2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)

Campus-Level Instagram Traffic: A Case Study

Steffen Berg Klenow
University of Southern Denmark
Odense, Denmark
sklen15@student.sdu.dk

Carey Williamson
University of Calgary
Calgary, AB, Canada
carey@cpsc.ucalgary.ca

Martin Arlitt
University of Calgary
Calgary, AB, Canada
marlitt@ucalgary.ca

Sina Keshvadi
University of Calgary
Calgary, AB, Canada
sina.keshvadi1@ucalgary.ca

***Abstract*—Instagram is a popular network application for photo sharing, video streaming, and online social media interaction. In this paper, we present results from an initial characterization study of Instagram network traffic, as viewed from a large campus edge network. Despite the challenges of NAT, DHCP, end-to-end encryption, and high traffic volume, we are able to identify key characteristics of Instagram traffic, which exceeds 1 TB per day. The main highlights from our study include classic observations such as diurnal usage patterns, Zipf-like distributions for IP frequency-rank profile, and heavy-tailed transfer size distributions.**

***Index Terms*—Network traffic measurement; Internet traffic characterization; online social networks; Instagram; TCP/IP**

I. Introduction

Internet traffic is always changing, as new network applications come to the forefront and grow in popularity. In some cases, these applications displace prior network services, which decrease in popularity since they are no longer in vogue. In other cases, the new applications just compete for our attention, and add to the time that people spend online.

On today's Internet, a lot of the traffic is for video streaming. Prominent examples include services like YouTube and Netflix, for our personal entertainment. However, these are not the only popular network applications. People also want to communicate with each other, whether it is through traditional email, or via the latest and greatest social media applications.

In the last few years, Instagram has grown rapidly in usage and popularity. It is a popular Internet service for photo sharing, video streaming, and online social media interaction. Instagram has arguably become the newest and hottest social media application, especially among high-school and college-age students, as well as the general public.

In this paper, we investigate the use of Instagram by our campus community at the University of Calgary. We believe that campus networks offer a rich and fertile environment for studying current trends in network application usage, because of the high-speed network connectivity, the flexible usage policies, and the large cohort of young and technically-savvy users, who are often referred to as "digital natives" because of how they spent their growing up years online [31]. Furthermore, such studies can help identify the performance implications of these network applications on future enterprise networks, whether in academia or industry.

As motivational context for our work, it is important to consider previous studies of Internet traffic. Prior researchers have looked at the emergence of YouTube video streaming for user-generated content [8], [11], [15], the usage of wireless LANs [10], [22], and Netflix video streaming traffic [1], [2], [19]. More recent papers from our own group at the University of Calgary have looked at Learning Management System (LMS) traffic, and Outlook (Office 365) email traffic [34]. All of these studies have offered insights into the usage and performance of current network applications, and ways to improve them in the future.

One of the primary technical challenges in traffic characterization studies of this type is the growing use of end-to-end encryption on the Internet. While encryption is essential for the privacy and security of online users, it also obfuscates several aspects of the traffic, such as file names, types, and content popularity. A secondary challenge is the growing complexity of campus enterprise networks, which often use middle-boxes (e.g., wireless APs, DHCP, NAT, VPN) to support flexible BYOD (Bring Your Own Device) networks. These technologies also enhance privacy by obscuring aspects of the traffic (e.g., number of users, user location, mapping between IP addresses and end-user devices), making session identification and user modeling difficult. Nonetheless, key characteristics can still be discerned from the traffic, such as diurnal usage patterns, content sizes, and bandwidth consumption.

The two main research questions in our work are:

- What are the key characteristics of Instagram traffic?
- What are its network performance implications?

Many of the techniques used in this paper are inspired by prior works on network traffic measurement, power-laws, and heavy-tailed distributions, such as [2], [9], [21], [24]. We seek to establish the existence of these properties (or not) in campus-level Instagram traffic.

The rest of this paper is organized as follows. Section II provides some background on Instagram, and discusses prior research on Internet traffic measurement. Section III describes our research methodology, measurement infrastructure, and software tools for data collection and analysis. Section IV presents the traffic characterization results. Finally, Section V concludes the paper.

II. Background and Related Work

A. Instagram

Instagram is a photo and video sharing service owned by Facebook [32]. The platform was originally developed by Kevin Systrom and Mike Krieger, and its initial release was

2375-0227/19/$31.00 ©2019 IEEE
DOI 10.1109/MASCOTS.2019.00032

on October 6, 2010. The platform quickly grew in popularity, with 25,000 signups on the very first day [17].

Using Instagram, people can share their photos and videos with the world, categorizing them using hashtags, and describing the photos using a short text. A user can browse through postings from other users, either by user or by hashtag, and "like" the posts of other users to show their support for it [32].

