Modeling the Dropbox Client Behavior

Glauber Gonçalves*, Idilio Drago, Ana Paula Couto da Silva*, Alex Borges Vieira‡, Jussara M. Almeida

*Universidade Federal de Minas Gerais, Brazil ‡Universidade Federal de Juiz de Fora, Brazil

ggoncalves@dcc.ufmg.br; idiliod@gmail.com; ana.coutosilva@dcc.ufmg.br; alex.borges@ufjf.edu.br; jussara@dcc.ufmg.br

***Abstract*—Cloud storage systems are currently very popular, generating a large amount of Internet traffic. Indeed, many companies offer this kind of service, including worldwide providers such as Dropbox, Microsoft and Google. These companies as well as new providers entering the cloud storage market could greatly benefit from knowing typical workload patterns that their services have to face in order to develop more cost-effective solutions. However, despite recent analyses of typical usage patterns and possible performance bottlenecks, no previous work investigated the underlying client processes that generate workload to the system. In this context, this paper proposes a hierarchical two-layer model for representing the Dropbox client behavior. We characterize the statistical parameters of the model using passive measurements gathered in 3 different network vantage points. Our contributions can be applied to support the design of realistic synthetic workloads, thus helping in the development and evaluation of new, well-performing personal cloud storage services.**

I. Introduction

Cloud computing [1] attracts a large interest from both industry and academia, serving as architectural platform for a variety of services. Cloud storage services [2], in particular, are gaining popularity among both domestic and enterprise users as a simple, practical and safe mechanism to store data. Such popularity continues to increase with the recent entrance of big players, such as Google and Microsoft, to the cloud storage market. As consequence, the volume of traffic generated by cloud storage applications is growing at a fast rate. For example, Dropbox, currently the most popular cloud storage provider, claims to serve 1 billion file uploads in a daily basis.[1]

Both established providers and new players entering the cloud storage market need a deep understanding of the typical workload patterns that cloud storage services have to face in order to develop cost-effective solutions. However, several aspects make the analysis of cloud storage services a challenge. As the stored content is private and synchronization protocols are mostly proprietary, the knowledge of how these applications work is limited. Moreover, the use of encryption for both data and control messages makes the analysis of such services a hard task. Thus, despite their high popularity, only recent works have started analyzing characteristics of cloud storage services [3], [4], focusing either on architectural design aspects [5], data security and privacy related issues [6], or benchmark-driven performance studies [7], [8], [9]. Although the typical usage and possible performance bottlenecks of Dropbox have been investigated in [2], a characterization of underlying client processes that generate workload to the system is still lacking. Such knowledge is key to drive future system optimizations as well as the design of similar services.

[1]https://www.dropbox.com/news/company-info

This paper proposes a two-layer hierarchical model that represents the behavior of clients in successive Dropbox sessions. The higher *session level* captures the multiple Dropbox sessions that a client may have in a given period, whereas the lower *data transmission level* captures the client interactions with Dropbox servers while it stores or retrieves files during a session. We then characterize a list of statistical parameters for the model at each level, including: (i) session durations and inter-session times; (ii) the number of data transfers per client session; (iii) data transfer durations (i.e., *On* times); (iv) the time between consecutive transfers within a single session (i.e., *Off* times); (v) the number of data flows per transfer; and (vi) flow durations and transfer volume. We learn the statistical parameters of our model by analyzing datasets consisting of network traffic generated by Dropbox in three different university campuses. Key observations from our analysis of workload patterns include:

- Typically, Dropbox sessions last for only some minutes and can be well modeled by the Weibull distribution.
- A non-negligible number of sessions without any data transmissions can be seem in the network. For example, almost 80 % of client sessions present no data transmission.
- When users start 2 consecutive sessions, they do it in a relatively short period. In this case, about 80 % of consecutive sessions from the same client occur within a few minutes.

Our proposed model, as well as the results of characterizing the model parameters from network traces, provide data to support the generation of synthetic workloads. Our contributions, therefore, can help both in the evaluation of existing cloud storage services and in the design of new cloud storage applications.

The remainder of this paper is organized as follows. We discuss further related work in Section II. The essential background on Dropbox, our proposed client behavior model and our data collection methodology are presented in Section III. Section IV presents the characterization of our model parameters. Finally, our findings, their implications and directions for future work are offered in Section V.

