Measurement and Gender-Specific Analysis of User Publishing Characteristics on MySpace

William Gauvin*, Bruno Ribeiro†, Benyuan Liu*, Don Towsley†, Jie Wang*

* Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854, USA
{wgauvin, bliu, wang}@cs.uml.edu

† Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01002, USA
{ribeiro, towsley}@cs.umass.edu

Abstract—Online social networks have become popular platforms for people to make connections, share information, and interact with each other. In an online social network, user publishing activities such as sending messages and posting photos represent online interactions between friends that involve the use of network and system resources. As more and more businesses use online social networks as a means to promote their “brand names”, a good understanding of user publishing characteristics is important not only for the capacity planning (server scaling estimation, network bandwidth provisioning and performance tuning), but also for marketing analysis and security measures of online social networks. Recently there have been many efforts to measure and analyze various characteristics of online social networks. Most of these studies have focused on the network graph properties such as node degree distribution, clustering coefficient, connected components, etc. In this work, we measure and analyze the gender-specific user publishing characteristics on MySpace. Our results show that there are recognizable patterns with respect to profile attributes, user interactions, temporal usages, and blog content for users of different genders. In particular, gender neutral profiles (music bands, TV shows, and other commercial sites) differ significantly from normal male and female profiles in several publishing patterns.

I. INTRODUCTION

Online social networking (OSN) websites such as Facebook, MySpace, and LinkedIn, have become popular platforms for people to connect to and interact with each other. The OSN sites provide functionalities for users to build online profiles, establish friendships, send messages, post photos, advertise events, comment on friends profiles, create and join online communities of common interests. With the fast growing number of users and the deployment of new services and applications, online social networking has become a vital part of the Internet ecosystem. Businesses have been quick in identifying OSN sites as an effective means to promote their “brand names”. Many of these sites offer additional plug-ins and the success of a site is based on the number of “endorsements” that its friends publish on the site.

In an online social network, user publishing activities such as sending messages and posting photos represent online interactions between friends that involve the use of network and system resources such as network bandwidth and storage space. The characteristics of these interactions have profound implications on the design and performance of online social network systems and applications. The ability to identify commercial sites used both for well established businesses and “Profile Spam”, and validate the attributes used to represent profile owners is highly desirable. A good understanding of the publishing patterns is important, not only for the capacity planning (server scaling estimation, network bandwidth provisioning and performance tuning), but also for marketing analysis and security measures of online social networks.

Recently there have been a flurry of efforts to measure and analyze various characteristics of online social networks (e.g., [1], [2], [3], [4]). However, most of these studies have focused on the properties of the friendship graph including node degree distribution, clustering coefficient, assortativity, among others. There has been limited coverage on the use publishing behaviors on OSNs. In this work we measure and analyze the gender-specific user publishing characteristics of MySpace. Specifically, we will address the following questions: who publishes on a user's wall? when do they publish? what do they publish? and how do the publishing characteristics depend on the gender of the users?

MySpace is a major online social networking site. It was launched in 2004 and has since attracted a large
number of users including especially teenagers, music bands, artists, and other commercial sites. Currently it has more than 100 million users around the world and is particularly popular among young users and music communities. Each MySpace user owns a profile containing basic personal information such as name, age, gender, location, and so on. We define profiles whose gender is neither female nor male as gender “neutral” profiles. In our traces a neutral profile occurs when we cannot ascertain the gender or the gender is not specified. Many of the gender neutral profiles are commercial sites such as music bands and TV shows. Despite the commercial importance of these profiles, to the best of our knowledge there is no work in the literature that looks specifically at gender neutral profiles. One of our findings is that the friend and many publishing characteristics of gender neutral profiles differ significantly from the profiles that declare their gender as male or female.

In this work we provide measurement and analysis that allows us to look at the following publishing characteristics. We also examine if users of different genders exhibit different publishing patterns, and provide possible explanations.

- **Member Profile Pattern:** What is the distribution between private and public profiles for users of different genders? What is the correlation between the number of friends and corresponding number of publishers of a user profile?
- **Temporal Pattern:** What are the temporal publishing patterns of different profile types? How do they differ in terms of time of the day, day of the week, and months of the year?
- **Content Pattern:** What types of content do people tend to publish on MySpace? What are the relative occurrence frequencies between different account types?