Currently, Instagram has over one billion users who are active at least monthly. Each day has more than 500 million users active, posting more than 400 million stories [17].

Instagram keeps expanding its functionality beyond the posting/liking of photos and videos. Some of the newer features are [32]:

- Instagram Direct: users can send messages directly to other users.
- Instagram Stories: a Snapchat-like feature, where users can share photos/videos publicly for 24 hours.
- Instagram TV (IGTV): supports uploads of longer videos filmed in different formats (e.g., smartphone).
- Live Streaming: users can send/receive live video streaming sessions with multiple other users.

B. *Related Work*

Internet traffic measurement is a well-known technique to facilitate the study of network-based applications [12]. The two main methodological approaches are *passive* and *active* network measurement. Passive approaches observe a network system without perturbing it, while active approaches use test sessions or probe packets to see how a server or a network handles certain requests. These two basic techniques can be used to collect and analyze data about network systems. A tutorial on network traffic measurement is available in [33].

A primary goal of network measurements is traffic characterization. Web browsing [4], [9], peer-to-peer applications [20], [6], and video streaming services [8], [11], [15] are examples of network applications that have been studied in the past. Other works have focused on smartphone traffic [14], Learning Management Systems [28], and online social networks [7], [18], [16], [23], [25], [26], [29].

Network traffic measurement studies are especially valuable when they identify performance problems with particular protocols, applications, or services, or can predict potential performance implications in the future. For example, Adhikari *et al.* studied Netflix [1], [2] and the evolution of its CDN infrastructure to support large-scale video streaming services. In our present paper, our measurements help provide insight into the potential performance implications of Instagram traffic on our campus network, as well as on Instagram's infrastructure itself. For example, our measurements show that Instagram now accounts for over half of the total Facebook traffic seen on our campus network.

III. METHODOLOGY

Our traffic characterization study is conducted using a combination of passive and active network measurement. All the measurements are captured at the University of Calgary in Calgary, Alberta, Canada.

The primary datasets used in our paper are collected using passive traffic measurement. At the edge router connecting the University of Calgary to the Internet, we have installed specialized hardware for Internet traffic measurement. Our Endace DAG packet capture card sees all inbound and outbound network packets in a mirrored stream from the campus edge router. For privacy reasons, we only process packet headers, and not packet payloads (which are often encrypted anyway). To conserve on storage space requirements for long-term data collection, the packet streams are processed using Zeek (formerly called Bro [27]) to produce TCP connection-level summaries.

Each connection is summarized into a one-line entry in a log, as shown in the example in Figure 1. We use the default Zeek connection log format. The most relevant fields for our needs are the timestamp, the source IP address and port, the destination IP address and port, the TCP connection duration, the TCP connection state, and the counts of packets and bytes sent and received on each TCP connection.

We pre-process the logs to extract only the Instagram traffic of interest. Looking at the IP addresses used by Instagram and Facebook, there are many different addresses, and they change frequently, since these are cloud-hosted services in AWS, with lots of DNS round-robin for load-balancing across servers. However, through some active measurements of Instagram test sessions locally (using an Android smartphone, the Instagram mobile app, the Charles proxy software, and experiments with login/logout, messaging, photo sharing, live streaming, and IGTV), we identified one particular IP address (157.240.3.63) that was used consistently in over 90% of the Instagram requests seen. The DNS host names associated with this IP are `i.instagram.com`, `platform.instagram.com`, `instagram.c10r.facebook.com`, `scontent-sea1-1.cdninstagram.com`, and `graph.instagram.com`, all of which resolve to the same IP address. The data we study in this paper is only for connections to this address.

In this paper, we focus on a single week of data from Sunday March 3, 2019 to Saturday March 9, 2019. This week is from the middle of the academic semester, when many students, staff, and faculty are around campus, and thus provides a representative sample of Instagram activities. The data contains just over 13 million TCP connections between the University of Calgary and Instagram.

IV. EMPIRICAL MEASUREMENT RESULTS

This section presents the Instagram traffic measurements from our campus edge network. Specifically, we focus on traffic destined to IP address 157.240.3.63, which is known as `i.instagram.com`. While there are several other IP addresses involved in an Instagram session (e.g., `b.i.instagram.com` 157.240.3.174 and `graph.facebook.com` 157.240.3.20), this IP appears to