II. Related Work

This work studies the *underlying client processes* that generate workload in personal cloud storage services, complementing various recent related efforts. For example, the authors of [2] present an extensive characterization of Dropbox, describing typical usage, traffic patterns, and possible performance bottlenecks. Our work relies on the methodology of [2] to collect data about Dropbox usage, and to understand its client. However, unlike [2], we propose a model

of client behavior, characterizing model parameters from passive measurements, and shedding light on the statistical distributions governing the workload of cloud storage systems.

Other previous efforts analyze specific cloud storage solutions [3] or compare alternative providers [8], focusing on aspects related to performance, security and privacy of cloud storage. For example, the system architecture and synchronization performance of 5 popular services are evaluated in [7], while Hu *et al.* [10] study the backup and restore performance as well as privacy related issues of 4 cloud storage services. Gracia-Tinedo *et al.* [4] present an active measurement study of 3 different systems, providing statistical distributions that model various key performance aspects, such as transfer speed and failure rate. None of these prior studies characterize client behavior and how it affects the workload on the system.

Finally, some other related studies [11], [12] note the existence of performance bottlenecks in cloud storage services, and propose new mechanisms to overcome such limitations. Our work provides new elements that can be used to develop realistic synthetic workloads, thus contributing to the efforts to develop new, well-performing cloud storage services.

III. Background and Methodology

In this section, we first briefly review the background on Dropbox (Section III-A). Then, we present the hierarchical model we propose to characterize its client behavior (Section III-B). Finally, we present the methodology adopted to collect our datasets of Dropbox usage, which are used to learn the parameters of our model (Section III-C).

A. Dropbox Background

Dropbox is currently one of the major players in the cloud storage market. According to the Google Trends,[2] the volume of searches for Dropbox has surpassed the search for other similar services since 2010, suggesting that Dropbox is currently the most widely used cloud storage service. The volume of traffic generated by the application is also increasing at a fast rate. For example, as reported in [2], Dropbox already accounts for about 4 % of the total traffic in some networks (i.e., around one third of the YouTube traffic). Given its current importance, we here focus only on Dropbox in our analyses.

Two major components can be identified in Dropbox architecture: (i) *control* servers, which are controlled by Dropbox; and (ii) *data storage* servers, which are outsourced to Amazon. Hence, Dropbox stores client files always in the Amazon cloud. In both cases, sub-domains of `dropbox.com` are used to identify the different parts of the service offering a specific functionality.

Files transferred between Dropbox clients and servers are compressed on the client side in order to reduce transfer time [10]. Similarly, only the difference between 2 consecutive versions of the same file is exchanged, and duplicated files are transferred only once. Finally, all transfers are encrypted with TLS/SSL. We refer to [2], [7] for more information about the Dropbox protocol as well as for an analysis of capabilities found in the Dropbox client.

[2]http://www.google.com/trends/

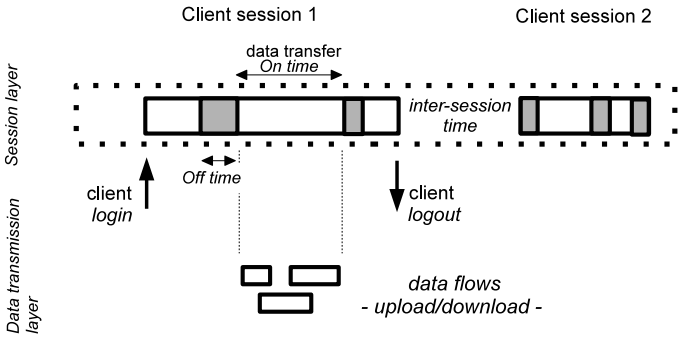

Fig. 1. Dropbox client behavior model.

B. Hierarchical Model of Client Behavior

In order to characterize the Dropbox client synchronization behavior, we propose a two-level hierarchical model to deconstruct the observed workload into a collection of client sessions, further breaking each of them into a sequence of data flows. Figure 1 provides a graphical view of our proposed client behavior model.

A client session starts with the *login action* from a particular device, identified by an IP address,[3] and ends with the *logout action*. We refer to the time between 2 consecutive sessions from the same IP address as *inter-session time*. A Dropbox client keeps continuously opened a TCP connection to a notification server (e.g., `notify1.dropbox.com`) used for receiving information about changes performed elsewhere. Moreover, clients start *data transfers* always over another TCP connection. Hence, IP addresses of notification servers can be used to identify the client presence in the network.

