A

Workload Characterization: A Survey Revisited

MARIA CARLA CALZAROSSA and LUISA MASSARI, Università di Pavia
DANIELE TESSERA, Università Cattolica del Sacro Cuore

Workload characterization is a well established discipline that plays a key role in many performance engineering studies. The large scale social behavior inherent in the applications and services being deployed nowadays leads to rapid changes in workload intensity and characteristics and opens new challenging management and performance issues. A deep understanding of the user behavior and workload properties and patterns is therefore compelling. This paper presents a comprehensive survey of the state of the art of workload characterization by addressing its exploitation in some popular application domains. In particular, we focus on conventional Web workloads as well as on the workloads associated with social network, mobile and video services and cloud computing infrastructures. We discuss the peculiarities of these workloads and present the methodological approaches and modeling techniques applied for their characterization. The role of workload models in frameworks, such as performance evaluation, capacity planning, content distribution and resource provisioning, is also explored.

Categories and Subject Descriptors: C.4 [**Performance of Systems**]: Measurement techniques; C.4 [**Performance of Systems**]: Modeling techniques; D.4.8 [**Operating Systems**]: Performance—*Modeling and prediction*

General Terms: Experimentation, Measurement

Additional Key Words and Phrases: Workload characterization, user behavior, statistical techniques, graph analysis, performance evaluation, Web workload, online social networks, mobile services, video services, cloud computing

1. INTRODUCTION

The term workload refers to all inputs received by a given technological infrastructure. Understanding the properties and behavior of the workloads is essential for performance engineering studies dealing with the design and capacity planning of the infrastructures and the optimization of their cost. More generally, the evaluation of the Quality of Service (QoS) and of the Quality of Experience (QoE) perceived by the users requires a deep knowledge of the underlying workloads. Similarly, workload characterization is the basis for devising efficient resource provisioning, power management and energy conservation strategies, content distribution policies, security mechanisms, recommendation systems and marketing strategies.

Workload characterization is a well established discipline that has been extensively studied since early 1970's (see, e.g., [Ferrari 1972] and [Ferrari et al. 1983] Chapter 2). During all these years it has been continuously updated to cope with the numerous advances in the technological infrastructures and the ways users interact between each other and with the underlying infrastructures. In particular, the advent of Web 2.0 has opened the way to

Authors' addresses: M. Calzarossa and L. Massari, Dipartimento di Ingegneria Industriale e Informazione, Università di Pavia, via Ferrata 5 – I-27100 Pavia, Italy. E-mail:{mcc,massari}@unipv.it; D. Tessera, Dipartimento di Matematica e Fisica, Università Cattolica del Sacro Cuore, via Musei 41 – I-25121 Brescia, Italy. E-mail: daniele.tessera@unicatt.it.

DOI:http://dx.doi.org/10.1145/0000000.0000000

the deployment of a large variety of innovative services. These services are widely exploited thanks the increased pervasiveness and ubiquity of mobile devices equipped with geolocation capabilities and always connected to the Internet. Moreover, these services provide functionalities (e.g., communication, entertainment, social networking, information retrieval) characterized by different requirements in terms of performance and QoS. In these scenarios, users often play an active role exploited by a wide range of one-to-one and many-to-many relationships. In addition, the network effects inherent in the social nature of a large number of services can lead to sudden and unpredictable changes of their popularity, e.g., flash crowd phenomena. These effects in turn produce load surges, increase its burstiness and cause system performance degradation.

The models obtained as a result of the characterization process summarize and explain the main properties of the workloads, In addition they enable the generation of realistic synthetic workloads for performance evaluation studies of complex technological infrastructures. Workload models are also frequently exploited as input of analytic and simulation system models and for the definition of benchmarking experiments. When building workload models, particular care has to be put to their accuracy and representativeness, that is, their ability to capture and reproduce the most relevant characteristics of the workloads as well as how users behave.

In this paper we revisit and complement the surveys on workload characterization published some years ago (see [Calzarossa et al. 2000; Calzarossa and Serazzi 1993]) by focusing on five application domains selected according to their relevance and popularity. In particular, we address conventional Web workloads, as well as the workloads associated with emerging scenarios, i.e., online social networks, mobile devices, video services, cloud computing. For each domain we present the peculiarities and commonalities of their workloads and we discuss the state of the art in terms of methodological approaches and techniques applied for workload characterization. The major findings and their performance implications are also offered.

The paper is organized as follows. Section 2 presents a comprehensive overview of workload characterization methodologies by focusing on the measurement process and the analysis techniques employed for this purpose. Section 3 summarizes the main results of the characterization of Web workloads. The behavior of the users of online social networks is discussed in Sections 4, whereas the characterization of the usage patterns of the apps deployed on mobile devices is addressed in Section 5. The state of the art of video service workloads is covered in Section 6, while Section 7 presents a survey of workload characterization of cloud computing environments. Finally, some concluding remarks are given in Section 8.

2. WORKLOAD CHARACTERIZATION METHODOLOGY

Workload characterization relies on experimental approaches based on the analysis of measurements collected on the technological infrastructures while they are operating, that is, under their real workloads. As pointed out in [Calzarossa et al. 2000], the complexity of the infrastructures and of their workloads makes these approaches rather challenging. Hence, it is necessary to devise systematic and thorough methodologies.

2.1. Measurement techniques

Measurements provide qualitative and quantitative information about the individual workload components (i.e., the basic units of work being processed) as well as about the users. The quality of the measurements being collected is a necessary ingredient to ensure the accuracy of the workload models. This means that measurements have to specifically refer to the workload components of interest and capture their static and dynamic properties and the behavior of the users. Moreover, it is important to take into account the hierarchical nature typical of the workloads of the new emerging infrastructures and services (e.g., e-business, video streaming). Therefore, as outlined in [Calzarossa et al. 1988], measurements have to be collected at each individual layer and the mapping and transformations between layers have also to be considered.

Passive monitoring tools (e.g., network sniffers, profilers, accounting tools and logging facilities that are part of the operating systems and application services) are particularly suitable for obtaining workload measurements (see [Crovella and Krishnamurthy 2006] and [Feitelson 2015] Chapter 2). In particular, sniffers provide detailed measurements about the network traffic and allow for discovering the behavior of the users, e.g., clickstreams. In the framework of logging facilities, Web servers record in their access logs measurements for each HTTP transaction (e.g., the IP address of the client that issued the HTTP request, the time stamp of the transaction, the method and resource requested, the status code of the server response, the number of bytes transmitted by the server, the referrer of the page previously visited by the client, and the user agent used by the client to issue the request). Figure 1 shows three records of an access log collected by an Apache Web server. To cope with privacy issues, the IP addresses of the clients that sent the HTTP requests have been anonymized.

```
7517553923 - - [02/Jul/2015:14:22:59 +0200] "GET /robots.txt HTTP/1.1" 200 1032 "-"
"Mozilla/5.0 (iPhone; CPU iPhone OS 7_0 like Mac OS X) AppleWebKit/537.51.1 (KHTML,
like Gecko) Version/7.0 Mobile/11A465 Safari/9537.53 (compatible; bingbot/2.0;
http://www.bing.com/bingbot.htm)"
9126678821 - - [02/Jul/2015:14:23:39 +0200] "GET /news.html HTTP/1.1" 200 2488
"http://peg.unipv.it/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/43.0.2357.130 Safari/537.36"
3761105124 - - [02/Jul/2015:14:25:10 +0200] "GET /wp//wp-login.php HTTP/1.1" 404
292 "-" "Mozilla/5.0 (X11; U; Linux i686; pt-BR; rv:1.9.0.15) Gecko/2009102815
Ubuntu/9.04 (jaunty) Firefox/3.0.15"
```

Fig. 1: Example of an anonymized Apache access log.

Similarly, media server access logs store information about the individual requests of the clients (e.g., IP address, time stamp, advertized duration of the media file) and about the responses of the server (e.g., available bandwidth while the file was playing, number of bytes sent by the server, elapsed time of the requested file). Data center traces contain measurements related to capacity and load of the individual servers (e.g., number of cores, clock speed, allocated memory) and to the resources used by jobs and tasks (e.g., CPU and memory usage, disk IO time.) Of course, the choice of the parameters to be considered for the characterization depends on the objectives and on the nature and type of workload to be analyzed (e.g., batch/background vs interactive/real-time).

Whenever suitable datasets (e.g., anonymized access logs) are not publicly available, crawlers are often exploited to obtain some aggregated workload measurements [Olston and Najork 2010]. These software agents traverse the Web and take multiple snapshots of the websites by periodically downloading the entire sites, individual Web pages or their metadata. Workload measurements are then extracted by parsing these snapshots. For instance, in the case of online social networks, parsers extract information, such as user identifiers, time stamps, number and text of comments and hyperlinks, number of “likes” and location tags. In addition, to reduce the overhead due to these crawling activities, providers frequently offer APIs for querying some specific information about their services. An example of information collected by crawling an online social network is shown in

Figure 2.

userID	time	latitude	longitude	location id
60514	2015-06-24T13:44:46	45.452176	9.276308	376497
96422	2015-06-24T13:44:58	45.464098	9.191927	175991
16151	2015-06-24T13:45:06	45.485888	9.204283	376191

Fig. 2: Example of information collected by crawling an online social network.

Some monitoring tools allow for online (i.e., real-time) data collection and analysis, whereas others are intended for data collection and storage only, thus the data analysis has to be performed offline. In general, the amount of data being collected can become quite large and sometimes even intractable. This is the case, for example, of large-scale geographically distributed infrastructures being monitored under heavy load conditions. Similarly, log files can grow extremely large, thus producing significant overhead for their storage and management. Thereby, as a general rule, it is advisable to avoid excessive monitoring activities because of the additional load exercised on the technological infrastructures. Moreover, to reduce potential perturbations and overhead and overcome at the same time "big data" issues, appropriate sampling techniques have to be applied. As there might be the danger of ignoring events referring to rare workload components or specific users, it is very important to ensure the representativeness of the data sample being considered.

In next section, we present the state of the art of the techniques applied for offline, i.e., post mortem, analysis of the measurements.

2.2. Analysis techniques

Once empirical measurements have been collected and the parameters describing each workload component selected, it is necessary to undertake a systematic exploratory data analysis. This analysis highlights the properties and behavior of the workloads and their patterns. Hence, it has to be seen as the basis for building workload models.

The exploratory data analysis relies on the application of statistical and visualization techniques. More specifically, descriptive statistics and measures of dispersion (e.g., mean, range, variance, coefficient of variation, skewness, median, percentiles) are very useful to summarize the properties of each parameter. In addition, parametric statistics allow for exploring the strength of the relations between the parameters. For example, the Pearson's correlation coefficient ρ_{xy} provides a quantitative measure of the extent to which parameter x is positively or negatively related to parameter y, that is:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

where σ_{xy} denotes the covariance and σ_x and σ_y the standard deviations of the parameters x and y, respectively.

Furthermore, even though the number of parameters and workload components can be quite large, it is worth to inspect the data by means of visualization methods. Diagrams (e.g., histograms, scatter plots, box plots) work well for this purpose in that they ease the interpretation of the data. In particular, scatter plots highlight the correlations between parameters, whereas box plots summarize their distributions and are very useful for identifying the outliers. The term *outlier* denotes the workload components characterized by an atypical behavior of one or more parameters. It is critical to take the right approach towards outliers because of their potential effects on the workload models. Outliers could indicate phenomena or properties previously unknown, thus worth exploring. On the contrary, they

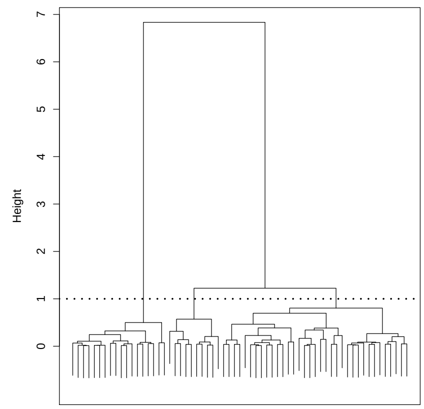

Fig. 3: Dendrogram displaying the results of the agglomeration process.

could correspond to anomalous operating conditions of the infrastructures or even errors in the measurements, thus worth discarding.

To derive models able to capture and summarize the overall properties of the workloads, a further step of the workload characterization methodology deals with the analysis of the components in the multi-dimensional space of their parameters. In this respect, multivariate analysis techniques and in particular clustering techniques have been extensively applied.

Clustering [Jain et al. 1999] is an unsupervised process that subdivides a set of observations (i.e., workload components) into homogeneous groups (i.e., clusters). The components of each group are very similar, whereas the components across groups are quite distinct. The centroids (i.e., the geometric centers of the clusters) are often used as representatives of the groups. Distance-based clustering techniques differ for the algorithms applied (e.g., hierarchical, iterative) and their similarity measures (e.g., Euclidean distance, Manhattan distance).

The k-means algorithm is a very popular non hierarchical clustering algorithm that iteratively partitions the data into k clusters by assigning each observation to the cluster C_j that minimizes the objective function, that is:

$$\sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - c_j||^2$$

where x_i and c_j denote the m coordinates of the i-th observation and the j-th cluster centroid, m being the number of parameters describing the observations, and $||.||$ refers to the Euclidean distance. As a result, the intra-cluster distance is minimized, while the inter-cluster distance is maximized.

In the case of hierarchical algorithms, cluster agglomeration relies on methods, such as single linkage, complete linkage and Ward method [Gan et al. 2007]. For the single and complete linkage methods the concept of distance between two clusters is defined as the minimum and maximum distance between their components, respectively. On the contrary, the Ward method is based on the analysis of the variance, that is, the total sum of squared deviations from the mean of a cluster. The visualization of the agglomeration process applied for building the clusters relies on a dendrogram, that is, a hierarchical tree. The dendrogram shown in Figure 3 displays the observations - whose labels have been omitted for legibility - and the sequence of clusters. The heights of the tree represent the distances between the clusters and are proportional to the intra-clusters variance. The branches corresponding to

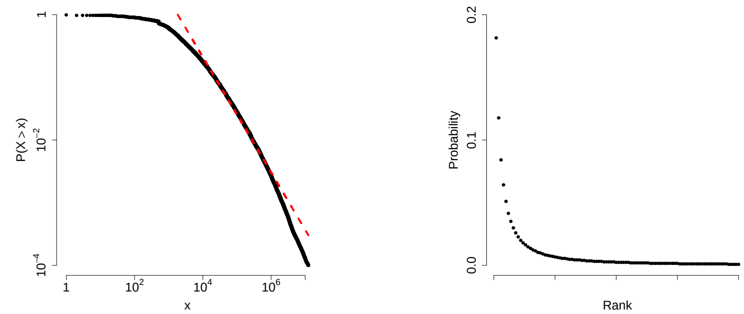

Fig. 4: Pareto ....

the clusters are obtained by pruning the tree at a given level. The horizontal dotted line drawn in the figure at height one corresponds to a pruning level. The result of this cut are three subtrees, each representing a cluster.

