Characterization of Large-Scale SMTP Traffic: the Coexistence of the Poisson Process and Self-Similarity *

Youngjae Lee and Jin-Soo Kim
Computer Science Department
Korea Advanced Institute of Science and Technology(KAIST), South Korea
yjlee@camars.kaist.ac.kr, jinsoo@cs.kaist.ac.kr

Abstract

Network traffic such as Ethernet, Internet, World Wide Web, and TCP/UDP protocols has been extensively studied, with efforts focusing on the Poisson process and self-similarity. However, although SMTP (Simple Mail Transfer Protocol) occupies a significant portion of Internet traffic, it has attracted little attention. This paper shows that large-scale SMTP traffic possesses both the characteristics of the Poisson process and self-similarity through an analysis of high quality SMTP traces collected from one of the largest Web portal sites in South Korea over a period of almost nine months.

First, we show that, at small (several-second) time scales, interarrival times of SMTP session arrivals are exponentially distributed and independent of each other, which makes it possible to model SMTP session arrivals as a Poisson process. On the contrary, at large (several-month) time scales, SMTP session arrivals exhibit self-similarity. They are strongly autocorrelated across time scales of days and their Hurst parameters are estimated to be somewhere between 0.85 to 0.97. In addition, we find that SMTP traffic consists of many individual ON/OFF sources whose distributions of OFF-period lengths are heavy-tailed, which confirms self-similarity of SMTP traffic.

1. Introduction

Network traffic has been extensively studied with high quality network traces, whose time unit is less than milliseconds and coverage is multiple days. Most previous researches focused on the question of whether network traffic can be modeled as Poisson processes or whether it shows characteristics of self-similarity/long-range dependence.

Although for decades many Poisson-based network traffic models have been developed in favor of their attractive theoretical properties, Paxson et al. [20] attempted to show the failure of modeling network arrivals as Poisson processes. Karagiannis et al. [15] examined the possibility of modeling Internet backbone traffic as Poisson packet arrivals at various time scales. Self-similarity of network traffic has been investigated empirically and theoretically [19] since the pioneering work of Leland et al. [16]. In [16, 24, 25], the authors presented that Ethernet traffic exhibits self-similarity and long-range dependence, meaning burstiness occurs at every time scale. Further studies show that the superposition of ON/OFF sources based on the packet trains model [14, 18], each of which exhibits the "*Noah Effect*", results in self-similar traffic. The Internet [7] and World Wide Web [9, 10] are also shown to be long-range dependent and self-similar.

Poisson processes and self-similar processes, which have been key considerations in the analysis of network traffic, show very different theoretical properties. Traffic generated by the former has properties that its interarrival times have an exponential distribution and they are independent of each other. Such short-range dependent traffic becomes smoother as the time scale increases. On the contrary, self-similar traffic is dependent on each other and has heavy-tailed distributions with high variance. Such long-range dependent traffic always has burstiness at every time scale, never becoming stable. In spite of these differences, further study on the Poisson process presented impressive findings that there exist both a Poisson process and long-range dependence in heavy Internet backbone traffic [15]. The authors showed that TCP/UDP packets obey a Poisson process at sub-second time scales, while they are long-range dependent at large time scales. This suggests that relatively simple statistical theories of the Poisson process can still be applicable to the design and optimization of network systems.

*This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST) (No.R01-2007-000-11832-0)

SMTP (Simple Mail Transfer Protocol) [1] is an Internet standard protocol based on a server-client model for transferring emails. The protocol, which rules messages exchanged between the sender and the receiver, is one of the most popular protocols occupying a significant portion of Internet traffic. However, most existing studies on the Poisson process and self-similarity of network traffic have centered on traffic such as Ethernet, Internet, World Wide Web, and TCP/UDP protocols. Only a few studies [4–6, 11, 12] have been devoted to the analysis of SMTP traffic. These studies, however, did not conduct a comprehensive analysis on the characteristics of SMTP traffic, since they were performed based on small-scale traces with a few hundred thousand SMTP requests.

To address these limitations of previous work, this paper attempts to provide an extensive empirical study of large-scale SMTP traffic. Our work is based on traces collected from one of the largest Web portal sites in South Korea, which provides the general webmail service to more than twenty million users. These traces cover almost nine months (from January to September 2007) and contain more than a billion SMTP requests. The received SMTP commands are logged with timestamps accurate to a millisecond. A single SMTP session from `HELO` to `QUIT` SMTP command has been considered as a unit of SMTP traffic.

The contribution of this paper is an intensive analysis of large-scale SMTP traffic to highlight the coexistence of the Poisson process and self-similarity. This analysis consists of two parts. First, we present that SMTP traffic can be modeled as a Poisson process at small time scales. Using quantile-quantile plots and linear regression method, we show that interarrival times of SMTP session arrivals follow an exponential distribution. In addition, we present p-values of fitting interarrival times to the exponential distribution with a Chi-square test, and, using autocorrelation functions, show that most interarrival times are independent of each other.

Second, at large time scales, we present self-similarity of SMTP traffic and its possible cause. Through statistical tests such as a variance-time plot and R/S analysis, the Hurst parameter is estimated to be somewhere between 0.85 to 0.97, which implies that SMTP traffic is self-similar. We also group SMTP sessions by each source IP address and show that their OFF-period lengths have heavy-tailed distributions using complementary cumulative distribution function plots and Hill's estimation.

The remainder of this paper is organized as follows. In section 2, we provide a brief introduction to the Poisson process and self-similarity. Section 3 overviews the structure of an SMTP session and traces used in this paper. Section 4 presents the Poisson characteristics of SMTP traffic at small time scales. Section 5 demonstrates self-similarity and individual heavy-tailed ON/OFF sources of SMTP traffic. Section 6 summarizes previous work and Section 7 concludes the paper.

