Measurement and Modeling of Tumblr Traffic

Rachel Mclean, Mehdi Karamollahi, and Carey Williamson

University of Calgary, Calgary, AB, Canada
{rarmclea,mehdi.karamollahi,cwill}@ucalgary.ca

Abstract. Tumblr is a popular microblogging platform that allows users to share content and interact with other users. This paper focuses on the measurement and modeling of Tumblr network traffic characteristics, since few studies have focused on Tumblr from this perspective. Our work uses a combination of active and passive approaches to network traffic measurement. Using Wireshark and mitmproxy, we identify the primary hosts associated with Tumblr traffic, the traffic patterns associated with specific user actions, and the TCP connection behaviour. We then study Tumblr usage by our campus community for one week, using passively collected connection summaries. As a frame of reference, we also compare this traffic with several other popular social media platforms with user-generated content, namely Facebook, Instagram, and Twitter. Our work identifies several similarities and differences in the network traffic patterns for these social networking sites. We also develop and calibrate a synthetic workload model for Tumblr network traffic.

Keywords: Network traffic measurement · Internet traffic characterization · Workload modeling · Online social networks · Tumblr · TCP/IP

1 Introduction

Today's Internet supports many different online social network (OSN) communities. Facebook is the most well-known OSN, with nearly 26 billion monthly visits worldwide. Its familiar structure focuses on creating networks of people, organizing social events, and providing user-specific services. However, there is a wide variety of other social media platforms, including microblogging platforms like Twitter and Tumblr. The latter two examples focus more on providing a platform for curating and sharing user-generated content.

OSN sites with user-generated content have experienced tremendous growth in popularity over the past decade. Instagram [16] is a particularly interesting example; owned by Facebook, Instagram began as a photo and video sharing site, but has since grown into an immensely popular social media platform, with recently added functionality for voice calls, instant messaging, and live video streaming. Currently, Instagram has over one billion users who are active on at least a monthly basis. A typical day has more than 500 million users active, posting more than 400 million stories [6].

Tumblr is another example of a social media platform with user-generated content. Tumblr is organized around the concept of the blogosphere, but with

microblogging as its key idea. On Tumblr, users can create and curate their own microblogs, with any topic or media content type of their choosing. The microblogs in Tumblr span many diverse topics, including anime, cooking, fitness, gardening, hiking, movies, music, sports, yoga, and more.

In this paper, we focus on characterizing the usage of Tumblr by our campus community (i.e., faculty, staff, and students), and comparing its usage patterns with those of Facebook, Instagram, and Twitter, which have been well-studied in the prior literature. One motivation for our work is the lack of recent measurement studies of Tumblr, especially from a networking viewpoint. Another motivation is a desire to compare and contrast Tumblr with other OSN applications. Although existing social media platforms each serve different niches, we find several similarities in their underlying usage. For example, OSNs tend to generate long-duration sessions, often involving media objects with heavy-tailed transfer sizes.

The primary research questions examined in this paper are:

- What are the key characteristics of Tumblr traffic?
- How does Tumblr compare to other OSN applications in terms of network traffic usage patterns?

For our study, we collected information about Tumblr and other OSN sites for a one-week period in February 2020. We analyze this traffic in terms of usage patterns at the application, transport, and network layers. First, we characterize the network traffic patterns for Tumblr and other popular OSN sites. Second, we characterize the TCP connections and transfer sizes used. Third, we identify several characteristics that appear similar or different across these OSN sites. Finally, we design and implement a synthetic workload model for Tumblr traffic that can be used in network simulations or capacity planning studies.

The remainder of this paper is organized as follows. Section 2 provides background information on Tumblr, and reviews related research on Internet traffic measurement. Section 3 describes our measurement methodology. Section 4 provides a workload characterization of Tumblr traffic on our campus network, and compares these findings to the other OSN traffic observed. Section 5 presents our synthetic workload model for Tumblr network traffic. Finally, Section 6 concludes the paper.

2 Background and Related Work

This section provides some background information on Tumblr and network traffic measurement research.

2.1 Tumblr

Tumblr is a short-form blogging platform [15] that was launched in 2007. It currently has nearly 500 million blogs and 17 million daily posts [14].

Tumblr shares many features with Twitter and Instagram, but with fewer limitations on post type and length, allowing for highly diverse content. Each user has at least one dedicated blog with its own associated Tumblr URL.

Tumblr blogs are generally accessible to anyone, including non-Tumblr users. These blogs typically contain a chronological list of all posts published on that blog (including original and reblogged content), along with a brief user bio, a customized theme, and a set of links for navigating that user's blog. However, for Tumblr users, most of the interaction with other users is not done via their individual blogs, but via the functionality in the Tumblr 'dashboard'. The dashboard is only accessible to authenticated Tumblr users, and contains a time-ordered feed of content from all the blogs being followed.

Tumblr allows several different types of original posts, including text, audio, video, images, and external links. However, text posts may (and frequently do) include photos, links, and videos. Similarly, any user may append additional content of various types as a comment when reblogging a post.

