Characterizing Cyberlocker Traffic Flows

Aniket Mahanti[1], Niklas Carlsson[2], Martin Arlitt[3,4], Carey Williamson[3]
[1] Department of Computer Science, University of Auckland, New Zealand
[2] Department of Computer and Information Science, Linköping University, Sweden
[3] Department of Computer Science, University of Calgary, Canada
[4] Sustainable Ecosystem Research Group, HP Labs, USA

Abstract—**Cyberlockers have recently become a very popular means of distributing content. Today, cyberlocker traffic accounts for a non-negligible fraction of the total Internet traffic volume, and is forecasted to grow significantly in the future. The underlying protocol used in cyberlockers is HTTP, and increased usage of these services could drastically alter the characteristics of Web traffic. In light of the evolving nature of Web traffic, updated traffic models are required to capture this change. Despite their popularity, there has been limited work on understanding the characteristics of traffic flows originating from cyberlockers. Using a year-long trace collected from a large campus network, we present a comprehensive characterization study of cyberlocker traffic at the transport layer. We use a combination of flow-level and host-level characteristics to provide insights into the behavior of cyberlockers and their impact on networks. We also develop statistical models that capture the salient features of cyberlocker traffic. Studying the transport-layer interaction is important for analyzing reliability, congestion, flow control, and impact on other layers as well as Internet hosts. Our results can be used in developing improved traffic simulation models that can aid in capacity planning and network traffic management.**

I. Introduction

Over the years, there have been a range of services such as YouTube and BitTorrent that have allowed an average user to share content with others. While YouTube is the dominant platform for sharing user-generated video files, BitTorrent is the leading protocol for sharing large multimedia content. The emergence of cyberlockers is changing this content-sharing landscape. In contrast to BitTorrent, which splits files into small pieces that can be downloaded in parallel from many peers, these services rely on centralized infrastructure to serve each client individually. As clients typically retrieve all their content from a single server, this can result in much larger data volumes being transferred over a single connection. Since cyberlockers use the same underlying protocol (HTTP) as other Web applications, their increased popularity can significantly alter the known characteristics of Web traffic.

A rich content repository and simple (one-click) user interface that does not require installation of additional software (as in BitTorrent) have made cyberlockers very popular for content sharing. These services provide users with a Web interface to upload, manage, and share files in the cloud. When a file is uploaded to a cyberlocker site, a unique URL is generated that can be used for downloading the file. The user may then make the link public for sharing content. Cyberlockers offer two-levels of service: free and premium. Free users have mandatory wait times before a download begins, and their download rates are throttled. These restrictions are removed for premium users for a subscription fee. Well-known cyberlockers include MediaFire and RapidShare, which are among the top-200 most visited Web sites in the world (according to a July 2011 report by Google[1]).

The surge in popularity of cyberlocker traffic has been reported in the literature. Gehlen *et al.* [6] found that a cyberlocker was among the top-10 Web applications and constituted 5% of the total Web traffic. Maier *et al.* [8] reported that a single cyberlocker consumed 15% of total Web bandwidth in a large residential network. Allot [1] reported that cyberlocker traffic accounted for 19% of the total mobile broadband traffic, while Cisco[2] forecasts that cyberlocker traffic will grow at an annual rate of 47% in the coming years.

While usage, content characteristics, performance, and infrastructure of selected cyberlockers have been analyzed, not much is known about the flow-level characteristics of cyberlocker traffic and their impact on edge networks. Previous research on LAN and Web traffic have reported the presence of heavy-tailed properties (see [4], [15] and the references therein). Heavy tails have several implications on network traffic including load balancing, server scheduling, traffic routing, and caching. The tail behavior of cyberlocker traffic has not been investigated yet.

Using a year-long trace collected from a large campus network, we analyze flows originating from several cyberlockers, and study their properties at the transport layer. Studying the transport-layer interaction is important for analyzing reliability, congestion, flow control, and impact on other layers as well as of Internet hosts. These results provide insights into current network traffic and may be useful in developing cyberlocker traffic simulation models that can aid in capacity planning and network traffic management.

We use a combination of *flow-level* and *host-level* characteristics to provide insights into the behavior of cyberlockers and their impact on networks. We analyze four flow-level characteristics (*flow size*, *flow duration*, *flow inter-arrival time*, and *flow rate*), two host-level characteristics (*transfer volume* and *on-time*), and develop statistical models that capture the salient features of cyberlocker traffic. While several studies have analyzed these characteristics for Web and P2P traffic (e.g., [3], [13]), none exist for cyberlockers. A major contri-

[1] http://www.google.com/adplanner/static/top1000/
[2] http://www.cisco.com/web/MT/news/10/news_100610.html

bution of this work is the analysis and statistical modeling of transport layer characteristics of cyberlocker traffic.

The rest of the paper is organized as follows. Section II discusses prior work. Section III describes our objectives, metrics used, and trace collection methodology. Section IV presents an overview of the dataset. Sections V and VI present our characterization results and statistical models for cyberlocker flows and hosts, respectively. Section VII concludes the paper.

II. Related Work

RapidShare service architecture, usage patterns, and content characteristics were studied by Antoniades *et al.* [2], with the traces collected from two academic networks. They used active measurements to compare RapidShare with BitTorrent in terms of user-perceived throughput and content similarity. Most RapidShare files on the academic networks were requested once. Through targeted experiments, they found that RapidShare evenly distributed load across storage servers.

Cuxart *et al.* [12] analyzed RapidShare and Megaupload traffic using traces collected from a research network. They studied traffic properties, usage, content distribution, and server infrastructure. They noted that RapidShare and Megaupload were responsible for a significant fraction of the total traffic, and relied on a huge server infrastructure. A non-negligible percentage of users paid for premium accounts.