2. Background

This section presents technical background on the Poisson process and self-similarity which will be discussed in this paper. These are brief descriptions that are necessary to understand the following sections. More details can be found in [16, 19, 22, 24].

2.1. The Poisson process

A stochastic process $\{N(t), t \geq 0\}$ is called a *Poisson process* with rate λ if the interarrival times $X_1, X_2, \ldots$ have a common exponential distribution function as follows [22].

$$P\{X_n \leq x\} = 1 - e^{\lambda x},\ x \geq 0 \tag{1}$$

In Equation 1, the average interarrival time is $\frac{1}{\lambda}$. 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.

2.2. Self-similarity

Let $X(t)$ be a stochastic process defined for $t = (0, 1, 2, \ldots)$ with mean μ and variance σ^2. $X(t)$ can be translated in various ways. For instance, $X(t)$ may represent the traffic volume measured in packets, bytes, or sessions at time instance t. For $X(t)$, the autocorrelation function $r(k)$ and autocovariance function $\gamma(k)$ are defined for $k \geq 0$ as follows.

$$r(k) = E\{\big(X(t) - \mu\big)\big(X(t+k) - \mu\big)\} \tag{2}$$

$$\gamma(k) = \sigma^2 r(k) \tag{3}$$

In particular, if $X(t)$ has an autocovariance function of the form $\gamma(k) = \frac{\sigma^2}{2}\big((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\big)$, $X(t)$ is said to be *exactly second-order self-similar* with the Hurst parameter H ($1/2<H<1$). An aggregated process $X^{(m)}(t)$[1], which is obtained by averaging the original process $X(t)$ over non-overlapping blocks of size m, has the same autocovariance as $X(t)$, i.e. $\gamma^{(m)}(k) = \gamma(k)$. Practically, if the weaker condition

$$\lim_{m\to\infty} \gamma^{(m)}(k) = \frac{\sigma^2}{2}\big((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\big) \tag{4}$$

is valid, $X(t)$ is *asymptotically second-order self-similar* with the Hurst parameter H ($1/2<H<1$). An asymptotically self-similar process $X^{(m)}$ has a fixed correlation structure for large m.

[1] $X^{(m)}(k) = \big(X(km-m+1) + \ldots + X(km)\big)/m,\ k = 1, 2, \ldots$

From Equation 4, for $0< H <1,\ H \neq 1/2$, we can derive

$$r(k) \approx H(2H-1)k^{2H-2},\ k \to \infty. \quad (5)$$

In particular, if $1/2< H <1$, $r(k)$ asymptotically behaves as $ck^{-\beta}$ where $c>0$ is a constant. In this case, we can rewrite Equation 5 as

$$r(k) \approx H(2H-1)k^{2H-2} \approx ck^{-\beta}, 0 < \beta < 1,\ k \to \infty. \quad (6)$$

Note that, in Equation 6, $r(k)$ is non-summable, i.e., $\sum_{k=-\infty}^{\infty} r(k) = \infty$. We say that such an autocorrelation function decays hyperbolically and the corresponding process $X(t)$ is *long-range dependent*. In contrast, the autocorrelation function of a Poisson process decays exponentially and is summable; that is $\sum_{k=-\infty}^{\infty} r(k) = 0$. Such a process is said to be *short-range dependent*.

The (asymptotically) second-order self-similar process $X(t)$ satisfying Equation 6 has a property [8] such that

$$Var(X^{(m)}) \approx m^{-\beta},\ \beta = 2-2H. \quad (7)$$

Therefore, plotting $log(Var(X^{(m)}))$ versus $log(m)$ for each aggregation level m results in a line whose slope $(-\beta)$ is between -1/2 and 0. This plot is called a *variance-time plot* and we can estimate the Hurst parameter $(H = 1-\beta/2)$ by this plot.

As explained in [16, 24], the rescaled adjusted range (R/S) statistic for a set of observations $(X(t) : t = 1,2,\ldots,n)$ with sample mean $\overline{X}(n)$ and sample variance $S^2(n)$ is given by

$$\frac{R(n)}{S(n)} = 1/S(n)[max(0,W_1,W_2,\ldots,W_n) - min(0,W_1,W_2,\ldots,W_n)] \quad (8)$$

with $W_k = (X_1 + X_2 + \ldots + X_k) - k\overline{X}(n),\ k \geq 1$. If $X(t)$ is long-range dependent, as self-similar processes are, the following equation holds

$$E[\frac{R(n)}{S(n)}] \approx cn^H,\ 1/2 < H < 1 \quad (9)$$

where H is the Hurst parameter of $X(t)$ and c is a constant. This is known as the *Hurst effect*.

3. SMTP Traces

This section introduces SMTP and the basic structure of an SMTP session. We also briefly describe SMTP traces analyzed in this paper and the methodology used to obtain them.

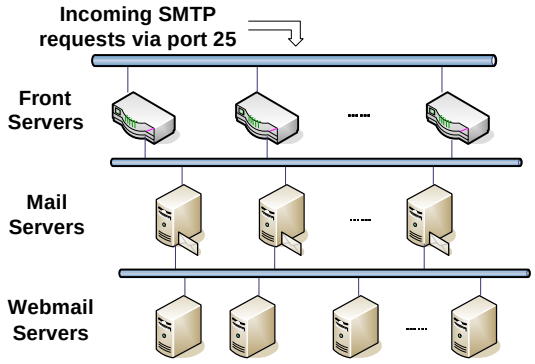


Figure 1. The architecture of the webmail service system

Table 1. Summary of SMTP traces

Trace Set	Period Start	Period End	Accepted Requests	Accepted Sessions
A	25 Jan	27 Feb	472,985,233	73,045,392
B	23 Mar	13 Apr	373,438,992	56,889,377
C	10 May	7 Sep	1,845,470,166	274,218,856
D	31 Aug		15,390,362	3,722,104

3.1. SMTP Sessions

The SMTP [1] is an Internet standard protocol based on a server-client model for transferring emails. The protocol prescribes messages exchanged between the sender (mail clients and mail servers) and the receiver (mail servers). In order to deliver an email, the sender first establishes a session by transmitting an SMTP request to the receiver. An SMTP session is initiated with either a `HELO` or an `EHLO` command, followed by a list of commands, such as `MAIL FROM`, `RCPT TO`, and `DATA`, that must be applied in sequence. After completing the transmission of the mail data, the sender closes the current SMTP session using the `QUIT` command.

