Noname manuscript No.
(will be inserted by the editor)

Workload Characterization of a Location-Based Social Network

Theo Lins · Adriano C. M. Pereira · Fabrício Benevenuto

Received: date / Accepted: date

Abstract Recently, there has been a large popularization of location-based social networks, such as FourSquare and Apontador, in which users can share their current locations, upload tips and make comments about places. Part of this popularity is due to facility access to the Internet through mobile devices with GPS. Despite the various efforts towards understanding characteristics of these systems, little is known about the access pattern of users in these systems. Providers of this kind of services need to deal with different challenges that could benefit of such understanding, such as content storage, performance and scalability of servers, personalization and service differentiation for users. This article aims at characterizing and modeling the patterns of requests that reach a server of a location-based social network. To do that, we use a dataset obtained from Apontador, a Brazilian system with characteristics similar to FourSquare and Gowalla, where users share information about their locations and can navigate on existent system locations. As results, we identified models that describe unique characteristics of the user sessions on this kind of system, patterns in which requests arrive on the server as well as the access profile of users in the system.

Keywords Workload Characterization · Location-Based Social Networks · Web 2.0

1 Introduction

Since its beginning, the Internet has received a large diversity of applications, including the Web and Peer-to-Peer networks, in which the different traffic patterns helped reshaping its infrastructure. Recently, online social networking applications have emerged as extremely popular applications. According to Alexa.com, social networks like Facebook and Twitter are among the top 10 most visited websites in the world, both in terms of unique users and time spent on these websites. With

Federal University of Minas Gerais (UFMG)
Computer Science Department (DCC)
Belo Horizonte, MG, Brazil
E-mail: {theo, adrianoc, fabricio}@dcc.ufmg.br

750 million users, if Facebook was a country, it would be the third most populous country in the world [20].

Several online social networks allow some features in common. Generally, they allow users to share information with friends and have a page with the user profile that can post or update any content. Content varies from simple text messages to multimedia files, such as photos or videos. In order to encourage users to share content, social networks make updates available to users immediately after their friends share the content. Thus, not only users spend much time in these systems, but they also create huge amounts of content. As an example, the photo-sharing service on Facebook is the largest repository of photos on the Web, containing more than 60 billion images [19]. And YouTube receives 24 hours of video per minute [21].

In particular, there is one special type of social network system namely Location Based Social Networks (LBSNs), which are attracting new users in exponential rates. LBSNs, like Foursquare, allow users to share their geographic location with friends through smartphones equipped with GPS, search for interesting places, as well as posting tips about existing locations. It has been reported that, nowadays, nearly one in five smartphone owners access this kind of service via their mobile devices [15].

Intuitively, there is one crucial difference between traditional publishing of content on the Web and share content through social networks and location based social networks. When people share content on the Web, they typically make the content accessible to anyone. When users share content in typical online social networks they often intend to reach a certain audience, like friends or followers. Finally, when users share content in LBSNs they often intend to reach a local or regional audience, which might include or not friends. Thus, LSBNs provides a new social structure made up of individuals connected by the interdependency derived from their locations in the physical world as well as their location-tagged media content, such as photos, video, and texts.

This crucial difference might affect important properties of the traffic that reaches LBSN systems, which in turn might impact different aspects of system design, such as caching and content distribution. More important, given the exponential growth of several social system, it is reasonable to consider that these systems have the power to reshape the Internet traffic in the Future. In fact, social networking has been a major topic of discussion nowadays, known as the **Future Internet**, a movement that aims at formulating and evaluating alternative architectures for the changes that the Internet might need in the Future [23]. Despite considerable interest, little is known about patterns of access on new social network systems like LBSNs and how they differ from the access patterns of traditional systems.

This work aims at giving the first step on this direction, by providing a wide workload characterization of a location based social network very popular in Brazil, namely Apontador[1]. Apontador includes the main features of systems like Foursquare and Gowalla. It allows users to search for places, register new locations, post tips to existing places, and check-in in locations using mobile devices, such as smartphones and tablets.

[1] www.apontador.com.br

Through a large data set obtained from Apontador, we present an in-depth workload characterization of requests and sessions that reaches this LBSN server. We obtained a clickstream dataset, which described session-level summaries of over 64,309,252 million HTTP requests extracted during one month.

Using this dataset, we provide a series of analyses that provide a definition for user session in this context and model traffic and session patterns of the workload. Particularly, we examined how frequently people connect to LBSNs, for how long and how users interact across locations. Then we provide best fit models for a number of request and session level measures, such as session inter-arrival times and session length distributions and also characterize user navigation within sessions. Our study provides many interesting findings, including:

- A typical user session of a location based social network system lasts 30 minutes, a 3 times value in comparison with traditional Web systems.
- The popularity distributions of accesses to locations follow Lognormal distribution. In comparison other systems follows a Zipf distribution.
- The rankings of user activity in terms of the number of requests sent and sessions created follow a Weibull distribution, and power-law distributions.
- The arrival request rate at the system presents a periodic pattern with higher intensity during the day and smaller intensity during the night.
- The distributions of inter-request time and inter-session time can be modeled by Weibull and Gamma distributions.

The remainder of this work is organized as follows. The next section describes related efforts. Section 3 shows some statistics about the workload of the Apontador LBSN, as well as the dataset we crawled. Section 4 presents a wide workload characterization of our dataset, describing the main findings of this investigation. Finally, Section 5 concludes the paper and present directions for future work.

2 Related Work

Workload characterization is important for the understanding and improving web systems. There are several studies that present workload characterizations of different systems. Particularly, a seminal study of Web servers was presented in [2]. In this work, the data used were extracted from logs of the 1998's World cup Web page, where most of the accesses were directed to a small set of static files, making caching strategies quite efficient. Still related to web servers, [4] propose a methodology to identify invariants on the data, which are useful to derive models able to represent the workload of web services. Based on the obtained models, the authors also propose improvements on issues related to server performance and cache design. Another important effort from that period [6] used the models extracted from World cup characterization to create a realistic workload generation tool, which mimics a set of real users accessing a server.

Some years later, there were several approaches that attempt to characterize electronic commerce services [28,33,3], which identified models and patterns of arrival of requests and user sessions, useful to determined the impact on the performance and scalability of the systems. These efforts also highlight the importance of caching to ensure scalability of large e-commerce systems. In the context of video services on demand, [16] and [41] presented analyses of the object popularity and user activity, highlighting the workload differences of these systems to common static Web servers. [27] showed an automated approach for constructing synthetic workloads of session-based systems. The authors conducted an experimental study that investigates the impact of workload and various features that influence the performance of systems based on sessions.