Users may interact with a post in many different ways:

- *reblog*: copying a post to the user's own blog [19];
- *queue*: adding a copy of the post to a queue for later publication according to some user-specified time interval;
- *schedule*: setting a specific day and time for the post to be published to the user's blog;
- *save draft*: adding a copy of the post to a collection of drafts to be reviewed, posted, or deleted at a later time;
- *share*: allowing users to share a link to the post via Twitter or Facebook, or by sending a direct link to another Tumblr user using Tumblr's instant messenger;
- *like*: adding the post to a collection of 'liked' posts; and
- *reply*: adding a message that appears in the notifications for the post creator, and in the post's public history, but is not itself published to anyone's dashboard.

2.2 Related Work

Many papers in the literature have explored the complexities of online social networks [3–5, 8, 9, 13]. As the total population of social media users continues to grow, the ability to accurately characterize OSN user behaviour is increasingly valuable. Developing a clear understanding of network usage, number of requests, and data traffic volume can provide useful insights on how to improve protocol efficiency and user experience on a given platform or network.

One common approach to OSN research is to focus on the social aspects of interactions. For example, researchers have examined the structure and behaviour of social networks [9], analyzed click-stream data of browsing sessions [13], and characterized the behaviour of the users themselves [3, 8].

Relatively few papers have dealt with Tumblr specifically. In 2014, Xu *et al.* [20] analyzed 23.2 million users and 10.2 billion posts over four months.

They found that the majority of content in Tumblr is recirculated in the form of reblogged posts. They also found that Tumblr posts tended to have a longer lifespan on average than posts on other social media and microblogging platforms. By cross-referencing both implicit and explicit links on Twitter and Tumblr, they identified more than 6.5 million cross-linked pairs of users on the two platforms. Also in 2014, Chang *et al.* [4] characterized Tumblr in terms of user content, connections, and activity. A year later, in 2015, Alrajebah [1] examined content propagation across Tumblr by characterizing the cascade structure of reblogs.

While many of these papers prioritize analysis of OSN content, a higher level analysis of traffic patterns in terms of volume and connection characteristics can reveal useful insights into network performance. In 2018, Roy *et al.* [12] conducted a network measurement study of Learning Management System (LMS) traffic, identifying several issues at the transport layer that resulted in sluggish network performance. More recently, in 2019, Klenow *et al.* [7] measured Instagram traffic on a campus network, showing that this traffic averaged approximately 1 TB of data per day, and had very consistent usage patterns from one weekday to the next. Our work is similar in flavour to these latter two studies, but with a focus on Tumblr network traffic.

3 Methodology

Our research methodology involved a combination of active and passive approaches to network traffic measurement. The active approach was applied to study micro-scale aspects of Tumblr traffic for specific user test sessions conducted by us. The passive approach was used to provide a macro-scale view of Tumblr usage by our campus community as a whole.

3.1 Active Measurements

We conducted active measurements using our own client laptop in order to test Tumblr features and study browsing sessions on both Google Chrome and Mozilla Firefox. During these scripted test sessions, two existing network traffic analysis tools were used to passively capture network-level data, namely Wireshark and mitmproxy.

Wireshark [18] is an open-source network protocol analyzer. It captures packets as they pass through the network and displays them in a human-readable format, including IP addresses, port numbers, content length, and flags. Wireshark has powerful filtering capability and statistical analysis tools, which are helpful for identifying TCP connection behaviour associated with Tumblr and specific user actions.

Figure 1 shows an example of a Tumblr browsing session, based on a Wireshark capture that lasted just over one minute. This graph shows a time-series representation of the user activity on a single TCP connection to Tumblr. The vertical axis shows the bursts of network traffic (in bytes) as Tumblr pages and objects are accessed, while the horizontal axis illustrates the timing structure

for the user's interactions as indicated on the graph. Tumblr sessions generate bursty on-off patterns in the network traffic because of the think times between user interactions, such as page downloads, uploads, or reblogging events.

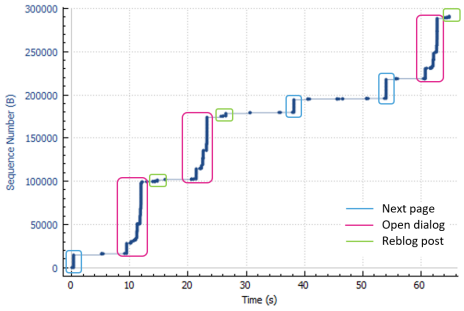


Fig. 1. Annotated TCP sequence number plot of a Tumblr browsing session

Because Tumblr's network traffic is encrypted (HTTPS), data collection was supplemented by the use of mitmproxy [10], a tool for intercepting secure network traffic between the server and client. We used this tool to identify traffic to and from Tumblr. During these test sessions, two IP addresses (152.199.24.192 and 152.195.50.59) were identified as responsible for the majority of Tumblr traffic.

Table 1 summarizes all of the Tumblr-related IP addresses identified during our active measurements. Some of the domain names (e.g., Yahoo, Oath, Verizon) and IP addresses reflect the historical evolution of Tumblr as a social media platform [14, 15].

