creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Master's Thesis

Real-Time Estimation of Self-Similarity in Wireless Traffic

Cong Liu（劉 聰）

Department of Computer Science and Engineering

Pohang University of Science and Technology

2012

dCollection @ postech

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Real-Time Estimation of Self-Similarity in Wireless Traffic

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Real-Time Estimation of Self-Similarity in Wireless Traffic

by

Cong Liu
Department of Computer Science and Engineering
Pohang University of Science and Technology

A thesis submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Computer Science and Engineering

Pohang, Korea
6. 29. 2012
Approved by

Young-Joo Suh
Academic Advisor

Real-Time Estimation of Self-Similarity in Wireless Traffic

Cong Liu

The undersigned have examined this thesis and hereby certify that it is worthy of acceptance for a master's degree from POSTECH

06/29/2012

Committee Chair Young-Joo Suh (Seal)
Member Chansu Charles Yu (Seal)
Member Raouf Boutaba (Seal)

For Many Dreams

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

MECE
20101012

Cong Liu, 劉聰, Real-Time Estimation of Self-Similarity in Wireless Traffic. Department of Computer Science and Engineering, 2012, 53P, Advisor: Young-Joo Suh. Text in English.

ABSTRACT

Self-Similarity (S-S) is a valuable property of various networks traffic being abundantly studied for both analysis and applications. It is promising to test the S-S property for recent network traffic and potential discoveries would contribute to more useful applications. Besides, although off-line analysis of the S-S property has predominated since 1990s, there are recent studies showing that real-time estimation of the S-S property can be applied to some tailor-made applications. For example, it may help detect some network attacks because the S-S property parameter reacts quickly to those attacks. We first demonstrate the existence of the S-S property of a recent collection of 802.11 wireless network data traces, and then present a faster implementation of an existing linear-modeling method Variance-Time Plot (VTP) for those data traces, as well as a non-linear modeling of VTP plot being examined and shown to represent the S-S property more appropriate and accurate for testing data traces. Results showed that our proposed method achieved the processing time much shorter than original VTP method, while the accuracy remained in an accepted range. This performance would enable several real-time applications which require swift response to the change of the S-S property. For the non-linear modeling of VTP plot, the accuracy was shown to be as high as the proposed modified VTP method. Meanwhile, the results showed that the new parameter set of proposed method would enhance the understanding and performance of original VTP method.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Contents

I. INTRODUCTION 1
  1.1. Overview of Self-Similarity 1
  1.2. Motivation 3
  1.3. This Thesis 5
  1.4. Thesis organization 5
II. METHODS 6
  2.1. Mathematical definitions 6
    2.1.1. Network traffic modelling 6
    2.1.2. Self-Similarity 12
    2.1.3. Long Range Dependence 14
    2.1.4. Short Range Dependence 14
  2.2. Variance-Time Plot 17
    2.2.1. Original version 17
    2.2.2. Real-time version 20
  2.3. Proposed method 20
    2.3.1. Concept 20
    2.3.2. Algorithm 24
III. EXPERIMENT AND RESULTS 30
  3.1. Experiment setup 30

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

3.1.1. Bellcore Ethernet data sets 31

3.1.2. Portland wireless data sets 32

3.2. Metrics and results 33

3.2.1. Time 34

3.2.2. Error rate 36

3.2.3. Confidence intervals 40

IV. DISCUSSION 41

4.1. Choosing $m_{maximum}$ 41

4.2. Estimation of $m_{critical}$ 45

4.3. Choosing δ_{time} 45

4.4. Validity of result metrics 46

4.5. Validity of estimating $\widehat{H}$ using $\widehat{m_{stable}}$ 46

4.6. Reasons of curve fluctuation 46

4.6.1. Defect in Variance-Time Plot realization 47

4.6.2. Mixed properties for the data traces 47

4.7. Non-linear representation of Variance-Time Plot 48

4.7.1. Polynomial trend-line 48

4.7.2. Exponential trend-line 51

V. CONCLUSION 53

LIST OF FIGURES iv

LIST OF TABLES v

REFERENCES 54

ABSTRACT i

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

ACKNOWLEDGMENT 60

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

LIST OF FIGURES

Figure 1. An example of aggregated series 9

Figure 2. An example of the log-log plot 11

Figure 3. The log-log plot of an S-S time series...13

Figure 4. The log-log plot of an SRD time series...16

Figure 5. An illustration of the original VTP method 19

Figure 6. The concept of proposed method illustrated 23

Figure 7. The pseudo-code of the Step 3..27

Figure 8. The processing time comparison chart..38

Figure 9. The Variance-Time Plot for Portland powells sub data traces........39

Figure 10. Errors occur when $m_{maximum} = \left\lceil \frac{n}{m} \right\rceil$ 42

Figure 11. Polynomial third degree trend-line test...49

Figure 12. Polynomial second degree trend-line test 50

Figure 13. Exponential trend-line test 52

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

LIST OF TABLES

Table 1. The results of Bellcore data analysis 37

Table 2. The results of Portland data tests 37

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Chapter I

INTRODUCTION

In this chapter, we introduce the overview of Self-Similarity, the motivation of this study, the topic of this paper, and the organization of this paper.

1.1 Overview of Self-Similarity

In some network traffic traces, the presence of “burstiness” of packet sizes exists across a wide range of time scales [3]. In particular, for consecutive hours of monitored network traffic, it can be observed that the large time unit (e.g. minute) traffic-time plot looks similar to small time unit (e.g. millisecond) plot. This observation exists within an obvious daily cycle of the network traffics. This behavior is so-named as *Self-Similarity* (S-S) because the similarity of randomness and “burstiness” cannot be distinguished by human eye across different scales.

For convenience, we adopt the term “zoom-in” for changing scale from large time-unit to small one, and “zoom-out” for the opposite way. A network traffic with S-S behavior can better represented with S-S model other than previously suggested stochastic models such as Poisson model. The major difference lies in the situation when the network traffic-

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

time plot has been zoomed out by averaging or aggregating some significant time units, the aggregating packet size (i.e. the vertical axis value) can be of little difference with each other over the whole time scale [3].

The S-S is based on *Long Range Dependence* (LRD) [1], [2]. The LRD property implies a dependent relationship existing between two distant packet-size sample points within a restrained time domain. On the contrary, some stochastic models, such as Markov process, suggest independence between sample points, and other models suggest independence between sample points that are sufficiently far away from each other [1]. This LRD property contributes to the self-similar behavior described in above-mentioned phenomenon.

The S-S property are important for design, control, and analysis of high-speed, cell-based networks, and that aggregating streams of such traffic typically intensifies the "burstiness" instead of smoothing it. It also help analyze the network traffic in unique perspectives, and even provide innovative solutions to some existing problems.

The S-S property is valuable for many kind of networks: Ethernet Local Area Networks (LANs) [3], World Wide Web [4], [5], Wireless LANs [6], [7], [8], some 802.11 wireless LAN (WLAN) networks [9], voice and video telephony networks such as Voice over Internet Protocol [10], [11] and Peer-to-Peer networks such as BitTorrent and Botnet networks [12], [13].

