Revisiting Coexistence of Poissonity and Self-Similarity in Internet Traffic

Himanshu Gupta[‡,⋆]
[‡]IBM Research Laboratory
New Delhi, India
higupta8@in.ibm.com

Anirban Mahanti
NICTA
Australia
anirban.mahanti@nicta.com.au

Vinay J. Ribeiro
[⋆]Indian Institute of Technology Delhi
New Delhi, India
vinay@cse.iitd.ac.in

Abstract—**The immense popularity of new-age "Web 2.0" applications such as YouTube, Flickr, and Facebook, and non-Web applications such as Peer-to-Peer (P2P) file sharing, Voice over IP, online games, and media streaming have significantly altered the composition of Internet traffic with respect to what it was a few years ago. In light of these changes, this paper revisits Internet traffic characteristics and models that were proposed when "traditional" Web traffic was the largest contributor to Internet traffic. Specifically, we study whether or not the following characteristics, namely: (1) traffic is *self-similar* and *long-range* dependent, and (2) traffic can be approximated by *Poisson* at smaller time scales, are still valid. Our experiments on recent traces show that these traffic characteristics continue to hold. We further argue that current Internet traffic can be viewed to have two key constituents, namely Web+ and P2P+; Web+ traffic consists of traffic from both Web 1.0 and Web 2.0 applications; P2P+ traffic consists largely of traffic from P2P applications and other non-Web applications excluding applications on well-known ports such as FTP and SMTP. We then show that both Web+ and P2P+ components exhibit self-similar behavior and can be approximated by Poisson at smaller time scales.**

I. Introduction

Traditionally, data traffic was modeled by Poisson processes. Poisson processes are characterized by packet interarrival times that are distributed exponentially and are independent of each other. In addition, the number of Poisson arrivals in non-overlapping time segments are independent. When such a process is aggregated to large time-scales, the law of large numbers applies and the aggregated process tends to the mean quickly. Visually the aggregated process appears "smooth" and non-bursty.

Since the early 1990s, numerous studies have demonstrated that Internet traffic when aggregated to large time-scales does not appear smooth and is in fact quite bursty [18]–[21], [23]. Such behavior is typical of *long-range dependent* (LRD) processes where there is a strong (non-summable) correlation between number of arrivals in different time segments. This strong correlation inhibits the smoothing that normally takes place with aggregation of a weakly correlated process.

The seminal work of Leland et al. showed that traffic was in fact well-modeled by a sub-class of LRD processes called second-order self-similar processes [20], [21]. These processes essentially "look-similarly bursty" at different time-scales. Willinger et al. [27] unraveled the physical causes of self-similarity in traffic; their work showed that Internet traffic could be viewed as a superposition of heavy tailed ON/OFF processes which gives rise to self-similarity [20]. Crovella et al. [12] later supported these findings by analyzing Web browsing traffic collected in 1994-95.

However, despite the ubiquitous presence of LRD, Karagiannis et al. demonstrated that Internet traces collected in 2002-03 do appear to possess certain Poisson-like characteristics at small timescales such as packet interarrivals that are exponential and have autocorrelation close to zero [18]. This finding of exponential inter-arrival times is in contrast to earlier ones which showed that the distribution of packet interarrivals on the Internet are far from exponential [13], [16] and implies that simple Poisson models can still be used for design and optimization of network systems.

The primary objective of this paper is to revisit and investigate the coexistence of self-similarity and Poissonity in recent Internet traffic. We want to revisit these models as the composition of Internet traffic has changed in the last six years. Today's Internet traffic consists of various modern applications such as P2P (BitTorrent, Gnutella) file sharing, VoIP (Skype), online games, and video sharing portals (YouTube, Hulu). P2P traffic has steadily increased and this traffic has become a major constituent of Internet traffic [6], [14], [17]. The emergence and vast popularity of video portals like YouTube and Hulu, however, has likely restored the balance and some studies report that Web traffic has once again overtaken P2P traffic [4], [5] in 2007. Skype traffic has increased from 6.9 billion minutes in the forth quarter of 2006 to 20.5 billion minutes during the same time frame in 2008; Skype is adding 30 million new subscriptions every quarter and Skype minutes are only expected to grow [3]. Similarly, online gaming market is expected to hit $4.4 billion in 2010 up from $1.7 billion in 2006 [2]. A multitude of other applications like newsgroups, Facebook, and Flickr have become very popular in last few years and thus a snapshot of today's Internet traffic looks quite different to five years back [6], [7], [9], [10], [15].

Traffic of these applications may have very different properties as compared to traditional Web traffic and can thus affect Internet traffic characteristics. For example, P2P traffic introduces many elephant flows ($\geq$ 5 MB) [6] while HTTP traffic does not, P2P connections are long lived [6] while HTTP connections are short, video transfer sizes from YouTube are orders of magnitude larger than non-video content transfers [15], and so on. Secondly, overall traffic and as a result link utilization has gone up in last six years primarily due to wide usage of P2P applications and video sharing portals. High link utilization may also result in different Internet traffic models and characteristics.

The primary contribution of this paper is an intensive analysis of recent traces (publicly available) from a high speed backbone link to establish the coexistence of Poissonity and self-similarity. These traces differ in two significant ways from traces used by Karagiannis et al. [18] wherein the coexistence of Poissonity and long-range dependence was first demonstrated. First, compared to earlier traces, these traces have been collected from a backbone link with significantly higher utilization (≈3 times). Second, earlier traces were collected in the pre-YouTube era[1] while our traces are likely to contain YouTube content because such traffic is widely believed to be a significant part of current Internet traffic [4], [5]. Overall, our traces reflect, as discussed above, the changed composition of Internet traffic.

Our analysis finds that recent Internet traffic exhibits biscaling behavior in Hurst parameter. Specifically, we find that the Hurst parameter is significantly smaller at small time scales as compared to the Hurst parameter at larger time scales, with the point of transition being around 100 ms. We further notice that the behavior of recent traces over an interval of few seconds (≈ 5 s) can be nicely modeled by Poisson (i.e., packet interarrivals are found to be independent and exponential).

In addition, we notice that modern Internet traffic can be viewed to consist of two main components, namely Web+ and P2P+. P2P+ traffic is defined as P2P traffic as well as traffic due to other applications such as online games and Skype which also use dynamic and random ports rather than well defined default ports. Web+ consists of both traditional Web browsing traffic and traffic owing to audio/video sharing applications such as YouTube. We extract Web+ and P2P+ components and establish coexistence of Poissonity and self-similarity in Web+ and P2P+ traffic components as well. To the best of our knowledge, no study has looked at recent Web traffic (Web+) for verifying the assumption of coexistence of Poissonity and self-similarity. Some studies have investigated self-similarity in individual components of P2P+ traffic such as BitTorrent [22] and individual online games e.g. 'Call of Duty' [8]; however, to the best of our knowledge this is the first study which analyzes P2P+ traffic for the coexistence of Poissonity and self-similarity.