More recently, Mahanti *et al.* [7] analyzed the usage behavior, infrastructure properties, content characteristics, and user-perceived performance of five services: RapidShare, Megaupload, zSHARE, MediaFire, and Hotfile. They observed positive growth trends for cyberlockers as well as a large number of premium downloads. Most services had their servers in a centralized location, with the exception of Megaupload. Premium users achieved higher throughputs than free users, with both user types getting better rates than P2P transfers.

Our work in this paper complements these prior works. We collect longitudinal data from a large campus network, and analyze flow-level and host-level characteristics of cyberlocker traffic. To the best of our knowledge, this is the first work to analyze transport layer characteristics of cyberlocker traffic and produce statistical models that capture their salient features. Where possible, we compare these statistical models to models previously proposed for Web and P2P traffic.

III. Methodology

A. Objectives and Metrics

The transport layer characteristics of network traffic can significantly impact the effectiveness of networks. For example, the presence of many small flows or large flows can directly impact the effectiveness of load balancing, traffic management, prioritization algorithms, among other things.

To better understand the inherent behaviors of cyberlocker traffic, we apply a number of characterization metrics. While our primary focus is on *flow metrics* that capture the characteristics of a flow (defined as the set of packets being sent over a TCP connection, established between two host-port pairs[3]), we also consider *host metrics* that captures the characteristics of a host and its usage of these services. The flow-level metrics considered are listed and defined as follows:

- *Flow size:* The total number of bytes transferred within a single TCP flow. Flow size is computed as the total bytes transferred between connection establishment (SYN packets) and connection teardown (FIN packets) of a flow. Typically, TCP flows are bi-directional and flow size accounts for the bytes in both directions.
- *Flow duration:* The time between the establishment and termination of a single TCP flow. In our analysis, duration is computed as the time difference between the connection establishment (SYN packets) and connection teardown (FIN packets) of a flow.
- *Flow rate:* The average data transfer rate of a TCP connection. We compute flow rate by dividing the flow size by its duration.
- *Flow inter-arrival times:* The time interval between two consecutive flow arrivals.

In addition to understanding the traffic patterns, it is also important to understand the impact that individual hosts (or users) may have on the observed traffic patterns. To provide some insights to these characteristics, we consider (and define) the following host metrics:

- *Transfer volume:* The total traffic volume transferred by a campus host during the trace period. Transfer volume of a host is the sum of the sizes of all flows associated with an internal campus host.
- *On-time:* The total time the host was active during the trace period. We calculate a host's on-time as the sum of the duration of all flows associated with an internal campus host.

B. Distribution Fitting and Model Selection

To capture the above metrics and their traffic characteristics, we use both summary statistics (such as mean, median, standard deviation, Inter Quartile Range (IQR), and skewness), as well as empirical distributions. We primarily use the summary statistics to capture trends of the data.

We refer to values in the upper (biggest or most popular, for example) 10% of the empirical distribution as the tail; the remaining 90% of the distribution is referred to as the body. We use the Cumulative Distribution Function (CDF) to characterize the body of a distribution, and the Complimentary Cumulative Distribution Function (CCDF) to analyze the tail of a distribution.

Both the tail and the body of the distributions for the above metrics allow us to obtain further insights into the traffic characteristics of these services. In particular, the shape of the tails can have big impact on performance. For this purpose, CCDF tails are often studied to determine how quickly or slowly they decay. A distribution where the tail decays more

[3] Our analysis focus on TCP, as these services are HTTP-based and do not use UDP.

slowly than an exponential distribution is called *heavy-tailed*, with a sub-class called *long-tailed*.

We present statistical models that capture the salient features seen in the data. While we used a large set of distributional models, many of them are generalizations or variations of either the *Lognormal*, *Pareto*, or the (generalized) *Gamma* distribution. In the following, we define the CDF F_X for these distributions:

Lognormal: $F_X(x;\mu,\sigma^2) = 0.5 + 0.5\text{erf}\left(\frac{log(x)-\mu}{\sqrt{2\sigma^2}}\right)$

Pareto: $F_X(x;\alpha,\beta) = 1 - \left(\frac{\alpha}{x}\right)^{\beta}$

(Generalized) Gamma: $F_X(x;a,b,c) = \frac{\gamma(\frac{a}{c},(\frac{x}{b})^c)}{\Gamma(\frac{a}{c})}$

Here, erf is the error function, Γ is the gamma function, γ is the lower incomplete gamma function, and μ, σ^2, α, β, a, b, and c are the fitted parameters of the respective distributions. The complexity of the empirical distribution required us to apply hybrid fits of candidate distributions, where we fit the empirical distributions piece-wise. Each empirical distribution was divided into pieces based on manual inspection. We fitted seven well-known non-negative candidate statistical distributions (*Lognormal, Pareto, Gamma, Weibull, Levy,* and *Log Logistic*) to each piece and calculated the non-linear sum of least square error. The statistical distribution with the lowest error was chosen. After fitting all the pieces of the empirical distribution, we generated the P-P and Q-Q plots. The goodness of the fit was determined by manually inspecting these plots. These plots are omitted from the paper for brevity. While the piece-wise model fitting method allows the use of additional parameters, our models are in general able to capture the behavior of the entire empirical distribution (as exemplified by the good fit in both the CDF and CCDF). In contrast, most other works focus only on the tail behavior.

C. Data Collection

We used measurements collected from a large campus network's 400 Mbps full-duplex link to the Internet. The campus network has over 30,000 users. The dataset analyzed in this paper spans from January 1, 2009 through December 31, 2009.

The collected dataset is a trace of connection summaries. We refer to this data as the flow trace. We used the `conn` feature of *Bro* to collect these summaries. Each connection summary contained information such as the source and destination IP addresses, source and destination port numbers, the number of bytes transferred in each direction, and the *state* of the connection. A detailed description of the connection states is provided in the Bro documentation[4]. We only focus on connections with normal establishment and termination (represented by the *SF* flow state in Bro). Figure 1 illustrates our trace collection and analysis methodology.