In this paper, we consider a single SMTP session, instead of each individual SMTP request, as a unit of SMTP traffic. This is because the order of SMTP requests within an SMTP session is ruled by the SMTP standard and shows relatively deterministic behavior in any SMTP traffic.

3.2. The Collection of SMTP Traces

Our SMTP traces are collected from one of the largest Web portal sites in South Korea. The site provides an integrated search service as well as a wide range of Internet services including webmail, blogs, communities, games, etc. to more than twenty million users.

The webmail service system of the Web portal consists of a set of the front servers, mail servers, and web servers,

as shown in Figure 1. The incoming SMTP requests are first forwarded to one of the front servers. Abnormal requests are dropped out at front servers and only normal requests are delivered to mail servers. Mail transport agents (MTAs) running on mail servers deal with a sequence of SMTP requests and store the received mails to the storage. Later those mails are retrieved by the users through the web interface provided by web servers.

Our traces are collected from one of the front servers from January to September 2007. The selected front server is configured to log SMTP events and inspection results for all the incoming SMTP requests it receives with timestamps accurate to a millisecond. Each log record is composed of an SMTP command and source IP address. The inspection result for an SMTP request includes flags indicating whether the request is subject to spam or virus emails. For the privacy of users, all contents of traces are anonymized so that they do not reveal any private information.

Table 1 summarizes SMTP traces used in this paper. On average, 36–55% of SMTP sessions were refused daily at the front server, either because their source IP addresses were listed in RBL(Realtime Black List)s or the rate of connection requests were too high. We exclude refused SMTP sessions from our analysis as the front server rejects connections with abnormal senders before any SMTP session is established.

4. The Poisson Process

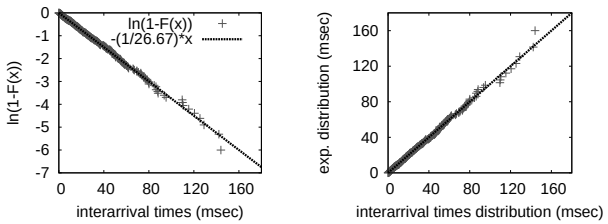
In this section, we show that SMTP traffic follows a Poisson process at small (several-second) time scales. Section 4.1 describes that the interarrival times of SMTP session arrivals fit an exponential distribution and they are independent of each other. Section 4.2 shows that at larger time scales, however, it becomes increasingly difficult to model SMTP session arrivals as a Poisson process.

4.1. Interarrival Times

As mentioned in section 2.1, the interarrival times of a Poisson process form an exponential distribution and they are independent of each other. In order to see if the arrivals of SMTP sessions can be modeled as a Poisson process, we have investigated the interarrival times and their autocorrelation functions using trace D obtained on August 31, 2007 (cf. Table 1). We selected SMTP sessions that arrived within the first ten seconds every hour.

By taking logarithms on both sides of Equation 1, we obtain the following equation.

$$ln(1 - P\{X_n \leq x\}) = -\lambda x \qquad (10)$$



(a) CCDF plot (b) quantile-quantile plot

Figure 2. CCDF plot and quantile-quantile plot for interarrival times during ten seconds after 6 PM in trace D

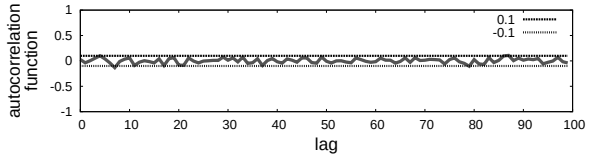


Figure 3. Autocorrelation functions for interarrival times during ten seconds after 6 PM in trace D

Thus, plotting $ln(1 - P\{X_n \leq x\})$ versus x for each interarrival time of a Poisson process results in a linear series of points with slope $-\lambda$. From this *complementary cumulative distribution function (CCDF)* plot, we can conjecture whether the distribution can be approximated by an exponential distribution. If it is deemed possible, we can then obtain the mean rate λ of the distribution by a linear regression method.

Figure 2(a) depicts the CCDF plot and the corresponding linear regression result for interarrival times during ten seconds after 6 PM in trace D. As can be seen in Figure 2(a), the configuration of the points is almost straight, along a 45-degree line. The linear regression yields a slope of $\frac{1}{26.67}$ with a R-squared value $R^2 = 99.88$. This suggests that the interarrival times of SMTP sessions can be closely approximated by an exponential distribution, with the average being 26.67 milliseconds.