Since the overall number of observations and the number of their characterizing parameters are often quite large, to reduce the data dimensionality, Principal Component Analysis (PCA) is applied, sometimes even in conjunction with clustering [Johnson and Wichern 2007; Jolliffe 2002]. The PCA linearly transforms the potentially correlated parameters into a set of uncorrelated parameters, namely, the principal components. In general, few principal components are sufficient to explain and summarize most of the variability in the original data.

To study the dynamics of the workloads and model their temporal patterns, numerical fitting techniques are usually applied [Draper and Smith 1998]. More specifically, the estimation of the parameters of the function that best fits the empirical data is based on methods (e.g., least squares, Maximum Likelihood Estimation). The goodness of fit is evaluated by means of statistics (e.g., coefficient of determination R^2, Kolmogorov-Smirnov, Anderson-Darling and F tests).

Particular attention has to be paid for discerning the shapes of the empirical data and quantifying whether they are drawn from well known probabilistic distributions (e.g., exponential, Weibull, Gamma, lognormal, binomial). It has been shown in several papers (see, e.g., [Mahanti et al. 2013]) that the distributions of many workload properties are right skewed, i.e., large values can occur with a non-negligible probability. Therefore, their behavior is described by power laws, that is: $p(x) = Cx^{-\alpha}$, where C and α are positive constants, α being the scaling exponent [Clauset et al. 2009]. As a consequence, instead of considering the extreme values as outliers, it is worth studying their properties in details. Pareto and Zipf distributions are classical examples of power law distributions (see Fig. 4). In general, the power law behavior seldom applies to the entire distribution. On the contrary, it often refers to its tail only, thus leading to the so-called heavy tail or long tail effects. We recall that a power law distribution appears as a straight line when plotted in a double logarithmic scale.

Furthermore, the time-varying properties of the workloads can be represented and modeled by means of stochastic processes [Trivedi 2002]. For instance, the models of the burstiness typically found in the workload traffic often rely on Markov modulated Poisson Processes [Fischer and Meier-Hellstern 1993], that is, generalized Poisson processes where the arrival rates change over time. In addition, in the framework of workload modeling and forecasting spectral analysis [Priestley 1981] and time series analysis [Box et al. 2008] are particularly useful. These approaches are frequently devised to represent complex arrival patterns characterized by some periodicity. Various methods (e.g., nonparametric filtering, wavelets) are applied to identify the structure of the predictive models, extract the corresponding components (i.e., trend and seasonal components) and estimate their parameters. These models are particularly relevant for studies dealing with online capacity planning and resource management.

Another interesting aspect taken into account in the analysis of workload dynamics is represented by the behavior of the users and, in particular, the sequences of requests - often referred to as user sessions - generated by the users towards services or infrastructures. The access patterns of the users and their sessions are frequently characterized in terms of activity (ON) and inactivity (OFF) periods, whose durations offer some valuable insights into the behavior of the users and their impact on the underlying infrastructure.

Ferrari first introduced the concept of user behavior graph to model the behavior of interactive users (see, e.g., [Calzarossa and Ferrari 1986; Ferrari 1984]). Some years later, Menascé et al. and Menascé and Almeida [Menascé et al. 1999; Menascé and Almeida 2001] extended these graph-based representations to study e-business workloads. In particular, the so-called Customer Behavior Model Graphs are used to model the navigational patterns of the customers of e-commerce websites. The nodes of these probabilistic graphs represent different command or request types issued by the users and the edges correspond to the transitions between nodes. A transition probability matrix is associated with each graph. Additional characteristics considered in the description of the nodes are their sojourn times, that include both the think times of the users and the response times of the system.

Graphs are also used to study the structure and connectivity of Web workloads as well as its evolution and dynamics [Broder et al. 2000]. Moreover, in the framework of online social networks graphs describe the connections and relationships existing and being developed in the virtual communities. In particular, in a directed graph $G = \{V, E\}$, each node $v \in V$ represents the entity of interest (e.g., Web page, user) and a directed edge $e \in E$ captures the relationship between nodes. Various metrics are used to explore the topology and properties of the graphs and of their nodes [West 2001]. For example, the paths between nodes are analyzed in terms of the shortest path distance, namely, the minimum number of hops to reach from a node every other node. Other metrics refer to the eccentricity and characteristic path length of a node. From these metrics, the radius, diameter and characteristic path length of the entire graph are derived. Similarly, the in-degree and out-degree of a node characterize the number of incoming and outgoing edges of a node. Moreover, the clustering coefficient of a node (i.e., the fraction of pairs of nodes that are connected to each other by edges) allows for exploring how densely the neighborhood of a node is connected.

The rest of the paper offers a comprehensive state of the art of the methodological approaches applied for the characterization of the workloads of five popular application domains. The peculiarities and commonalities of these workloads are discussed in detail. Moreover, an overview of the state of the art in terms of data collection and analysis methods is provided in each section (see tables I to III).

3. WEB WORKLOADS

Web workloads have been extensively studied since Web inception. The characteristics and nature of these workloads have evolved significantly. Conventional Web workloads mainly consist of the HTTP requests issued by the clients towards Web or proxy servers to download pages. On the contrary, the recent technological advances allow users to search for information, conduct business as well as disseminate and share their own textual and multimedia content. This leads to a different perception and participation of the users in the Web coupled with an explosive growth of content that changes frequently and rapidly. In this framework, to assess the impact on scalability and resource provisioning and more generally on the QoS and QoE requirements, it has become very important to analyze and understand the characteristics of Web 2.0 workloads [Cormode and Krishnamurthy 2008].

In this section we focus on the characterization of the different flavors of conventional Web workloads, whereas the characterization of novel emerging services (i.e., online social networks, mobile applications, video services, cloud computing applications) is addressed in sections 4 to 7.

3.1. Conventional Web workloads

Table I presents an overview of the state of the art in the area of Web workload characterization. As can be seen, the most relevant aspects considered in the framework of conventional Web workloads refer to:

— page properties
— traffic properties
— access patterns
— user behavior.

All these aspects have been analyzed in the literature under different perspectives (e.g., client, server) and with different approaches (e.g., statistical, hierarchical) and objectives (e.g., prefetching, caching, scalability). The measurements collected by the Web servers in their access logs (see Fig. 1) are typically the basis of these analyses.

	Target	Focus	Data collection		Data analysis	
			How	Who	What	How
[Pitkow 1999]	Web	Workload characterization			SURVEY	
[Arlitt and Williamson 1996]	Web	Invariants	Logging	Authors	HTTP request properties	EDA
[Arlitt and Williamson 1997]	Web	Invariants	Logging	Mixed	HTTP request properties, location	EDA
[Williams et al. 2005]	Web	Invariants (evolution)	Logging	Authors	HTTP request properties	EDA
[Mahanti et al. 2009]	Web	Traffic profile, usage patterns	Logging	Authors	Traffic/visitor properties	EDA, fi
[Bent et al. 2006]	Web	Cacheability, cookie usage	Sniffing	Authors	HTTP traffic headers	EDA
[Gill et al. 2011]	Web services	Service characterization	Logging	Authors	Service properties	EDA, fi
[Ihm and Pai 2011]	Web traffic	Web traffic evolution	Logging	Public	HTTP request properties	EDA
[Butkiewicz et al. 2011]	Web page	Web page complexity	Crawling	Authors	Object properties	EDA, fi
[Arlitt et al. 2001]	E-commerce	Multi-tier archtecture scalability	Logging	Authors	HTTP request/session properties	EDA, fi
[Menascé 2003]	E-business	Methodological approach	Logging	-	HTTP request/session/function properties	EDA, fi
[Menascé et al. 2003]	E-business	Hierarchical multiscale characterization	Logging	Provider	HTTP request/session/function properties	EDA, fi
[Akula and Menascé 2007]	Auction	Usage patterns	Crawling	Authors	Bidding properties	EDA, fi
[Doran and Gokhale 2011]	Web	Web robot detection survey	-	-	-	-
[Calzarossa and Massari 2012]	Web	Web robot access patterns	Logging	Authors	HTTP request properties	EDA
[Calzarossa et al. 2013]	Web	Web robot classification	Logging	Authors	HTTP request properties	EDA, ti
[Dikaiakos et al. 2005]	Web	Web robot behavior	Logging	Authors	HTTP request properties	EDA, fi
[Stassopoulou and Dikaiakos 2009]	Web	Web robot detection	Logging	Authors	HTTP request properties	Bayesia
[Doran et al. 2013]	Web	Web robot characterization	Logging	Authors	HTTP request properties	EDA, fi
[Lee et al. 2009]	Web	Web robot characterization	Logging	Provider	HTTP request properties	EDA
[Tan and Kumar 2002]	Web	Navigational patterns	Logging	Provider	HTTP request/session properties	Machine
[Adar et al. 2009]	Web content	Content evolution	Crawling	Authors	Change properties	EDA
[Brewington and Cybenko 2000]	Web content	Dynamics	Crawling	Authors	Change properties, age	EDA, fi
[Calzarossa and Tessera 2008]	Web content	Website evolution	Crawling	Authors	Web content properties	EDA, fi
[Calzarossa and Tessera 2015]	Web page	Dynamics prediction	Crawling	Authors	Change temporal properties	Time se
[Cho and Garcia-Molina 2003]	Web page	Change prediction	Mixed	Authors	Change properties	Fitting,
[Fetterly et al. 2004]	Web page	Content evolution	Crawling	Authors	Page properties	EDA
[Radinsky and Bennett 2013]	Web page	Predicting content changes	Crawling	Authors	Change properties	Machine

Table I: Summary of the state of the art of Web workload characterization.

The summary of the earliest studies of the characterization of conventional Web workloads presented in [Pitkow 1999] emphasizes that many workload properties follow regular and predictable patterns. For example, the page popularity at client, proxy and server levels is described by heavy tailed Zipf distributions (see Fig. 4(b)). Similarly, the HTTP traffic is characterized by a periodic nature and a self-similar behavior. In addition, it has been shown (e.g., [Arlitt and Williamson 1996; 1997; Williams et al. 2005]) that some workload properties and characteristics are *invariant*, namely is, they do not change across websites of different providers and are likely to persist over time. These invariants include, among the others, document size distribution, document referencing behavior, inter-reference times and one timers. All these findings play an important role in studies dealing with Web server performance and capacity planning.

Another interesting study in the domain of conventional Web workloads is represented by the work of Mahanti et al. [Mahanti et al. 2009] who characterize the load of the Web server of a large conference from both the client and the server perspectives. For the analysis of the usage patterns of the site, the study employs the logs recorded by the server, whereas the reports produced by the Google Analytics service provide complementary information for tracking user activity. The characterization relies on a multi-layer approach. The high level look of the traffic outlines its profile and trends and more specifically the distributions of the file types requested and the response codes of the server. Additionally, the visitor trends are examined in terms of various characteristics, such as usage patterns of the website, frequency of visits, visit durations and page depth per visit. The study confirms the long tail nature of the distributions that describe most of these characteristics. Other aspects considered in this investigation deal with the load patterns of the website across days of the week and the locality properties of the visitors. In particular, the analysis of the traffic sources shows the increased importance played by search engines that nowadays represent for many users their primary entry point to the Web.

The impact exercised by cookies and cacheability attributes on Web content caching and delivery is analyzed in [Bent et al. 2006]. This comprehensive investigation focuses on a server farm hosting a large number of commercial websites. Measurements refer to the network footprint of the HTTP traffic collected by sniffing the TCP data packets related to HTTP transactions. The analysis of the HTTP header lines shows the tendency of many websites to make an indiscriminate use of cookies. It is also very common to find uncacheable HTTP responses and responses requiring mandatory validations that prevent the optimal deployment of content delivery mechanisms.

A new dimension to Web workload characterization, Web-based services, is introduced by Gill et al. [Gill et al. 2011]. The study presents a methodological approach to identify the service properties (e.g., providers, instances, brands) and understand their usage. This characterization provides a comparison point for future studies on the evolution of the Web. In addition, by revisiting the traditional Web properties, it has been shown that - despite the significant transformation in scope undertaken by the Web - the underlying object access properties have not significantly changed.

The evolution of the nature of Web traffic is another important aspect to be taken into account when dealing with Web workload characterization. The major changes undertaken by this type of traffic have been investigated in [Ihm and Pai 2011] by analyzing a five years real Web traffic collected from a globally-distributed proxy system. This analysis offers some interesting insights in the new traffic characteristics. In particular, a good fraction of the overall traffic is due to client-side interactions. Moreover, as also pointed out in [Butkiewicz et al. 2011], the complexity of Web pages has grown in terms of both the number of objects and their size. Despite these new properties, page loading latency has dropped because of the increased number of concurrent connections opened by the browsers and the improved caching behavior.

3.2. Shopping service workloads

In the framework of shopping services exploited on the Web, workload characterization is a fundamental component to evaluate the impact of the user requests and sessions on the performance of the infrastructures. Moreover, it is very useful for assessing the scalability of these multi-tier architectures. While the clustering techniques applied in [Arlitt et al. 2001] identify classes of user sessions for capacity planning purposes, a hierarchical and multiple time scale approach is adopted by Menascé [Menascé 2003] and Menascé et al. [Menascé et al. 2003] to study the interactions of the customers with e-commerce websites. In details, this hierarchical multi-scale perspective relies on the decomposition of the workloads into various levels, namely:

— business
— session
— function
— protocol

each described in terms of some level specific features. The business level takes into account the overall characteristics of the workload and their impact on the interactions of the customers with the sites. The session level considers the sequences of requests originating from individual customers during a single visit to the e-business site. The workload at this level is modeled in terms of session length and number of sessions initiated per time unit. In particular, the study suggests that the number of requests per session follows a heavy-tailed distribution. The next lower level focuses on the nature and popularity of the e-business functions invoked by the customers during a session. It has been shown that a significant fraction of the functions requested by customers refers to product selection as opposed to product ordering. Finally, to capture and model the properties of the arrival process and the patterns of the corresponding interarrival times, the characterization considers the HTTP requests arriving at the Web servers. More specifically, a statistical analysis of the number of requests complemented with their visual inspection across various time scales allows for detecting correlations and quantifying their strength.

