Mass-Count Disparity in Mobile Traffic

Jaeyoon Chung, Youngjoon Won, Byungchul Park, and James Won-Ki Hong

***Abstract*—Mass-count disparity is a basis of the elephants and mice phenomenon in Internet traffic analysis. Mobile applications tend to minimize transmission overhead by reducing object size. Assuming the mice get smaller, we first look at properties of mass-count disparity in the smart device traffic. We find the existence of elephants and represent an accurate inequality measure using the Gini coefficient where the heavy-tail property is not clearly visible in the trace. The Gini coefficients range from 0.71 (web) to 0.98 (application market traffic), implying a heavy inequality distribution toward 1. The cutoff point of elephants is 1.55 MB that is even comparable with small size photos. Our hypothesis from the early analysis indicates that every mobile user is potentially generating elephant flows. We observe that a significant stance of application market traffic is responsible for such phenomenon.**

***Index Terms*—Measurement, mobile traffic.**

I. Introduction

MASS-COUNT disparity is known as the *elephant and mice phenomenon* of the wired Internet traffic. It indicates that a majority of traffic is carried over a small number of connections [1]. However, there is no fixed boundary of determining elephant flows. It all differs case-by-case depending on the strategy of how to handle such flows in the network [2]. A few previous studies focused on identifying elephants in the sampled packets against its sampling ratio [3] and defining a correlation boundary between size and duration [4].

The Cisco VNI 2014-2019 reports that global mobile data traffic will grow three times faster than fixed IP traffic [5]. In this letter, we look at the properties of mass-count disparity in mobile device traffic with respect to its application type. We collected hourly WiFi traces at the Internet junction of an university in Korea. Our dataset was the captured traffic behind roughly 1,000 NAT-enabled access points. MAC address information is not revealed which leaves us blind to the number of distinct users in the dataset. Thus, we conducted application-based traffic classification rather than host-based, by using deep packet inspection (DPI). First, we discriminate the smart device traffic only from the trace by looking up the HTTP *user-agent* field as classifier [6] [7]. (e.g., 'User-Agent: AppleCoreMedia.*iPad'). Second, we relied on payload signatures [8] to identify the non-HTTP traffic, such as over-the-top contents, with invalid user-agent.

Manuscript received July 14, 2015; revised October 20, 2015; accepted October 22, 2015. Date of publication October 30, 2015; date of current version January 7, 2016. This work was supported by the National Research Foundation of Korea under Grant NRF-2014R1A1A2057301 and ICT R&D program of MSIP/IITP [B0190-15-2011]. The associate editor coordinating the review of this paper and approving it for publication was G. Reali.

J. Chung is with the Princeton University, Princeton, NJ 08540 USA.

Y. Won is with the Department of Information System, Hanyang University, Seoul, South Korea (email: youngjoon@hanyang.ac.kr).

B. Park is with the University of Toronto, Toronto, ON M5S 3G4, Canada.

J. Won-Ki Hong is with the POSTECH, Pohang, South Korea.

Digital Object Identifier 10.1109/LCOMM.2015.2496594

We investigate any existence of elephant and represent an accurate inequality measure, the Gini coefficient, where the heavy-tail property is not clearly visible in the trace. Our hypothesis from the early analysis indicates that every mobile user is potentially generating elephant flows. Due to the nature of the mobile ecosystem [7], we observe that a significant stance of 'app market' traffic contributes to such phenomenon.

The organization of this letter is as follows. We present the properties revealing mass-count disparity in Section II. Section III discusses elephants in smart device traffic. Section IV concludes and discusses further work.