Using the average interarrival time obtained by the linear regression method, we can draw a quantile-quantile plot and perform a Chi-square test whose null hypothesis is that these interarrival times come from an exponential distribution. We observe again that the interarrival times form points along a straight line at 45 degrees, as shown in Figure 2(b). In fact, the Chi-square test presents that the interarrival times are not statistically significant, as the p-value is estimated as 0.9414. Figure 3 illustrates autocorrelation functions with lags between 0 and 100 for the same trace used in Figure 2. Most autocorrelation functions lie be-

Table 2. Observations of interarrival times for SMTP sessions during the first ten seconds of selected hours in trace D (time unit is millisecond)

Hour	Average	Standard Deviation	Coefficient of Variation	p-value	Sample Autocorrelation Function(lag)					
					1	20	40	60	80	100
0	37.271	37.703	1.011	0.5862	0.101	0.007	0.150	0.058	0.120	0.001
2	53.476	53.366	0.997	0.4853	0.132	-0.115	0.007	-0.047	0.018	-0.038
4	64.694	66.411	1.042	0.9350	0.053	-0.065	0.015	-0.001	-0.061	-0.225
6	92.592	90.007	0.972	0.5333	-0.090	-0.045	-0.166	-0.097	-0.107	0.184
8	75.757	80.171	1.058	0.0328	0.068	-0.003	-0.011	-0.109	0.153	-0.068
10	79.365	81.979	1.032	0.7221	0.070	0.116	-0.228	-0.051	0.105	0.082
12	54.644	55.580	1.017	0.9033	0.003	-0.014	-0.010	0.125	-0.100	-0.046
14	43.290	46.247	1.068	0.9927	0.027	-0.058	-0.039	-0.015	0.026	-0.086
16	56.497	56.304	0.996	0.0242	-0.087	0.024	0.092	-0.051	0.108	0.051
18	26.666	26.867	1.007	0.9414	0.028	0.065	0.046	-0.027	-0.105	-0.04
20	25.445	24.017	0.943	0.3024	-0.026	0.001	0.034	-0.062	0.031	-0.083
22	29.761	34.074	1.144	0.0010	0.064	0.072	0.047	0.021	-0.023	0.012

tween -0.1 and 0.1, meaning that the interarrival times are almost independent of each other.

Although we are unable to display all the linear regression plots and quantile-quantile plots here due to limited space, similar observations hold for other traces at different hours. Table 2 summarizes average interarrival times, standard deviations, coefficients of variation, p-values of the Chi-square test, and sample autocorrelation functions for SMTP sessions arrived during ten seconds at selected hours in trace D. Note that the coefficients of variation are close to 1 and the most p-values are greater than 0.05. Sample autocorrelation functions also indicate the independence of the interarrival times.

4.2. Time Scales vs. Tailed Interarrival Times

Section 4.1 shows that SMTP session arrivals follow a Poisson process at small time scales. As the time scale increases, however, SMTP session arrivals do not obey a Poisson process.

Figure 4 depicts quantile-quantile plots of interarrival times in various intervals after 6 PM of trace D. As the interval becomes greater than ten seconds, points of the quantile-quantile plots deviate from a 45-degree line, indicating that the distribution of interarrival times does not follow an exponential distribution. Although points for small interarrival times still matches the diagonal line, other points corresponding to large interarrival times form a curve. As the interval becomes larger, points for large values move farther away from the line.

In the quantile-quantile plots shown in Figure 4, points away from the 45-degree line denote that the distribution has large values of interarrival times, called *tailed interarrival times*, with the probability of their appearance being extremely low. Due to these tailed interarrival times, the interarrival times of SMTP traffic do not conform to an exponential distribution at larger time scales. In terms of the number of interarrival times, the portion of such tailed interarrival times is less than 0.5%. When we investigate SMTP session arrivals in smaller time scales, the number of tailed interarrival times is too small to be significant. This makes it possible for SMTP session arrivals to be modeled as a Poisson process for shorter time intervals, as shown in Table 2. The existence of tailed interarrival times in SMTP traffic, however, plays an important role related to self-similarity, which will be discussed in the next section.

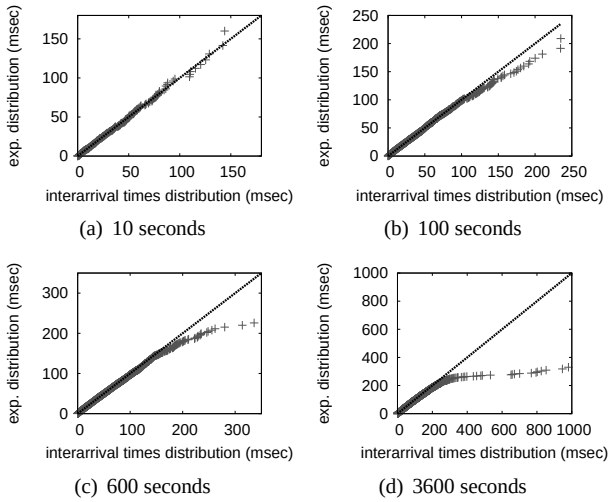


Figure 4. Quantile-quantile plots of interarrival times during (a) 10 seconds, (b) 100 seconds, (c) 600 seconds, and (d) 3600 seconds after 6 PM of trace D

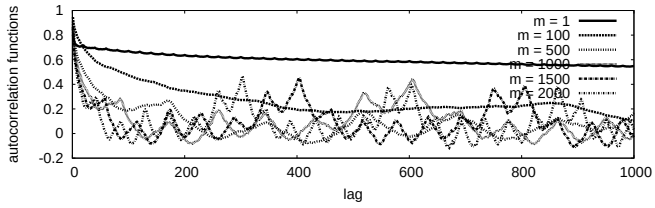
5. Self-similarity

In this section, we demonstrate self-similarity of SMTP traffic at large time scales using traces A, B, and C that span several months (cf. Table 1). Section 5.1 shows the presence of self-similarity by means of traffic visualization and autocorrelation functions. In section 5.2, we estimate the Hurst parameters using two rigorous statistical methods, variance-time plot and R/S analysis. In section 5.3, we investigate the distribution of ON/OFF-period lengths of individual SMTP sources to find out possible causes of self-similarity.

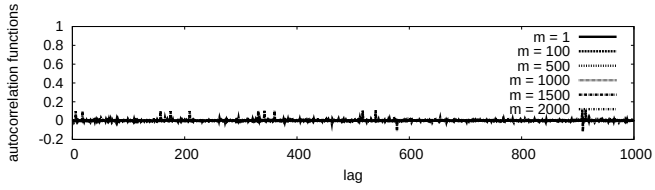
5.1. The Presence of Self-similarity

As described in section 2.2, for large aggregation levels, autocorrelation functions of the aggregated time series of a Poisson process converge to zero. However, the aggregated time series of a self-similar process have the asymptotically same non-degenerate autocorrelation function. This means that a self-similar process always has burstiness at every time scale, which is caused by scale-invariant variance. We can recognize the presence of self-similarity in SMTP traffic in Figures 5 and 6.

