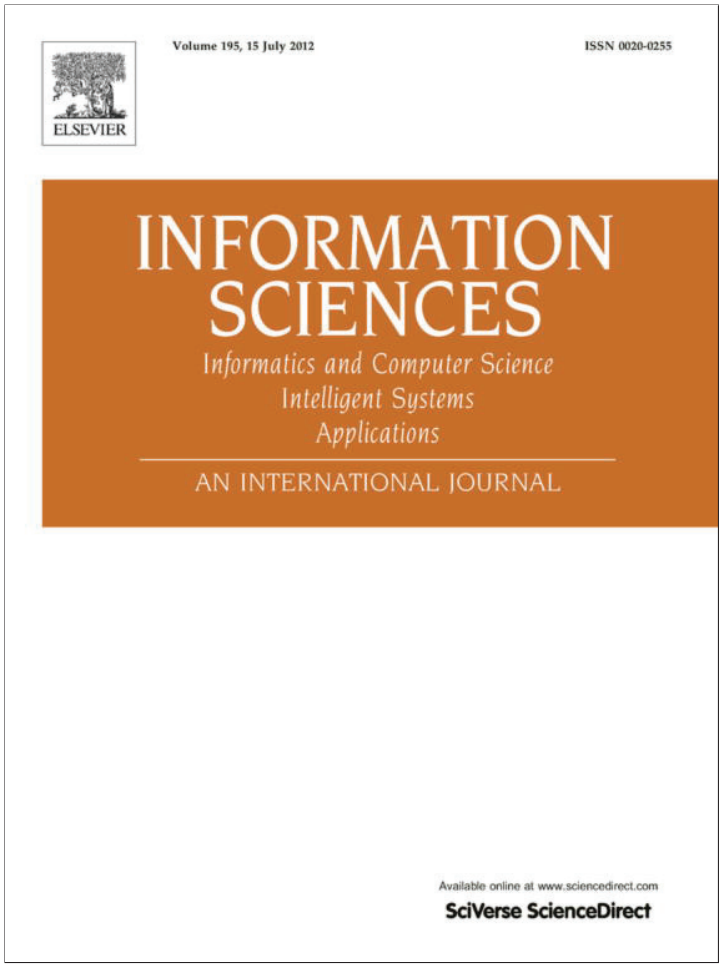
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Information Sciences 195 (2012) 1–24

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Characterizing user navigation and interactions in online social networks

Fabrício Benevenuto [a,*], Tiago Rodrigues [b], Meeyoung Cha [c], Virgílio Almeida [b]

[a] Computer Science Department, Federal University of Ouro Preto, Campus Universitário - Morro do Cruzeiro, 35400-000 Ouro Preto, Brazil
[b] Computer Science Department, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, 31270-010 Belo Horizonte, Brazil
[c] Graduate School of Culture Technology, KAIST, 373-1 Guseong-Dong, Yuseong-Gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 16 February 2011
Received in revised form 24 November 2011
Accepted 2 December 2011
Available online 13 December 2011

Keywords:
Online social networks
User behavior
Session
Clickstream
Social network aggregator
Browsing

ABSTRACT

Understanding how users navigate and interact 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 an in-depth analysis of user workloads in online social networks. This 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 gather 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' and non-immediate friends' pages. Results show 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, silent interactions like browsing friends' pages increase the measured level of interaction among users. Additionally, we find that friends requesting content are often within close geographical proximity of the uploader. We also discuss a series of implications of our findings for efficient system and interface design as well as for advertisement placement in online social networks.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Online social networks (OSNs) have become extremely popular. According to Nielsen Online's research [7], 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 [54,12] and adver-

* Corresponding author.
E-mail address: fabricio@dcc.ufmg.br (F. Benevenuto).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.12.009

tisement placement policies [35]. 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 [42,34,35]. 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 [41,33].

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 [54,19,52,28]. As a result, these studies reconstruct user actions from "visible" artifacts like messages and comments. While these 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 [18]. 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 an in-depth 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. 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 types of analyses. First, we characterized traffic and session patterns of OSN workloads (Section 4). 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 5). 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 collected the Orkut website and analyzed user activity along the social graph (Section 6). Our analysis reveals how often users visit other people's online profiles, photos, and videos. We also show that, in terms of physical distance, users usually interact mainly with local friends.

This paper provides many interesting findings:

(1) Session duration, inter-request time, and inter-session time are heavy-tailed, indicating large variations in the OSN usage among users. We provide best-fit distributions for these measures in order to provide models able to reproduce activity in Orkut sessions.
(2) Using clickstream data, we present the frequency, sequence, and duration of user activities in Orkut. We find that browsing, which cannot be inferred from publicly available data, is the most dominant behavior (92%). We also noted that users tend to repeat activities and perform a small subset of related activities per session.
(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%).
(5) The analysis of user interaction along the physical distance suggests that users mostly interact with users located within a close geographical distance. This means that while content in OSNs is created across geographically diverse regions, it is consumed locally.

In summary, our study provides an in-depth 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 and interface design as well as for advertisement placement in OSNs.

2. Related work

There is a rich set of related efforts towards characterizing user interaction and navigation in OSNs as well as exploring geo-location of users. Next we survey such efforts.

2.1. Interactions in OSNs

There has been a number of efforts that study the properties of user interactions in OSNs. A network formed by textual interactions in Cyworld was recently approached [19]. They compare the explicit friend relationship network with the implicit network created by messages exchanged on Cyworld's guestbook, finding several similarities in terms of network

structure. Particularly, they found that the in-degree and out-degree are close to each other and social interaction through the guestbook is highly reciprocated. Wilson et al. [54] use 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. Recently, Gilbert and Karahalios [22] used Facebook data to demostrated that the ''strength of ties'' varies widely, ranging from pairs of users who are best friends to pairs of users who even wished they were not friends. Cha et al. [16] studied information propagation by analyzing the spread of favorite-marking of Flickr photos. They showed that social links are a primary way users find and share information in social media (as opposed to other features such as search and hot lists). Valafar et al. [50] 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. Complementarily, Viswanath et al. studied the evolution of activity between users at Facebook, investigating how pairs of users in a social network interact and examining how the varying patterns of interaction affect the overall structure of the activity network.

More recently, Burke et al. [13] studied the roles of user interactions on Facebook, analyzing the activities of 1,193 recruited users. They quantified the usage of visible actions such as wall posts and comments and also silent actions such as consumption of friend's content. By correlating their measures with results from a survey with the volunteers, they investigated the roles that these two forms of interaction play. They showed that, different from high levels of content consumption, high levels of direct communication among users is usually associated with greater feelings of emotional support from close friends. Finally, the Facebook data team recently showed that a typical Facebook user communicates with a small subset of their entire friend network, but maintains relationships with a group that is two times larger [37].

Finally, a number of efforts have attempted to characterize and detect malicious forms of interactions in online social networks, including Facebook [21], YouTube [9], Twitter [8,25], and Foursquare [51]. In particular, Caverlee et al. [14] studies a number of vulnerabilities inherent in online social networks, and propose a framework for supporting trust establishment in these systems.

2.2. User navigation

There is also a rich set of studies on analyzing user navigation and usage on OSN websites. Through interviews with Facebook users, Joinson [31] identified seven unique reasons for users to use Facebook: social connection, shared identities, content, social investigation, social network surfing, and status updating. Burke et al. [12] studied user motivations for contributing in OSN sites based on data from Facebook. They found that newcomers who see their friends contributing 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 [17] 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. Caverlee and Webb [15] studied the characteristics of large OSNs analyzing over 1.9 million MySpace profiles as an effort to understand who is using these networks and how they are being used.

There has been a few efforts that used clickstream data to analyze user navigation in OSNs. Schneider et al. [46] analyzed OSN clickstream data extracted from network traffic and identifying typical user navigation patterns in OSNs, such as Facebook. Jiang et al. [30] used traces from a chinese social network to obtain statistics of profile visits on the network and showed that latent (or silent) interactions are much more prevalent and frequent than visible events, non-reciprocal in nature, and that profile popularity are uncorrelated with the frequency of content updates in that system.

