A Large-Scale Study of MySpace: Observations and Implications for Online Social Networks

James Caverlee
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
caverlee@cs.tamu.edu

Steve Webb
College of Computing
Georgia Tech
Atlanta, GA 30332-0280
webb@cc.gatech.edu

Abstract

We study the characteristics of large online social networks through an extensive analysis of over 1.9 million MySpace profiles in an effort to understand who is using these networks and how they are being used. We study MySpace through a comparative study over three different, but related, facets: (i) the sociability of users in MySpace based on relationship, messaging, and group participation; (ii) the demographic characteristics of MySpace users in terms of age, gender, and location, and a study of how these factors correlate with their privacy preferences; and (iii) the text artifacts of MySpace users, which can be used to construct language models that distinguish MySpace users not just by who they say they are but also by the language model they employ. We find a number of surprising results and conjecture several potential research directions based on our observations.

Introduction

Online communities are the fastest growing phenomenon on the Web, enabling millions of users to discover and explore community-based knowledge spaces and engage in new modes of social interaction. Sites like Bebo, Facebook, MySpace, Orkut, and LinkedIn have grown tremendously in the past few years, garnering increased media and popular awareness.

As online social networks continue to grow, evolve and develop, an important challenge we face is how to maintain the incredible success of Web 2.0 going forward. There is a growing demand for understanding this new social phenomenon, understanding the process by which communities come together, attract new members and develop over time, and understanding what it takes to empower the online communities with the ability to attract and retain a core of members who participate actively (Backstrom & others 2006; Coleman 1990).

As a step toward these goals, we present in this paper the results of a large-scale study over MySpace, the largest and most active online social network. By studying the characteristics of MySpace, we hope to provide insight into the types of users using these online social networks, how the network is organized, and the important text artifacts that distinguish these users. In particular, we study over 1.9 million real social network profiles, with an emphasis on:

- The sociability of users in MySpace based on relationship, messaging, and group participation.
- The demographic characteristics of MySpace users in terms of age, gender, and location, and a study of how these factors correlate with their privacy preferences.
- The text artifacts of MySpace users, which can be used to construct emergent language models that distinguish MySpace users not just by who they say they are but also by the language model they employ.

By studying how MySpace users participate in the social network (*sociability*), how they describe themselves (*demographics*), and how they communicate their personal interests and feelings (*language model*), we hope to encourage the development of new models, algorithms, and approaches for the further enhancement and continued success of online social networks. The core findings of our study are:

- Nearly half of the profiles on MySpace have been abandoned, meaning that the overall growth and explosive rate of user interest in social networks may need to be tempered; but we also identify a large core of active users within MySpace who account for the vast majority of friends, comments, and group activity.
- While young users (in their teens and 20s) are most prevalent on MySpace, women who are most prevalent at the youngest ages (14 to 20), whereas men are most prevalent for all other ages (21 and up).
- There are clear patterns of language use for users based on their age, location, and gender, which is useful both for text mining and characterization applications. We identify class-specific distinguishing terms and language model clusters that could be used to identify deceptive users who misrepresent their demographics.
- Overall, the fraction of private profiles is increasing with time, indicating that new adopters of social networks may be more attuned to the inherent privacy risks of adopting a public Web presence. We find that women favor private profiles 2-to-1 over men, and that (perhaps, counterintuitively) younger users are more likely to adopt a private

profile than older users. We also find that the more connected a user is in the social network, the more likely she is to adopt a private profile.

Related Work

The study of social networks has a rich history (Milgram 1967), and the recent rise of online social networks has seen renewed interest in this area. For example, a number of previous studies have examined the nature and structure of online social networks, including social networks derived from blogspaces (Backstrom & others 2006; Liben-Nowell *et al.* 2005), email networks (Adamic & Adar 2005), online forums (Zhang, Ackerman, & Adamic 2007), photo sharing sites (Kumar, Novak, & Tomkins 2006), among many others.

With respect to online social networks like MySpace and Facebook, there has been some research interest, but most studies have been on a smaller scale. In one study, researchers analyzed the relationship between a user's profile and friendships over 31,000 Facebook profiles (Lampe, Ellison, & Steinfeld 2007). Social capital has been studied over several hundred Facebook users in (Ellison, Steinfield, & Lampe 2006), and the privacy attitudes of 7,000 Facebook users was studied in (Acquisti & Gross 2006). (Dwyer, Hiltz, & Passerini 2007) surveyed a number of trust-related issues of over 100 MySpace and Facebook users. Personal information revelation among 10,000 young people on MySpace was studied in (Hinduja & Patchin 2008). One study considered membership formation for 200,000 Orkut members (Spertus, Sahami, & Buyukkokten 2005) and another looked exclusively at the messaging characteristic of 4 million Facebook users (Golder, Wilkinson, & Huberman 2007).

In comparison with previous work, we provide the first large-scale demographic study over millions of real social network profiles with respect to age, gender, and location, and we study how these factors correlate with their privacy preferences. We compare two sampling approaches for extracting social network data, and we provide a unique analysis of text artifacts to distinguish users.