II. Properties of Mass-Count Disparity

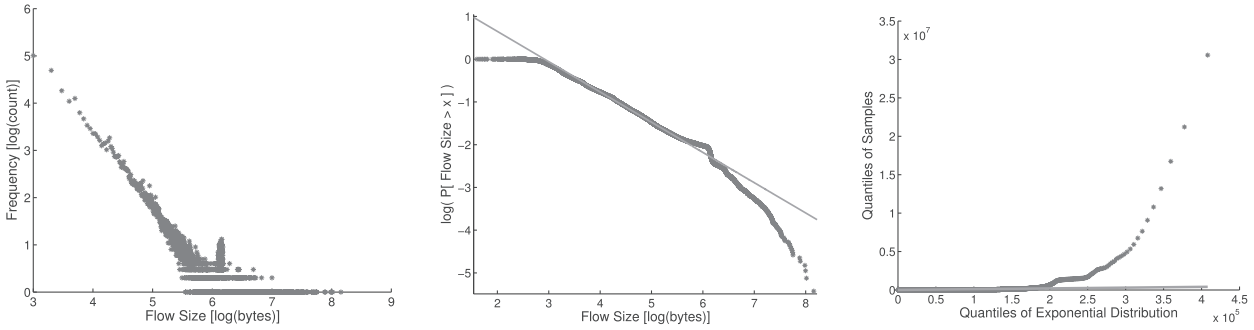
A. Fitting Heavy-Tail Distributions

The power law states that $y = bx^{-a}$ where y is the number of flows whose size is x. If flow size distribution follows the power law, the distribution is a straight line with slope $-a$ in log-log scale. Figure 1(a) is a log-log scale plot of the flow size distribution with 1 KB bin size. It is not clear to fit a linear function to this sample group due to a lack of valid samples in the greater flow size ranges. We used the Pareto cumulative distribution such that $P[X > x] \sim x^{-a}$ to smooth the tail (Figure 1(b)). Thus, the slope is -0.708 with 95% confidence and the confidence bound is $(-0.7088, -0.7072)$.

In the wired traffic, the flow size followed the heavy-tail distribution with infinite mean (since $a < 2$) [9]. Figure 1(b) is not scale-free due to multiple possible slopes, other than a single fitted line with 95% confidence. A quantile-quantile (Q-Q) plot is to compare a theoretical distribution. Figure 1(c) shows that the measured flow size (dots) does not fit an exponential distribution (gray line). If the samples are linearly related with the distribution, those in the Q-Q plot should be lined parallel to the diagonal line (gray). When the flow size is large, the measured flow size does not matched with an exponential distribution where the heavy-tali property is not clearly visible.

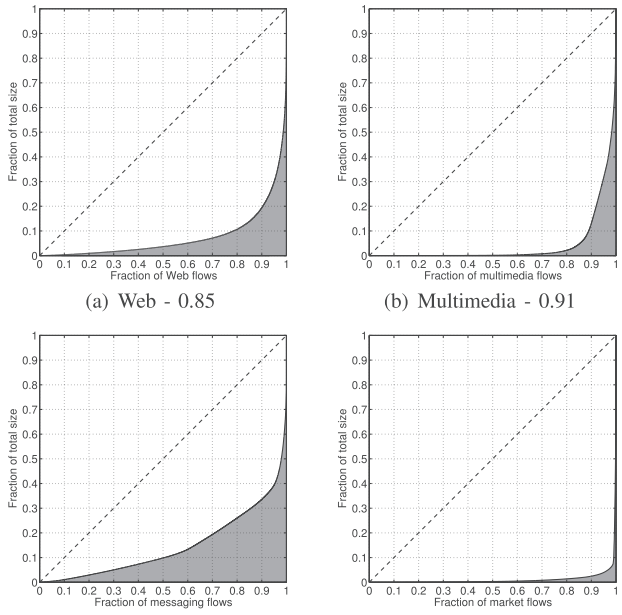
B. Mass-Count Measures

Figure 2 presents the mass-count disparity of mobile traffic using the Gini coefficient and the Lorenz curve. Gini coefficient is commonly used in the field of economics to explain inequality of wealth. It ranges from 0 to 1 where higher values indicate more unequal distribution with 1 being complete inequality. The Lorenz curve illustrates the cumulative distribution function of inequality [10]. For web, the top 20% flows occupy 90% of the traffic volume (Figure 2(a)). For multi-media, the top 20% flows occupy more than 95% of the traffic volume, and the bottom 60% of multimedia flows occupies almost 0%



(a) flow size and frequency distributions (b) CDF of flow size; slope $= -0.708$. (c) Q-Q plot; flow size and its exponential distribution

Fig. 1. Heavy-tail distributions of mobile traffic.