Table 1. IP Addresses Observed for Tumblr Traffic

Domain Name	IP Address
www.tumblr.com assets.tumblr.com px.srvcs.tumblr.com api.tumblr.com	152.199.24.192
66.media.tumblr.com static.tumblr.com	152.195.50.59
tspmagic.tumblr.com	35.211.245.42
fc.yahoo.com	216.115.100.124
opus.analytics.yahoo.com	152.199.24.48
consent.cmp.oath.com	152.195.55.192
(unknown Verizon/ANS)	152.195.14.41

3.2 Passive Measurements

The primary network traffic dataset for our research was obtained using the connection logs from Zeek (formerly known as Bro [11]). This network monitor passively records connection-level summaries of traffic between our campus network and the Internet. These summaries do not include packet payloads, but do include source and destination IP addresses, port numbers, connection duration, connection state, as well as the number of packets and bytes that are sent and received on each TCP connection.

Our research used connection logs from a one-week period between Sunday, February 2 and Saturday, February 8, 2020. This period is well into the regular Winter semester, but before the COVID-19 pandemic that led to the University of Calgary switching to distance learning mode on March 15, 2020. These logs were filtered by IP address to consider only the relevant addresses that were identified during active measurements.

4 Network Traffic Characterization

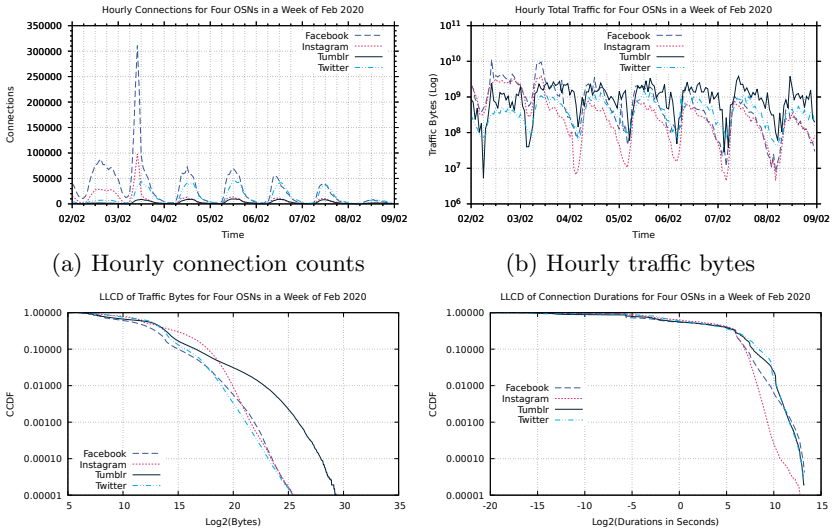
This section presents our workload characterization of Tumblr network traffic. We begin with an overview of the OSN traffic on our campus network, and then proceed to study Tumblr's diurnal traffic patterns, connection-level characteristics, and session-level characteristics. We also highlight similarities and differences compared to other OSN traffic.

4.1 OSN Traffic Overview

A measurement study of Tumblr and other OSN applications provides an opportunity to compare their characteristics and to gain a better perspective on Tumblr's traffic. For this purpose, we selected three of the most popular OSNs (Facebook, Instagram, and Twitter), and collected measurements for the exact same one-week period (February 2-8, 2020). Prior to collecting this empirical data, we used active measurements to determine the main IP addresses used by these OSNs, whether connecting via a Web browser or their mobile applications.

Unlike Tumblr, most other OSNs use cloud-based services like Amazon Web Services (AWS) and/or Content Distribution Networks (CDNs) to deliver a lot of their content, such as multimedia files. It is non-trivial to find the exact IP addresses of the CDNs used by these OSNs, since depending on the time of the day, type of content, and other characteristics of the content, the IPs change frequently. For example, depending on the Twitter account page from which content is being retrieved, the CDN may vary. Furthermore, it is difficult to determine whether the observed IPs are being used for any other non-OSN services on the Internet. Therefore, for this study, we have only used the IP addresses that are owned and managed by these OSNs, and used consistently in all the traces.

Figure 2(a) shows a comparison between the four OSN sites in terms of the number of TCP connections per hour. As one might expect, Facebook received



(a) Hourly connection counts

(b) Hourly traffic bytes

(c) LLCD of bytes per connection

(d) LLCD of connection durations

Fig. 2. Comparison between traffic characteristics of four OSNs over one week (Feb 2-8, 2020)