Another popular method of shopping on the Web is represented by online auctions. Despite classic e-commerce websites, the activities of auction websites are characterized by very peculiar features (e.g., spikes during the closing minute, drops after the closing of each auction). Hence, a successful deployment of these services in terms of revenues and user experience requires a good understanding of their workloads. The characterization approach proposed in [Akula and Menascé 2007] is based on a two level view of the workload, namely, site viewpoint and user viewpoint. The analysis at the site level covers the properties and temporal behavior of the workload as seen at the auction site. In particular, clustering allows auction classification, whereas a multi-scale analysis is performed to identify the activities at different time scales. On the contrary, the user level is examined in terms of various attributes (e.g., number of bids/bidders, closing prices, auction duration, times between consecutive bids) and of the popularity properties of the various actors (e.g., auctions, winners, sellers, bidders). It has been shown that the bidder popularity follows a power law distribution, thus indicating that the majority of bids are placed by few unique bidders. Moreover, the number of bids placed throughout the lifetime of an auction does not depend on the auction duration.

3.3. Web robot traffic

When studying Web workloads particular attention has to be paid to the requests generated by Web robots. Nowadays Web robots are extensively deployed for the periodic crawling activities of search engines as well as for malicious purposes (e.g., exploiting website vulnerabilities, gathering website business intelligence). Since robot requests represent a good fraction of the overall Web traffic, it is very important to identify their presence and assess their impact on the performance of the infrastructure.

Several studies address the characterization of Web robot traffic (see, e.g., [Calzarossa and Massari 2012; Calzarossa et al. 2013; Dikaiakos et al. 2005; Doran et al. 2013; Lee et al. 2009; Stassopoulou and Dikaiakos 2009; Tan and Kumar 2002]). These papers investigate the behavior and navigational styles of various Web robots by analyzing the qualitative and quantitative properties of the corresponding HTTP requests (e.g., resource type and size, method, user agent, referrer, status code). In particular, to discover similarities and contrasts in the behavior of human and robots, several metrics (e.g., popularity index, coverage) have been derived from the HTTP traffic properties and types of requested resources [Dikaiakos et al. 2005; Lee et al. 2009]. Conversely, the temporal properties of the browsing and crawling patterns are modeled by applying fitting techniques and time series analysis [Calzarossa and Massari 2012; Calzarossa et al. 2013; Doran et al. 2013]. Another important aspect covered in the framework of Web traffic refers to the identification and classification of the sessions as being generated by robots or human [Stassopoulou and Dikaiakos 2009; Tan and Kumar 2002]. A session is defined as the sequence of HTTP requests that originate from the same user agent and arrive at a Web server within a given time frame. Decision trees and Bayesian networks combined with machine learning approaches are usually applied for these classification purposes. An interesting survey of the prevalent Web detection approaches is presented in [Doran and Gokhale 2011].

3.4. Web content

Web workloads have also been investigated from the perspective of the website content by focusing on various aspects related to content characteristics and dynamics (see, e.g., [Adar et al. 2009; Brewington and Cybenko 2000; Calzarossa and Tessera 2008; 2010; 2015; Cho and Garcia-Molina 2003; Fetterly et al. 2004; Radinsky and Bennett 2013]). Crawling is the data collection approach commonly adopted in this framework. More specifically, measurements are collected by taking multiple snapshots of Web pages or websites selected according to the objective of the investigation. The analysis of these samples allows for assessing whether and to what extent a page changed.

The large-scale study of the evolution of Web pages presented in [Fetterly et al. 2004] considers the page properties (e.g., size, number of words) and the degree of their changes. The statistical observations of the measurements show that page size is a strong predictor of both frequency and degree of change. Similarly, even though the average degree of change varies across websites, changes are rather correlated. This means that it is possible to predict future changes to a page from its past changes.

Various similarity measures (e.g., edit distance, Dice coefficient, cosine coefficient of similarity) are introduced in [Adar et al. 2009; Calzarossa and Tessera 2008; 2010] to quantify the degree of change to a page. These measures allow for the identification of stable and dynamic content within individual pages. Moreover, the problem of estimating and predicting the change frequency of Web content - even in presence of an incomplete change history - has been addressed in [Brewington and Cybenko 2000; Cho and Garcia-Molina 2003]. Machine learning approaches, numerical fitting techniques as well as time series analysis are applied to characterize and model the temporal evolution of the content and predict the change dynamics [Calzarossa and Tessera 2015; Radinsky and Bennett 2013]. All these models and findings have important implications for devising tuning strategies aimed at reducing page load latency and designing technologies for content discovery, retrieval and management.

4. ONLINE SOCIAL NETWORK WORKLOADS

The increased popularity of online social networks (OSN) exploited for both personal and professional activities opens new challenges related to the management and provisioning of the infrastructures for their deployment. These networks, frequently built around specific themes, integrate a large variety of technologies for providing users with sophisticated services, e.g., collaboration, communication, geolocalization. In particular, OSNs allow individuals to keep in touch with family, friends and customers, disseminate information and do business. In the virtual communities being formed within OSNs, users play a central role. They connect between each other, upload and share their own content, access, comment and rate content posted by other users. It is then important to understand the behavior of the users and their complex interactions, that is, the OSN workloads [Cormode et al. 2010], as to guarantee the QoS and QoE requirements foreseen for these types of services. In addition, this knowledge is fundamental for developing effective marketing strategies aimed at maximizing revenues and for identifying anomalous behaviors.

A large body of the literature addresses the characterization of OSN workloads (see, e.g., [Ahn et al. 2007; Benevenuto et al. 2009; Cha et al. 2008; Cha et al. 2012b; Gonçalves et al. 2010; Guo et al. 2009; Jeon et al. 2012; Kumar et al. 2006; Massari 2010a; Viswanath et al. 2009]). Most studies rely on measurements obtained by periodically crawling the OSN websites, whereas only few are based on datasets made available by the providers themselves. It is worth noting that providers are seldom willing to disclose their own access logs for competitive reasons and because of confidentiality and privacy issues towards their users. Nevertheless, to reduce the burden, e.g., bandwidth and overhead, of crawling activities, providers often offer APIs for querying some specific information about the services being deployed. However, since the scope of these APIs is often quite limited, this information is frequently integrated and complemented with additional measurements obtained by scraping the OSN websites. Moreover, to capture the OSN dynamics and evolution, consecutive timestamped snapshots of the websites are collected through repeated crawling activities. In general, the crawling process is combined with a parsing process whose goal is to extract the attributes describing the properties of the OSN workloads as well as the behavior of the users and their social interactions.

It is worth mentioning that particular care must be paid to crawling the social graphs associated with OSNs, that is, the graphs describing the relationships being developed within a network. This process is very challenging and cumbersome in that it requires many crawling iterations [De Choudhury et al. 2010; Gjoka et al. 2011]. In addition, to keep the size of the graphs manageable, it is necessary to devise appropriate sampling techniques, e.g., random walks, graph traversal methods, and to obtain at the same time representative, e.g., unbiased, graph samples.

A typical approach applied in the framework of OSN workload characterization is based on the analysis of user profiles, i.e., the customized Web pages containing personal information, preferences and various types of multimedia content, associated with users upon their subscription to OSNs (see [Farahbakhsh et al. 2013; Lampe et al. 2007; Massari 2010a; 2010b; van Dam and van de Velden 2015]). Profiles store a rich variety of information about the users, their activities and relationships. The analysis of the attributes extracted from the profiles, e.g., gender, age, location, education, job, cultural interests, number of friends, number of comments, number of uploads, community memberships, provides interesting qualitative and quantitative insights into the overall characteristics and behavior of the users. More specifically, the correlations and associations among these attributes highlight their role in the formation of online connections. In addition, the application of clustering techniques allows the identification of groups of users with similar behavior, e.g., level of participation to the OSN.

A complementary approach towards the characterization of OSN workloads relies on the analysis of the user clickstreams collected by monitoring the HTTP traffic at the network edges (see [Benevenuto et al. 2009; 2012; Schneider et al. 2009]). This approach captures the social interactions of the users with the OSNs by considering some specific activities that cannot be extracted from user profiles, e.g., profile and home page browsing and editing, messaging, search. In particular, the analysis of the clickstreams allows the identification of the dominant activities of the users. Further details about the dynamics of the user behavior are obtained by studying the transitions between pairs of activities and the associated rates. Markovian models are usually adopted for this purpose. In addition, for assessing the impact of the user activities, clickstreams are grouped into sessions, that is, sequences of requests issued by a given user during a single visit to the OSN. Sessions are described by several parameters, such as, duration, frequency, inter-session time, number of bytes transferred, that summarize how often and how long users connect to the OSN as well as their degree of concurrency.

To better understand the phenomena driving the social interactions and assess their impact on the technological infrastructures, the structure of various OSNs, e.g., Facebook[1], Google+[2], LinkedIn[3], MySpace[4], Renren[5], has been extensively studied by analyzing the corresponding graphs (see, e.g., [Ahn et al. 2007; Jiang et al. 2010; Leskovec et al. 2008; Mislove et al. 2007; Nazir et al. 2008; Traud et al. 2011; Wilson et al. 2012]). In general, the nodes of these graphs refer to the individuals, while the edges correspond to the various types of direct or latent social relationships being investigated, e.g., friendship, interaction. Well-known metrics, such as, out-degree and in-degree of a node, path length, reciprocity, clustering and assortativity coefficients, are used to describe the topological characteristics of the graphs and highlight properties, such as, node popularity, small-world phenomena, potential extent of content propagation. In particular, it has been shown that the degree distributions are heavy tailed due to the presence of a small number of nodes characterized by a very large number of edges. Moreover, there is the tendency of high degree nodes to connect to each other.

Other phenomena, such as, network growth and formation strategies, migration patterns among communities, have been investigated by considering the structure of the OSNs from an evolutionary perspective that takes into account temporal information related to nodes and edges (see, e.g., [Gong et al. 2012; Kumar et al. 2006; Leskovec et al. 2008]).

In what follows we present the state of the art of workload characterization of some specific categories of social network services, that is, blog and microblog services, visual content sharing services, location based services and collaborative networks.

Blog platforms, e.g., LiveJournal[6], Tumblr[7], play an important role in the framework of social network services in that they allow users to create and disseminate textual and multimedia content. Understanding the blogosphere and in particular blog popularity, behavior of bloggers and effects of their interactions is of paramount importance for an effective exploitation of these services and for devising optimization strategies of the technological infrastructures. Some studies, e.g., [Duarte et al. 2007; Jeon et al. 2012], characterize the workloads as seen at the blog server side. In particular, the characterization focuses on different views, that is, aggregate access patterns of all bloggers, their individual sessions as well as access patterns of each blog. In addition, blogs are analyzed in terms of properties, such as, reference behavior, temporal locality, level of blogger involvement. Other studies, e.g., [Chi et al. 2007; Goldberg et al. 2009; Gonçalves et al. 2010; Kumar et al. 2003; Leskovec et al. 2007; Mitrović and Tadić 2010], focus on the social structure of the blogs and on information propagation. Graphs are used to represent the communication and temporal patterns and more generally the social interactions being developed between bloggers. Note that the nodes of the graphs represent the bloggers, and the edges denote their

[1]http://www.facebook.com
[2]https://plus.google.com
[3]http://www.linkedin.com
[4]http://www.myspace.com
[5]http://www.renren.com
[6]http://www.livejournal.com
[7]https://www.tumblr.com/

relationships. Different levels of detail, e.g., individual nodes, communication edges, entire network, are adopted for investigating the properties of the graphs. It has been shown that these properties are rather stable, that is, they are not significantly affected by the blog evolution. Hence, they are particularly useful for identifying anomalous behaviors. Moreover, information propagation on the blogspace is illustrated by cascades, i.e., conversation trees, whose properties and topological patterns provide useful insights in the characteristics of the underlying social networks as well as for predicting the spread of ideas and opinions.

Similar approaches (see e.g., [Cha et al. 2012b; Fu and Shen 2014; Gabielkov et al. 2014; Gao et al. 2012; Java et al. 2007; Jiali et al. 2012; Krishnamurthy et al. 2008; Kwak et al. 2010; Li and Cardie 2014; Yan et al. 2013]) are applied to characterize microblogging services, e.g., Twitter[8], Sina Weibo[9], where users communicate and share information by posting up to 140 characters long messages, i.e., tweets. The mix of conventional blogging features with OSN features makes these user-to-user interactions rather complex. Let us remark that because of the "follower" concept and the directed nature of the relationships among users, tweets can reach a large audience without any manual user intervention. To investigate the nature of the connections between users, early studies [Java et al. 2007; Krishnamurthy et al. 2008] analyze the topological and geographical properties of the network and describe the behavior of the users in terms of some broad attributes, e.g., number of followers, number of following users. More recently, the social network attributes of microblog users, e.g., geographic location, language, number of followers, number of tweets, number of friends, number of favorites, have been complemented by considering the characteristics of the posts, e.g., number of hashtags, number of mentions, number of URLs. In particular, it has been observed that hashtags and URLs are more likely to be found in messages posted via desktop applications than via mobile applications. Moreover, the relationships between the number of tweets and the number of followings/followers are used to explain the behavior of the users. In general, it has been shown that a small fraction of the users posts tweets, whereas the majority is passive and simply receives/reads the tweets being posted. In addition, the topological structure of the social graphs built around these microblogging services underlines interesting properties among users, e.g., reciprocity, degree of separation, homophily.

Furthermore, the analysis of the "retweeting" phenomenon, that is, message relaying beyond adjacent neighbors, explains how information propagates in the network in relation to the user characteristics and across topics and time (see [Cataldi and Aufaure 2014; Lerman and Ghosh 2010; Yang et al. 2010]). In particular, the retweet trees, i.e., the subgraphs extracted from the entire network, highlight the presence of multi-hop chains connecting users, and of specific patterns, such as, retweeting the same tweet, retweeting each other. Moreover, the speed of information propagation is measured by studying the evolution of the retweets over time. As a result of these analyses, categories of popular and influential users, e.g., mass media, celebrities, politicians, who are the seeds for information diffusion, are clearly identified.

In the framework of visual content sharing, social media platforms, such as, Flickr[10], Foursquare[11], Instagram[12], Picasa[13], Pinterest[14], offer opportunities of interaction and collaboration. These platforms exploit a wide range of OSN features that allow users to upload, organize, disseminate and rate content, e.g., photos, images, by sharing at the same time the

[8] http://www.twitter.com
[9] http://d.weibo.com
[10] http://www.flickr.com
[11] https://foursquare.com
[12] http://instagram.com
[13] http://picasa.google.com
[14] http://www.pinterest.com

geographic position of their mobile devices. In the literature, the workloads of these types of platforms have been investigated under different perspectives, by considering aspects related to user behavior as well as to content usage and propagation patterns.