Data and Setup

To study the characteristics of large online social networks, we selected MySpace as our target social network. MySpace is the largest social networking site, the 6th most visited Web destination according to Compete.com, and one that has received a tremendous amount of media coverage. In addition to these appealing characteristics, MySpace is one of the few online social networks that provides open access to user profiles. Many other sites require a user account and, even then, access to the entire social network can be restricted.

On MySpace, as on most online social networks, the most basic element is a *profile*. A profile is a user-controlled Web page that includes some descriptive information about the person it represents. Profiles connect to other profiles through explicitly declared *friend relationships* and numerous messaging mechanisms. MySpace allows users to choose between making their profiles publicly viewable (the default option) or private. If a user's profile is designated as private, only the user's friends are allowed to view the profile's detailed personal information (e.g., the user's interests, blog entries). However, a private profile still reveals the user's name, picture, headline, gender, age, location, and last login date.[1]

Since extracting and analyzing all 250 million MySpace profiles would place resource and network burdens on both MySpace and our local infrastructure, we adopted a sampling-based approach to extract representative samples from MySpace for further study. We consider two approaches – random-sampling and relationship-based (or snowball) sampling:

- *Random Sampling:* MySpace profiles are sequentially numbered and made publicly Web accessible by constructing a URL containing the profile's unique profile ID. Hence, we can randomly sample from the space of all MySpace profiles by randomly generating profile IDs. By construction, we expect a random sample of MySpace profiles to provide perspective on the global characteristics of the entire MySpace social network.
- *Relationship-Based Sampling:* Unlike random sampling, the second approach leverages the relationship structure of the social network to select profiles from the social network. We begin by generating a set of randomly selected seed profiles. We extract the IDs of their friends, add these friend IDs to the queue of profiles to sample, and continue in a breadth-first traversal of the social network. When the queue is empty, we generate a new random profile ID and continue the process. In contrast to the random sampling approach, we expect the profiles extracted through this sampling approach to provide a more focused perspective on the active portion of the social network.

In practice, we collected two representative datasets from MySpace: the **Random Dataset** using Random Sampling and the **Connected Dataset** using the Relationship-Based Sampling. We wrote two MySpace-specific crawlers (based on Perl's LWP::UserAgent and HTML::Parser modules). Both crawlers disregarded invalid profile IDs (i.e., profiles that were deleted or undergoing maintenance at the time of the crawl) and entertainment profile IDs (i.e., profiles that were associated with bands, comedians, etc.), to focus our collections on active profiles that belong to regular individuals. In June 2006, we deployed ten instances of the Random Sampling crawler in parallel across ten different servers, collecting profiles for about a week. We repeated this setup in September 2006 with the Relationship-Based Sampling crawler. Summary statistics for each dataset are listed in Table 1.

We wrote a custom MySpace parser to extract the name, age, and other pertinent information from each collected profile. Some of these features are self-described by the owner of the profile – like age and gender – and so these features may or may not be truthful. Other features – like number of friends – are maintained by MySpace and are

[1] MySpace also provides a few finer-grained privacy mechanisms for limiting IMs, comments, and for blocking specific users.

	Public Profiles	Private Profiles	Total Profiles	Size
Random	859,347	101,158	960,505	52 GB
Connected	717,337	173,830	891,167	98 GB

Table 1: Summary Statistics for the Two MySpace Datasets

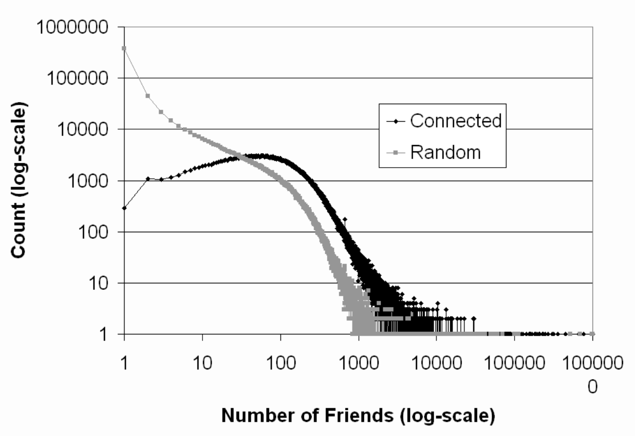


Figure 1: Distribution of Friends: The x-axis is the number of friends a user may have; the y-axis is a count of the number of users with a particular number of friends.

expected to be correct. Note that MySpace has a limited validation process; thus, we have no assurances that a self-described 20-year old male from Texas is really who he says he is. Having said that, we do believe there is significant value in studying demographics in the aggregate, and as we will see in the following section, certain text artifacts specific to certain groups on the social network could be used to mitigate deceptive profiles.

Results and Observations

In this section we present the main findings of our study through a series of characterizations: sociability, demographics, language models, and privacy preferences.

Sociability Characterization

We begin our characterization of MySpace by examining the social aspects of users in the network. Since online social networks derive their value from users actively participating in relationships with others users, we are interested to observe to what degree users actually take advantage of these social aspects. To examine this sociability over both datasets, we measure the number of friends, the number of comments, and the number of groups a user participates in. Note that these values are only available for public profiles.

