Sizing Up Online Social Networks

Reza Rejaie, Mojtaba Torkjazi, and Masoud Valafar, University of Oregon
Walter Willinger, AT&T Labs-Research

Abstract

While the size of popular online social networks such as MySpace and Twitter has been reported to be in the tens or hundreds of millions of users (and growing), little is known about the fraction of users who have either deleted or abandoned their accounts. Therefore, the growth of an OSN's overall user population and, more important, its population of active users cannot easily be determined. In this article we describe a measurement technique to infer the fine-grained growth in the total number of allocated accounts for a class of OSNs that include MySpace and Twitter and are characterized by two features. First, they assign numerical user IDs using a format and allocation strategy that can be determined. Second, a fraction of their users have abandoned these OSNs shortly after creating their accounts, and these short-lived users (called "tourists") are scattered across the ID space. By exploiting these two properties, we are able to determine the growth in total and valid user accounts for MySpace and Twitter since their inception. For valid user accounts, we also derive the fraction of active users in the system at the time of our experiment, where we define the activity of a user in terms of the recentness of her last visit to the OSN. In the case of MySpace and Twitter, our results show that the active population of these OSNs is typically an order of magnitude smaller than the reported (total) population.

The size of popular online social networks (OSNs) such as MySpace (www.myspace.com) and Twitter (www.twitter.com) has been reported to be tens or hundreds of millions of users, and growing. Their enormous size and astounding growth reflect their societal impact and can in turn affect their further growth. As a result, OSNs have attracted considerable attention from all areas of society, including researchers, advertisers, and politicians.

Most of the reported information about popular OSNs focuses on their large size and rapid growth for good reasons. First, to avoid negative publicity and possible ripple effects on other users, popular OSNs are generally silent about negative statistics related to their growth, or the fraction of users who have deleted or abandoned their accounts. In fact, they often tend to obscure any information that would enable third parties to derive such statistics. For example, Facebook (www.facebook.com) does not explicitly notify a user when her friends delete their accounts or remove their friendship links. Second, OSNs are typically studied and characterized when they are very popular, and the fraction of departing or inactive users is likely to be small. In short, it is difficult to obtain accurate information about the actual growth in user population or the number of active users in an OSN; yet it is such information that provides a meaningful way to assess an OSN's societal impact.

In this article, after a brief discussion of the challenges associated with estimating the population of an OSN, we describe a simple measurement technique for accurately inferring the fine-grained growth in user population of certain OSNs from their inception to the time of our measurement experiments in early 2010. The technique was originally described in [1] and applies to OSNs that have two particular features. First, their user ID allocation strategy can be determined and used to estimate the total number of allocated accounts each time a new user joins the system. Second, there is a fraction of short-lived users, called *tourists*, and they are scattered across the entire ID space. The key observation regarding these tourists is that their account creation time can be estimated by their last login time and can subsequently be used to estimate the account creation time of all users in the system. Using the proposed technique, we conduct an empirical study of two major OSNs that have these features, MySpace and Twitter, and examine the growth in their total user populations. We also distinguish between departed users (i.e., deleted accounts) and existing users (i.e., valid accounts), and further divide valid accounts into public and private groups based on their account settings as well as their levels of activity. Here we equate activity with active use of the OSN (e.g., logging in or sending a tweet) and measure it in terms of the recentness of a user's last visit to the system.

Our analysis shows that a non-negligible fraction of users have removed their accounts, and a significant fraction of existing users have either abandoned these OSNs shortly after creating their accounts (i.e., are tourists) or have not logged in for a long period of time (i.e., are not active). To size up an OSN's user population that "counts," we apply a definition of activity that targets very active users who have visited their account during the last 10 days. Our results reveal that *the active population of MySpace and Twitter is typically an order of magnitude smaller than the reported (total) population.* When we relax the definition and target more moderately active users who visited their account within the last 100 days, the fraction of active accounts roughly doubles in both OSNs.

This work is supported by the NSF award IIS-0917381.