During a session, the client alternates between data transfers and idle times. Data transfers start with the client contacting a specific Dropbox sub-domain to open the file synchronization process. A data transfer is further broken into multiple data flows that start within a very short time interval after the file synchronization startup. Furthermore, idle data flows are kept open waiting for possible new files only for a short time interval (i.e., 60 seconds). Thus, we here define a minimum time threshold between consecutive data flows to identify different data transfers within the same client session: consecutive flows from the same IP address within a time interval below 60 seconds are grouped into a single data transfer. The total synchronization time during a data transfer is referred to as *On Time*, whereas the time interval between consecutive data transfers within the same session is referred to as *intra-session Time* (or *Off Time*).

In sum, our client behavior model has several components. At the higher session layer, client behavior is characterized in terms of session duration, inter-session time, number of data transfers as well as *On* and *Off* times. At the lower data transmission layer, the number of flows per data transfer as well as flow duration and volume are the important parameters. Next, we discuss how we collect data about Dropbox usage and how we use these data to infer the parameters of our model.

[3]Note that, in presence of Network Address Translation (NAT), we are not able to distinguish sessions from different devices sharing the same IP address.

TABLE I. DATASET OVERVIEW.

Name	Total traffic	Dropbox traffic	Period
Campus 1	526.297 TB	12.193 TB	Mar 6th - May 9th 2013
Campus 2	38.864 TB	1.296 TB	Fev 19th - Mar 14th 2013
Campus 3	30.839 TB	0.655 TB	Mar 6th - May 6th 2013

TABLE II. TOTAL DROPBOX TRAFFIC.

Name	# Unique IPs	# Sessions	# Data Flows	Volume Data Flows
Campus 1	17,457	718,631	1,752,516	10.804 TB
Campus 2	4,637	98,789	132,672	1.077 TB
Campus 3	155	10,823	74,558	0.564 TB

C. Datasets

Our data collection methodology follows the one proposed in [2]. Specifically, we rely on passive measurements to analyze the behavior of the Dropbox client. We use the open source Tstat tool [13], installed on different vantage points, to monitor and collect information regarding all TCP connections in the network, including client and server IP addresses and the volume of exchanged data. We apply the same heuristics of [2] to identify and classify Dropbox traffic. For example, we use both the string `*.dropbox.com` found in TLS/SSL certificates and the Fully Qualified Domain Name (FQDN) that clients request to DNS servers to classify Dropbox traffic among the different Dropbox functionalities (i.e., control, data storage etc.). A complete list of domain names used by Dropbox as well as further details about the methodology to isolate and classify Dropbox traffic can be found in [2].[4]

We run Tstat on border routers of 3 large university campuses, here referred to as Campus 1, Campus 2 and Campus 3. The Campus 1 dataset consists of all traffic generated in an European university, with an official population of around 13 thousand people, including students, faculties and staff. Campus 2 and Campus 3 consist of the incoming and outgoing traffic of 2 Brazilian universities, with populations of 57 and 20 thousand people, respectively. All 3 datasets include traffic generated by wired workstations in research and administrative offices as well as wireless access points, whereas the Campus 1 dataset also includes traffic from/to student houses. Table I provides an overview of our datasets, showing the total collected traffic, the traffic from/to Dropbox, as well as the data collection period.

As previously mentioned, we apply heuristics to filter out data that are not related to Dropbox. Driven by our client behavior model, we restrict our focus to data and notification flows – i.e., flows related to *data transfers* and *sessions* in our model. Traffic related to other Dropbox user interfaces, such as the Dropbox Web interface, are discarded. This decision is justified by the fact that the vast majority of Dropbox traffic is produced by the Dropbox client application [2], which we could also confirm in our datasets. Finally, we discarded flows with duration under 2 seconds or volume below 5 kB, as they mostly reflect communication problems in the monitored networks (e.g., failed TCP connection attempts).

We group multiple flows into the same client session, according to our proposed model, by evaluating the client IP address and the start and end times associated with sessions and flows. A flow f is considered part of a session s if the client IP addresses of both f and s are the same, $start(f) \geq start(s)$, and $end(f) \leq end(s)$.

However, we notice some exceptions, such as sessions starting before the previous one from the same IP address is ended. In such cases of overlap between sessions, we cannot assign data flows occurring during the overlap to a unique session. Such overlaps might be due to (i) the use of NAT, which makes sessions and data flows originated from multiple clients to appear with a single source IP address; or (ii) communication failures between Dropbox clients and servers, which make the Dropbox client to open a new session before the previous one is terminated. In the latter case, we notice that the overlap between successive sessions is shorter.