In Figure 1, we present the distribution of the number of friends for both datasets on a log-log scale. For the Random dataset, we see a heavy-tailed distribution – that is, most users have very few friends, but a few users have many friends. Such a heavy-tailed distribution has been observed in a number of related domains, and observing it here is no surprise. What is surprising is the number of users with zero or only one friend: 426,926 or 50% of the public profiles in the Random dataset. Since MySpace provides each new user with a single default friend, we surmise that more than half of MySpace users created an account and subsequently abandoned it.

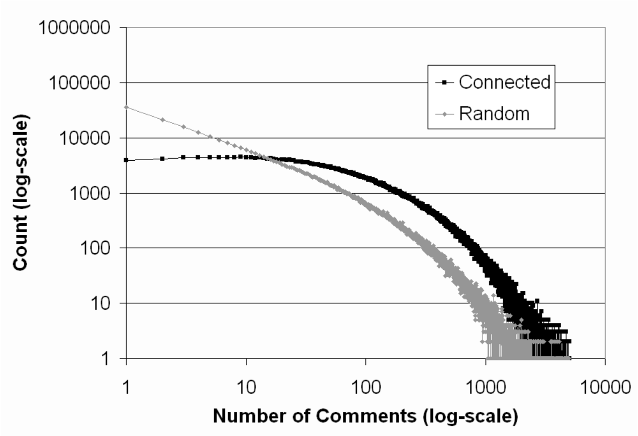


Figure 2: Distribution of Comments: The x-axis is the number of comments a user may have posted on her profile; the y-axis is a count of the number of users with a particular number of comments posted on her profile.

In contrast, we see that for the Connected dataset, most users have many friends and are actively participating in the social network. Only 2.5% of the public profiles in the Connected Dataset have zero or one friend. By construction, the Connected dataset favors users with many friends.

To further validate the sociability divide, we show in Figure 2 the distribution of the number of comments posted to a user's profile for both datasets. The commenting feature of MySpace is one of several avenues for users to communicate with other users; comments written to a particular user are posted on that user's profile, so we would anticipate that users with many comments are well-known and active in the social network. Again, we see the heavy-tailed distribution for the Random dataset, whereas the Connected dataset shows more skew, since it is by construction more connected.

Group participation is another metric of the sociability of a social network. While over 80% of the users in the Random dataset (and hence, we can extrapolate for MySpace as a whole) participate in no groups, we find that slightly less than half of the users in the Connected dataset belong to at least one group and that nearly 20% of users in the Connected dataset belong to at least 8 groups. This evidence further confirms what we observed with the friend and comment measures of sociability: most MySpace users have effectively abandoned their online profiles, but there is a large core of active users within MySpace who account for the vast majority of friends, comments, and group activity.

But who are these active users? In an effort to understand if some users are more likely to be active than others, we considered a number of features, including the age, location, gender, and length of time a profile had existed on My-

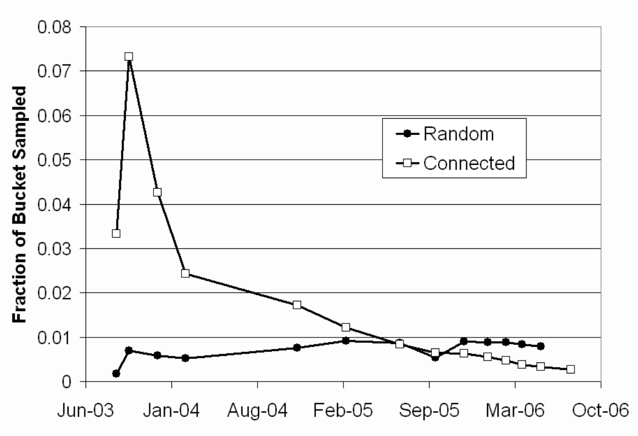


Figure 3: Sampling By Date: The x-axis shows buckets of profiles organized by the date of their creation; the y-axis shows the fraction of all profiles created within a bucket's range that were sampled.

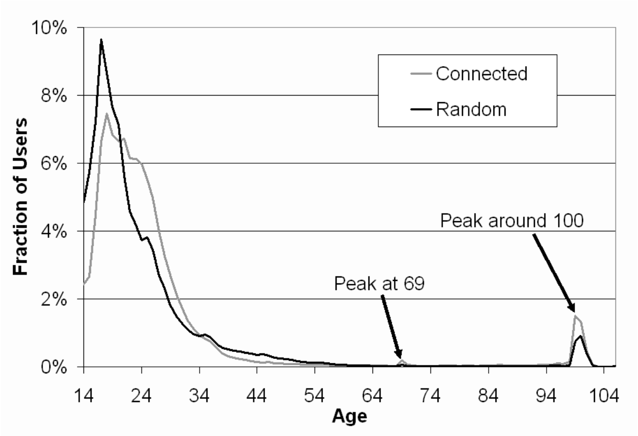


Figure 4: Distribution of Ages: The x-axis is the self-reported age on a profile; the y-axis is the fraction of all profiles declaring a particular age.

