Characterizing User Behavior in Online Social Networks

Fabrício Benevenuto[†] Tiago Rodrigues[†] Meeyoung Cha[*] Virgílio Almeida[†]

[†]Computer Science Department, Federal University of Minas Gerais, Brazil
[*]Max Plank Institute for Software Systems (MPI-SWS), Kaiserslautern/Saarbrücken, Germany

ABSTRACT

Understanding how users behave when they connect to social networking sites creates opportunities for better interface design, richer studies of social interactions, and improved design of content distribution systems. In this paper, we present a first of a kind analysis of user workloads in online social networks. Our study is based on detailed clickstream data, collected over a 12-day period, summarizing HTTP sessions of 37,024 users who accessed four popular social networks: Orkut, MySpace, Hi5, and LinkedIn. The data were collected from a social network aggregator website in Brazil, which enables users to connect to multiple social networks with a single authentication. Our analysis of the clickstream data reveals key features of the social network workloads, such as how frequently people connect to social networks and for how long, as well as the types and sequences of activities that users conduct on these sites. Additionally, we crawled the social network topology of Orkut, so that we could analyze user interaction data in light of the social graph. Our data analysis suggests insights into how users interact with friends in Orkut, such as how frequently users visit their friends' or non-immediate friends' pages. In summary, our analysis demonstrates the power of using clickstream data in identifying patterns in social network workloads and social interactions. Our analysis shows that browsing, which cannot be inferred from crawling publicly available data, accounts for 92% of all user activities. Consequently, compared to using only crawled data, considering silent interactions like browsing friends' pages increases the measured level of interaction among users.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

IMC'09, November 4–6, 2009, Chicago, Illinois, USA.

General Terms

Human Factors, Measurement

Keywords

Online social networks, user behavior, session, clickstream, social network aggregator, browsing, silent activity

1. INTRODUCTION

Online social networks (OSNs) have become extremely popular. According to Nielsen Online's latest research [23], social media have pulled ahead of email as the most popular online activity. More than two-thirds of the global online population visit and participate in social networks and blogs. In fact, social networking and blogging account for nearly 10% of all time spent on the Internet. These statistics suggest that OSNs have become a fundamental part of the global online experience.

Through OSNs, users connect with each other, share and find content, and disseminate information. Numerous sites provide social links, for example, networks of professionals and contacts (e.g., LinkedIn, Facebook, MySpace) and networks for sharing content (e.g., Flickr, YouTube).

Understanding how users behave when they connect to these sites is important for a number of reasons. First, studies of user behaviors allow the performance of existing systems to be evaluated and lead to better site design [3, 33] and advertisement placement policies [2]. Second, accurate models of user behavior in OSNs are crucial in social studies as well as in viral marketing. For instance, viral marketers might want to exploit models of user interaction to spread their content or promotions quickly and widely [18, 31]. Third, understanding how the workload of social networks is re-shaping the Internet traffic is valuable in designing the next-generation Internet infrastructure and content distribution systems [16, 26].

Despite the potential benefits, little is known about social network workloads. A few recent studies examined the patterns using data that can be gathered from OSN sites, for instance, writing messages to other users [8, 14, 30, 33] or accessing third party applications [10, 22]. As a result, these studies reconstruct user actions from "visible" artifacts like messages and comments. While these initial studies yield insights into social network workload, they do not provide a global picture of the range and frequency of activities that users conduct when they connect to these sites.

A complementary approach to study OSN workloads is to use traces such as clickstream data that capture *all* activities of users [7]. Since clickstream data include not only visible interactions, but also "silent" user actions like browsing a profile page or viewing a photo, they can provide a more accurate and comprehensive view of the OSN workload.

In this paper we present a first of a kind analysis of OSN workloads based on a clickstream dataset collected from a social network aggregator. Social network aggregators are one-stop shopping sites for OSNs and provide users with a common interface for accessing multiple social networks [25]. Because social network aggregators are an excellent measurement point for studying workloads across various OSNs, we collaborated with a popular social network aggregator in Brazil for this study. We obtained a clickstream dataset, which described session-level summaries of over 4 million HTTP requests during a 12-day period in 2009. The dataset included activity data for a total of 37,024 users who accessed various OSNs through the social network aggregator.

Using the clickstream data, we conducted three sets of analyses. First, we characterized the traffic and session patterns of OSN workloads (Section 3). We examined how frequently people connect to OSN sites and for how long. Based on the data, we provide best fit models of session inter-arrival times and session length distributions. Second, we developed a new analysis strategy, which we call the *clickstream model*, to characterize user activity in OSNs (Section 4). The clickstream model captures dominant user activities and the transition rates between activities. We profiled user activities for four OSN services: Orkut, MySpace, Hi5, and LinkedIn. Third, to gain insight into how users interact within a given social network, we additionally crawled the Orkut website and analyzed user activity along the social graph (Section 5). Our analysis reveals how often users visit other people's online profiles, photos, and videos.

Our study provides many interesting findings:

1) Session duration is heavy-tailed, indicating large variations in the OSN usage among users. We provide a best-fit distribution function for the Orkut sessions.

2) Using clickstream data, we present the frequencies and sequences of user activities in Orkut (Table 2 and Figure 6). We find that browsing, which cannot be inferred from publicly available data, is the most dominant behavior (92%).

3) When we consider silent interactions like browsing friends' pages, the number of friends a user interacts with increases by an order magnitude, compared to only considering visible interactions.

4) Analysis of user interaction along the social graph shows that Orkut users not only interact with 1-hop friends, but also have significant exposure to friends that are 2 or more hops away (22%).

In summary, our study provides a first look into the usage of OSN services from the viewpoint of a social network aggregator. The clickstream data analyzed in the paper provides an accurate view of how users behave when they connect to OSN sites. Furthermore, our data analysis suggests several interesting insights into how users interact with friends in Orkut. We believe that our findings have implications for efficient system design.

2. DATASET

We use two datasets in this paper. The first is a clickstream dataset that is collected and provided by a social network aggregator site. The second is the Orkut social network topology that we crawled. These two datasets provide complementary types of information that we correlate in Section 5. Below we describe both datasets and our methodology for crawling Orkut. We also discuss some limitations of these datasets.

2.1 Clickstream data

We describe how social network aggregators operate and introduce the clickstream dataset we obtained and analyzed.

2.1.1 *Social network aggregator*

Social network aggregators pull content from multiple social networking sites to a single location, thereby helping users who belong to multiple networks manage diverse profiles more easily [25, 27]. Upon logging into a social network aggregator, users can access their social network accounts through a common interface, without having to login to each OSN site separately. This is done by a two-level real-time HTTP connection: the first level is between a user and a social network aggregator site and the second is between the social network aggregator site and the OSN sites. Social network aggregators typically communicate with OSN sites using Open APIs that OSN sites provide [12]. All content from OSN sites are shown to users through a social network aggregator's interface. Figure 1 depicts the scheme interaction among users, a social network aggregator site, and OSN sites. Through the interface of the social network aggregator, a user can enjoy all features that are provided by OSN sites, for instance, checking updates from friends, sending messages, and sharing photos.

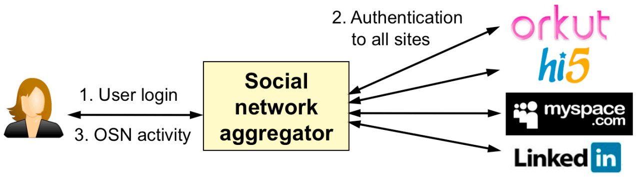


Figure 1: Illustration of a user connecting to multiple OSNs through the social network aggregator

2.1.2 *Data description*

The clickstream data that we analyzed were collected over a 12-day period (March 26 through April 6, 2009). The data consist of summaries of HTTP header information for traffic exchanged between the social network aggregator server and users. The dataset summarizes 4,894,924 HTTP requests, including information about time stamp, HTTP status, IP address of the user, login ID in the social network aggregator site, URL of the social network site, login ID within the social network site, session cookies, and the traffic bytes sent and received. After discarding events with missing fields or HTTP status associated with error codes (e.g., 301, 302), there were 4,649,595 valid HTTP requests. HTTP requests in the trace are grouped into sessions, where a session represents the sequence of a user's requests during a single visit to the social network aggregator. The trace included 77,407 sessions, covering 16,175 distinct user IP addresses and 37,137 distinct login IDs in the social network aggregator site.