More recently, [11] presented the characterization of an online travel and booking Web site, and identified that the workload exhibits different properties between day and night traffic in terms of request type mix. They also showed that the user session length covers a wide range of durations, besides observing that the response time grows proportionally to server load. In [44], they study three search query traces collected from real world web search engines in three different search service providers. To study the traces, they build a timing model to measure the rate variation of queries and sessions, a semantic model to measure the frequency of queries and terms, a locality model to measure the temporal locality of queries. [26] showed that their method helps to understand group-level workload characteristics, which can provide more accurate predictions on workload changes in a cloud computing environment.

When it comes to the first social systems, [17] presented a wide characterization of the access patterns in blogs, highlighting differences on the use of these systems in terms of the interactions between users and objects in comparison with the traditional Web content. [8] provided the first workload of a video-sharing service from the point of view of the web server, presenting a characterization of sessions and user navigation profiles. The results provide a better understanding of the access pattern of users to video sharing systems and identify different profiles, useful for service differentiation.

Among the many contributions of the above efforts, we highlight the creation of valuable models able to describe the workload that arrives on different servers, essential for generating synthetic workload that, in turn, enables experimentation and simulation based on realistic distributions. In our work, we present a characterization of a novel workload from the point of view of the server, providing models that describe how requests reach a typical location-based social network and also covering aspects that are uniquely related to location-based social networks.

In the context of social networks, [9] used data from clicks of Orkut users to characterize the forms of navigation and user interaction on these systems. Similarly, [37] presented a study of users navigation in Facebook. In a more recent study, [10] measured the physical distance and topological interactions between users of Orkut, showing that the content of these systems is in most produced and consumed locally. [18] provides characterization of a Twitter data set, out of which they create a framework for generating synthetic activities writing twitter. [24] characterized YouTube user sessions and compared the results with the traditional sessions and it has been identified that YouTube users transfer more data and takes more on each session than the Traditional Web work. These differences have implications for capacity planning of the system.

Moreover, there are several related to characterization of LBSN. [35] analyzed data crawled from three LBSNs, or Bright-kite, Foursquare, Gowalla. They found strong heterogeneity among users with different geographic scales in terms of interaction through social ties, with the high likelihood of a social tie between users within close geographical distances. Recent efforts [31] analyzed the dynamics of check-ins, showing spatio temporal patterns on user mobility in urban spaces. More recently, [39, 38] characterized how users interact with each other using tips, by collecting their Foursquare profiles. Tips are hints on a particular sites and can be marked as dones if a user agrees with its contents. Additionally, [30] used a spectral clustering algorithm to group users based on patterns of check-ins. Based on the attributes of regions and two users metropolitan cities, they were able to identify groups of users who visit similar categories of posts and characterize the type of activity that happens in each region of the city. [13] studied Gowalla, Brightkite and cell phone data, reporting that a long distance trip is more influenced by social friendship, while movements with short distances are not influenced by social networks. Finally, Rodrigues [34] studies how users access URLs through online social networks and show that most of the cascades spread in well defined geographical locations. More recently, Scellato *et. al* propose an approach to improve Content Delivery Networks (CDNs) by keeping track of the location of social cascades [36].

Unlike these efforts, our work aims at characterizing and understanding how requests arrive to a server of a location-based social networks, a type of system that has not been investigated from this perspective.

3 Dataset Description

This section presents the different datasets used throughout this article. Much of the datasets that will be described have been used in previous efforts. Therefore, only the most important features for the scope of work will be discussed.

3.1 Data from Apontador

In our study, we analyze the workload of the website Apontador[2]. Apontador is a Brazilian location based social network with georeferenced dataset containing approximately seven million locations. Each location has a page in the site where information is presented, such as: name, address, latitude, longitude, category and phone location. The users accessing this information can do so anonymously or recorded (logged). In addition to search and view the information in these sites, users can also recommend, evaluate, insert photos and register new locations. However, to create a new location, or evaluate an existing associate a photo to the site, users must be logged into the site. The same locations on the site are also available in applications for mobile platforms iPhone, Android or BlackBerry. In these applications, a registered user can check-in in places as well as take pictures and associate them to places.

The logs used correspond to the period of one month, from 01 to 31 of October, 2011. Table 1 shows that have been accounted for a total of 64,309,252 requests, 51,914,221 coming from different users. Each record of workload is a request sent by a user to Apontador. The following information is available for every request timestamp, user, object, type, and location. The *timestamp* field is the time in which the request was received by the server. The *user* field corresponds a cookie identifier of the user who generated the request. The *object* is the unique code

[2] http://www.apontador.com.br

Description	Distinct	Requests
Logged users	38,053	603,696
Not logged users	51,876,168	63,705,556
All users	51,914,221	64,309,252
Location accessed	**2,679,533**	**27,499,263**

Table 1 Statistics about the Apontador data set

to identify the request. The field *type* represent the actions that a person can be perform in one location. The *local* field is the location requested in the request made by the user.

Group Name	Number of Requests	Percentage
Visit	53,623,387	83,3800 %
Phone	9,225,458	14.3400 %
Site	1,160,655	1.8000 %
Thumbs up	242,937	0.3700 %
Thumbs down	49,604	0.0770 %
Send photo	3,941	0.0060 %
Focus email	669	0.0010 %
Facebook	655	0.0010 %
Email	630	0.0009 %
Focus phone	547	0.0008 %
Orkut	343	0.0005 %
Wigdet	235	0.0003 %
Focus copoun	125	0.0001 %
Twitter	66	0.0001 %

Table 2 Types of actions

As Table 2, there are several actions that a person can execute in one place. Examples of actions include: accessing the page of a location (visit), click on the phone of the location (phone) [3] clicking on the button “recommend” the location (thumbs up), click on the button “do not recommend” location (thumbs down), click on the button to go to the site location (site); upload a photo related to the site (send photo), share the location in Facebook (facebook), shares the location in Orkut (Orkut), share the location in Twitter (twitter), click the e-mail provided on the location (email), and, when the person requests the widget with the location's map (widget). Besides these actions described, there are other actions that are monitored when the site is sponsored, which are: when the person requesting the printing of a promotional coupon (focus coupon), when one visualizes the phone's location (focus phone); and when the person is viewing the e-mail site (focus email).