2.3. Geo-location of OSN users

Geographical aspects of social interactions have also been studied recently. Liben-Nowell et al. [36] showed a strong correlation between friendship links and geographic location of those friends for LiveJournal users. With the recent widespread adoption of the use of online social networks through mobile devices, a key spatial dimension has been added to the study of online social networking [44]. In particular, some articles demonstrated that geographically identified social content, like chatter from Twitter, can be used to monitor real-world events and create interesting applications. Particularly, Gomide et al. [24] proposed a spatio-temporal approach to identify potential dengue epidemics, whereas Sakaki et al. [43] proposed to treat Twitter users as sensors and use them to create a mechanism for earthquake detection of earthquakes. They showed that their approach is able to send alerts faster than meteorological agencies.

Finally, Rodrigues et al. [42] studied URL propagation in Twitter and showed that there is a significant correlation between propagation and physical proximity. In particular, they show that content tends to spread for short distances only on the first hops away from the content creator. Scellato et al. [45] also studied how geographic information extracted from social cascades can be exploited to improve caching of multimedia files in a Content Delivery Network (CDN). Their evaluation showed that cache hits can be improved with respect to cache policies without geographic and social information.

Compared to the all the above efforts, our work focuses on characterizing the user navigation across *all* activities, beyond the use of a single application and including silent activities like browsing. To the best of our knowledge, we present a pioneer research on user navigation and interaction in OSNs. The present work greatly builds on our preliminary effort to characterize user navigation and interaction in OSNs [11], providing a much more thorough investigation. Particularly, we have addressed three new aspects in this work, the characterization of different types of sessions (Section 5.4), an investigation of

how interaction patterns (considering visible and silent interactions) affect content popularity (Section 6.2.3), and also a study of social interactions along the physical distance (Section 6.3).

3. 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 collected. These two datasets provide complementary types of information that we correlate in Section 6. Below we describe both datasets and our methodology for gathering Orkut data. We also discuss some limitations of these datasets.

3.1. Clickstream data

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

3.1.1. 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 [47,32]. 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 two-level real-time HTTP connections: 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 [2]. All content from OSN sites are shown to users through a social network aggregator's interface. Fig. 1 depicts the scheme of 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.

3.1.2. 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 such as 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 entries 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

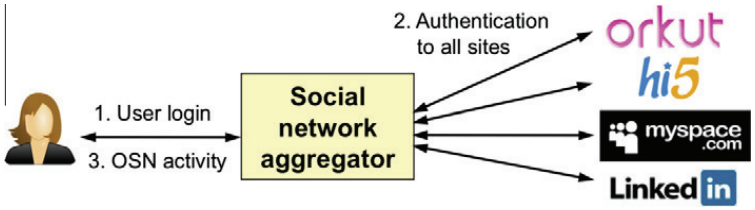


Fig. 1. Illustration of a user connecting to multiple OSNs through the social network aggregator.

Table 1
Summary of the clickstream data.

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

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, followed by Hi5 with 14,532 requests.

3.2. Social network topology of Orkut

To gain insight into user behavior over the social graph, we gather profile information from the Orkut users that appear in the clickstream dataset. We used the Orkut user IDs that appear in the trace. 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. For each user inspected, we retrieved only the information that was publicly available.

We gathered the profile information of 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.

3.2.1. 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.

The social network topology of Orkut was crawled under the supervision of the online social network aggregator service, and the user IDs in the social graph were also anonymized in the same way as the clickstream data.

3.3. 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 of the social network aggregator portal. The social network aggregator had more users accessing Orkut than other OSN sites like Hi5, MySpace, and LinkedIn. Even among the users who accessed Orkut, we could see bias in their demographics. To examine the geographical distribution of users, we used the GeoIP database [38] to identify the location of 16,175 IP addresses that appeared in the trace. Table 2 shows the location of the social network aggregator users, IP address, their requests. These users were located across all continents in the world, spanning several countries. However, certain geographical locations contained more users than others. Particularly, Brazil had the highest presence based on the number unique user IDs (68%), number of IP addresses (71%) and the number of the HTTP requests (70%). The second largest user base came from India and accounted for 15% of the users, 12% of the IP addresses, and 14% of the requests. The third and fourth most common location were the United States and South Africa. Hence, we can characterize user behavior in Brazil and India with higher accuracy.

Second, user behavior in a given social networking site is influenced by the specific mechanisms and services 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, our dataset contains only hundreds of users for three of the OSNs. Because of the small sample, we do not provide an exhaustive comparison of user behavior across different OSN sites. Additionally, we are not able to study how users inter-

Table 2
Location of the social network aggregator users.

Country	Users (%)	IP addresses (%)	Requests (%)
Brazil	68	71	70
India	15	12	14
United States	2	1	1
South Africa	1	1	1

act with their social contacts over a long period of time (e.g., several months or years), since the data were collected over a 12-day period. However, we expect that how users navigate OSN sites does not change much over time.

4. Connection pattern analysis

In this section, we characterize OSN workloads at the session level. We briefly describe how sessions are identified in the social network aggregator, then examine the duration and frequency of connections to OSN services.

4.1. Defining a session

The social network aggregator considers the following events to determine the 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 min 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 min. For 22% of all sessions (generated by 13% of all users), an earlier session by the same user ended less than 20 min prior (i.e., 22% of sessions were solely identified by events of closing of web browsers or logging out). For our analysis, we used the session information that is identified by the social network aggregator.

Using the session information, we first examined the number of concurrent users (i.e., concurrent sessions) that accessed any of the four OSN sites (Fig. 2). The beginning of each day is marked in the horizontal axis. We see a diurnal pattern with strong peaks around 3:00 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 tenfold 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 [23] and in the content generation of blog posts, bookmarks, and answers in user generated content (UGC) websites [26,20].

To see the usage pattern of heavy OSN users, we also show in Fig. 2 the number of users who stayed online for more than 1 h 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.

4.2. OSN session characteristics

So, how often 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. 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 Fig. 3(a). On one hand, 51% of the users spent no more than 10 min at the social network

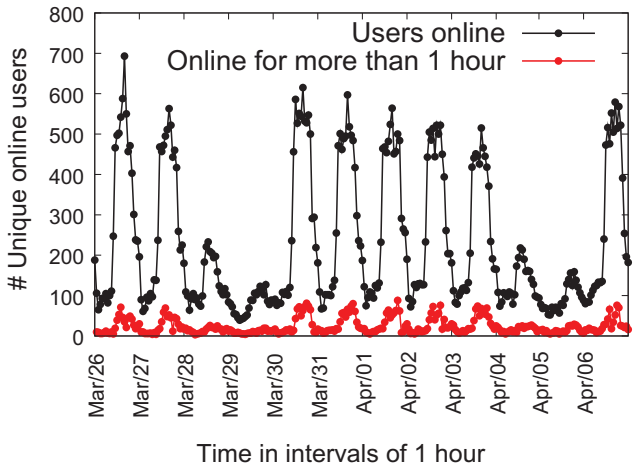
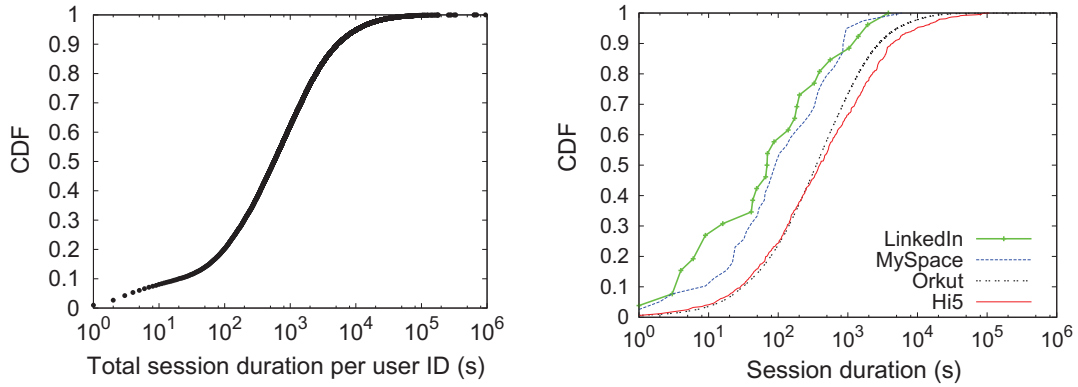


Fig. 2. Number of online users over time.



(a) Total time spent on OSNs per user (b) Session duration per OSN site

Fig. 3. Session level characteristics of OSN workload.

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 logs into 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 min (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. Fig. 3(b) 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 min, 2.7 min and 24 s, respectively, indicating 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 s). In the following section, we take a deeper look into which activities are popular across these sites.