0890-8044/10/$25.00 © 2010 IEEE

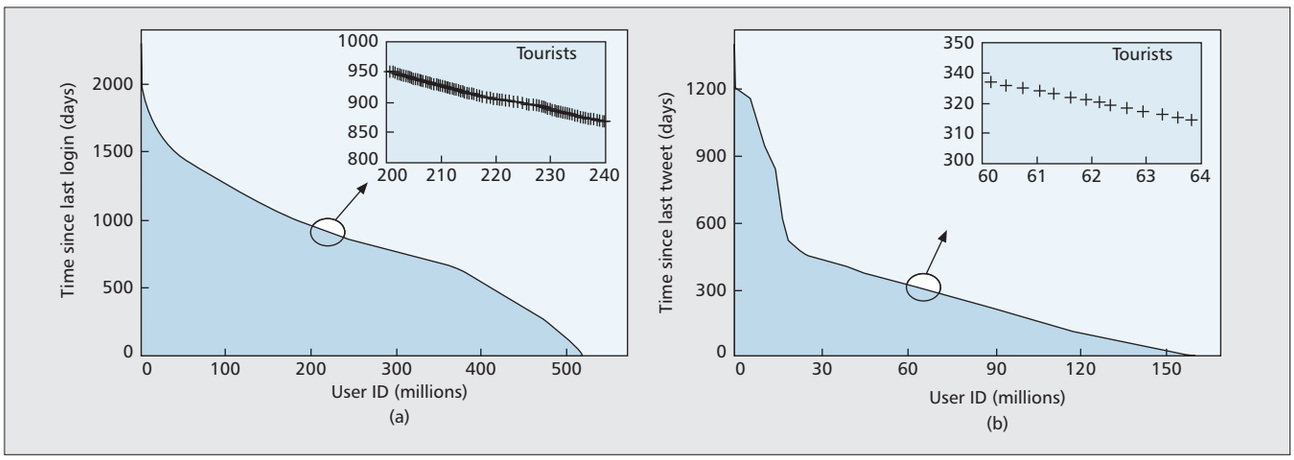

Figure 1. *Time since last login/tweet vs. user ID for MySpace and Twitter users. Zoom-in views showi the time since last login/tweet for tourists.*

Examination of the growth in user population of these OSNs over time indicates that their growth goes through different phases. This is indicative of the popularity of these OSNs within the larger OSN ecosystem. We discuss the implications of our findings for characterizing OSNs and conclude with a number of interesting open problems.

Challenges in Assessing Growth

Since popular OSNs typically do not report detailed information about their growth in user population, measurement is the most promising alternative for third parties to collect this information. The most obvious approach is to use an OSN-specific crawler to capture multiple snapshots of the OSN in question, estimate the population of users in each snapshot, and then determine the evolution of the OSN user population during the measurement period. However, this straightforward approach has major limitations. First, given the large size of and various crawling constraints imposed by popular OSNs, it is prohibitively expensive to capture complete and accurate snapshots of their user population. For example, Twitter allows each user to submit only 150 queries per hour to the server. Given the structure of its API, this constraint limits the rate of discovery to about 50 users per hour. Even with multiple concurrent crawlers, capturing a snapshot of Twitter's user population at such a low rate could take months. Second, this crawling-based approach enables us to capture the growth in user population only for the duration of the measurement experiment and only at a coarse granularity (i.e., time between consecutive snapshots).

Various companies, such as Alexa (www.alexa.com), monitor the traffic of OSN web sites, and some studies use this information to estimate the relative popularity of different OSNs over time [2, 3]. While such information is clearly worthwhile, it only represents an aggregate measure of an OSN's popularity and provides no insight into user-level characteristics. For example, the aggregate rate of accesses to MySpace's web site does not reveal the rate at which new users join or existing users leave MySpace. Given these limitations, we need to devise alternative techniques for estimating the growth in OSN user population.

Our Approach

To study the fine-grained (i.e., per user) evolution of an OSN user population since its inception, our approach is to focus on a random sample of users in the OSN, estimate their account creation times, infer the user population when each of the sampled users joins the system, and then relate user population to time to assess the population's growth.