(a) Web - 0.85 (b) Multimedia - 0.91

(c) Messaging - 0.711 (d) App Market - 0.98

Fig. 2. Lorenz curve with the Gini coefficients for smart device traffic. 0 means all flows have the same flow size. 1 means complete inequality, i.e. a single flow occupies entire traffic volume.

of the volume of traffic (Figure 2(b)). Conversely, we observed that the flows for messaging applications consist of small size flows. Figure 2(c) shows that the top 30% of messaging flows accounts for 80% of the traffic volume. The Gini coefficient of app market traffic is over 0.98 and the top 1% of flows accounts for over 91% of the traffic volume (Figure 2(d)). The app market download traffic shows the highest inequality, and the messaging traffic shows relatively lower than the others. *Joint ratio* is the simplest metric to measure the degree of inequality such as the 90/10 rule and the 80/20 rule. The *joint ratio* of flow size in the mobile traffic ranges from 75/25 to 95/5.

C. Revisiting Elephants in Mobile Traffic

We define elephants as a subset of mobile traffic where the flow size is greater than the mean plus two standard deviations of flow size ($\mu_{size} + 2\sigma_{size} = 1.55$ MB) [11]. The empirical analysis was necessary to cope with our mobile traffic trace. We applied mean plus two standard deviation while three standard

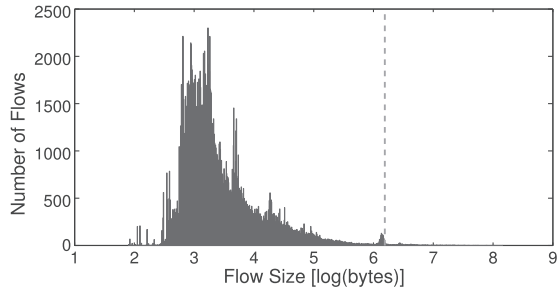


Fig. 3. Log scaled histogram of flow size in mobile traffic; the vertical dashed line at 6.19 ($= \log 1.55MB$) indicates a cut-off point for elephants.

deviation showed the statistically insignificance, showing too low target flow population. Figure 3 presents that the log-scale flow size distribution is right-tailed whose mean and median are 44.28 KB and 1.797 KB respectively. The vertical dashed line at 6.19 ($= \log 1.55MB$) is the cut-off point between elephants and non-elephants. Due to the high inequality of flow size, the statistical difference between them is clearly observed. The mean and median of elephants are 6.26 MB and 3.06 MB, and those of non-elephants are 19.8 KB and 1.80 KB.

Table I and II show any (non-)elephants among all flows sorted by application types. The first column indicates whether the application traffic of each row belongs to elepahnts, long-durations, or both. The occurrence ratio indicates a portion of flows by count; the occupation ratio is by traffic volume. The elephants take 0.39% of all flows but are responsible for 55.49% of traffic volume. The web browsing and messaging applications generate a larger volume of non-elephants, but even more traffic by multimedia streaming and market download applications is transferred via elephants. The occupation ratios of elephants by market and web browsers are close to each other, 5.93% and 5.32%. The occurrence ratio of elephants by market is 5 times less than that of web browsers. 74.8% of elephants are generated by multimedia, which is 6 times higher than web browsers.

We performed a similar breakdown for flow duration. We define *long-duration flow* in which its duration is longer than the mean plus two standard deviations ($\mu_{duration} + 2\sigma_{duration} = 73.21$ sec). Long-duration flows are 2.03% of all flows but occupy 22.57% in volume. Its count is 5.21 times higher than that of elephants, but the volume is 2.45 times less. For multimedia and market, the percentages of long-duration

TABLE I
OCCURRENCE RATIOS OF ELEPHANT (E) AND LONG-DURATION (L) FLOWS (UNIT: %)