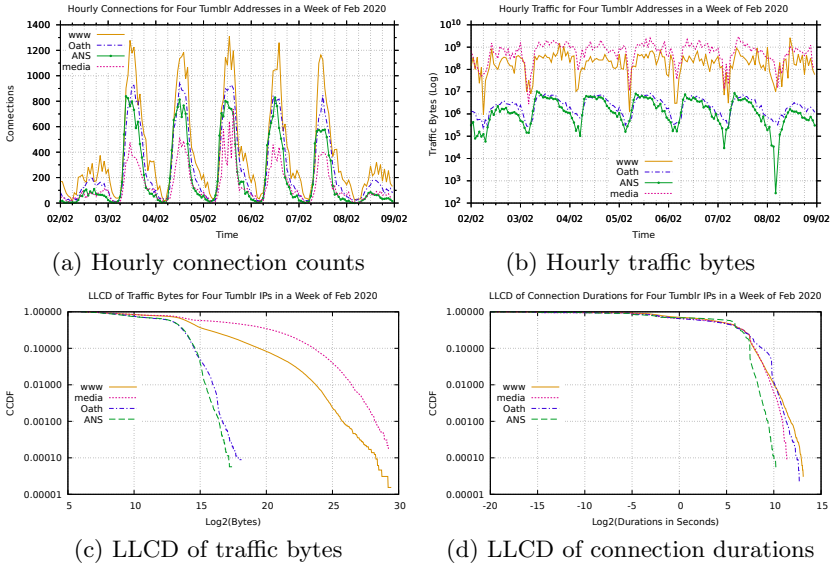
more connections than the other three OSNs in this week. Twitter had the second most connections, with Instagram third, and Tumblr having the fewest connections. Facebook also had the greatest variability in its connection activity, with a pronounced spike on the Monday, and a slight decline in activity throughout the week. The other three OSNs are much more consistent in their day-to-day traffic patterns, except for the weekends.

Figure 2(b) shows the hourly data traffic volume (in bytes) for the four OSNs under consideration. Surprisingly, the data traffic volume for Tumblr is comparable to, and sometimes higher than, the data volumes for the other three OSNs during the week, even though the number of TCP connections is much lower for Tumblr. This observation reflects larger media objects being transferred over Tumblr, as is evident from the Log-Log Complementary Distribution (LLCD) plot of transfer sizes in Figure 2(c). Specifically, Tumblr has some transfer sizes that exceed 500 MB, while the transfer sizes for the other three OSNs rarely exceed 50 MB. Another contributing factor is the use of CDNs (e.g., Akamai, Fastly) for storing and delivering large media objects for some OSNs, like Facebook. At the time of our study, Tumblr did not seem to use any CDNs at all.

Figure 2(d) shows the LLCD plots of connection durations for the four OSNs. These distributions look quite similar across the four OSN sites, even in the tails. However, Instagram has a slightly shorter tail to the distribution than Tumblr.

4.2 Tumblr Traffic Overview

Figure 3 provides a graphical overview of the Tumblr traffic characteristics, using the same format as Figure 2. The four lines on these graphs correspond to the four main IP addresses associated with Tumblr, namely the Web server (152.199.24.192), the media server (152.195.50.59), oath.com (152.195.55.192), and Verizon/ANS (152.195.14.41). We discuss these traffic characteristics over the next few subsections.



(a) Hourly connection counts

(b) Hourly traffic bytes

(c) LLCD of traffic bytes

(d) LLCD of connection durations

Fig. 3. Tumblr traffic characteristics for one week (Feb 2-8, 2020)

Table 2 shows the total volume of data transferred to and from the two primary Tumblr IP addresses identified earlier in Section 3. Two observations are evident from this table. First, the volume of inbound traffic dominates the outbound traffic, with nearly 98% of Tumblr traffic being inbound. This asymmetric traffic pattern is similar to that observed for other OSNs, such as Instagram [7]. Second, the media server for Tumblr (IP 152.195.50.59) is responsible for approximately three times as much data traffic volume as the main Web server (IP 152.199.24.192), despite having fewer TCP connections.

4.3 Tumblr Traffic Patterns

We next study the pattern of Tumblr traffic over time, in terms of hourly connections and hourly data traffic volume.

Figure 3(a) plots the number of new TCP connections initiated to Tumblr in every one-hour interval for one week of observation. These plots show clear

Table 2. Summary of Tumblr Traffic Asymmetry

IP Address	Pkts Out	Bytes Out	Pkts In	Bytes In
152.199.24.192	21,335,822	1.7 GB	33,635,987	43.8 GB
152.195.50.59	51,700,867	2.4 GB	100,808,104	143.4 GB
Total	73,036,689	4.1 GB	134,444,091	187.2 GB

diurnal patterns: the number of connections is lowest in the early hours of the morning, rises sharply to a peak around noon, then falls again in the evening. The number of Tumblr connections drops off markedly on the weekends, since fewer people are on campus.

The weekday traffic for Tumblr is fairly consistent on a day-to-day basis, suggesting that Tumblr users are creatures of habit. This consistency of usage is stronger than that seen in our earlier studies of LMS traffic on our campus network [12], but not quite as pronounced as the consistency seen for Instagram traffic [7]. One small difference in the Tumblr traffic is the "shoulder" effect in the late evenings, which is most evident for the Web server traffic. This plateau is possibly due to students in the campus residences who access Tumblr after classes have ended for the day.