The proposed approach is feasible for a class of OSNs that exhibit the following two properties:

- An ability to reverse-engineer the ID allocation strategy of the OSN
- The availability of account creation time for a subset of users who are scattered across the OSN's ID space

Knowing the ID allocation strategy enables us to relate the ID of each user to the total size of user population at the time that user joins the OSN. The presence of "short-lived" users coupled with the (direct or indirect) availability of last login information provides an opportunity to accurately estimate the account creation time for these short-lived users. If these short-lived users are scattered across the ID space, we can use their account creation times along with the knowledge of the ID allocation strategy to approximate the account creation time for other users. For example, consider an OSN with numerical user IDs and a sequential allocation strategy, where the account creation times for two users with IDs id_1 and id_2 are ct_1 and ct_2, respectively. The account creation time for id_x ($id_1 \le id_x \le id_2$) can be interpolated simply as

$$ct_x = ct_1 + \frac{(id_x - id_1)(ct_2 - ct_1)}{id_2 - id_1}.$$

To demonstrate our proposed methodology, we examine the relationship between user ID and the elapsed time since the user's last login for two major OSNs that have the required features, MySpace and Twitter. To this end, we first identify the format and valid range of user IDs for these OSNs by careful inspection of a large number of users across the ID space. Next we use this information to generate random user IDs and obtain a representative group of users. The time of last login for individual users is explicitly reported in the profile of MySpace users, but it must be implicitly inferred from the time of the last tweet for Twitter users. Table 1 summarizes the main characteristics of the 239,000 and 895,000 randomly selected MySpace and Twitter users, respectively.[1]

Figures 1a and 1b show scatter plots of the days since the last login of user with ID id_x (y-axis) vs. user ID id_x (x-axis) for the sets of randomly selected MySpace and Twitter users, respectively. The most striking feature of these figures is the appearance of a monotonically decreasing "clean edge." We

[1] *MySpace users were sampled over a 12-h period on 01/16/2010. Twitter users were sampled between 06/06/2010 and 06/17/2010.*

OSN	Total	Deleted	Public	Private
MySpace	239,000	42.3%	41.4%	16.3%
Twitter	895,000	23.3%	70.7%	6.0%

Table 1. *Statistics on sampled user accounts.*

claim that a plausible explanation for this feature is the presence of "short-lived" users, along with a monotonically increasing ID allocation strategy in an OSN. Since the last login of these short-lived users occurs shortly after they created their account, we refer to them as tourists. Importantly, their last login provides an accurate estimate for their account creation time. The formation of the monotonically decreasing clean edge implies that the time since the account creation time for tourist u_t must be larger than that of any user whose ID is larger than u_t's ID. Clearly, this is the case when the ID allocation strategy assigns monotonically increasing user IDs.

We repeatedly conducted several tests [1] to gain insight into the micro-level pattern of the ID allocation strategies used by MySpace and Twitter. For example, we identified the largest allocated ID and monitored the short-term pattern of allocation for larger IDs. We also examined the distribution of the gap between adjacent deleted IDs within a given block of allocated IDs. When combined with results from a previous Twitter study [4], these tests support the claim that *MySpace and Twitter assign numerical user IDs in a strictly sequential order and do not recycle IDs of deleted accounts.*[2]

We consider a user as a tourist if her last login occurs within a day from her account creation time. Using our randomly selected MySpace and Twitter users, we found that 20 percent of valid MySpace users and 18 percent of public Twitter users are indeed tourists. Moreover, Figs. 1a and 1b depict a zoom-in view of the placement of these tourists across an arbitrarily selected segment of the ID space and support the claim that these tourists are indeed scattered across the OSN's entire ID space. The relatively large fraction of tourists in MySpace and Twitter along with their dispersed locations across the ID space enable us to interpolate the account creation time of any user from its user ID based on the account creation times of adjacent tourists.

Presence of Tourists in OSNs

Given the importance of tourists for our technique, a key question is whether the presence of tourists is a common phenomenon in most OSNs or whether it is specific to MySpace and Twitter. While several prior studies have commented on the rate of departing users in different OSNs (e.g., [5, 6]), there are a number of possible reasons for tourist-like behavior among OSN users. First, the low (often zero) cost of joining an OSN may encourage users to join an OSN out of curiosity. For example, close to 90 percent of identified tourists in MySpace with private profiles are users who declared themselves to be less than 16 years old. This suggests that the phenomenon of tourists among users with private profiles may be partly due to young users who join a new OSN to check it out (and for privacy reasons are often required to establish a private account), but do not find their experience with the OSN all that exciting and thus abandon the OSN or join another system. Second, some users create (possibly many) account(s) to use the "credit" that various web sites offer to an OSN account owner or to use the account for sending a couple of messages. Finally, OSNs actively monitor user behaviors to quickly identify and remove misbehaving users who violate their terms of use. While these short-lived users could contribute to the tourist population, our technique cannot identify them as tourists because their accounts have been deleted.