Figure 5 compares SMTP traffic with artificial traffic across six-order magnitude of time scales from 10 milliseconds to 1000 seconds. The graphs on the left side visualize the number of SMTP sessions per each bucket of trace C, while the graphs on the right illustrate the number of artificial sessions per each bucket. The artificial sessions are generated by compound Poisson processes whose mean arrival rates are adjusted to that of trace C. The bucket size of the top graphs is ten milliseconds and every successive graph has ten times larger bucket size than the previous one. All graphs have ten thousand buckets and each graph is the refinement of the successive one. Both of two top graphs, Figure 5(a) and (a′), have burstiness. However, as the time scale increases, while right graphs become extremely smooth and predictable, burstiness of left graphs still remains indicating self-similarity of SMTP traffic intuitively.



(a) Trace C



(b) Poisson traffic

Figure 6. Autocorrelation functions of the aggregated time series for (a) trace C and (b) artificial Poisson traffic

Self-similarity of SMTP traffic can be also seen in Figure 6, which depicts autocorrelation functions of the aggregated time series for trace C (Figure 6(a)) and the artificial Poisson traffic (Figure 6(b)). Each plot represents autocorrelation functions whose aggregation levels (m) are 1, 100, 500, 1000, 1500, and 2000 with lags from 1 to 1000. As shown in Figure 6(b), autocorrelation functions of the artificial Poisson traffic series are almost zero and relatively very small for all m. However, Figure 6(a) presents those of SMTP traffic series have the asymptotically same structure which is heavy-tailed and fluctuating, neither decaying exponentially nor converging to zero. We can observe that autocorrelation functions of SMTP traffic behave similar to self-similar processes and exhibit extensive long-range dependence.

5.2. Estimation of the Hurst Parameter

In this section, we estimate the Hurst parameter (H) of SMTP traffic using traces A, B, and C. We employ well-known graphical tools described in section 2.2, namely a variance-time plot and R/S analysis, to determine the Hurst parameters of these traces.

We first plot $log\big(Var(X^{(m)})\big)$ versus $log(m)$ for each trace varying $log(m)$ from 0 to 5, and obtain the slope (β) of each variance-time plot using a linear regression method. The corresponding H-value can be calculated as $1-\frac{\beta}{2}$.

Figure 7 illustrates variance-time plots and linear regression results for each trace. Against our expectation, points of each variance-time plot form a line that is piecewise linear, not totally. For small m, slopes of all plots are about -0.05, resulting in the estimated H-value of about 0.97 for all traces. As m increases, the estimated H-values shrink and become stable; when m is larger than 10000, H-values are reduced to 0.87, 0.85, and 0.84, for traces A, B, and C, respectively. These results are consistent with those shown in Figure 6. For small m, SMTP traffic exhibits high degree of burstiness and long-range dependence. When m becomes large, autocorrelation functions decrease and they have the asymptotically same structure. Such nonstationary H-values are due to the fact that SMTP traffic is not globally self-similar, failing to satisfy Equation 6 exactly. Nevertheless, all estimated H-values are noticeably larger than 1/2, and this implies self-similarity of SMTP traffic.