Previously the Poisson model was assumed to be an appropriate stochastic model for network traffic arrival process, and later, the S-S model for the network traffic are proven to be more appropriate for some network traffics than the Poisson model [3]. An S-S

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

behavior is distinguished from the Poisson-modeled behavior when the traffic-time plot is zoomed out to large time-unit scale. The S-S phenomenon might be caused by different reasons for a specific network [2], [5].

1.2 Motivation

First motivation is to examine the S-S behavior in recent wireless network.

The S-S property of a variety of network traffics has been studied since 1990s and those studies become intensive recently [3], [2], [31], [6]. Some of the 802.11 wireless network traffics were believed to have the S-S property based on some empirical studies [9]. It is beneficial to continue exploring S-S property in recent wireless networks. For example, we can have more useful applications based on this property: it helps improve the efficiency and performance of network planning, resource management, Media Access Control (MAC)-layer algorithms [9], and network control, such as allocating output bandwidth [17] and measurement-based admission control [32].

Second motivation is the demand of real-time estimation of the S-S property in some applications.

In early times, off-line methods would suffice in most applications of the S-S property. The S-S property indicator (the H parameter) was calculated off-line, i.e., the estimation was processed after finishing collecting all the traffic data. These off-line methods include Variance-Time Plot (VTP) [3], rescaled analysis [3], [19], periodogram analysis [19], Whittle's estimation [19], and wavelet analysis [19]. They served as accurate and acceptable models for estimation given enough analysis time and facilities, because the

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

capability of computers in early times is limited and H calculation requires large traffic trace sampling data.

However, as time goes by, the real-time methods for the S-S property analysis have recently emerged as an interesting topic. The “real-time” means that the estimation work is conducted in parallel as the monitoring is going on.

This yields quick responses, so, comparing to off-line methods, real-time calculation of H can be served for some special applications, such as protection against Denial of Service (DoS), Distributed Denial of Service (DDoS) or Botnet attacks [12], [14], [15], [20], [31]. That is because, during those attacks, H will increase suddenly given such attacks and drop back to normal quickly too after the attacks [31].

Other applications include (1) network queue and loss analysis in the routers/switches with low CPU consumption [16]; (2) predicting the size of the data buffer from the degree of S-S [18], we could give a rough estimation of a minimum buffer size according to the tolerance coefficient defined in specific situation tailored to individual customer’s needs; and (3) exploiting characteristics of network traffic management, for example, admission control, rate adaption, and network monitoring [21].

There real-line methods such as: online VTP [17], online rescaled analysis [20], online wavelet analysis [19], and modified embedded branching process [18], [22]. Each of them has both advantages and disadvantages. We chose the VTP method was chosen as baseline method of the proposed method in the paper among other off-line methods, because it has some advantages over the other related methods, such as meaningful visual expression and intuitive mathematical concepts.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

1.3 This thesis

We demonstrate the existence of the S-S property a specific recent collection of 802.11 wireless network traffic traces, and then present a VTP-based technique to enable real-time estimation of Hurst parameter, as well as a non-linear modeling of VTP plot.

The proposed method is modified from VTP, a graphical method, to shorten the estimation time of H by some approximation within acceptable accuracy. We introduce refine the original VTP method by some pruning technique. Contrary to previous real-time H estimation works [17], [18], we tried to show that real-time estimation may be achieved by exploring the property of

Experimental results showed that the processing time can be reduced by about 90% in relative comparison from the original VTP method while the accuracy remained in accepted range.

1.4 Thesis organization

This thesis is organized as following: Chapter I briefly introduces the background information and motivation of the proposed method; Chapter II presents the proposed method in detail; Chapter III lists the experimental results; Chapter IV hosts some technical discussion; and Chapter V concludes this thesis.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Chapter II

METHODS

In this Chapter, we first introduce basic mathematical background related to this paper. Secondly we explain the concepts of classical linear-modeling VTP method. Then we move to the details of proposed method. A faster implementation of the classical VTP method is proposed as a way to realize the S-S property parameter real-time estimation.

2.1 Mathematical definitions

2.1.1 Network traffic modelling

We consider a series of the size of packet arrivals from a network traffic trace as a time series (a discrete stochastic process):

$$X \overset{\text{def}}{=} \{X_t : t \in \mathbb{N}, t \leq n\}. \qquad (2\text{–}1)$$

In Equation (2–1), subscript t refers to the "time" of the series or the numbering of the series; $\mathbb{N}$ is the set of natural numbers; n is the number of samples in that series. Relating to this time series, a few properties are denoted as follows. The *mean* (*expectation*) in time t is denoted by:

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

$$\mu_t \overset{\text{def}}{=} E[X_t]. \tag{2–2}$$

The *variance* in time t is denoted by:

$$\sigma_t^2 \overset{\text{def}}{=} \mathrm{Var}[X_t] = E[(X_t - \mu_t)^2]. \tag{2–3}$$

The *autocorrelation function* (ACF) is denoted by:

$$r(t_1, t_2) \overset{\text{def}}{=} E\left[X_{t_1} \cdot X_{t_2}\right]. \tag{2–4}$$

In Equation (2–4), both t_1 and t_2 are time indices.

Usually, it is assumed that the time series of the size of packet arrivals of a network traffic traces follows these conditions of Equations (2–5), (2–6), and (2–7):

$$\mu(t) = E[X_t] \text{ is independent of t;} \tag{2–5}$$

$$r(t_1, t_2) = r(0, t_2 - t_1) = r(\tau),\ t_2 \geq t_1 \geq 0,\ \tau \overset{\text{def}}{=} t_2 - t_1; \tag{2–6}$$

$$r(0) = E[(X_t)^2] < \infty, \text{ (i.e., finite second moment).} \tag{2–7}$$

When this simplified time series for network traffic packet arrivals met the three conditions of Equations (2–5), (2–6), and (2–7), it is called a *wide-sense stationary time series* for convenience [23], [24]. The “stationary” condition means the distribution of X in each time t is time-invariant under shifts of the time origin, as Equation (2–5) shows. These conditions would lay a basis of assumptions for the following properties of interest. The “wide-sense stationary” property implies the ACF can be determined by the time difference only, Equation (2–6). In many industrial applications including this thesis’s topic, it is usually assumed that this time-invariant property exists [3], [19]. While time series which is not wide-sense stationary would have an ACF such as Equation (2–4). The ACF is a measure of dependence among random variables of X [23].