It is also interesting to note that reasons for tourist-like behavior have been reported in the literature on non-virtual networks (e.g., perceived usefulness or ease of use for adopting or abandoning a technology in information technology systems [7]). However, whether or not tourists are a typical feature of today's OSNs remains an open question that deserves further study.

Growth of User Population

We use our technique to characterize the growth in user population of MySpace and Twitter. Our crawlers collect the following information associated with each sampled user:

- Whether a user account is deleted (via proper error message)[3]
- Whether the encountered user profile is private or public
- The time of the user's last login

The latter is of particular interest because it enables us to assess the level of user engagement with the OSN. Table 1 summarizes the pertinent information for our random sets of MySpace and Twitter users.

Using these datasets, we first identify the tourists for each OSN and exploit their presence to determine the account creation time for all non-tourist users as described earlier. Given the user ID and account creation time for all users, we finally determine the growth in both total user population over time and its breakdown into deleted, public, and private accounts. Since we cannot determine the deletion time of deleted accounts, our analysis is retrospectively based on the status of sampled accounts at the time of our measurements by associating each account with its creation time. Figures 2a and 2b depict the growth in the total number of created accounts in MySpace and Twitter from system inception until the time we performed our measurement experiments. The slope of the top line in these figures represents the rate of growth.

These figures illustrate two interesting properties of MySpace and Twitter. First, the growth in user population appears to exhibit distinct phases. For example, in the case of MySpace we observe three such phases:

- *Initial phase*, when the OSN is relatively unknown and thus the rate of growth in user population is low (e.g., from late 2003 to early 2005)
- *Expansion phase*, when there is significant hype and excitement about the OSN, which leads to a large and possibly increasing rate of growth (e.g., from early 2005 to early 2008)
- *Maturity phase*, when interest in the OSN is fading and growth in user population exhibits a moderate and possibly decreasing rate (e.g., from early 2008 to late 2009)

Second, MySpace and Twitter appear to be in different phases of their growth. For example, focusing on 2009, My-

[2] *Twitter initially used a "mod* n*" strategy (with different* n *values over various parts of ID space) for allocating IDs (i.e., assigning only monotonically increasing IDs whose mod* n *value is zero) [4]. However, that strategy was replaced by a sequential strategy for user IDs greater than 14,000,000 around February 2008. The lower density of points for small user IDs in Twitter (Fig. 1b) is due to the lower utilization of the ID space caused by this allocation strategy in Twitter.*

[3] *We note that an account may be removed either directly by the user or by the system administrator in case of a violation of the terms of use, among other things. We are unable to determine the actual cause that led to the termination of individual accounts or the actual time of removal.*