Space. We find that California and other western U.S. states dominate the total number of profiles on MySpace, but that these differences are minor across the Connected and Random datasets, which means location is not a strong indicator of sociability. Likewise, we find little evidence that a profile's self-declared age or gender impacts its relative sociability. In contrast, we find that the length of time a user has participated is a strong indicator of sociability. To measure participation length on MySpace, we augment our original sampling process. The profile creation date is listed for each profile on a separate blog page linked off the public profile Web page. This requires accessing one additional page per profile sampled. In an effort to avoid burdening MySpace with a doubling of page requests, we sampled a handful of profiles (e.g., profile 10,000, profile 100,000) and their creation date to create a time series. Since MySpace profile IDs are assigned sequentially, we can interpolate the date of creation for each profile sampled.

Hence, in Figure 3 each point represents a bucket of all profiles created before that date back to the previous point. The y-axis measures the rate of sampling for each bucket. As expected, the random sampling approach nearly uniformly samples from each bucket (One caveat: we see a hiccup at the beginning since the bucket is so small, and at the end because the sampling periods are slightly different). In contrast, the relationship-based sampling used to create the Connected dataset identifies users who joined overwhelmingly at an earlier date. These long-lived users are presumably more plugged-in and active participants in the social network.

Demographic Characterization

In the previous section, we studied the sociability of MySpace users – how active are they and to what degree are they connecting to other users? In this section, we expand our analysis of MySpace to consider the demography of participants. How old are they? Are they predominately male or female? Where are they located? The answers to the questions can provide us with added insight into how a social network grows, what features are attractive to certain participants, and other interesting avenues.

Recall that both public and private profiles on MySpace list basic demographic information. We find that nearly all MySpace users ($> 99.9\%$) provide some age, gender, or location information. Only 1,311 profiles in the Random Dataset declare no age, gender, or location; in the Connected Dataset, only 1,203 profiles declare nothing.

Figure 4 shows the distribution of ages in both datasets. As expected, MySpace is dominated by the young, with a peak at 17 years of age for the Random Dataset. Nearly 85% of the users on MySpace are 30 or younger. Interestingly, we observe that the Random dataset skews slightly younger than the Connected dataset, indicating that the most active users on MySpace may in fact be users in their 20s. We also observe a peak at the age of 69 – presumably either a joke age or an age intentionally selected by users interested in sex to find one another through the age-based search facility available on MySpace (Scalet 2007). We also observe a peak around 100, but we can presume that most of these self-reported ages are false.

In Table 2 we show the gender breakdown for each dataset: the split between male and female is nearly even: 52% male and 48% female in the Random dataset versus 50% male and 50% female in the Connected dataset. The "Other" gender is a placeholder for profiles that list either no gender information or non-standard gender information. In Figure 5 we consider the gender distribution across both datasets. The results are intriguing: women are more prevalent at the youngest ages, whereas men are more prevalent for all other ages (barring a few hiccups at the older end where the data is sparser).

Why are women more active participants at younger ages? Perhaps women intentionally self-report a younger age, or men intentionally self-report an older age. Perhaps there are clear gender differences in how users participate in a social

	Random	Connected
Male	505,357	440,330
Female	452,240	448,920
Other	2,908	1,917

Table 2: Gender breakdown for each dataset.

network, so that younger women are more attracted to certain social aspects than their male counterparts? These are interesting and open questions that deserve further exploration.

Finally, we studied the self-reported location information for each profile. MySpace users hail from all fifty U.S. states, and a significant fraction come from other countries. Not all profiles list an intelligible location (e.g. "Somewhere Over the Rainbow"), and some list multiple locations (e.g., "Honolulu and Metro DC"), so we built a best-effort parser. Based on our initial analysis, we find that the U.S. is by far the most prevalent location, followed by the United Kingdom and Canada; thus, we shall focus solely on U.S. states for the rest of this study. For the Random Dataset, we find that 77% list a U.S. state in the location, and for the Conected Dataset, we find that 87% list a U.S. state.

In Table 3, we report the top-5 states that are over-represented on MySpace relative to their actual population as well as the top-5 most under-represented states. We measure the relative presence of a state i on MySpace versus its relative share of the U.S. population as:

$$rel_i = 1 - \frac{pop_{i,MySpace}}{\sum_j pop_{j,MySpace}} / \frac{pop_{i,US}}{\sum_j pop_{j,US}}$$

where $pop_{i,US}$ is the population of state i based on the latest U.S. Census data and $pop_{i,MySpace}$ is the number of profiles in our dataset that declare state i as their location. For the Random dataset, we see in Table 3 that California and other western U.S. states are the most over-represented on MySpace relative to their actual population. Southern and mid-west states tend to lag relative to their actual population.

Most Over-represented	Most Under-represented
Hawaii [+115%]	Mississippi [-58%]
California [+61%]	West Virginia [-53%]
Washington [+41%]	Arkansas [-52%]
Alaska [+40%]	Missouri [-49%]
Nevada [+39%]	South Dakota [-48%]

Table 3: The states that are most over-represented and most under-represented on MySpace relative to their actual U.S. Census population. [Random Dataset]

We attribute much of this geographic discrepancy to MySpace's initial launch by a California-based company and success with Los Angeles area bands (Rosenbush 2005). Although California accounts for only 12% of the U.S. population, users from California dominate the early adopters of MySpace.