Category	W	MM	MSG	MK	Misc.	Total
E	0.05	0.31	<0.01	0.01	0.02	0.39
$\neg E$	79.78	7.50	9.31	1.00	2.01	99.61
L	0.45	0.67	0.62	0.06	0.21	2.03
$\neg L$	79.38	7.14	8.69	0.95	1.82	97.97
$E \wedge L$	0.01	0.04	N/A	<0.01	<0.01	0.06
$E \wedge \neg L$	0.04	0.27	<0.01	0.01	0.02	0.34
$\neg E \wedge L$	0.45	0.63	0.62	0.06	0.21	1.97
$\neg E \wedge \neg L$	79.34	6.87	8.69	0.94	1.80	97.64

W: Web; MM: Multimedia; MSG: Messaging; MK: Market;

TABLE II
OCCUPATION RATIOS OF ELEPHANT (E) AND LONG-DURATION (L) FLOWS (UNIT: %)

Category	W	MM	MSG	MK	Misc.	Total
E	5.32	41.52	0.28	5.93	2.44	55.49
$\neg E$	19.10	23.22	1.27	0.50	0.42	44.51
L	2.43	17.78	0.09	1.78	0.49	22.57
$\neg L$	21.99	46.96	1.46	4.65	2.37	77.43
$E \wedge L$	2.10	16.59	N/A	1.73	0.42	20.85
$E \wedge \neg L$	3.22	24.93	0.28	4.20	2.02	34.64
$\neg E \wedge L$	0.33	1.18	0.09	0.04	0.07	1.72
$\neg E \wedge \neg L$	18.77	22.03	1.18	0.45	0.35	42.79

W: Web; MM: Multimedia; MSG: Messaging; MK: Market;

flows are 27.46% and 27.68%, respectively. These two are higher than web browsers (9.95%) and messaging (5.81%). Thus, we observe a higher involvement of long-duration flows in multimedia and market.

Our traffic breakdown showed that 84.6% of elephants are not-long-duration flows (Table I and II). Flows belonging to both elephants and long-duration are 15.4% of elephants and 3.00% of long-duration flows in the occurrence ratios. Zhang *et. al* [12] analyzed the correlations among flow size, duration, and bandwidth of non-mobile traffic only. They showed that the correlation of size–duration is *very weak* (0.100 to 0.296), the negative correlation of bandwidth–duration is *fairly weak* (-0.453 to -0.187), and the correlation of size–bandwidth is *very strong* (0.835 to 0.885). In contrast to the characteristics of non-mobile traffic, Table III shows the correlations of flow size, duration, and bandwidth of (non-)elephants in our dataset. The strongest but *fairly weak* correlation is shown between size and duration of elephants where its correlation coefficient is 0.4455. Excluding the correlation of size–duration, the correlations of bandwidth–duration and size–bandwidth in elephants are *very weak* showing -0.2802 and 0.1493, respectively. Non-elephants do not show any significant correlation among size, duration, and bandwidth.

To show a pattern of bandwidth consumption, Figure 4 illustrates empirical CDFs of bandwidth of application types in different categories. We noticed that the bandwidth of elephants is significantly higher than that of non-elephants (Figure 4(a) and 4(b)). Over 90% of elephants consume over 100 kbps, and the CDF curves rapidly increase around 1 Mbps. In contrast to elephants, the bandwidth distribution of most non-elephants ranges from 100 bps to 1 Mbps. Its CDFs are gradually increased between 1 kbps and 1 Mbps regardless of application types. Figure 4(c) and 4(d) show the bandwidth distributions

TABLE III
CORRELATIONS OF SIZE, DURATION, AND BANDWIDTH IN (NON-)ELEPHANTS

Elephant			
	Size	Duration	Bandwidth
Size	1	0.4455	0.1493
Duration	0.4455	1	-0.2802
Bandwidth	0.1493	-0.2802	1
Non-Elephant			
	Size	Duration	Bandwidth
Size	1	0.0392	0.0043
Duration	0.0392	1	-0.0019
Bandwidth	0.0043	-0.0019	1

of long-duration and not-long-duration flows. Except for the multimedia traffic, the bandwidth of long-duration flows ranges from 10 bps to 10 kbps, which is significantly lower than elephants. Multimedia flows with long-duration can be classified into three ranges: 100 bps, 10 kbps, and 1 Mbps. This observation is an evidence for multiple streaming strategies for video applications. It implies that elephants are more significant than long-duration flows in bandwidth management.