TABLE I
SUMMARY OF FLOW TRACE

Service	Campus Hosts	Flows	Bytes
Mega Network (%)	75.1	42.7	68.2
RapidShare (%)	40.5	41.7	13.2
zSHARE (%)	34.6	4.0	8.1
MediaFire (%)	34.3	8.2	3.2
Hotfile (%)	5.2	0.4	2.4
Enterupload (%)	30.2	0.8	2.1
Sendspace (%)	10.7	1.0	0.8
2Shared (%)	6.9	0.2	0.8
Depositfiles (%)	7.5	0.7	0.8
Uploading (%)	5.2	0.4	0.4
Top-10 Services	13,223	7,300,774	21,810 GB
HTTP	30,225	3,876,358,538	482,647 GB

D. Mapping Network Flows to Cyberlockers

We used a complementary trace containing HTTP transactions (HTTP trace) that was captured concurrently with the flow trace. We used the HTTP trace to extract the IP addresses associated with the *top-10* services in terms of traffic volume. These services were as follows: *RapidShare*, *Mega Network*, *zSHARE*, *MediaFire*, *Hotfile*, *Enterupload*, *Sendspace*, *2Shared*, *Depositfiles*, and *Uploading*. For every transaction in the HTTP trace, we extracted the IP addresses associated with these 10 services, using the `Host` header. Note that Mega Network refers to the conglomeration of all the services under the *Megaupload* banner such as *Megavideo* and *Megalive*. We noticed that Megaupload sister sites used the same IP addresses as used by Megaupload. Thus, we were not able to separate network flows associated with Megaupload alone; hence, we labeled them as Mega Network. We aggregated all HTTP flows (flows using TCP port 80/8080), and studied their properties. Analyzing the aggregate cyberlocker flow traffic allows us to understand their impact on the network from a capacity planning and network management viewpoint.

IV. Dataset Overview and Preliminary Analysis

A. Trace Summary

The flow trace consists of 1 TB of compressed logs containing connection summaries of almost 4 billion HTTP flows. We aggregated the trace containing connection summaries for the top-10 cyberlocker services. During the 1-year period, the top-10 cyberlockers generated over 7 million flows from over 13,000 campus hosts. While cyberlocker flows represented only 0.19% of the total HTTP flow count, these flows accounted for 22 TB of traffic volume (4.5% of the total HTTP traffic volume). Over 40% of the total campus hosts accessed the top-10 services during the trace duration. We observed over 10,000 external cyberlocker servers, and these servers represented 0.06% of the total external HTTP servers.

Table I shows a summary of the flow trace. Note that the campus host count, number of flows, and bytes transferred for the individual services are reported as a percentage of the top-10 services total. The last two rows show the absolute quantities observed on the network. We wanted to distinguish *cyberlocker content flows* that transferred actual content from

[4] http://www-old.bro-ids.org/wiki/index.php/Reference_Manual:_Analyzers_and_Events#Connection_summaries

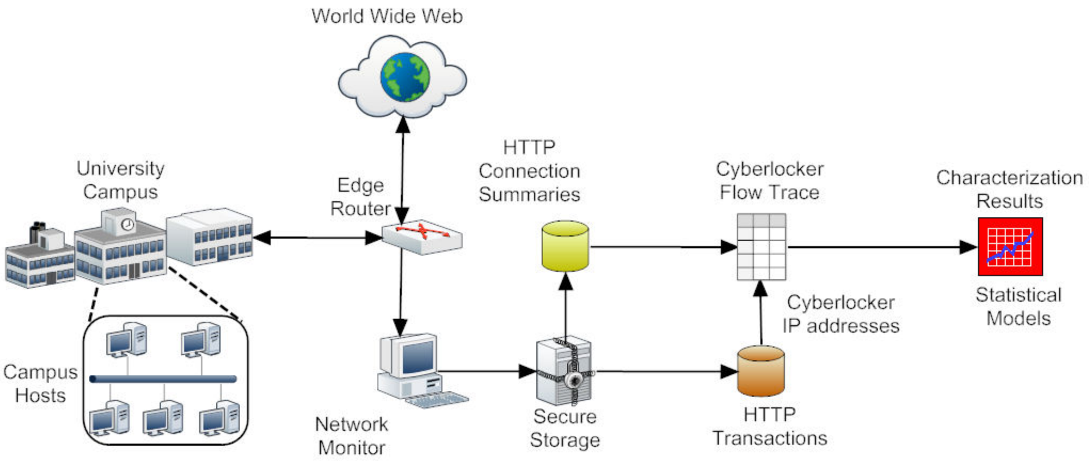


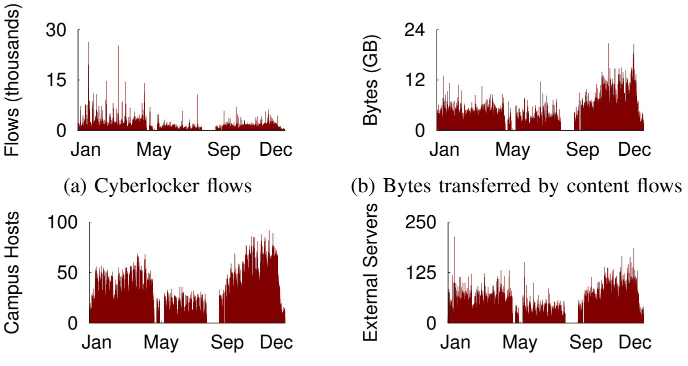
Fig. 1. Flow trace collection and analysis methodology

ordinary browsing flows. Antoniades *et al.* [2] used a threshold flow size of 150 KB to distinguish between content flows and browsing flows. We used a threshold flow size of 200 KB because we analyze more than one cyberlocker. Cyberlocker flows that transfer more than 200 KB of data are labeled as content flows. Note that content flows are a subset of the cyberlocker flows. Although content flows represent about 5% of the cyberlocker flows, they consume over 99% of the total traffic volume. About 81% of the campus hosts generated content flows.