4.3. 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 $t(i)$, $i = 1,2,3,\ldots$ 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. Fig. 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

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

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. Fig. 4(b) shows the 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. [27] 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. Fig. 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 var-

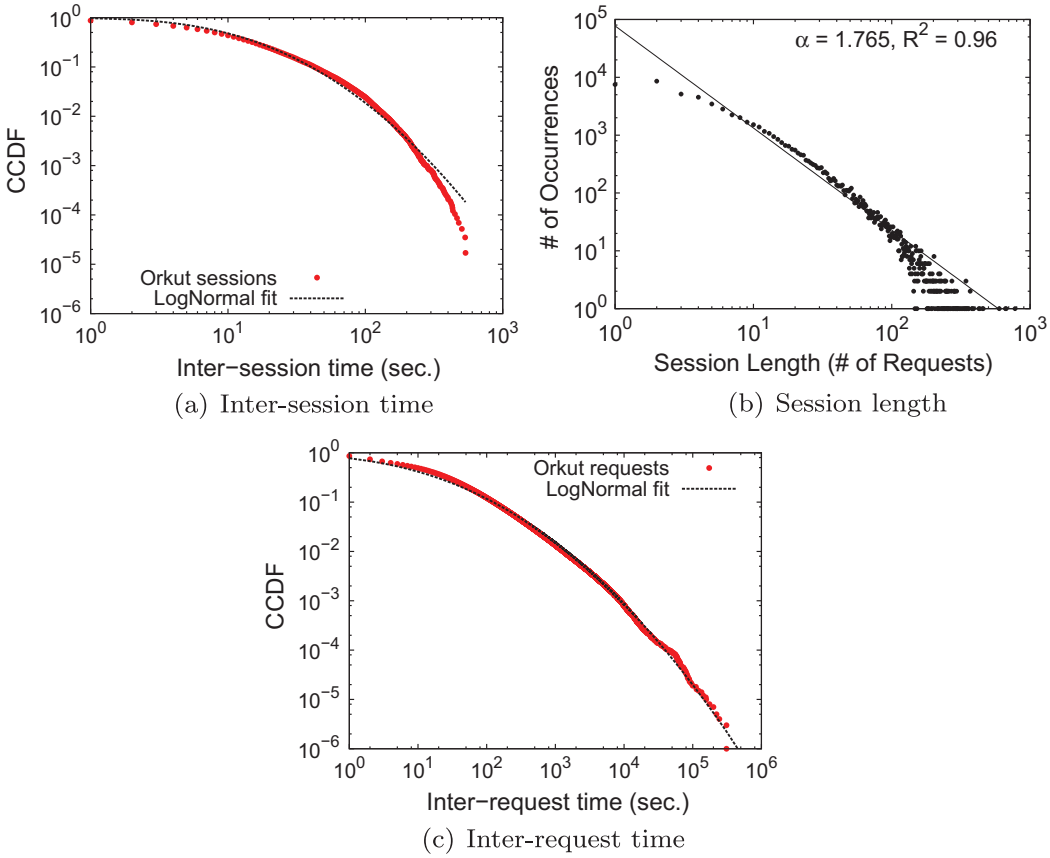


Fig. 4. Characteristics of Orkut sessions and the best fit functions.

iability 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.

5. User navigation patterns in OSNs

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., the 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.

5.1. 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 3 displays the list of 41 activities with the number and fraction of users who engaged in the corresponding activity at least once, the number and fraction of HTTP requests, and the total traffic volume both received and sent by users.

The activity categories listed in Table 3 represent the following features in Orkut, which are described in more detail in [3,4]:

- *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 3 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 3 shows that the messages feature is not widely used in Orkut.
- *Testimonials* (activities 6–8) are comments 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

Table 3
Enumeration of all activities in Orkut and their occurrences in the clickstream data.

Category	ID	Description of activity	# Users	**(%)**	# Requests	(%)	Size (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 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, and 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

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 3 suggest interesting trends in the usage of Orkut. First, we can note that the most popular activities, both in terms of the number of users and the request volume, are related to Profile & Friends. In fact, Orkut interface is designed in a way that users need to browse a user profile in order to do some activities, such as view the scrapbook of that user. Additionally, the user homepage works like the main portal in the social network where users can check updates from their friends, new messages received, upcoming birthdays, etc. Thus, it is natural to expect that users visit their homepage to check updates and spend more time on this page due to the high number of available information. The second and third most popular groups of activities are related to photos and scrapbook. Interestingly, Schneider et al. [46] made very similar observations for the Hi5 social network. However, they show a different trend for Facebook, where messages and applications tend to be the most popular activity.

Note that browsing is the most common user behavior across all categories of activities. Thus, we take a closer look at browsing related activities. We categorize them into four types: (i) browsing of media content such as photos and videos, (ii) profile content (both one's own and others'), (iii) text messages of testimonials, scraps, and messages, and (iv) community content which belongs to not a user but a community within Orkut. Table 4 displays the popularity of these categories based on the fraction of associated requests and the average time spent on each state. In total, browsing accounted for 92.7% of all requests! Compared to other non-browsing activities in the same category, browsing typically engaged 2–100 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.

Admittedly, these activities are not independent but are interrelated. Certain updates from friends can lead to interaction and search will be followed by browsing activity. Therefore, it is very important to understand the relationships among these activities, which are studied in the next subsection.

5.2. 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.

Fig. 5 shows the transition probability between all pairs of activities. A dark 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 3. We 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.

Table 4
Browsing activity in Orkut.

Browsing activity	Request (%)	Time spent (min)
Browsing media	30.8	11.2
Browsing profile	39.1	15.4
Browsing text messages	19.8	9.6
Browsing community	3	13.3

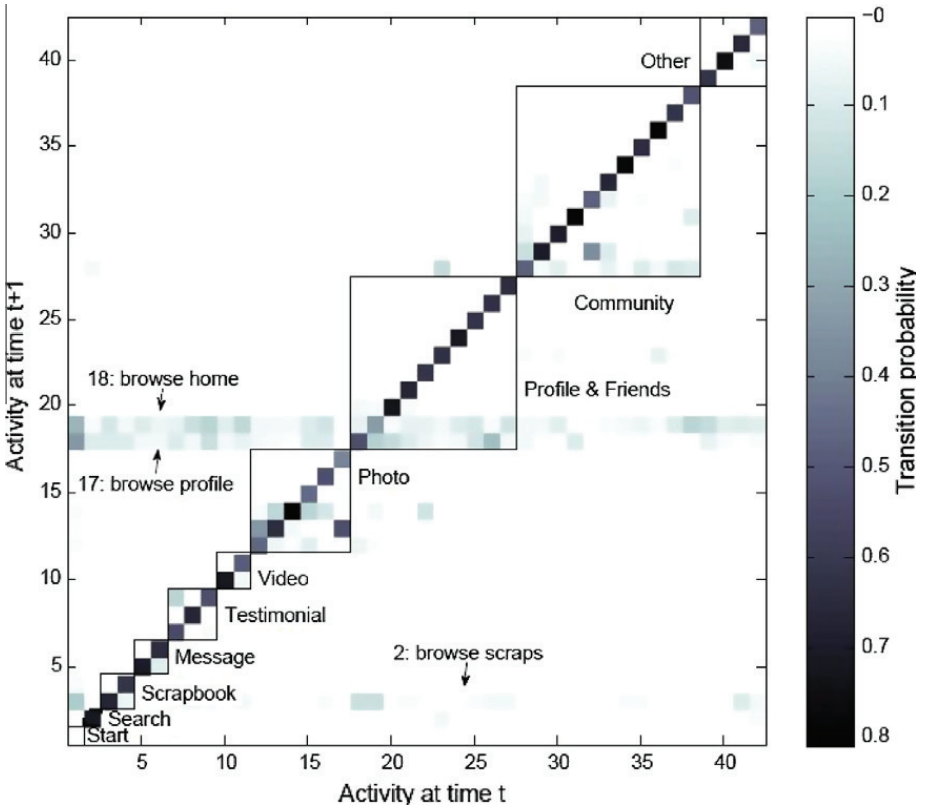


Fig. 5. Transition probability among activities in the clickstream model for Orkut.

When users log into 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 Fig. 5. 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 was repeated.

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.