III. DISCUSSION

A. Spotting Elephants From Every User

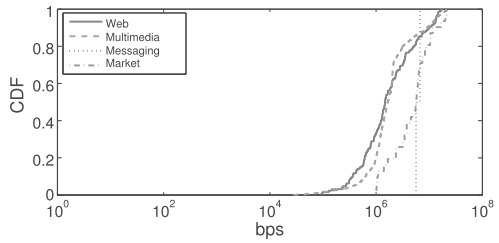
The number of daily downloads from the U.S. App Store in May 2015 was 6.7 million. We analyzed the size distribution of 5,519 top ranking applications and showed that the flow size via application market was equal the application size in download. Assuming 95% of market flows were classified as elephants (see Table I), mobile networks should handle over 6.3 millions of burst download flows. Since the mobile market is the only source of application download, there is a greater chance that every user becomes a heavy hitter. The increase of application size [13] itself also contributes to this observation.

The coexistence of mobile and non-mobile contents is another cause of elephants. Mobile webpages remove embedded objects such as images and Flash ad to minimize traffic volume as well as to fit into the smaller displays of mobile devices [14]. We examined popular web sites, such as YouTube, Facebook, and CNN, to compare the average size of desktop-version and mobile-version webpages. The average size for desktop and mobile devices are 1,898 KB and 195 KB, respectively. It implies that downloading relatively large objects in desktop-version webpages from mobile device is also relevant to elephants.

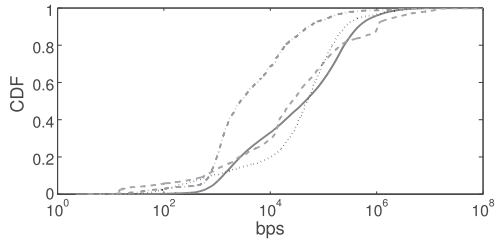
B. Handling Elephants With Caution

For now, traffic offloading from 3G/4G to WiFi or vice versa can be manually triggered by user. For example, video streaming applications issue a traffic overuse warning when user is on 3G/4G. *Instapaper*[1] and *Pocket* (previously *Read It Later*)[2] induce users to pre-fetch the web contents via WiFi. On-going flows should be closed and re-established later through a newly connected and preferred network. Although most mobile

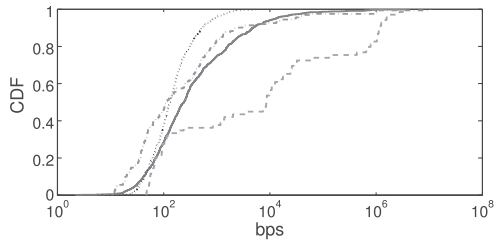
[1]https://www.instapaper.com/
[2]https://getpocket.com/



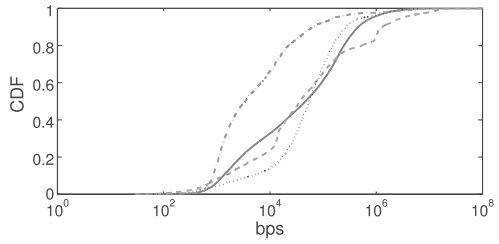
(a) Elephant



(b) Non-elephant



(c) Long-duration



(d) Not-long-duration

Fig. 4. Empirical CDFs of bandwidth according to flow size and duration breakdown: Elephant vs. Non-elephant vs. Long-duration vs. Not-long-duration.

devices support simultaneous multi-access of 3G/4G and WiFi, the network selection for offloading is still dependent on user decision.