Characterizing Language Models

In our study so far, we have characterized how users participate in the social network (e.g., friendships, comments) and how users describe themselves (e.g., male, 24, from California). In this section, we examine what users are saying on their profiles through an analysis of the "language models" of social network users. Our goal is to understand how language use varies by class. For example, do women express themselves differently from men? Do older MySpace users describe themselves differently from younger MySpace users?

We begin with some definitions. We treat each profile as a sequence of terms drawn from a vocabulary set $V = \{t_1, t_{,2}, ..., t_{|V|}\}$. We consider all terms on a profile that are generated by the user (e.g., "About Me", "Interests"), and we exclude all terms most likely generated by other users (e.g., terms in comments). Following the standard information retrieval approach, we can describe the language model of all profiles as a probability distribution over the terms in the profiles according to a unigram language model:

$$\{p(t)\}_{t \in V} \text{ s.t. } \sum_{t \in V} p(t) = 1$$

Terms with high probability are more likely to be observed on a profile than low probability terms. We can compute $p(t)$ as a function of the count $count(t)$ of profiles containing term t relative to the total number of profiles n: $p(t) = count(t)/n$. For example, the top-5 most probable terms in the Connected Dataset are: *the, and, straight, friends, with*. These common terms provide little insight, and hence, we augment the basic language model by identifying *class-specific distinguishing terms* for classes based on age, gender, and location. Our goal is to identify terms that are more likely to be generated by a certain class of users: (e.g., by women).

To identify *class-specific distinguishing terms*, we rely on an information theoretic measure – Mutual Information – for assessing the importance of a term to a particular class.[2] Mutual Information between a term and a class is defined as:

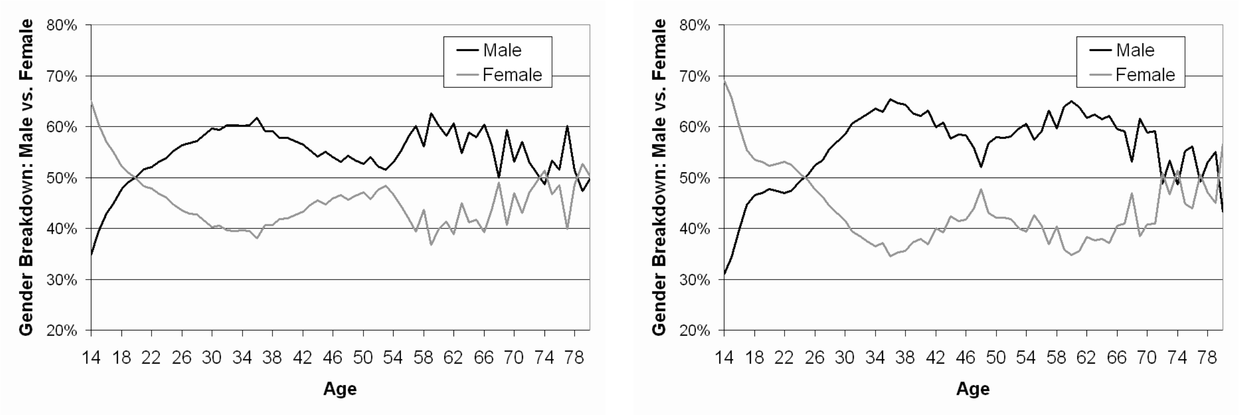
$$MI(t, c) = p(t|c)p(c) \log \frac{p(t|c)}{p(t)}$$

where $p(t|c)$ is the probability that a profile contains term t given that it belongs to class c, $p(c)$ is the probability that a profile belongs to class c, and $p(t)$ is the unigram language model described above for the probability of term t across all profiles. Letting $count(c)$ denote the count of profiles belonging to class c and letting $count(c, t)$ denote the count of profiles containing term t that belong to class c, we have:

$$p(t|c) = \frac{count(c, t)}{count(c)} \text{ and } p(c) = \frac{count(c)}{n}$$

Mutual Information measures how much information a particular term t tells us about class c. Higher MI values

[2] Since this work is exploratory in nature, we choose to keep all common words (like "the" and "and") that are often eliminated for text mining. In the same spirit, we perform no stemming.



(a) Random Dataset

(b) Connected Dataset

Figure 5: Gender Breakdown by Age: The x-axis is the self-reported gender on a profile; the y-axis is the fraction of all profiles of a particular age declaring a particular gender.

indicate stronger association. In this raw form, however, rare terms that by chance happen to occur only in profiles belonging to a particular class will score highly by Mutual Information. Hence, a natural correction is to replace $p(t|c)$ with a "smoothed" version that gives every term a non-zero probability of occurrence across all classes:

$$p^*(t|c) = \alpha p(t|c) + (1 - \alpha)p(t)$$

where $0 \leq \alpha \leq 1$. In practice we select a smoothing factor of $\alpha = 0.9$. We can interpret $\{p^*(t|c)\}_{t \in V}$ s.t. $\sum_{t \in V} p^*(t|c) = 1$ as a class-specific language model.

Class-Specific Distinguishing Terms Given the Mutual Information measure for identifying distinguishing terms, we next explore the language models of MySpace users according to three characteristics: gender, location, and age. Since we are primarily interested in users who are actively using the social network, we report results from the Connected Dataset. Superficially, we see many similarities with the Random Dataset in the presence of distinguishing terms. Note that only public profiles are included in this analysis since the contents of private profiles are hidden.