The remainder of this paper is organized as followed. Section II gives a brief introduction of Poissonity, self-similarity, and statistical tests to identify them. Section III summarizes related work. Section IV describes methodology and the traces used in this paper. Section V analyzes the recent traces as well as Web+ and P2P+ components and shows that they can be approximated by Poisson distribution when analyzed over time-intervals of few seconds. Section VI analyzes the recent traces and shows that they display self-similar behavior and models recent traces using heavy tailed ON/OFF process to explain the self-similar behavior. Web+ and P2P+ components are also shown to display self-similar behavior. Section VII summarizes the paper.

II. Background

A. Poisson Process and Statistical Tests

A stochastic arrival process is said to be Poisson having rate $\lambda, \lambda > 0$, if the interarrival times X_1, X_2,... have a common exponential distribution function [24]:

$$P(X_n \leq t) = 1 - e^{-\lambda t}, t \geq 0. \quad (1)$$

The average interarrival time is given by $\frac{1}{\lambda}$. All the arrivals are independent of each other and the number of arrivals occurring in a given time interval depends only on the length of the interval.

To evaluate whether a process is Poisson or not [18], we need to test whether the process is exponentially distributed and is consistent with independent arrivals. A linear behavior of Complementary Cumulative Distribution Function (CCDF) with y-axis on log scale indicates an exponential distribution. To check whether arrivals are uncorrelated, we compute autocorrelation coefficients (ACF) at various lags. ACF of a time series $X_n; n = 1, 2, \ldots, \infty$ at lag k, is defined as its normalized auto-covariance:

$$r(n,k) = \frac{Cov(X_n, X_{n+k})}{Var(X)} = \frac{E[X_n X_{n+k}] - E^2[X]}{Var(X)} \quad (2)$$

If the arrivals are truly uncorrelated, the estimated ACF is approximately normally distributed with mean 0 and variance $\frac{1}{N}$ and hence most of ACF values lie within 95% confidence interval $\pm \frac{2}{\sqrt{N}}$ where N is the number of packet inter-arrivals.

Index of Dispersion for intervals (IDI): Let S_k denote the sum of k consecutive inter-arrival times $S_k = X_1 + X_2 + ... + X_k$. The IDI or k-interval squared coefficient of variation is defined as:

$$c_k^2 = \frac{kVar(S_k)}{[E(S_k)]^2} \quad (3)$$

IDI of an ideal Poisson process is equal to 1 for all k. If the arrival process has higher variance than Poisson at some time scale, then the index tends to increase as a function of k. c_k^2 only depends upon the length of the series used, not on any specific part of the trace [25].

B. Self-Similarity, Long-Range Dependence, and Statistical Tests

Self-similarity and long-range dependence (LRD) are closely related phenomenon. Self-similarity refers to the phenomenon where a process aggregated at different time scales has similar structure and various statistical properties such as mean, variance, and marginal distribution remain the same (under a transformation). Note that for a process X, its aggregated process $X^{(m)}$ is given as $X^{(m)}(k) = \frac{1}{m}\sum_{i=km-m+1}^{i=km} X(i), k = 1, 2, \ldots, \infty$. LRD refers to the phenomenon where the correlations in data across large lags,

[1] YouTube was launched in Nov 2005.

though decreasing, never become insignificant. Precise definitions of self-similarity and LRD follow.

A process X is called exactly second-order self-similar with self-similarity parameter $H = 1 - \frac{\beta}{2}$ if

- $X(n)$ is wide sense stationary (WSS). A process is called WSS if its first two moments (mean and variance) do not vary with time.
- for all $m = 1, 2, \ldots, \infty$

$$Var(X^{(m)}) = \sigma^2 m^{-\beta} \tag{4}$$

A WSS process is LRD if its autocorrelation is non-summable, that is $\sum_k r(k) = \infty$.

Second-order self-similar processes with $H > 0.5$ manifest equivalent properties, such as (a) slowly decaying variance, i.e., the variance with scale $Var(X^{(m)})$ decreases very slowly, (b) long-range dependence, i.e., the ACF decays very slowly toward zero and is non-summable, and (c) power spectrum decays in a $1/f$ fashion near the zero frequency. These properties are not shared by Poisson processes.

Several proposed estimators for Hurst parameter are discussed next:

R/S Estimator: R/S statistic (rescaled adjusted range) for the process X_n with mean $\bar{X}(n)$ and variance $S^2(n)$ given as:

$$\begin{aligned} \tfrac{R(n)}{S(n)} &= \tfrac{1}{S(n)}[max(0, W_1, W_2, ..., W_n) \\ &\quad -min(0, W_1, W_2, ..., W_n)], \end{aligned} \tag{5}$$

where $W_k = (X_1 + X_2 + ... + X_k) - k\bar{X}(n), k \geq 1$. For an asymptotic second-order self-similar process X_n the following condition holds:

$$E\left[\frac{R(n)}{S(n)}\right] \approx cn^H, 0.5 < H < 1. \tag{6}$$

For a second-order self-similar process, a plot of $log(E[\frac{R(n)}{S(n)}])$ with $log(n)$ will be linear with slope $H, 0.5 < H < 1$.

Variance-Time Estimator: This plots the variance of aggregated process versus aggregation level on a log-log plot. From (4) we see that such a plot is linear with slope $-\beta$. The Hurst parameter can be estimated as $H = 1 - \frac{\beta}{2}$.

Wavelet Estimator: Veitch and Abry [26] describe a semi-parametric estimator of H based on the discrete wavelet transform (DWT). They plot the logarithm of variance of the wavelet coefficients obtained after taking the DWT of the original process against scale j (also called *octave*). The slope of this plot γ obtained by linear regression gives an estimate of $H(= \frac{1+\gamma}{2})$. Wavelet estimators are generally preferred as they are immune to smooth polynomial trends in the data and hence are comparatively robust to nonstationarity in the data. For further details please refer to [26].

C. Heavy Tailed Processes and Statistical Tests

A random variable X is said to be heavy tailed if $P[X > x]$ is proportional to $x^{-\alpha}$ as $x \rightarrow \infty, 0 < \alpha < 2$. A heavy tailed distribution has a heavier tail than an exponential distribution (i.e., large x values occur with non-negligible probability in heavy tailed distributions). To check whether or not a distribution is heavy tailed we look at the distribution's log-log complementary distribution (LLCD) plot. LLCD plot graphs logarithm of complementary cumulative distribution (CDF) against $log(x)$. The parameter α is called the tail index and is equal to the slope of the tail on LLCD plot. Hence if a distribution is heavy tailed, the tail on LLCD plot will appear linear with a slope between 0 and 2. Superposition of many heavy tailed ON-OFF processes gives rise to a self-similar process [12], [19], [20].