Thus, we employ the following heuristic to deal with overlaps between sessions from the same IP address. Since Campus 1 does not have NAT in its sub-networks, all overlaps are likely caused by communication failures. Yet, overlaps are observed in 42 % of the Dropbox sessions in this campus. By analyzing the distribution of overlap durations, we can see a clear knee at around 140 seconds – thus, this value is used as a threshold to identify overlaps caused by communication failures. In all 3 datasets, sessions with overlaps lasting for up to 140 seconds are combined into a single one. This merge operation was performed in 36 %, 38 % and 50 % of the sessions collected in campuses 1, 2 and 3, respectively. Sessions with longer overlaps are discarded, as we are not able to uniquely assign data flows to them. In total, 5 %, 15 % and 45 % of the sessions collected from campuses 1, 2 and 3 have been discarded. We note the larger fractions of discarded sessions in the datasets collected from Campus 2 and Campus 3, where NAT is known for being widely deployed.

Table II summarizes some characteristics of the 3 datasets *after* the aforementioned filters have been applied. It presents the numbers of unique client IP addresses, sessions, data flows and the total traffic volume in data flows.

IV. CLIENT BEHAVIOR CHARACTERIZATION

We now characterize the Dropbox client behavior according to our two-layer model, presenting, for each model component, the statistical distribution that best fits the measured data. The best-fitted distribution is determined by comparing the Kolmogorov-Smirnov statistic [14] (for continuous distributions) and the least square errors (LSE) [15] (for discrete distributions) of the best-fitted curves for a number of commonly used distribution models. The Maximum-Likelihood Estimation (MLE) method [16] is used to estimate models parameters. We visually compare the curve fittings both at the body (small values in the x-axis) and at the tail (large values in the x-axis) of the measured data to support our fitting decisions.

The following distribution models are considered as candidates for best fit for continuous variables: Normal, Log-Normal, Exponential, Cauchy, Gamma, Logistic, Beta, Uniform, Weibull, Pareto. For discrete variables, we considered: Poisson, Binomial, Negative Binomial, Geometric and Hypergeometric.

[4]See also http://www.simpleweb.org/wiki/Dropbox_Traces

A. Session Layer

We first investigate the client *session duration*. Figure 2 shows the Cumulative Distribution Functions (CDF) of session durations in the 3 datasets. To make visual inspection clearer, we plot this figure in log scale. In general, client sessions tend to be short, although clearly longer in campus 1. For instance, the fraction of sessions longer than 200 minutes is 17% in campus 1 but only 8% in the other two. Similarly, the average session durations are 143.95, 84.65, and 93.75 minutes for campuses 1, 2 and 3, respectively, although the distributions present high variability, with coefficients of variation (CV)[5] ranging from 3.9 to 4.8. Recall that NAT is often used in campuses 2 and 3. Moreover, both campuses, particularly campus 3, experienced some degree of network instability during the monitored period. In campus 1, instead, users are connected to more stable networks with public IP addresses, and clients may remain connected to a Drobpox server throughout the period the device is turned on. In the other campuses, users may often change their IPs (due to NAT) or turn their devices offs, when their Dropbox clients are disconnected. Nevertheless, Figure 2 also shows that the measured data is best-fitted by Weibull distribution, which is a statistical distribution that has been used to model client active periods in other systems (e.g., active periods in live streaming in Peer-to-Peer systems [17]). We note that, despite the differences in parameter values (see caption), all 3 datasets are well fitted by Weibull distributions.

During a session, Dropbox clients may alter between active (*On*) and inactive (*Off*) periods. During an *On* period, clients upload/download data to/from Dropbox storage servers. Figure 3 presents the CDFs of the number of data transfers (*On Times*) during a single Dropbox client session, in the 3 datasets. We find a large fraction of sessions without any data transfer, in all 3 campuses, but particularly in campuses 2 and 3 (85% of the sessions). In those cases, clients connect to Dropbox servers, synchronize their account information but do not transfer any file. Like observed for session durations, we clearly note that users in campus 1 tend to perform more data transfers: as clients remain connected for longer, they have more opportunity to bundle and thus transmit more data. On average, clients perform 1.3, 0.56 and 0.47 data transfers per session in campuses 1, 2 and 3, respectively, whereas corresponding CVs are 3.7, 5.5, 6.1, indicating high variability. Despite differences, once again, we find that the same distribution - Negative Binomial in this case - is the best fit for the 3 campuses.