Timestamp	UID	Src_IP	SPort	Dest_IP	DPort	Prot	Svc	Duration	TCPout	TCPin	State	IPout	B_out	IPin	B_in
15628.248886	CDu29N3WgQZb	1.2.3.4	50468	157.240.3.63	443	tcp	ssl	165.901378	9053	86515	S3	100	14297	98	90892
15628.250997	CThhn41tYm27	1.2.3.4	50470	157.240.3.63	443	tcp	ssl	3.334059	489	447	RSTO	11	1093	7	1133
15628.301082	Cz1RCa39Ralf	1.2.3.5	50040	157.240.3.63	443	tcp	ssl	329.763400	425964	45413	SF	737	468313	538	73893
15628.307782	CwGcdG3hgX7e	1.2.3.6	62558	157.240.3.63	443	tcp	ssl	0.004667	39	39	SF	4	247	5	263
15628.316061	CLkeo71eiHx1	1.2.3.7	57396	157.240.3.63	443	tcp	ssl	209.412519	5239	91968	SF	67	9983	83	94918
15628.348089	COhKpI2ASJOg	1.2.3.7	57397	157.240.3.63	443	tcp	ssl	209.388824	7489	914542	S3	454	31736	682	933553
15628.502214	CVh9Ev3pUSga	1.2.4.1	52990	157.240.3.63	443	tcp	ssl	8.459407	1703	126326	SF	83	6031	99	131482
15628.504240	CpDJJC4XfYE8	1.2.4.1	52991	157.240.3.63	443	tcp	ssl	8.457381	2222	962556	SF	431	24646	719	998714

Fig. 1. Example of Selected Fields from the Zeek Connection Log Format for Empirical Analysis of Instagram Traffic (anonymized source IPs)

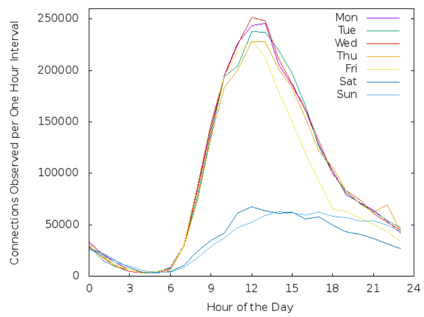


Fig. 2. Daily Patterns for Instagram Traffic (March 3-9, 2019)

be the main entry point into Instagram, and accounts for over 90% of the HTTP(S) requests during our test sessions. We thus focus on this single IP, with the caveat that our results may slightly *underestimate* the total Instagram traffic.

A. Overview

Table I provides a statistical summary of our week-long dataset. The table shows the daily totals for connections, packets, and bytes for Instagram traffic, as well as some structural properties of the data, such as the number of distinct local IP addresses and /24 subnets observed. The primary observations from Table I are the high data volumes generated by Instagram traffic (e.g., about 1 TB per day), the asymmetry of this traffic (e.g., received bytes exceed sent bytes by about a factor of 20), and the large client base (e.g., several thousand distinct IP addresses observed, many of which are NAT addresses with multiple users behind them). What is also remarkable is the consistency in the traffic from one day to the next. A typical weekday involves about 2 million TCP connections to Instagram from 1,600 different IP addresses across about fifty /24 subnets, exchanging well over 1 TB of data. Each TCP connection lasts about 72 seconds on average, though this is slightly higher on weekends.

Figure 2 provides an graphical view of the daily Instagram traffic. The Instagram traffic shows a strong diurnal pattern, as is common for many Web-based applications and services. Traffic activity rises quickly each morning, peaks near mid-day, and then declines gradually in the mid-to-late afternoon and evenings. The Instagram connection counts are quite consistent on each weekday, but drop by about 60% on the weekends. This pattern reflects the diurnal activities of the faculty, staff, and students on campus. The network traffic is higher when more people are present on campus, although Instagram traffic continues late into the evening, perhaps reflecting students in labs, libraries, coffee shops, or dormitories.

Another observation from Figure 2 is the consistency of the weekday traffic from Monday to Thursday, despite the varying class lecture schedule (e.g., MWF versus TTh). There is a slight decline on the Friday afternoon (yellow), and lower activity levels on weekends (since no lectures take place, and fewer people are on campus). There are also some subtle differences between Saturday (dark blue, with slightly lower traffic, especially in the evening) and Sunday (light blue, with slightly higher traffic in the evening). These patterns are consistent with the intuition of Friday and Saturday evenings for social outings, with Sundays and "school nights" for catching up on academic and/or online pursuits.

B. TCP Connection State

Our next analysis focuses on TCP connection state. As indicated in the sample log data in Figure 1, some connections have TCP's normal opening handshake (SYN) and closing handshake (FIN), resulting in state SF, while some do not. For example, some SYN connection requests are never answered (S0), and some TCP connections are aborted with a reset, either by the originator (client) or the responder (server). Furthermore, some connections might last so long that they commence in one (3 hour) log and finish in a different log. These partially observed connections can have several different states in the connection log, such as S1, S2, S3, or OTH.