Mega Network and RapidShare were the largest cyberlockers in terms of number of hosts, flows, and bytes transferred. Mega Network hosts transferred an order of magnitude more bytes than RapidShare. This is mostly due to streaming services that are part of Mega Network, such as Megavideo. These services offer easier viewing for hosts that may not wish to download the content. Additionally, video files uploaded to Megaupload were also available for viewing on Megavideo. The bottom-5 cyberlockers among the top-10 services had far fewer hosts that generated content flows. For example, only 25% of Enterupload hosts generated content flows.

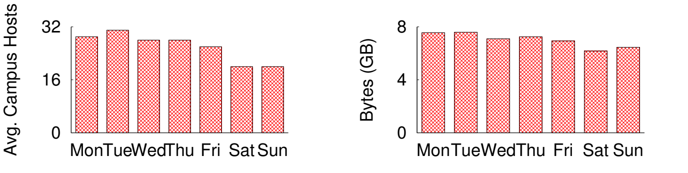
B. Campus Usage Trends

Figure 2 shows the activity trends of cyberlocker traffic in the campus network using 1-hour time bins. There is a gap in the trace between August 7 and August 31, when we were not able to collect data due to technical problems. Figure 2(a) shows that 50% of the total cyberlocker flows occurred during the Winter 2009 term (Jan-Apr), while 33% of total flows were observed in the Fall 2009 term (Sep-Dec). We also found that more content flows were noticed in the fall term (52%) as opposed to the winter term (36%). Figure 2(c) shows an increase in the number of campus hosts connecting to the cyberlockers towards the end of the year with up to 92 hosts active during an hour. Figure 2(d) shows the number of external cyberlocker content servers contacted by the campus hosts. As many as 214 external cyberlocker servers were contacted per hour by content flows.



(a) Cyberlocker flows
(b) Bytes transferred by content flows
(c) Content flow hosts
(d) Content flow external servers

Fig. 2. Flow and host activity trends



(a) Average hosts per day
(b) Average bytes transferred per day

Fig. 3. Day of week activity for campus hosts

Figure 3 shows the day of week activity of campus hosts. From Figure 3(a), we observe that on average up to 32 campus hosts per day contact cyberlockers for downloading content. There is more user activity during weekdays than during the weekend. Figure 3(b) shows a steadier trend in the amount of content flows during any day of the week. On average, up to 8 GB of content flow traffic volume is generated during any day of the week.

V. Flow-level Characterization

In this section, we present our flow-level characterization results and present distributional models from these results. Summary statistics for cyberlocker traffic are presented in Table II. We will refer to this table throughout our discussion of

TABLE II
FLOW-LEVEL SUMMARY OF CYBERLOCKERS

Type	Property	Mean	Median	Std.Dev	IQR	Skewness
Cyberlocker Flows	Flow Size (KB)	3,153.8	1.4	26,125.3	7.8	16.7
	Flow Duration (sec)	45.7	0.8	305.2	6.3	20.4
	Flow Rate (KB/sec)	11.1	2.0	58.9	6.4	34.8
	Flow IAT (sec)	4.3	0.4	835.6	3.0	2,235.5
Cyberlocker Content Flows	Flow Size (KB)	62,289.5	24,520.4	99,127.5	93,072.9	4.1
	Flow Duration (sec)	743.5	279.9	1,154.1	996.9	5.5
	Flow Rate (KB/sec)	109.2	67.9	230.7	75.0	9.6
	Flow IAT (sec)	86.0	25.8	3,736.3	66.6	499.3

results. We also compare our presented models with statistical models proposed in prior works for Web and P2P traffic.

A. Flow Size

From Table II, we note that the average flow size of the top-10 services is around 3 MB, which is an order of magnitude larger than HTTP flows in the trace (130 KB). We also observe that the content flows are orders of magnitude larger because these flows transfer the large content items typically hosted on these sites. The high mean and low median indicate the presence of some large-sized flows and many small-sized flows. Figure 4 confirms our hypothesis. From Figure 4(a), we observe that over 90% of all flows were smaller than 30 KB. Over 50% of the content flows were larger than 23 MB. These flows constitute video, audio, and executable files, which are much larger than a typical Web object [5], [6].

Figure 4(b) shows the tail of the flow size distribution. We notice that the tail for cyberlocker flows decays slowly, while the tail for the content flows decays more sharply. We analyzed the tail using the Pareto fitting and found cyberlocker flow sizes to be heavy-tailed ($\beta \sim 1.8$), however, the content flows were long-tailed ($\beta \sim 2.2$). Mori *et al.* [9] have reported similar results for video sharing workloads.

Models: We present statistical models that describe the body and the tail of flow size (S) distribution. These models may be used to generate transfer sizes of TCP flows in network simulations. Cyberlocker flows ($\mathbb{C}$) can be well-modeled by a mix of Lognormal and Pareto distributions:

$$F_{\mathbb{C}}(S) = \begin{cases} \mathit{Lognormal}(0.43, 0.37) & S < 1 \text{ KB} \\ \mathit{Lognormal}(0.48, 2.32) & 1 \text{ KB} \leq S < 67 \text{ KB} \\ \mathit{Lognormal}(-20.55, 15.44) & 67 \text{ KB} \leq S < 69 \text{ MB} \\ \mathit{Pareto}(8201.35, 1.82) & 69 \text{ MB} \leq S \end{cases}$$

Cyberlocker content flows ($\mathbb{CC}$) are well-modeled by a bounded Lognormal distribution:

$$F_{\mathbb{CC}}(S) = \begin{cases} \mathit{Lognormal}(6.02, 0.30) & S < 329 \text{ KB} \\ \mathit{Lognormal}(10.20, 5.98) & 329 \text{ KB} \leq S < 45 \text{ MB} \\ \mathit{Lognormal}(10.66, 1.07) & 45 \text{ MB} \leq S \end{cases}$$

Similar heavy-tailed distributions have been used to capture Web and P2P traffic. For example, Pustisek *et al.* [11] report Web flow sizes to follow a Lognormal distribution, while P2P flow sizes could be modeled by Pareto distribution.