Next, we look at the durations of the data transfers. As shown in Figure 4, all 3 campuses present very similar distributions of *On* times, with most transfers occurring within very short intervals. For example, in at least 74% of the cases when a user exchanges data with a Dropbox storage server, it takes at most 200 seconds. We also note a knee in the curves around 60 seconds, which, we conjecture, is a default value for transmission timeout applied by Dropbox (corroborating the results found in [2]). On average, data transfers last for 192, 247 and 179 seconds in campuses 1, 2 and 3, with CVs falling around 2.5-3.7. All 3 distributions are well fitted by a Log-Normal distribution. The Log-Normal distribution has been previously used to model transfers duration (ON times) in other contexts such as in a web live streaming [18].

[5]Ratio of standard deviation to the average

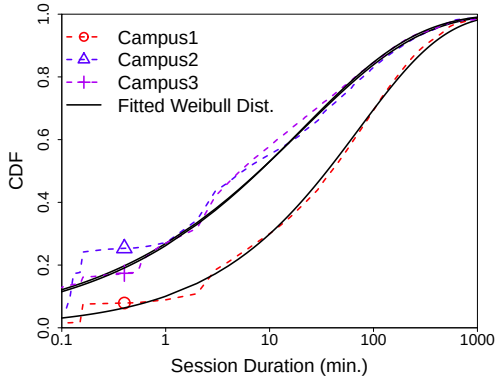

Fig. 2. Distributions of session durations. The probability density function (PDF) of Weibull distribution is $p_X(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}e^{-(x/\beta)^{\alpha}}$. Parameter values of best fit for campuses 1, 2 and 3 are: $\alpha = 0.525; \beta = 71.788$; $\alpha = 0.383; \beta = 20.776$; and $\alpha = 0.395; \beta = 20.366$.

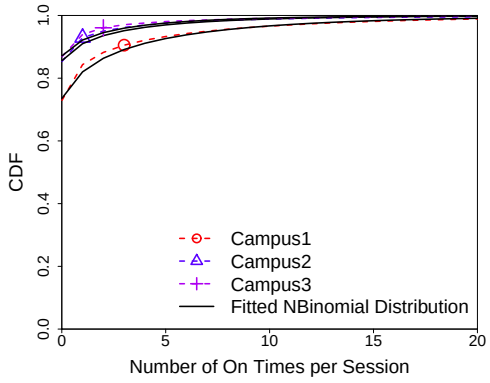

Fig. 3. Distributions of number of data transfer (On Times) per session. The probability density function (PDF) of Negative Binomial distribution is $p_X(x) = \frac{\Gamma(x+r)}{\Gamma(r)x!}p^r(1-p)^x$ for the probability p and Gamma function Γ. Parameter values of best fit for campuses 1, 2 and 3 are: $(r = 0.125; p = 0.086)$; $(r = 0.071; p = 0.112)$ and $(r = 0.066; p = 0.124)$.

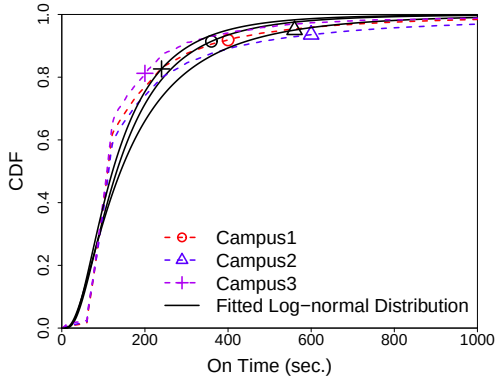

Fig. 4. Distributions of On times. The probability density function (PDF) of Log-normal distribution is $p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$. Parameters values of best fit for campus 1, 2 and 3 are: $(\mu = 4.889; \sigma = 0.712)$; $(\mu = 4.953; \sigma = 0.835)$ and $(\mu = 4.805; \sigma = 0.719)$.

We now turn to the periods of client inactivity (*Off times*). Figure 5 shows their distributions and best fits for the three campuses. On average, a client remains idle for 29, 39 and 81 minutes between consecutive data transfers in campuses 1, 2 and 3, with CVs ranging from 3.3 to 6.5. As expected, *Off* times are much longer than *On* times, as users spend much more time with their local jobs (file creation and editing) than transferring file updates from/to servers. Moreover, some users may temporarily disable the client synchronization option to avoid transferring all file updates. Also, Dropbox presents a file-bundling strategy in which file updates are delayed: files are bundled and pipelined in order to reduce latency and control overhead, as pointed out in [7]. One key reason for a larger Off time period on campus 3 is the large number of peers under NAT and dynamic IPs, which makes harder to identify 2 consecutives sessions from the same user (defined by its IP address). Despite the differences observed in the data measured in the 3 campuses, a Pareto distribution fits well the 3 curves.