Table II provides a statistical summary of the TCP connection states observed in our dataset. Approximately half (47.9%) have the normal SF state, while the other half do not. Among the latter, the most prevalent is a reset of a successful connection by the originating client (RSTO). There could be many reasons for this, including a user aborting slow content, changing pages prematurely, or deactivating their mobile device. It could also reflect how some Web browsers handle idle TCP connections [5]. Next most prevalent are partial connections (S1, S2, or S3), many of which are long-lasting and exchange a lot of data. Two other reset types are also seen, either for unsuccessful connections (RSTS0), or for successful connections reset by the server (RSTR). The next state of interest is S0, for unsuccessful TCP connection attempts. Finally, there are a half-dozen other unusual states (e.g., half-open connections, REJECT, etc) that account for a very small proportion of the total connections and bytes.

TABLE I
OVERVIEW OF EMPIRICAL DATASET FOR INSTAGRAM TRAFFIC ANALYSIS (UNIVERSITY OF CALGARY, MARCH 3-9, 2019)

Item Description	Sun Mar 3	Mon Mar 4	Tue Mar 5	Wed Mar 6	Thu Mar 7	Fri Mar 8	Sat Mar 9	Overall
TCP Connections	896,849	2,355,640	2,313,701	2,352,614	2,253,556	2,055,827	853,820	13.1 M
Mean Duration	78.7 s	72.1 s	71.9 s	72.0 s	72.3 s	73.4 s	76.7 s	72.3 s
Packets Sent	264.3 M	565.3 M	565.2 M	561.9 M	550.3 M	509.0 M	283.3 M	3.3 B
Packets Received	550.9 M	1,003 M	953.9 M	931.1 M	950.7 M	910.2 M	589.9 M	5.9 B
Bytes Sent	32.2 GB	63.4 GB	60.4 GB	60.2 GB	60.0 GB	57.3 GB	33.3 GB	367 GB
Bytes Received	695 GB	1,259 GB	1,196 GB	1,167 GB	1,193 GB	1,141 GB	744.5 GB	7.2 TB
Client IP Addresses	1,450	1,679	1,605	1,532	1,621	1,547	1,449	3,498
IP Subnets	31	60	53	49	59	52	49	81

TABLE II
SUMMARY OF TCP CONNECTION STATES OBSERVED

State Description	Conns	%Conns	Bytes	%Bytes
SF: SYN-FIN	6,265,336	47.88%	3.78 TB	52.55%
RSTO: origin reset	2,487,505	19.01%	1.74 TB	22.91%
S3: no FIN seen	1,554,591	11.88%	879.9 GB	11.21%
S2: client FIN only	595,772	4.55%	340.1 GB	4.38%
S1: server FIN only	498,635	3.81%	189.7 GB	2.33%
RSTOS0: fail/RSTO	354,775	2.71%	222.9 GB	2.87%
RSTR: rcvr reset	335,304	2.56%	49.2 GB	0.63%
SH: no SYN-ACK	294,300	2.25%	107.1 GB	1.37%
SHR: no SYN seen	273,951	2.09%	57.3 GB	0.74%
OTH: other state	201,788	1.54%	71.3 GB	0.92%
S0: failed setup	166,822	1.27%	0.03 GB	$< 0.01\%$
REJ: rejected	37,455	0.29%	4.5 GB	0.06%
RSTRH: rcvr reset	20,329	0.16%	2.0 GB	0.03%
Total	13,086,563	100.0%	7.5 TB	100.0%

Figure 3 provides a more detailed look at the TCP connection state. This is a time series plot, with a one-minute time granularity for the week, and a small tick mark at midnight as a demarcation between each day. The graph shows the relative proportion of each (color-coded) TCP connection state in each one-minute interval. The purple colors represent SF (dark purple), S1, S2, and S3 (light purple), while S0 is yellow.

Two observations are evident from Figure 3. First, there is a strong diurnal pattern in the TCP connection states. At night time, the SF state dominates, while when the traffic load is higher during the day, many other TCP states are observed. Second, the downward "fingers" along the top of the graph represent the busiest parts of each day. When the traffic load is near its peak, there are more instances of unsuccessful connections (RSTOS0 in dark green) and server-side resets (RSTR in orange). There are also more half-open connections (SH in light blue, and SHR in dark blue), but these could be attributable either to the server, our monitor, or Internet congestion (i.e., packet losses) in transit.

C. TCP Connection Duration

The next analysis focuses on how long each TCP connection lasts. The TCP connection duration is reported in the logs, and represents the elapsed time between the first packet (usually a SYN) and the last packet (usually a FIN or a FIN/ACK) observed for a given TCP connection.