First, we consider the class distinction by gender – male and female. In Table 4, we report the top-16 class-specific distinguishing terms for profiles declared to be male and for profiles declared to be female. The differences are stark.

Male		Female	
dating	sport	love	people
networking	metal	dancing	life
serious	football	shopping	can
relationships	s***	girl	family
single	wars	hearts	being
straight	band	have	notebook
video	f***	are	dance
guitar	gay	favorite	things

Table 4: Distinguishing Terms by Gender (Ranked by MI)

Second, we consider class distinction by location for all fifty U.S. states. In Table 5 we report representative results from three states representing distant geographic regions of the U.S.: the south, pacific northwest, and northeast. We see an interesting mix of geography-specific identifiers (e.g., *protestant* in Alabama versus *catholic* in Connecticut), interests (e.g., *football* in Alabama versus *camping* in Oregon), and word constructions (e.g., *yall* versus *rad* versus *sneakers*).

Alabama	Oregon	Connecticut
christian	camping	catholic
african-descent	pdx	yankees
tide	hiking	nyc
jesus	northwest	uconn
football	pixies	hispanic
bama	snowboarding	bronx
church	coast	boston
christ	rafting	sox
protestant	floater	nas
gospel	rad	italian
yall	wine	goodfellas
nascar	vegan	sneakers

Table 5: Distinguishing Terms for Three Representative Locations (Ranked by MI): Popular location names (e.g., Birmingham, Portland) within each state are excluded.

Finally, we consider how the language model of MySpace users varies by age. In Table 6, we report the distinguishing terms for ages ranging from 16 to 100. We see how the language model shifts in focus with age based on education (e.g., from *high school* to *college* to *graduate* to *retired*). Also, older members use terms like *married*, *parent*, and *proud*, whereas younger members user terms like *single*, *friend*, and *love*.

We next consider a few notes about the older (and perhaps, less truthful ages). The 69-year olds have a clearly-expressed interest in sex. The odd language model of 80-

16	18	20	25	30	40	60	69	80	100
high	high	college	graduate	networking	parent	parent	networking	scudda	swinger
school	school	someday	college	graduate	proud	proud	swinger	mortenson	our
hearts	someday	student	networking	parent	married	president	sex	gable	kids
junior	love	love	grad	proud	networking	swinger	a**	jeane	capricorn
single	best	straight	professional	married	kids	his	f***	showgirl	networking
best	boy	caucasian	relationship	grad	great	married	rock	asphalt	virgo
hair	ever	white	traveling	professional	our	kids	islander	dimaggio	artists
friend	hair	like	some	art	divorced	united	real	dougherty	their
lol	lol	girl	reading	cure	daughter	began	our	harlow	please
play	single	know	working	travel	years	retired	night	actress	official

Table 6: Distinguishing Terms by Age (Ranked by MI)

year olds is skewed by the presence of many Marilyn Monroe tribute profiles (who would have been 80 at the time); all of the terms are relevant to her movie career and relationships. The 100-year olds display a less coherent language model, perhaps due to the diversity of users declaring such an age.

Identifying Language Model Clusters In the previous section, we saw how certain classes of MySpace users can be described by distinguishing terms that are relatively strong indicators of class membership. In this section, we continue this analysis by considering clusters of related classes. For example, given that most self-declared 100 year-old members of MySpace are not actually 100, what is their true age? MySpace has made some effort to remove self-declared older members (Scalet 2007) through manual inspection. Can the language models provide us with a scalable solution?

We begin with the class-specific language models of interest (e.g., by age: $\{p^*(t|c = 16)\}_{t \in V}$, $\{p^*(t|c = 17)\}_{t \in V}$, and so on). Are there clusters of language models by age or by location? In this initial study, we consider a similarity measure for determining the "closeness" of two language models based on the Kullbeck-Leibler divergence (or relative entropy). KL-divergence measures the difference between two probability distributions p and q over an event space X:

$$KL(p, q) = \sum_{x \in X} p(x) \cdot \log(p(x)/q(x))$$

Intuitively, the KL-divergence indicates the inefficiency (in terms of wasted bits) of using the q distribution to encode the p distribution. In this case, we can measure the divergence of two class-specific language models (i.e. $p = \{p^*(t|c = 16)\}_{t \in V}$ and $q = \{p^*(t|c = 17)\}_{t \in V}$). Note that KL-divergence is not symmetric so we will typically find $KL(p, q) \neq KL(q, p)$.

First, we report the KL-divergence in Figure 6 for 16-year olds versus other ages, for 20-year olds versus other ages, and for 30-year olds versus other ages. Since there are very few profiles listing an older age, we omit these from the graph.

Note that the KL-divergence of 16-year olds is lowest for profiles closest in age, which means the language model of a 16-year old is closest to a 17-year old, then an 18-year old, and so on. A similar pattern holds for the 20-year old language model and for the 30-year old language model. There are clear clusters based on age.

What do we observe when we consider profiles that are more likely to be deceptive about their true age? As an illustration, we show in Table 7 the closest language models for profiles listing an age of 69 and profiles with an age of 100.