5.3. Transition from one category to another

Finally, we examined the sequence of user activities at the level of categories (Fig. 6). 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 0.04% 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 Fig. 5, 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 Community 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 scraps events (activity

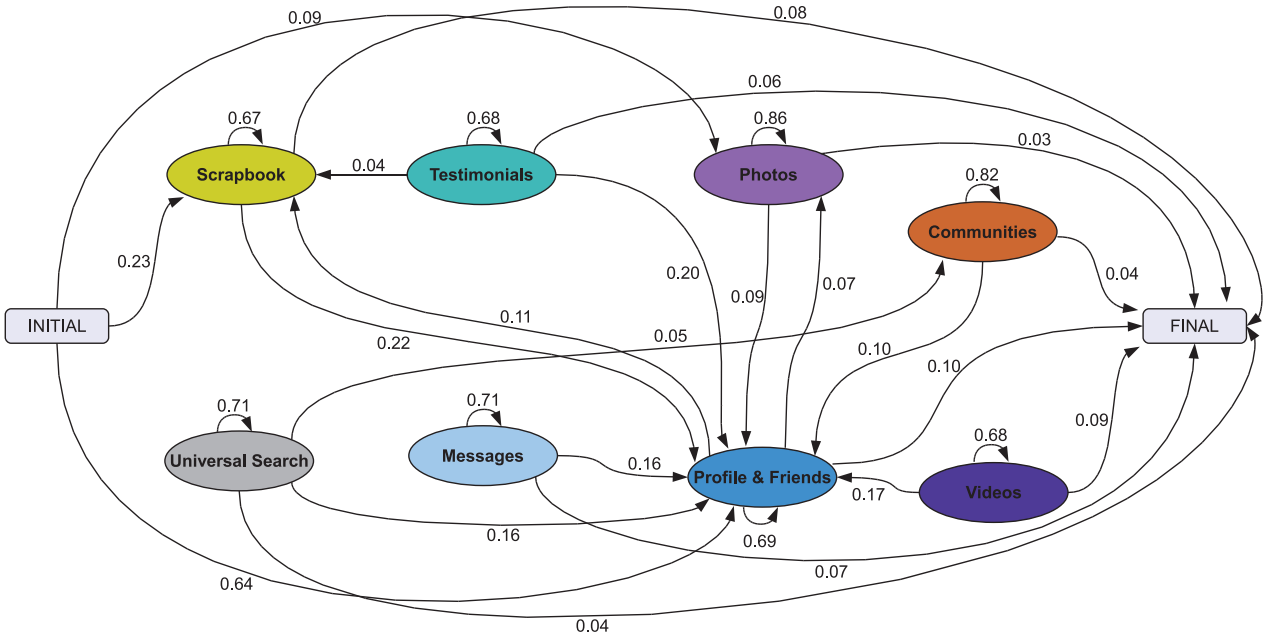


Fig. 6. Transition probability among categories in the clickstream model for Orkut.

Table 5
Differences between transition probabilities over different log periods.

	Entire log (12 days)	Days 1–4	Days 5–8	Days 9–12
Search → Search	0.71	−0.02	−0.01	+0.03
Scrapbook → Scrapbook	0.67	0.00	−0.01	0.00
Messages → Messages	0.71	−0.03	0.00	+0.02
Testimonials → Testimonials	0.68	+0.01	−0.02	0.00
Videos → Videos	0.68	0.00	+0.01	−0.02
Photos → Photos	0.86	0.00	0.00	+0.01
Profile → Profile	0.69	0.00	0.00	+0.01
Communities → Communities	0.82	+0.01	0.00	0.00
Initial → Profile	0.64	0.00	0.00	+0.01
Initial → Scrapbook	0.23	+0.01	0.00	0.00
Search → Profile	0.16	+0.01	0.00	−0.01
Messages → Profile	0.16	0.00	0.00	−0.01
Testimonials → Profile	0.20	0.00	+0.01	0.00
Videos → Profile	0.17	−0.01	0.00	+0.01
Profile → Scrapbook	0.11	−0.01	0.00	0.00
Scrapbook → Profile	0.22	−0.01	0.00	0.00

2). Except for self loops, Profile & Friends was the most common preceding state for most activities. Similar observations were made in [46] for other social networks such as Facebook.

In order to verify how effective our clickstream data are at capturing the large scale statistical behavior, we conduct the following analysis. We divided our 12-day log into three parts, each one containing four consecutive days of the original log. Then, we computed the first order Markov probabilities for each category of activities.

Table 5 displays the original transition probabilities as well as the difference between the probabilities of each part and the original probabilities. For instance, the probability for the transition Search → Search is 0.71 for the 12-day log, but it is 0.69 for the first 4 days of logs. Then, we present in Table 5 the difference between the two probabilities, i.e., 0.69 − 0.71 = −0.02. We can see that these probabilities computed for different periods of the log does not differ considerably from the results we presented considering the entire log. Table 5 displays the transition probabilities of only those edges with probability higher than 0.10. Overall, the transition probability remains rather stable across different time periods; the highest probability difference is 0.03. Such small variability across the days suggests that the clickstream dataset can effectively capture large-scale behavior of users in a social network.

5.4. Finding typical sessions

So far we have examined the overall navigation patterns of users across all sessions. In this section, we investigate *typical* or representative sessions in OSN sites. In order to find typical sessions, we used a clustering algorithm and grouped sessions of similar characteristics.

For clustering, we first defined each session as a vector of activities. Then the clustering algorithm calculated similarity between two sessions as the distance between two vectors. In more detail, a vector was used to represent each session, where each position in the vector contained the probability of a user navigating from one activity category to another (see Table 3 for the types of categories). In other words, each session was represented by a vector containing the probabilities of all arcs in Fig. 6.

We used *X*-means [40] clustering algorithm, which extends the popular *K*-means [29] algorithm. A key advantage of *X*-means over *K*-means is that the algorithm not only provides the clusters, but also estimates the best possible number of clusters. Therefore, we do not have to decide a priori the number of typical sessions. *X*-means algorithm finds clusters by minimizing the sum of the squared distances between each vector and the cluster's centroid, a vector that represents the averaged properties of each group. The distance between two vectors is computed by the Euclidean distance as follows:

$$D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{2}$$

where n is the size of any vector and x and y are the two vectors.

We used the implementation of *X*-means available on the tool Weka [55] and set the maximum number of groups to 10. The *X*-means algorithm indicated that four distinct groups was the best choice to fit our dataset, indicating that there are four typical groups of sessions in Orkut.

Table 6 displays the fraction of the sessions that were clustered in each cluster. The table also shows the mean session durations of each cluster. The first-order Markov representations of two of the typical sessions are represented in Fig. 7. For clarity, only predominant arcs and states are represented in the figures. Arcs with probability smaller than 3% are omitted.

The first typical session in Fig. 7(a) accounted for 65% of all sessions in the dataset. In this session, a user typically begins by engaging in the Profiles & Friends activity and then leaves the system or engages in another activities such as Scrapbook and Photos. The mean session duration of this cluster (22 min) is higher than those of other clusters. This is because many sessions in this cluster involved Photo activities. However, we also see very short sessions that lasted less than one minute. The short duration of such sessions typically involved a user leaving the Orkut page immediately after checking Profiles & Friends page. Nearly 40% of the sessions in this cluster had such light usage pattern.

The second typical session in Fig. 7(b) accounted for 25% of all sessions. In this session, a user typically begins by checking the Scrapbook. Note that users can access various features of Orkut from the social network aggregator. Therefore, it is possible to start an Orkut session from Scrapbook. In 24% of the sessions in this cluster, a user only visited Scrapbook before

Table 6
The frequency and the mean duration of typical sessions in Orkut.