Hill Estimator Test: This is a more rigorous test for computing heavy tail index α [27]. Let $X_1, X_2, ..., X_n$ be the values for a stochastic process X_t. Rearranging the individual values in increasing order i.e., $X_{(1)} \leq X_{(2)} \leq X_{(3)} ... \leq X_{(n)}$. Hill's estimate α_n is then given by:

$$\alpha_n = \left(\frac{1}{k} \sum_{i=0}^{i=k-1} (logX_{(n-i)} - logX_{(n-k)}) \right)^{-1}, \tag{7}$$

where k denotes how many of largest observations enter into the calculation of Equation 7. A plot of Hill's estimate α_n is drawn as a function of k for a range of k-values. In the presence of a heavy tail, Hill's plot may vary considerably for small values of k but stabilizes as more points are added. Hill's estimate of heavy tail index α then can be read from the y-axis where the plot becomes stable. If Hill's plot does not become stable, it indicates an absence of heavy tail.

III. Related Work

Many studies have concluded that network behavior is characterized by the presence of long-range dependence, scaling phenomenon, and heavy tailed distributions. The presence of long-range dependence and self-similarity in network traces was first shown by the seminal work carried out by Leland et al. [20]. Later Willinger et al. [27] showed that the self-similarity can be explained due to superposition of ON/OFF sources based on Packet Trains Model whose period lengths have heavy tailed distributions. For recent traces we verify these findings.

Self-similarity has also been studied in the context of Web traffic. Crovella et al. [12] verified that Web traces collected in 1995 at Boston university were self-similar. We verify self-similar behavior for recent Web traces (Web+). Additionally we also show that recent Web+ traffic when analyzed locally shows Poisson characteristics over time-intervals of few seconds.

Karagiannis et al. [18] analyzed network traces collected in 2002-03 and revisited Internet traffic characteristics. The rationale to revisit these traffic models was that there had been a tremendous growth (more than three orders of magnitudes) of Internet backbone link speeds and number of Internet connected hosts in the previous few years. As a result large number of flows, elephants as well as mice, got multiplexed within the core and this huge traffic multiplexing might have resulted in characteristics not captured well by the traffic

TABLE I
SUMMARY OF TRACES

Trace	Period (JST)	Packets (M)	Avg Traffic Rate (Mbps)	Link Utilization (%)	Flows	Web+ Packets (%)	P2P+ Packets (%)	Web+ Bytes (%)	P2P+ Bytes (%)
SPJ07	Jan 01, 07 (0700-0930)	114.6	70.87	70.87	5311006	43.09	32.51	57.16	29.41
SPM08	Mar 18, 08 (0900-1130)	128.6	73.36	48.90	7557911	41.31	29.99	53.64	30.10
SPJ09	Jan 22, 09 (1400-1415)	20.5	124.95	82.66	2194862	41.36	45.15	41.43	53.74

TABLE II
SUMMARY OF P2P+ COMPONENT

Trace	P2P Packets (%)	P2P Bytes (%)	Gaming Packets (%)	Gaming Bytes (%)
SPJ07	95.63	98.49	2.85	1.10
SPM08	86.60	96.15	9.10	2.84
SPJ09	93.84	98.09	4.56	1.57

models in vogue. Contrary to the general understanding, the authors first showed the coexistence of Poissonity and long-range dependence. We verify these findings on recent traces.

Some recent studies have analyzed self-similarity of traffic from few individual P2P+ applications. Liu et al. [22] showed the self-similarity of BitTorrent traces and explained it due to heavy-tailed distributions of BitTorrent transmission times and quiet times. Cevizci et al. [8] revealed the self-similar behavior of an online game 'Call of Duty'. using two PCs in a laboratory where one machine played the game while the other one collected the trace. In this paper, we look at P2P+ traffic extracted from recent traces and show the coexistence of Poissonity and self-similarity.

IV. METHODOLOGY

This work uses publicly available traces collected from WIDE backbone [1], [11]. For each day a 15 minute extract is made public for download after anonymizing IP addresses. For our analysis we have chosen one such 15 minute extract collected in Jan 2009. One such 15 min extract was used by Karagiannis et al. [18] as well. Additionally two longer traces of 48 and 72 hour long duration are available which were collected in January 2007 and March 2008, respectively. From these two traces, we choose two 150 minutes long extracts.

Next we describe how we extract P2P+ and Web+ components. As discussed above, the P2P+ component consists of applications that use random ports rather than well defined default ports. P2P forms the major component of P2P+ traffic. To extract P2P traffic we make use of techniques that identify P2P traffic at transport layer [17]. These techniques identify P2P traffic based on flow connection patterns of P2P traffic and do not inspect packet payloads (see [17] for details). We also implement heuristics to extract gaming traffic provided in [17]. We identify the rest of the P2P+ traffic as packets with both source and destination port greater than 1023. Port numbers in range 0-1023 are reserved and hence most of the P2P+ traffic is found on ports greater than 1023.

To extract Web+ component we inspect traffic on ports 80, 443 and 8080 (i.e. traffic using the 'http' protocol). After extracting out P2P+ traffic, every packet is inspected and if either of the source or destination port is 80, 443 or 8080 the packet is classified as Web+. P2P+ traffic is first extracted out as a part of P2P traffic may be on 'http' ports. Further classifying a Web+ packet whether it belongs to traditional Web page download or non traditional video streaming portal is difficult as it requires checking the payload and these traces do not capture any payload. However, due to popularity of video sharing portals worldwide, we can safely assume that any recent Web+ traffic should have significant component emanating from audio/video portals [4], [5].

Table I presents the summary of three selected traces[2]. Web+ and P2P+ components together are found to constitute majority of the aggregate traffic ($\approx$ 80%) for each of the three traces. This provides evidence to support the assertion that Web+ and P2P+ form the major components of recent traffic. Table II presents the breakdown of the P2P+ component. We find that P2P traffic constitutes majority of P2P+ traffic ($\approx 95\%$). Link utilization of these three traces approximately varies between 50% to 80% which is significantly greater than traces used in [18] where utilization was between 10% and 35%.

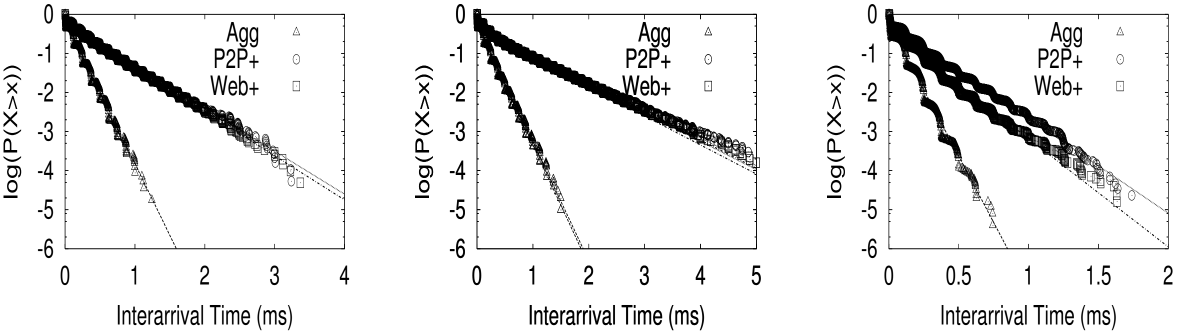
Analyzing a trace at timescale $t\ ms$ means we look at the series obtained by averaging the base series over consecutive blocks of $t\ ms$. In context of Poissonity, however, we use the terminology timescale to mean the analysis of a trace extract of length $t\ ms$. Timescale $t\ ms$ is called small or large based on whether t is small or large depending on the context.