B. Flow Duration

Table II shows that cyberlocker flows have a mix of short-lived and long-lived flows. Short-lived flows are associated with delivering objects from the sites, and since both the external servers and the campus hosts are well-provisioned, the time to transfer these objects is low. Content flows are long-lived, with the median being 280 seconds. Content flows can be long-lived due to the wait times instituted by the services on the free users, as well as bandwidth throttling.

Figure 5(a) shows the CDF of the flow durations. We notice a sharp increase at the 4 second mark for the cyberlocker flows. This increase is caused by RapidShare flows. We conjecture these to be aborted attempts by hosts to download content from the site or an artifact on the part of RapidShare. Note that this sharp increase is not visible for content flows. We notice that about 75% of the flows last less than 10 seconds. Analyzing the content flows, we notice that about 60% of these flows have durations less than 10 minutes. These correspond to downloading medium-sized content from these services.

We analyze the tail of the flow duration in Figure 5(b). We find that cyberlocker flow durations are long-tailed ($\beta \sim 3.2$). Content flows are also long-tailed as can be seen with the quicker tail decline.

Models: Cyberlocker flow duration (τ) can be well-modeled by a hybrid Gamma, Lognormal, and Pareto distribution:

$$F_{\mathbb{C}}(\tau) = \begin{cases} \mathit{Gamma}(3.64, 0.07, 0.70) & \tau < 6 \text{ sec} \\ \mathit{Lognormal}(-3.82, 8.51) & 6 \text{ sec} \leq \tau < 54 \text{ sec} \\ \mathit{Pareto}(0.14, 0.48) & 54 \text{ sec} \leq \tau < 30 \text{ min} \\ \mathit{Pareto}(423.1, 3.16) & 30 \text{ min} \leq \tau \end{cases}$$

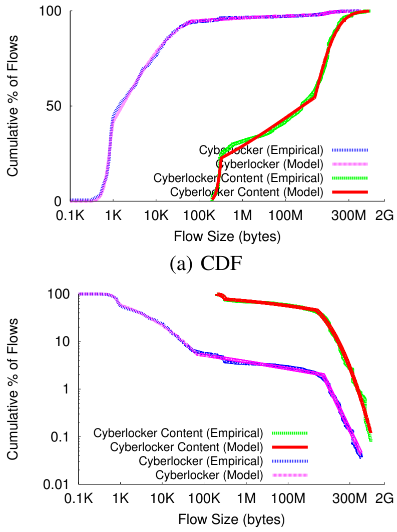
Content flows can be well-modeled by a combination of Lognormal and Gamma distribution:

$$F_{\mathbb{CC}}(\tau) = \begin{cases} \mathit{Lognormal}(4.13, 1.06) & \tau < 34 \text{ sec} \\ \mathit{Gamma}(0.36, 515.960, 0.33) & 34 \text{ sec} \leq \tau < 56 \text{ min} \\ \mathit{Lognormal}(6.18, 1.01) & 56 \text{ min} \leq \tau \end{cases}$$

These results concur with heavy tails previously reported for Megaupload flow durations [6]. Pustisek *et al.* [11] modeled Web and P2P flow durations using a Lognormal distribution.

C. Flow Rate

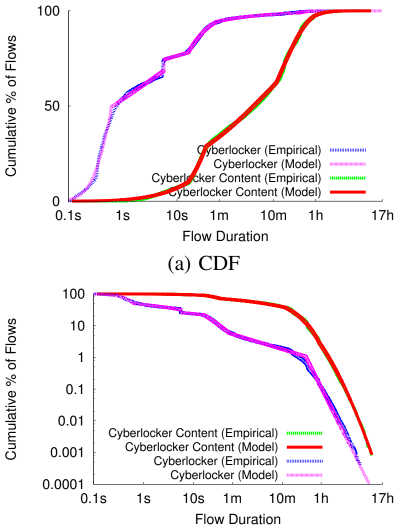
Table II shows that cyberlocker flows have much less dispersion than other characterization metrics. The average cyberlocker flow rate is 11 KB/sec, while the content flow rate average is much higher at 109 KB/sec. The median and mean



(a) CDF

(b) CCDF

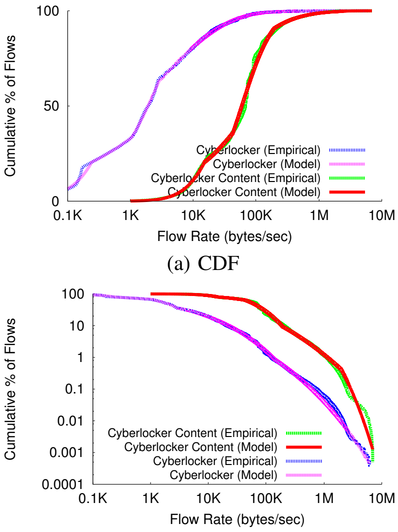
Fig. 4. Flow size distribution



(a) CDF

(b) CCDF

Fig. 5. Flow duration distribution



(a) CDF

(b) CCDF

Fig. 6. Flow rate distribution

content flow rates are much closer, also indicating less dispersion. It is interesting to note that cyberlocker content flows are larger and long-lived, yet they receive higher flow rates. The long-lived large flows extend well beyond TCP's slow-start phase, and thus achieve high-rate steady state throughput in TCP's congestion avoidance phase.

Figure 6(a) shows the CDF of flow rate. The content flow rate curve rises steadily and has a pronounced tail. These graphs include the presence of both free and premium hosts that download content from the services. Since the data shows the flow rates for several services, we do not observe steep increases at specific threshold values imposed by services for throttling free users. Since many flows are due to the Mega Network, which does not impose fixed throttling, we observe a much smoother trend.

Inspecting the tails in Figure 6(b), we notice that they decline slowly. The tail shows the higher end of the flow rates, which likely are for premium users. We analyzed the tails and found them to be heavy-tailed similar to as reported by Gehlen *et al.* [6] for Megaupload.