We define an *aggregated series* of X as $X^{(m)}$, such that:

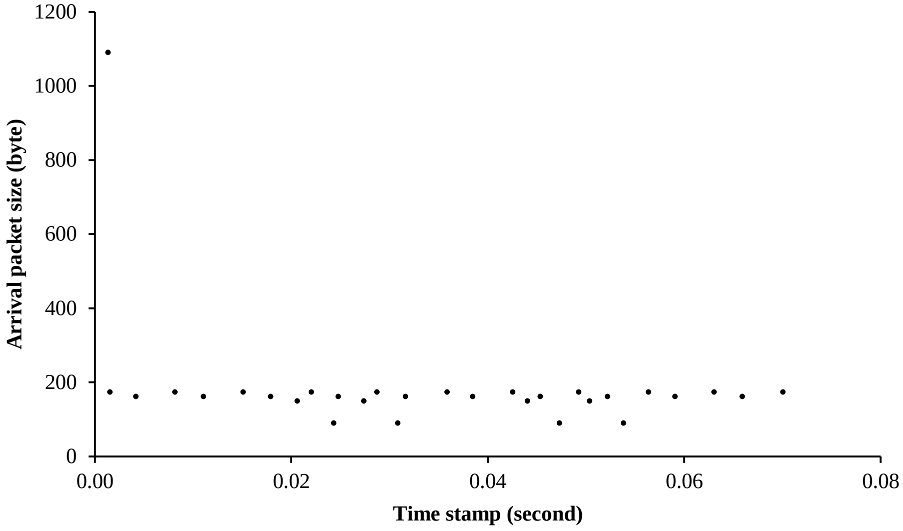
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

$$X^{(m)} = \left\{ \begin{matrix} X_t^{(m)}: X_t^{(m)} = \frac{1}{m}\sum_{i=(t-1)m+1}^{tm} X_i \,, \\ t \in \mathbb{N}, t \leq \frac{n}{m}, m \in \mathbb{N}, 1 < m \leq \frac{n}{2} \end{matrix} \right\}. \tag{2–8}$$

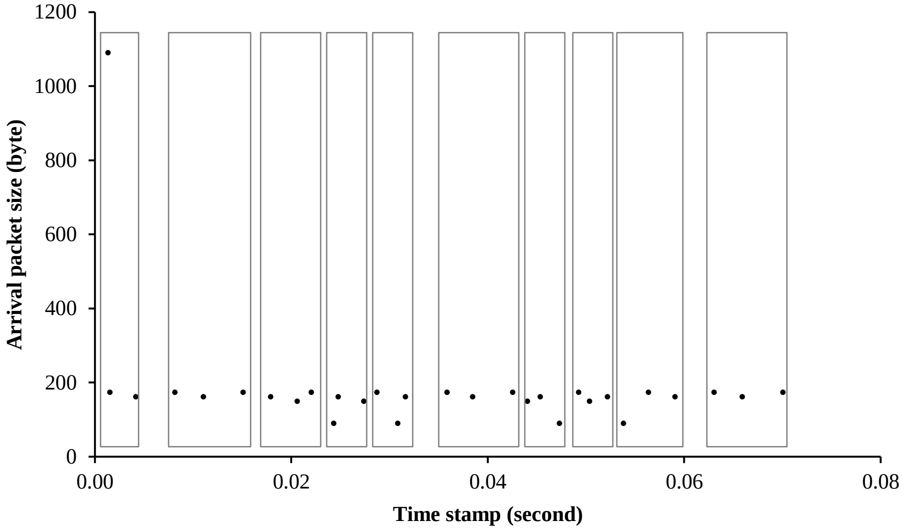
The m is called *aggregation level* or *aggregation index*. This aggregated series $X^{(m)}$ represents some "zoomed-in" versions of X. An example of artificial data is given for illustration (Fig. 1).

Then we introduce the "log-log plot" (Fig. 2). The $\mathrm{Var}\left[X^{(m)}\right]$ v.s. m plot (in logarithm scale) to be used in this paper is henceforth referred to as *the log-log plot* unless otherwise specified. The $\mathrm{Var}\left[X^{(m)}\right]$ is the variance value of aggregated series $X^{(m)}$. As the horizontal axis value m increases, the point with a larger m corresponds to an aggregated series that is a zoomed-in version of the one represented by a smaller m. The trend of the plot curve indicates the trend of variance change as the time series X being zoomed in. Let k be the slope of the trend line, we can calculate the degree of the S-S property by a relation between k and the Hurst parameter.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

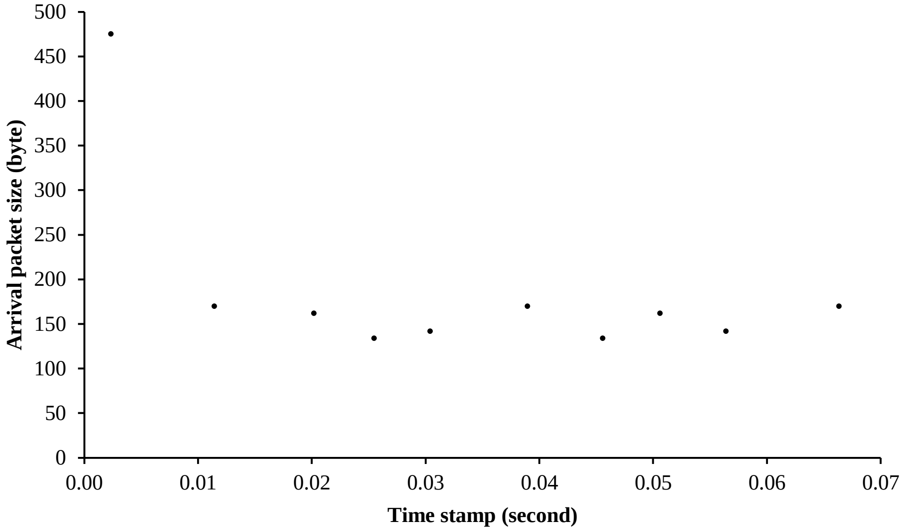


(a)



(b)

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교



(c)

Figure 1. An example of aggregated series. (a) an example 30-packet arrivals traffic traces, denoted as time series $X = \{X_1, X_2, \cdots, X_{30}\}$ (b) blocks showing the averaging process (c) for $m = 3$, the plot of aggregated series $X^{(3)} = \{X_1^{(3)}, X_2^{(3)}, \cdots, X_{10}^{(3)}\}$

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

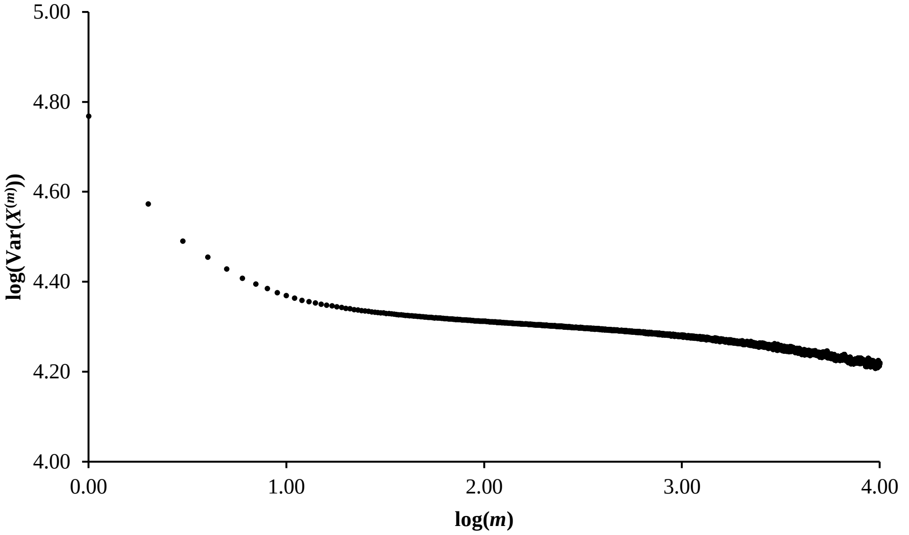


Figure 2. An illustration of the log-log plot. The data come from BC data (Chapter III) sub set Oct89Ext4. Packet arrival sample population $n = 1{,}000{,}000$, aggregation level $2 \leq m \leq 10{,}000$.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

2.1.2 Self-Similarity

There are two kinds of self-similarity (S-S) property relating to our discussion. One is exactly self-similarity. An *exactly self-similar* time series X satisfies:

$$X = m^{1-H} \cdot X^{(m)}. \tag{2–9}$$

The other one is second-order self-similarity. A time series X is a *second-order self-similar* time series if X and $m^{1-H}X^{(m)}$ share the same variance and auto-correlation.