In the remainder of the paper we use the following notations: (a) All the plots for components Web+ and P2P+ are titled 'Web+' and 'P2P+' respectively, (b) All the plots for complete traces are titled 'Agg', (c) All logarithms are taken for base 10 unless stated otherwise and (d) The terms 'complete trace' and 'aggregate component', 'packet interarrival times' and 'interarrival times', 'milliseconds' and 'ms' are used interchangeably. For lack of space the results of all three traces are not shown for every test; however, the results presented apply to all three traces unless stated otherwise.

V. POISSONITY AT TIME SCALES OF FEW SECONDS

In this section we show that recent traffic can be approximated by Poisson process at time scales of few seconds($\approx$

[2] This link was upgraded from 100Mbps to 150Mbps in June 2007. That is why SPM08 has higher average rate but lesser link utilization as compared to SPJ07.



(a) SPJ07 (b) SPM08 (c) SPJ09

Fig. 1. Distribution of Packet InterArrival Times (duration - 5 seconds)

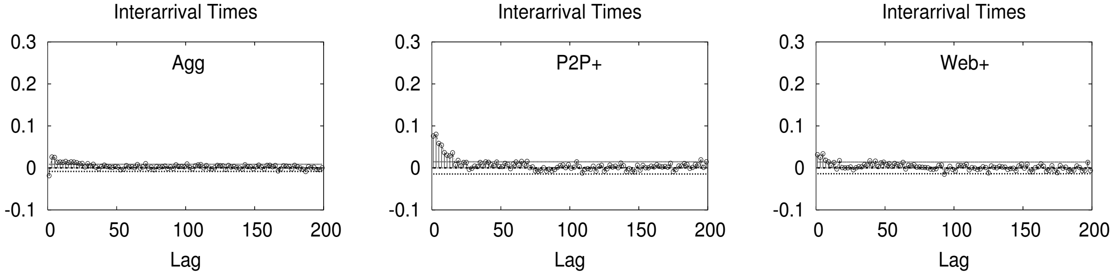


Fig. 2. AutoCorrelation Coefficients for Packet Interarrival Times and Packet (duration - 5 seconds) for trace SPJ07

5s). We show that packet interarrivals follow exponential distribution and are uncorrelated. Our discussion uses tests described in Section II.

A. Distribution of Interarrival Times

Figure 1 shows CCDF plot for a 5 second portion of all the three traces. We can see that distribution of interarrival times for the aggregate as well as Web+ and P2P+ components can be nicely modeled as exponential. The CCDF of interarrival times when Y axis is plotted on log scale shows approximately linear behavior across entire range except for small interarrival times (less than 7 μs). When a linear regression is carried out, the coefficient of determination[3] R^2 is found to be more than 99% which shows a good fit to the data points. Other packet arrivals like TCP packets, specific sized packets (e.g 1518 bytes) etc. are similarly found to have exponentially distributed interarrivals. A similar behavior of CCDF plot is observed for any 5 second extract of a trace.

A little deviation at small interarrival times (less than 7 μs) can be explained due to the effect of back-to-back packets. In a heavily utilized link interarrival times are a function of packet sizes as many packets are sent back-to-back, hence as a result may not contain any idle time. Additionally we can see that interarrivals of SPJ09 trace are comparatively shorter when compared to traces SPJ07 and SPM08. Trace SPJ09 has higher link utilization and hence successive packets contain lesser idle time when compared to SPJ07 and SPM08[4].

Interestingly we find that for traces SPJ07 and SPM08, CCDF plots of P2P+ and Web+ components lie almost on top of each other while for trace SPJ09 they are pretty close. Plot for aggregate component is below that of Web+ and P2P+ component which is intuitive. As number of packets increase, successive packets contain lesser idle time and as a result the probability of occurrence of a large interarrival time is quite low.

B. Independence

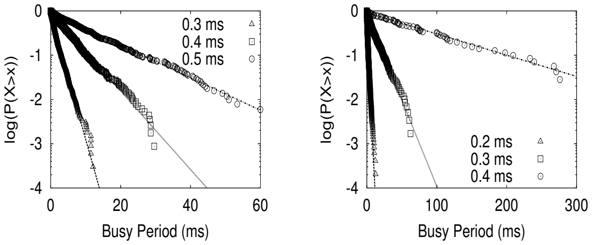
Autocorrelation Function: Figure 2 plots the autocorrelation coefficients for interarrival times at varying lags (between 0 and 200) for trace SPJ07. Identical 5 second portions have been used as those used to draw the CCDF plots in Figure 1. Horizontal lines represent 95% confidence intervals. As can be seen, autocorrelation coefficients for most lags lie between 95% confidence interval. This shows that interarrival times are almost independent of each other and we can conclude that aggregate traffic as well as Web+ and P2P+ components can be approximated by Poisson distribution at a time scale of few seconds.

Burst Sizes: To further strengthen the claim of memoryless Poisson arrivals and independence, we look at packet bursts. A

[3] Coefficient of determination R^2 is defined as the square of the sample correlation coefficient between the outcomes and their predicted values.

[4] Looking closely at Figure 1 we see that CCDF plot consists of a series of small steps. This wavy behavior is more pronounced for traces with higher link utilization (SPJ09). Traces with low link utilization collected few years back from the same link do not exhibit such a behavior. We plan to study this behavior more closely.

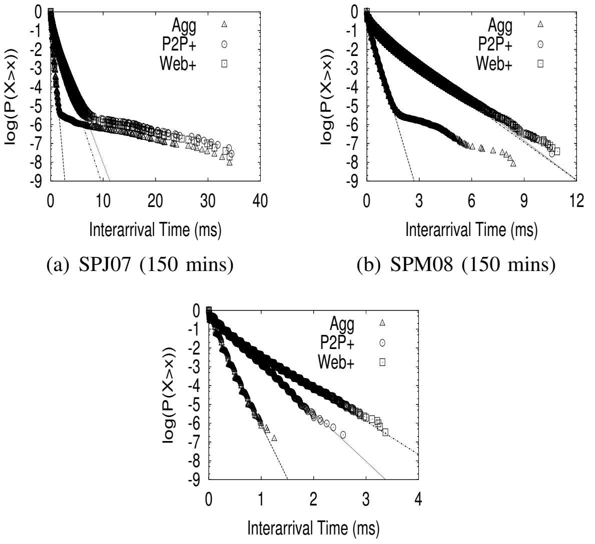
burst is defined as a sequence of successive packets with each packet interarrival less than a threshold value. If the arrival process is memoryless, the characteristics of the burst will remain the same irrespective of the threshold chosen [18]. Figure 3 plots the CCDF of burst busy periods (time span between first and last packet arrival of a burst) for 5 seconds extracts of traces SPJ07 and SPJ09 using three different thresholds. We see that CCDF plots nicely fit a straight line thereby showing that packet arrivals can be nicely approximated by Poisson. Distribution of burst idle periods (burst interarrivals) is similarly exponential.