Models: Cyberlocker flow rates ($\Re$) are well-modeled by a two-mode Gamma distribution:

$$F_{\mathbb{C}}(\Re) = \begin{cases} \mathit{Gamma}(1.29, 0.87, 1.29) & \Re < 140\text{KB/sec} \\ \mathit{Gamma}(1.49, 1.08, 7.67) & 140 \text{ KB/sec} \leq \Re \end{cases}$$

Cyberlocker content flow rates can be well-modeled by a combination of Gamma and Lognormal distribution:

$$F_{\mathbb{CC}}(\Re) = \begin{cases} \mathit{Gamma}(2.92, 0.50, 0.43) & \Re < 43 \text{ KB/sec} \\ \mathit{Lognormal}(2.92, 1.78) & 43 \text{ KB/sec} \leq \Re < 1.9 \text{ MB/sec} \\ \mathit{Gamma}(1.79, 2.27, 0.37) & 1.9 \text{ MB/sec} \leq \Re \end{cases}$$

In contrast, Mori *et al.* found that video sharing workloads were well-modeled by Lognormal distribution [9]. These differences may be due to the time scale of the traces. Mori *et al.* measured network traffic for less than 10 hours, but from a more diverse set of users, whereas our trace captures the campus usage over a full year.

D. Flow Inter-arrival

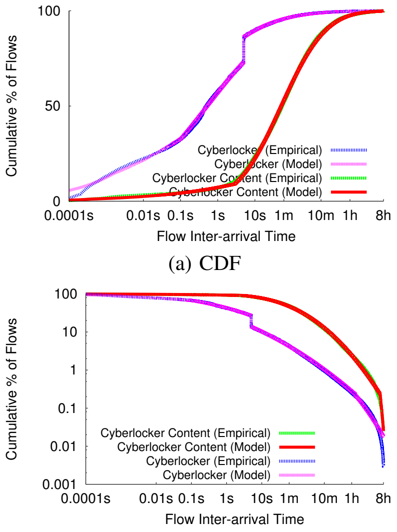
Analysis of our data (see Table II) shows that cyberlocker flow inter-arrival times are long. Content flows have even longer inter-arrival times. Figure 7(a) shows the CDF of cyberlocker flow inter-arrival times. Most of the inter-arrivals are within 10 seconds. We again notice the steep increase at the 4 second mark, which is due to RapidShare flows. Web hosts often maintain more than one concurrent TCP connection. Web browsers often initiate multiple concurrent connections to transfer content in parallel. This parallel download feature increases the degree of flow concurrency in HTTP-based applications; thus, reducing flow inter-arrival times. In case of cyberlockers, we found that these concurrent connections are not made to distinct IP addresses. Furthermore, cyberlocker flows have much lower flow arrival rate than HTTP flows, which further increases the inter-arrival times of cyberlocker flows.

Looking closely at the inter-arrival times of the content flows, we notice two regions. The region representing inter-arrivals less than 10 seconds could be due to download managers that initiate several parallel connections to download content. The region with inter-arrivals greater than 10 seconds increases steadily. These represent user-initiated downloads over a period of time. Content flow inter-arrivals are longer because there are far fewer such flows; most of the flows are due to objects being retrieved from sites.

Analyzing the tails in Figure 7(b), we find that cyberlocker inter-arrival times are heavy-tailed. Content flow inter-arrival times are heavy-tailed as well.

Models: Cyberlocker flow inter-arrival times (Δ) can be well-modeled by a hybrid Lognormal and Gamma distribution:

$$F_{\mathbb{C}}(\Delta) = \begin{cases} \mathit{Lognormal}(0.73, 6.79) & \Delta < 0.1 \text{ sec} \\ \mathit{Lognormal}(-2.64, 3.91) & 0.1 \text{ sec} \leq \Delta < 1.3 \text{ hours} \\ \mathit{Gamma}(0.15, 0.84, 0.19) & 1.3 \text{ hours} \leq \Delta \end{cases}$$



(a) CDF

(b) CCDF

Fig. 7. Flow inter-arrival time distribution

Content flow inter-arrival times can be well-modeled by a hybrid Gamma and Lognormal distribution:

$$F_{\mathbb{CC}}(\Delta) = \begin{cases} Gamma(2.27, 5.62, 0.09) & \Delta < 3 \text{ sec} \\ Lognormal(3.93, 2.15) & 3 \text{ sec} \leq \Delta \end{cases}$$

In contrast, Web flow inter-arrival times have been previously shown to be well-modeled by Weibull distribution [3], [11]. Both distributions significantly differ relative to the exponential distribution often used in analytic models and simulations (which assumes a Poisson arrival process), and may increase the amount of concurrency among flows due to such models.

VI. Host-level Characterization

This section presents a host-level characterization of cyberlocker traffic. This characterization provides information to network administrators for tasks such as bandwidth management and capacity planning. The results presented here may also be used to develop synthetic workloads and design realistic network simulations. Table III shows the summary statistics for the host-level characterization metrics. We refer to this table throughout this section. Note that cyberlocker campus hosts are those hosts who generated at least one flow by accessing any of the top-10 services. Cyberlocker content campus hosts are hosts that generated a content flow to any of the top-10 services. Campus hosts are identified by their IP addresses, which makes host-level analysis sensitive to DHCP address assignments. For example, wireless users may obtain new IP addresses using DHCP each time they connect, whereas campus residence users may have static addresses.

A. Transfer Volume

Table III shows that campus hosts transfer on average 1.6 GB of data from cyberlockers, although the median is much lower at 135 MB. This indicates the presence of some hosts that transfer a lot of data as well as hosts that transfer less data. Most of the transfer volume is due to content flows. On average, HTTP hosts transfer 31 MB of data. These results show that campus hosts using the top-10 cyberlockers have a very different traffic profile than typical campus hosts browsing the Web.