Figure 4 shows the results from our analysis of TCP connection duration. Specifically, Figure 4(a) shows the cumulative distribution function (CDF) on a linear scale, while Figure 4(b) shows the pdf on a log scale, and Figure 4(c) shows the log-log complementary distribution (LLCD).

There are several idiosyncracies in the connection duration distribution. First, about 12% of the connections have a duration of zero, since they consist of only a single packet. Second, there is a small peak near 4 seconds, since many of the failed (S0) connections give up after several unsuccessful retransmission attempts. Third, there is a large peak at 65 seconds; we attribute this to a default persistent connection timeout value for an idle TCP connection. Finally, there is another peak at 185 seconds. Again, this is due to a persistent connection timeout value used by the Instagram site. Through active measurements, we have determined that the 65-second timeout is used to terminate persistent connections when the Instagram app is closed, and the 185-second timeout is used when the app is still running in the background. Facebook's proxygen HTTP server also uses a 185-second timeout.

The CDF plot in Figure 4(a) shows a more detailed breakdown of the connection duration distribution based on TCP connection state. The S0 connections are the shortest, typically lasting 2-5 seconds, and appear in the upper left part of the plot. Successful (SF) connections have a wide range of durations, with the median near 65 seconds, the average near 83 seconds, and the longest observed connection (9,923 seconds) lasting almost 3 hours. The distributions for S1/S2/S3 connections are similar in shape to SF, though S1 connections tend to be much shorter, S2 connections only slightly shorter, and S3 connections tend to be much longer than SF. The "kinks" in these CDF plots for SF and S1/S2/S3 connections align with the persistent connection timeouts mentioned earlier, namely at 65 seconds and 185 seconds. Furthermore, many of the RSTR connections occur at exactly 65 seconds, suggesting that the reset is a mechanism for the server to reclaim needed resources. Finally, the RSTO connections tend to be shorter; this line falls between those for S0 and S1.

Since the TCP connection durations vary so widely, we apply a log-transform (base 2) to the duration data, and re-plot the distribution in Figure 4(b). Note that the vertical scale now is also logarithmic. This graph shows a wide-ranging distribution, from the single-packet connections with near-zero duration, to the connection that lasted almost 3 hours. The tallest peak in this distribution represents durations of 64 to 127 seconds.

The LLCD plot in Figure 4(c) provides a closer look at the tail of the distribution, on a log-log scale. In this graph,

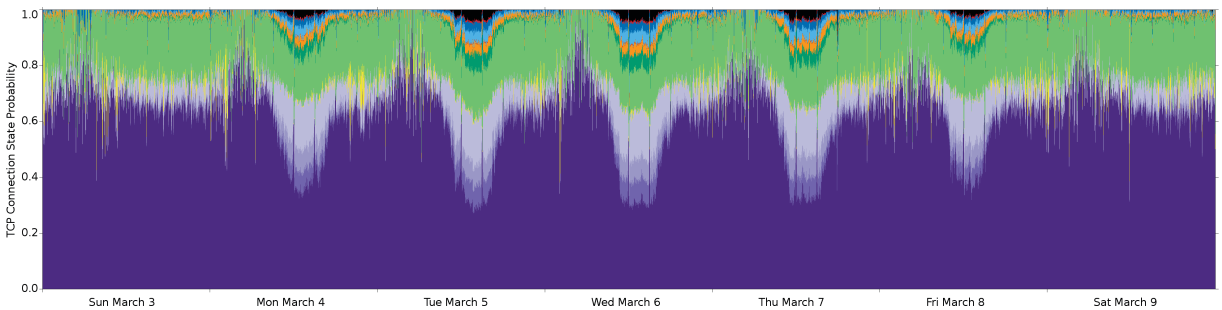
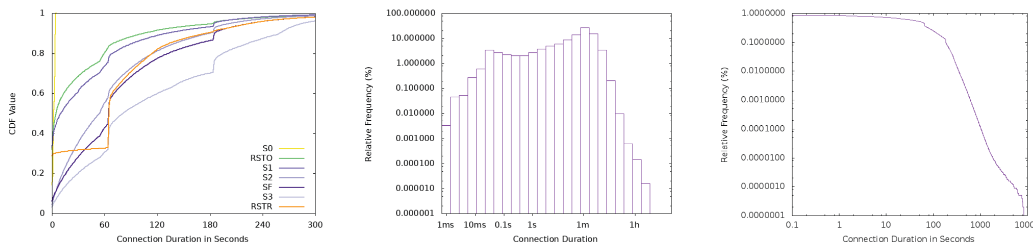


Fig. 3. Time Series Illustration of TCP Connection State in One-Minute Intervals for Instagram Traffic (March 3-9, 2019)



(a) CDF (linear scale) (b) pdf (log2 transform) (c) LLCD

Fig. 4. Distribution of TCP Connection Duration for Instagram Traffic

the straight line (slope -3.15, $R^2 = 0.9748$) indicates a power-law structure, implying a heavy-tailed distribution for the TCP connection duration. The tail of the distribution starts near 100 seconds, and spans almost all the way to 10,000 seconds.