In the log-log plot interpretation, the above property is translated into a slow decrease of the aggregated series variance as the aggregation level m increases. An example of an artificial S-S time series is illustrated afterward to show how the trend of aggregated series variance goes when m increases (Fig. 3).

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

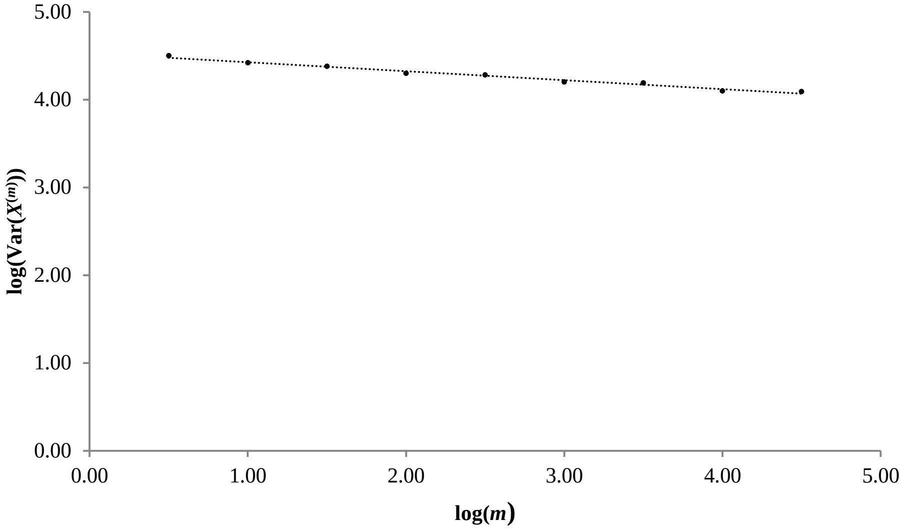


Figure 3. The log-log plot of an S-S time series. This artificial S-S time series X generates a nine-point (nine m values) log-log plot with a trend line slope $k \in (-1,0)$. The dotted line is the trend line. The trend line equation is $y = -0.1027x + 4.5301, r^2 = 0.982$.

2.1.3 Long Range Dependence

As a special case of S-S property, the Long Range Dependence (LRD) is also known as asymptotically second-order self-similar.

The LRD is the loosened version of exact self-similar time series. Comparing to Equation (2–9), if X and $m^{1-H}X^{(m)}$ share the same variance and auto-correlation, as $m \to \infty$, the X is said to have LRD property [2].

The LRD is also the opposite of SRD, characterizing those time series having a strong interdependence between distant samples [1]. Similar to SRD, the $\log\left(\mathrm{Var}\left[X^{(m)}\right]\right)$ v.s. $\log(m)$ plot gives a visualization of the LRD property: the plot of an LRD X has a trend-line flat slope k such that $0 < k < -1$. If $k \to 0$, X is bearing LRD property in a perfect sense [19].

The S-S property and LRD property are used interchangeably in many publications in literature, and the S-S property was sometimes interpreted as the LRD property [2]. The VTP method holds for both the exact S-S definition and the LRD definition [3].

2.1.4 Short Range Dependence

As a comparison to LRD property, the *Short Range Dependenc*e (SRD) property suggests that time series X has an ACF with exponential decrease as time increases:

$$r(\tau) \sim \rho^{\tau}, \rho \in (0,1), \tag{2–10}$$

where τ denotes the time difference $t_2 - t_1$ as in Equation (2–6). In other words, Equation (2–10) can also be represented as:

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

$$\sum_{\tau} r(\tau) < \infty, \tag{2–11}$$

The SRD property can be visualized by a relatively steep trend-line slope of the log-log plot. For example, for the previously mentioned Poisson process (a stationary process), the slope value of trend line $k = -1$, so this process bears SRD property [17], [3].

An illustration of an example SRD time series is followed, using artificial data (Fig. 4).

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

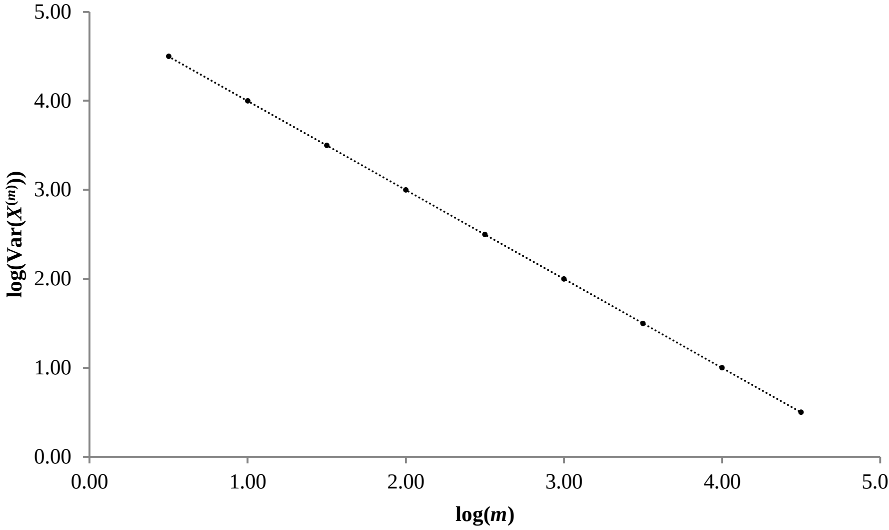


Figure 4. The log-log plot of an SRD time series. This artificial SRD time series X has a nine-point (nine m values) log-log plot with a trend line slope $k = -1$.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

2.2 Variance-Time Plot

2.2.1 Original version

The VTP method was originated from an equivalent interpretation of the LRD definition:

$$\mathrm{Var}\left(X^{(m)}\right) \sim c m^{-1}, \text{ as } m \to \infty \tag{2–12}$$

where c is a finite positive constant. The steps for the original VTP method are as follows:

First, calculate $X^{(m)}$ for all possible aggregation level m according to Equation (2–8), and obtain the corresponding log-log plot.

Then from the points of log-log plot, a linear trend line is estimated by the commonly used ordinary least square method. Record the slope value k of the trend line.