3.2 Crawling Locations

The data obtained from the users clicks on Apontador contains only the identifier of the locations stored in the system. Thus, information such as address, geolocation

[3] Intentionally, the phone number of the site is partially hidden to force a person to do an extra click to view the number.

and category of the locations are not available in the logs of servers from the Apontador. However, from the identifier of the site is possible to collect such information through the API of the Apontador[4].

To collect this extra information, we developed a crawler in Python that retrieved information from all available locations in our data set of users clicks. Below we present the characteristics of the locations collected.

Description	Distinct	Percentage
Locations accessed	2,679,533	100
Locations Collected in XML successfully	2,672,353	99.8

Table 3 Locations Collected

Table 3 shows the characteristics of the collected data set. In the total, it was possible to retrieve information from 99.8% of distinct accessed locations. Each site in XML format has the following information: identification unique name, description, counter clicks, number of reviews, number recommendations, site category, address, phone number, latitude, longitude, website address and location information of the user who created the site. Through the address field, we list the most frequent states of the distinct locations accessed only in one month, according to Table 4.

The six most frequent states corresponds to 70.86% the total of sites, which three of them belong to the southeast and the other three to the south region of the country.

State	**Number of Distinct Location**	**Percentage**
São Paulo	796,181	29.79 %
Minas Gerais	279,772	10.47 %
Rio de Janeiro	251,029	9.39 %
Rio Grande do Sul	224,546	8.40 %
Paraná	195,554	7.32 %
Santa Catarina	146,524	5.48 %
Bahia	121,633	4.55 %
Pernambuco	88,383	3.31 %
Ceará	76,121	2.85 %
Goias	74,561	2.79 %
Espirito Santos	53,533	2.00 %
Mato Grosso	41,134	1.54 %
Distrito Federal	40,255	1.51 %
Mato Grosso do Sul	39,138	1.47 %
Pará	34,820	1.30 %
Rio Grande do Norte	32,976	1.23 %
Paraíba	29,901	1.12 %
Others	146,292	5.48 %

Table 4 States with highest number of accesses

The category field identifies what type of establishment or service offered by the site. Table 5 shows the most frequent categories of distinct locations accessed within one month.

[4] http://api.apontador.com.br/

Category	Number of Distincts Location	Percentage
Addresses Corporate	254,468	9.52 %
Cars and Vehicles	82,677	3.09 %
Confections and Clothing	77,130	2.89 %
Construction	67,927	2.54 %
Beauty	54,168	2.03 %
Home Decoration	53,703	2.01 %
Medicine and Health	52,579	1.97 %
Banks and Financial Institutions	44,900	1.68 %
Food	44,251	1.66 %
Associations and Unions	43,663	1.63 %
Gas Stations	43,483	1.63 %
Restaurants	41,931	1.57 %

Table 5 Popular Categories

Table 6 shows the categories of the 10 Locations with the highest number of user sessions in a month.

Category	#Sessions
General Services	5,660
Laboratories	5,283
Consulates and Embassies	4,684
Food	3,782
Postal	3,688
Confections and Clothing	3,427
Transport	3,403
Public Schools	3,146
Transport	3,009
Transport	2,979

Table 6 Categories of 10 Locations with most Sessions

3.3 Other Systems

Next, we describe other systems we use to compare properties with the patterns we uncover from Apontador's data. Table 7 shows the summary of these systems.

System/Dataset	Object	Type	Reference
Web Site World Cup 1998	Images and Html content	Visits	[1]
Youtube	Videos	Views	[12]
Orkut	Photo and Profile	Clicks	[9]
Uol Mais	Videos and Html content	Views	[8]
LBSN Apontador	**Location**	**Visits**	

Table 7 Summary information and references for the Datasets used

3.3.1 Web server of the 1998's World Cup

Ideally, we would like to compare data obtained from existing social networking with Web 1.0 data, consisting mostly of servers containing static pages where Web users were mere spectators. A data set that meets these requirements and is publicly available. We will use anonymised data public Web server World Cup 1998 [2], which had averaged 11,000 visits per minute and 40 MB of data transferred per minute to the users.

In particular, we use the log of 30 days (from May 24 to June 24 1998), containing 69,747 unique objects and 681,469,425 registered requests for these objects.

Table 8 shows that almost all requests from users (98%) were for HTML content or an image file. A typical feature observed in workloads of Web servers.

Type	**% of requests**
Images	88.16
HTML	9.85
Java	0.82
Compressed	0.08
Audio	0.02
Video	0.00
Dynamic	0.02
Others	1.05

Table 8 Distribution by File Type - World Cup 1998

3.3.2 Orkut

We used data from Orkut collected and characterized in a previous work [9]. These data were collected from a aggregator of social networking and has the record of all objects accessed from different social networks by 36,309 users who used the system during the monitored period. To carry out our analysis, we used only access from the photos of Orkut in order to measure the popularity of photos shared in this system. In total this dataset contains 23,764 photos in our logs, accessed 121,939 times.

3.3.3 YouTube

Among the current social systems, a major traffic is associated to the distribution of videos. In order to compare the popularity of videos to the popularity of other objects of Web 2.0 and Web 1.0 objects, we are going to use a dataset containing 1,666,226 YouTube videos collected in December 2006 [12]. To each video, this dataset contains the number of views of the videos, and that in total the videos received 369,762,000,000,000 base hits.

3.3.4 Uol Mais

Our base of YouTube videos contain only numbers on the popularity of videos. However, systems for sharing videos receive requests for other pictures representing the videos or requests for search and navigation systems. The Types of requests are presented in Table 9. To study the popularity of all objects accessed and not only the videos, we also are going to use a dataset from UOL Mais, a system of video sharing from UOL. A detailed description of this data base can be obtained by reference in [8]. The log used in this work was obtained the period from 12 December 2007 to 07 January 2008, has 109.239 objects and 3,613,935 requests of access to these objects.