	Request (%)	Session duration (mean) (min)
Cluster 1	65.4	22
Cluster 2	25.2	2
Cluster 3	5.2	10
Cluster 4	4.2	2

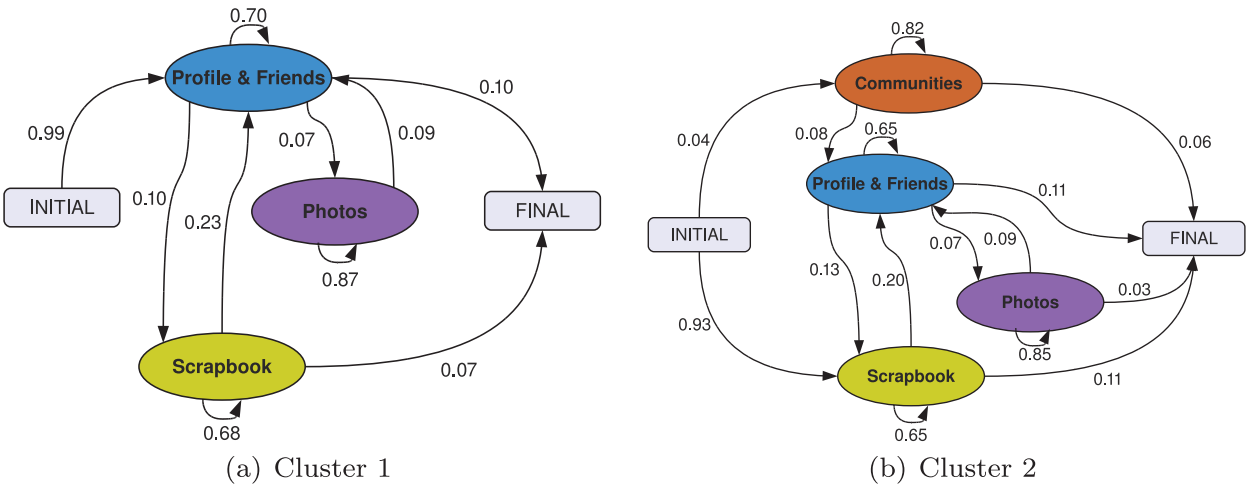


Fig. 7. Markov representation of the two most popular types of sessions in Orkut.

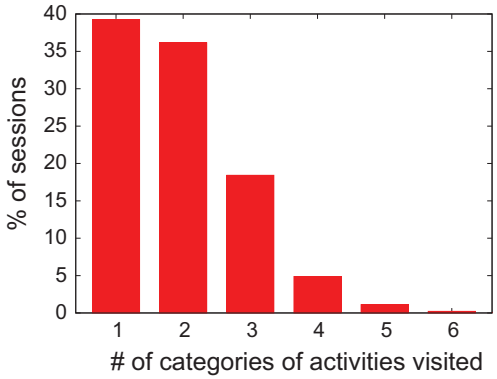


Fig. 8. The number of different types of activities users conduct in a session.

leaving the Orkut webpage. This suggests that, for a non negligible fraction of time, users just check for new messages on their scrapbook and leave the system.

Interestingly, we see a recurrent pattern of bidirectional links between Profile & Friends and Scrapbook in both of the clusters in Fig. 7. These two categories are related for the following two reasons. First, upon receiving a scrap from any other user, a link to that user's profile appears in Scrapbook. Therefore, it is natural to expect navigation from scrapbook to Profile & Friends. Second, in order to access other user's scrapbook and start a new thread of conversation, it is necessary to access that user's profile page first. This again explains the high transition probabilities between the two categories.

Having found that a large fraction of user sessions only involve one type of activity, we next investigate the number of different types of activities users conduct in an arbitrary session. Fig. 8 shows an histogram of the number of categories a user engaged in per session. Overall, users engage in very few types of activities: 75% of user sessions involved at most two types of activity categories. In 39% of the sessions, users just perform one type of activity category. This observation implies that an arbitrary session in OSN involves very few types of activity categories.

5.5. 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 min, (b) 1–10 min, (c) 10–20 min, and (d) 20 min 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.

Fig. 9 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 min), users spent 90% of their time on these activities. However, even for a long session (i.e., 20 min 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

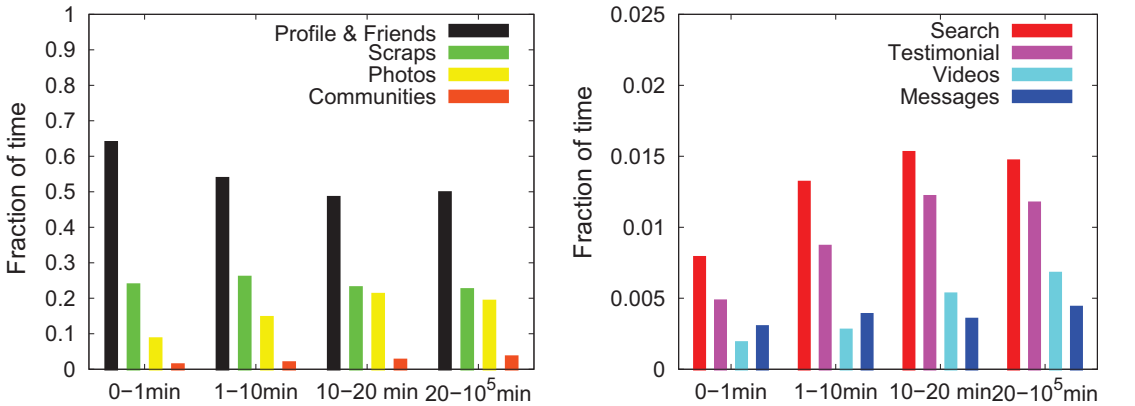


Fig. 9. Probability of the user activity as a function of session duration.

by a factor of 2 when comparing sessions shorter than 1 min to those longer than 20 min. The probability of seeing Community activity also increased with the session duration.

5.6. Comparison of user activity across OSNs

To get perspective on how user behaviors vary across different social networks, we repeated the analysis shown in Table 3 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 7 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. 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 the services provided by both OSNs 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.

Finally, Fig. 10 depicts the first-order Markov chain of user activity in the category level for Hi5, which is the second largest OSN in our dataset. Similar to what we have observed for Orkut in Fig. 8, self-loops are prevalent. In Hi5, Profile & Friends also plays a central role in connecting the other categories of activities. We expect to see similar usage trends for other social networks that possess similar service characteristics to Orkut and Hi5.

Table 7
Comparison of popular user activities across four OSN sites.

Rank	Orkut		MySpace		LinkedIn		Hi5	
	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

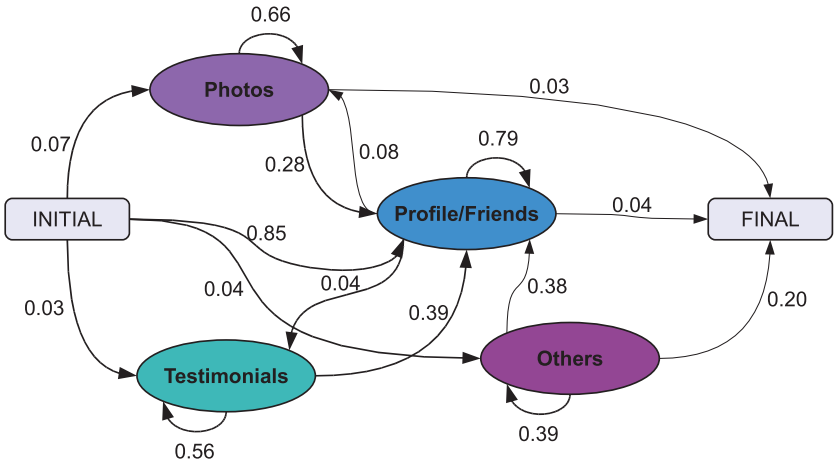


Fig. 10. Transition probability among categories in the clickstream model for Hi5.

6. 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, considering the social network distance as well as the physical distance.

6.1. Overview

Understanding social interactions have 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 [54,19,52,28,10]. Although these initial studies have identified several important properties of social interaction, there are behaviors of users that cannot be obtained from 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 understanding the social behaviors of users.

This section provides a complete view of user interactions in social networks, by considering both visible interactions and silent interactions. Particularly, we are interested in three sets of analysis: (a) We would like to know what fraction of user interactions is silent, compared to visible. If we consider browsing a friend's profile or photos 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? (c) In the physical world we have local and distant friends. Here we want to know if users mostly interact with friends located in a close geographical region.

6.2. Interaction over social network distance

Marlow et al. [37] defined as *passive engagement* the relationships in which one keeps up with friends only by reading feeds about them, without any form of explicity communication. Similarly, we considered explicitly visiting another user's page to be a 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), 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 3.2, we gathered information about the list of friends for all users in the workload trace by collecting the public data on Orkut website.

6.2.1. Webpage access patterns

To investigate the patterns of interactions 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 interactions 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 3. We excluded all activities related to search, communities, and others in Table 3.

Fig. 11 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 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,

[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.