Finally, the Hurst parameter H can be obtained by the relation between H and the trend-line slope k of the log-log plot yielded from Equation (2–12) (Fig.):

$$H = 1 + \frac{k}{2}. \tag{2–13}$$

For S-S time series X, the Hurst parameter follows:

$$H \in (½, 1). \tag{2–14}$$

If $H \to 1$, the time series X is a perfectly self-similar one, i.e.,

$$X = m^{1-H} X^{(m)} \text{ for all } m. \tag{2–15}$$

If $H \to ½$, the time series is degraded to an SRD one [1], [5], [19], [25].

An illustration of the original VTP method is followed (Fig. 5).

The advantage of the original VTP method first lies in the visualization. The trend as the variance of aggregated time series decreases as the aggregation level m increases.

Since VTP is used as an off-line analysis tool in early times, a disadvantage of this

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

original VTP is the slow processing time. To obtain a thorough analysis of the log-log plot, all the possible values of aggregation level m (from 2 to $m_{maximum} = \left\lfloor \frac{n}{2} \right\rfloor$) would be put into consideration. This is a brute-force realization following the definitions of aggregated series and the S-S property. Thus, optimization of the VTP method is a way to improve the processing time and enable the real-time version of VTP.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

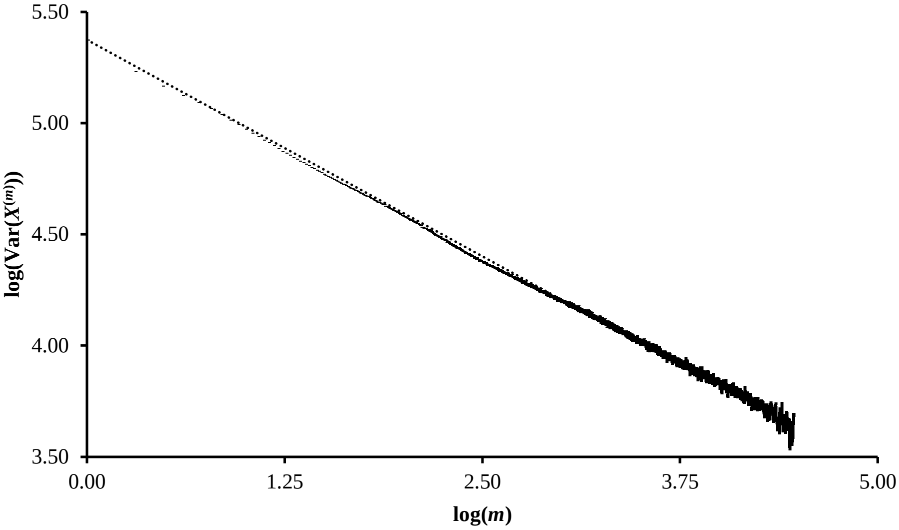


Figure 5. An illustration of the original VTP method. The data are adopted from BC data sub set “pAug89” with one million packet arrivals samples, and $m_{maximum} = 29{,}000$. There are many points of variance values for each aggregation level m (thick dash marks). The trend line is the light dotted line, having an expression of $y = -0.3894x + 5.3748$, $r^2 = 0.9792$. $\widehat{H} = 0.8053$.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

2.2.2 Real-time versions

First, we choose the VTP method as the baseline method for real-time estimation method development. This is because the VTP method is a visual method that requires less complex mathematical background, and it is clearer and more intuitive than most of other off-line methods.

There are some other real-time estimation of Hurst parameters. Some authors developed their real-time estimation method (based on VTP) using truncated historical packet arrival samples [17]. This approximation by reducing sample sizes may be inappropriate, because the S-S property is scattered among a large set of samples, although the limited experimental results showed no hazard. Sample size can be important to the estimation of H, as shown previously in terms of high-precision timestamps (equivalently the large sample size) [3].

2.3 Proposed method

2.3.1 Concept

The proposed method is based on the original VTP method. In the original VTP method, the Hurst parameter H can be estimated by the slope of log-log plot trend line by the relation of Equation (2–13). We try to reasonably reduce the points of the log-log plot while maintaining the accuracy of estimated value of H. Here is the major steps:

Firstly, stop generating the plot points when m is large enough. This “large enough” m is denoted as $m_{critical}$. A relatively small portion of the set of m is proposed to be

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

enough to get accurate value of estimated H. If m is too large, it may trigger significant error to the variance values according to analysis result of pre-processing (Chapter IV).

Thus this first step is summarized as “to determine $m_{critical}$”.

Secondly, check whether the chosen $m_{critical}$ is appropriate. We define the initial value of m_{stable} as the middle point between 2 and the estimated $m_{critical}$. This checking process is performed by comparing the accuracy of slope values yielded from the estimated $m_{critical}$ and the estimated m_{stable} values using divide-and-conquer algorithm to find out the most appropriate m_{stable} value in between the domain of m, $(2, \widehat{m_{critical}})$ (the angle hat symbol means an estimated value). The criterion is that the best $\widehat{m_{stable}}$ gives the least error of slope value comparing to that using $\widehat{m_{critical}}$.

In this second step, if the estimated $m_{critical}$ is proven to be too small, then go back to the pre-processing phase to adjust the tolerance coefficient higher.

Thus this second step is summarized as “to determine m_{stable}”.

Finally, an estimated H value is calculated based on the most accurate estimation of m_{stable} in previous step. This final step is summarized as “to estimate H”.

To support the above concept of proposed method, we have developed a theoretical background:

The log-log plot can be divided into three segments according to our own analysis on the BC data. We number these three segments as segments S_1, S_2, and S_3, in the order as the aggregation level m increases. Then let $m'_{critical}$ be the value of m that lies on the border between segments S_1 and S_2, and let $m_{critical}$ be the value of m that lies on border between segments S_2 and S_3. Let m_{stable} be the value of m in the middle of

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

segment S_2 (Fig. 6).

Segment S_1 lasts for a small range of m starting from $m = 1$, and follows a quick decrease on the variance of aggregated series, even though sometimes not so distinguishable from segment S_2. In practice, segment S_1 is a small portion comparing to the whole set of m values, so we merge segments S_1 and S_2, and regard $m'_{critical}$ as equivalent as $m_{minimum}$. By the definition of aggregated series $X^{(m)}$ in Equation (2–8), $m_{minimum} = 2$. Segment S_2 is a relatively stable linear trend-line part comparing to segments S_1 and S_3. In segment S_2, the middle m values is referred to as m_{stable} for checking the accuracy of estimated $m_{critical}$. Segment S_3 is the most time-consuming and non-linear part. In segment S_3, the trend line would begin to be degraded into fluctuating curve, because of the co-existence of SRD property of the traffic traces [26], [27], the precision of timestamps [3], as well as the errors become more and more significant as $m \to \frac{n}{2}$. In Chapter IV, there is some discussion about latter cause.