Group Name	Request Type	#Request	Percentage
View	View a video	2,758,883	74.94 %
User	Video list of a user	218,335	5.93%
	Video list of a user with a certain tag	75,583	2.05%
Lists	List of top videos	55,307	1.50%
	List of related videos of a video	32,838	0.89 %
Interactions	Video evaluation	22,038	0.60%
	Video comment	14,131	0.38%
	Favorite video	10,774	0.29%
Search	Search	1,625	0.04%
	List of videos with a certain tag	421,700	11.46%
Others	Main page	2,679	0.07%
	Error requests or unformatted registry	67,339	1.82%

Table 9 Request Groups of UOL Mais

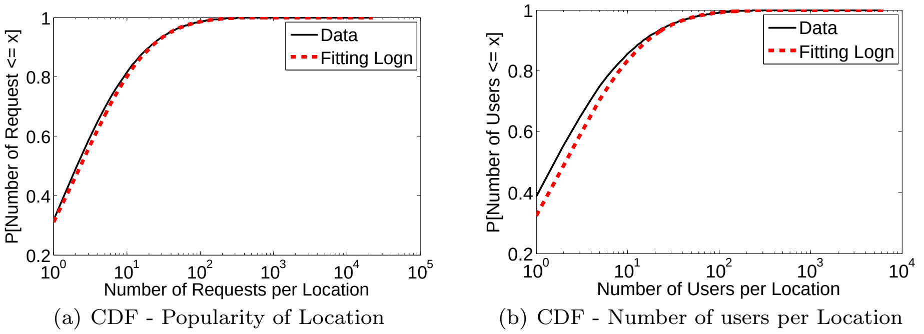
4 Characterization of Workload

In this session, we present a characterization of the workload of Apontador from different perspectives, showing various aspects and distributions. Moreover, we describe the main findings from these characterization analyses.

4.1 Popularity of Locations

First we evaluate the popularity of the locations (or places), with the objective of verify if it follows a specific distribution. To check the accuracy of the proposed models, we use the coefficient of determination, which is the proportion of variability in a dataset that can be explained by the statistical model. R^2 can take 0 and 1 as minimum and maximum, respectively. In all the models presented in the work, the values of R^2 are above 0.96. When the value of R^2 is equal to 1 means that there are no differences between the model and the actual workload.

Next we analyze if the distributions of popularity of Locations follow some distribution. Figure 1(a) shows the cumulative distribution function (CDF) of the number of requests for locations. We note that there is a small amount of places with a great amount of hits and a great quantity of locations with just a few. For example over 80% of the sites have up to 10 requests. This observation is important because it shows the great potential for local caching that the system has. In fact, this distribution is well modeled with a lognormal distribution with $\mu = 0.849$, $\sigma = 1.720$ and $R^2 = 0.996$, where the parameters denoted μ and σ are the mean and standard deviation of the variables in natural logarithm, respectively.



(a) CDF - Popularity of Location

(b) CDF - Number of users per Location

Fig. 1 Number of Requests and Users per Location

As the distribution of requests for location, Figure 1(b) shows cumulative distribution function (CDF), which follows a lognormal distribution, whereas show the number of unique users per location with $\mu = 0.741218$, $\sigma = 1.617501$ and $R^2 = 0.979$.

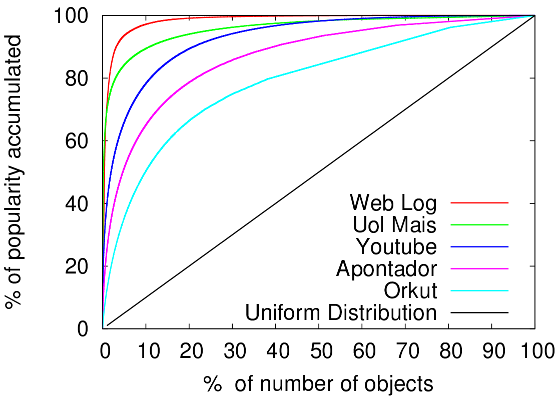


Fig. 2 Normalized accesses to objects in different Web systems

On the Web, the idea of having a large concentration of popularity in a few objects is the basis for the construction of hierarchical caching systems and has been widely applied in the systems of caches projects, in a very recent past [7, 5, 22, 42]. This occurs because the fact that the relations of friendships are more influential in the social networking difficult the objects to become very popular, as well as in places where the distance also influences.

Next we analyze the characteristics of the popularity of content in different systems as an attempt to quantify how patterns of interactions of social networks affect the popularity of content in these systems. Figure 2 shows these normalized distributions for different systems discussed in section 3. The x-axis represents the ranking of content in percentage, the ranking where 10% represents 10% of the first objects of each analyzed dataset. The y-axis represents the cumulative percentage of popularity, ie, for 10% of the first objects of the ranking, the y-axis shows what fraction of the accesses those 10% received. We can note the large difference in concentration of popularity that every curve presents and shows that the social curves are much more distributed compared with the concentration of the popularity of the data objects from the Web server of the 98's World Cup. As an example, while 10% of objects from the Web server of the World Cup concentrate 97.18% of accesses, 10% of objects from Orkut received only 50.33% of accesses.

In the other networks we can see that the concentration of popularity also is always smaller compared to the server of the World Cup. The UOL-mais, for being a video server also receives requests related to images (thumbnails) representing the videos or search requests and navigating the system, is the one with the curve that is the closest to the Web server of World Cup 98. On the data from YouTube, which account only the popularity of access to videos, we can notice a greater spread of access to objects. On Apontador, the objects analyzed are locations and the curve represents the popularity of access to different locations. We note that the concentration of popularity is even lower, reflecting the local interest by different objects in this kind of system. On Orkut, photos and their popularity are analyzed. The concentration of popularity proved the least, since Orkut's typically users only access pictures of their friends, which hinders the formation of objects very popular in the system.

With this analysis we can conclude that the system Apontador at level of popularity of objects is between videos sharing systems and orkut and possibly this is due to the physical proximity of the users accessing the Apontador's objects.

To examine more thoroughly the differences in popularity, we studied the datasets of various systems to measure the disparity between these measures. The measure of disparity is well known in economics to measure differences between rich and poor in a country. This measure is called *range ratios* and it is computed by dividing a value at one predetermined percentile by the value at a lower predetermined percentile. Typically, the 95th and the 5th percentiles are compared.

Table 10 shows the range ratios for different distributions. The range rate between the 95th and the 5th percentiles is 20 for Orkut and 45,831 for the server of world cup 98. Even when we compare the disparity of other distributions with Web distribution, we note that the range ratio on the Web workload is in a greater magnitude than that of social systems.