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

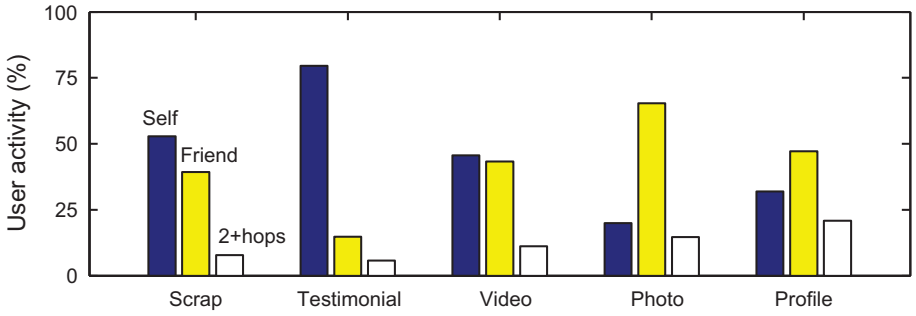


Fig. 11. Webpage accesses along the social network distance.

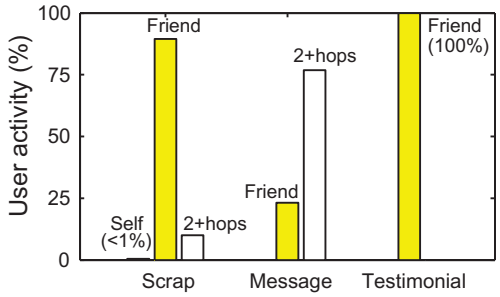


Fig. 12. Interaction in writing.

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, Fig. 11 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.

Fig. 12 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 was 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.

6.2.2. 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 8 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 8 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 3). The remaining accesses occurred when the user was navigating on other user's webpages; accesses to a 1-hop 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 3), 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 5.1, 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.

Table 8
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.

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

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 3). This reinforces the previous findings that users in social networks find new content and contacts through their 1-hop friends [42,16]. 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 8. 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.

6.2.3. How interaction patterns affect content popularity

Previously, we noted that most part of the requests registered in our Orkut data correspond to users accessing their own content or accessing 1-hop friend's content. Intuitively, there is a crucial difference between publishing content on the traditional web and sharing content over OSNs. When people publish content on the Web, they typically do so to make the content accessible to Internet users everywhere. On the contrary, when users publish content on OSNs they often have an intended audience, namely, their friends. Sometimes the audience is explicitly defined by the user or the site's policy. For example, content in Orkut's photo and video sharing service is by default accessible to only the immediate friends of the content uploader. At other times, the audience is implicitly limited by the inherent nature of the content. For example, a picture of a user and her friends at a birthday party is likely to be of interest only to people close to the user in the social network.

Next, we analyze the characteristics of the content popularity as an attempt to quantify how those interaction patterns affect content popularity within the Orkut. In Orkut, 13,095 users accessed photos during a 12-day period. We used the number of requests per photo as a measure of content popularity. For comparison, we used a publicly available web workload from the 1998 World Cup web site [6] over a period of 92 days. We consider the number of requests to each web page within the World Cup site as a measure of content popularity. We treated the World Cup dataset in the same way as above by picking a random 12-day period. We repeated our analysis over multiple 12-day samples and obtained consistent results.

We first compare the popularity distributions of social and web content in Fig. 13. The *x*-axis represents the content rank in percentile, where rank of 1% depicts the popularity of the content in position 1%. The *y*-axis represents percentage of all requests the given content instance received during a 12 day period. The noticeably different slopes of the plots indicates that popularity is much more skewed in web workloads than in social workloads.

To further examine the differences in the popularity distributions of social and non-social workloads, we use the disparity measure. Disparity is widely used in economics to measure the difference in household incomes. Typically, the 95th percentile and 5th percentile are compared. Table 9 displays the disparity quantities for the three distributions. The disparity between the 95th percentile and the 5th percentile is 4.0–7.0 in social workloads, while disparity is over 700 in the web. Even

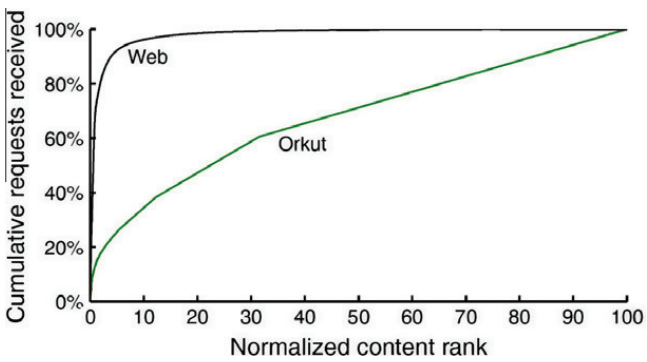


Fig. 13. Content popularity of Orkut photos in comparison with a traditional web server.

Table 9
Popularity disparity.

Ratio	Orkut	Web
99th/1th	10.0	10225.0
95th/5th	4.0	715.0
90th/10th	3.0	185.0

when we compare the 90th percentile against the 10th percentile and the 95th percentile against the median (50th percentile), we see two orders of magnitude higher disparity for non-social content compared to Orkut content. This clearly illustrates that content popularity is less skewed in Orkut than in web workloads.

Our finding that Orkut content is not dominated by few hits raises important questions about the effectiveness of traditional content distribution infrastructures. This is because the current content delivery infrastructures rely on caching a relatively low number of popular pieces of content that dominate the workload. However, the absence of these extremely popular objects in social workloads might call for a reexamination of infrastructure being used to deliver online social network content. Indeed, recent efforts showed that Facebook user requests can be processed 79% faster and use 91% less bandwidth [56].

6.3. Interaction over physical distance

So far we have investigated the user interaction patterns along the social network distance. Next, we study geographical aspects of social interactions. Social interactions among individuals located within a short physical proximity has been used to explain a number of phenomena in society, such as the proliferation of specific industries in a certain region [48] and individuals employment status [49]. Here, we are particularly interested in understanding how far (i.e., over what distance) social content is generated and consumed. Given that OSNs use the infrastructure originally designed for web workloads to deliver social content, understanding geographical aspects of social interactions can unveil potential opportunities for improvements on the current designing of OSN content delivery.

In order to investigate the physical proximity of users in OSN we need to know the location of users involved in the interaction. In our dataset, we can use the IP address to identify the location of users that sent requests to the social network aggregator. However, we cannot identify the users that are in the receiving part of the interaction (e.g., users that received a scrap, users that had a photo browsed, etc.). Thus, in order to obtain the location of these users we further gathered the Orkut public profile information of these users.

The location information available in user profiles is in free text form and often contained invalid location like "Mars". We used the Yahoo Maps Web service Geocoding API [1] to filter out invalid locations and infer user locations. In this way, we identified the location of 276,558 users, of which we know the location at the city level for 128,836 users and at the country level for 276,558 of the users. These users correspond to 42% of the users identified in our dataset. In total, the identified users were located in 4,297 different cities across 226 countries.

Fig. 14 shows how the probability of interaction varies as a function of the physical distance between two users. Physical distance between users is computed based on their longitude and latitude. We grouped distances in units of 10 km, so that 0 km means a distance between (0,10 km). The graph shows the probability for each distance d, which is the physical distance among all pairs u, v of users who interacted in our dataset. For comparison, we also show the probability of friendship over physical distance between users.

We can see that there is significant correlation between friendship, interactions, and physical proximity in the Orkut social network. Although Orkut works as a means to bridge the distance between friends, two users within a short distance (e.g., 10 km) have a high probability of interacting. Current OSNs infrastructure could exploit the physical proximity between

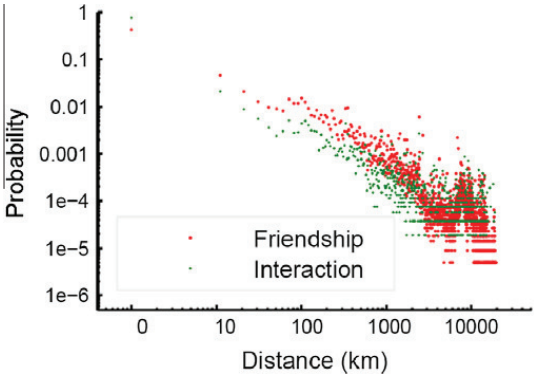


Fig. 14. Interaction across physical distance.