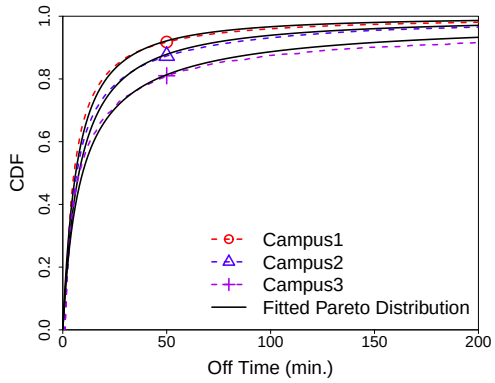

Fig. 5. Distributions of Off times. The probability density function (PDF) of Pareto distribution is $p_X(x) = \frac{\alpha\kappa^\alpha}{(x+\kappa)^{\alpha+1}}$. Parameters values of best fit for campus 1, 2 and 3 are: $(\alpha = 1.382; \kappa = 9.624)$; $(\alpha = 1.131; \kappa = 9.179)$ and $(\alpha = 0.790; \kappa = 6.781)$.

Finally, we turn to the last component of the session layer, i.e., the inter-session times. We found that no single distribution provided a good fit for the data, in any dataset. Thus, we opted for breaking the measured data into ranges, and determining the best fit for each range. Figure 6 shows the empirical distributions and best fits for measured times less than 720 minutes, which are the majority of all measurements (69%, 81% and 79% for campuses 1, 2 and 3). Once again, in order to make visual inspection clearer, we plot the curves in this figura in log scale. Inter-session times tend to be short, implying that users who leave the Dropbox service and later reconnect tend to do it quickly. This occurs more often in campuses 2 and 3 where the use of NAT and more unstable networks cause disconnections more often. For example, 52% (campuses 2 and 3) and 27% (campus 1) of the inter-session times are under 5 minutes. We find that a Log-Normal distribution is the best fit for all three campuses for this range of measured inter-session times as well as for the other considered ranges (below and above 2000 minutes[6]).

[6]In campuses 2 and 3, around 12% of the inter-session times are between 720 and 2000 minutes, leaving a fraction of 8% and 9% for the third range. In campus 1, 14% of the measured times are between 720 and 2000 minutes, and 17% of them are above 2000 minutes.

These other results are omitted due to space constraints.

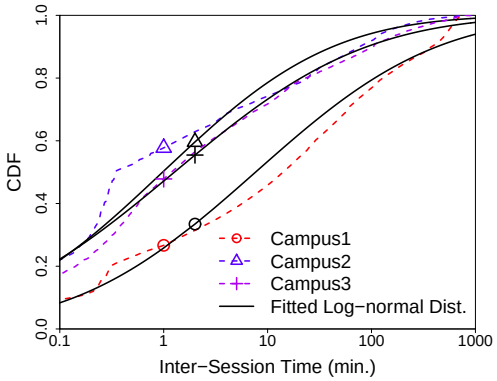

Fig. 6. Inter-session time distribution. The probability density function (PDF) of Log-normal distribution is $p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$. Parameters values of best fit for campus 1, 2 and 3 are: $(\mu = 2.035; \sigma = 3.137)$; $(\mu = -0.025; \sigma = 2.942)$ and $(\mu = 0.237; \sigma = 3.328)$.

B. Data Transmission Layer

The data transmission layer regards the multiple data flows that a data transfer (*On time*) may have. In this layer, we characterize the number of flows per data transfer as well as flow duration and volume.

Figure 7 shows that the distributions of number of flows per data transfer are very similar in all three campuses, with the vast majority (at least 71%) of the data transfers containing only 1 flow. On average this number is 1.62, 1.64, 1.36 in campuses 1, 2 and 3, respectively. Thus, in most transfers, a single data flow carries all the data required to synchronize the Dropbox folder. The Geometric distribution was the best fit, among all tested distributions, in the 3 datasets, although it does somewhat overestimates the number of flows per transfer. However, we point out that, to drive performance studies (e.g., capacity management and planning efforts), it is preferably to overestimate than underestimate the number of data flows as the former may lead to more conservative decisions.