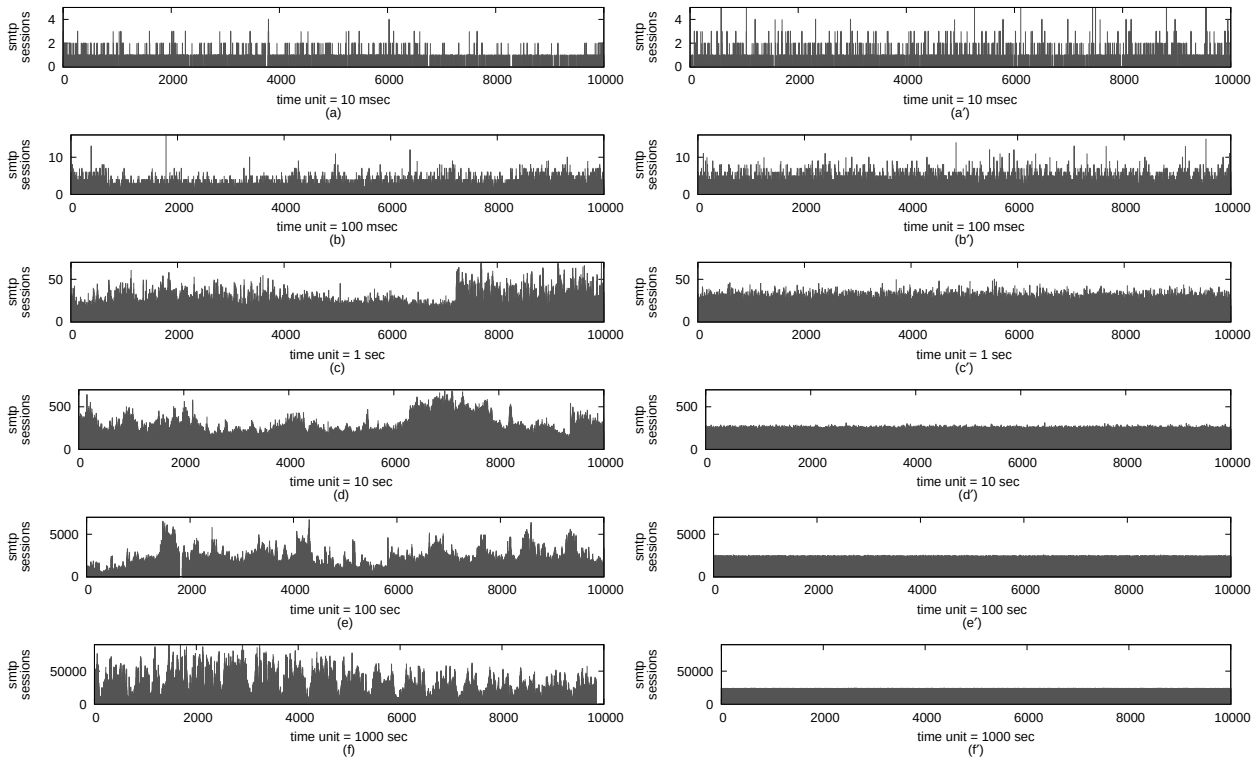


Figure 5. Visualization of trace C: Left figures illustrate the number of SMTP sessions per time unit for six different time scales (a)-(f). For comparison, right figures present artificial traffic which is generated by compound Poisson processes on the same six different time scales (a′)-(f′). Each figure has ten thousand buckets.

Note that the Hurst parameter of non-self-similar process such as a Poisson process is 1/2.

The result of R/S analysis also presents self-similarity of SMTP traffic. This analysis is originally described in [17] and detailed explanations can be found in [24]. In order to obtain H-value from Equation 9, we depict $log\big(E[\frac{R(n)}{S(n)}\big)$ versus $log(n)$ for non-overlapping blocks of each trace; this plot is called a *pox plot*. For a self-similar process, the pox plot produces a roughly linear graph with a slope equal to H. Given pox plots, we can compute their slopes by a linear regression method.

Figure 8 displays pox plots and linear regression results for traces A, B, and C. The slopes of approximated lines lie between 0.5 (the lower line) and 1.0 (the upper line), which suggest again self-similarity of SMTP traffic. Using pox plots, H-values are estimated to 0.89, 0.91, and 0.89, for traces A, B, and C, respectively. These results agree with those obtained previously from variance-time plots.

Table 3 summarizes estimated H-values of trace A, B, and C obtained through variance-time plots and R/S analysis. Although there are slight variations in the obtained H-values, all the Hurst parameters lie between 1/2 and 1, which confirms the presence of self-similarity in SMTP traffic at large time scales.

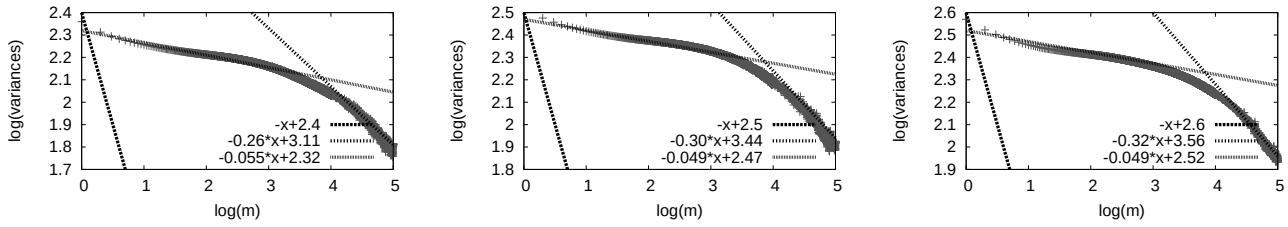
Table 3. Summary of estimated H-values

	Trace A	Trace B	Trace C
Variance-time plot	0.87∼ 0.97	0.85∼ 0.97	0.84∼ 0.97
R/S analysis	0.89	0.91	0.89

5.3. Heavy-tailed ON/OFF Sources

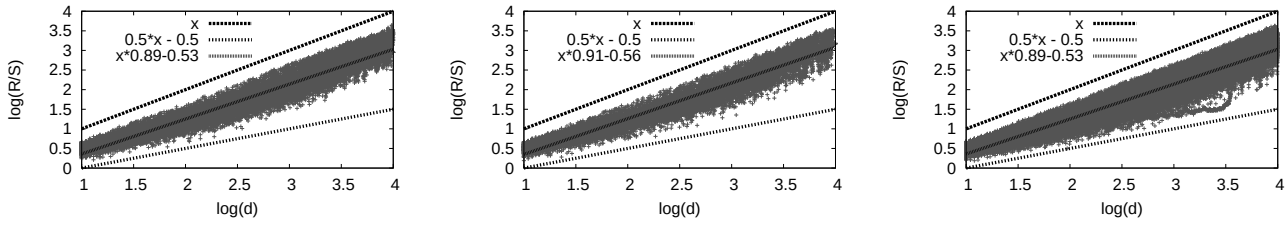
The previous sections show the presence of self-similarity in SMTP traffic by various analytical methods. In this section, we explore the reason of self-similarity in SMTP traffic in more detail based on the ON/OFF source model [9, 13, 14, 18, 21, 25]. To analyze individual sources of SMTP traffic, we apply similar methods described in [25] to the series of SMTP session arrivals whose time unit is millisecond. For this analysis, we have used a portion of trace C collected from August 22 to September 7.