In the case of Flickr, a pioneer of these platforms, the characterization focuses on the properties of the content, i.e., photos, being uploaded and shared, and its temporal evolution (see, e.g., [Cha et al. 2008; Cha et al. 2009; Cha et al. 2012a; van Zwol 2007]). The popularity of a photo is described in terms of well known metrics, e.g., number of views, number of favorite marks, number of comments. In particular, the age of the photos is an important indicator for the identification of the patterns describing the temporal evolution and growth of the popularity. Despite other OSNs, the main driver of the relationships being developed within Flickr among users who post photos and users who mark the photos as favorite is the content itself rather than the friendship. Hence, the characterization of these types of indirect interactions, i.e., interactions among users through photos, relies on the corresponding interaction graphs whose analysis offers valuable insights on the level and extent of the interactions. Moreover, similarly to other OSNs, social cascades illustrate the information propagation, whose spread is estimated using epidemiological models.

Some recent studies, e.g., [Bernardini et al. 2014; Gilbert et al. 2013; Han et al. 2014; Ottoni et al. 2013], investigate the usage patterns and the social interactions being developed within Pinterest, a social content curation website mainly devoted to leisure and entertainment. In detail, the activity on Pinterest, i.e., the amount of content generated, is examined from a statistical perspective by considering both quantitative, e.g., number of "pins" and "repins", number of comments, number of followers, number of boards, and categorical attributes, e.g., gender, country. In addition, the role played by the gender in these social interactions is investigated in terms of the user characteristics, such as, repinning, interests and language used. To further explore the user behavior, the activity patterns of the users are modeled as Markov chains, whose states correspond to the activity types, e.g., pin, repin, like, and whose edges, with their associated probabilities, represent the transitions between states. Similarly, the pin propagation is described by a tree, i.e., a directed graph representing the users and their repins. The temporal and structural properties of the pin trees, e.g., inter-pin times, maximum depth and width, are the basis for estimating the pin propagation speed and patterns. As a result of these analyses, it has been observed that Pinterest mainly appears as a "retransmission network" mostly driven by the pin properties, that is, their topic and content.

As an attempt to bridge the gap between the physical world and the virtual world of the OSNs, several characterization studies take into account the new dimension of user participation introduced by geo-social services, where users share the current physical location of their mobile device and geo-tag their media content, e.g., text, photo, video, accordingly (see, e.g., [Allamanis et al. 2012; Cheng et al. 2011; Cranshaw et al. 2010; Le et al. 2014; Lins et al. 2014; Pelechrinis and Lappas 2014; Preoţiuc-Pietro and Cohn 2013; Silva et al. 2013; Vasconcelos et al. 2012]). These studies investigate the role of these features, their impact on the existing social ties as well as the new connections derived from the geographical locations of the users and from their geo-tagged content. In particular, the popularity of a location is analyzed by considering properties, such as, the number of users sharing a location at a given time and the amount of tagged content at a specific location. Furthermore, various types of graphs are introduced to explore the social, spatial and temporal aspects of the user behaviors especially in relation to the content being shared. The user graphs represent the social relationships between users, namely, the conventional relations existing in a social network and the additional relations that may be induced by the locations visited by the users. Similarly, the location graphs, derived from the sequences of locations visited by the users, exploit the connections between geographical locations. Moreover, the relations between the users and the locations being visited, or the content being shared are exploited by the user-location graphs. Note that these bipartite graphs are characterized by two types of nodes representing the users and the locations, respectively. The structural properties of all these graphs provide a detailed description of the visiting patterns of the users and allow the identification of influential users and points of interest within a certain geographical area. In addition, the analysis of the hidden relations between time and location provides further insights in the user behavior, e.g., periodic patterns in user mobility and activities performed at a specific location. All these results reveal very important for studying the evolution of the social network and devising traffic forecasting and urban planning techniques, recommendation systems, as well as epidemiological models of disease spread.

Finally, in the framework of collaborative networks, it is worth mentioning the characterization studies on Wikipedia[15] workload. The knowledge of the overall properties of the articles and in particular of their access patterns and popularity is fundamental for designing effective replication and distribution strategies as well as scheduling and caching algorithms. In the literature, Wikipedia workload has been characterized by considering its evolution and the user behavior (see, e.g., [Almeida et al. 2007; Brandes et al. 2009; Eldin et al. 2014; Urdaneta et al. 2009]). This workload is described in term of various quantitative features related to the individual articles, e.g., number of entries, their history, e.g., number of revisions, and their graph structure, e.g., number of links to other articles. Moreover, the dynamics of the requests to the various articles are represented as time series whose analysis provides models able to explain and predict the long term evolution and short term trends of the workload.

5. MOBILE DEVICE WORKLOADS

Mobile devices have become closely integrated in our personal and professional lives and nowadays represent a very popular gateway for most types of social interactions. These devices deploy many diverse mobile applications, that is, apps, either native, i.e., factory installed, or downloaded and installed directly by the users via dedicated stores, e.g., Apple App Store[16], Google Play Store[17]. Apps allow users to create some complex interactions with the devices for accessing and sharing content and resources. Service and content providers as well as network operators and software developers need to cope with these challenging scenarios and optimize the QoE of the users by adopting effective content optimization and placement strategies as well as efficient power management and dynamic resource provisioning policies. Hence, one important question to be addressed deals with the app usage patterns, that is, how, where and when apps are used.

Table **??** summarizes the

[15] http://www.wikepedia.org
[16] http://store.apple.com
[17] http://play.google.com

	Target	Focus	Data collection		Data analysis	
			How	Who	What	How
[Xu et al. 2011]	Apps	Usage patterns	-	Provider	Traffic/app s properties	EDA
[Böhmer et al. 2011]	Apps	Contextual usage	Sensing, sampling	Authors	Context ual properties (& location)	EDA
[**?**]	Users	Smartphone contextual usage	Sensing	Authors	Physical/ social contextual properties	EDA
[Do and Gatica-Perez 2014]	Users	User behavior prediction	Sensing	Authors	Physical/soci al contextual properties	Fitting, machin
[Liao et al. 2013]	Apps	App usage prediction	-	Provider	Usage properties	Probabilistic
[Patro et al. 2013]	Apps	User experience	Logging	Authors	Resource usage, devi ce properties	EDA
[Yan et al. 2012]	Apps	App Launch predictor	-	Public	Contextual propertie s	Machine learnin
[Yang et al. 2015]	Users	App usage behavior	-	Provider	Resource usage	Clustering
[Falaki et al. 2010]	Users	Smartphone use	Logging	Authors	Traffic/app proper ties	EDA, fitting
[Shin et al. 2012]	Apps	App usage, prediction	Sensing, sampling	Authors	Cont extual properties	Machine learnin
[Petsas et al. 2013]	Apps	App popularity trends	Crawling, sampling	Authors	Usag e/download properties	EDA, fitting
[**?**]	Apps	Usage patterns	Logging	Authors	Temporal/spatial p roperties	EDA, machine

Table II: Summary of the state of the art of mobile app characterization.

The comprehensive investigation presented in [Xu et al. 2011] covers these issues by examining aspects, such as, spatial and temporal prevalence, locality and correlation of smartphone apps. The anonymized dataset analyzed in this study refers to the traffic of a tier-1 cellular network provider. From these measurements, apps are identified using signatures based on the HTTP headers and in particular the `User-Agent` field values. The features taken into account to describe the usage patterns of the individual apps refer to traffic volume, access time, unique subscribers and locations. The analysis of these features highlights some interesting relationships among apps, e.g., co-occurrence and overlap between pairs of apps, whose mix might exert a significant impact on the performance. Similarly, the mobility patterns inferred from the network access patterns provide a clue for coping with the connectivity and bandwidth variability issues experienced by bandwidth sensitive apps and once more for improving the user experience.

Other important properties considered in the characterization of mobile device workloads deal with the context-aware behavior of the users and their interactions with the devices. As discussed in several papers (see, e.g., [Böhmer et al. 2011; Do and Gatica-Perez 2014; Liao et al. 2013; Malmi 2014; Patro et al. 2013; Yan et al. 2012; Yang et al. 2015]), these aspects affect performance and user experience and, as such, they have to be taken into account when designing mobile apps and operating systems and for devising optimization mechanisms aimed, for example, at reducing the launch delays typically faced by many apps. Hence, to better understand and predict the app usage patterns, it is compelling to fully exploit contextual information, e.g., time, location, user profile configuration, last-used app. Time of day and location strongly affect app popularity and play a central role in assessing their usage. For instance, it has been shown that news apps are more popular in the morning, whereas gaming apps over night. In addition, the analysis of user sessions, that is, the sequences of apps opened by the users, highlights the relationships and dependencies in app usage, e.g., bundling effects, and offer valuable insights for making predictions about what app will be opened next. Similarly, the characterization of the apps along dimensions, such as, network behavior, app usage and footprints, device capability and settings, user actions, provides a deep understanding of their impact on network resources, application usage, revenue generation for developers and user experience.

Smartphone usage is described in [Falaki et al. 2010] in terms of several features, e.g., session length, interarrival time between consecutive sessions, application popularity, that summarize the user activities. Simple analytical models are derived for each feature. For instance, a mixture of exponential and Pareto distributions captures the variability of the session length, whereas a Weibull distribution accurately explains the OFF times of smartphone screens.

A machine learning approach is proposed in [Shin et al. 2012] to obtain personalized app predictions based on a wide range of contextual information. This information is collected on the smartphones by sampling various signals produced by built-in sensors, e.g., device status, battery status, running apps, cellular network location. The predictions of the apps associated with a given context are obtained by means of an inference model based on simple probabilistic classifiers. Similarly, to predict which apps are most likely to be launched at a given time, probabilistic models are exploited [Liao et al. 2013]. These models rely on temporal-based features, e.g., global, temporal and periodic app usage, obtained by mining app usage traces.

An alternative perspective for the characterization of the app usage patterns is adopted in [Petsas et al. 2013]. This perspective covers the entire app ecosystem from the point of view of the dedicated app stores. More specifically, by periodically crawling these stores, various information referring to the characteristics of each individual app, e.g., number of downloads, version, category, price, developer, are collected. These characteristics are the basis for investigating the app behavior, namely, download patterns, popularity trends and pricing strategies. It is interesting to point out that, despite what discovered for other workload types, e.g. Web workloads, app popularity deviates from Zipf-like models as it appears truncated at both edges. In particular, the tail truncation is easily explained by the clustering effects, that is, the strong temporal affinity of user downloads to app categories. On the contrary, the head truncation is due to the “fetch-at-most-once” property, typically found in file sharing systems and also observed in the app ecosystem. This is due to the limited number of updates undertaken by individual apps. Moreover, the popularity is characterized by a highly skewed Pareto effect, with 10% of the apps accounting for 70–90% of the total downloads. These properties have some significant implications when devising strategies aimed at improving delivery performance and increasing developers revenues as well as for designing effective recommendation systems.

6. VIDEO SERVICE WORKLOADS

Video service workloads refer to the load produced by the users who access and share content either supplied by media producers or self generated. A good understanding of the characteristics of these workloads is very beneficial in many contexts (e.g., design and development of content distribution systems, resource management and provisioning).

The most relevant aspects of video service workloads considered in the literature (see Table III for an overview) refer to the characteristics of the various media types as well as to the interactions of the users with the media content and the usage of Web 2.0 features, namely:

— media properties
— traffic properties
— user behavior
— social sharing properties.

	Target	Focus	Data collection		Data analysis	
			How	Who	What	How
[Almeida et al. 2001]	VoD	Educational media server workload characterization	Logging	Authors	Media/client properties	EDA, fitti
[Chen et al. 2014]	VoD	User behavior	-	Provider	Browsing properties	EDA, fitti
[Cherkasova and Gupta 2004]	VoD	Enterprise media server workload characterization	Logging	Provider	Media/client properties	EDA, fitti
[García et al. 2007]	VoD	User behavior	Logging	Authors	User/traffic properties	Fitting, U
[Guo et al. 2005]	Media service	Multimedia workload characterization	Logging	Authors	Media properties	EDA
[Li and Ma 2014]	VoD	Video popularity	Logging	Authors	Video properties	EDA, fitti
[Li et al. 2005]	Media service	Video popularity	Crawling	Authors	Media properties	EDA, fitti
[Liu et al. 2014a]	IPTV	User behavior	Logging	Authors	**Behavioral** properties	EDA, fitti
[Veloso et al. 2006]	Live streaming	Hierarchical characterization	Logging	Provider	Video properties	EDA, fitti
[Yu et al. 2006]	VoD	User behavior	Logging	Provider	Video properties	EDA, fitti
[Gopalakrishnan et al. 2011]	IPTV	User behavior	Logging	Provider	Video/stream control properties	Fitting, F
[Silva et al. 2011]	Live streaming	Usage patterns	API& sampling	Authors	Transmission/channel properties	EDA, fitti
[Borghol et al. 2012]	UGC	Popularity factors	APIs	Authors	Social sharing	PCA
[Brodersen et al. 2012]	UGC	Geographic popularity	-	Provider	Referrer, location	EDA
[Cha et al. 2007]	UGC	Popularity, evolution	Crawling	Authors	Properties, social sharing	EDA, Fitt
[Cha et al. 2009]	UGC	User participation	Crawling	Authors	Social sharing	EDA, fitti
[Chatzopoulou et al. 2010]	UGC	Popularity	APIs	Authors	Social sharing	Graph ana
[Cheng et al. 2013]	UGC	Growth trend	APIs, scraping	Authors	Properties, Social sharing	Graph ana
[Ding et al. 2011]	UGC	Uploading patterns	Crawling	Authors	Demographic	EDA
[Figueiredo et al. 2011]	UGC	Popularity growth	APIs, sampling	Authors	Social sharing, referrer	EDA
[Figueiredo et al. 2014]	UGC	Popularity evolution	-	Public	Social sharing, referrer	Clustering
[Gill et al. 2007]	UGC	Traffic and usage patterns	Sniffing	Authors	Properties, HTTP	EDA
[Gürsun et al. 2011]	UGC	Access patterns	Logging	Provider	Properties	PCA, AR
[Kang et al. 2010]	UGC	Workload characterization	Crawling	Authors	Properties	EDA, fitti
[Islam et al. 2013]	UGC	Popularity evolution	-	Mixed	Social sharing	EDA, fitti
[Li et al. 2014]	UGC	Propagation, user behavior	Logging	Provider	Properties	Graph ana
[Maia et al. 2008]	UGC	User behavior	Crawling, sampling	Authors	User properties	Clustering
[Mitra et al. 2011]	UGC	Invariants	Crawling	Authors	Social sharing	EDA, fitti
[Siersdorfer et al. 2014]	UGC	Commenting/rating behavior	Crawling, sampling	Authors	Social sharing	EDA
[Wattenhofer et al. 2012]	UGC	Social interactions	-	Provider	Social sharing	Graph ana
[Zink et al. 2009]	UGC	Traffic nature	Sniffer	Authors	Properties, HTTP	EDA

Table III: Summary of the state of the art of video service characterization.