Not all log entry in the trace were related to accessing OSNs. Some log entries reflect users accessing non-OSN features of the aggregator site, such as listening to an Internet radio or watching videos. Other log entries result from the automatic display of advertisements and the aggregator site's website logo. After discarding non-OSN related log entries, 802,574 or 17% of the HTTP requests were related to accessing the following four OSNs: Orkut, Hi5, MySpace, and LinkedIn. The remainder of this paper focuses on these HTTP requests related to accessing OSNs.

Table 1 displays the number of users, sessions, and HTTP requests for these OSNs. Among them, Orkut had the largest number of users and accounted for nearly 98% of all HTTP requests. Although the remaining OSN sites take up only 2% of the trace, the data contain sufficient number of users for each of these sites. Therefore, we can identify meaningful user behaviors for these OSNs.

OSNs	# users	# sessions	# requests
Orkut	36,309	57,927	787,276
Hi5	515	723	14,532
MySpace	115	119	542
LinkedIn	85	91	224
Total	37,024	58,860	802,574

Table 1: Summary of the clickstream data

2.1.3 Data anonymization

The social network aggregator anonymized any sensitive information that might reveal a user's identity prior to our analysis. There were three types of information that were anonymized. First is the user login IDs in the social network aggregator site. Second is the user IDs in the social network site. Third is the IDs of web content that users accessed. We could determine the content ID only if the content ID appeared in the URL of the fetched webpage. For example, when a user browses a particular photo, content information like the photo ID, the uploader ID, and the album ID appears on the URL of the fetched webpage, and was therefore logged and anonymized. On the other hand, when a user browses his or her own homepage and sees update feeds from friends, information about these web objects does not appear in the URL of the fetched webpage, and was therefore not logged.

2.2 Social network topology of Orkut

To gain insight into user behaviors over the social graph, we crawled the largest OSN site in the trace, Orkut. Because of the sheer size of the Orkut network, we decided to crawl friendship information for only those users that appear in the clickstream dataset. We used the Orkut user IDs that appear in the trace, prior to anonymization. We implemented a crawler which downloaded the profile page of each of these Orkut users. A profile page contained a variety of information about users. Certain profile information is made publicly available to all Orkut users, for instance, the list of friends, the list of community memberships, name, gender, and country. On the other hand, other information like email, phone number, and age is set private and is shown only to friends by default. When crawling Orkut, we stored all profile information that is made publicly available.

We gathered the profile information of the 36,309 Orkut users the week after the clickstream data were gathered, during April 10–17, 2009. The average number of friends was 211.4 and the median number of friends was 152. Some users had no listed friends at all, while the user with the highest number had 998 friends. Orkut allows a user to have at most 1,000 friends. Later we examine what fraction of friends a user visibly or silently interacts with. The IDs of users in the crawled social graph were anonymized in the same way as the clickstream data.

2.3 Data limitations

Although the clickstream data give us a unique opportunity to study user activities across multiple OSNs, the dataset has limitations.

First, the dataset is biased towards the set of users in the social network aggregator portal. One evident bias is the demographics of users in Orkut. To examine the geographical distribution of users, we used the GeoIP database [20] to identify the location of 16,175 IP addresses that appeared in the trace. These users were located across all continents in the world, spanning 90 countries. However, certain geographical locations contained more users than others. Brazil had the highest presence both based on the number of IP addresses (71%) and the number of the HTTP requests (70%). The second largest user base came from India and accounted for 12% of the IP addresses and 14% of the requests. The third most common location was the United States. The bias in user samples may raise a concern about how representative our results are for other social networks in the data, i.e., Hi5, MySpace, and LinkedIn.

Second, user behavior in a given social networking site is influenced by the specific mechanisms the site provides. Therefore, our findings about user activity may change as new features are added to social networking sites. To examine the set of user behaviors that are relatively oblivious to the specific design of websites, we studied user behaviors across *multiple* social networks and tried to look for patterns that remain consistent across multiple services.

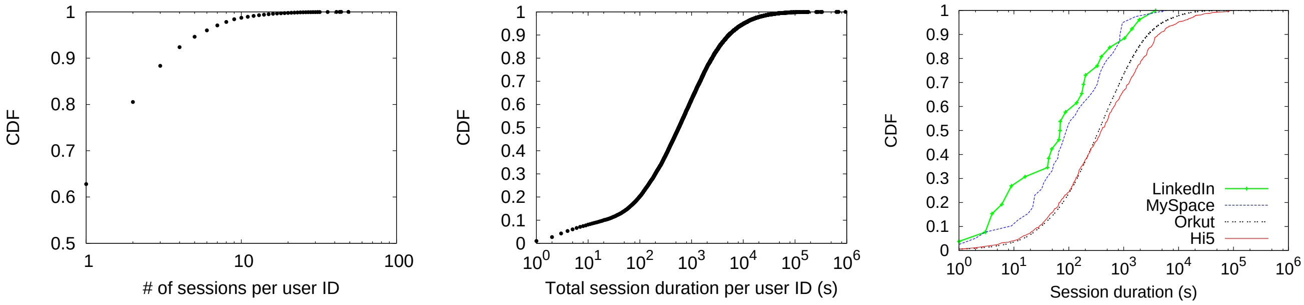
Third, we are not able to infer behaviors of users over a long term period (e.g., several months) since the data were collected only over a 12-day period.

3. CONNECTION PATTERN ANALYSIS

In this section, we characterize OSN workloads at the session level. We first describe how sessions are identified in the social network aggregator, then examine the duration and frequency of connections to OSN services. We also model two key session characteristics from a system's perspective: inter-arrival times and session length distribution.

3.1 Defining a session

The social network aggregator considers the following events to determine end of a session (a) when a user closes the web browser or logs out or (b) when a user does not engage in any action for more than an arbitrarily set period of time. The system uses a 20 minute threshold. To check the sensitivity of this session threshold, we examined whether any two consecutive sessions of the same user had a shorter interval than 20 minutes. For 22% of all sessions (generated by 13% of all users), an earlier session by the same



(a) Frequency of accesses to OSN sites (b) Total time spent on OSNs per user (c) Session duration per OSN site

Figure 3: Session level characteristics of OSN workload

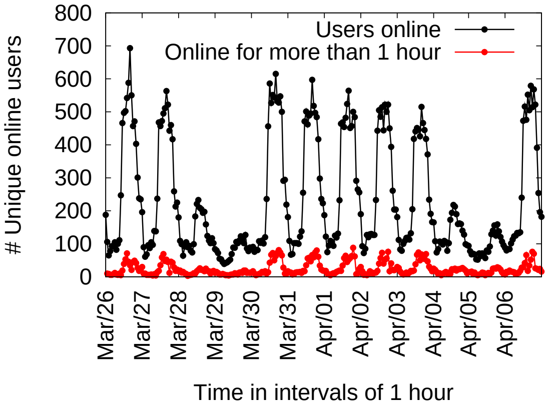


Figure 2: Number of online users over time

user ended less than 20 minutes prior (i.e., 22% of sessions were solely identified by events of closing of web browsers or logging out). For analysis, we used the session information that is identified by the social network aggregator.

Utilizing the session information, we first examined the number of concurrent users (i.e., concurrent sessions) that accessed any of the four OSN sites (Figure 2). The beginning of each day is marked in the horizontal axis. We see a diurnal pattern with strong peaks around 3 PM (in Brazil). At all times, there are at least 50 people who are using the social network aggregator service. At peak times, the number of concurrent users surpasses 700, more than a 10-fold increase over the minimum. Drops in usage on certain days indicate clear weekly patterns, where weekends showed a much lower usage than weekdays. The strong diurnal pattern in OSN workloads has also been observed in accessing messages and applications on Facebook [11] and in the content generation of blog posts, bookmarks, and answers in user generated content (UGC) websites [9, 13].

To see the usage pattern of heavy OSN users, we also show in Figure 2 the number of users who stayed online for more than 1 hour at any given point in time. The daily peaks for heavy users coincide with the peaks from all users. The total number of online users and the number of heavy users showed a strong correlation; the Pearson's correlation coefficient was 0.84. This indicates that the ratio between the heavy users and all users is oblivious to the time of day. The gap between the two data points in the figure also indicates that there are users who login and connect for less than an hour throughout the day.