(a) SPJ07 (5 sec)

(b) SPJ09 (5 sec)

Fig. 3. Distribution of Burst Busy Periods



(a) SPJ07 (150 mins)

(b) SPM08 (150 mins)

(c) SPJ09 (15 mins)

Fig. 4. Distribution of Packet Interarrival Times

C. Deviation from Poisson at large timescales

Next we show Internet traffic when analyzed over large timescales deviates from Poisson behavior.

Tailed Interarrival Times: Figure 4 plots the CCDF plot of interarrival times for the entire duration of the traces. We find that CCDF plot does not follow an exponential distribution. Large interarrival times (also called tailed interarrival times) deviate considerably from a straight line although their probability of occurrence is quite low. A similar deviation is observed for busy periods of packet bursts when computed for the entire duration of the traces. This shows that for larger time scales (more than a few milliseconds), an exponential and hence Poisson model cannot be applied. When we investigate the packet arrivals on a 5 second extract, the number of tailed interarrival times is very small and hence not significant. As a result we find that packet arrivals can be modeled as exponential for short time intervals. Though not shown here, we find that as the timescale of analysis increases, the deviation from straight line increases more and more. Similar deviation for large values of interarrival times has been observed for SMTP sessions [19]. Here we observe such a behavior for recent network traces.

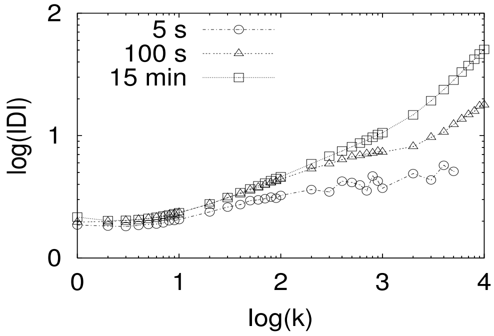


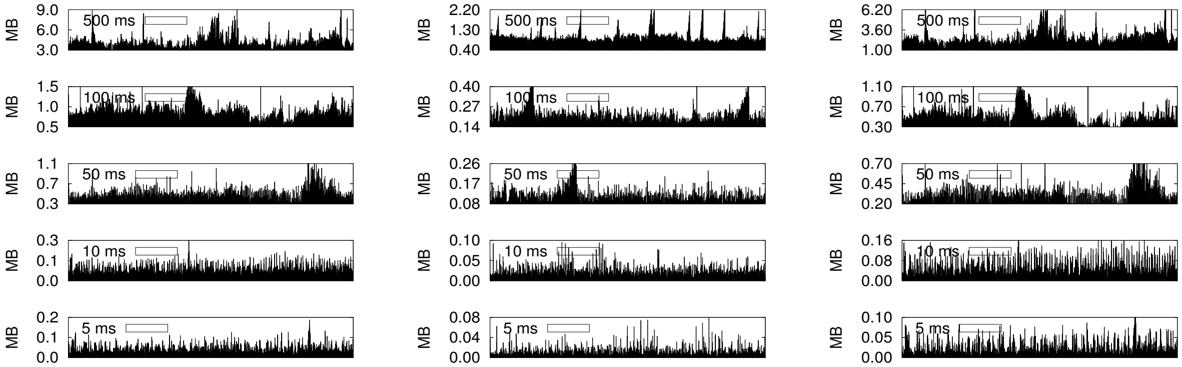
Fig. 5. IDI Plot : SPJ07

Once again the CCDF plots for Web+ and P2P+ components are found to lie on top of each other for traces SPJ07 and SPM08. Additionally, we notice that interarrival times for the aggregate component deviate earlier than Web+ and P2P+ components. All interarrivals greater than 1.7 ms for SPJ07 show a deviation from a straight line. The corresponding number for Web+ and P2P+ components is approximately 4 and 5 ms. Similar behavior is observed for trace SPM08 as well. However no deviation (except $\leq 7\mu$s) from a straight line behavior is found for trace SPJ09. This again can be attributed to higher link utilization of trace SPJ09 because of which no large interarrival is observed.

Index of Dispersion for Intervals: Similar deviation at large scales can also be visualized using the Index of dispersion for interval (IDI) plot (cf. section II). Figure 5 shows the IDI plot for three extracts of different time intervals for trace SPJ07. An ideal Poisson process has $c_k^2 \equiv 1$ for all k. If arrival process has higher variance at some timescale, c_k^2 increases with an increase in k; Figure 5 shows such behavior. Values of c_k^2 for all k are close to 1 for 5 second extract. However, the values of c_k^2 for larger timescales (100 sec and 15 mins) quickly diverge as k increases which indicates a deviation from Poisson behavior. Web+ and P2P+ components are found to behave similarly.

VI. Biscaling Behavior of Hurst Parameter and Self-Similarity

In this section we first show that recent traces can be modeled as piecewise wide sense stationary. Then we investigate the self-similarity of recent traces as well as Web+ and P2P+



(a) Agg (b) P2P+ (c) Web+

Fig. 6. Visualization of trace SPM08. Number of Bytes transferred are plotted for different timescales.

components. First we show the presence of self-similarity by means of traffic visualization and autocorrelation functions. Next we augment the evidence by computing Hurst parameter and showing it lies between 0.5 and 1. Finally we model the traces using a Packet Train Model to explain why recent traces display self-similar behavior despite an increase in non-traditional traffic.

A. Piecewise Wide Sense Stationarity

Figure 7 plots for trace SPJ07, the average number of bytes transferred in a millisecond within every 10 second window. This results in 900 data points for the 2.5 hour long SPJ07 trace. Figure 7 clearly shows the non-stationary nature of aggregate and Web+ traffic. However traffic can be nicely modeled as piecewise wide sense stationary. Looking closely at Figure 7 we see that mean and variance of aggregate and Web+ traffic can be considered stationary for first 500 data points, next 200 data points and last 200 data points. Interestingly, P2P+ traffic is found to be wide sense stationary across entire duration. Similar observations also hold true for trace SPM08.