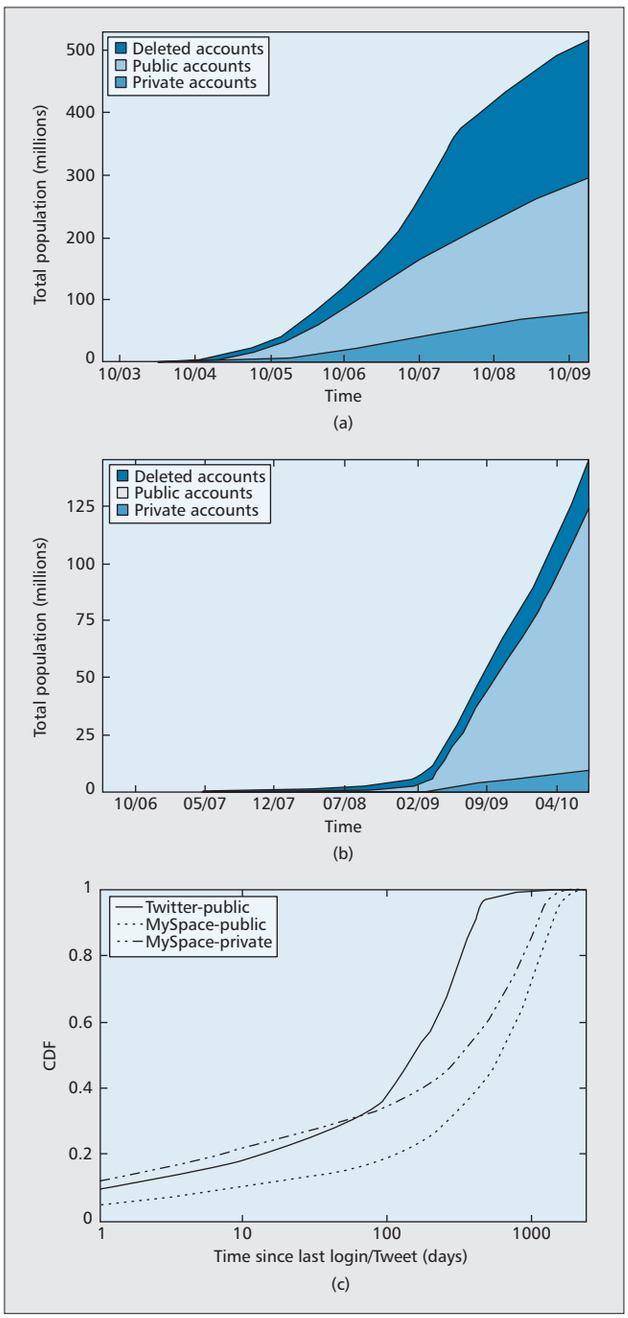

Figure 2. *a, b) Growth in user population of MySpace and Twitter; c) distribution of time since last login/tweet.*

Space has apparently reached its mature phase while Twitter is still in its expansion phase.

To examine category-specific growth of user population, Figs. 2a and 2b present a breakdown of the total number of accounts by deleted and valid accounts, and further divide the valid accounts into public and private profiles. In essence, the growth in the number of valid accounts provides a more realistic view of the user population of an OSN over time. While the relative growth in the population of different groups of users over time shows that the population of valid accounts (both public and private) in MySpace has increased at a rather constant rate, the growth rate for deleted accounts initially increased and then stabilized in early 2008. This implies that the increasing growth rate during the expansion phase can primarily be attributed to the deleted accounts. In contrast, the fraction of deleted accounts in Twitter is much smaller than in MySpace.[4] Finally, we observe that the fraction of deleted accounts in Twitter is much smaller than in MySpace, and a much larger fraction of valid accounts have public profiles. We hypothesize that the observed smaller fraction of users with private profiles in Twitter is because they have a more limited domain of information sharing than users with public profiles.

Factors that Affect Growth

Given the observed growth of the user population for MySpace and Twitter, it is natural to hypothesize what factors affect the overall or group-specific growth of an OSN's user population and cause OSNs to exhibit such distinctly different phases of growth. Intuitively, the ability of an OSN to keep its existing users interested and actively engaged while attracting new users depends on at least two sets of factors:

- *Internal factors* such as introducing exciting new features, limiting the volume of spam, or offering new security or privacy solutions
- *External factors*, such as increasing popularity and hype caused by another up-and-coming OSN or negative public opinion about an OSN

Not only is it generally difficult to reliably assess the impact of each of these factors on an OSN's user population, but they are likely to impact different OSNs in different ways. An easier task is to identify major factors that result in a sudden and significant change in the popularity of an OSN. For example, the pronounced decrease in the rate of growth in MySpace's user population around April 2008 in Fig. 2a is likely due to a significant drop in the arrival rate of new users. A reasonable hypothesis for this observed drop is that the emergence of Facebook had a significant impact on the arrival of new users to MySpace (and possibly other OSNs). To verify this hypothesis, at least qualitatively, we examined the number of daily accesses to the web servers for MySpace, Facebook, and Orkut (another popular but largely regional OSN) as reported by Alexa [1]. For the time period of interest, we observed that the increasing trend in the access rate of Facebook coincides with a decrease in popularity of both MySpace and Orkut (www.orkut.com). This shows that the aggregate popularity of different OSNs is strongly correlated, which lends some credibility to our hypothesis.