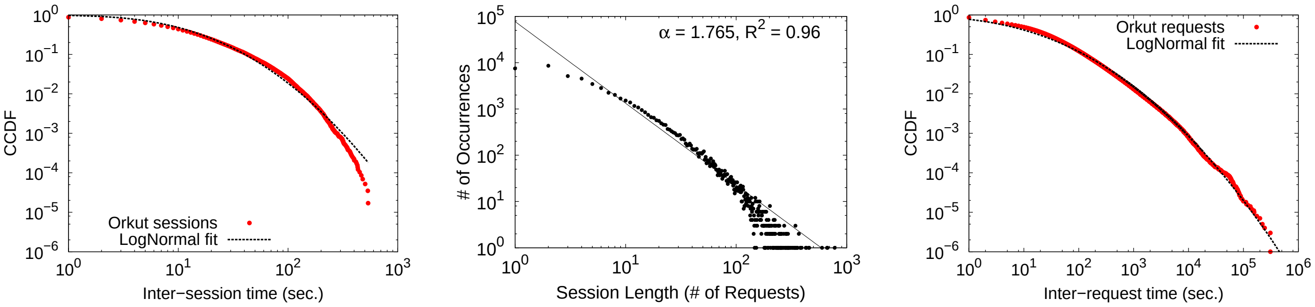
3.2 OSN session characteristics

So, how ofen and for how long do people connect to OSN sites? To estimate these quantities, we measure the frequency and duration of sessions for each user. We calculate session duration as the time interval between the first and the last HTTP requests within a session. This approach allows us to infer the duration of any session with two or more HTTP requests. 87% of all sessions in the dataset contained at least two HTTP requests.

Individuals varied widely in the frequency with which they accessed social networks. Figure 3(a) shows the cumulative distribution function (CDF) of the total number of sessions per user. The majority of users (63%) accessed the social network aggregator's site only once during the 12-day period. The most frequently logging in user accessed the social network aggregator's site on average 4.1 times a day. The total time spent accessing social networks also varied largely per individual, as shown in Figure 3(b). On one hand, 51% of the users spent no more than 10 minutes at the social network aggregator's site over the 12 days. On the other hand, 14% of the active users spent in total more than an hour and the most active 2% of the users spent more than 12 hours (i.e., an average of an hour a day).

Across all users, we did not see a high correlation between the frequency and duration of OSN accesses (correlation coefficient 0.27). This means that the amount of time a user spends on social networks is not strongly correlated to the specific number of times that the user logins to social networks. We also did not see a strong correlation between a session duration and the number of HTTP requests made during the session (correlation coefficient 0.16). The correlation became relatively stronger when we considered relatively short sessions that lasted less than 20 minutes (correlation coefficient 0.49). This may suggest that long sessions tend to have idle users. For short sessions, the longer the session duration, the more activities the session contains.

In addition to widely varying OSN usage per individual, session durations also varied widely across the four OSN sites. Figure 3(c) shows the CDF of the session durations for each OSN site. All four OSN sites exhibit a consistent heavy-tailed pattern in their session durations. However, the median session durations vary across OSNs. The median session durations of Orkut, Hi5, and MySpace are 13.4 minutes, 2.7 minutes and 24 seconds, respectively, indicat-



(a) Inter-session time

(b) Session length

(c) Inter-request time

Figure 4: Characteristics of Orkut sessions and the best fit functions

ing that users likely engage in a series of activities when they connect to these sites. In contrast, the median session duration of LinkedIn is very short (3 seconds). In the following section, we take a deeper look into which activities are popular across these sites.

3.3 Modeling Orkut sessions

To understand the dynamics of user arrival and departure processes from a system's perspective, we measure the session inter-arrival times. Here, we present a case study for Orkut. More formally, we utilize a time series $t(i), i = 1, 2, 3, ...$ to denote the arrival time of the ith session in the trace. The time series $a(i)$ is defined as $t(i+1) - t(i)$ and it denotes the inter-arrival time of the ith and $i+1$th sessions, where sessions may belong to different users. Figure 4(a) shows the complementary cumulative distribution function (CCDF) of $a(i)$, which we fitted to a Lognormal distribution. The probability distribution function for the lognormal distribution is given by:

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-(log(x)-\mu)^2}/2\sigma^2 \qquad (1)$$

with parameters $\mu = 2.245$ and $\sigma = 1.133$.

To characterize the period of time during which a session is active, we use a time series $l(i)$ which denotes the length of the ith session in the trace, defined as the number of requests in that session. Figure 4(b) shows the frequency marginal distribution of $l(i)$ for all sessions identified in the Orkut trace. We observe a heavy-tail distribution; most of the sessions involve very few HTTP requests, while a small number of sessions involve a large number of HTTP requests. This implies significant deviations in the number of actions (or clicks) users make in a single session.

The distribution was fitted to a Zipf distribution of the form $\beta x^{-\alpha}$ with parameters $\alpha = 1.765$ and $\beta = 4.888$. A Zipf-like distribution suggests that session lengths are highly variable when users connect to online social networks. Such high variability is in line with the patterns seen in web surfing. Huberman et al. [15] also found strong variability in the number of clicks a user exhibits in a session, as well as when navigating a given website.

The last variable we characterize at the session layer is the inter-arrival time between requests within a single session. Figure 4(c) displays the CCDF distribution that was fitted to a Lognormal distribution, with parameters $\mu = 1.789$ and $\sigma = 2.366$. Large inter-arrivals would correspond to users leaving Orkut pages to spend time on other social networks or other features of the social network aggregator then returning back to Orkut. On the other hand, small inter-arrivals would correspond to users constantly interacting with the social networking site. We found that the average session lengths and the session starting times are not correlated (the Pearson's correlation coefficient is -0.027). This suggests that the high variability in session length is not due to diurnal pattern in user behaviors (as was the case with the number of active clients), but rather it is a fundamental property of the interaction of OSN users.

The combination of request inter-arrival time and session length provides an important model for understanding the behavior of OSN users, for the two quantities reflect the inherent nature of OSN users and are not related to load (e.g., the number of active sessions) or time of the day. The best fit distribution functions presented in this section can be used to generate synthetic (parameterizable) traces, that mimic actual OSN workloads.

4. THE CLICKSTREAM MODEL

In this section we present a comprehensive view of user behavior in OSNs by characterizing the type, frequency, and sequence of activities users engage in. We developed a new analysis strategy, which we call the *clickstream model*, to identify and describe representative user behaviors in OSNs based on clickstream data.

The modeling of the system implies two steps. The first step is to identify dominant user activities in clickstreams. This step involves enumerating all features users engaged in on OSNs at the level of basic unit, which we call *user activity*. We manually annotated each log entry of the clickstream data with the appropriate activity class (e.g., friend invitation, browsing photos), based on the information available in the HTTP header. Because a user can conduct a wide range of activities in a typical OSN site, we further tried to group semantically similar activities into a *category* by utilizing the webpage structure of OSN sites (i.e., which set of activities can be conducted in a single page) and manually grouping related activities into categories.

The second step of modeling is to compute the transition rates between activities. To represent the sequence in which activities are conducted, we built a first-order Markov chain of user activities and compute the probability transition between every pair of activity states. To gain a holistic view, we built a Markov chain that describes how users transition from actions in one category to another.

Different OSNs provide different features, potentially leading to a substantial variation in the set of popular user activities. Our analysis in this section highlights the similarities and differences in user behaviors across four different social networks in the trace. Below we present the full clickstream model only for Orkut, which is the most accessed OSN in the trace.

4.1 User activities in Orkut

In the first step of modeling, we identified 41 activities with at least one HTTP request in the clickstream data. We grouped these activities into the following categories: Search, Scrapbook, Messages, Testimonials, Videos, Photos, Profile & Friends, Communities, and Other. Table 2 displays the list of 41 activities with the number and share of users who engaged in the corresponding activity at least once, the number and share of HTTP requests, and the total traffic volume both received and sent by users.

The activity categories listed in Table 2 represent the following features in Orkut, which are described in more detail in [24, 32]:

- **Universal search** (activity 1) allows users to search for other people's profiles, communities, and community topics (or forums) in the entire Orkut website. A search box appears at the upper right corner of every Orkut page, allowing users to engage in the search feature from any page.

- **Scrapbook** (activities 2 and 3) displays all text messages sent to a given user. Unlike personal messaging or email, Scrapbook entries are public, meaning that anyone with an Orkut account can read others' *scraps.* By default, anyone can leave a scrap in a user's scrapbook. However, users can set their scrapbook to be private, so that only friends or friends of friends in the network can leave a scrap. Table 2 shows that browsing and writing scraps is one of the most popular forms of user interaction in Orkut.