Ratio	*Web*	*UOL Mais*	*YouTube*	*Apontador*	*Orkut*
1º / 99º	703,959	334	15,410,5	128	46
5º / 95º	45,831	52	979.62	39	20
10º / 90º	15,119	24	214.61	21	12

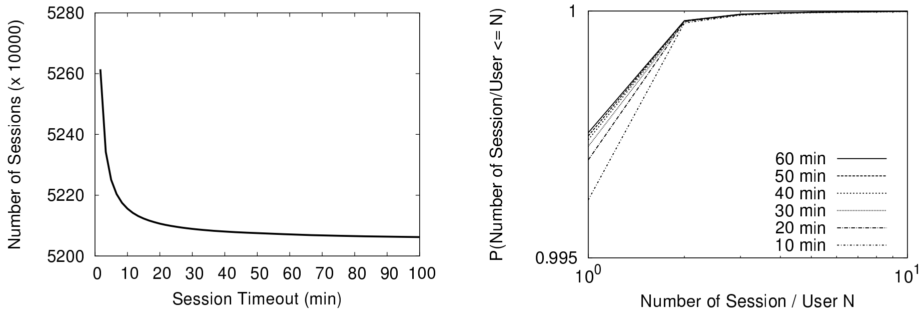
Table 10 Disparity in Object Popularity

Our observations that distributions of accesses to objects in social systems are far less concentrated than in data from a typical 1.0 web server that raise important questions about the effectiveness of the traditional infrastructure for content distribution today, and especially the case in the future growth expectations and even greater popularization of social systems is confirmed. This is because the current infrastructure is based on caching a small fraction of objects that dominate content. The lack of objects extremely popular in sequences on the Web requests suggests that it may be necessary to reexamine the infrastructure for distributing social content in the future. In fact, it is not surprising that recent studies have shown that the content of Facebook could be processed 79% faster and use 91% less in bandwidth [43].

4.2 Definition of sessions

A session of a user is defined as a series of requests made by the user to a site over a given period of time [29,1]. In environments of location search, a user session includes access to the site, access to the website, telephone access and all the actions mentioned in section 3. These types of requisitions are significantly different from the user sessions of conventional websites, which do not have the same degree of user interaction systems of Web 2.0.

The determination of the beginning and end of a session in search applications of locations requires a specific analysis of the time between requests in order to measure the inactivity of the user, since most of the sessions do not present an explicit registration of login and logout operations. Therefore, it is necessary to perform an analysis to identify a limit value of time between requests for them to be considered as being of the same session. Thus, two consecutive requests are considered in the same session if the time between them is smaller than this limit, called the session expiration time.



(a) Session Timeout

(b) CDF - Number of Session per User

Fig. 3 Definition of sessions

It is important to choose an adequate expiration time in order to not generate sessions that do not represent the use of the service by users, avoiding unite different times of use of the service or fragment a navigation performed by the user. Next we following the methodology proposed in [29] to define an appropriate session timeout period. It consists of conducting an empirical evaluation of how a chosen expiration time would affect the number of sessions in the workload.

Figure 3(a) shows the total number of sessions for different values expiration time. An extremely small value (e.g., 1 minute) results in an extremely high amount of session (over 50 million sessions), generating almost only sessions with one request. As the value of the expiration time increases, the number of sessions continuously reduces until this decrease stabilizes. This stability occurs around

30 minutes, indicating that this value is a limit suitable to be adopted as the expiration time of the session.

To test this hypothesis we generated a cumulative distribution function (CDF) of the number of sessions per user for various values of expiry time of the session, as shown in Figure 3(b). The difference between distributions for different values of expiration time is greater to smaller values, making it more consistent after 30 minutes. Therefore, we adopted 30 minutes as timeout of sessions to our analyzes, obtaining a total of 52,089,255 user sessions in our workload.

It is interesting to note in Table 11 that this result is similar to the analyzes achieved in the work [17], a little lower than in the workloads characterized in the references [8,25]. When compared with the results that characterize traditional Websites sessions [1,32], the value of expiration time of the session obtained here is 3 times larger than the 10 minutes typically observed.

This is due to the longer time the user spends to view places with your details and related services, which may lead users to stay longer in their navigation through the system.

System/Description	**Timeout (min)**	**Year Collection**
Web Site World Cup 1998 [1]	10	1998
E-Comerce [3]	15	2001
LBSN Apontador	**30**	**2011**
Blogosphere [17]	30	2006
Youtube, Video Sharing Systems [25]	40	2007
Uol Mais, Video Sharing Systems [8]	40	2008
Live Streaming Media [41]	60	2002
Twitter, Write Activity [14]	167	2009

Table 11 Session Timeout (minutes)

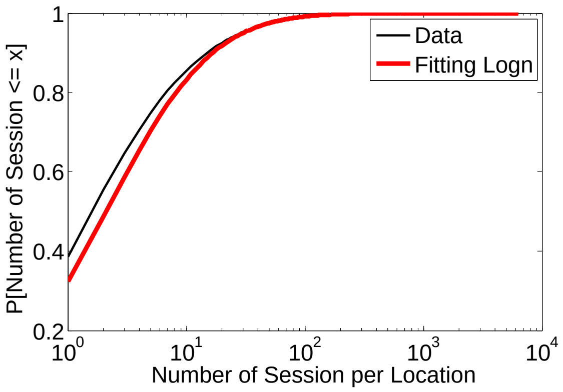


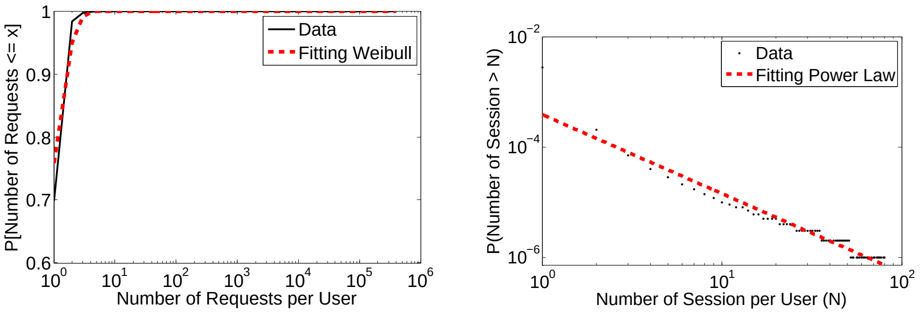
Fig. 4 CDF of the Number of Sessions per Location

Figure 4 shows a lognormal distribution for the number of sessions per Location with $\mu = -4.524$, $\sigma = 3.018$ and $R^2 = 0.979$.

4.3 Activity Level of Users

We also analyzed the level of user activity. We know that users can access the service of location search repeatedly within the same session or return to the system constantly, generating a large number of sessions. Thus, we characterize the ranking of users in terms of the number of requests sent and in terms of the number of sessions created in the system. We call user each anonymised IP address of the workload.