Users that "Count"

The number of valid accounts provides a more realistic metric for the user population of an OSN. However, users are likely to exhibit different levels of activity. Given that it is precisely those *active* users who determine the societal importance of an OSN, estimating the number of active users could arguably provide an even more meaningful measure of user population for an OSN. Clearly, to assess the level of activity or engagement of a user, one can consider different user properties (e.g., average rate of tweets or rate of posted content). In our analysis we focus on user participation (i.e., frequency of visit) in an OSN as a measure of a user's activity, mainly because it can be viewed as an indication of a user's interest in (or dependency on) the OSN. However, since we cannot obtain the information about a user's individual visits to an OSN, we use the time of last visit as an estimate for the level of activity. Thus, our intuition is that the longer a user does not visit an

[4] We observed that roughly 25 percent of the deleted accounts in Twitter are due to "suspicious behavior" by the corresponding user and have been removed by the OSN.

OSN	Total accounts	Valid	Moderately active	Very active
MySpace	518 million	57.7%	13.4%	7.9%
Twitter	144 million	76.7%	~37.9%	~18.5%

Table 2. *Total, valid, and active population of MySpace and Twitter.*

OSN, the less likely she will visit that OSN in the future.

For MySpace where the time of last login is explicitly available, Fig. 2c depicts the cumulative distributed function (CDF) of the duration of time (in days) between a user's last login and the time of our measurement experiments for our randomly selected set of users with valid accounts who have been in the system for more than 100 days. We further divide the valid MySpace accounts into public and private profiles. The figure reveals that only about 19 percent of the users with public profiles and about 34 percent of users with private profiles on MySpace have logged in within the last 100 days. We call these *moderately active users*. These numbers drop to about 10 and 22 percent for those users with public and private profiles who have logged in within the last 10 days. We call these *very active users*. The figure also indicates that users with private profiles are relatively more active than users with public profiles. For Twitter, we use the time of the last tweet of a user with a public profile as a surrogate measurement for the user's time of last login. If we simply assume that Twitter users with private profiles have a similar level of activity, Fig. 2c shows that a lower-bound for the fraction of moderately active and very active Twitter users is 37.9 and 18.5 percent, respectively.

Our above analysis shows that:

- A non-negligible fraction of accounts in MySpace and Twitter are invalid, most likely due to misbehavior or departure of the corresponding users.
- A large fraction of valid accounts in these OSNs have not been visited for a relatively long time, indicating moderate to low interest and thus moderate to low level of activity by the corresponding users.

To put this analysis into perspective, we summarize the main results in Table 2. Focusing on MySpace, we observe that the total number of assigned MySpace accounts at the time of our measurement was around 518 million. Based on our random samples, 42.3 percent (219.1 million) of all accounts have been deleted, and 57.7 percent (298.9 million) are valid.

The estimated fraction of moderately active MySpace users (both public and private) at the time of our experiment is 13.4 percent (69.4 million). This estimate drops to 7.9 percent (40.9 million) if we only count very active users. These observations suggest that a large fraction of users with a valid profile are not actively using MySpace and might very well have abandoned the system. For all practical purposes, they should be ignored when trying to size up MySpace. For Twitter, 45 percent of users have not sent any tweets. We cannot estimate the time of last login for these users and consider them nonactive. Of the 55 percent of users who have sent at least one tweet, 37.9 percent are moderately active and 18.5 percent are very active. Since we use the time of a user's last tweet to estimate the time of her last login, a fraction of users who tweet infrequently might still periodically log into their Twitter accounts to check the tweets they received. These users are not actively generating tweets or generate them only infrequently, but their use of Twitter is mainly for receiving or reading tweets. In short, the reported fraction of active users in Twitter directly depends on the notion of "activity" we use.

Implications and Open Problems

The observed growth in user population at possibly different rates over time coupled with the high percentage of tourists in MySpace and Twitter suggests that important characteristics of these systems (e.g., user connectivity) may vary significantly over time or evolve in substantially different ways during the different phases of these systems. Thus, rather than trying to study these OSNs as a whole at a given point in time, it may be more informative to perform empirical studies that focus on different groups of users and their behaviors during a specific life cycle of an OSN. On one hand, this suggestion adds significantly more burden on future empirical OSN studies because it requires more detailed measurements, more sophisticated analysis of the data, and a more intricate characterization effort compared to past system-wide investigations. At the same time, the observed skewed nature of activity among OSN users implies potentially significant opportunities for future empirical research on OSN characterization. More specifically, exploring the often much smaller portion of active OSN users (compared to all users) looms as a definite possibility, and could result in substantial efficiencies in future OSN measurement and analysis. This observation is based on the premise that the non-negligible fraction of users that do not actively participate in the system can be viewed as noise and ignored for all practical purposes.