Intuitively, each of the three segment of the log-log plot needs at least one parameter to represent its property: $m'_{critical}$ (border of segment S_1 and S_2) for the first and second segment, m_{stable} (an appropriate sample in segment S_2) for the second segment, $m_{critical}$ (border of segment S_2 and S_3) for the second and third segment, and k (slope) for the whole curve. Thus this four parameter set is chosen for the three-segment log-log plot curve.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

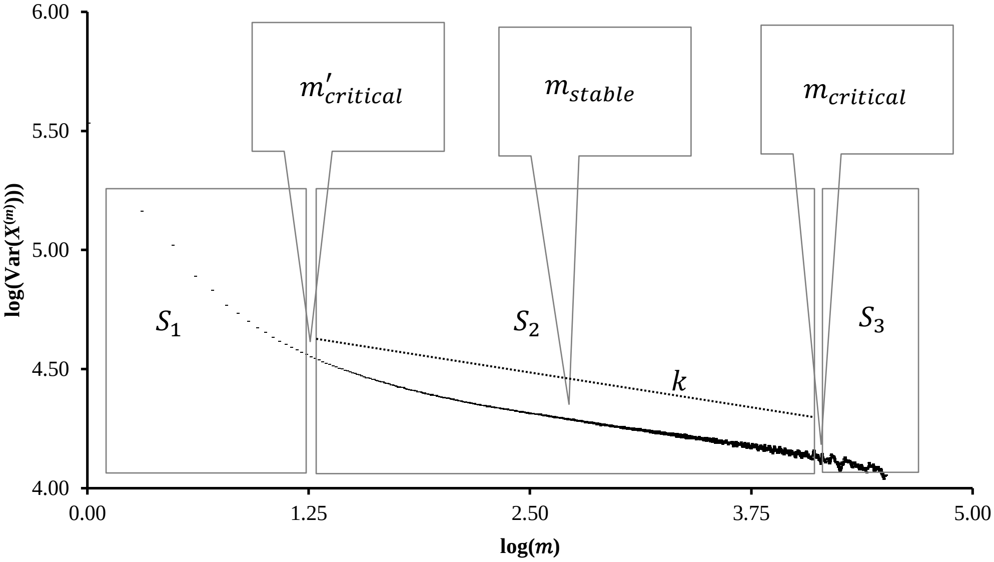


Figure 6. The concept of proposed method illustrated. There are three segments (S_1, S_2 and S_3) dividing the VTP log-log plot. To describe the three segments, we choose three special value of the aggregation level m and the slope k of the trend line (the light dotted line above the curve segment S_2) for the middle segment S_2.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

2.3.2 Algorithm

Previous section briefly introduce the main concept of proposed method. Here shows some major steps on how the $m_{critical}$ and m_{stable} can be estimated, then the Hurst parameter H:

Step 1: Define tolerance coefficient (α_{tc})

Step 2: Estimate $m_{critical}$

Step 3: Estimate m_{stable} and H

- Step 3.1 Initialization for while loop
- Step 3.2 Running while loop
- Step 3.3 Pin down $\widehat{\mathrm{m_{stable}}}$ and $\widehat{\mathrm{k_{stable}}}$, and $\widehat{\mathrm{H}}$

The complexity of the proposed algorithm is $O\left(\log(n)+\widehat{m_{critical}}^{2}\right)$, with time spent primarily on Steps 2 and 3. In the worst case, if the tolerance coefficient is not properly determined, the complexity of proposed method can be degraded back to $O(n^2)$, same as the original VTP. The detailed algorithm is listed below:

Step 1: Define a tolerance coefficient α_{tc}

We define a *tolerance coefficient*, denoted as α_{tc}. Considering the sample population in our experimental data and the definition of aggregated series $X^{(m)}$ in original VTP, this α_{tc} can be set as 0.01 and we used it in our Portland data experiments. For detailed discussion, see Chapter IV.

In particular, if we have:

$$\alpha_t \coloneqq 0.01, \tag{2–16}$$

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

where the symbol “≔” stands for “is assigned with”, it is suggest that the difference between an error and the true value less than one percent of the true value can be tolerated, and the error with more than one percent difference may be too significant to be ignored.

The α_{tc} value can be determined in training or pre-processing period, i.e., the beginning period of monitoring. For example, some significant candidates of α_{tc} values are 0.1, 0.01, 0.001 or 0.0001. However, when $\alpha_{tc} \coloneqq 0.1$, the tail of the log-log plot still remain too many samples that slow down the whole calculation.

The algorithm can be added with a section to determine the most suitable value of α_{tc} by comparing the estimation result of H and processing time for those preset candidates of α_{tc}. This adaption part was not tested in the experiment part.

Step 2: Define a threshold $\boldsymbol{m_{\widehat{critical}}}$

Find out a threshold value of m, denoted as $m_{critical}$. According the tolerance coefficient α_{tc}, the relation follows:

$$m_{\widehat{critical}} \coloneqq \alpha_{tc} \cdot m_{\widehat{maximum}},$$

where $m_{\widehat{maximum}} = \left\lfloor \frac{n}{2} \right\rfloor$. The n is the number of samples or sample population, and $m_{maximum}$ is the theoretical maximum value of m, by the definition of aggregated series $X^{(m)}$.

A discussion of using floor ($\left\lfloor \frac{n}{2} \right\rfloor$) or ceiling ($\left\lceil \frac{n}{2} \right\rceil$) for the maximum value of m is placed in Chapter IV.

Step 3: Estimate $\boldsymbol{m_{stable}}$ and $\boldsymbol{H}$

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Estimate the slope of trend line, within tolerance, for the log-log plot using a divide-and-conquer strategy, under a condition to decide the end of loops, which is guaranteed to converge as each loop reduces half of m values (Fig. 7).

Practically, $\widehat{m'_{critical}}$ is replaced with $m_{minimum}$ in the algorithm (by lines 4 and 5) because $m'_{critical}$ was usually a small number comparing to $m_{critical}$ from the analysis of BC data. Some reasoning and discussion on each step of the following algorithm is put in Chapter IV.

Step 3.1 Initialization for while loop

Find the estimated log-log plot trend slope, $\hat{k}$, using $\widehat{m_{critical}}$ as new $\widehat{m_{maximum}}$. Mark this special $\hat{k}$ as $k_{reference}$, which would be used as the reference k in Step 3, whose accuracy is assured by the definition of the tolerance coefficient defined in Step 1.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

```
1. $m_{left} := m_{minimum} \overset{\text{def}}{=} 0$
   $m_{right} := m_{\widehat{critical}}$
   $\delta_k := \alpha_{tc}$
   $m_{previous} := m_{\widehat{critical}}$
   $k_{previous} := k_{reference} \overset{\text{def}}{=} VTP(m_{\widehat{critical}})$
   $m_{current} := \left[\frac{m_{left} + m_{right}}{2}\right]$
   $k_{current} := VTP(m_{current})$
   $\delta_{time} := \frac{3}{4}$

2. while $(m_{left} \le m_{right})$ do 3.–8.

3.     $m_{previous} := m_{current}$
       $k_{previous} := k_{current}$
       $m_{current} := \left[\frac{m_{left} + m_{right}}{2}\right]$
       $k_{current} := VTP(m_{current})$
       $\varepsilon_{time} := \frac{t_{previous} - t_{current}}{t_{previous}}$

4.     if $(\varepsilon_{time} \le \delta_{time})$ then goto 9.

5.     else if $(|\varepsilon_k| > \delta_k)$ then do 6.

6.             $m_{left} := m_{current}$
               $\delta_{time} := \frac{7}{16}$        (Case i)

7.         else do 8.

8.             $m_{right} := m_{current}$
               $\delta_{time} := \frac{3}{4}$        (Case ii)

9. $\widehat{m_{stable}} := m_{current}$
   $\widehat{k_{stable}} := k_{current}$
   $\widehat{H} := 1 + \frac{\widehat{k_{stable}}}{2}$
```

Figure 7. The pseudo-code of the Step 3.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Prepare for the Step 3.2 with k and m, using expressions in line 1, where $VTP(m_\beta)$ is the standard VTP procedure with $m_\beta = m_{maximum}$ and $m_{minimum} = 0$, corresponds to the original time series.