Figure 5(a) shows the cumulative distribution function (CDF) of number of requests sent to the server by users. We note that there is a small amount of users that make many requests to the server and a lot of users who make few requests. 69% of users have one request, and over 99% of users have up to 5 requests. We identify that a Weibull distribution fits well this behavior, with parameters $\alpha = 0.345$, $\beta = 2.683$ and $R^2 = 0.967$, where the parameters α and β are known as the shape and scale parameters, respectively.



(a) CDF of the Number of Requests per Users

(b) CCDF of the Number of Sessions per Users

Fig. 5 Activity Level of Users

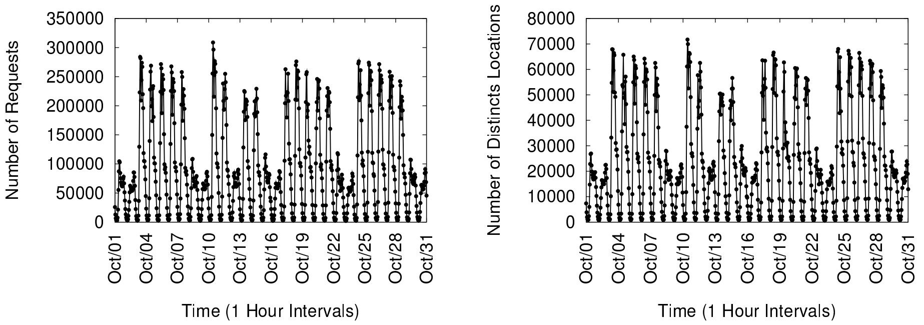
In terms of sessions created in the server, as shown in Figure 5(b), we obtain a function that follows the Power Law to model the distribution of the ranking of sessions with $\alpha = 0.0007$ and $R^2 = 0.984$. This result emphasizes the behavior that few users have many sessions as many have just a few sessions.

Compared to other work already done [8,41], in terms of the ranking of requests by users, the distributions follow Zipf. In the ranking of sessions per user, they follow a Zipf and Exponential distribution, respectively.

4.4 Temporal patterns of access

In this section we analyze the number of incoming requests to the server over time. Figure 6 shows the number of requests arriving at server at intervals of one hour. The curve has a periodic pattern, with a higher access intensity during the day and at lower intensity at night. We note that during the weekends and holidays, as for example, the holiday of October 12 access falls occur to the system. As discussed peaks that typically go 250,000 requests in 1 hour on weekdays, weekends and holidays are around 100,000 requests in one hour, a drop of over 50%. Considering only the locations, the requests follow the same pattern as can be seen in Figure 6(b).

These patterns are similar to those described in studies on traditional Web servers [40,4] as well as other types of servers as the weblog [17], sharing videos [8], electronic commerce [3], on demand videos [41]. Differing only for special occasions that may be a large increase in requests, such as [2], occurred an increase in demand in key games of the World Cup 1998, as well as special events that may affect the e-commerce sites with advertising campaigns, special promotions, or the approach of holidays like Valentine's Day, Easter, Mother's Day, Father's Day and Christmas.



(a) Requests in 1h intervals

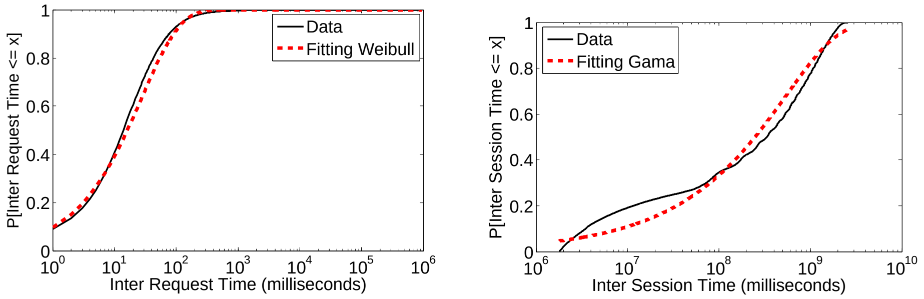
(b) Locations in 1h intervals

Fig. 6 Number of Requests and Locations in 1h Intervals

To analyze the participation of visiting users of the system, we characterize the time between arrivals of requests and sessions to the system. Here we show in Figures 7(a) and 7(b) cumulative probability (CDF) to the time intervals between requests and sessions, respectively. We note that the probability on the interval of time between requests that is greater than 500 milliseconds is less than 3%, and 78% of the requests arrive at the server at intervals smaller than 100 millisecond. Similarly, about 99% of the intervals between sessions are less than 1 second. And analyzing the interval between sessions we notice that the probability of being less than 1h is 20%.

The distribution of the interval between requests is best approximated by Weibull distribution, where $\alpha = 0.049$, $\beta = 0.710$ and $R2 = 0.983$. For the distribution of the time interval between sessions was used a Gamma distribution with $\alpha = 0.360$, $\beta = 1023168222$ and $R2 = 0.961$.

Comparing with other work, we show that the distribution of the interval between requests is similar to [3,17,6], which all follow a Weibull distribution. In contrast we show that when we compare the distribution of the interval between sessions we identify a Weibull distribution, [17] modeled it as an Exponential distribution and [9] follows a Lognormal distribution.



(a) CDF - Inter Request Time

(b) CDF - Inter Session Time

Fig. 7 Temporal Patterns of Access

4.5 Distance and Categories of Locations

In this section we investigate the distances between locations that appear consecutively on user requests. In order to measure the distance between each pair of locations in a session, we reconstruct the user sessions and consider only sessions with more than 2 requests to different locations. Then we compute the distance between each pair using the location data (latitude and longitude) of these locations. Figure 8 shows the cumulative distribution function (CDF) of the average distances between locations visited during the same user session, where we can see that 50% of the distances are below 18 kilometers.