Rank	Age 69	Age 100
1.	100 [0.017]	99 [0.047]
2.	99 [0.021]	101 [0.103]
3.	101 [0.047]	30 [0.105]
4.	33 [0.068]	31 [0.105]
5.	31 [0.072]	29 [0.106]

Table 7: Identifying Outliers: Which language model most closely matches the language model of the self-described 69-year olds? And of the 100-year olds? [KL-divergence]

For the 69-year olds, we see that the closest matches are other outlier ages – 100, 99, and 101. This gives us some evidence that the type of user who lies about his age is bound by some common language model cues. The next two closest matches are in their 30s. This is a bit surprising; we would have expected teenagers to be more likely to engage in such behavior. For the 100-year olds, we see a similar pattern: close matches with other outlier ages (99 and 101) and then close matches with younger profiles that are presumably more likely to declare true ages. We believe this line of inquiry could be extended along a number of fruitful directions.

Privacy Preferences

Finally, we turn our attention in this study to the important issue of privacy in social networks. A number of researchers have examined some of the aspects impacting privacy on social networks (e.g., (Barnes 2006; Boyd 2007; Nussbaum 2007)) in an effort to understand user's understanding of privacy and the limits of privacy controls, and so on. In this section, we examine the privacy choices of members of MySpace through the lens of our demographic study. Recall that MySpace users can elect to declare their profile as public or private. A private profile displays only limited

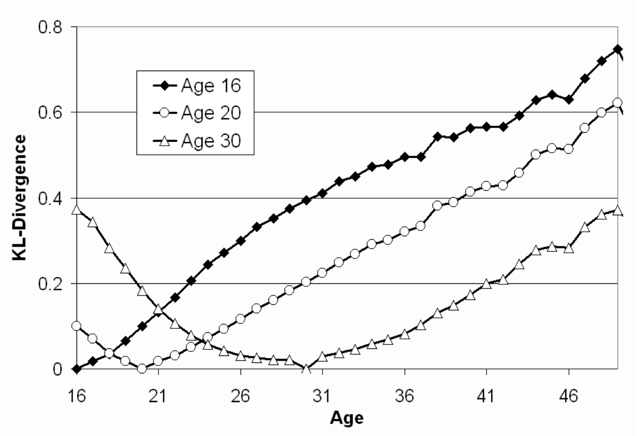


Figure 6: KL-Divergence by Age: We compare the class-specific language model using KL-divergence (lower is better).

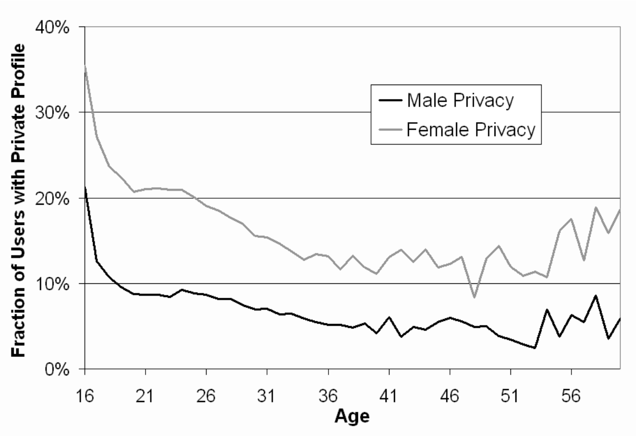


Figure 7: Privacy Breakdown by Age: The x-axis is the self-reported age on a profile; the y-axis is the fraction of all profiles declaring a particular age that are private. [Connected]

information like name, age, gender, and location. Especially young members of MySpace (14 and 15-year olds) are required to have a private profile.

First, we report in Table 8 the privacy preferences of the randomly selected MySpace users of the Random Dataset (which is intended to reflect the overall MySpace population) versus the privacy preferences of the more sociable members of the Connected Dataset. Members of the Connected Dataset select private profiles by nearly 2-to-1 over the average MySpace user. These findings are especially surprising since the relationship-based sampling technique used to extract the Connected Dataset relies on the friendships declared on public profiles to identify profiles to sample; private profiles reveal no friendships and so the sampling terminates when it arrives at a private profile. We further examined the private profiles in each dataset and found that nearly all (99.9%) of the private profiles in the Random Dataset belong to 14 and 15-year olds (see Table 9). In contrast, we find that over 73.7% of the private profiles in the Connected Dataset are of the age 16 or higher.

	Random	Connected
Private	101,158 (10.5%)	173,830 (19.5%)
Public	859,357 (89.5%)	717,337 (80.5%)
Total	960,505	891,167

Table 8: Privacy preferences for each dataset.

	Random	Connected
14/15 Years Old	101,017 (99.9%)	45,633 (26.3%)
All Other Ages (16+)	141 (00.1%)	128,197 (73.7%)
Total	101,158	173,337

Table 9: Privacy preferences by age for each dataset.

Overall, very few users elect private profiles when given the opportunity (00.1%), but of users who actively use the social network, we see a much larger fraction. These results also lend credence to the hypothesis that more sociable members tend to be more likely to choose private profiles.