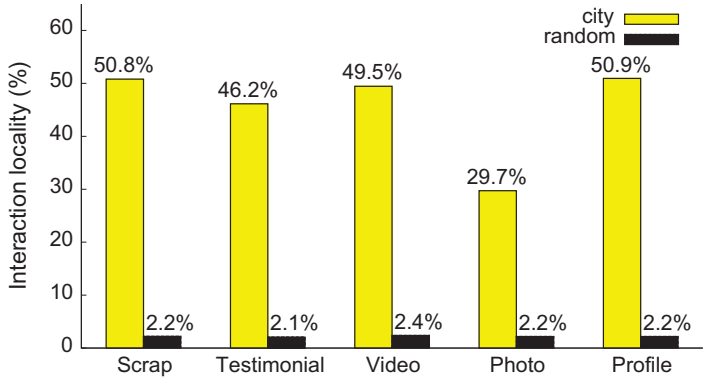


Fig. 15. Locality between content producers and consumers.

content producers and consumers. We also observe a strong correlation between the probability of interaction and the probability of forming friendship links. This is expected as users tend to interact more with their friends in the social graph. Although this could be a result of the OSN's interface design (e.g., updates from a user are pushed only to the user's friends), even when there are no such limitations, the very nature of social content (e.g., personal pictures) might make it appealing to a small set of people. For example, in Orkut users can access any other user's profiles by default. However, our data show that users mostly accessed profiles of their immediate 1-hop friends in the social network. This suggests that users in the social network tend to be geographically closer to each other when the interaction occurs mainly due to the presence of social links. In fact, it worth mentioning that Liben-Nowell et al. [36] also found a strong correlation between friendship and geographic location for LiveJournal users. Additionally, Rodrigues et al. [42] showed that URLs posted in Twitter tend to spread for short distances only on the first hops away from the content creator.

The observations that social content is primarily of interest to friends of the uploader and friends are usually located in a close physical distance could be exploited for caching design and Content Delivery Networks (CDNs). Thus, we next explore the characteristics of locality of content producers and content consumers. Fig. 15 shows the probability that content producers and consumers are located in the same city. In order to test if these results are affected by a biased sample of users, we randomly picked 500,000 pairs of users and computed the cities they are located. The results are shown as "random" in the figure. We can see that the fraction of interaction between users located in the same city is much smaller for the randomized set of pairs of users, for all the types of interaction analyzed. Consequently, in contrast to geographically-diverse content uploads, our findings indicate that social content is consumed *locally* in Orkut.

6.4. Number of friends interacted with

Finally we investigated how silent interactions affect the level of user interactions along the social network topology. We compute the number of friends (including multi-hop friends) a user interacts when we consider all activities performed by users, including silent interactions such as browsing, and we compared with the number of friends a user interacts when we consider only visible activities.

Fig. 16 shows these quantities as a function of the number of friends users have in the social graph. 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. [54], 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.

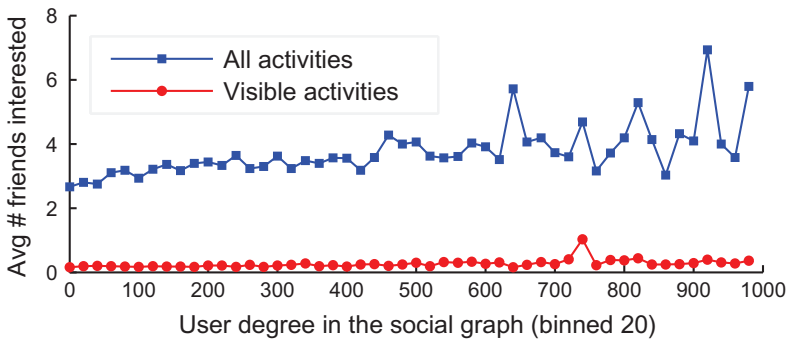


Fig. 16. Comparison of the Orkut social graph degree and interaction degree.

Interestingly, even for a short trace period, the degree of all interactions 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. Similar observations were recently provided by Jiang et al. [30]. The authors obtained data from a Chinese social network containing detailed statistics of profile visits on the network. They also showed that silent interactions (namely as latent interactions in their work) are much more prevalent and frequent than visible events in the analyzed Chinese system.

Finally, another trend that we observe in Fig. 16 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 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. Thus, if one were to measure the strength of social ties based only on visible traces, such analysis would be biased because 85% (=47/55) of the users would be completely disregarded.

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, compared to only considering visible interactions. Furthermore, we showed that if one were to measure the strength of social ties based on visible traces, 85% of the users would be disregarded.

7. Discussion

Our measurement analysis provides many interesting findings that are useful in various ways. We discuss implications of the findings below.

7.1. Modeling OSN sessions

In Section 4, we characterized the properties of user sessions in OSN workloads. Among various findings, we obtained models for several session features like inter-session times, session lengths, and inter-request times. As example, this is useful to show that 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 [27].

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. Such distribution functions can be used to generate synthetic (parameterizable) traces, that mimic actual OSN workloads. The statistics summarized in the paper and the session modeling are valuable in evaluating OSN services and testing potential system design alternatives. Additionally, the models describing typical types of sessions in Section 5.4 are useful for researchers interested in simulating user navigation in a given system.

7.2. Understanding user activity in OSNs

In Section 5, 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. These analyses demonstrated that browsing, which cannot be identified from visible data, is the most dominant behavior (92%). Given the large fraction of silent activities in Orkut, our findings highlight the potential bias in studies of user interactions that use only data containing visible activities.

Understanding user navigation is important for OSN service providers and portals [54,12] as well as for advertising agencies [53]. Frequently repeated activities (e.g., browsing home, browsing scraps) naturally serve as good targets for advertisements and can be exploited to improve the website design. Furthermore, the fact that users frequently check updates from their social contacts without involving any other action naturally makes OSNs a great place for advertisement.

Another 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 or even future activities of a user in a session. 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.

7.3. Interaction over the social graph

In Section 6 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 [39,5,54], 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 evidence of word-of-mouth-based information propagation.

Especially when it comes to rich media content like videos and photos, more than 80% of content was found through a 1-hop friend (Fig. 11). This finding confirms 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*) [42,16]. As OSN traffic is expected to grow rapidly, the patterns of social interaction and information flow can be valuable in designing the next-generation Internet infrastructure and content distribution systems [41,33]. 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 inferences about the geographical regions to which particular piece of content will likely spread.

7.4. Interaction over the physical distance

In Section 6 we also studied the physical distance between users as well as the distance between content producers and consumers. Our results suggest that users in a social network tend to be geographically closer to each other when the interaction occurs mainly due to the presence of social links. Additionally, we showed that social content is mostly consumed *locally* in Orkut.

These results call for a reexamination of today's content distribution infrastructures, which does not exploit the physical proximity between content producers and consumers in OSNs. Today, OSNs are using infrastructures originally designed for web workloads to deliver social content [41,56]. Considering that content in online social networks is typically produced by geographically-diverse users but consumed locally, one could allow users to upload content to a local server in the corresponding geographical area, like a city. Such mechanism could significantly reduce the amount of wide-area bandwidth needed compared to an upload to a centralized, remote server. The local server can then handle requests coming from users in the same geographical area and content needs to leave the local area only when a remote user requests it. However, this is expected to happen infrequently. In particular, high locality in content access at the city level in our dataset indicates that placing a server in every city can reduce the amount of cross-city traffic. The exact benefit depends on the size of a city, with more bandwidth savings attained in larger urban areas.

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.

This 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 investigate the impact of friends on the behavior of user of social networks. 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 distribute it to another sites.

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

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web Grant No. 573871/2008-6). Meeyoung Cha was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0012988). Finally, we are grateful to the social network aggregator for providing us with the clickstream data.