This concept of describing network behavior as a series of piecewise stationary intervals (also called *change free regions*) has been applied elsewhere as well [18], [28]. It allows us to extract out a change free region and analyze it without the interference of underlying nonstationary. For trace SPJ07 we extract out the traffic component spanning the first 500 data points (i.e., 5000 sec long) and carry out a self-similarity analysis on this component. We similarly find trace SPM08 to be WSS between data points 100 and 500 and extract this component for further analysis. Trace SPJ09 is 15 min long and is found to be WSS for entire duration. All plots in the rest of the paper have been drawn over these extracts, however are labeled with original trace-names.

B. Presence of Self-Similarity

As discussed above a self-similar traffic exhibits burstiness at various levels of aggregation which is caused by a scale-invariant variance property. We can aggregate either the number of bytes transferred or the number of packets observed. We find that all the results are similar for both these formulations. Hence for the rest of the paper we have shown results only for the formulation where we aggregate bytes sent over a time-interval. Figure 6 plots the number of MBs sent per time unit at various timescales for the trace SPM08. Both P2P+ and Web+ components are shown as well. Timescales plotted are 5, 10, 50, 100 and 500 ms. All the plots have 1500 data-points. We can see that recent traffic as well as both Web+ and P2P+ components exhibit burstiness at all timescales thereby indicating self-similar behavior.

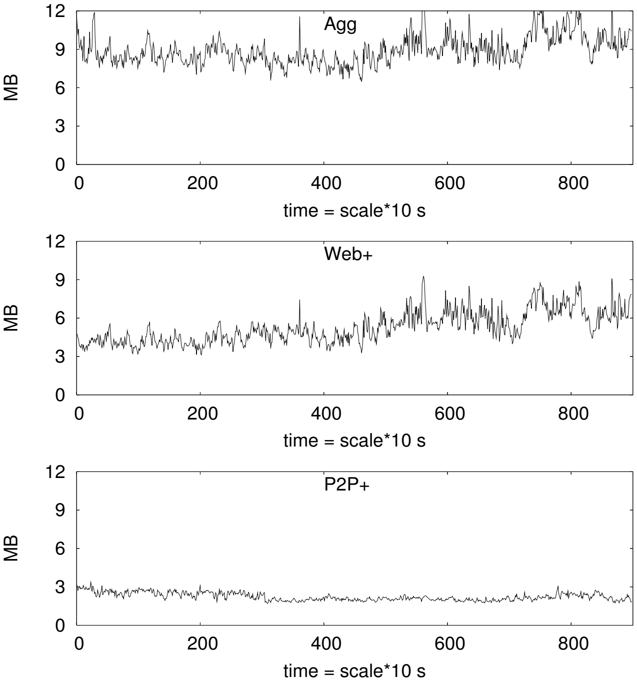


Fig. 7. Nonstationary nature of trace SPJ07

Figure 8 plots autocorrelation coefficients for the aggregated sizes series for trace SPM08 at various time scales. Plots for aggregation level 1,10, 50 and 100 ms are shown. Unlike Poisson process the values of autocorrelation coefficient, for

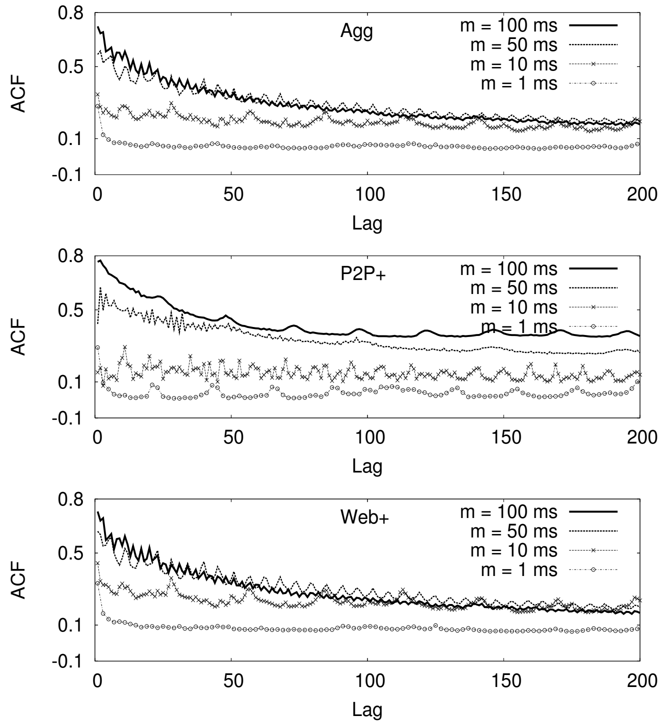


Fig. 8. AutoCorrelation Coefficients for trace SPM08 at various aggregation levels (in ms).

aggregation level larger than 1ms, are not close to zero showing thereby a dependence of data values at different lags. A non-zero value of autocorrelation coefficient at large lags shows the presence of long-range dependence. We see that all the curves are fluctuating, neither decaying exponentially nor converging to zero, thereby behaving like self-similar processes. Next we estimate Hurst parameter using various estimators described in section II.

C. Estimation of Hurst Parameter

First we employ variance-time estimator. For each trace we construct a sequence where each value in the sequence represents the number of bytes sent over the link every millisecond. From this sequence X we construct various aggregated series X^m for different values of m. Next we plot $log(Var(X^m))$ versus $log(m)$ for each trace by varying $log(m)$ from 0 to 4 and obtain the slope $-\beta$ of each variance-time plot using a linear regression plot. Finally Hurst parameter for the trace can be calculated as $1 - \frac{\beta}{2}$.

Figure 9 shows the variance time plots and linear regression results for the trace SPJ07. We observe that instead of being linear across all time scales, variance time plots are piecewise linear. This dichotomy has been observed elsewhere as well [18], [19]. This dichotomy happens because the traffic is not globally self-similar and the value of Hurst parameter depends on the scale at which the traffic is viewed. At scales below a certain threshold the Hurst parameter is smaller and at larger scales it increases. The threshold and the values of Hurst Parameter vary depending on the trace. Table III summarizes the results obtained. Values of Hurst parameter are found to be noticeably larger than 0.5 and less than 1 which suggests the LRD behavior of the traces. Web+ and P2P+ components are also found to be displaying self-similarity and similar dichotomy in Hurst parameter value.

Result of R/S estimator also shows self-similar behavior of recent traces. A plot of $log(E(\frac{R(n)}{S(n)}))$ vs. $log(n)$ is taken with increasing values of n where n represents the size of a non-overlapping block. A roughly linear plot shows the self-similarity of the distribution. Slope of the linear plot estimates the Hurst Parameter. Figure 10 shows R/S plots (also known as pox plots) for trace SPJ07. A similar dichotomy in values of Hurst Parameter can be observed here as well.

Finally we apply wavelet estimator for computing the value of Hurst parameter. Figure 11 presents the Logscale diagrams as obtained for trace SPJ07. Logscale diagrams also nicely bring out the biscaling behavior of current traces with the change of point being around 100 ms. Both variance-time and R/S estimators also agree with this observation.