A tool was developed to sample MySpace, uniformly at random, to collect user profile (name, gender, age, location, friends, etc) and their corresponding blog information. We randomly scanned 6 million possible profiles in the time period from September 2008 to May 2009. When blogs for these random profiles were found, each blog was scanned to obtain the publisher profile, time stamp, and content information. MySpace provides a standard template to display profile information. A profile may be public (default) or private depending on its privacy setting. Private profiles only share the owner’s name, gender, age, and location. No blog or friend information can be retrieved. As a result, most of our analysis is limited to the blogs posted on public profiles. Although we cannot examine the blogs of a private profile, we can extract many other publishing characteristics of private profile members by observing their interactions on public sites.

Our major findings are listed as follows:

- The majority of profile owners and publishers are within the age range of 15 to 30. Beyond this range the density falls off rapidly, only to increase around 69 and 100, likely due to a common practice among under age users to elude the age restriction by reporting an exaggerated age. Female users are more privacy aware than males. Nearly all of the neutral profiles are public, as these are often music band or other commercial sites trying to attract people to visit their contents. Users tend to interact with friends of similar ages, regardless of the account type being private or public. A large fraction (75%) of the publishing activities occurs between people within a five-year age difference.
- The distribution of the number of publishers for male and female profiles nearly coincide with each other except for the tail part, while the distribution for gender neutral profiles lies far above. This indicates that gender neutral profiles have more friends and publishers than normal male and female users. This is consistent with our observation that neutral profiles generally correspond to band or other commercial sites with a large number of friends who tend to publish once to comment on the music, movie, or product being promoted.
- Friendship does not necessarily reflect the real online interactions between users. Not all friends publish on a user’s blog wall. For many users on MySpace the number of distinct publishers constitutes only a small fraction of the friends, consistent with the observations made in [3] for Facebook. The difference is even more pronounced for gender neutral profiles. A possible explanation is that many people join the friend lists of commercial sites just to view the information.
- Temporal patterns closely resemble the Instant Message usage [5] for both private and public profiles, indicating a level of multitasking during business hours and times of leisure. There exist significant disparities in publishing activities between different hours of the day, days of a week, and months of a year. There is a sharp increase in publishing activities from morning to afternoon with a significant drop around midnight. High publishing
activities are observed for calendar events such as Valentine’s Day, Mother’s Day, and Christmas. Lower traffic patterns are observed during weekends than weekdays.

- There are four main content types in the blogs: words, hyper-link references (HREFs), images, and objects such as ColdFusion scripts or other animated objects. Users tend to post short messages. Only 10% of the blogs have more than 50 words; 4% of the blogs have 3 or more images. In a blog, word count dominates those of the other types, with object count being the smallest, reference and image counts similar to each other and lying in-between. There are a number of distinct patterns with respect to the count of these content types for different genders. The word count distribution of gender neutral profiles drops off drastically around 410, which may be caused by the need for commercial sites to have very specific and focused contents (e.g., advertisements). Another recognizable pattern is that females tend to publish more HREFs and images than male and gender neutral profiles.

II. RELATED WORK

Recently there have been a flurry of studies on the measurement and analysis various characteristics of online social networks. In [1], the authors present a large-scale measurement study and analysis of the graph structure of four online social networks, namely, Flickr, YouTube, LiveJournal, and Orkut. In Chun et al. [2] the authors study Cyworld’s (a Korean online social networking site) wall posts, directed activity graph, communication motifs, and the evolution of these metrics over time. In [4] the authors have looked at various statistics of male and female profiles in MySpace, including the friend degree distribution, user demographics, language and privacy preference.

Wilson et al. [3] study the relationship between friends and publishers and whether social links (social network friendships) are valid indicators of real user interaction. Their study builds a graph based on users interactions (messages exchanged between two users). The authors observe that the interaction graph has much lower average node degrees as compared to the Facebook friends graph. In this work we make a similar observation that only a small fraction of the friends actually write to a user’s blog wall.

Caverlee and Webb [4] analyzes word frequencies on MySpace wall posts, grouped by gender and age. Our work goes a step further and analyzes access patterns pertaining to posted content such as hyper-links, images, and objects. Temporal patterns of posts have also been studied in (pure) blog websites [8]. In [9] the authors break blog word usages into age groups, and gender. Several models have been proposed to describe the stochastic process of blog posts, please refer to [10] and the references therein. In Lento et al. [11] the connection between blogs, user activity, and user retention is studied in the Wallop on-line social network.