Figure 8(a) shows the CDF of host transfer volume. About 30% of the hosts transfer less than 50 MB of data. Another 35% of the hosts transfer between 50 MB to 2 GB of data. Large transfer volumes due to cyberlockers have also been reported lately by other studies [14]. The CDF has a pronounced tail that shows the presence of heavy hitters. We study the behavior of heavy hitters later. Figure 8(b) shows the presence of a heavy tail in the empirical distribution. Hosts who are engaged in active downloading of content have a long tail instead.

Model: We only present models for transfer volumes ($\mathcal{V}$) of cyberlocker campus hosts. Note that the campus hosts with content flows are responsible for the bulk of the transfer volume. Cyberlocker campus hosts can be modeled using a hybrid Lognormal and Pareto distribution:

$$F_{\mathbb{C}}(\mathcal{V}) = \begin{cases} Lognormal(4.15, 4.02) & \mathcal{V} < 24 \text{ MB} \\ Pareto(2821.8, 1.45) & 24 \text{ MB} \leq \mathcal{V} \end{cases}$$

Behavior of Heavy Hitters: Figure 9 shows the fraction of the total transfer volume at campus associated with individual campus hosts. Here, hosts are ranked according to their individual transfer volume, with the host responsible for the largest volume having rank 1. We find that a small number of hosts account for much of the volume transferred; we call these hosts heavy-hitters.

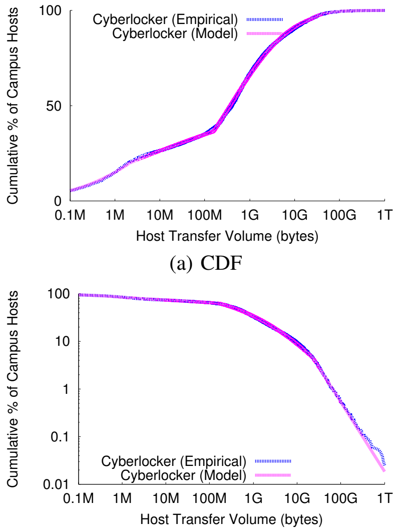
Figure 9(a) shows that there is a high skew in the cyberlocker usage on campus. For example, the top-10 ranked hosts account for about 40% of the total cyberlocker traffic. The top-100 ranked hosts account for more than 85% of the traffic volume. This distribution can be well-modeled by the Lavalette distribution[5], which captures the sharp decline of the hosts that transfer low data volumes. These hosts represent occasional users who browse the site, but do not download any content.

In Figure 9(b), we focus on hosts that acquire content from the services. Again, we notice that the top-10 ranked hosts account for over 40% of the traffic volume. The top-100 ranked hosts accounted for more than 91% of the bandwidth consumed. This distribution can be modeled by a hybrid Lavalette and Tsallis distribution. Ranked hosts 1-300 are modeled by the Lavalette distribution because it captures the decrease towards the lower ranked hosts well, while the Tsallis distribution models the volume transfer of hosts ranked higher than 300 (These hosts account for only 0.4% of the total cyberlocker content traffic volume.). Sen *et al.* [13] found that the top-10% of the P2P heavy hitters were heavy-tailed, but did

[5] Power-law distributions such as Lavalette and Tsallis are variations of the Zipf distribution [10]. These distributions provide more flexibility in fitting without increasing the number of fitting parameters significantly. These distributions are often referred as non-linear power-law distributions as they capture the drop towards the lower ranking items in the empirical data. This is in contrast to the Zipf distribution that appears as a straight line in log-log rank-frequency plot.

TABLE III
HOST-LEVEL SUMMARY OF CYBERLOCKERS

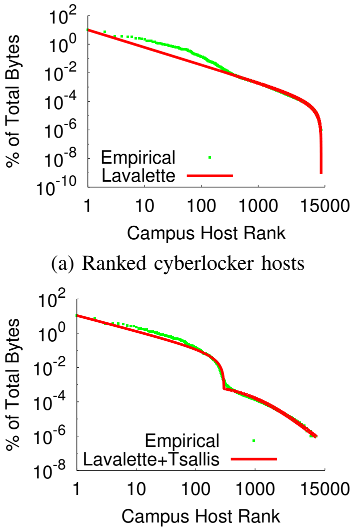
Type	Property	Mean	Median	Std.Dev	IQR	Skewness
Cyberlocker Campus Hosts	Transfer Volume (MB)	1,691.5	135.7	11,425.5	599.5	29.0
	On-time (sec)	25,240.2	2,761.8	129,598.5	10,202.5	24.7
Cyberlocker Content Campus Hosts	Transfer Volume (MB)	2,061.6	219.7	12,578.3	827.6	26.4
	On-time (sec)	25,198.5	3,277.7	115,605.7	11,198.4	22.0



(a) CDF

(b) CCDF

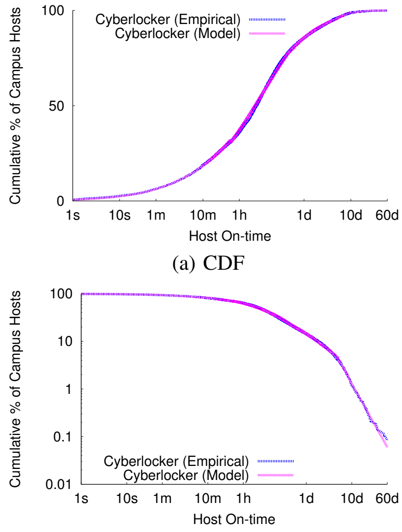
Fig. 8. Host transfer volume distribution



(a) Ranked cyberlocker hosts

(b) Ranked cyberlocker content hosts

Fig. 9. Ranked campus host transfer volumes



(a) CDF

(b) CCDF

Fig. 10. Campus host on-time distribution

not follow the Zipf distribution. Basher *et al.* [3] showed that the top-10% of Web and P2P heavy-hitters followed power-law characteristics, but this did not apply to the rest of the distribution. Our results are in contrast to prior work, however, these studies did not fit the empirical data to any non-linear power-law distributions.