- **Messages** (activities 4 and 5) are a private way to communicate. Messages can be sent by anyone. Table 2 shows that the messages feature is not widely used in Orkut.

- **Testimonials** (activities 6 to 8) are a commentary that users leave about his or her friends. Testimonials can only be written by friends, but can be viewed by anyone by default. A user can set options so that testimonials are kept private, and only the user's friends can view the testimonial page. Compared to the interaction through scrapbook, we see much less interaction through testimonials.

- The **Videos** (activities 9 and 10) and **Photos** (activity 11–16) categories incorporate all activities in which users share multimedia content. The photos category is another popular activity in Orkut. A photo can be tagged and commented on only by friends. However, a photo can be viewed by anyone by default. To share a video, Orkut asks users to first upload their videos to YouTube then to add the video URLs at the Orkut's video page.

- **Profile & Friends** (activities 17–26) represent all activities in which users manage their own profiles or visit other people's profiles. Orkut allows anyone to visit anyone's profile, unless a potential visitor is on the "Ignore List" (a list where a user specifies other users who he or she wants to block from any form of interaction). Users can customize their profile preferences and can restrict the information that appears on their profile page from other users.
  A user's homepage displays a short list of updates about the user's friends. The homepage also displays a short list of friends ordered by login time, where the first person is the one who logged in most recently.

- **Communities** (activities 27–37) can be created by anyone with an Orkut account. Community members can post topics, inform other members about an event, ask questions, or play games. Users can freely join any public community, while a moderated community requires explicit approval. Invitations to join a community are sent through messages.

The statistics of user activity in Table 2 suggest interesting trends in the usage of Orkut. Browsing (marked with a * sign) is the most common user behavior, both in terms of the number of users and the request volume. In fact, browsing accounted for 92% of all requests! Compared to other non-browsing activities in the same category, browsing typically engaged 2 to 100 times more users. For instance, the number of users who ever browsed messages was 13 times larger than those who sent messages. In fact, other behaviors that require more user engagement were less prominent in the trace; time-intensive behaviors like browse a favorite video (activity 10) and participation-oriented behaviors like posting in a community topic (activity 32) are not popular.

Our findings demonstrate that many Orkut users primarily use the service for passive interactions such as browsing updates from their friends through homepage, profile pages, and scrapbook, while occasionally engaging in more active interaction such as writing scraps, searching, editing photos, and accessing applications.

4.2 Comparison of user activity across OSNs

To get perspective on how user behaviors vary across different social networks, we repeated the analysis in Table 2 for other social networks that appear in the trace (i.e., MySpace, LinkedIn, and Hi5). All four OSNs exhibited a common pattern in that the most popular activity was browsing profiles. Some activities, however, could only be observed in a subset of these four networks, because the four social networks provided different features to users. For example, MySpace uniquely provided Blogs and News pages and LinkedIn uniquely provided Jobs and Companies pages. Also video and photo features are not supported in LinkedIn.

Table 3 displays for all four social networks the top five categories based on the number of HTTP requests and the share of corresponding HTTP requests. The statistics are normalized for each social network, so that the sum of share of all activity categories is 100% for each social network.

We make several observations. First, the Profile & Friends category is the most popular across all social networks. Users commonly browsed profiles, homepage, and the list of friends across all four networks.

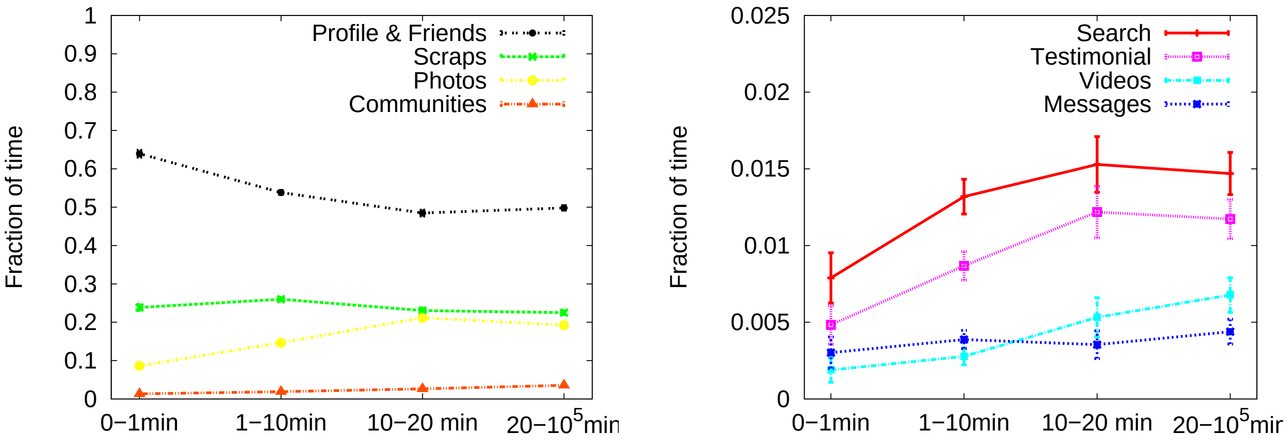
Second, LinkedIn shows a much lower degree of interaction among users using messages than Orkut. Only 4% of the requests in LinkedIn are related to messaging between users.

Category	ID	Description of activity	# Users	(%)	# Requests	(%)	Bytes (MB)
Search	1	Universal search	2,383	(2.1)	15,409	(2.0)	287
Scrapbook	2	*Browse scraps	17,753	(15.9)	147,249	(18.7)	2,740
	3	Write scraps	2,307	(2.1)	7,623	(1.0)	113
Messages	4	*Browse messages	931	(0.8)	3,905	(0.5)	64
	5	Write messages	70	(0.1)	289	(<0.1)	5
Testimonials	6	*Browse testimonials received	1,085	(1.0)	3,402	(0.4)	57
	7	Write testimonials	911	(0.8)	4,128	(0.5)	65
	8	*Browse testimonials written	540	(0.5)	1,633	(0.2)	26
Videos	9	*Browse the list of favorite videos	494	(0.4)	2,262	(0.3)	44
	10	*Browse a favorite video	390	(0.3)	862	(0.1)	13
Photos	11	*Browse a list of albums	8,769	(7.8)	43,743	(5.6)	871
	12	*Browse photo albums	8,201	(7.3)	70,329	(8.9)	2,313
	13	*Browse photos	8,176	(7.3)	122,152	(15.5)	1,147
	14	*Browse photos the user was tagged	1,217	(1.1)	3,004	(0.4)	47
	15	*Browse photo comments	355	(0.3)	842	(0.1)	16
	16	Edit and organize photos	82	(0.1)	266	(0.0)	3
Profile & Friends	17	*Browse profiles	19,984	(17.9)	149,402	(19.0)	3,534
	18	*Browse homepage	18,868	(16.9)	92,699	(11.8)	3,866
	19	*Browse the list of friends	6,364	(5.7)	50,537	(6.4)	1,032
	20	Manage friend invitations	1,656	(1.5)	8,517	(1.1)	144
	21	*Browse friend updates	1,601	(1.4)	6,644	(0.8)	200
	22	*Browse member communities	1,455	(1.3)	6,963	(0.9)	133
	23	Profile editing	1,293	(1.2)	7,054	(0.9)	369
	24	*Browse fans	361	(0.3)	1,103	(0.1)	17
	25	*Browse user lists	126	(0.1)	626	(0.1)	9
	26	Manage user events	44	(<0.1)	129	(<0.1)	2
Communities	27	*Browse a community	2,109	(1.9)	8,850	(1.1)	164
	28	*Browse a topic in a community	926	(0.8)	9,454	(1.2)	143
	29	Join or leave communities	523	(0.5)	3,043	(0.4)	43
	30	*Browse members in communities	415	(0.4)	3,639	(0.5)	56
	31	*Browse the list of community topics	412	(0.4)	2,066	(0.3)	38
	32	Post in a community topic	227	(0.2)	1,680	(0.2)	24
	33	Community management	105	(0.1)	682	(0.1)	12
	34	Accessing polls in communities	99	(0.1)	360	(<0.1)	6
	35	*Browse the list of communities	47	(<0.1)	337	(<0.1)	8
	36	Manage community invitations	20	(<0.1)	63	(<0.1)	1
	37	Community events	19	(<0.1)	41	(<0.1)	1
Other	38	Accessing applications	1,092	(1.0)	4,043	(0.5)	61
	39	User settings	403	(0.4)	2,020	(0.3)	32
	40	Spam folder, feeds, captcha	48	(<0.1)	150	(<0.1)	2
	41	Account login and deletion	39	(<0.1)	76	(<0.1)	1
Total			36,309 (distinct)		787,276		17.3 GB

Table 2: Enumeration of all activities in Orkut and their occurrences in the clickstream data. Events related to browsing are marked with a (*) sign.