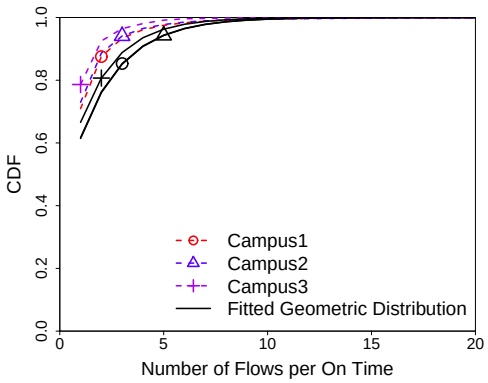

Fig. 7. Number of flows per On time. The probability density function (PDF) of Geometric distribution is $p_X(x) = p(1-p)^x$ for the probability p. Parameters values of best fit for campus 1, 2 and 3 are: $(p = 0.382)$; $(p = 0.379)$ and $(p = 0.422)$.

As shown in Figure 8, all 3 campuses present very similar distributions of flow volume. Although around 50% of the

flows carry less than 0.05 MB, at least 5% of them have more than 10 MB. On average, flow volume is about 6 MB, 9 MB and 8 MB for campuses 1, 2 and 3, but the variability is very high (CVs equal to 6.2, 5.2 and 5.3). The empirical distributions are clearly heavy tailed (note the log scale on the x-axis), being well fitted by Pareto distributions.

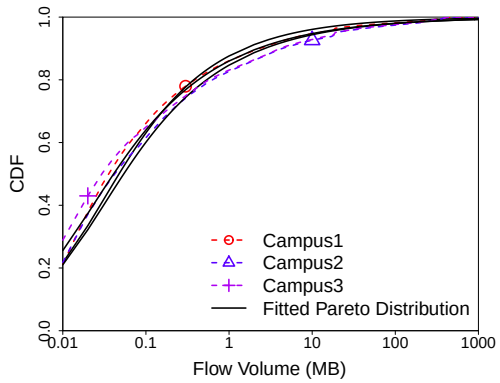

Fig. 8. Data flow volume. The probability density function (PDF) of Pareto distribution is $p_X(x) = \frac{\alpha\kappa^\alpha}{(x+\kappa)^{\alpha+1}}$. Parameters values of best fit for campus 1, 2 and 3 are: $(\alpha = 0.504; \kappa = 0.016)$; $(\alpha = 0.438; \kappa = 0.014)$ and $(\alpha = 0.426; \kappa = 0.010)$.

Finally, Figure 9 shows the distributions of flow durations. Since most data transfers consist of a single flow, the distributions of flow durations are similar to those of *On times* (Figure 4), being also well fitted by Log-Normal distributions.

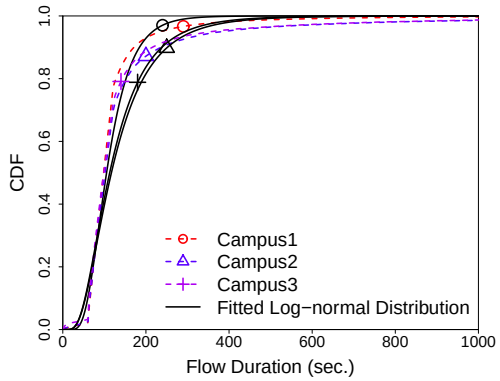

Fig. 9. Data flows duration. The probability density function (PDF) of Log-normal distribution is $p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$. Parameters values of best fit for campus 1, 2 and 3 are: $(\sigma = 4.629; \mu = 0.501)$; $(\sigma = 4.717; \mu = 0.673)$ and $(\sigma = 4.646; \mu = 0.753)$.

V. Summary and Future Work

In this paper, we have presented a characterization of the Dropbox client behavior using data collected from 3 campuses.

Our characterization was driven by a hierarchical model that captures client behavior at both session and data transmission levels. For each component of our model, we provided best-fitted statistical distributions, which can be used to support the generation of realistic synthetic workloads. A summary of our results is presented in Table III. Based on our findings, we make the following observations:

TABLE III. Hierarchical Model of Dropbox client behavior: summary.

Hierarchy Level	Model Component	Distribution
Session Layer	Session Duration	Weibull
	# On Times per Session	Neg. Binomial
	On Time	Log-normal
	Off Time	Pareto
	Inter-Session	Log-normal
Data Transfer. Layer	# Flows per On Time	Geometric
	Flow Volume	Pareto
	Flow Duration	Log-normal

First, we found close agreement across all monitored campuses for all components of our client behavior model, implying that the same distributions provide a reasonably good fit for all such campuses.

Second, in all 3 campuses, there is a large number of users with short sessions as well as a large fraction of session with no data transmission. In this case, users usually start their Dropbox clients, check for updates, and then, close their application. This behavior suggests that the use of client-side caching during sessions might be of limited benefit.