To further explore the impact of demographics, we present in Figure 7 the fraction of private profiles in the Connected Dataset by age and gender. We truncate the graph over the age of 60 since there are very few profiles at those ages and hence we see more noise. We find that women favor private profiles 2-to-1 over men and that (perhaps, counterintuitively) younger users are more likely to adopt a private profile than older users. Why is this? Perhaps older users are less technically savvy and have difficulty understanding how to set up the privacy setting; perhaps younger users are more attuned to the privacy and security concerns of social networks. We believe this is an area deserving more attention.

Finally, we consider how privacy preferences have changed over time. In Figure 8 each point represents a bucket of all profiles created before that date back to the previous point. The y-axis measures the fraction of profiles created within that bucket that are private (again, relying on MySpace's use of sequential IDs to interpolate profile creation dates).[3] After an initial drop in privacy rate, we see a fairly steady growth of privacy adoption for new members. Overall, the fraction of private profiles is increasing with time, indicating that new adopters of social networks tend to be more attuned to the inherent privacy risks of adopting a public Web presence. We also investigated privacy preferences by location, but find no dramatic swings from state-to-state.

Conclusions

In this paper, we have presented a large-scale study over MySpace in an effort to better understand this new social phenomenon. Our comparative study differs from previous

[3] We assume the choice of public/private is a one-time decision. In practice, users can modify their privacy settings at any time.

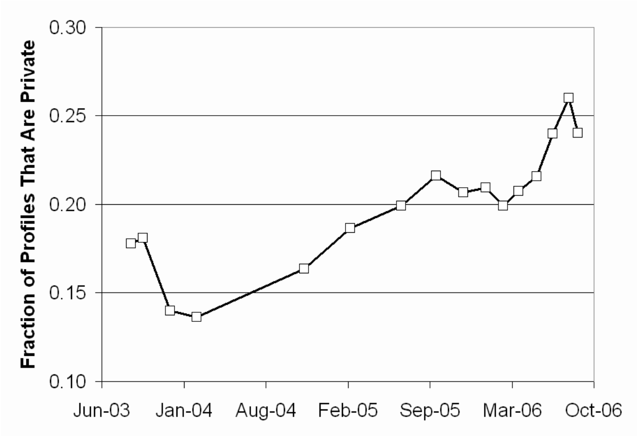


Figure 8: Privacy Over Time: The x-axis shows buckets of profiles organized by the date of their creation; the y-axis shows the fraction of all profiles created within a bucket's range that are private. [Connected Dataset]

work in its scale (over 1.9 million profiles) and in its breadth. In particular, we have examined how MySpace users participate in the social network (*sociability*), how they describe themselves (*demographics*), and how they communicate their personal interests and feelings (*language model*). We have identified a number of surprising and interesting features that motivate our continuing research. In particular, we are interested in augmenting and extending models of social network growth to incorporate the demographic variations we have observed. Along this line, we believe finer-grained language models that move beyond age, gender, and location to capture user interest and user expectations of the social network (e.g., for business-development networking, for making friends) could be beneficial.

References

Acquisti, A., and Gross, R. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *6th Workshop on Privacy Enhancing Technologies (PET)*.

Adamic, L. A., and Adar, E. 2005. How to search a social network. *Social Networks* 27(3):187–203.

Backstrom, L., et al. 2006. Group formation in large social networks. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Barnes, S. B. 2006. A privacy paradox: Social networking in the United States. *First Monday* 11(9).

Boyd, D. 2007. Social network sites: Public, private, or what? *The Knowledge Tree: An e-Journal of Learning Innovation*.

Coleman, J. 1990. *Foundations of Social Theory*. Harvard University Press.

Dwyer, C.; Hiltz, S. R.; and Passerini, K. 2007. Trust and privacy concern within social networking sites. In *Proceedings of the Thirteenth Americas Conference on Information Systems*.

Ellison, N.; Steinfield, C.; and Lampe, C. 2006. Spatially bounded online social networks and social capital. In *International Communication Association*.

Golder, S. A.; Wilkinson, D.; and Huberman, B. 2007. Rhythms of social interaction: messaging within a massive online network. In *Third International Conference on Communities and Technologies*.

Hinduja, S., and Patchin, J. W. 2008. Personal information of adolescents on the Internet: A quantitative content analysis of MySpace. *Journal of Adolescence*.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Lampe, C.; Ellison, N.; and Steinfeld, C. 2007. Profile elements as signals in an online social network. In *Conference on Human Factors in Computing Systems*.

Liben-Nowell, D.; Novak, J.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences* 102(33):11623–1162.

Milgram, S. 1967. The small-world problem. *Psychology Today* 60 – 67.

Nussbaum, E. 2007. Kids, the Internet, and the end of privacy. *New York Magazine*.

Rosenbush, S. 2005. News Corp.'s place in MySpace. *Business Week*.

Scalet, S. D. 2007. MySpace cracks down on 69-year-old members. *CSO Online*.

Spertus, E.; Sahami, M.; and Buyukkokten, O. 2005. Evaluating similarity measures: A large-scale study in the Orkut social network. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Zhang, J.; Ackerman, M.; and Adamic, L. 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the International World Wide Web Conference (WWW)*.