D. Data Transfer Sizes

We next look at the data volumes exchanged on TCP connections, either inbound or outbound. We do so based on both packet counts and byte counts, as a sanity check on our data. The smallest packet sizes observed are 40 bytes, and the average packet size on the large transfers is around 1,440 bytes. Both these values make sense for typical TCP/IP implementations on the Internet.

Figure 5 shows LLCD plots for the number of packets sent and received on each TCP connection. Since this number varies widely, we apply a log-transform to this data, using base 2. In general, the received packet counts are slightly higher, though the two values are comparable since TCP uses ACK packets to ensure reliable data transfer. With TCP's delayed ACK strategy, the number of ACKs is typically half the number of data packets. The tail of the distribution has several connections with well over 100,000 packets, possibly for photos or streaming videos. The graph suggests that both distributions are heavy-tailed, based on the straight-line behavior in the tail of the distribution on the log-log plot (similar slopes for both: -1.85, $R^2 = 0.9887$).

Figure 6 shows the results for the number of bytes sent and received on each TCP connection. This number varies widely, from zero bytes to 600 MB, so we again apply a log-transform (base 2) to this data. Figure 6(a) shows the pdf of the resulting distribution for bytes sent, while Figure 6(b) shows the pdf

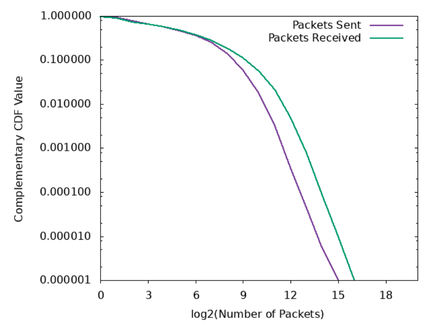


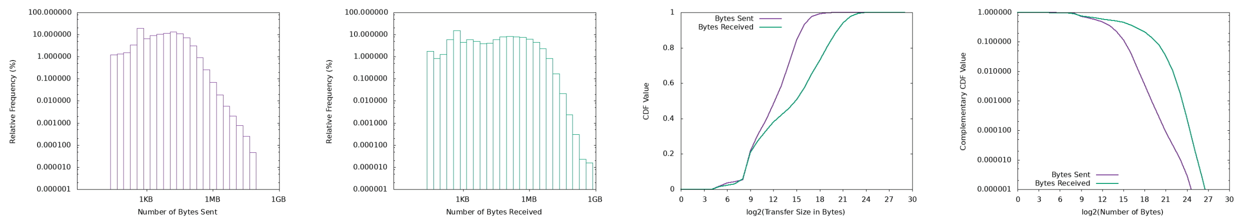
Fig. 5. Distribution of Transfer Sizes in Packets

for bytes received. The tallest bars are for connections with less than 1 KB of data, although the distribution continues well to the right. Two connections sent over 100 MB of data (possibly for live streaming or for video uploads), and two connections received over 600 MB of data. In general, the received byte counts are about an order of magnitude larger than the bytes sent, though this does not hold true for individual TCP connections that are uploading lots of data.

Figure 6(c) shows the CDFs of the two distributions, while Figure 6(d) compares their LLCD plots. Both distributions are heavy-tailed, with the bytes received (slope -1.53, $R^2 = 0.9633$) having a longer tail (slope -1.1, $R^2 = 0.9856$).

E. TCP Throughput

From the TCP connection durations and transfer sizes, it is possible to analyze the TCP throughput achieved, both for inbound and outbound data transfers. For smaller transfers,



(a) Bytes Sent pdf (b) Bytes Received pdf (c) CDF (d) LLCD

Fig. 6. Distribution of Transfer Sizes in Bytes

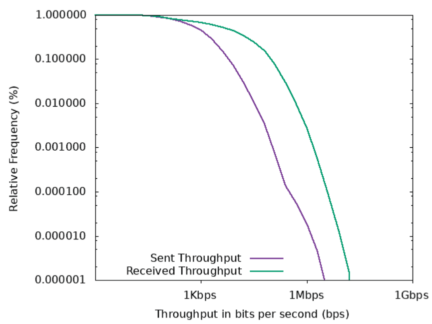


Fig. 7. LLCD Plot for TCP Connection Throughput

persistent connection timeouts bloat the duration, resulting in low average throughput. For larger transfers, however, this metric provides a good assessment of TCP performance, and the demands being placed upon the campus network and the Instagram servers.

