examples, or by combining over- and under-sampling. We use 70%–30% training–testing split. For the testing dataset, we have drawn randomly 30% of the users from the original dataset keeping the same top retained versus regular ratio of the Blogster dataset. Finally, using ROSE, we over-sample the top retained class and under-sample the regular class to draw 10 000 samples for training from the training dataset (70% of the remaining users that are not included in the testing dataset).

We use various classification algorithms, including support vector machines with radial basis function kernel (SVM-RBF), naive bayes, K-nearest neighbors (KNN), boosted logistic regression and SGBT and find that the SGBT shows the best performance. SGBT offers a prediction model in the form of an ensemble of weak prediction models [50]. Table IX shows a summary of our testbed and experimental setup. For evaluation, we used widely used metrics in classification problems: Accuracy, Precision, Recall, F1-score, and area under the receiver–operator characteristic curve (AUC). We use all the features used in the regression analysis as features for the classification problem.

Tables X shows the performance comparison of various classifiers when all features are used. All the classifiers perform very well, but SGBT performs the best in terms of all performance measures. The SGBT classifier is able to classify 93.62% of the instances correctly with lower false-positive and false-negative rates (see the confusion matrix in Table XI).
TABLE X
PERFORMANCE OF THE VARIOUS CLASSIFIERS

<table>
<thead>
<tr>
<th>Classifier name</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive bayes</td>
<td>89.13</td>
<td>98.08</td>
<td>89.65</td>
<td>93.67</td>
<td>73.55</td>
</tr>
<tr>
<td>Boosted logistic regression</td>
<td>90.90</td>
<td>99.14</td>
<td>90.66</td>
<td>94.71</td>
<td>76.60</td>
</tr>
<tr>
<td>KNN</td>
<td>92.67</td>
<td>98.88</td>
<td>92.90</td>
<td>95.79</td>
<td>79.40</td>
</tr>
<tr>
<td>SVM-RBF</td>
<td>92.67</td>
<td>98.88</td>
<td>92.90</td>
<td>95.79</td>
<td>79.40</td>
</tr>
<tr>
<td>SGBT</td>
<td>93.62</td>
<td>99.53</td>
<td>93.34</td>
<td>96.33</td>
<td>81.32</td>
</tr>
</tbody>
</table>

TABLE XI
CONFUSION MATRIX

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Regular</th>
<th>Top retained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>93.33%</td>
<td>3.87%</td>
</tr>
<tr>
<td></td>
<td>6.66%</td>
<td>96.12%</td>
</tr>
</tbody>
</table>

Fig. 17. Relative importance of top 10 features in classification.

Fig. 17 shows the most important features in classification. We can observe that each of the features has contribution in classification. However, blog traffic and other users’ comments appear to be the best predictors of top retained users.

VII. RELATED WORK

As the popularity of community blogs has increased, a wide body of research has been done on the community blogs. Researchers attempted to classify blogs using tags [51], identify spam in blogs [52], [53], identify influential bloggers [54] [55], and analyze blog sentiment [56]. Also, research has been done to understand why bloggers write blogs or upload contents and why they continue to publish content in the long run. Nardi et al. [13] conducted interviews with 23 bloggers in and around Stanford University and discovered five major motivations for blogging: documenting one’s life; providing commentary and opinions; expressing deeply felt emotions; articulating ideas through writing; and forming and maintaining community forums. Java et al. [12] explored these interview findings on a Twitter (a microblogging platform) dataset and found that people use small blog posts (e.g., Twitter updates) to talk about their daily activities and to seek or share information. Also, analyzing the user intentions associated at a community level, they showed that users with similar intentions connect with each other.

Researchers also explored the factors that contribute to increased continued participation or higher retention of users. Joyce and Kraut [9] 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. [57] found that new users contribute more content 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. However, outside of the blog, literature suggests that social relationships are important for continued participation. Pre-existing social relationships have been linked with recruitment on political and social movements [58], [59]. Analysis of online special interest groups (e.g., newsgroup) shows that group members with a strong sense of attachment to a group are more likely to participate [60] and closely connected groups are more supportive of members [61]. Stiggelbout et al. [62] investigated factors to continued exercise participation among older adults. They found female sex, younger age, being married, being a nonsmoker, 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 is a small and implicit network that the authors have built from users’ activity traces (e.g., comments, who invites whom to join). So, the network is highly sparse (3119 nodes, 4323 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 these data were 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.

VIII. CONCLUSION

In this paper, we asked what factors are related to continued user participation in the Blogster community blogging platform. 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 relate 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 their higher explicit and implicit interactions with other community members. Using the variables that relate to bloggers retention, we built prediction models to predict retention and top retained bloggers. We examined a number of widely accepted machine learning algorithms and found that our models can accurately predict retention (adjusted $R^2 = 0.84$) and top retained bloggers (accuracy = 93.62%). 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. We acknowledge that our study is observational; hence, we can only associate some of the factors that relate to user retention. In the absence of controlled experimental ground-truth data, we cannot draw causal conclusions regarding whether these factors are necessary and sufficient for user retention.

REFERENCES

Imrul Kayes received the B.Sc. (Engg.) degree in computer science and engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2009, the M.Sc. degree in computer science from the University of South Florida, Tampa, FL, USA, in 2014, and he is currently pursuing the Ph.D. degree in computer science and engineering at the University of South Florida, Tampa, FL, USA.

Before joining the Ph.D. program, he worked more than three years in software development industry. He has worked as a Software Engineer for SoftwarePeople, Delta life Insurance, and Binary Solutions. He spent the fall of 2013 as an intern at Yahoo Research Labs, Barcelona, Spain. His research interests include privacy, abuse, user influence, and retention in large, complex networks such as online social, community Q/A, and blogging networks.

Jacob Chakareski (SM’14) completed the M.Sc. and Ph.D. degrees in electrical and computer engineering at the Worcester Polytechnic Institute, Worcester, MA, USA, Rice University, Houston, TX, USA, and Stanford University, Stanford, CA, USA. He was a Senior Scientist at École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. He has held research positions with Microsoft, Hewlett-Packard, and Vidyo, a leading provider of Internet telepresence solutions. He is currently an Assistant Professor of Electrical and Computer Engineering at the University of Alabama. He has authored one monograph, three book chapters, and over 100 international publications, and holds five U.S. patents. He was a recipient of the Technical University Munich Mobility Fellowship, the University of Edinburgh Chancellor’s Fellowship, and fellowships from the Soros Foundation and the Macedonian Ministry of Science. He was the recipient of the Texas Instruments Graduate Research Fellowship at Rice University, the Swiss NSF Ambizione Career Development Award, the Best Student Paper Award at the SPIE VCIP 2004 Conference, and the Best Paper Award of the Stanford Electrical Engineering and Computer Science Research Journal for 2004.

Dr. Chakareski regularly participates on the technical program and organizing committees of multiple IEEE conferences. He is an Advisory Board Member of Mainframe2, an innovative visual cloud computing start-up with a bright future. For more information, please visit http://www.jakov.org.