Another interesting observation from Figure 3 is the relative number of connections to each Tumblr server address. The main Web server has the highest number of connections. The numbers of connections to Oath and Verizon/ANS are smaller; furthermore, the activity to each of these two sites seems to move in tandem, suggesting that they are closely related. Finally, the media server consistently has fewer daily TCP connections than the other three Tumblr addresses, although it is responsible for the largest proportion of the byte traffic in Figure 3(b). The latter observation implies that it tends to deliver larger objects; this observation will be confirmed shortly in our upcoming traffic analyses.

Figure 3(b) plots the hourly total data volume (in bytes) for Tumblr traffic. The total amount of data transferred varies widely across the four main IP addresses, so the number of traffic bytes is shown on a log scale (base 10), to more clearly display the hourly traffic patterns. As mentioned earlier, the Web server (yellow line) and media object server (pink line) account for substantially more data volume than the other two server addresses (shown in blue and green).

4.4 Tumblr Transfer Sizes

Figure 3(c) shows LLCD plots of the bytes per connection for the four Tumblr-associated IP addresses. This value demonstrates substantial variation, with some connections transferring very few bytes, and others approaching 1 GB. In the graph, the number of bytes per connection is shown on a log scale (base 2). This figure shows clear differences between the primary IP addresses (shown in pink and yellow) and the secondary addresses (shown in blue and green). The primary addresses have a pronounced tail to the transfer size distribution, with

a slow and gradual decline similar to a LogNormal distribution. In contrast, the data volumes for the secondary addresses decline earlier and more sharply, suggesting a lighter tail to these distributions.

4.5 Tumblr Connection Durations

We next examine the duration of TCP connections to Tumblr recorded in the Zeek connection logs. This value measures the time elapsed between the first and last observed packet over a single connection. The TCP connection durations vary widely, with many connections lasting less than a second while others last a few hours. The average connection durations for the primary IP addresses were 100.2 seconds for the media server, and 110.1 seconds for the Web server. These durations are even longer than the 72-second average observed for Instagram, which also uses persistent connections [7].

Figure 3(d) shows LLCD plots of TCP connection duration for the four main addresses associated with Tumblr, with the durations shown on a log scale (base 2). These plots show that the distribution of connection duration is similar for all four addresses measured, suggesting that Tumblr connection durations are not directly related to transfer sizes. A more detailed statistical analysis (not shown here) confirms that the correlation between transfer size and connection duration is rather weak.

Two particularly interesting observations from our Tumblr datasets are the large sizes of some of the transfers for a "microblogging" site (e.g., 320 MB), and the extremely low throughputs achieved (e.g., about 5 Mbps). Since the transfers are encrypted, we do not know the types of the objects involved. We speculate that the low throughput is attributable to both the persistent connection timeouts being used, as well as the window-limited TCP performance between Calgary and Tumblr (e.g., 64 KB of data every RTT).

To better understand the long-lasting TCP connections, some additional active measurement experiments were performed. These test sessions revealed a regular "API ping" between the client and the Tumblr Web server every 30 seconds, to update status information for the user. This "keep-alive" feature is unique to Tumblr, and helps explain some of the long-lasting connections with very little data volume.

4.6 Tumblr Connection State

We next analyze the TCP connection states recorded in the connection logs. A typical TCP connection, opened with a SYN flag and terminated with a FIN flag, will have a recorded final state SF, while a connection that is terminated with a RST flag may have a recorded final state of RSTO (reset by the originator) or RSTR (reset by the responder). Other connections may be only partially[1] observed, including connections that are attempted but not established (S0),

[1] Our network traffic monitor is restarted every 3 hours to reduce the risks of data loss.

connections that are established but not terminated (S1), or connections where only midstream traffic is observed without opening or closing handshakes (OTH).

Table 3 shows the relative frequency of different connection states in the dataset of Tumblr connections. SF connections are the most common, accounting for approximately 25% of connections to the Web server and 30% of those to the media server. Among the remaining connections, RSTO, OTH, and S1 are the most frequent final states for both addresses. The RSTO connections may be the due to user actions interrupting TCP connections, or resource management policies that use resets to terminate idle TCP connections. One explanation for the unusually high frequency of OTH and S1 states may be the repeated API pings and long connection durations described in Section 4.5.

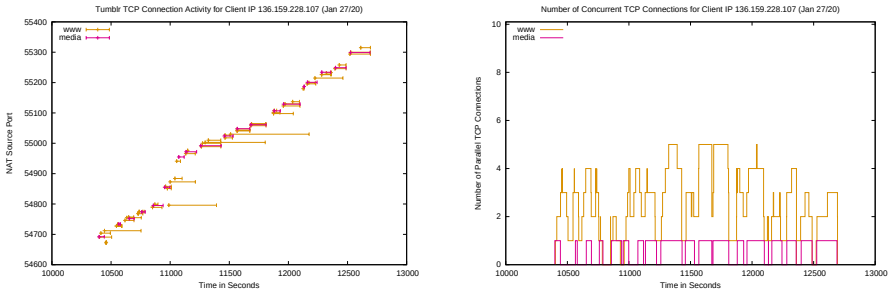
Table 3. TCP Connection States for Tumblr Traffic