Our findings also raise a number of new open problems that deserve further consideration. First, it is important to determine the underlying social, economical, or technological factors that affect the growth of an OSN's user population. This in turn enables us to meaningfully study the structure and evolution of these systems with the goal to develop valid models with high predictive power. Only an in-depth understanding of the root causes underlying the different aspects of the observed growth in OSN user populations will enable insightful what-if type studies that may, for example, provide explanations for the rise and fall of existing or future OSNs. Second, given the likely migration of users among OSNs over time as well as the concurrent participation of some users in multiple OSNs, it is important to capture and characterize OSNs in the bigger context of the Internet's OSN ecosystem. However, this is an extremely challenging task, mainly because of the existing difficulties in accurately tracking the identity of individual users across multiple OSNs.

References

[1] M. Torkjazi, R. Rejaie, and W. Willinger, "Hot Today, Gone Tomorrow: On the Migration of MySpace Users," *ACM SIGCOMM Wksp. Social Networks (WOSN)*, 2009.
[2] Knowledge@Wharton, "MySpace, Facebook and other Social Networking Sites: Hot Today, Gone Tomorrow?," *Wharton J.*, 2006.
[3] B. Tancer, "Facebook: More Popular than Porn," Time, 2007; http://www.time.com/time/business/article/0,8599,1678586,00.html[4] B. Krishnamurthy, P. Gill, M. Arlitt, "A Few Chirps about Twitter," *ACM SIGCOMM Wksp. Social Networks*, 2008.
[5] J. Golbeck, "The Dynamics of Web-Based Social Networks: Membership, Relationships, and Change," *Sunbelt XXVIII Int'l. Sunbelt Social Network Conf.*, 2008.
[6] R. Kumar, J. Novak, and A. Tomkins, "Structure and Evolution of Online Social Networks," *Int'l. Conf. Knowledge Discovery and Data Mining*, 2006.
[7] F. D. Davis, "Perceived USefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, 1989.

Biographies

REZA REJAIE [SM'06/ACM, SM'06] (reza@cs.uoregon.edu) received his M.S. and Ph.D. degrees from the University of Southern California in 1996 and 1999, and his B.S. degree from Sharif University of Technology in 1991. He is currently an associate professor at the University of Oregon. From 1999 to 2002 he was a senior technical staff member at AT&T Labs-Research, Menlo Park, California. His research interests include P2P networking, network measurement, and multimedia networking. He received an NSF CAREER Award for his work on P2P

streaming in 2005.

MOJTABA TORKJAZI (moji@cs.uoregon.edu) is currently a Ph.D. student in the Computer Science Department at the University of Oregon. He received his B.Sc. degree in computer science from Sharif University of Technology, Iran, in 2006. His research interests include online social network characterization and network measurement.

MASOUD VALAFAR (masoud@cs.uoregon.edu) is pursuing a Ph.D. degree in computer science at the University of Oregon. He received his B.Sc. degree in information technology from Sharif University of Technology, Iran, in 2007. His research interests include online social network characterization and information flow in online social networks.

WALTER WILLINGER [F'05](walter@research.att.com) received his Diplom (Dipl. Math.) from ETH Zurich, Switzerland, and M.S. and Ph.D. degrees from the School of ORIE, Cornell University. He is currently a member of the Information and Software Systems Research Center at AT&T Labs-Research, and before that he was a member of technical staff at Bellcore Applied Research (1986-1996). He is a Fellow of ACM (2005). For his work on the self-similar ("fractal") nature of Internet traffic, he received the 1996 IEEE W.R.G. Baker Prize Award, the 1994 W. R. Bennett Prize Paper Award, and the 2005 ACM/SIGCOMM "Test of Time" Paper Award. This work is supported by NSF award IIS-0917381.