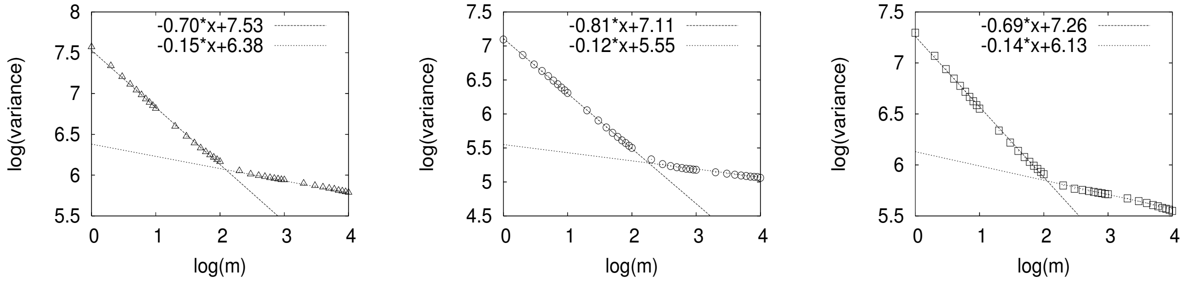
Table III enlists the values of Hurst parameter both at below and above the point of change(100-200 ms) as predicted by variance-time, R/S and wavelet estimators. Predictions of all three estimators match nicely. Interestingly we find that values of Hurst Parameter (both at small and large scales) of aggregate, Web+ and P2P+ components are roughly similar for all traces.

TABLE III
ESTIMATION OF HURST EXPONENT

Trace	Estimator		
	Variance-Time	R/S	Wavelet
Agg			
SPJ07	0.65,0.93	0.62,0.92	0.63,0.98
SPM08	0.71,0.91	0.57,0.98	0.59,0.98
SPJ09	0.65,0.89	0.65,0.93	0.69,0.97
P2P+			
SPJ07	0.61,0.94	0.62,0.86	0.59,0.98
SPM08	0.60,0.93	0.56,0.99	0.62,0.99
SPJ09	0.61,0.81	0.60,0.92	0.70,0.79
Web+			
SPJ07	0.65,0.93	0.61,0.94	0.64,1.03
SPM08	0.73,0.92	0.59,0.98	0.67,0.98
SPJ09	0.63,0.88	0.59,0.97	0.65,0.89

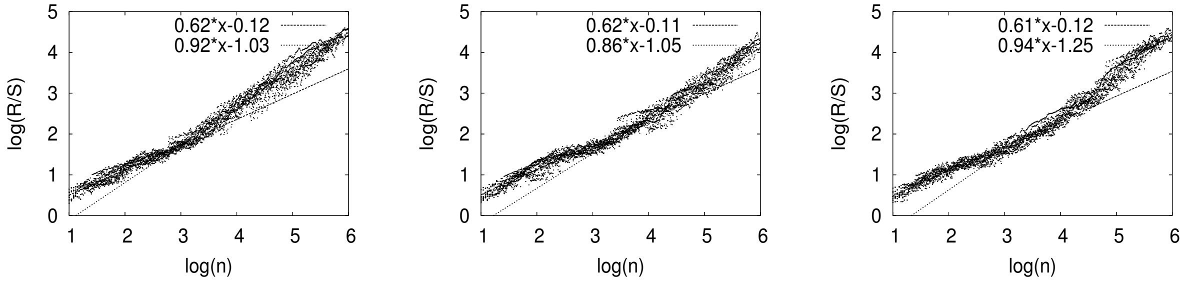
D. Heavy Tailed ON/OFF sources

Next we investigate the reasons behind self-similarity using ON/OFF model [12], [27]. To validate ON/OFF modeling we group the traffic by source IP address and for each source an ON/OFF process is constructed. An ON period is defined as a packet train which has burst of packets arriving from the same source. If the interval between two packets exceeds a predefined threshold, they are said to belong to different packet trains. Lengths of these ON periods constitutes the ON process. The spacing between two packet trains constitutes the OFF process. During OFF periods there is no packet arrival. The threshold is a system parameter dependent on



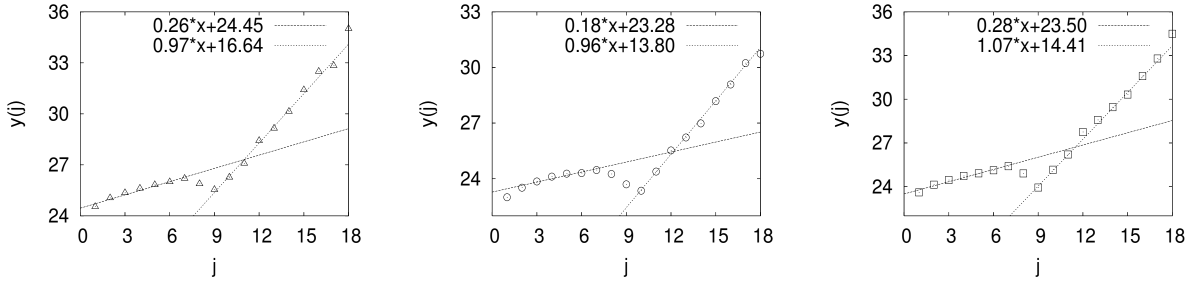
(a) Agg (b) P2P+ (c) Web+

Fig. 9. Variance Time Plots for trace SPJ07. m represents the aggregation level in millisecond.



(a) Agg (b) P2P+ (c) Web+

Fig. 10. Pox Plots for trace SPJ07. n represents the size of non-overlapping block at which R/S statistic is computed.

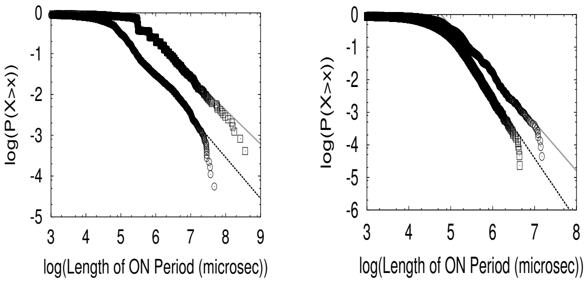


(a) Agg (b) P2P+ (c) Web+

Fig. 11. Logscale Diagrams for trace SPJ07. j represents the octave (time = 2^j ms)

the frequency with which applications use the network. If the distribution of the lengths of either ON or OFF periods is heavy tailed (Noah effect), the superposition of these processes will result in a self-similar process (Joseph effect) [27].