Connection	**152.199.24.192 (www)**		**152.195.50.59 (media)**	
State	***Conns (%)***	***Bytes (%)***	***Conns (%)***	***Bytes (%)***
SF	26.36%	39.42%	32.23%	34.35%
RSTO	22.10%	24.11%	19.36%	22.98%
OTH	16.56%	12.24%	15.83%	14.75%
S1	15.29%	12.74%	15.04%	16.01%
RSTOS0	7.01%	4.08%	5.43%	5.88%
RSTRH	3.61%	0.49%	1.08%	0.65%
SHR	2.65%	0.52%	6.45%	0.45%
RSTR	2.62%	2.96%	1.61%	2.54%
S3	1.78%	2.98%	1.53%	1.90%
S0	1.24%	0.00%	0.59%	0.00%
S2	0.42%	0.31%	0.52%	0.34%
SH	0.27%	0.09%	0.28%	0.15%
REJ	0.09%	0.05%	0.04%	0.01%
Total	100.0%	100.0%	100.0%	100.0%

4.7 Session-Level Characteristics

Figure 4(a) shows an example of the Tumblr connection activity for one user during a session that lasted about 45 minutes. The horizontal axis is time, and the vertical axis shows the TCP source port number used by the campus NAT. In general, the port numbers increase monotonically upward with time, until they reach the maximum possible port value in the range, and wrap around to the lower end of the range again. Each '+' on the plot indicates the start time of a TCP connection to a Tumblr server (www or media). The horizontal lines, when present, indicate the time duration of the connection. A solid line is used for connections to the Tumblr Web server, and a dashed line for connections to the Tumblr media server.

Several observations are evident from Figure 4(a). First, about 75% of the 78 connections in this session were to the Tumblr Web server, with the rest to the Tumblr media server. Second, most of the connection durations are short, but there are a few long ones, as evident from the lines on the graph. Third, there are several examples of multiple TCP connections in parallel, either to the Web server, or to the media server, or to both servers concurrently. In most cases, there are clear timing dependencies between these connections, which either start at very similar times, or end at very similar times.



(a) Example of a real user session (b) Concurrent connections in real session

Fig. 4. Example of connection-level characteristics for a real Tumblr session

Figure 4(b) provides a detailed look at the use of parallel TCP connections during this user session with the Tumblr site. This user maintains up to five TCP connections in parallel with the Tumblr Web server, and at most one TCP connection with the media server. The number of concurrent connections fluctuates with time, as the user navigates to different pages and takes different actions on the Tumblr site.

5 Tumblr Traffic Model

As the final component of our work, we have designed and implemented a synthetic workload model for Tumblr network traffic. The model is written in C, and consists of just over 300 lines of code.

Our synthetic workload model for Tumblr is conceptually similar to Web browsing models from the mid-1990's [2]. Specifically, we use a hierarchical workload model, with three conceptual layers. The topmost layer models a single user *session* in Tumblr. This session consists of one or more *conversations* with a Tumblr server[2] (e.g., 75% to Web server and 25% to media server) at the intermediate

[2] We ignore the Oath and Verizon/ANS servers, which contribute negligibly to the connection count and data volume in the empirical Tumblr traffic.

layer, with random think times in between. Each conversation with a server in turn involves one or more TCP *connections*, with either independent or correlated start times, and randomly generated transfer sizes. The TCP connection layer constitutes the lowest layer of the Tumblr model; we do not model the IP packet layer, or the dynamics of TCP congestion control. Concurrent TCP connections are allowed to both of the Tumblr servers, with connection start times slightly staggered to reflect processing overheads and non-deterministic user interactions.

The Tumblr workload model has been calibrated based on the empirical measurement data reported in the previous section. We use a geometric distribution for the number of connections, and hybrid distributions for the numbers of bytes sent and received on each connection. Connection durations depend on data transfer sizes, network bandwidth for uploading/downloading, TCP handshaking, and the timeout values used for persistent connections. Table 4 provides a summary of the main parameters in our Tumblr model, and the default settings for these parameters. The default settings produce Tumblr sessions with an average of 50 TCP connections, and lasting just under half an hour on average.

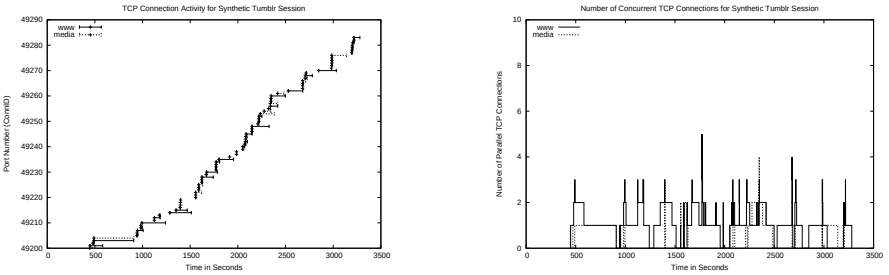
Table 4. Parameters and Settings for Tumblr Traffic Model

Parameter	Setting
Session IAT	Exponential(120)
Conversations	Geometric(5)
Web Server Prob	0.70
Media Server Prob	0.20
Dual Server Prob	0.10
Connections	Geometric(10)
Persistent Conn Prob	0.20
Persistent Conn Timeout	60 s
Tail Prob	0.50
Bytes Sent	LogNormal(12,2)
Bytes Received (www)	LogNormal(14,2)
Bytes Received (media)	LogNormal(15,2)
Upload Bandwidth	1.5 Mbps
Download Bandwidth	4.0 Mbps
User Think Time	Uniform(0,60)

When building a synthetic workload model for Tumblr traffic, it is important to model the cross-correlations in TCP connections, which are clearly not independent. For this purpose, the conversation model allows some shared state between Web and media server connections, with the data volumes randomly split between the two connections, while the connection durations are harmonized.