To validate the ON/OFF modeling assumption for individual sources, we group SMTP traffic series by source IP addresses which represent SMTP servers at the client side



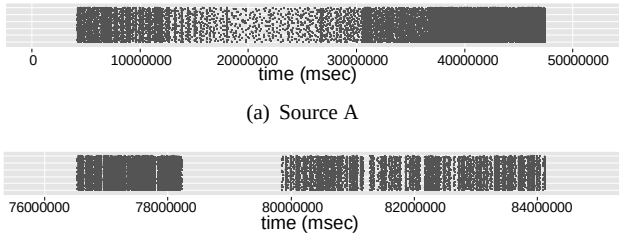
(a) Trace A (b) Trace B (c) Trace C

Figure 7. Variance-time plots and linear regression results for traces (a) A, (b) B, and (c) C



(a) Trace A (b) Trace B (c) Trace C

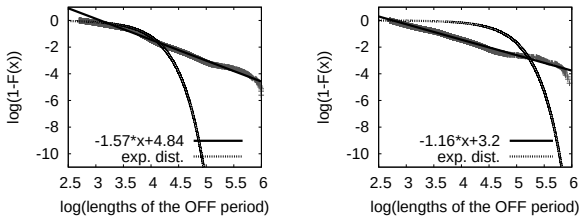
Figure 8. Pox plots and linear regression results for trace (a) A, (b) B, and (c) C



(a) Source A

(b) Source B

Figure 9. Textured dot strip plots for SMTP session arrivals of (a) Source A and (b) Source B



(a) Source A (b) Source B

Figure 10. CCDF plots for OFF-period lengths of two popular sources

and depict SMTP session arrivals of each source using a *textured dot strip plot* [23, 25]. Figure 9 shows the textured dot strip plots for two sources whose IP addresses are most popular. The plots in Figure 9 contain about ten thousand of SMTP session arrivals on August 23. Dots in each vertical column correspond to the density of arrivals per one millisecond. The *x*-axis represents the arrival time in millisecond. It is clear that both active and inactive regions are revealed in Figure 9, satisfying the assumption of the ON/OFF model. Other sources exhibit similar behavior.

We divide the series of SMTP session arrivals of each source into ON and OFF-periods based on the packet trains model [14] to verify the heavy-tailed distribution. ON-periods have SMTP session arrivals which occur within a threshold value of 500 milliseconds and there is no arrival in OFF-periods. For a given random variable X that has a heavy-tailed distribution, its cumulative distribution function should be described as follows.

$$F(X < x) \approx 1 - cx^{-\alpha}, x \to \infty \ \ where \ \ 1 < \alpha < 2 \quad (11)$$

Therefore, we examine the distribution of ON/OFF-period lengths in accordance with Equation 11.

Equation 11 suggests that plotting $log(1-F(x))$ versus $log(x)$ for a heavy-tailed distribution results in a roughly linear line whose slope is equal to $-\alpha$ for large *x*-values. Figure 10 depicts such CCDF plots with linear regression results for the OFF-period lengths of two selected sources. The points in two graphs construct approximately straight lines whose slopes are estimated to -1.57 ($\alpha = 1.57$) and -1.16 ($\alpha = 1.16$) for source A and B, respectively. The α-values between 1 and 2 denote that the distribution of OFF-period lengths is heavy-tailed, i.e., exhibits hyperbolic tails satisfying Equation 11. In Figure 10, the dotted line represents a CCDF plot for an exponential distribution whose average is same to the corresponding OFF-period lengths. Compared to the exponential distribution, we can observe

that the distribution of OFF-period lengths is heavy-tailed. For large *x*-values, while curves of the exponential distribution fall off, points of OFF-period lengths remain linear and their *y*-values are significantly larger than those of the exponential distribution. This means that the distribution of OFF-period length contains more large values (called *tailed values*) than the exponential distribution.

Alternatively, we can use Hill's estimator [25] to examine whether these distributions are heavy-tailed or not. Hill's estimation leads to α = 1.48 for source A, and α = 1.13 for source B, which are very close to α-values obtained from Figure 10.

We can also apply these analysis methods to ON-periods of the same sources. However, results of CCDF plots and Hill's estimation indicate that the distribution of ON-period lengths is similar to an exponential distribution rather than follows a heavy-tailed distribution. In spite of this, the heavy-tailed OFF-period lengths contribute to the self-similar nature of SMTP traffic.

6. Related Work

A solid understanding of realistic workload characteristics is essential for design and/or optimization of the underlying system. For the workload of email service systems, Bertolotti et al. [4, 5] presented the distribution of interarrival times for SMTP requests and the characteristics of emails such as message sizes and the number of senders/recipients, which have been the basis of the SPECmail benchmark [2]. Gomes et al. [11] demonstrated an intensive analysis on both spam and legitimate emails, including arrival processes, interarrival times, and the number of senders/recipients, in order to evaluate the influence of spam emails to aggregated email traffic. However, as their traces contain only a few hundred thousand SMTP requests with several thousand users taken over about two weeks, it is difficult to ensure that they provide a comprehensive analysis on the characteristics of SMTP traffic.

In this paper, we quantitatively analyze large-scale SMTP traces that span almost nine months with more than two hundred million of SMTP sessions. In particular, our traces are collected from one of the largest Web portal sites in South Korea which provides a webmail service to more than twenty million users. We believe we were the first to analyze this scale of SMTP traces obtained from a commercial Internet site.

Traditionally, the Poisson process has been considered when studying not only network traffic but also general workloads that can be treated as arrival processes. Paxson et al. [20] examined the feasibility of modeling numerous wide-area TCP arrival processes as a Poisson process, including TELNET, FTP, NNTP, and SMTP. They have found that user-initiated traffic such as TELNET connection and FTP session arrivals can be modeled as Poisson processes with fixed rates, whereas machine-generated bulk transfers such as SMTP, FTPDATA, and NNTP traffic can not. In contrast, our results show that SMTP session arrivals obey a Poisson process at small time scales.

Our results are similar to those obtained by Karagiannis et al. for TCP/UDP packet arrivals [15]. They showed that the packet arrivals on heavy Internet backbone traffic construct a Poisson process at sub-second time scales, but they exhibit long-range dependence at large time scales. These findings revisited the validity of the Poisson process after the discovery of self-similarity of network traffic.

Since the pioneering work of Leland et al. [16, 24], self-similarity has been one of the key considerations of heavy network traffic research. In [16, 24], the authors pointed out that Ethernet LAN traffic has self-similar nature through the estimation of Hurst parameters using several graphical tools. Similarly, analyses of variable-bit-rate (VBR) video [3] and World Wide Web [9] traffic have shown that they exhibit long-range dependence. In this paper, we examined self-similarity of SMTP traffic, which has been little attention in the previous work.