In the we explore in detail the workloads associated with media services and with video sharing services (e.g., UGC).

6.1. Media services

Media services provide audio/video content prerecorded and stored on servers as well as their live counterparts. The interaction of the users with this content differs in many respects. Access to stored content is on demand, namely, driven by the users who play an active role by sending their requests at any time and interacting with the media stream during the playback. Conversely, in the case live streaming services, the same content is simultaneously received by multiple users who play a passive role.

In a good number of papers, the characterization of the workloads of video services focuses on the properties of the various media types as well as the user behavior and usage of Web 2.0 features (see, e.g., [Almeida et al. 2001; Chen et al. 2014; Cherkasova and Gupta 2004; García et al. 2007; Guo et al. 2005; Kang et al. 2010; Li and Ma 2014; Li et al. 2005; Liu et al. 2014a; Mitra et al. 2011; Veloso et al. 2006; Yu et al. 2006]). Most of these studies rely on large sets of empirical data collected on the infrastructures where the services are deployed. These studies play a key role in the design of the infrastructures as well as for evaluating and predicting their performance, optimizing storage and content management and delivery and more generally for enhancing the viewing experience of the users.

The workloads of live and Video on Demand (VoD) services have been characterized and modeled under different perspectives by focusing on aspects, such as, arrival process of the requests, video popularity, content access patterns, interactive behavior of the users. In [Gopalakrishnan et al. 2011], the request arrival process in an IPTV environment is modeled in the frequency domain by using a Fast Fourier Transform with few parameters that preserves the diurnal pattern of the traffic and its periodic bursty nature. Moreover, user interactions are studied in terms of the stream control operations, e.g., start, pause, play, generated by the users during a viewing session. The dynamics of these interactions is modeled using a Finite State Machine, whose states correspond to the various control features. Let us remark that separate state transition probability matrices have been derived for accurately reproducing the weekends and weekdays workloads as it has been observed that the characteristics of the corresponding interactions significantly differ. In addition, two alternative approaches are proposed to model the state sojourn time distributions. In particular, under the assumption of independent distributions, a semi-Markov model is obtained. On the contrary, to take into account the overall session duration distribution, a constrained model is derived and the functions that best fit the empirical distributions are identified.

The viewing behavior of the users of an IPTV deployment that offers both live and VoD services is analyzed in [Liu et al. 2014a] with the objective of comparing the access patterns of the two services. The study considers various characteristics, such as, bitrate, video length, request rate, holding time, user mode. Fitting techniques are applied to investigate the arrival process of the requests to the two types of service and derive the corresponding models. Moreover, from the analysis of the holding times for live TV and VoD valuable insights for characterizing the browsing/surfing and viewing behavior of the users are obtained.

Live streaming media workload has also been analyzed under a hierarchical perspective. In [Veloso et al. 2006] this workload is described at three levels of abstraction, that is, client, session and transfer layers, each characterized by an increasing granularity. The top layer of the hierarchy, i.e., the client layer, focuses on the entire client population of the streaming service, by considering properties, such as, client concurrency and interest profiles and interarrival times. The behavior of the individual clients is then mapped at the next lower layer into the corresponding sessions described in terms of activity and inactivity periods, i.e., ON and OFF times. Finally, the bottom layer of the hierarchy takes into account the data transfer streams resulting from the client actions. The autocorrelation function is used to investigate the periodic nature and stationarity of the time series representing the temporal patterns of the workloads. In addition, fitting procedures allow the identification of the distributions, e.g., Pareto, exponential, lognormal, that summarize the workload properties. It is interesting to point out that the client interactions with the live content as well the nature of the content itself influence the variability of some properties, e.g., diurnal patterns. These results play an important role when dealing with capacity planning of live content delivery infrastructures. Furthermore, we emphasize that the approaches based on the hierarchical decomposition of the workloads are particularly powerful in that changes to the characteristics of one layer are automatically mapped and captured by the others.

6.2. Video sharing

Content creation and consumption are rapidly growing in size and popularity. This is due to the increased pervasiveness of high performance mobile devices deployed in personal and professional environments.

The characterization of user generated content (UGC) and, in particular, of YouTube[18] and YouTube-like workloads has been addressed by a large body of the literature (see, e.g., [Borghol et al. 2012; Brodersen et al. 2012; Cha et al. 2007; 2009; Cheng et al. 2013; Ding et al. 2011; Figueiredo et al. 2011; Gill et al. 2007; Gürsun et al. 2011; Maia et al. 2008; Siersdorfer et al. 2014; Silva et al. 2011; Zink et al. 2009]). These investigations examine the nature of the traffic and the properties of the videos being uploaded/viewed. Some of these studies rely on data observed on the infrastructure where the video service is deployed, others on measurements collected by monitoring the traffic over campus networks, and still others gather data by directly crawling the video websites. Let us remark that sampling techniques are frequently applied to avoid bias in the collected data especially towards popular videos.

The network traffic associated with video streams is analyzed in terms of characteristics, such as, duration of transport sessions, payload size, data rate (see, e.g., [Gill et al. 2007; Zink et al. 2009]). On the contrary, the viewing patterns and social interactions created by videos are investigated in terms of features, such as, video and channel popularity, referencing characteristics, usage patterns, file properties, video length, age, transfer behavior (see, e.g., [Cha et al. 2007; Silva et al. 2011]). Various statistical techniques are applied to explore in details these features. In particular, the coefficients of correlation highlight the relationships existing between the features. Moreover, to capture and summarize their behavior, probabilistic distributions are employed. Standard fitting techniques are applied to identify the parameters of the distributions that fit the empirical data. In general, it has been shown that most of the properties of this workload type are described by power law distributions. More specifically, the video popularity is characterized by a power law with a truncated tail, that is, few videos accumulate very many views and the vast majority of videos only attracts a small number of views. In addition, the daily access patterns of the videos are described by simple models, e.g., Autoregressive Moving Average models, that can be used for forecasting, that is, to predict the number of accesses a video will have in the near future. Note that all these models can be exploited for multiple purposes, e.g., capacity planning, designing content delivery architectures, managing disk storage.

Within video sharing services, where users upload their own video clips and watch clips uploaded by other users, Mitra et al. [Mitra et al. 2011] identified several key invariants that describe the workload properties. As already pointed out for other workload types, the presence of characteristics that hold across services deployed by different providers has some important implications when devising content distribution services and video search and recommendation systems. The invariants are derived from the metadata of the videos

[18] http://www.youtube.com

collected by crawling some video websites and refer to features, such as, popularity, rating, comments, bookmarks. We outline that most of these features are characterized by highly skewed distributions described by the Pareto principle, commonly encountered in other workloads. According to this principle, the top 20% uploaders contribute for approximately 80% of the total number of uploaded videos. Another important aspect taken into account in this study is the behavior of the users with respect to their tendency of being passive, i.e., to watch videos, other than active, i.e., to upload and rate videos.

As many online social networks build around video sharing services, the video popularity is also evaluated in terms of other aggregated metrics specifically referring to the social mechanisms for community interaction and participation, e.g., number of comments posted for each video, number of users marking the video as favorite, number of "stars", number of "likes" (see, e.g., [Chatzopoulou et al. 2010; Cheng et al. 2013; Wattenhofer et al. 2012]). It is interesting to analyze the virtual communities being created across videos. These networks, described by means of graphs, are usually the result of interactions associated with the videos other than the result to pre-existing social relationships among users.

In addition, it has been shown that the attributes referring to the social interactions of the users, e.g., reciprocity, clustering coefficient, in contrast to the attributes related to individual users, e.g., number of uploads, number of views, are a good discriminator for identifying relevant user behaviors. Furthermore, the analysis of the comments and comment ratings associated with videos provides additional insights on the level of participation of the users and the community dynamics [Siersdorfer et al. 2014].

The evolution over time of the popularity of a video since its upload as well as the links, i.e., referrers, used to reach the videos, are governed by complex processes that depend on several endogenous and exogenous factors (see, e.g., [Cha et al. 2009; Figueiredo et al. 2014; Li et al. 2014; Liu et al. 2014b]). A good understanding of these dynamic properties offers some important insights into the mechanisms that attract users to videos and contribute to their views and propagation. In particular, starting from the observation that individual video popularity is highly unstable and unpredictable, the three-phase characterization proposed in [Borghol et al. 2011] partitions videos into three disjoint groups based on their age and popularity peak. From the analysis of the videos of each group, the distributions describing the key properties of popularity evolution, e.g., weekly viewing rate and total accumulated views to videos at a particular age, movements of videos between phases, are derived. These distributions represent the basis of the synthetic workload generation models that capture the popularity evolution of newly uploaded videos. It is interesting to point out that these types of model yield accurate predictions even when evaluated over some multi-year measurements of video view counts [Islam et al. 2013].

The analysis of video popularity provides some additional insights for discovering the presence of "information bottlenecks", that is, poor search and recommendation mechanisms that tend to hide the content (see, e.g., [Brodersen et al. 2012; Cha et al. 2009]). More specifically, properties, such as, the geographic locality of interest in online videos, i.e., views arising from a confined spatial area, the geographic relevance and the proximity between users, have a strong impact on video popularity. Hence, the analysis of these geographic patterns is useful for devising caching mechanisms aimed at improving delivery performance and for customizing the content presented to the users.

7. CLOUD WORKLOADS

The cloud computing paradigm with its promise of lower costs and better efficiency opens interesting challenges related to the effective management of the technological infrastructures, the optimization of their performance and the reduction of the operational costs [Armbrust et al. 2010]. Virtualization technologies are at the core of this paradigm in that they allow the deployment of abstract shared resources. Although these concepts date back to the 70's (see, e.g., [Goldberg 1974]), nowadays virtualization benefits of a much greater flexibility obtained at the expenses of an increased complexity. Hence, for achieving an optimal tradeoff among cloud performance, QoS requirements and Service Level Agreement obligations, it is important to gain a solid understanding of the characteristics and dynamics of the service being deployed in the clouds, that is, their workloads. Moreover, this knowledge is fundamental for performance engineering studies dealing with the capacity planning of the infrastructures as well as for devising resource and service allocation strategies and consolidation policies aimed at energy saving [Jennings and Stadler 2014].

In the literature, the characterization of cloud workloads has been recently addressed by focusing on various aspects related to their resource usage. The basic workload components considered in these studies refer to the applications, e.g., jobs, tasks, submitted by the cloud users, and to the Virtual Machines (VM) instantiated by cloud providers for the deployment of these applications. Workload measurements are collected at various levels of detail by instrumenting the cloud infrastructures and applying customized monitoring and profiling solutions [Weingärtner et al. 2015]. For example, the Virtual Machine Monitor captures low level events, e.g., page faults, disk and network IO accesses. For the individual VMs, the measures refer to demands of the physical and virtual resources, e.g., processor, memory. Let us remark that, to reduce the intrusiveness and overhead of the monitoring process, sampling techniques are frequently applied.

To investigate the characteristics and behaviors of cloud workloads and assess their impact on scheduling and resource management policies, several papers focus on measurements made recently available by some cloud providers (see, e.g., [Chen et al. 2010; Di et al. 2014; Ghorbani et al. 2014; Liu and Cho 2012; Mishra et al. 2010; Solis Moreno et al. 2014; Reiss et al. 2012; Sharma et al. 2011]). In particular, from the analysis of the Google cluster datasets the heterogeneity and dynamicity of the tasks are assessed in terms of their resource usage as a function of time. In addition, to discover similarities in the workload composition, jobs and tasks are classified according to high level events, such as, submission, schedule, eviction, fail, kill, as well as attributes related to their execution e.g., length, resource usage. Furthermore, starting from the observation that users are responsible for driving the workload in terms of task volume and resource requirements, task models are complemented by models of the user behavior. More specifically, clustering techniques are applied to quantify the diversity of user behavioral patterns, e.g., submission rate, CPU and memory demands, and uncover the relationships between user behavior and tasks. From the analysis of the cluster composition combined with fitting techniques, the parameters of the statistical distributions describing the characteristics of both tasks and users are derived.

Let us remark that to devise efficient resource allocation policies for cloud environments, it is important to take into consideration the fluctuations and time-varying characteristics of the arrival patterns of their workloads. Markovian models work well for this purpose in that they provide a compact description of these phenomena and allow for capturing their heavy-tailed and bursty nature [Casale et al. 2012; Pacheco-Sanchez et al. 2011; Yin et al. 2014]. These models reveal particularly useful for generating synthetic workloads to be used for simulation and benchmarking experiments.

The comparison of cloud and Grid workloads presented in [Di et al. 2012] shows that their characteristics are significantly different especially with respect to job length and priority, submission frequency and resource utilization. Moreover, the jobs deployed in the cloud have usually lower resource usage for CPU and memory than Grid jobs because of their interactive and real-time nature in contrast to the batch nature of the scientific jobs typically deployed in Grid environments.

In the framework of virtualization technologies, workload characterization focuses on issues related to demand and usage patterns of the VMs with the aim of deriving predictive models to be used for virtual resource provisioning and for evaluating cloud performance (see, e.g., [Azmandian et al. 2011; Birke et al. 2013; Calheiros et al. 2014; Gmach et al. 2007; Iosup et al. 2011; Khan et al. 2012; Kochut and Beaty 2007; Wolski and Brevik 2014; Wood et al. 2008]). These issues are addressed under different perspectives by considering both static and dynamic properties of VM workload. In particular, the characteristics of this type of load, e.g., interarrival time of the requests issued by cloud users, VM lifetime, number of VMs in the request, number of CPU cores requested for each VM, are modeled by statistical distributions obtained through the application of fitting techniques. Conversely, the life cycles of the Virtual Machines are described in terms of their ON/OFF activity, namely, the duration and temporal patterns of ON/OFF states, whose dynamics offers valuable insights on the virtualization deployment. It has been shown that, in spite of the large flexibility provided by these technologies, in private clouds a significant percentage of VMs are never turned off. Alternative approaches for the characterization of VM workloads are based on time series and spectral analysis exploited on historical data. These approaches allow for devising models able to capture and predict the peculiarities of the resource demands, e.g., periodicity, fluctuations. Similarly, Markovian models are applied to characterize the temporal correlations across VMs and predict the variations of the corresponding workload patterns.