Step 3.2 Running while loop

We represent how much time can be saved in current process by defining ε_{time}, the difference of previous and current processing time over the previous one, as:

$$\varepsilon_{time} \overset{\text{def}}{=} \frac{t_{previous} - t_{current}}{t_{previous}}. \tag{2–16}$$

We define δ_{time} as the ratio of expected time to be saved from the complexity theory (a threshold set for ε_{time} to prune useless branches). Since the complexity of $VTP(\cdot)$ procedure is $O(n^2)$, we can define δ_{time} according to the proportion of its complexity:

$$\delta_{time} = \begin{cases} \dfrac{(m_\beta)^2 - \left(\frac{m_\beta}{2}\right)^2}{(m_\beta)^2} = \dfrac{3}{4} \quad (Case\ i) \\ \dfrac{(m_\beta)^2 - \left(\frac{\frac{m_\beta}{2} + m_\beta}{2}\right)^2}{(m_\beta)^2} = \dfrac{7}{16} \quad (Case\ ii) \end{cases}. \tag{2–17}$$

We represent the difference of current and reference slope over the reference by defining ε_k as:

$$\varepsilon_k \overset{\text{def}}{=} \frac{k_{current} - k_{reference}}{k_{reference}}. \tag{2–18}$$

Then, if $\varepsilon_{time} > \delta_{time}$ (the reduced processing time is more than the time that can be reduced from the complexity theory), update these variables:

$\{k_{current}, k_{previous}, m_{current}, m_{previous}, \varepsilon_k, \varepsilon_{time}\}$.

(Case i) When $|\varepsilon_k|$ is larger than a threshold δ_k according to the tolerance coefficient,

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

i.e., $|\varepsilon_k| > \delta_k = \alpha_{tc}$, indicating that $\widehat{m_{stable}}$ is too small, and a larger one is needed. In our experiment, $\alpha_{tc} := 0.01$ by analysis from BC data (see Assumption 2 in Chapter IV for discussion). Then update variables as in line 6 of pseudo-code. Then repeat the while loop in Step 3.2.

(Case ii) When $|\varepsilon_k| < \delta_k = \alpha_{tc}$, implying a smaller m as a more accurate $\widehat{m_{stable}}$. Then update variables as line 8 of the pseudo-code. Repeat the while loop in Step 3.2.

Then back to outer loop, if $\varepsilon_{time} \leq \delta_{time}$, go to Step 3.3.

Step 3.3 Pin down $\widehat{m_{stable}}$ and $\widehat{k_{stable}}$, and $\widehat{H}$

The H can be estimated from line 9 of pseudo-code, as the most accurate estimation candidate.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Chapter III

EXPERIMENT AND RESULTS

In this chapter, we explain the experiment data traces and show results of experiments. The Bellcore (BC) Ethernet data traces are used for analysis, and the Portland wireless network traffic traces were tested for the presence of S-S property first and were tested to see how effective our proposed method is by the three metrics: time, error rate, and confident interval.

3.1 Experiment setup

The Bellcore (BC) Ethernet data traces are used for analysis, because each of the available sub data sets contains well estimated reference value for the Hurst parameter metrics. Then the Portland wireless data traces for testing our proposed method.

The BC data sets are considered necessary for analyzing usage because its authors provided a comprehensive investigation on the Ethernet Self-Similarity study, though it is a rudimentary research back in early 1990s. We can compare our estimated H with BC data reference value, an accurate enough estimation of H.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

The wireless data sets tests are our main concern. We aim to check whether it follows the assumption that this specific wireless network traffic should be self-similar, given some evidence in other researchers' works [9].

3.1.1 Bellcore Ethernet data sets

Used for analyzing the proposed method, a classical data is the BC traces from Bellcore (BC) Morristown Research and Engineering facility. There are Ethernet LAN traffic data (i.e. size of packet arrivals) with well-estimated reference value of Hurst parameter [3]. This analysis experiment is designed to give a comparison of the S-S property behavior across Ethernet and wireless data to better understand the behavior of the S-S property in the given wireless traffic environment and to show that the proposed method works in both Ethernet and wireless network environment.

BC data traces consists of four separate data traces, which is a portion of the whole BC data traces the authors collected from 1989 to 1992. Each of the four sub-set of traces (code-named as pAug, pOct, OctExt, and OctExt4) contains one million samples of Ethernet packet arrivals. The “p” means the “purple cable” where the packet arrival was captured; “Aug” means the August 1989; “Oct” means October 1989.

Each line of data file contains a floating-point timestamp (representing the time in seconds since the start of a trace) and an integer length (representing the Ethernet data length in bytes). This is a high-accurate one-microsecond timestamp (six places after the decimal point), though some limitation reduce resolution to roughly 10 microseconds. In particular, each of these four contains: 3,142.82 seconds for pAug, 1,759.62 seconds for

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

pOct, 122,797.83 seconds for OctExt, and 75,943.08 seconds for OctExt4.

The mother traces of each those four last for: one day (pAug), one day (pOct), 35 hours (OctExt), and 307 hours (OctExt4). The pAug, pOct, and OctExt are the first one million samples of their mother traces; OctExt4 starts from about 215 hours after the first sample of its mother trace. We define $\widehat{H}_r$ as the reference value of estimated H of one given trace. Only overall $\widehat{H}_r$ of each mother traces is provided in the reference paper. We still use that $\widehat{H}_r$ to compare with our estimation because we assume the downloadable portions of traces reflect the S-S property as precise as their whole mother traces. Our assumption should be valid because most of the mother traces are within the daily cycle, that is, where the self-similarity holds.

3.1.2 Portland wireless data sets

Used for testing the proposed method, a recent data from CRAWDAD is the Portland data traces consisting of six wireless LAN traffic traces of various public places in Portland, Oregon in 2006 and updated in 2009, using a commercial packet analyzer VWave which has a nano-second time resolution [8], [29].

We tested the six the specific wireless network environment to show their potential behaviors of the S-S property. The six different locations include three on-campus and three off-campus places: (1) Portland State University (PSU) Computer Science Department near faculty Offices in Networking Closet, codenamed as “psu_cs”; (2) PSU Library, third floor, codenamed as “library”; (3) PSU Cafeteria, codenamed as “cafe”; (4) Office overlooking “Pioneer Square” from the second floor, codenamed as “pioneer”; (5)

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Urban Grind Coffee, codenamed as “ug”; and (6) Worldcup Coffee at Powell's Books, codenamed as “powells” [29].