Willinger et al. [25] demonstrated that self-similarity is due to the superposition of many ON/OFF sources whose period lengths have a heavy-tailed distribution. They divided each individual source of self-similar Ethernet traffic into ON and OFF-periods based on the packet trains model [14]; an ON-period contains much activity, while there is no activity in an OFF-period. Then, it is shown that OFF-period lengths follow a heavy-tailed distribution. Other studies have supported these findings by showing that various kinds of self-similar traffic consist of many heavy-tailed ON/OFF sources [9, 18]. Detailed explanations and mathematical proofs of the relationship between heavy-tailed ON/OFF sources and self-similarity can be found in [19, 21, 25]. This paper confirms that individual sources composing SMTP traffic exhibit ON/OFF behavior and the distribution of their OFF-period lengths is heavy-tailed.

7. Conclusion

This paper investigates the characteristics of SMTP traffic based on large-scale SMTP traces collected from one of the largest Web portal sites in South Korea, which cover several months and more than a billion SMTP requests. Analyses of these traces have shown the coexistence of the Poisson process and self-similarity in SMTP traffic.

At small time scales, we find that SMTP session arrivals follow a Poisson process. The quantile-quantile plot and Chi-square test demonstrate that interarrival times of SMTP session arrivals, measured in ten-second intervals, obey an exponential distribution. Autocorrelation func-

tions of interarrival times also show that they are independent of each other. As the time interval of observation becomes large, interarrival times do not follow an exponential distribution. Instead, SMTP session arrivals exhibit self-similarity and long-range dependence at large time scales. The Hurst parameters of SMTP traces are estimated to lie somewhere between 0.85 to 0.97 by variance-time plots and R/S analysis. The fact that SMTP traffic consists of many individual ON/OFF sources whose distributions of OFF-period lengths are heavy-tailed confirms self-similarity of SMTP traffic.

Our findings can be actually applied when designing and optimizing network systems associated with SMTP traffic. For instance, simple statistical theories related to the Poisson process suggest that multiplexing SMTP traffic by a simple division results in a fair workload balancing at small time scales. We can also construct benchmarks for large-scale email server systems by the methodology based on the superposition of many individual ON/OFF sources whose OFF-period lengths obey heavy-tailed distributions such as the Pareto distribution.

For future work, we will investigate more specific causes explaining why SMTP traffic consists of individual sources whose OFF-period lengths follow heavy-tailed distributions. In addition, we plan to study the relationship between the time scale and the appearance of tailed interarrival times. This will be helpful to understand why the Poisson process and self-similarity coexist in SMTP traffic.

References

[1] Simple Mail Transfer Protocol(SMTP), http://www.ietf.org/rfc/rfc0821.txt.

[2] SPECmail2001, http://www.spec.org/osg/mail2001.

[3] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transections on Communications*, 43(2-4):1566–1579, 1995.

[4] L. Bertolotti and M. C. Calzarossa. Workload characterizations of mail servers. In *proceedings of the SPECT'2000*, 2000.

[5] L. Bertolotti and M. C. Calzarossa. Models of mail server workloads. *Performance Evaluation*, 46(2-3):65–76, 2001.

[6] M. C. Calzarossa. Performance evaluatoin of mail systems. *LNCS 2965*, 2001.

[7] D. Chakraborty, A. Ashir, T. Suganuma, G. M. Keeni, T. Roy, and N. Shiratori. Self-similar and fractal nature of internet traffic. *International Journal of Network Management*, 14:119–129, 2004.

[8] D. R. Cox. *Statistics: An Appraisal*. Iowa State Univ. Press, 1984.

[9] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[10] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, 2002.

[11] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida, and J. Wagner Meira. Characterizing a spam traffic. In *proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 356–369. ACM, 2004.

[12] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida, and J. Wagner Meira. Workload models of spam and legitimate e-mails. *Performance Evaluation*, 64(7-8):690–714, 2007.

[13] S. D. Gribble, G. S. Manku, D. Roselli, E. A. Brewer, T. J. Gibson, and E. L. Miller. Self-similarity in file systems. In *proceedings of the 1998 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 141–150, New York, NY, USA, 1998. ACM.

[14] R. JAIN and S. A. Routhier. Packet trains-measurement and a new model for computer network traffic. *IEEE Journal of selected areas in communications*, 4(6):986–995, September 1986.

[15] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A nonstationary poisson view of internet traffic. In *proceedings of the IEEE INFOCOM 2004*, pages 84–89. IEEE, March 2004.

[16] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994.

[17] B. B. MandelBrot and J. R. Wallis. Computer experiments with fractional gaussian noises. part 2: Rescaled bridge range and pox diagrams. *Water Resources Research*, 5:228–241, 1969.

[18] K. Park, G. Kim, and M. Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *proceedings of the 1996 International Conference on Network Protocols (ICNP '96)*, page 171, Washington, DC, USA, 1996. IEEE Computer Society.

[19] K. Park and W. Willinger. *Self-similar Network Traffic and Performance Evaluation*. Wiley, 2000.

[20] V. Paxson and S. Floyd. Wide-area traffic: the failure of poisson modeling. In *proceedings of the conference on Communications architectures, protocols and applications*, pages 257–268, New York, NY, USA, 1994. ACM.

[21] M. S. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *SIGCOMM Computer Communication Review*, 27(2):5–23, 1997.

[22] H. Tijms. *A First Course in Stochastic Models*. Wiley, 2003.

[23] J. W. Tukey and P. A. Tukey. Strips displaying empirical distributions: I. textured dot strips. *Bellcore Technical Memorandum*, 1990.

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

[25] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, 1997.