Furthermore, it is worth mentioning another emerging deployment of virtualization technologies, namely, desktop cloud. Understanding the peculiarities of the workloads associated with this new delivery model is very important for devising mechanisms aimed at improving the efficiency of virtual desktops (see, e.g., [Jiang et al. 2012; Kochut et al. 2010]). These workloads are analyzed in terms of low level properties, such as, instructions per clock cycle, cache and TLB misses, CPU and memory usage. In particular, the autocorrelations of the resource usage highlight the dynamic properties of the workloads at the level of individual and aggregated desktops.

Another important standpoint taken into account in the characterization of cloud workloads is the network traffic exploited within and between data centers (see, e.g., [Benson et al. 2010; Chen et al. 2011; Kandula et al. 2009]). More specifically, traffic dynamics are investigated from macroscopic viewpoints by looking at communication patterns between servers and across data centers, and from microscopic viewpoints by focusing on the traffic characteristics at flow and packet levels. Let us remark that the analysis of the temporal and spatial variations of attributes, such as, flow duration and size, flow and packet interarrival times, packet size, allows for assessing their impact on network performance and the evaluation of traffic engineering strategies.

Similarly, the workloads of personal cloud storage systems, such as, Dropbox[19], are characterized from the network traffic viewpoint by focusing on the control and data storage flows, such as, notifications, retrieve operations, store operations (see, e.g., [Drago et al. 2012; Gonçalves et al. 2014]). A deep knowledge of the patterns of this workload type is very useful for designing cost effective solutions. Flow properties, e.g., Round Trip Time, flow size, number of chunks per batch, describe the traffic patterns, whereas attributes, e.g., daily usage patterns, session duration, intersession time, storage volume, number of active devices, characterize the usage patterns. Moreover, to further investigate the behavior of the users, their workload generation process is captured by applying a hierarchical approach that takes into account the relationships between user sessions and the corresponding data transmission flows.

8. CONCLUSIONS

The deployment of sophisticated services and the increased pervasiveness of mobile devices always connected to the Internet produce rapid changes in workload intensity and characteristics. To meet user expectations as well as QoS requirements and SLA obligations, system designers and providers need to cope with these issues by taking into account the fundamental properties and patterns of the workloads as well as the user behavior.

[19]http://www.dropbox.com

Workload characterization plays a key role in many performance engineering studies. As witnessed by the very large number of papers appeared in the literature, workload characterization is a live research field that has continuously evolved during the past four decades to exploit the new technological developments.

In this paper, we presented the state of the art of workload characterization by focusing on some application scenarios selected according to their relevance and popularity. In particular, we examined the distinctive characteristics of the workloads associated with Web, social network services, mobile devices, video services and cloud computing infrastructures. The peculiarities of each workload type have been analyzed from various perspectives by considering both qualitative and quantitative aspects related to the technological infrastructures as well as the interactions of the users with services and applications. It is interesting to point out that despite the complexity of these new workload types, the methodological approaches and the modeling techniques adopted for traditional workloads have been successfully exploited for the characterization of the workloads of the scenarios considered in this survey. Moreover, we have observed that a major issue faced by workload characterization is the lack of publicly available workload data. This issue has become even more critical nowadays due to the bigger complexity of the technological infrastructures and the richer user interactions. In general, providers are very reluctant to disclose data of their workloads to prevent leakage of confidential and competitive information. Nevertheless, understanding the behavior of their own workloads would be very beneficial in many respects, e.g., for devising cost effective solutions. In addition, this will allow for significant advancements of the scientific knowledge.

The presence of workloads with unique, nontrivial and not fully understood properties makes workload characterization a research topic of timely relevance. Open research issues deal with the workloads of emerging scenarios, such as, fog and vehicular clouds, big data, anonymous social networks, sensor networks, to name a few.

REFERENCES

E. Adar, J. Teevan, S.T. Dumais, and J.L. Elsas. 2009. The Web Changes Everything: Understanding the Dynamics of Web Content. In *Proc. of the 2nd ACM Int. Conf. on Web Search and Data Mining - WSDM'09*. ACM, 282–291.

Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. 2007. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proc. of the 16th Int. Conf. on World Wide Web - WWW'07*. ACM, 835–844.

V. Akula and D.A. Menascé. 2007. Two-level workload characterization of online auctions. *Electronic Commerce Research and Applications* 6, 2 (2007), 192–208.

M. Allamanis, S. Scellato, and C. Mascolo. 2012. Evolution of a Location-based Online Social Network: Analysis and Models. In *Proc. of the 12th ACM SIGCOMM Conf. on Internet Measurement - IMC'12*. ACM, 145–158.

J.M. Almeida, J. Krueger, D.L. Eager, and M.K. Vernon. 2001. Analysis of Educational Media Server Workloads. In *Proc. of the 11th Int. Workshop on Network and Operating Systems Support for Digital Audio and Video - NOSSDAV'01*. ACM, 21–30.

R.B. Almeida, B. Mozafari, and J. Cho. 2007. On the Evolution of Wikipedia. In *Proc. of the 1st Int. AAAI Conf. on Weblogs and Social Media - ICWSM'07*.

M. Arlitt, D. Krishnamurthy, and J. Rolia. 2001. Characterizing the Scalability of a Large Web-based Shopping System. *ACM Transactions on Internet Technology* 1, 1 (2001), 44–69.

M. Arlitt and C. Williamson. 1996. Web Server Workload Characterization: The Search for Invariants. In *Proc. of the ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems*. ACM, 126–137.

Martin F. Arlitt and Carey L. Williamson. 1997. Internet Web Servers: Workload Characterization and Performance Implications. *IEEE/ACM Transactions on Networking* 5, 5 (1997), 631–645.

M. Armbrust et al. 2010. A View of Cloud Computing. *Commununications of the ACM* 53, 4 (2010), 50–58.

F. Azmandian, M. Moffie, J.G. Dy, J.A. Aslam, and D.R. Kaeli. 2011. Workload Characterization at the Virtualization Layer. In *Proc. of the 19th Int. Symp. on Modeling, Analysis Simulation of Computer and Telecommunication Systems - MASCOTS'11*. IEEE, 63–72.

F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. 2009. Characterizing User Behavior in Online Social Networks. In *Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement - IMC'09*. ACM, 49–62.

F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. 2012. Characterizing user navigation and interactions in online social networks. *Information Sciences* 195 (2012), 1–24.

T. Benson, A. Akella, and D.A. Maltz. 2010. Network Traffic Characteristics of Data Centers in the Wild. In *Proc. of the 10th ACM SIGCOMM Conf. on Internet Measurement - IMC'10*. ACM, 267–280.

L. Bent, M. Rabinovich, G.M. Voelker, and Z. Xiao. 2006. Characterization of a Large Web Site Population with Implications for Content Delivery. *World Wide Web Journal* 9, 4 (2006), 505–536.

C. Bernardini, T. Silverston, and O. Festor. 2014. A Pin is Worth a Thousand Words: Characterization of Publications in Pinterest. In *Proc. of the 5th Int. Workshop on TRaffic Analysis and Characterization - TRAC'14*. IEEE, 322–327.

R. Birke, A. Podzimek, L.Y. Chen, and E. Smirni. 2013. State-of-the-practice in data center virtualization: Toward a better understanding of VM usage. In *Proc. of the 43rd Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks - DSN'13*. 1–12.

M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. 2011. Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage. In *Proc. of the 13th Int. Conf. on Human Computer Interaction with Mobile Devices and Services - MobileHCI'11*. ACM, 47–56.

Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. 2012. The Untold Story of the Clones: Content-agnostic Factors That Impact YouTube Video Popularity. In *Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD'12*. ACM, 1186–1194.

Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. 2011. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation* 68, 11 (2011), 1037–1055.

G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. 2008. *Time Series Analysis - Forecasting and Control* (Fourth ed.). Wiley.

U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. 2009. Network Analysis of Collaboration Structure in Wikipedia. In *Proc. of the 18th Int. Conf. on World Wide Web - WWW'09*. ACM, 731–740.

B.E. Brewington and G. Cybenko. 2000. How Dynamic is the Web? *Computer Networks* 33, 1–6 (2000), 257–276.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 2000. Graph structure in the Web. *Computer Networks* 33, 1–6 (2000), 309–320.

A. Brodersen, S. Scellato, and M. Wattenhofer. 2012. YouTube Around the World: Geographic Popularity of Videos. In *Proc. of the 21st Int. Conf. on World Wide Web - WWW'12*. ACM, 241–250.

M. Butkiewicz, H.V. Madhyastha, and V. Sekar. 2011. Understanding website complexity: Measurements, metrics, and implications. In *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - IMC'11*. ACM, 313–328.

R. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya. 2014. Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications' QoS. *IEEE Transactions on Cloud Computing* (2014). DOI:http://dx.doi.org/10.1109/TCC.2014.2350475

M. Calzarossa and D. Ferrari. 1986. A sensitivity study of the clustering approach to workload modeling. *Performance Evaluation* 6 (1986), 25–33.

M. Calzarossa, G. Haring, and G. Serazzi. 1988. On workload modeling for computer networks. In *Architektur und Betrieb von Rechensystemen*, U. Kastens and F.J. Rammig (Eds.). Springer-Verlag, 324–339.

M. Calzarossa and L. Massari. 2012. Temporal analysis of crawling activities of commercial Web robots. In *Computer and Information Sciences III*, E. Gelenbe and R. Lent (Eds.). Lecture Notes in Electrical Engineering, Vol. 264. Springer, 429–436.

M. Calzarossa, L. Massari, and D. Tessera. 2000. Workload Characterization: Issues and Methodologies. In *Performance Evaluation - Origins and Directions (Lecture Notes in Computer Science)*, G. Haring, C. Lindemann, and M. Reiser (Eds.), Vol. 1769. Springer, 459–484.

M. Calzarossa, L. Massari, and D. Tessera. 2013. An extensive study of Web robots traffic. In *Proc. of Int. Conf. on Information Integration and Web-based Applications & Services - IIWAS'13*. ACM, 410–4417.

M. Calzarossa and G. Serazzi. 1993. Workload Characterization: a Survey. *Proc. of the IEEE* 8, 81 (1993), 1136–1150.

M. Calzarossa and D. Tessera. 2008. Characterization of the evolution a news Web site. *Journal of Systems and Software* 81, 12 (2008), 2236–2344.

M. Calzarossa and D. Tessera. 2010. An exploratory analysis of the novelty of a news Web site. In *Proc. of the Int. Symp. on Performance Evaluation of Computer and Telecommunication Systems - SPECTS 2010*. IEEE, 399–404.

M. Calzarossa and D. Tessera. 2015. Modeling and Predicting Temporal Patterns of Web Content Changes. *Journal of Network and Computer Applications* (2015).

G. Casale, Ningfang M., L. Cherkasova, and E. Smirni. 2012. Dealing with Burstiness in Multi-Tier Applications: Models and Their Parameterization. *IEEE Transactions on Software Engineering* 38, 5 (2012), 1040–1053.

M. Cataldi and M.A. Aufaure. 2014. The 10 million follower fallacy: audience size does not prove domain-influence on Twitter. *Knowledge and Information Systems* (2014). DOI:http://dx.doi.org/10.1007/s10115-014-0773-8

M. Cha, F. Benevenuto, Y.-Y. Ahn, and K.P. Gummadi. 2012a. Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Networks* 56, 3 (2012), 1066–1076.

M. Cha, F. Benevenuto, H. Haddadi, and K.P. Gummadi. 2012b. The World of Connections and Information Flow in Twitter. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 42, 4 (2012), 991–998.

M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. 2007. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. of the 7th ACM SIGCOMM Conf. on Internet Measurement - IMC'07*. ACM, 1–14.

M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. 2009. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking* 17, 5 (2009), 1357–1370.

M. Cha, A. Mislove, B. Adams, and K.P. Gummadi. 2008. Characterizing Social Cascades in Flickr. In *Proc. of the 1st Workshop on Online Social Networks - WOSN'08*. ACM, 13–18.

M. Cha, A. Mislove, and K.P. Gummadi. 2009. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proc. of the 18th Int. Conf. on World Wide Web - WWW'09*. ACM, 721–730.

G. Chatzopoulou, Cheng S., and M. Faloutsos. 2010. A First Step Towards Understanding Popularity in YouTube. In *Proc. IEEE INFOCOM Conf. on Computer Communications Workshops*. 1–6.

L. Chen, Y. Zhou, and D.M. Chiu. 2014. A study of user behavior in online VoD services. *Computer Communications* 46 (2014), 66–75.

Y. Chen, A.S. Ganapathi, R. Griffith, and R.H. Katz. 2010. *Analysis and Lessons from a Publicly Available Google Cluster Trace*. Technical Report UCB/EECS-2010-95. Electrical Engineering and Computer Sciences, University of California at Berkeley.

Y. Chen, S. Jain, V.K. Adhikari, Z. Zhang, and K. Xu. 2011. A First Look at Inter-Data Center Traffic Characteristics via Yahoo! Datasets. In *Proc. of the Int. Conf. on Computer Communications - INFOCOM*. IEEE, 1620–1628.

X. Cheng, J. Liu, and C. Dale. 2013. Understanding the Characteristics of Internet Short Video Sharing: A YouTube-Based Measurement Study. *IEEE Transactions on Multimedia* 15, 5 (2013), 1184–1194.

Z. Cheng, J. Caverlee, K. Lee, and D. Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. In *Proc. of the 5th Int. AAAI Conf. on Weblogs and Social Media - ICWSM'11*. 81–88.

L. Cherkasova and M. Gupta. 2004. Analysis of Enterprise Media Server Workloads: Access Patterns, Locality, Content Evolution, and Rates of Change. *IEEE/ACM Transactions on Networking* 12, 5 (2004), 781–794.

Y. Chi, S. Zhu, X. Song, J. Tatemura, and B.L. Tseng. 2007. Structural and temporal analysis of the blogosphere through community factorization. In *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD'07*. 163–172.

J. Cho and H. Garcia-Molina. 2003. Estimating frequency of change. *ACM Transactions on Internet Technology* 3, 3 (2003), 256–290.