Figure 7 shows the results from this analysis. Specifically, we only consider SF connections that last at least 1.0 seconds, of which there are 5.9 million. We again use our log transform (base 2) on the throughputs, which vary widely. In general, the received throughputs are much higher than the sending throughputs. The highest value observed for sending throughput was just over 10 Mbps, for a 1.6 MB transfer that completed successfully in 1.32 seconds (Tuesday 9:28am). For receiving throughput, the highest value observed was 65 Mbps, for a 12 MB transfer that completed successfully in 1.54 seconds (Monday 2:45am).

Figure 7 shows the LLCD plots for throughput. Both distributions have a pronounced tail. The receive throughputs are about an order of magnitude higher than the sending throughput for the largest transfers observed. Through active measurements, we have determined that the server supports the TCP window scaling option, which enables higher throughputs (but only if the client supports it as well).

F. Summary

Our characterization study of Instagram traffic has provided several interesting insights. First, the sheer volume of this traffic is staggering. On our campus network, the Instagram traffic averages over 1 TB of data downloaded per day on a normal weekday. This level of usage is third behind video streaming services such as Netflix (6 TB per day) and YouTube (4 TB per day), which together account for over half of the daily inbound traffic on our campus network [19]. Second, there is surprising consistency in the traffic from day to day, suggesting that Instagram users are creatures of habit. Third, the Instagram traffic exhibits power-law properties and heavy-tailed distributions like other information systems. For example, the TCP connection durations and the byte transfer size distribution are heavy-tailed. While these characteristics are similar for many Web and media streaming services, some features of Instagram traffic also appear to be unique (e.g., TCP states, connection durations).

V. Conclusion

This paper has presented a network traffic characterization study of Instagram, based on one week of data collected from a campus edge network. We studied the traffic profile, TCP connection states, transfer sizes, and throughput. We identified several trends in the time series data, such as diurnal patterns, consistency from day to day, and a noticable decline on weekends. In many of our results, we have found skewed distributions with high variability (e.g., transfer sizes, throughputs), and heavy-tails (e.g., connection durations, transfer sizes). These characteristics can have a large impact on a campus edge network.

There are several interesting future directions to consider. These include session-level characterization (without compromising any user identities), and the dynamics of user mobility. Finding an effective way to monitor network applications that use many dynamic IPs in a cloud-based infrastructure is also a measurement challenge.

More generally, Instagram is still a rather new social media service, and is still developing as an app and influencing user behavior online. Studies of the traffic/usage of Instagram as reported in this paper can provide a baseline for future studies of the social aspects of this technology. It will be interesting to see how the behavioral tendencies evolve over time, particularly as even more new features are added.

Acknowledgements

Financial support for this work was provided by Canada'a Natural Sciences and Engineering Research Council (NSERC). The authors thank University of Calgary Information Technologies (UCIT) for enabling the collection of our network traffic measurement data.

REFERENCES

[1] V. Adhikari, Y. Guo, F. Hao, V. Hilt, Z-L. Zhang, M. Varvello, and M. Steiner, "Unreeling Netflix: Understanding and Improving Multi-CDN Movie Delivery", *Proceedings of IEEE INFOCOM*, Orlando, FL, pp. 1620-1628, March 2012.

[2] V. Adhikari, Y. Guo, F. Hao, V. Hilt, Z-L. Zhang, M. Varvello, and M. Steiner, "Measurement Study of Netflix, Hulu, and a Tale of Three CDNs", *IEEE/ACM Transactions on Networking*, Vol. 23, No. 6, pp. 1984-1997, December 2015.

[3] J. Alstott, E. Bullmore, and D. Plenz, owerlaw: A Python Package for Analysis of Heavy-Tailed Distributions", 2014. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085777

[4] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization and Performance Implications" , *IEEE/ACM Transactions on Networking*, Vol. 5, No. 5, pp. 631–645, 1997.

[5] M. Arlitt and C. Williamson, "An Analysis of TCP Reset Behaviour on the Internet", *ACM Computer Communication Review*, Vol. 35, No. 1, pp. 37-44, January 2005.

[6] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, and M. Arlitt, "A Comparative Analysis of Web and Peer-fo-Peer Traffic", *Proceedings of WWW*, pp. 287-296, Beijing, China, April 2008.

[7] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing User Behavior in Online Social Networks", *Proceedings of ACM IMC*, pp. 49-62, Chicago, IL, November 2009.

[8] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and Modeling Popularity of User-generated Videos", *Proceedings of IFIP Performance*, Amsterdam, Netherlands, October 2011.

[9] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications", *Proceedings of IEEE INFOCOM*, New York, NY, pp. 126–134, March 1999.