III. DATA ANALYSIS

We randomly sampled 6 million possible user profiles in the possible profile ID range from 1 to 1,500,000,000 between September 2008 and May 2009. The results of this process produced 546,829 public and 68,141 custom profiles of which 92,653 sets of blogs where obtained. There were a total number of 1,867,299 blogs collected. The number of blogs contained within a randomly scanned profile ranged from 0 to 85,643. To determine the publisher characteristics, 1,768,884 profiles were scanned.

![Fig. 1. MySpace Profile Distribution](image-url)

1) Gender and Privacy Distributions:

A. Member Profile Patterns

The distributions of the valid profiles between different genders are shown in Figure 1. Female, male, and neutral profiles each constitute 52.8%, 37.7%, and 9.5% of the total valid IDs, respectively. A larger percent of female users (57.1%) are more privacy aware than males (39.2%). Almost all (99.999%) of the gender neutral profiles are public as these are usually commercial sites trying to attract people to visit their pages.

Figure 2 shows the age distribution of private account profiles. The majority of profile owners are within the range of 15 to 30. Beyond this range the distribution falls off rapidly, only to increase around age 69 and 100. This is likely due to a common practice among under
age users to elude the age restriction by reporting an exaggerated age.

Two age ranges are important to describe: the first is the absence of ages from 1 to 12. By policy, MySpace protects users within this age range and does not allow profiles within this range to be created. Another policy that affects the results for age analysis is the MySpace policy for ages under 16 to be set to private. This restricts the information which can be obtained for these sites. It can be noted that there is a large number of zero-age profiles due to the under age accounts which are kept private.

![MySpace Private Account and Age Distribution](image)

**Fig. 2.** MySpace Private Account and Age Distribution

1) **Publishers versus Friends:** In an online social network, friend relationship does not necessarily reflect the real online interactions between users. It is interesting to see among all the friends of a user, how many of them actually post blogs on his/her wall.

![Distribution of Number of Friends and Publishers](image)

**Fig. 3.** Distribution of Number of Friends and Publishers

Figure 3 plots two sets of complementary cumulative distribution functions (CCDF), one set corresponding to the number of friends and the other set corresponding to the number of distinct publishers among all friends. In each set, a CCDF is obtained for each profile gender: female, male, and neutral.

It can be observed that the distributions for male and female profiles exhibit similar behaviors except for the tail part where the female distribution extends farther and the male distribution falls off sharply after a certain point. The distribution for gender neutral profiles lies above the male and female distributions. This indicates that neutral profiles tend to have more friends and publishers than normal male and female profiles. This trend is consistent with our observation that the neutral profiles are generally music band or other commercial sites with a large number of friends that tend to publish once to comment on the music, movie, or product being promoted.

![Distribution of Friends Verse Publishers](image)

**Fig. 4.** Distribution of Friends Verse Publishers

Figure 4 plots the number of distinct publishers versus the number of friends of MySpace users. It can be observed that for most users, as the number of friends increases, only a small fraction of them actually interact with them through blogs. This is consistent with the observations made in [3] for Facebook.

Interestingly, it can be observed that the envelopes of the publisher-versus-friend plot for male and female profiles nearly coincide with each other, while the envelope for the gender neutral profiles falls far below. This indicates that although gender neutral profiles tend to have more friends than normal male and female profiles, a smaller fraction of these friends actually write on the their blog walls. A possible explanation for this difference is that many people join the friend lists of commercial sites just to browse or receive information from the sites. The relationship between friends and gender neutral commercial sites are less "personal" than that between normal users.
It is worth noting that some profiles contain more distinct publishers than friends. This is a result of profile owners removing friends from their profiles, and occurs rather rarely.

2) Profile Owner and Publisher Age Difference: Figure 5 shows a density graph of the ages of publishers versus those of the profile owners. It can be observed that the density is highest along the diagonal line (zero age difference between publishers and profile owners) and drops quickly as the age difference increases. This indicates that users tend to interact with friends of similar ages. A large fraction (75%) of the publishing activities occurs between people within a five-year age difference. The age difference distribution can be well modeled by an exponential function in the form of \( f(x) = a \exp(b|x|) \) where \( a = 0.198 \), \( b = -0.4017 \). Note that the other small islands in the figure are artifacts caused by the following two groups of users, under-age users below 14 whose age information is not available and set to 0, and users with exaggerated ages around 69 and 100.

Fig. 5. Publisher Age Verse Owners

B. Temporal Usage Patterns

Three specific elements of time are analyzed. They are Time Of Day, Day Of Week, and Monthly Usage Patterns, as shown in Figure 6.