Figure 5 shows an example of the output from this model for a 50-minute user session with about 84 TCP connections. About two-thirds of these connections

go to the Web server, and about one-third to the media server. These graphs use the same style and format as Figure 4. Specifically, Figure 5(a) represents the time series evolution of TCP connection usage, while Figure 5(b) shows the concurrent connection usage across the two main Tumblr servers. These graphs are visually similar to those for the empirical user session shown in Figure 4. However, we have not modeled the API ping feature, which likely causes some of the longer Web server connections in Figure 4(a).



(a) Example of a synthetic user session (b) Concurrent connections in synthetic session

Fig. 5. Example of connection-level characteristics for a synthetic Tumblr session

Figure 6 provides a more detailed look at how transfer sizes for connections are modeled. In the example shown here, we use a hybrid distribution, with 50% of the transfer sizes being in the body of the distribution (e.g., less than 64 KB), and 50% of the transfer sizes being in the tail of the distribution. The tail is modeled using a LogNormal distribution, as indicated earlier. We explicitly model the asymmetry of the traffic, with received bytes on average being about four times larger than sent bytes. We also increase the average transfer size for the media server, which has a more pronounced tail to the distribution for received bytes. The modeling results in Figure 6 are structurally similar to those shown for the empirical workload in Figure 3(c), though the latter did not explicitly separate the two directions of traffic.

By combining the foregoing Tumblr session model with a time-varying Poisson arrival process, we have generated one synthetic week of Tumblr traffic, as shown in Figure 7. In this particular example, we used a mean session arrival rate of 30 Tumblr sessions per hour during the main part of the work day (8:00am to 4:00pm), but only 20% of this rate in the evening, and only 10% of the base rate in the early morning hours. (We also ignored the notion of weekends.) The graph shows the instantaneous number of Tumblr sessions that are concurrently active at each time throughout the week. With these example settings, there are about 15 active Tumblr sessions in steady state during the main part of each day.

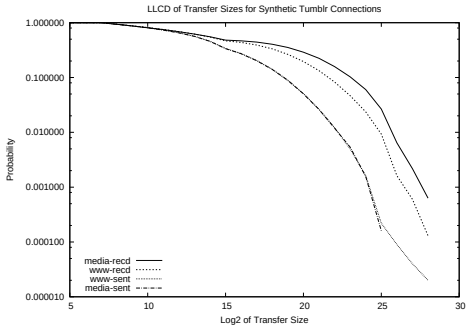


Fig. 6. Distribution of transfer sizes for synthetic Tumblr connections

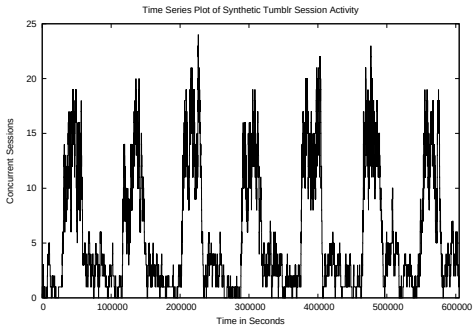
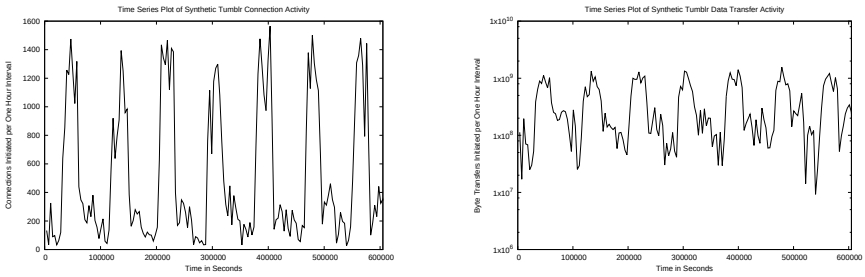


Fig. 7. Diurnal profile for one week of synthetic Tumblr sessions

Figure 8 provides a more detailed breakdown for our synthetic week of Tumblr traffic, in a format similar to that of Figure 3(a) and (b). Figure 8(a) shows the synthetic connection arrival pattern on a per-hour basis, while Figure 8(b) shows the corresponding byte transfer information, again on a per-hour basis. This aggregate model captures the diurnal structure well, while still reflecting the stochastic nature of connection arrivals and transfer size variability.