Finally, some components of our client behavior model, notably session durations and data transfer time, present similarities, in terms of distribution models, with other multimedia and web systems [17], [18]. However, we emphasize that parameter values are very different. For example, unlike in most Web systems, where clients interact for a few seconds and transmit a few kBytes, Dropbox client sessions tend to be much longer (few minutes) and transfer much more data (on the order of MBytes). These characteristics may deeply impact capacity planning and management decisions.

Future work includes extending the characterization to include other datasets as well as other aspects such as characteristics of the stored contents. Moreover, we intent to build a realistic synthetic workload generator for cloud storage applications.

ACKNOWLEDGMENTS

This research is partially funded by the authors' individual grants from CNPq, CAPES, FAPEMIG as well as by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web Grant No. 573871/2008-6) and the EU-IP project mPlane (n-318627).

References

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud Computing: State-of-the-Art and Research Challenges," *Journal of Internet Services and Applications*, vol. 1, pp. 7–18, 2010.

[2] I. Drago, M. Mellia, M. M. Munafò, A. Sperotto, R. Sadre, and A. Pras, "Inside Dropbox: Understanding Personal Cloud Storage Services," in *Proc. of the 12th ACM Internet Measurement Conference*, 2012.

[3] T. Mager, E. Biersack, and P. Michiardi, "A Measurement Study of the Wuala On-line Storage Service," in *Proc. of the IEEE 12th International Conference on Peer-to-Peer Computing*, 2012.

[4] R. Gracia-Tinedo, M. Sánchez-Artigas, A. Moreno-Martínez, C. Cotes-González, and P. García-López, "Actively Measuring Personal Cloud Storage," in *Proc. of the IEEE CLOUD'13*, 2013.

[5] A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm, "What's inside the Cloud? An architectural map of the Cloud landscape," in *Proc. of the ICSE Workshop on Soft. Eng. Challenges of Cloud Computing*, 2009.

[6] M. Zhou, R. Zhang, W. Xie, W. Qian, and A. Zhou, "Security and privacy in cloud computing: A survey," in *Proc. of the International Conference on Semantics, Knowledge and Grid*, 2010.

[7] I. Drago, E. Bocchi, M. Mellia, H. Slatman, and A. Pras, "Benchmarking personal cloud storage," in *Proc. of IMC*, 2013.

[8] A. Li, X. Yang, S. Kandula, and M. Zhang, "Cloudcmp: comparing public cloud providers," in *Proc. of the 10th ACM SIGCOMM conference on Internet measurement*, 2010.

[9] G. Wang and T. S. E. Ng, "The impact of virtualization on network performance of amazon ec2 data center," in *Proc. of the 29th Conference on Information Communications - INFOCOM*, 2010.

[10] W. Hu, T. Yang, and J. N. Matthews, "The Good, the Bad and the Ugly of Consumer Cloud Storage," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, pp. 110–115, 2010.

[11] H. Wang, R. Shea, F. Wang, and J. Liu, "On the Impact of Virtualization on Dropbox-Like Cloud File Storage/Synchronization Services," in *Proc. of the IEEE 20th Int. Workshop on Quality of Service*, 2012.

[12] Z. Li, C. Wilson, Z. Jiang, Y. Liu, B. Zhao, C. Jin, Z.-L. Zhang, and Y. Dai, "Efficient Batched Synchronization in Dropbox-like Cloud Storage Services," in *Proc. of the ACM/IFIP/USENIX Middleware Conference*, 2013.

[13] A. Finamore, M. Mellia, M. Meo, M. M. Munafò, and D. Rossi, "Experiences of Internet traffic monitoring with tstat," *IEEE Network*, vol. 25, no. 3, pp. 8–14, 2011.

[14] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit Techniques*. New York, USA: Marcel Dekker, 1986.

[15] R. K. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. New York, USA: John Wiley & Sons, 1991.

[16] W. N. Venables and B. D. Ripley, *Modern applied statistic with S*. New York, USA: Springer, 2002.

[17] A. Borges, P. Gomes, J. Nacif, R. Mantini, J. M. Almeida, and S. Campos, "Characterizing SopCast Client Behavior," *Computer Communications*, vol. 35, no. 8, pp. 1004–1016, 2012.

[18] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin, "A Hierarchical characterization of a live streaming media workload," in *Proc. of the ACM SIGCOMM Workshop on Internet Measurment*, 2002.