A. Clauset, C.R. Shalizi, and M.E.J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Rev.* 51, 4 (2009), 661–703.

G. Cormode and B. Krishnamurthy. 2008. Key Differences between Web 1.0 and Web 2.0. *First Monday* 13, 6 (2008).

G. Cormode, B. Krishnamurthy, and W. Willinger. 2010. A manifesto for modeling and measurement in social media. *First Monday* 15, 9 (2010).

J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. 2010. Bridging the Gap Between Physical Location and Online Social Networks. In *Proc. of the 12th ACM Int. Conf. on Ubiquitous Computing - UbiComp'10*. ACM, 119–128.

M. Crovella and B. Krishnamurthy. 2006. *Internet measurement: infrastructure, traffic & applications.* Wiley.

M. De Choudhury, Y.-R. Lin, H. Sundaram, K.S. Candan, L. Xie, and A. Kelliher. 2010. How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?. In *Proc. of the 4th Int. AAAI Conf. on Weblogs and Social Media - ICWSM'10.* 34–41.

S. Di, D. Kondo, and F. Cappello. 2014. Characterizing and modeling cloud applications/jobs on a Google data center. *The Journal of Supercomputing* 69, 1 (2014), 139–160.

S. Di, D. Kondo, and W. Cirne. 2012. Characterization and Comparison of Cloud Versus Grid Workloads. In *Proc. of the 2012 Int. Conf. on Cluster Computing - CLUSTER'12.* IEEE, 230–238.

M.D. Dikaiakos, A. Stassopoulou, and L. Papageorgiou. 2005. An investigation of web crawler behavior: characterization and metrics. *Computer Communications* 28, 8 (2005), 880–897.

Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose. 2011. Broadcast Yourself: Understanding YouTube Uploaders. In *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - IMC'11.* ACM, 361–370.

T.M.T. Do and D. Gatica-Perez. 2014. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing* 12 (2014), 79–91.

D. Doran and S.S. Gokhale. 2011. Web robot detection techniques: Overview and limitations. *Data Mining and Knowledge Discovery* 22, 1–2 (2011), 183–210.

D. Doran, K. Morillo, and S.S. Gokhale. 2013. A Comparison of Web Robot and Human Requests. In *Proc. of the 2013 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining - ASONAM'13.* ACM, 1374–1380.

I. Drago, M. Mellia, M.M. Munafò, A. Sperotto, R. Sadre, and A. Pras. 2012. Inside Dropbox: Understanding Personal Cloud Storage Services. In *Proc. of the 12th ACM SIGCOMM Conf. on Internet Measurement - IMC'12.* ACM, 481–494.

N.R. Draper and H. Smith. 1998. *Applied Regression Analysis* (Third ed.). Wiley.

F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. 2007. Traffic Characteristics and Communication Patterns in Blogosphere. In *Proc. of the 1st Int. AAAI Conf. on Weblogs and Social Media - ICWSM'07.*

A.A. Eldin, A. Rezaie, A. Mehta, S. Razroev, S.S. Luna, O. Seleznjev, J. Tordsson, and E. Elmroth. 2014. How will your workload look like in 6 years? Analyzing Wikimedia's workload. In *Proc. of the 2014 Int. Conf. on Cloud Engineering - IC2E'14.* IEEE, 349–354.

H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. 2010. Diversity in Smartphone Usage. In *Proc. of the 8th Int. Conf. on Mobile Systems, Applications and Services - MobiSys'10.* ACM, 179–194.

R. Farahbakhsh, X. Han, A. Cuevas, and N. Crespi. 2013. Analysis of publicly disclosed information in Facebook profiles. In *Proc. of the IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining - ASONAM'13.* 699–705.

D.G. Feitelson. 2015. *Workload Modeling for Computer Systems Performance.* Cambridge University Press.

D. Ferrari. 1972. Workload Characterization and Selection in Computer Performance Measurement. *Computer* 5, 4 (1972), 18–24.

D. Ferrari. 1984. On the Foundations of Artificial Workload Design. In *Proc. of the ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems.* ACM, 8–14.

D. Ferrari, G. Serazzi, and A. Zeigner. 1983. *Measurement and Tuning of Computer Systems.* Prentice-Hall.

D. Fetterly, M. Manasse, M. Najork, and J. Wiener. 2004. A Large-Scale Study of the Evolution of Web Pages. *Software: Practice & Experience* 34, 2 (2004), 213–237.

F. Figueiredo, J.M. Almeida, M.A. Gonçalves, and F. Benevenuto. 2014. On the Dynamics of Social Media Popularity: A YouTube Case Study. *ACM Transaction on Internet Technology* 14, 4 (2014), 24:1–24:23.

F. Figueiredo, F. Benevenuto, and J.M. Almeida. 2011. The Tube over Time: Characterizing Popularity Growth of YouTube Videos. In *Proc. of the 4th ACM Int. Conf. on Web Search and Data Mining - WSDM'11.* ACM, 745–754.

W. Fischer and K. Meier-Hellstern. 1993. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation* 18, 2 (1993), 149–171.

X. Fu and Y. Shen. 2014. Study of collective user behaviour in Twitter: a fuzzy approach. *Neural Computing and Applications* 25, 7–8 (2014), 1603–1614.

M. Gabielkov, A. Rao, and A. Legout. 2014. Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph. In *Proc. of the ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems.* ACM, 277–288.

G. Gan, C. Ma, and J. Wu. 2007. *Data Clustering: Theory, Algorithms, and Applications*. SIAM.

Q. Gao, F. Abel, G.J. Houben, and Y. Yu. 2012. A Comparative Study of Users' Microblogging Behavior on Sina Weibo and Twitter. In *User Modeling, Adaptation, and Personalization*, J. Masthoff, B. Mobasher, M.C. Desmarais, and R. Nkambou (Eds.). Lecture Notes in Computer Science, Vol. 7379. Springer, 88–101.

R. García, X.G. Pañeda, V. García, D. Melendi, and M. Vilas. 2007. Statistical characterization of a real video on demand service: User behaviour and streaming-media workload analysis. *Simulation Modelling Practice and Theory* 15, 6 (2007), 672–689.

M. Ghorbani, Y. Wang, Y. Xue, M. Pedram, and P. Bogdan. 2014. Prediction and Control of Bursty Cloud Workloads: A Fractal Framework. In *Proc. of the 2014 Int. Conf. on Hardware/Software Codesign and System Synthesis - CODES'14*. ACM, 12:1–12:9.

E. Gilbert, S. Bakhshi, S. Chang, and L. Terveen. 2013. "I Need to Try This!": A Statistical Overview of Pinterest. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems - CHI'13*. ACM, 2427–2436.

P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson. 2011. Characterizing Organizational Use of Web-Based Services: Methodology, Challenges, Observations, and Insights. *ACM Transactions on the Web* 5, 4 (2011), 19:1–19:23.

P. Gill, M. Arlitt, Z. Li, and A. Mahanti. 2007. YouTube Traffic Characterization: A View from the Edge. In *Proc. of the 7th ACM SIGCOMM Conf. on Internet Measurement - IMC'07*. ACM, 15–28.

M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. 2011. Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas Communications* 29, 9 (2011), 1872–1892.

D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper. 2007. Workload Analysis and Demand Prediction of Enterprise Data Center Applications. In *Proc. of the Int. Symp. on Workload Characterization - IISWC'07*. IEEE, 171–180.

M. Goldberg, M. Magdon-Ismail, S. Kelley, and K. Mertsalov. 2009. Stable Statistics of the Blogograph. In *Protecting Persons While Protecting the People*, C.S. Gal, P.B. Kantor, and M.E. Lesk (Eds.). Lecture Notes in Computer Science, Vol. 5661. Springer, 104–114.

R.P. Goldberg. 1974. Survey of Virtual Machine Research. *Computer* 7, 9 (1974), 34–45.

G. Gonçalves, I. Drago, A.P. Couto da Silva, A.B. Vieira, and J.M. Almeida. 2014. Modeling the Dropbox Client Behavior. In *Proc. of the Int. Conf. on Communications - ICC'14*. IEEE, 1332–1337.

M.A. Gonçalves, J.M. Almeida, L.G.P. dos Santos, A.H.F. Laender, and V. Almeida. 2010. On Popularity in the Blogosphere. *IEEE Internet Computing* 14, 3 (2010), 42–49.

N.Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song. 2012. Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications using Google+. In *Proc. of the 12th ACM SIGCOMM Conf. on Internet Measurement - IMC'12*. ACM, 131–144.

V. Gopalakrishnan, R. Jana, K.K. Ramakrishnan, D.F. Swayne, and V.A. Vaishampayan. 2011. Understanding Couch Potatoes: Measurement and Modeling of Interactive Usage of IPTV at Large Scale. In *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - ICM'11*. ACM, 225–242.

L. Guo, S. Chen, Z. Xiao, and X. Zhang. 2005. Analysis of Multimedia Workloads with Implications for Internet Streaming. In *Proc. of the 14th Int. Conf. on World Wide Web (WWW'05)*. ACM, 519–528.

L. Guo, E. Tan, S. Chen, X. Zhang, and Y. Zhao. 2009. Analyzing Patterns of User Content Generation in Online Social Networks. In *Proc. of the 15th ACM Int. Conf. on Knowledge Discovery and Data Mining - KDD'09*. ACM, 369–378.

G. Gürsun, M. Crovella, and I. Matta. 2011. Describing and forecasting video access patterns. In *Proc. of IEEE INFOCOM*. 16–20.

J. Han, D. Choi, B.G. Chun, T. Kwon, H.C. Kim, and Y. Choi. 2014. Collecting, Organizing, and Sharing Pins in Pinterest: Interest-driven or Social-driven?. In *Proc. of the ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems*. ACM, 15–27.

S. Ihm and V.S. Pai. 2011. Towards understanding modern Web traffic. In *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - IMC'11*. ACM, 295–312.

A. Iosup, S. Ostermann, M.N. Yigitbasi, R. Prodan, T. Fahringer, and D.H.J. Epema. 2011. Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing. *IEEE Transactions on Parallel and Distributed Systems* 22, 6 (2011), 931–945.

M.A. Islam, D. Eager, N. Carlsson, and A. Mahanti. 2013. Revisiting Popularity Characterization and Modeling of User-Generated Videos. In *Proc. of the 21st Int. Symp. on Modeling, Analysis Simulation of Computer and Telecommunication Systems - MASCOTS'13*. IEEE Computer Society Press, 350–354.

A.K. Jain, M.N. Murty, and P.J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys* 31, 3 (1999), 264–323.

A. Java, X. Song, T. Finin, and B. Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. ACM, 56–65.

B. Jennings and R. Stadler. 2014. Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management* (2014). DOI:http://dx.doi.org/10.1007/s10922-014-9307-7

M. Jeon, Y. Kim, J. Hwang, J. Lee, and E. Seo. 2012. Workload Characterization and Performance Implications of Large-Scale Blog Servers. *ACM Transactions on the Web* 6, 4 (2012), 16:1–16:26.

L. Jiali, L. Zhenyu, W. Dong, K. Salamatian, and X. Gaogang. 2012. Analysis and Comparison of Interaction Patterns in Online Social Network and Social Media. In *Proc. of the 21st Int. Conf. on Computer Communications and Networks - ICCCN'12*. IEEE, 1–7.

J. Jiang, C. Wilson, X. Wang, W. Sha, P. Huang, Y. Dai, and B.Y. Zhao. 2010. Understanding Latent Interactions in Online Social Networks. In *Proc. of the 10th ACM SIGCOMM Conf. on Internet Measurement - IMC'10*. ACM, 369–382.

T. Jiang, R. Hou, L. Zhang, K. Zhang, L. Chen, M. Chen, and N. Sun. 2012. Micro-architectural Characterization of Desktop Cloud Workloads. In *Proc. of the Int. Symp. on Workload Characterization - IISWC'12*. IEEE, 131–140.

R.A. Johnson and D.W. Wichern. 2007. *Applied Multivariate Statistical Data Analysis* (Sixth ed.). Pearson Prentice Hall.

I.T. Jolliffe. 2002. *Principal Component Analysis* (Second ed.). Springer.

S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken. 2009. The Nature of Data Center Traffic: Measurements & Analysis. In *Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement - IMC'09*. ACM, 202–208.

X. Kang, H. Zhang, G. Jiang, H. Chen, X. Meng, and K. Yoshihira. 2010. Understanding Internet Video sharing site workload: A view from data center design. *Journal of Visual Communication and Image Representation* 21, 2 (2010), 129–138.

A. Khan, X. Yan, Shu Tao, and N. Anerousis. 2012. Workload characterization and prediction in the cloud: A multiple time series approach. In *Proc. of the 13th Network Operations and Management Symposium - NOMS'12*. IEEE, 1287–1294.

A. Kochut and K. Beaty. 2007. On Strategies for Dynamic Resource Management in Virtualized Server Environments. In *Proc. of the 15th Int. Symp. on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems - MASCOTS'07*. IEEE, 193–200.

A. Kochut, K. Beaty, H. Shaikh, and D.G. Shea. 2010. Desktop Workload Study with Implications for Desktop Cloud Resource Optimization. In *Proc. of the Int. Symp. on Parallel Distributed Processing, Workshops and Phd Forum - IPDPSW'10*. 1–8.

B. Krishnamurthy, P. Gill, and M. Arlitt. 2008. A Few Chirps About Twitter. In *Proc. of the 1st Workshop on Online Social Networks - WOSN'08*. ACM, 19–24.

R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. 2003. On the Bursty Evolution of Blogspace. In *Proc. of the 12th Int. Conf. on World Wide Web - WWW'03*. ACM, 568–576.

R. Kumar, J. Novak, and A. Tomkins. 2006. Structure and Evolution of Online Social Networks. In *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ACM, 611–617.

H. Kwak, C. Lee, H. Park, and S. Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proc. of the 19th Int. Conf. on World Wide Web - WWW'10*. ACM, 591–600.

C.A.C. Lampe, N. Ellison, and C. Steinfield. 2007. A Familiar Face(book): Profile Elements as Signals in an Online Social Network. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems - CHI'07*. ACM, 435–444.

A. Le, K. Pelechrinis, and P. Krishnamurthy. 2014. Country-level Spatial Dynamics of User Activity: A Case Study in Location-based Social Networks. In *Proc. of the ACM Web Science Conf. - WebSci'14*. ACM, 71–80.

J. Lee, S. Cha, D. Lee, and H. Lee. 2009. Classification of web robots: An empirical study based on over one billion requests. *Computers & Security* 28, 8 (2009), 795–802.