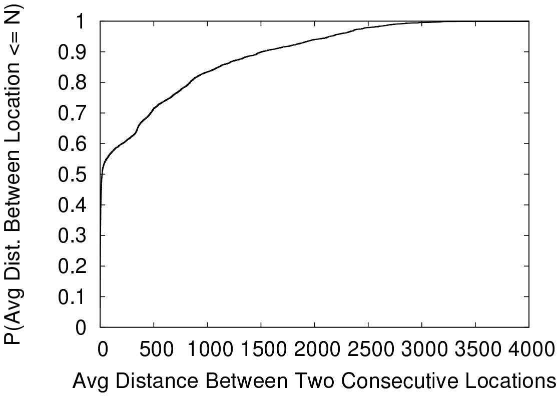


Fig. 8 CDF of the average distances between locations in a user session

The high distance values occur because the user is redirected to the Web *site* of Apontador through another search engine. For example, if he is looking for a site called "Minerais Itaguacu", the search engine returns several *links* of locations in the Web *site* of Apontador that corresponds to that name, but one is located in the State of São Paulo, another one in Rio de Janeiro, etc. And to identify the location that the user really wants, he would access all suggested locations.

Another reason for the some high distance values would be bots, which access the Web *site* of Apontador in order to collect data. As an example, there are some sessions that contain access to thousands of locations in different parts of the country, performed in few minutes. In fact, all location content shared by Apontador can be freely accessed by bots and indexed by search engines, as can be confirmed by rules established in the robots.txt of Apontador Web site[5].

We also compared the distances between locations of users requests, considering the category of the location, whether two location belong to the same category or not, as shown by Figure 9. We note that about 45% of locations have equal distances between categories, and 66% of them have distances below 15 kilometers between each other. For different categories, which corresponds to 55% of the total, 37% of the occurrences present distances up to 15 kilometers. Thus, we can observe that when the user is looking for a location with the same categories, the probability that these locations are very close to each other is higher than when looking for locations which belong to different categories. This kind of finding can be useful for a system to recommend a location to a user.

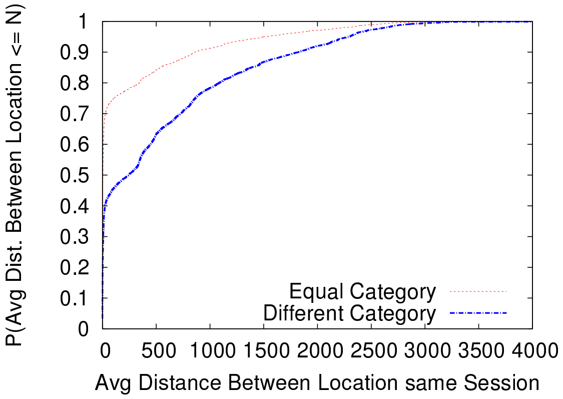


Fig. 9 CDF of the average distances between locations in each user session for equal and different categories

Moreover, we perform an analysis of the number of distinct categories per session, as showed in Figure 10. We analyzed only sessions with more than 2 requests, and we can observe that more than 99% of the sessions have only one distinct category. Furthermore we analyzed sessions that have more than one distinct category. The percentage of the predominant category in the session is presented by Figure 11, where 86% of the sessions have a category that is prevalent with up to 66% of requests in the user session.

Figure 12 shows the number of requests or occurrences for the average distance between locations. We can observe that for most sessions the number of requests is below 100, independently of the average distance. In addition, there are some user sessions, particularly with the average distance in the range 1000-1500 km, which have many requests, with values greater than 300. The fact that some sessions are

[5] www.apontador.com.br/robots.txt

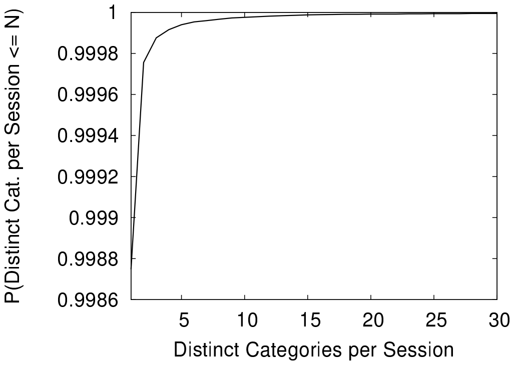


Fig. 10 CDF of distinct categories per user sessions

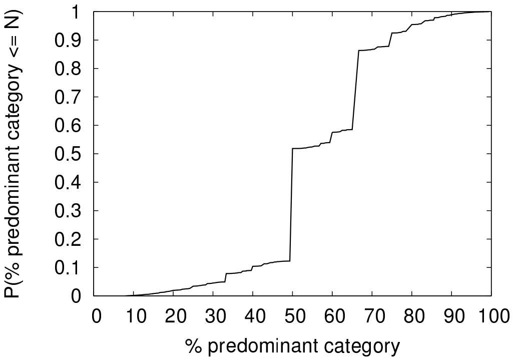


Fig. 11 CDF of the percentage of the predominant category in user sessions

very large indicates the existence of bots, which also explains some large values of distances between two visited locations.

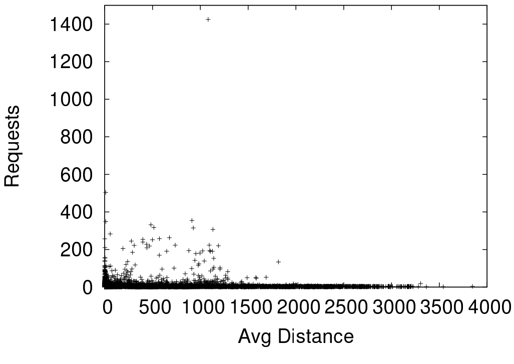


Fig. 12 Average distance between locations per requests

5 Conclusions

Since the launch of the first online social networks, these systems have continuously gaining popularity. Follow updates from friends is today one of the most popular Internet activities. This new paradigm of access to data on the Web is changing the way content is consumed on the Web.

Using data from different social networks, in this article, we investigate properties of access to the objects of these systems and discuss implications for the Future Internet.

Furthermore, in this work we provide an in-depth workload characterization of a LBSN, obtaining models that describe these data as well as uncovering user access patterns within sessions when they log into these systems. Particularly, we provided statistical models to represent object popularity and users activity level, time between arrival of requests and sessions, etc. The present study is innovative because it is the first to analyze a location-based social network from the point of view of the server.

The models presented are useful not only for generating synthetic workload but also for the design and creation of new infrastructure for this type of service. As future work we intend to create a synthetic workload generator for multiple online social networks, thus the models generated in this work will be useful in order to create the LBSN module of this workload generator.

Acknowledgments

This research has been supported by CAPES, CNPq, and Fapemig. The authors also would like to thank Apontador Inc. for the data provided.