Our measurement showed that which IP flows should be offloaded. If we can offload elephants contributing to 0.39% of all flows, the load shifting ratio can reach upto 55.49%. The need for flow-level offloading was also discussed in a 3GPP specification, Network based IP Flow Mobility [15]. However, flow selection strategy has not been discussed yet where most ISP's management policies were designed for profit-maximization purposes [16].

One challenge for offloading is the early detection of elephants. Before transmitting entire packets over a single connection, we require an accurate and efficient method to determine elephants in an early stage.

IV. Conclusion

In this letter, we presented several properties of mass-count disparity in smart device traffic. Our measurement results strongly support the basic insight of mobile traffic characteristics in terms of flow size. We find a high inequality in flow size distribution for multimedia and application market; the Gini coefficients are 0.91 and 0.98, respectively. This inequality results in the existence of elephants suggesting a volume boundary of $\geq$ 26.59MB. Our analysis illustrates that a significant stance of application market traffic is responsible for elephants and more than a half of the whole bandwidth consumption. Elephants are more significant than long-duration flows in bandwidth management. An opportunity for ISP-friendly offloading elephants in WiFi/3G/LTE mix networks is present where 0.39% of all flows, elephants, can relieve 55.49% of the smart device traffic in volume.

References

[1] L. Guo and I. Matta, "The war between mice and elephants," in *Proc. 9th Int. Conf. Netw. Protocols (ICNP'01)*, 2001, pp. 180–188.

[2] C. Estan and G. Varghese, "New directions in traffic measurement and accounting," *SIGCOMM Comput. Commun. Rev.*, vol. 32, no. 4, pp. 323–336, Aug. 2002.

[3] T. Mori, M. Uchida, R. Kawahara, J. Pan, and S. Goto, "Identifying elephant flows through periodically sampled packets," in *Proc. 4th ACM SIGCOMM Conf. Internet Meas. (IMC'04)*, 2004, pp. 115–120.

[4] K.-C. Lan and J. Heidemann, "A measurement study of correlations of Internet flow characteristics," *Comput. Netw.*, vol. 50, no. 1, pp. 46–62, Jan. 2006.

[5] Cisco, "Cisco visual networking index: Forecast and methodology, 2014–2019," White Paper, May 2015, [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white\s\do5(p)aper\s\do5(c)11-481360.pdf.

[6] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas. (IMC'10)*, 2010, pp. 281–287.

[7] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf. (IMC'11)*, 2011, pp. 329–344.

[8] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of P2P traffic using application signatures," in *Proc. 13th Int. Conf. World Wide Web (WWW'04)*, 2004, pp. 512–521.

[9] M. Crovella and B. Krishnamurthy, *Internet Measurement: Infrastructure, Traffic and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[10] D. G. Feitelson, "Metrics for mass-count disparity," in *Proc. 14th IEEE Int. Symp. Model. Anal. Simul. (MASCOTS'06)*, 2006, pp. 61–68.

[11] T. Mori, R. Kawahara, S. Naito, and S. Goto, "On the characteristics of internet traffic variability: Spikes and elephants," in *Proc. Int. Symp. Appl. Internet*, Tokyo, Japan, 2004, pp. 99–106.

[12] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker, "On the characteristics and origins of Internet flow rates," *SIGCOMM Comput. Commun. Rev.*, vol. 32, no. 4, pp. 309–322, Aug. 2002.

[13] CNET. (2013). *Apple Quietly Doubles Cellular App Download Limit to 100 MB* [Online]. Available: http://www.cnet.com/news/apple-quietly-doubles-cellular-app-download-limit-to-100 mb/.

[14] Y. Chen, X. Xie, W.-Y. Ma, and H.-J. Zhang, "Adapting web pages for small-screen devices," *IEEE Internet Comput.*, vol. 9, no. 1, pp. 50–56, Jan. 2005.

[15] GT 23.261, "IP flow mobility and seamless wireless local area network (WLAN) offload; Stage 2," Version 12.0.0 Release 12, Sep. 2014.

[16] A. de la Oliva, C. Bernardos, M. Calderon, T. Melia, and J. Zuniga, "IP flow mobility: Smart traffic offload for future wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 10, pp. 124–132, Oct. 2011.