K. Lerman and R. Ghosh. 2010. Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Proc. of 4th Int. AAAI Conf. on Weblogs and Social Media - ICWSM'10*. 90–97.

J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. 2008. Microscopic Evolution of Social Networks. In *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD'08*. ACM, 462–470.

J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. 2007. Patterns of Cascading Behavior in Large Blog Graphs. In *Proc. of the 2007 SIAM Int. Conf. on Data Mining*. 551–556.

H. Li, X. Cheng, and J. Liu. 2014. Understanding Video Sharing Propagation in Social Networks: Measurement and Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 10, 4 (2014), 33:1–33:20.

J. Li and C. Cardie. 2014. Timeline Generation: Tracking Individuals on Twitter. In *Proc. of the 23rd Int. Conf. on World Wide Web - WWW'14*. 643–652.

J. Li and S. Ma. 2014. Characterization and modeling of video popularity. *International Journal of Communication Systems* 27 (2014), 2604–2615.

M. Li, M. Claypool, R. Kinicki, and J. Nichols. 2005. Characteristics of Streaming Media Stored on the Web. *ACM Transaction on Internet Technology* 5, 4 (2005), 601–626.

Z.-X. Liao, Y.-C. Pan, W.-C. Peng, and P.-R. Lei. 2013. On Mining Mobile Apps Usage Behavior for Predicting Apps Usage in Smartphones. In *Proc. of the 22nd ACM Int. Conf. on Information & Knowledge Management - CIKM'13*. ACM, 609–618.

T. Lins, A.C.M. Pereira, and F. Benevenuto. 2014. Workload characterization of a location-based social network. *Social Network Analysis and Mining* 4, 1 (2014), 208–221.

J. Liu, Y. Yang, Z. Huang, Y. Yang, and H.T. Shen. 2014b. On the Influence Propagation of Web Videos. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1961–1973.

N. Liu, H. Cui, S.-H.G. Chan, Z. Chen, and Y. Zhuang. 2014a. Dissecting User Behaviors for a Simultaneous Live and VoD IPTV System. *ACM Transactions on Multimedia Computing, Communications and Applications* 10, 3 (2014), 23:1–23:16.

Z. Liu and S. Cho. 2012. Characterizing Machines and Workloads on a Google Cluster. In *Proc. of the 41st Int. Conf. on Parallel Processing Workshops - ICPPW'12* . IEEE, 397–403.

A. Mahanti, N. Carlsson, A. Mahanti, M. Arlitt, and C. Williamson. 2013. A tale of the tails: Power-laws in Internet measurements. *IEEE Network* 27, 1 (2013), 59–64.

A. Mahanti, C. Williamson, and L. Wu. 2009. Workload Characterization of a Large Systems Conference Web Server. In *Proc. of the 7th Annual Conf. on Communication Networks and Services Research - CNSR'09*. IEEE, 55–64.

M. Maia, J. Almeida, and V. Almeida. 2008. Identifying User Behavior in Online Social Networks. In *Proc. of the 1st Workshop on Social Network Systems - SocialNets'08*. ACM, 1–6.

E. Malmi. 2014. Quality Matters: Usage-based App Popularity Prediction. In *Proc. of the 2014 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing - UbiComp'14 Adjunct*. ACM, 391–396.

L. Massari. 2010a. Analysis of MySpace user profiles. *Information Systems Frontiers* 12, 4 (2010), 361–367.

L. Massari. 2010b. What's inside MySpace comments?. In *Proc. of the Int. Symp. on Performance Evaluation of Computer and Telecommunication Systems - SPECTS 2010*. IEEE, 311–316.

D.A. Menascé. 2003. Workload characterization. *IEEE Internet Computing* 7, 5 (2003), 89–92.

D.A. Menascé and V.A.F. Almeida. 2001. *Capacity Planning for Web Services: Metrics, Models, and Methods*. Prentice Hall.

D.A. Menascé, V.A.F. Almeida, R. Fonseca, and M.A. Mendes. 1999. A Methodology for Workload Characterization of E-commerce Sites. In *Proc. of the 1st ACM Conf. on Electronic Commerce - EC'99*. ACM, 119–128.

D.A. Menascé, V.A.F. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. Meira Jr. 2003. A hierarchical and multiscale approach to analyze E-business workloads. *Performance Evaluation* 54, 1 (2003), 33–57.

A.K. Mishra, J.L. Hellerstein, W. Cirne, and C.R. Das. 2010. Towards Characterizing Cloud Backend Workloads: Insights from Google Compute Clusters. *ACM SIGMETRICS Performance Evaluation Review* 37, 4 (2010), 34–41.

A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM Conf. on Internet Measurement - IMC'07*. ACM, 29–42.

S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, and A. Mahanti. 2011. Characterizing Web-based video sharing workloads. *ACM Transactions on the Web* 5, 2 (2011), 8:1–8:27.

M. Mitrović and B. Tadić. 2010. Bloggers behavior and emergent communities in Blog space. *European Physical Journal B* 73, 2 (2010), 293–301.

A. Nazir, S. Raza, and C.N. Chuah. 2008. Unveiling Facebook: A Measurement Study of Social Network Based Applications. In *Proc. of the 8th ACM SIGCOMM Conf. on Internet Measurement - IMC'08*. ACM, 43–56.

C. Olston and M. Najork. 2010. Web Crawling. *Journal of Foundations and Trends in Information Retrieval* 4, 3 (2010), 175–246.

R. Ottoni, J.P. Pesce, D. Las Casas, G. Franciscani Jr., W. Meira Jr., P. Kumaraguru, and V. Almeida. 2013. Ladies First: Analyzing Gender Roles and Behaviors in Pinterest. In *Proc. of the 7th Int. AAAI Conf. on Weblogs and Social Media - ICWSM'13*. 457–465.

S. Pacheco-Sanchez, G. Casale, B. Scotney, S. McClean, G. Parr, and S. Dawson. 2011. Markovian Workload Characterization for QoS Prediction in the Cloud. In *Proc. of the 2011 IEEE Int. Conf. on Cloud Computing - CLOUD 2011*. IEEE, 147–154.

A. Patro, S. Rayanchu, M. Griepentrog, Y. Ma, and S. Banerjee. 2013. Capturing Mobile Experience in the Wild: A Tale of Two Apps. In *Proc. of the 9th ACM Conf. on Emerging Networking Experiments and Technologies - CoNEXT'13*. ACM, 199–210.

K. Pelechrinis and T. Lappas. 2014. Mining emerging user-centered network structures in location-based social networks. In *Proc. of the 6th Int. Workshop on Network Science for Communication Networks - NetSciCom'14*. 771–776.

T. Petsas, A. Papadogiannakis, M. Polychronakis, E.P. Markatos, and T. Karagiannis. 2013. Rise of the Planet of the Apps: A Systematic Study of the Mobile App Ecosystem. In *Proc. of the 13th ACM SIGCOMM Conf. on Internet Measurement - IMC'13*. ACM, 277–290.

J.E. Pitkow. 1999. Summary of WWW Characterizations. *World Wide Web* 2, 1–2 (1999), 3–13.

D. Preoţiuc-Pietro and T. Cohn. 2013. Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks. In *Proc. of the ACM Web Science Conf. - WebSci'13*. ACM, 306–315.

M.B. Priestley. 1981. *Spectral Analysis and Time Series*. Academic Press.

K. Radinsky and P.N. Bennett. 2013. Predicting Content Change on the Web. In *Proc. of the 6th ACM Int. Conf. on Web Search and Data Mining - WSDM'13*. ACM, 415–424.

C. Reiss, A. Tumanov, G.R. Ganger, R.H. Katz, and M.A. Kozuch. 2012. Heterogeneity and Dynamicity of Clouds at Scale: Google Trace Analysis. In *Proc. of the 3rd ACM Symposium on Cloud Computing - SoCC'12*. ACM, 7:1–7:13.

F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. 2009. Understanding online social network usage from a network perspective. In *Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement - IMC'09*. ACM, 35–48.

B. Sharma, V. Chudnovsky, L. Hellerstein, R. Rifaat, and C.R. Das. 2011. Modeling and synthesizing task placement constraints in Google compute clusters. In *Proc. of the 2nd ACM Symp. on Cloud Computing - SOCC'11*. ACM, 3–14.

C. Shin, J.-H. Hong, and A.K. Dey. 2012. Understanding and Prediction of Mobile Application Usage for Smart Phones. In *Proc. of the 14th ACM Int. Conf. on Ubiquitous Computing - UbiComp'12*. ACM, 173–182.

S. Siersdorfer, S. Chelaru, J.S. Pedro, I.S. Altingovde, and W. Nejdl. 2014. Analyzing and Mining Comments and Comment Ratings on the Social Web. *ACM Transactions on the Web* 8, 3 (2014), 17:1–17:39.

T. Silva, J.M. Almeida, and D. Guedes. 2011. Live streaming of user generated videos: Workload characterization and content delivery architectures. *Computer Networks* 55, 18 (2011), 4055–4068.

T.H. Silva, P.O.S. Vaz de Melo, J.M. Almeida, J. Salles, and A.A.F. Loureiro. 2013. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Proc. of the Int. Conf. on Distributed Computing in Sensor Systems - DCoSS'13*. IEEE, 123–132.

I. Solis Moreno, P. Garraghan, P. Townend, and J. Xu. 2014. Analysis, Modeling and Simulation of Workload Patterns in a Large-Scale Utility Cloud. *IEEE Transactions on Cloud Computing* 2, 2 (2014), 130–142.

A. Stassopoulou and M.D. Dikaiakos. 2009. Web robot detection: A probabilistic reasoning approach. *Computer Networks* 53, 3 (2009), 265–278.

P.-N. Tan and V. Kumar. 2002. Discovery of Web Robot Sessions Based on their Navigational Patterns. *Data Mining and Knowledge Discovery* 6, 1 (2002), 9–35.

A.L. Traud, E.D. Kelsic, P.J. Mucha, and M.A. Porter. 2011. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* 53, 3 (2011), 526–543.

K.S. Trivedi. 2002. *Probability and Statistics with Reliability, Queuing and Computer Science Applications* (Second ed.). Wiley.

G. Urdaneta, G. Pierre, and M. van Steen. 2009. Wikipedia workload analysis for decentralized hosting. *Computer Networks* 53, 11 (2009), 1830–1845.

J.W. van Dam and M. van de Velden. 2015. Online profiling and clustering of Facebook users. *Decision Support Systems* 70 (2015), 60–72.

R. van Zwol. 2007. Flickr: Who is Looking?. In *Proc. of the IEEE/ACM Int. Conf. on Web Intelligence - WI'07*. IEEE, 184–190.

M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. 2012. Tips, Dones and Todos: Uncovering User Profiles in Foursquare. In *Proc. of the 5th ACM Int. Conf. on Web Search and Data Mining - WSDM'12*. ACM, 653–662.

E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. 2006. A Hierarchical Characterization of a Live Streaming Media Workload. *IEEE/ACM Transactions on Networking* 14, 1 (2006), 133–146.

B. Viswanath, A. Mislove, M. Cha, and K.P. Gummadi. 2009. On the Evolution of User Interaction in Facebook. In *Proc. of the 2nd Workshop on Online Social Networks - WOSN'09*. ACM, 37–42.

M. Wattenhofer, R. Wattenhofer, and Z. Zhu. 2012. The YouTube Social Network. In *Proc. of the 6th Int. AAAI Conf. on Weblogs and Social Media - ICWSM'12*. 354–361.

R. Weingärtner, G.B. Bräscher, and C.B. Westphall. 2015. Cloud resource management: A survey on forecasting and profiling models. *Journal of Network and Computer Applications* 47 (2015), 99–106.

D.B. West. 2001. *Introduction to Graph Theory* (Second ed.). Prentice Hall.

A. Williams, M. Arlitt, C. Williamson, and K. Barker. 2005. Web Workload Characterization: Ten Years Later. In *Web Content Delivery*, X. Tang, J. Xu, and S.T. Chanson (Eds.). Web Information Systems Engineering and Internet Technologies Book Series, Vol. 2. Springer, 3–21.

C. Wilson, A. Sala, K.P.N. Puttaswamy, and B.Y. Zhao. 2012. Beyond Social Graphs: User Interactions in Online Social Networks and Their Implications. *ACM Transactions on the Web* 6, 4 (2012), 17:1–17:31.

R. Wolski and J. Brevik. 2014. Using Parametric Models to Represent Private Cloud Workloads. *IEEE Transactions on Services Computing* 7, 4 (2014), 714–725.

T. Wood, L. Cherkasova, K. Ozonat, and P. Shenoy. 2008. Profiling and Modeling Resource Usage of Virtualized Applications. In *Proc. of the 9th ACM/IFIP/USENIX Int. Conf. on Middleware - Middleware'08*. Springer, 366–387.

Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman. 2011. Identifying Diverse Usage Behaviors of Smartphone Apps. In *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - IMC'11*. ACM, 329–344.

Q. Yan, L. Wu, and L. Zheng. 2013. Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and its Applications* 392, 7 (2013), 1712–1723.

T. Yan, D. Chu, D. Ganesan, A. Kansal, and J. Liu. 2012. Fast App Launching for Mobile Devices Using Predictive User Context. In *Proc. of the 10th Int. Conf. on Mobile Systems, Applications, and Services - MobiSys'12*. ACM, 113–126.

J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng. 2015. Characterizing User Behavior in Mobile Internet. *IEEE Transactions on Emerging Topics in Computing* 3, 1 (2015), 95–106.

Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. 2010. Understanding Retweeting Behaviors in Social Networks. In *Proc. of the 19th Int. Conf. on Information and Knowledge Management - CIKM'10*. ACM, 1633–1636.

J. Yin, X. Lu, X. Zhao, H. Chen, and X. Liu. 2014. BURSE: A Bursty and Self-similar Workload Generator for Cloud Computing. *IEEE Transactions on Parallel and Distributed Systems* (2014). DOI:http://dx.doi.org/10.1109/TPDS.2014.2315204

H. Yu, D. Zheng, B.Y. Zhao, and W. Zheng. 2006. Understanding User Behavior in Large-scale Video-on-demand Systems. In *Proc. of the 1st ACM SIGOPS/EuroSys European Conf. on Computer Systems - EuroSys'06*. ACM, 333–344.

M. Zink, K. Suh, Y. Gu, and J. Kurose. 2009. Characteristics of YouTube network traffic at a campus network - Measurements, models, and implications. *Computer Networks* 53, 4 (2009), 501–514.