	Orkut		MySpace		LinkedIn		Hi5	
Rank	Category	Share	Category	Share	Category	Share	Category	Share
1	Profile & Friends	41%	Profile & Friends	88%	Profile & Friends	51%	Profile & Friends	67%
2	Photos	31%	Messages	5%	Other (login)	42%	Photos	18%
3	Scrapbook	20%	Photos	3%	Messages	4%	Comments	6%
4	Communities	4%	Other (login)	3%	Search	2%	Other (login)	4%
5	Search	2%	Communities	1%	Communities	<1%	Messages	3%

Table 3: Comparison of popular user activities across four OSN sites



(a) Probability of dominant activities

(b) Probability of subdominant activities

Figure 5: Probability of the user activity as a function of session duration (error bars indicate 95% confidence interval)

Because Linkedin is a network used mainly for professional networking (e.g., finding jobs or employees), it is natural to expect that users primarily browse profiles and create links with each other, rather than exchanging messages.

Third, MySpace showed a different profile from Orkut, despite the similarity of its service to that provided by Orkut. MySpace showed a much lower interaction through Photos. A detailed look into the data reveals that 90% of the MySpace users also accessed one of the other three social networks (75% accessed Orkut). Thus, it seems that users who accessed MySpace using the social network aggregator use Orkut as their primary social network and access MySpace to keep in touch with friends that use only MySpace.

Fourth, the popular user activities in Hi5 were similar to those of Orkut: the most frequent user activity involved browsing friends' updates through Profile & Friends and Photos. The next most popular user activity in both OSNs was a form of message interaction among users: Scrapbook in Orkut and Comments and Messages in Hi5. We expect to see similar usage trends for other social networks that possess similar service characteristics to Orkut.

4.3 Probability of activity over time

We next investigated whether there is any correlation between the occurrence of a particular activity and session duration. To check for such a correlation, we categorized user sessions into four non-overlapping classes based on their session durations: (a) less than 1 minute, (b) 1 to 10 minutes, (c) 10 to 20 minutes, and (d) 20 minutes or longer. For sessions belonging to each of these intervals, we examined the average proportion of the total session duration that a user spent on each activity.

Figure 5 shows the fraction of time spent on each activity as a function of session duration. The results are shown in two separate plots to more easily exhibit the trends for both dominant and subdominant activities. We found two key patterns. First, irrespective of session duration, users spent the most time on Profile & Friends and Scrapbook activities. In very short sessions (i.e., less than 1 minute), users spent 90% of their time on these activities. However, even for a long session (i.e., 20 minutes or longer), the two activities accounted for 75% of the total. Second, the remaining categories of activities became more prevalent for longer sessions. The fraction of time spent consuming media content (i.e., Photos and Videos activities) increased by a factor of 2 when comparing sessions shorter than 1 minute to those longer than 20 minutes. The probability of seeing Community activity also increased with the session duration.

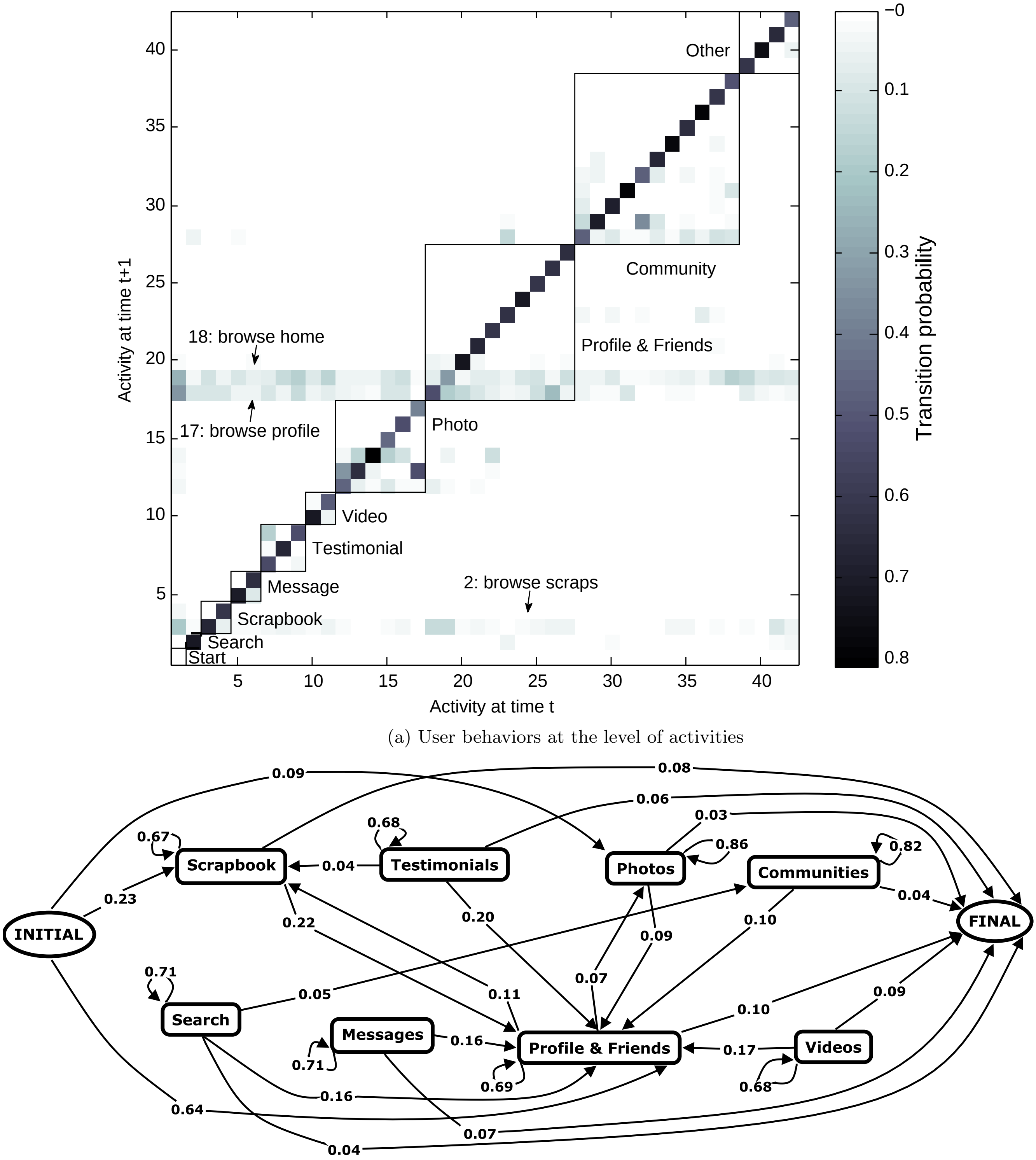
4.4 Transition from one activity to another

In the second step of modeling, we constructed a first-order Markov chain of user activity based on the sequence of activities seen from all sessions. We added two abstract states, *initial* and *final*, which we appended to the sequence of requests at the beginning and the end of the user sessions, respectively.

Figure 6(a) shows the transition probability between all pairs of activities. A color pixel at (x,y) represents the probability of transition from activity x in the horizontal axis to activity y in the vertical axis. Activity IDs in the figure are identical to the activity IDs in Table 2. We also visually show the boundaries for categories. Darker pixels indicate higher transition probability. For visual clarity, probabilities below 0.01 are shown as zero probability in the figure.

When users log in to the social network aggregator site, they are immediately exposed to a small selection of updates from all social networks. Users can then click on any of the displayed web objects or the logo of a social network to further browse a given social network. These events are shown as dark pixels on the first column in Figure 6(a). For example, x="Start" and y="browsing homepage" illustrates the case when a user clicked on the logo of a social network and the homepage of the social network was displayed. A typical session started with one of the following activities: browsing scrap, browsing profile, and browsing homepage.

Once a user engaged in a particular activity, the user was likely to repeat the same activity. This is shown by a strong linear trend in $y = x$. For instance, after browsing one photo, a user was likely to immediately browse other photos. In total, 67% of the user activities were repeated.