Figure 12 presents LLCD plots for distribution of ON period lengths for SPM08 and SPJ07 sources. Two popular sources are chosen for each trace. The threshold for separating packet trains is taken to be 40 ms. Tails for all the distributions are found to span two to three orders of magnitudes which indicates the presence of a heavy tail. A linear regression is carried out on the tail of each distribution. Table IV presents the values of heavy tail parameter α calculated using linear regression. We can see that the slopes of all the tails turn out to be between 1 and 2. We also use Hill's estimator to estimate the values of heavy tail parameter α. The α values as estimated by Hill's estimator are also given in Table IV and are found to be close to those estimated by linear regression.



(a) SPJ07 (b) SPM08

Fig. 12. LLCD plots for ON period lengths for two popular sources

Distribution of ON period lengths for P2P+ and Web+ sources is found to be heavy tailed as well.

TABLE IV
VALUES OF HEAVY TAIL PARAMETER α FOR ON TIME DISTRIBUTION

Test	Heavy Tail Parameter(α)	
	SPJ07	SPM08
Linear Regression	1.01,0.84	1.90,1.49
Hill Estimator	1.06,0.95	1.96,1.62

VII. CONCLUSIONS

In this paper we demonstrated the coexistence of Poissonity and self-similarity in recent Internet traces collected across a high speed Internet backbone with heavy utilization.

At a time-scale of few seconds, we found that packet interarrivals can be approximated by Poisson. Interarrival times are found to follow exponential distribution and are uncorrelated. At large timescales, however, we observed a deviation from Poisson behavior.

We further showed that recent traces exhibit self-similarity and long-range dependence. Values of Hurst parameter display a dichotomy. At small scales it lies between 0.6 and 0.75 while at large scales it lies between 0.85 and 0.98 with the point of change being between 100 to 200 ms. Distribution of ON period lengths for individual sources is found to be heavy tailed which explains the self-similarity of recent traces. These findings again emphasize the fact that choosing a proper timescale for traffic analysis is of paramount importance.

We also argued that recent Internet traffic can be looked to have two main constituents: Web+ and P2P+. Web+ encompasses traditional 'Web page downloads' as well as recent audio/video streaming traffic. P2P+ comprises traffic from applications which use random ports instead of well defined ports. We further show that both Web+ and P2P+ components exhibit coexistence of Poissonity and self-similarity.

Avenues for future research include verifying Internet traffic models and characteristics on traces collected from other links. Of particular interest also is studying flow and host-level properties of recent Internet traffic traces. Another interesting direction would be to study how flow and host-level characteristics have changed over the years as Internet traffic configuration changes.

ACKNOWLEDGEMENTS

We thank Martin Arlitt, Arzad Kherani, and Carey Williamson for discussions during early stages of this work. Thoughtful MASCOTS reviews have helped us improve the paper, and we thank the reviewers for their efforts.

REFERENCES

[1] MAWI working group traffic archive. http://tracer.csl.sony.co.jp/mawi.

[2] Online games market to hit $4.4 billion by 2010, http://www.clickz.com/3623306.

[3] Skype wikipedia page: http://en.wikipedia.org/wiki/skype.

[4] YouTube effect: HTTP traffic now eclipses P2P. http://arstechnica.com/old/content/2007/06/the-youtube-effect-http-traffic-now-eclipses-p2p.ars.

[5] YouTube is 10% of North American Internet traffic. http://www.last100.com/2007/06/27/youtube-represents-10-of-north-american-internet-traffic/.

[6] N. Basher, A. Mahanti, C. Williamson, and M. Arlitt. A comparative analysis of Web and Peer-to-Peer traffic. In *Proceedings of WWW*, pages 287–296, 2008.

[7] R. Birke, M. Mellia, M. Petracca, and D. Rossi. Understanding VoIP from backbone measurements. In *Proceedings of INFOCOM*, pages 2027–2035, 2007.

[8] I. Cevizci, M. Erol, and S. F. Oktug. Analysis of multi-player online game traffic based on self-similarity. In *Proceedings of NetGames*, page 25, 2006.

[9] C. Chambers, W. C. Feng, S. Sahu, and D. Saha. Measurement-based characterization of a collection of on-line games. In *Proceedings of IMC*, pages 1–14, 2005.

[10] K. T. Chen, P. Huang, C. Huang, and C. Lei. Game traffic analysis: An MMORPG perspective. In *Proceedings of NOSSDAV*, pages 19–24, 2005.

[11] K. Cho, K. Mitsuya, and A. Kato. Traffic data repository at the WIDE project. In *Proceedings of USENIX 2000 FREENIX Track*, 2000.

[12] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Transaction on Networking*, 5(6):835–846, 1997.

[13] P. Danzig, S. Jamin., R. Caceres, D. Mitzel, and D. Estrin. An empirical workload model for driving wide area TCP/IP network simulations. *Internetworking: Research and Experience*, 3(1):1–26, 1992.

[14] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson. Online/realtime network traffic classification using semi-supervised learning. *Journal of Performance Evaluation*, 64(9-12):1194–113, 2007.

[15] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from the edge. In *Proceedings of IMC*, pages 15–28, 2007.

[16] R. Gusella. A measurement study of diskless workstation traffic on an ethernet. *IEEE Transactions on Communications*, 38(9):1557–1568, 1994.

[17] T. Karagiannis, A. Broido, M. Faloutsos, and K. C. Claffy. Transport layer identification of P2P traffic. In *Proceedings of IMC*, pages 121–134, 2004.

[18] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A non-stationary Poisson view of Internet traffic. In *Proceedings of IEEE INFOCOM 2004*, pages 84–89, 2004.

[19] Y. Lee and J. S. Kim. Characterization of large scale SMTP traffic: the coexistance of the Poisson process and self similarity. In *Proceedings of MASCOTS 2008*, pages 143–152, 2008.

[20] W. E. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self similar nature of Ethernet traffic. *IEEE/ACM Transaction on Networking*, 2(1):1–15, 1994.

[21] W. E. Leland and D. V. Wilson. High time-resolution measurement and analysis of lan traffic: Implications for LAN interconnection. In *Proceedings of IEEE INFOCOM*, pages 1360–1366, 1991.

[22] G. Liu, M. Hu, B. Fang, and H. Zhang. Explaining BitTorrent traffic self similarity. In *Proceedings of PDCAT*, pages 839–843, 2004.

[23] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modelling. *IEEE/ACM Transaction on Networking*, 3(3):226–244, 1995.

[24] S. M. Ross. *Introduction to Probability Models Elsevier Inc.* John Wiley and Sons, Inc., 2006.

[25] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexors for voice and data. *IEEE Journal on Selected Areas in Communications*, 5(6):833–846, 1986.

[26] D. Veitch and P. Abry. A Wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Transactions on Information Theory*, 45(3):878–897, 1999.

[27] W. Willinger, M. S. Taqqu, W. E. Leland, and D. V. Wilson. Self-similarity in high-speed packet traffic: Analysis an modeling of Ethernet traffic measurements. *Statistical Science*, 10(1):67–85, 1995.

[28] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker. On the constancy of Internet path properties. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, 2004.