B. On-time

On-time is the sum of all periods during which a host is active in either transmitting or receiving data over a year. We use host on-time to understand the activity patterns of campus hosts when they use cyberlockers. Table III shows that on average campus hosts spent over 7 hours accessing cyberlockers. The median is around 46 minutes, which indicates dispersion in the data.

In contrast, HTTP hosts, on average, spent 84 minutes browsing the Web. The low median value for HTTP hosts indicates the presence of automated applications that occasionally connect to external servers for short periods of time, such as weather updates. There are many users that spend more time on cyberlockers, while there are others who spent less time. Users with premium subscription may spend less time since they can download more content in less time. We also observe that most of the time spent by hosts is for downloading content.

Figure 10 shows the distribution of host on-time. We observe that about 30% of the hosts spent 35 minutes on the top-10 services. About 20% of the hosts spent more than 15 hours on these services. We also analyzed the tails of the empirical distributions. On-times of cyberlocker hosts are heavy-tailed, while on-times of cyberlocker content hosts are long-tailed.

Models: We present a model for on-times ($\mathcal{T}$) of cyberlocker hosts. These hosts can be well-modeled by a bounded Gamma and Lognormal distribution:

$$F_{\mathbb{C}}(\mathcal{T}) = \begin{cases} Gamma(2.03, 2.27, 0.30) & \mathcal{T} < 98 \text{ sec} \\ Lognormal(10.03, 3.42) & 98 \text{ sec} \leq \mathcal{T} < 50 \text{ min} \\ Lognormal(9.19, 2.00) & 50 \text{ min} \leq \mathcal{T} \end{cases}$$

VII. CONCLUDING REMARKS

We analyzed the flow-level and host-level characteristics of cyberlockers in a large campus network. We observed that cyberlockers introduced many small and large flows. Most cyberlocker content flows are long-lived and durations follow a long-tailed distribution. We observed that cyberlocker flows achieved high transfer rates. Cyberlocker heavy-hitter transfers followed power-law distributions. Increased cyberlocker usage can have significant impact on edge networks. Long-lived content flows transferring large amounts of data can strain network resources. The heavy-tailed nature of cyberlocker flow sizes may require better network provisioning, however, presence of heavy-hitters among cyberlocker users may make traffic management easier.

Acknowledgements

The authors are grateful to the anonymous reviewers for their constructive suggestions, which helped improve the clarity of the paper. This work was supported by funding from the Natural Sciences and Engineering Research Council (NSERC) of Canada, Alberta Innovates Technology Futures (AITF) in the Province of Alberta, and CENIIT at Linköping University.

References

[1] Allot Communications, "MobileTrends: Global Mobile Broadband Traffic Report," White Paper, 2010, http://www.allot.com/mobiletrends.html.

[2] D. Antoniades, E. Markatos, and C. Dovrolis, "One-click Hosting Services: A File-sharing Hideout," in *Proc. ACM SIGCOMM Conference on Internet Measurement*, Chicago, USA, November 2009.

[3] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, and M. Arlitt, "A Comparative Analysis of Web and Peer-to-Peer Traffic," in *Proc. Conference on World Wide Web*, Beijing, China, April 2008.

[4] M. Crovella, M. Taqqu, and A. Bestavros, "A Practical Guide to Heavy Tails." Birkhauser, 1998, ch. Heavy-tailed Probability Distributions in the World Wide Web, pp. 3–25.

[5] S. Gebert, R. Pries, D. Schlosser, and K. Heck, "Internet Access Traffic Measurement and Analysis," in *Proc. Traffic Monitoring and Analysis Workshop*, Vienna, Austria, March 2012.

[6] V. Gehlen, A. Finamore, M. Mellia, and M. Munafo, "Uncovering the Big Players of the Web," in *Proc. Traffic Monitoring and Analysis Workshop*, Vienna, Austria, March 2012.

[7] A. Mahanti, C. Williamson, N. Carlsson, M. Arlitt, and A. Mahanti, "Characterizing the File Hosting Ecosystem: A View from the Edge," *Perform. Evaluation*, vol. 68, no. 11, pp. 1085–1102, November 2011.

[8] G. Maier, A. Feldmann, V. Paxson, and M. Allman, "On Dominant Characteristics of Residential Broadband Internet Traffic," in *Proc. ACM SIGCOMM Internet Measurement Conference*, Chicago, USA, November 2009.

[9] T. Mori, R. Kawahara, H. Hasegawa, and S. Shimogawa, "Characterizing Traffic Flows Originating from Large-Scale Video Sharing Services," in *Proc. Traffic Monitoring and Analysis Workshop*, Zurich, Switzerland, April 2010.

[10] I. Popescu, "On a Zipf's Law Extension to Impact Factors," *Glottometrics*, vol. 6, pp. 83–93, 2003.

[11] M. Pustisek, I. Humar, and J. Bester, "Empirical Analysis and Modeling of Peer-to-peer Traffic Flows," in *Proc. IEEE Mediterranean Electrotechnical Conference*, Ajaccio, France, May 2008.

[12] J. Sanjus-Cuxart, P. Barlet-Ros, and J. Sol-Pareta, "Measurement Based Analysis of One-Click File Hosting Services," *Journal of Network and Systems Management*, pp. 1–26, May 2011.

[13] S. Sen and J. Wang, "Analyzing Peer-to-Peer Traffic across Large Networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 2, pp. 219–232, April 2004.

[14] F. Wamser, R. Pries, D. Staehle, K. Heck, and P. Tran-Gia, "Traffic Characterization of a Residential Wireless Internet Access," *Telecommunication Systems*, vol. 48, no. 1-2, pp. 5–17, October 2011.

[15] W. Willinger, V. Paxson, and M. Taqqu, "A Practical Guide to Heavy Tails." Birkhauser, 1998, ch. Self-similarity and Heavy Tails: Structural Modeling of Network Traffic, pp. 27–53.