(a) User behaviors at the level of activities

(b) User behaviors at the level of categories

Figure 6: Transition probability in the clickstream model for Orkut

Next there were more transitions of activities within the same category (77%) than across categories (23%). This means that users typically conduct a sequence of activities that are conceptually related. For instance, a user is likely to browse photos immediately after browsing the list of photo albums, rather than after conducting a less related activity like accessing applications.

We also notice that popular activities like browsing homepage, browsing profiles, and browsing scraps display characteristic horizontal stripes in the graph. This is because every Orkut page embeds hyperlinks to a user's homepage, profile page, and scrapbook page. This suggests that providing a means for users to access a particular feature easily can motivate users to use the given feature frequently.

4.5 Transition from one category to another

Finally we examined the sequence of user activities at the level of categories (Figure 6(b)). Again we added two synthetic states, Initial and Final, at the beginning and the end of each session. Nodes now represent categories and directed edges represent the transition between two categories. Edges with probability smaller than 4% were removed to reduce the figure complexity. The sum of all outgoing probabilities (including the omitted edges) for each state is 1.0. Compared to Figure 6(a), user behaviors at the category level provide a more holistic view of OSN usage.

We observe that most users initiated their sessions from the Profile & Friends, Scrapbook, or Photos category, as mentioned earlier. We also observe that self loops are present in almost all states. For example, one Communities activity was followed by another Community activity with a probability of 0.82. Similarly, Photos activities showed high repetition with a probability of 0.86. Repetition also occurred in Search (probability 0.71). Repetition in Scrapbook was related to users replying to received scraps after browsing them. In Orkut, users can directly reply to an existing (received) scrap from one's own Scrapbook page. We found that 65% of write scrap events (activity 3) immediately followed browsing scrap events (activity 2). Except for self loops, Profile & Friends was the most common preceding state for most activities.

5. SOCIAL INTERACTIONS IN ORKUT

One crucial aspect of OSNs is the wide range of features that support communication between users. In this section, we investigate how users interact with each other through the various features OSNs provide.

5.1 Overview

Understanding social interactions has been of great interest in various research fields like sociology, economy, political science, and marketing. Until recently, obtaining large-scale data was one of the key challenges in studying social interactions. Nowadays, we get around this challenge by the wealth of OSN data available on the Internet. A few studies have used publicly crawled OSN data (e.g., comments, testimonials) to characterize social interactions [8,14,30,33]. Although these initial studies have identified several important properties of social interaction, there are behaviors of users that cannot be measured with datasets that contain only visible activity.

One such activity is browsing, which, as demonstrated in the previous section, is one of the most frequent activities in OSNs.[1] As opposed to "visible" interactions that are inferred from crawled data like writing a scrap, browsing a friend's web content can be considered "silent" social interaction. Although visible and silent interactions serve different purposes, both are interesting for the understanding the social interaction behaviors of users.

In this section we provide a complete view of user interaction in social networks, by considering both visible interaction and silent interaction. Our goal is two-fold: (a) We would like to know what fraction of user interaction is silent, compared to visible. If we consider visiting a friend' profile or photo pages as social interaction among users, how much increase would we observe in the number of friends a user typically interacts with? We highlight the potential bias in studies of user interactions using only visible data. (b) We are interested in knowing the interaction patterns among users along the social graph distance. In particular, how often do users visit their friends' profiles or even traverse multiple hops to visit the profile of friend of a friend?

5.2 Interaction over social network distance

We only considered explicitly visiting another user's page to be silent user interaction. It is possible that a user can silently "interact" with a friend by viewing the short list of updates about that friend that are automatically shown on the user's own homepage. However, we do not count these views as interaction, because we cannot be certain whether a user noticed these updates.[2] For example, a user may find a thumbnail of photo update from a friend at her homepage. Only when the user clicks on the photo (thereby visiting the friend's photo page), do we then consider the event as a valid social interaction with a 1-hop friend.

To gain a comprehensive understanding on the social behavior of a user, we needed an essential piece of information: the list of friends of a given user. The clickstream dataset does not include information about the list of friends. Therefore, as described in Section 2.3, we gathered information about the list of friends for all users in the workload trace by crawling the Orkut website.

5.2.1 Webpage access patterns

To investigate the patterns of interaction among users, we first examined how often users visit their friends' pages, compared to visiting their own. Not all accesses in the trace were related to interaction among users. Therefore, we focused on the following activities as a form of user interaction: scrapbook, messages, testimonials, videos, photos, and profile & friends. This list comprises activities from 2 to 26 in Table 2. We excluded all activities related to search, communities, and others in Table 2.

Figure 7 shows, for each category of user activity, the fraction of times a user was accessing one's own page (denoted *self* in the figure), a page of an immediate friend (denoted *friend*), or a page of a non-immediate friend (denoted *2+hops*). The result for Messages is omitted, because users can only access their own Messages page. Unless a user has explicitly restricted access, Orkut users can browse

[1] Most social networks do not log browsing events of users. However, one exception is Orkut. In Orkut, the list of "recent visitors" to every profile page is shown. Users can also turn this option off and hide their browsing patterns.

[2] User studies using eye tracking devices will be able to distinguish whether users noticed the exposed content or not.

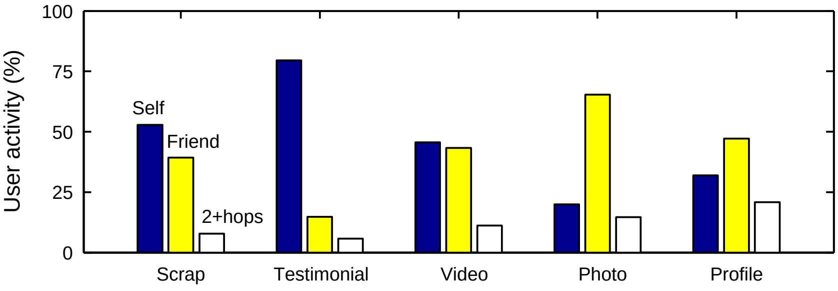


Figure 7: Webpage accesses along the social network distance

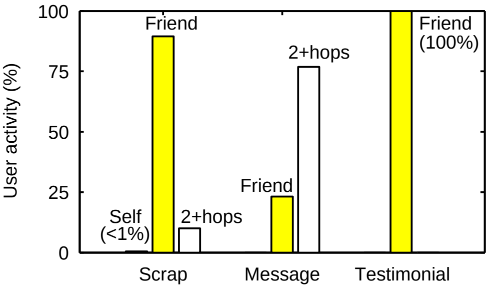


Figure 8: Interaction in writing

any other user's pages containing scrapbook, testimonials, video, photo, and profile. However, the bar chart shows that users mostly accessed pages of their own or their immediate friends; 80% of all accesses remain within a 1-hop neighborhood in the social network topology.

We examined each of the activity categories in detail. Users most frequently accessed their own pages when it comes to scrapbook and testimonials. Yet, users did visit scrapbook and testimonial pages of their 1-hop friends and read what messages are written about their friends. With a small probability, users also visited beyond the 1-hop neighborhood. In total, Orkut users accessed their friends' pages more frequently (59%)[3] than their own pages. When visiting friends' pages, Orkut users not only interacted with immediate friends, but also had significant exposure to non-immediate friends (22%=13/59).

Focusing on each category of interaction, Table 2 shows that users accessed their own video pages as often as they accessed their friends' video pages. On the other hand, in accessing photos, which is a popular activity in Orkut, users were more likely to access their friends' photo pages than their own. Accessing profile pages was well-divided among one's own, immediate friends, and non-immediate friends; 20% of the browsed profiles were 2 or more hops away.

Next we focused on visible interactions and examined which friends users interacted with. We considered the following three visible activities: write scraps (activity 3), write messages (activity 5), and write testimonials (activity 7), because for these activities we could determine the interaction partner from the URL of the trace.

Figure 8 shows the division of the times when a user wrote to oneself, a 1-hop friend, or a 2 or more hop away friend. When using the scrapbook feature, users mostly interacted with immediate friends. Self posts were rare (0.5%), but could serve as a broadcast message to everyone who visits the scrapbook. Interestingly, 10% of scraps were sent to users that are 2 or more hops away. On the other hand, users did not interact much with immediate friends through Messages. Instead, we observed frequent interaction with non-immediate friends through Messages (76%). Testimonials were only sent to immediate friends, as written in the Orkut policy. We discuss the implications of these findings in the following section.