[10] P. Cao, G. Li, A. Champion, D. Xuan, S. Romig, and W. Zhao, "On Human Mobility Predictability Via WLAN Logs", *Proceedings of IEEE INFOCOM*, Atlanta, GA, May 2017.

[11] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User-Generated Content Video System", *Proceedings of ACM IMC*, pp. 1-14, San Diego, CA, November 2007.

[12] M. Crovella and B. Krishnamurthy, *Internet Measurement: Infrastructure, Traffic and Applications*, John Wiley & Sons, 2006.

[13] Critical Stack LLC, "Bro Logs", 2014. http://gauss.ececs.uc.edu/Courses/c6055/pdf/bro_log_vars.pdf

[14] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A First Look at Traffic on Smartphones", *Proceedings of ACM IMC*, pp. 281-287, Melbourne, Australia, November 2010.

[15] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube Traffic: A View from the Edge", *Proceedings of ACM IMC*, pp. 15-28, San Diego, CA, November 2007.

[16] Y. Hu, L. Manikonda, and S. Kambhampati, "What We Instagram: A First Analysis of Instagram Photo Content and User Types", *Proceedings of 8th International AAAI Conference on Weblogs and Social Media*, pp. 595-598, Ann Arbor, MI, June 2014.

[17] Instagram, "A quick walk through our history as a company", March 2019. https://instagram-press.com/our-story/

[18] B. Krishnamurthy, P. Gill, and M. Arlitt, "A Few Chirps About Twitter", *Proceedings of Workshop on Online Social Networks*, pp. 19-24, Seattle, WA, August 2008.

[19] M. Laterman, M. Arlitt, and C. Williamson, "A Campus-Level View of Netflix and Twitch: Characterization and Performance Implications", *Proceedings of SCS SPECTS*, pp. 15-28, Seattle, WA, July 2017.

[20] A. Madhukar and C. Williamson, "A Longitudinal Study of Peer-to-Peer Traffic Classification", *Proceedings of IEEE/ACM MASCOTS*, Monterey, CA, USA, pp. 179-188, September 2006.

[21] A. Mahanti, N. Carlsson, A. Mahanti, M. Arlitt, and C. Williamson, "A Tale of the Tails: Power-Laws in Internet Measurements", *IEEE Network*, Vol. 27, No. 1, pp. 59-64, January 2013.

[22] A. Mahanti, C. Williamson, and M. Arlitt, "Remote Analysis of a Distributed WLAN Using Passive Wireless-Side Measurement" *Performance Evaluation*, Vol. 64, No. 9-12, pp. 909-932, October 2007.

[23] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", *Proceedings of ACM IMC*, pp. 29-42, San Diego, CA, October 2007.

[24] M. Mitzenmacher, "A Brief History of Generative Models for Power Law and Lognormal Distributions", *Internet Mathematics*, Vol. 1, No. 2, pp. 226-251, 2003.

[25] A. Nazir, S. Raza, and C. Chuah, "Unveiling Facebook: A Measurement Study of Social Network Based Applications", *Proceedings of ACM IMC*, pp. 43-56, Vouliagmeni, Greece, October 2008.

[26] A. Nazir, S. Raza, D. Gupta, C. Chuah, and B. Krishnamurthy, "Understanding Online Social Network Usage from a Network Perspective", *Proceedings of ACM IMC*, pp. 63-75, Chicago, IL, November 2009.

[27] V. Paxson, "Bro: A System for Detecting Network Intruders in Real-time", *Computer Networks*, Vol. 31, No. 23, pp. 2435-2463, Dec 1999.

[28] S. Roy and C. Williamson, "Why is my LMS so slow? A Case Study of D2L Performance Issues", *Proceedings of ISCA Computers and Their Applications (CATA)*, Las Vegas, NV, pp. 34-39, March 2018.

[29] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding Online Social Network Usage from a Network Perspective", *Proceedings of ACM IMC*, pp. 35-48, Chicago, IL, November 2009.

[30] Variety, "Facebook, Instagram Apps Suffer Hours-Long World-wide Outages for Some Users", https://variety.com/2019/digital/news/facebook-instagram-apps-outage-worldwide-1203162777/

[31] Wikipedia, "Digital Native", March 2019. https://en.wikipedia.org/wiki/Digital_native

[32] Wikipedia, "Instagram", March 2019. https://en.wikipedia.org/wiki/Instagram

[33] C. Williamson, "Internet Traffic Measurement", *IEEE Internet Computing*, Vol. 5, No. 6, pp. 70–74, November/December 2001.

[34] Z. Zhang and C. Williamson, "A Campus-level View of Outlook Email Traffic", *Proceedings of ICNCC*, Taipei, Taiwan, pp. 299–306, Dec 2018.