(a) Synthetic connections per hour (b) Synthetic bytes per hour

Fig. 8. Traffic profile for one week of synthetic Tumblr connections

In the future, we plan to incorporate our Tumblr model into network simulations of OSN usage on mobile wireless networks. Our synthetic traffic model for Tumblr is currently available online from the Web site of the third author (Williamson) at the University of Calgary.

6 Conclusions

In this paper, we have presented a detailed workload characterization study of Tumblr traffic on our campus network. Furthermore, we have built upon the insights gained from our study to identify similarities and differences compared to other popular social media applications.

The main highlights from our paper are summarized as follows. First, Tumblr usage is seemingly much lower than Instagram and other OSNs, when measured in users or TCP connections, but it is actually comparable in data traffic volume. Second, despite relative differences in popularity, the structural properties of network traffic for OSN sites are qualitatively similar in many ways, including diurnal profile, asymmetry, long-lived connections, and heavy-tailed transfer size distributions. Third, there are some distinct features of Tumblr traffic that differ from other OSNs. These include the session keep-alive behavior, dual server usage, low TCP throughput, and the absence of CDNs.

Our campus-level study provides a glimpse of possible future demands for OSN usage on enterprise, ISP, and mobile networks. We hope that our measurement and modeling results are of value to researchers, network operators,

protocol designers, and content providers as they consider how to handle future growth in OSN traffic, especially on mobile networks.

Acknowledgements

The authors thank the anonymous reviewers from IEEE MASCOTS 2020 for their constructive feedback and suggestions on an earlier version of this paper. Financial support for this research was provided in part by the Department of Computer Science at the University of Calgary, and by Canada's Natural Sciences and Engineering Research Council (NSERC). The authors are also grateful to University of Calgary Information Technologies (UCIT) for facilitating our collection and analysis of the campus-level network traffic.

References

1. N. Alrajebah, "Investigating the Structural Characteristics of Cascades on Tumblr", *Proceedings of IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, Paris, France, pp. 910–917, August 2015.
2. M. Arlitt and C. Williamson, "A Synthetic Workload Model for Internet Mosaic Traffic", *Proceedings of the 1995 Summer Computer Simulation Conference*, Ottawa, ON, Canada, pp. 852-857, July 1995.
3. F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing User Behavior in Online Social Networks", *Proceedings of the 9th ACM Internet Measurement Conference (IMC)*, Chicago, IL, pp. 49–62, November 2009.
4. Y. Chang, L. Tang, Y. Inagaki, and Y. Liu, "What is Tumblr: A Statistical Overview and Comparison", *ACM SIGKDD Explorations Newsletter*, Vol. 16, No. 1, pp. 21–29, September 2014.
5. Q. Deng, Z. Li, Q. Wu, C. Xu, and G. Xie, "An Empirical Study of the WeChat Mobile Instant Messaging Service", IEEE INFOCOM Workshops, pp. 390-395, Atlanta, USA, May 2017.
6. Instagram, "A quick walk through our history as a company", March 2019. `https://instagram-press.com/our-story`
7. S. Klenow, C. Williamson, M. Arlitt, and S. Keshvadi, "Campus-Level Instagram Traffic: A Case Study", *Proceedings of IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, Rennes, France, pp. 228–234, October 2019.
8. M. Maia, J. Almeida, and V. Almeida, "Identifying User Behavior in Online Social Networks", *Proceedings of 1st Workshop on Social Network Systems*, Glasgow, Scotland, pp. 1–6, April 2008.
9. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", *Proceedings of ACM Internet Measurement Conference (IMC)*, San Diego, CA, pp. 29–42, October 2007.
10. mitmproxy, `https://www.mitmproxy.org`, 2020.
11. V. Paxson, "Bro: A System for Detecting Network Intruders in Real Time", *Computer Networks*, Vol. 31, No. 23-24, pp. 2435–2463, December 1999.
12. S. Roy, C. Williamson, and R. Mclean, "LMS Performance Issues: A Case Study of D2L", *ISCA International Journal of Computers and Their Applications*, Vol. 25, No. 3, pp. 113–122, September 2018.

13. F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding Online Social Network Usage from a Network Perspective", *Proceedings of ACM Internet Measurement Conference (IMC)*, Chicago, IL, pp. 35–48, November 2009.
14. Tumblr, "About", May 2020. `https://www.tumblr.com/about`
15. Wikipedia, "Tumblr", May 2020. `https://en.wikipedia.org/wiki/Tumblr`
16. Wikipedia, "Instagram", May 2020. `https://en.wikipedia.org/wiki/Instagram`
17. C. Williamson, "Internet Traffic Measurement", *IEEE Internet Computing*, Vol. 5, No. 6, pp. 70–74, November/December 2001.
18. Wireshark, `https://www.wireshark.org`, 2020.
19. XKit, `https://new-xkit-extension.tumblr.com`, 2020.
20. J. Xu, R. Compton, T. Lu, and D. Allen, "Rolling Through Tumblr: Characterizing Behavioral Patterns of the Microblogging Platform", *Proceedings of ACM Conference on Web Science*, Bloomington, IN, pp. 13–22, June 2014.