5.2.2 *What leads users to visit other people's pages?*

Having studied the frequency at which users access their friends' pages, we now take a closer look at how a user navigates from one friend's page to another. Particularly, we are interested in understanding what activities lead users to visit a page of a friend or a non-friend. We performed the following analysis. Each time a user visited a page of a friend, we examined which preceding page the user was at: one's own page, an immediate friend's page, or a non-immediate friend's page? Table 4 shows the fraction of preceding locations for every first access to a friend's page in each session. In addition to the navigation statistics, Table 4 also shows the list of top activities that preceded the navigation event.

The majority of accesses (68%) to an immediate friend's webpage originated from browsing one's own webpage (the first row of Table 2). The remaining accesses occurred when the user was navigating the social network; accesses to an immediate friend's webpage were followed by browsing of another immediate friend's webpage (25%) or browsing of a non-immediate friend's webpage (7%). When it comes to visiting a non-immediate friend's webpage (the second row of Table 2), the preceding location of the user was well distributed across 0-hop, 1-hop, and 2 or more hops.

Interestingly, the most popular activity that leads a user to an immediate or non-immediate friend's webpage is browsing one's own homepage. As described in Section 4.1, a user's homepage contains a short list of updates from friends as well as a list of the subset of friends who recently logged in. Such updates can contain links to non-immediate friends when they interacted with mutual friends through photo comments, testimonials, or applications. Therefore, updates from friends can also drive users to visit the webpage of a friend of a friend.

Another interesting observation we make is the high fraction of accesses that originated from an immediate friend's webpage, which accounted for 25% of the accesses to another immediate friend's webpage and 30% of the accesses to a non-immediate friend's webpage (the third column of Table 2). This reinforces the previous findings that users in social networks find new content and contacts through their 1-hop friends [4,5,28]. Browsing an immediate friend's profile was the most common gateway that led users from one friend to another.

Lastly, we note that browsing scraps (activity 2) appears in the top three activities in all the rows of Table 4. This may mean that Orkut users are keen on reading other users' scrapbook content and also are curious about checking out new contacts that they encounter through such activity.

[3] This probability is computed as the count of activities at 1-hop divided by the total occurrences of all activities.

Current location (First access to)	Preceding location 0-hop		1-hop		2 or more hops	
Immediate friend's page (1-hop)	Total 68%		Total 25%		Total 7%	
	○ Browse homepage	(36%)	○ Browse profiles	(8%)	○ Browse profiles	(4%)
	○ Browse scraps	(12%)	○ Browse scraps	(6%)	○ Browse scraps	(1%)
	○ Browse friends list	(9%)	○ Browse photos	(3%)	○ Browse photos	(<1%)
Non-immediate friend's page (2 or more hops)	Total 37%		Total 30%		Total 33%	
	○ Browse homepage	(22%)	○ Browse profiles	(9%)	○ Browse profiles	(15%)
	○ Browse scraps	(6%)	○ Browse scraps	(9%)	○ Browse friends list	(5%)
	○ Browse profiles	(3%)	○ Browse friends list	(5%)	○ Browse scraps	(5%)

Table 4: How users arrive at other people's pages: preceding locations and activities for every first visit to an immediate and non-immediate friend's page

5.3 Number of friends interacted with

Finally we investigated how silent interactions affect the level of user interactions along the social network topology. We compare the number of friends (including multi-hop friends) a user interacts with through all activities with the number interacted with through only visible activities, as a function of the number of friends in the social graph.

Figure 9 shows these quantities. Overall, the degree of interaction is very low; the average user interacted (whether visibly or silently) with 3.2 friends in total over the 12-day period and interacted visibly with only 0.2 friends. This low level of interaction has also been observed in other work. According to Wilson et al. [33], in the Facebook social network nearly 60% of users exhibit no interaction at all over an entire year. Therefore, our workload trace of 12 days is expected to show a much lower level of interaction.

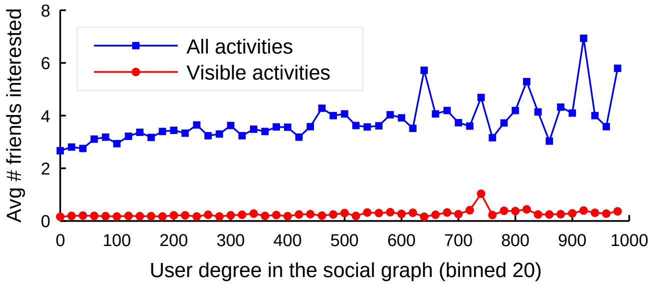


Figure 9: Comparison of the Orkut social graph degree and interaction degree

Interestingly, even for a short trace period, the degree of all interaction is 16 times or an order of magnitude greater than the degree of visible interaction. The stark difference in the two quantities may be because in OSN usage the majority of time is spent browsing, which cannot be captured by visible interactions. Another trend that we observe is that interaction degree does not grow rapidly with the user degree in the social graph; users with low degree interacted with a similar number of friends as users with high degree. This indicates that it is easier to form friend links than to actually interact with those friends.

In summary our analysis of social interaction in this section brought out many interesting findings. When we consider silent interactions like browsing friends' pages, the measured interaction among users significantly increased. In total, 55% of users in the workload trace interacted with at least one other user during the 12-day period; 8% showed at least one visible interaction and 47% showed only silent interactions. This means that if one were to measure the strength of social ties based on visible traces, such analysis would be biased because 85% (=47/55) of the users would be completely disregarded. Furthermore, considering silent interactions increased the number of friends a user interacts with by an order of magnitude over the 12-day period, compared to only considering visible interactions.

6. DISCUSSION

Our measurement analysis provides many interesting findings that we think will be useful in various ways. We discuss implications of the findings below.

Modeling of OSN sessions. In Section 3, we characterized the properties of individual session properties in OSN workloads. Among various findings, here we highlight that session quantities like inter-session times, session lengths, and inter-request times follow a heavy-tailed distribution. For example, the majority of the sessions remain short (on the order of tens of minutes), but some sessions last several hours to days. As a result of the asymmetry of the distribution, user behaviors cannot be represented as a normal distribution with comparable mean and variance. Also the typical behavior of users will not be the same as their average behavior [15].

To incorporate the large variation in user behaviors, we provided statistics for the average behavior as well as the best fit distribution functions that capture this asymmetry (Section 3.3). Such distribution functions can be used to generate synthetic (parameterizable) traces, that mimic actual OSN workloads. We hope that the statistics summarized in the paper and the session modeling will be valuable in evaluating and testing potential OSN services.

Understanding user activity in OSNs. In Section 4, we characterized the type, frequency, and sequence of user activities in OSNs. Using clickstream data, we presented a complete profile of user activity in Orkut in Table 2 and Figure 6. We also examined the differences and similarities in user activity across multiple OSN sites. Our analysis demonstrated that browsing, which cannot be identified from visible data, is the most dominant behavior (92%).

We believe that understanding user activity is important for OSN service providers and portals [3, 33] as well as for advertising agencies [2]. This is because frequently repeated activities (e.g., browsing home, browsing scraps) naturally serve as good targets for advertisements and the sequence of activities can be analyzed to improve the website design. One application of our analysis is that an OSN service provider may consider providing a personalized web interface for users based on the users' activity profiles. For example, a user login page can be reorganized so that frequently repeated activities are more easily accessible. OSN service providers may also use aggregate patterns in clickstreams to identify users with similar behaviors (e.g., belonging to the same communities, possessing similar profile description) and recommend popular content within the site.

Interaction over the social graph. In Section 5 we used both the clickstream data and the social graph topology to study how users interact with friends in OSNs. Among various findings, we observed that Orkut users not only interact with 1-hop friends, but also have substantial exposure to friends that are 2 or more hops away (22%). This exposure to friends' pages has significant implication for information propagation in OSNs: OSNs exhibit "small-world" properties [1, 21, 33], which means that the network structure has a potential to spread information quickly and widely. Our observation highlighted that users actively visiting immediate and non-immediate friends' pages serves as an empirical precondition for word-of-mouth-based information propagation.