References

[1] Yahoo! Maps Web Services-Geocoding API. <http://developer.yahoo.com/maps/rest/V1/geocode.html>, 2009 (accessed September 2010).
[2] Google OpenSocial. <http://code.google.com/apis/opensocial/>, 2010 (accessed in March 2010).
[3] Orkut help. <http://www.google.com/support/Orkut/>, 2010 (accessed in March 2010).
[4] Orkut on Wikipedia. <http://en.wikipedia.org/wiki/Orkut>, 2010 (accessed March 2010).
[5] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Analysis of topological characteristics of huge online social networking services, in: World Wide Web Conference (WWW), 2007, pp. 835–844.
[6] M. Arlitt, T. Jin, Workload Characterization of the 1998 World Cup Web Site.
[7] S. Bausch, M. McGiboney, Nielsen Online Report - Social Networks & Blogs Now 4th Most Popular Online Activity. <http://tinyurl.com/cfzjlt>, 2009 (accessed March 2010).
[8] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on Twitter, in: Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010.
[9] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, M. Gonalves, Detecting spammers and content promoters in online video social networks, in: International ACM Conference on Research and Development in Information Retrieval (SIGIR), 2009, pp. 620–627.
[10] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, K. Ross, Video interactions in online video social networks, ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP) 5 (4) (2009) 1–25.
[11] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, Characterizing user behavior in online social networks, in: ACM SIGCOMM Internet Measurement Conference (IMC), 2009, pp. 49–62.
[12] M. Burke, C. Marlow, T. Lento, Feed me: motivating newcomer contribution in social network sites, in: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), 2009, pp. 945–954.
[13] M. Burke, C. Marlow, T. Lento, Social network activity and social well-being, in: International Conference on Human Factors in Computing Systems (CHI), 2010, pp. 1909–1912.
[14] J. Caverlee, L. Liu, S. Webb, The socialtrust framework for trusted social information management: architecture and algorithms, Elsevier Information Sciences 180 (2010) 95–112.
[15] J. Caverlee, S. Webb, A large-scale study of MySpace: observations and implications for online social networks, in: AAAI Conference on Weblogs and Social Media (ICWSM), 2008.
[16] M. Cha, A. Mislove, K. Gummadi, A measurement-driven analysis of information propagation in the Flickr social network, in: World Wide Web Conference (WWW), 2009, pp. 721–730.
[17] C. Chapman, M. Lahav, International ethnographic observation of social networking sites, in: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), 2008, pp. 3123–3128.
[18] P. Chatterjee, D.L. Hoffman, T.P. Novak, Modeling the clickstream: implications for web-based advertising efforts, Marketing Science 22 (4) (2003) 520–541.
[19] H. Chun, H. Kwak, Y. Eom, Y.-Y. Ahn, S. Moon, H. Jeong, Comparison of online social relations in volume vs interaction: a case study of Cyworld, in: ACM SIGCOMM Internet Measurement Conference (IMC), 2008, pp. 57–70.
[20] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, J. Almeida, Traffic characteristics and communication patterns in blogosphere, in: Conference on Weblogs and Social Media (ICWSM), 2007.
[21] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, B.Y. Zhao, Detecting and characterizing social spam campaigns, in: ACM International Conference on Internet Measurement (IMC), Melbourne, Australia, 2010, pp. 35–47.
[22] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: ACM SIGCHI Conference on Human factors in Computing Systems (CHI), 2009, pp. 211–220.
[23] S. Golder, D. Wilkinson, B. Huberman, Rhythms of social interaction: messaging within a massive online network, in: Conference on Communities and Technologies (ICCT), 2007.
[24] J. Gomide, A. Veloso, W. Meira Jr., V. Almeida, F. Benevenuto, F. Ferraz, M. Teixeira, Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, in: ACM Web Science Conference (WebSci), 2011, pp. 1–8.
[25] C. Grier, K. Thomas, V. Paxson, M. Zhang, @spam: The underground on 140 characters or less, in: ACM International Conference on Computer and Communications Security (CCS), 2010, pp. 27–37.
[26] L. Guo, E. Tan, S. Chen, X. Zhang, Y.E. Zhao, Analyzing patterns of user content generation in online social networks, in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2009, pp. 369–378.
[27] B. Huberman, P. Pirolli, J. Pitkow, R. Lukose, Strong regularities in world wide web surfing, Science 280 (5360) (1998).
[28] B. Huberman, D. Romero, F. Wu, Social networks that matter: Twitter under the microscope, First Monday 14 (1) (2009).
[29] A. Jain, M. Murty, P. Flynn, Data clustering: a review, ACM Computing Surveys 31 (3) (1999) 264–323.
[30] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, B. Zhao, Understanding latent interactions in online social networks, in: ACM SIGCOMM Conference on Internet Measurement (IMC), 2010, pp. 369–382.
[31] A. Joinson, Looking at, looking up or keeping up with people? Motives and use of Facebook, in: ACM SIGCHI Conference on Human factors in Computing Systems (CHI), 2008, pp. 1027–1036.
[32] R. King, When your social sites need networking, BusinessWeek (2007). <http://tinyurl.com/o4myvu> (accessed March 2010).
[33] B. Krishnamurthy, A measure of online social networks, in: Conference on Communication Systems and Networks (COMSNETS), 2009.
[34] J. Leskovec, L.A. Adamic, B.A. Huberman, The dynamics of viral marketing, ACM Transactions on the Web (TWEB) 1 (1) (2007) 228–237.
[35] Y.-M. Li, C.-Y. Lai, C.-W. Chen, Discovering influencers for marketing in the blogosphere, Elsevier Information Sciences 181 (23) (2011) 5143–5157.
[36] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social network, in: Proceedings of the National Academy of Sciences (PNAS), vol. 102, 2005, pp. 11623–11628.
[37] C. Marlow, Maintained Relationships on Facebook. <http://overstated.net/2009/03/09/maintained-relationships-on-facebook>, 2009 (accessed August 2010).
[38] MaxMind, GeoIP Database. <http://www.maxmind.com/app/ip-location>, 2010 (accessed March 2010).
[39] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in: ACM SIGCOMM Conference on Internet Measurement (IMC), 2007, pp. 29–42.
[40] D. Pelleg, A. Moore, *X*-means: extending *K*-means with efficient estimation of the number of clusters, in: International Conference on Machine Learning (ICML), 2000.
[41] J. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, P. Rodriguez, The little engine(s) that could: scaling online social networks, in: ACM SIGCOMM Conference, 2010, pp. 375–386.
[42] T. Rodrigues, F. Benevenuto, M. Cha, K.P. Gummadi, V. Almeida, On word-of-mouth based discovery of the web, in: ACM SIGCOMM Internet Measurement Conference (IMC), 2011, pp. 381–393.
[43] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: International World Wide Web Conference (WWW), 2010, pp. 851–860.
[44] S. Scellato, Beyond the social web: the geo-social revolution, SIGWEB Newsletter (2011) 5:1–5:5.
[45] S. Scellato, C. Mascolo, M. Musolesi, J. Crowcroft, Track Globally, Deliver locally: improving content delivery networks by tracking geographic social cascades, in: International World Wide Web Conference (WWW), 2011, pp. 457–466.

[46] F. Schneider, A. Feldmann, B. Krishnamurthy, W. Willinger, Understanding online social network usage from a network perspective, in: ACM SIGCOMM Internet Measurement Conference (IMC), 2009, pp. 35–48.
[47] S. Schroeder, 20 Ways to aggregate your social networking profiles, Mashable (2007). <http://tinyurl.com/2ceus4> (accessed March 2010).
[48] O. Sorenson, Social networks and industrial geography, Journal of Evolutionary Economics 13 (5) (2003) 513–527.
[49] G. Topa, Social interactions, local spillovers and unemployment, Review of Economic Studies 68 (2) (2001) 261–295.
[50] M. Valafar, R. Rejaie, W. Willinger, Beyond friendship graphs: a study of user interactions in Flickr, in: ACM SIGCOMM Workshop on Online Social Networks (WOSN), 2009, pp. 25–30.
[51] M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, V. Almeida, Tips, dones and to-dos: uncovering user profiles in Foursquare, in: ACM International Conference of Web Search and Data Mining (WSDM), 2011.
[52] B. Viswanath, A. Mislove, M. Cha, K.P. Gummadi, On the evolution of user interaction in Facebook, in: ACM SIGCOMM Workshop on Online Social Networks (WOSN), 2009, pp. 37–42.
[53] B. Williamson, Social Network Marketing: Ad Spending and Usage. EMarketer Report. <http://tinyurl.com/2449xx>, 2007 (accessed March 2010).
[54] C. Wilson, B. Boe, A. Sala, K.P.N. Puttaswamy, B.Y. Zhao, User interactions in social networks and their implications, in: ACM European Professional Society on Computer Systems (EuroSys), 2009, pp. 205–218.
[55] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.
[56] M. Wittie, V. Pejovic, L. Deek, K. Almeroth, B. Zhao, Exploiting locality of interest in online social networks, in: ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT), 2010, pp. 1–12.