1) Hourly Publishing Pattern: All blog posting times are normalized to the local wall-clock time of the individual profiles. It can be observed from Figure 6(a) that there is a significant increase in publishing activity from morning to afternoon, with the peak usage time at 10pm and a significant drop at midnight. The low activity between midnight and 12pm may reflect the times of rest and busiest time for labor in the day as far as school and work are concerned.

2) Daily Publishing Pattern:

3) Monthly Publishing Pattern:

4) February Month Pattern:

Fig. 6. Temporal Publishing Patterns
2) Diurnal Publishing Pattern: Diurnal behaviors of posting habits exhibit some unexpected results, with respect to [6]. Weekdays characterized in Figure 6(b) have a higher degree of activity than weekend days. Prior to analysis, it was expected that weekend days would dominate, due to the available more free time to post new entries. As can be seen, this is not the case. Gender analysis for diurnal publishing patterns does not indicate any significant pattern difference between genders.

3) Monthly Publishing Pattern: The number of blogs posted per month in the period from January 2005 to December 2008 was collected, as shown in Figure 6(c). An interesting observation was that some months, especially December was not a high publishing month albeit its inclusion of two very popular holidays, Christmas and New Year's Eve. Further analysis for each specific month is required to determine the reason for the inconsistencies between human logical assumptions and substantiated patterns that are derived from the data set. The total number of blogs and content per day for each month was also collected. The reason for gathering this information was to ascertain whether trends in publishing patterns can be used to predict the load of a social network server and determine peak usage times. This information may be used to determine false positives during SPAM detection and other security filtering policies. Many months displayed evenly distributed activities on each day where the daily blog counts did not deviate by more than 200 blogs for the collection set analyzed. There were some months (February, November, December), however, that stood out with respect to predicting publishing patterns, mainly due to the specific calendar events within the months. An example is given in Figure 6(d) for February, which is a low publishing month in general, but there is a sharp contrast between the publishing activities on the 14th and those of the other days.

Similar high publishing activities are observed for other major calendar events including New Year's Day, Mother's Day, 4th of July, Halloween, Thanksgiving, and Christmas. Not answered in this paper is whether June/October weddings or April/May proms result in significant load increases. Observation of the April/May blog usage pattern indicates increased activities from previous months, but the exact reason was not a focus of this analysis.

C. Blog Contents Patterns

In a blog entry, there are four main content types, namely, words, hyper references (HREFs), images, and objects. In most cases the objects are ColdFusion scripts or other animated objects. Each blog is a composition of these four content types. The inclusion and makeup of these different content types is a reflection of each individual's style of writing. The goal of our analysis is to determine if specific content patterns can be identified.

Figure 7 depicts the complementary cumulative distribution functions (CCDF) of the counts of words, HREFs, images, and objects in a blog entry. It can be observed that the counts of these content types decay roughly according to power laws in certain regimes. In a blog entry, word count dominates the counts of other types. Object count is the smallest among all types. HREF and image counts are similar to each other and lie somewhere in-between. Of course, the counts of the different content types do not reflect their actual sizes (bytes). The size of an image is typically significantly larger than that of a word. In general, users tend to post short messages. For example, only 10% of the blog entries have more than 50 words; 4% of the blog entries have 3 or more images. The use of objects is even less frequent.

There are a number of recognizable patterns with respect to the count of these content types across different genders. The word count distribution of gender neutral profiles drops off drastically around word count 410. This may be caused by the need of the commercial sites to have very specific and focused contents (e.g., advertisements) that fit in people’s attention span. Another recognizable pattern is that females tend to publish more HREFs (hyperlink references) and images than male and neutral profiles.
IV. SUMMARY

In this article we present the measurement and analysis of gender-specific user publishing characteristics on MySpace. User publishing activities such as sending messages and posting photos represent online interactions between friends that involve the use of network and system resources. A good understanding of user publishing characteristics is important not only for the capacity planning, but also for marketing analysis and security measures of online social networks. Our results show that there are recognizable patterns with respect to profile attributes, user interactions, temporal usages, and blog contents for users of different genders. In particular, gender neutral profiles differ significantly from normal male and female profiles with respect to a number of publishing patterns. This is consistent with the observation that gender neutral profiles generally correspond to music band or commercial sites with a large number of followers who tend to publish once to comment on the music, movie, or product being promoted.

REFERENCES