Especially when it comes to rich media content like videos and photos, more than 85% of content was found through a 1-hop friend (Figure 7). This finding reinforces some of the recent studies that emphasize the impact of word-of-mouth-like information propagation through friends in social networks (the so called *social cascade*) [4,5,28]. As OSN traffic is expected to grow rapidly [23], the patterns of social interaction and information flow can be valuable in designing the next-generation Internet infrastructure and content distribution systems [16, 26]. For instance, by tracking down the patterns of social cascade in OSNs and correlating them with information about the geographical locations of users, we can make an educated guess about the geographical regions to which particular piece of content will likely spread. Such predictions will allow for the design of efficient content distribution systems.

7. RELATED WORK

There are a rich set of studies on analyzing the workloads of Web 2.0 services. Mislove *et al.* [21] studied graph theoretic properties of OSNs, based on the friends network of Orkut, Flickr, LiveJournal, and YouTube. They confirmed the power-law, small-world, and scale-free properties of these OSN services. Ahn *et al.* [1] studied the network properties of Cyworld, a popular OSN in South Korea. They compared the explicit friend relationship network with the implicit network created by messages exchanged on Cyworld's guestbook. They found similarities in both networks: the in-degree and out-degree were close to each other and social interaction through the guestbook was highly reciprocal.

Liben-Nowell *et al.* [19] analyzed the geographical location of LiveJournal users and found a strong correlation between friendship and geographic proximity. Krishnamurthy *et al.* [17] analyzed an OSN formed by users on Twitter. They examined geographical spread of Twitter usage and also analyzed user behavior in this environment. Huberman *et al.* [14] showed that Twitter users have a small number of friends compared to the number of followers they declare. Golder *et al.* [11] analyzed temporal access and social patterns in Facebook. They analyzed the message header exchanged by Facebook users, revealing periodic patterns in terms of messages exchanged on that network. Gjoka *et al.* [10] have studied application usage workloads in Facebook and the popularity of applications. Nazir *et al.* [22] similarly analyzed application characteristics in Facebook, by developing and launching their own applications.

Wilson *et al.* [33] proposed the use of interaction graphs to impart meaning to online social links by quantifying user interactions. They analyzed interaction graphs derived from Facebook user traces and showed that they exhibit significantly lower levels of the "small-world" properties shown in their social graph counterparts. Valafar *et al.* [29] conducted a measurement study of the Flickr OSN and showed that only a small fraction of users in the main component of the friendship graph is responsible for the vast majority of user interactions.

Burke *et al.* [3] studied user motivations for contributing in social networking sites, based on server log data from Facebook. They found that newcomers who see their friends contributing go on to share more content themselves. Furthermore, those who were initially inclined to contribute, receiving feedback and having a wide audience, were also predictors of increased sharing. Chapman and Lahav [6] conducted survey interviews and analysis of web browsing patterns of 36 users of four different nationalities to examine ethnographical differences in the usage of OSNs.

Compared to the studies above, we focused on characterizing the workload of *all* user activities, beyond use of a single application and including all silent activities like browsing.

8. CONCLUSION

In this paper we presented a thorough characterization of social network workloads, based on detailed clickstream data summarizing HTTP sessions over a 12-day period of 37,024 users. The data were collected from a social network aggregator website, which after a single authentication enables users to connect to multiple social networks: Orkut, MySpace, Hi5, and LinkedIn. We analyzed the statistical and distributional properties of most of the important variables of OSN sessions. We presented the clickstream model to characterize user behavior in online social networks.

Our study uncovered a number of interesting findings, some of which are related to the specific nature of social networking environments. Many previous social network studies reconstructed user actions from "visible" artifacts, such as comments and testimonials. Using the clickstream model, we underscored the presence of "silent" user actions, such as browsing a profile page or viewing a photo of a friend. These results led us to classify social interactions into two groups, composed of publicly visible activities and silent activities, respectively.

Our current and future work is focused on leveraging the results presented in this paper along three main directions.

First, we would like to to investigate the impact of friends on the behavior of user of social networks. The success of a social networking site is directly associated with the quality of content users share. Thus, in order to design social network services, it is key to understand factors that motivate users to join communities, become fans of something, and upload or retrieve media content.

Second, we are interested in understanding content distribution patterns across multiple OSNs. We would like to know to what extent content is shared across OSN sites as well as explore the impact of age, content, and geographical locality in object popularity. Given that users participate in multiple social networks, we expect that a user may share the same content across multiple sites. Answering these questions will let us explore opportunities for efficient content distribution, for example, caching and pre-fetching, as well as advertisement and recommendation strategies. For instance, certain types of content may be popular either in a specific geographical region or in a single social network, in which case advertisement algorithms should be based on this characteristic. On the other hand, if content is easily replicated across sites, then we can detect rising content from one social networking site and implant it into another site.

Lastly, based on our analysis, we plan to build a social network workload generator that incorporates many of our findings, including the statistical distributions of sessions and requests and the Markov models for user behavior.

Acknowledgments

We are grateful to the social network aggregator for providing us with the clickstream data. We sincerely thank Krishna Gummadi, Rodrigo Rodrigues, Hamed Haddadi, Rose Hoberman, and the anonymous reviewers, for their valuable comments and suggestions.

9. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW*, 2007.

[2] B. A. Williamson. Social network marketing: ad spending and usage. *EMarketer Report*, 2007. `http://tinyurl.com/2449xx`.

[3] M. Burke, C. Marlow, and T. Lento. Feed me: Motivating newcomer contribution in social network sites. In *ACM CHI*, 2009.

[4] M. Cha, A. Mislove, B. Adams, and K. Gummadi. Characterizing Social Cascades in Flickr. In *ACM SIGCOMM WOSN*, 2008.

[5] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the Flickr social network. In *WWW*, 2009.

[6] C. N. Chapman and M. Lahav. International ethnographic observation of social networking sites. In *ACM CHI Extended Abstracts*, 2008.

[7] P. Chatterjee, D. L. Hoffman, and T. P. Novak. Modeling the clickstream: implications for web-based advertising efforts. *Marketing Science*, 2003.

[8] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Online social networks: Sheer volume vs social interaction: a case study of Cyworld. In *ACM IMC*, 2008.

[9] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *AAAI ICWSM*, 2007.

[10] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang. Poking Facebook: characterization of OSN applications. In *ACM SIGCOMM WOSN*, 2008.

[11] S. Golder, D. Wilkinson, and B. Huberman. Rhythms of social interaction: messaging within a massive online network. In *ICCT*, 2007.

[12] Google OpenSocial. `http://code.google.com/apis/opensocial/`.

[13] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. (E.) Zhao. Analyzing patterns of user content generation in online social networks. In *ACM SIGKDD*, 2009.

[14] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 2009.

[15] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in world wide web surfing. *Science*, 1998.

[16] B. Krishnamurthy. A measure of online social networks. In *COMSNETS*, 2009.

[17] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *ACM SIGCOMM WOSN*, 2008.

[18] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM TWEB*, 2007.

[19] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social network. *PNAS*, 2005.

[20] MaxMind. GeoIP Database. `http://www.maxmind.com/app/ip-location`.

[21] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *ACM IMC*, 2007.

[22] A. Nazir, S. Raza, and C.-N. Chuah. Unveiling Facebook: a measurement study of social network based applications. In *ACM IMC*, 2008.

[23] Nielsen Online Report. Social networks & blogs now 4th most popular online activity, 2009. `http://tinyurl.com/cfzjlt`.

[24] Orkut Help. `http://www.google.com/support/orkut/`.

[25] R. King. When your social sites need networking, *BusinessWeek*, 2007. `http://tinyurl.com/o4myvu`.

[26] P. Rodriguez. Web infrastructure for the 21st century. *WWW'09 Keynote, 2009.* `http://tinyurl.com/mmmaa7`.

[27] S. Schroeder. 20 ways to aggregate your social networking profiles, *Mashable*, 2007. `http://tinyurl.com/2ceus4`.

[28] N. Sastry, E. Yoneki, and J. Crowcroft. Buzztraq: predicting geographical access patterns of social cascades using social networks. In *ACM EuroSys SNS Workshop*, 2009.

[29] M. Valafar, R. Rejaie, and W. Willinger. Beyond friendship graphs: a study of user interactions in Flickr. In *ACM SIGCOMM WOSN*, 2009.

[30] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in Facebook. In *ACM SIGCOMM WOSN*, 2009.

[31] D. J. Watts and J. Peretti. Viral marketing for the real world. *Harvard Business Review*, 2007.

[32] Wikipedia. Orkut. `http://en.wikipedia.org/wiki/Orkut`.

[33] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *ACM EuroSys*, 2009.