References

1. M. Arlitt. Characterizing web user sessions. *SIGMETRICS Performance Evaluation Review*, 28(2):50–63, 2000.
2. M. Arlitt and T. Jin. Workload characterization of the 1998 world cup web site. In *Technical Report HPL-1999-35R1*, 1999.
3. M. Arlitt, D. Krishnamurthy, and J. Rolia. Characterizing the scalability of a large web-based shopping system. In *ACM Transactions on Internet Technology*, pages 44–69, 2001.
4. M. Arlitt and C. Williamson. Web server workload characterization: the search for invariants. *SIGMETRICS Performance Evaluation Review*, 24(1):126–137, 1996.
5. P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns: Characteristics and caching implications. In *Proc. of Int'l Conference on World Wide Web (WWW)*, pages 15–28, 1999.
6. P. Barford and M. Crovella. Generating representative web workloads for network and server performance evaluation. In *ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, volume 26, pages 151–160, 1998.
7. F. Benevenuto, F. Duarte, V. Almeida, and J. Almeida. Web Cache Replacement Policies: Properties, Limitations and Implications. In *Proc. of Latin American Web Congress (La-Web)*, 2005.
8. F. Benevenuto, A. Pereira, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Characterization and analysis of user profiles in online video sharing systems. *Journal of Information and Data Management*, 1(2):115–129, 2010.
9. F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *ACM SIGCOMM conference on Internet measurement conference (IMC)*, pages 49–62, 2009.

10. F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user navigation and interactions in online social networks. *Information Sciences*, 195(15):1–24, 2012.
11. D. Carrera, R. Gavalda, J. Torres, and E. Ayguade. Characterization of workload and resource consumption for an online travel and booking site. *Proc. of IEEE International Symposium on Workload Characterization (IISWC)*, pages 1–10, 2010.
12. M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM Internet Measurement Conference*, 2007.
13. E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1082–1090, 2011.
14. G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *Proc. of the ACM conference on Hypertext and social media (HT)*, pages 123–132, 2012.
15. Inc. comScore. Nearly 1 in 5 smartphone owners access check-in services via their mobile device. `http://bit.ly/mgaCIG`, 2011.
16. C. Costa, I. Cunha, A. Vieira, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto. Analyzing client interactivity in streaming media. In *World Wide Web Conference (WWW)*, pages 534–543, 2004.
17. F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *Proc. Int'l Conference on Weblogs and Social Media (ICWSM)*, 2007.
18. V. Erramillia, X. Yanga, and P. Rodriguez. Explore what-if scenarios with song: Social network write generator. `http://arxiv.org/abs/1102.0699`, 2012.
19. Needle in a Haystack: Efficient Storage of Billions of Photos, 2009. Facebook Engineering Notes, `http://tinyurl.com/cju2og`.
20. Key Facts, Facebook Newsroom, 2012. `http://newsroom.fb.com/Key-Facts`.
21. YouTube Fact Sheet. `http://www.youtube.com/t/fact_sheet`. Acessado em Dezembro/2012, 2011.
22. L. Fan, P. Cao, J. Almeida, and A. Broder. Summary Cache: a Scalable Wide-area Web Cache Sharing Protocol. *IEEE / ACM Transactions on Networking*, 8(3):281–293, 2000.
23. A. Gavras, A. Karila, S. Fdida, M. May, and M. Potts. Future internet research and experimentation: the fire initiative. *SIGCOMM Comput. Commun. Rev.*, 37:89–92, July 2007.
24. P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *ACM SIGCOMM conference on Internet measurement (IMC)*, 2007.
25. P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Characterizing user sessions on youtube. In *IEEE Multimedia Computing and Networking (MMCN)*, 2008.
26. A. Khan, X. Yan, T. Shu, and N. Anerousis. Workload characterization and prediction in the cloud: A multiple time series approach. *IEEE Network Operations and Management Symposium (NOMS)*, pages 1287–1294, 2012.
27. D. Krishnamurthy, J. Rolia, and S. Majumdar. A synthetic workload generation technique for stress testing session-based systems. In *IEEE Trabsactions on software engineering*, volume 32, pages 868–882, 2006.
28. D. Menascé and V. Almeida. *Scaling for E Business: Technologies, Models, Performance, and Capacity Planning.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.
29. D. Menascé, V. Almeida, R. Fonseca, and M. Mendes. A methodology for workload characterization of e-commerce sites. In *ACM Conf. on Electronic Commerce (EC)*, 1999.
30. A. Noulas, C. Mascolo S. Scellato, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *SMW 2011*, 2011.
31. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *International Conference on Weblogs and Social Media*, 2011.
32. Ad. Oke and R. Bunt. Hierarchical workload characterization for a busy web server. In *Int'l Conf. on Computer Performance Evaluation, Modelling Techniques and Tools (TOOLS)*, 2002.
33. A. Pereira, L. Silva, and W. Meira Jr. Evaluating the impact of reactive workloads on the performance of web applications. In *Proceedings of the 25th IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Phoenix, Arizona, USA, 2006. IEEE CS.

34. T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 381–393, 2011.
35. S. Scellato. Beyond the social web: the geo-social revolution. *SIGWEB Newsletter*, pages 5:1–5:5, September 2011.
36. Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon Crowcroft. Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades. In *Proc. of Int'l Conference on World Wide Web (WWW)*, pages 457–466, 2011.
37. F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 35–48, 2009.
38. M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Caracterizacão e influência do uso de tips e dones no foursquare. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, 2012.
39. M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, dones and to-dos: Uncovering user profiles in foursquare. In *ACM International Conference of Web Search and Data Mining (WSDM)*, February 2012.
40. E. Veloso, V. Almeida, W. Meira Jr., A. Bestavros, and S. Jin. A hierarchical characterization of a live streaming media workload. *IEEE/ACM Transactions on Network*, 14(1):133–146, 2006.
41. E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. A hierarchical characterization of a live streaming media workload. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment (IMW)*, pages 117–130, 2002.
42. J. Wang. A Survey of Web Caching Schemes for the Internet. *ACM Computer Communication Review*, 25(9):36–46, 1999.
43. M. Wittie, V. Pejovic, L. Deek, K. Almeroth, and B. Zhao. Exploiting locality of interest in online social networks. In *Proc. of ACM Int'l Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, pages 1–12, 2010.
44. H. Xi, J. Zhan, Z. Jia, X. Hong, L. Wang, L. Zhang, N. Sun, and Gang Lu. Characterization of real workloads of web search engines. *Proc. of IEEE International Symposium on Workload Characterization (IISWC)*, pages 15–25, 2011.