One sub data traces of Portland data sets (under folder pdx/vwave/wlan_nano) are with fine time scale (precision at $1\ s$ or fraction of $1\ s$) with packet size as well as other information, where we only use the timestamps and packet sizes.

As to the 802.11 infrastructure of the network type methodology, these traces were collected using a packet analyzer called VeriWave WT20 Appliance, which consists of two 802.11 reference radios, real-time Linux, and two processors. The WT20 provides nano-second resolution timestamps and it logs the time when it began seeing a frame and the time when the frame finished arriving. The packet analyzer listens with two radios simultaneously on the same channel, recording frames to a per-radio, 256 MB, ring-buffer. The WT20’s firmware will discard any frames received with a signal less than -75 dBm, but the rest (Data and Management, but not Control) are logged without any scrubbing. Using a tclsh script, running on a laptop connected to the WT20 (via Ethernet), the contents of this ring-buffer are grabbed from each radio in-turn, every 10 seconds. This data is dumped and then converted on the fly. At the end of a four-hour capture there are 1,440 files which are stitched together using a custom program (because existing tools either contained bugs or did not work with 802.11 traces). Many small files are stitched together time-wise, omitting redundant portions by custom scripts [8], [29].

3.2 Metrics and Results

The metrics of our experiments include the processing time and accuracy measure.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Metrics for analysis (Table 1) and testing results (Table 2) are below with an example plot (Fig. 9). For those traces with reference H value H_r unavailable, ε_H was evaluated by comparing $\widehat{H}$ and H_r' obtained by using $m_{\widehat{critical}}$ as $m_{maximum}$. With precision of $10\,ms$, $t_{proposed}$ was compared with $t_{original}$, roughly estimated using $m_{maximum} = \left\lfloor\frac{n}{2}\right\rfloor$.

3.2.1 Time

We use relative timing to show how quick our proposed method can be applied into real-time estimation of the Hurst parameter in applications. Without conducting real-time environment simulation, we estimate the speed of real-time calculation from the ratio of our processing time and the monitoring time.

There are some notations used in Table 1 and 2:

The N represent the sample population (the number of packet arrivals) from the original data file; the n represents the first n number of packet arrivals that we choose to conduct each experiment (Table 1, 2).

The $\overline{sr}$ for each sub data is the average sampling rate in the unit of second. The BC data has a time precision of 10^{-6} second and the Portland data has nano-second time precision. The time intervals between neighboring packets are not evenly distributed in both BC and Portland data. We conduct the experiment by assuming that the small time intervals differences can be tolerated in the estimation of the S-S property (Hurst parameter).

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

The $t_{proposed}$ is the processing time using the proposed method and $t_{original}$ is the processing time using original VTP method. Both these two values do not include the pre-processing time. The time spent on determining the tolerance coefficient as the pre-processing time. The units of all the timing metrics are second. These two metrics can be put together to compare with each other because the proposed method outperforms the original VTP by pruning instead of having a lower order complexity function than the original VTP method.

The $\widehat{m_{critical}}$ is the estimated $m_{critical}$ and the $\widehat{m_{stable}}$ is the estimated m_{stable}. The estimation of $m'_{critical}$ is small enough to be ignored in the given data sets.

The $\widehat{H}$ is the estimated Hurst parameter. The error rate is defined as $\varepsilon_H = \frac{\widehat{H}-H_r}{H_r}$, where H_r is the reference value of Hurst parameter.

Our method can process H estimation of one-million samples of data traces around $100\ s$ (except sub-data cafe, Portland) (Table 1, 2), so the precision can be as accurate as $\frac{100\ s}{1{,}000{,}000} = 0.1\ (ms)$ when the sample population is one million.

We use $\frac{t_{proposed}}{t_{original}}$ ratio to denote the relative processing time. The average $\frac{t_{proposed}}{t_{original}}$ ratio is 3.78% for BC data and 6.30% for Portland data.

The results show that, there are on average about 96.22% (BC data) and 94.75% (Portland data) of processing time (excluding pre-processing time) saved from original VTP (Fig. 8).

From above mentioned statistics, we can infer that when the sample population is properly chosen, a sufficiently short processing time for each chunk of data can be

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

achieved to realize the real-time estimation.

3.2.2 Error rate

The error rate ε_H relies on H_r, but different authors have different definition of H_r [17]. We try to choose H_r from data author before generate from $VTP(m_{critical})$ (as for Portland data). H_r values of BC data are given but our data are only portions of the original data sets, which may not reflect precisely the S-S property of the whole data.

The average ε_H is 0.46% for BC data and -0.13% for Portland data.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

TABLE I

The results of Bellcore data analysis. The Unit of time measure is second.

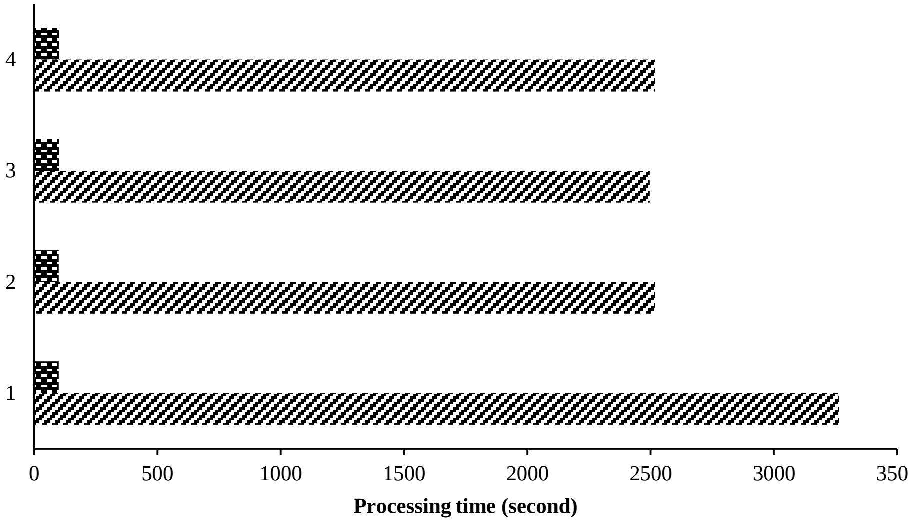
Metrics	**BC Ethernet Data Sets**			
	pAug89	***pOct89***	***Oct89Ext***	***Oct89Ext4***
N	1,000,000	1,000,000	1,000,000	1,000,000
n	1,000,000	1,000,000	1,000,000	1,000,000
$\overline{sr}$	0.003143	0.001760	0.122798	0.075943
$\widehat{m_{critical}}$	10,000	10,000	10,000	10,000
$\widehat{m_{stable}}$	5,000	5,000	5,000	5,000
$\widehat{H}$	0.8140	0.8604	0.9175	0.9758
ε_H	1.75%	−1.89%	1.56%	0.42%
$t_{original}$	3263.51	2517.49	2496.01	2517.85
$t_{proposed}$	99.79	100.25	100.98	101.21

TABLE II

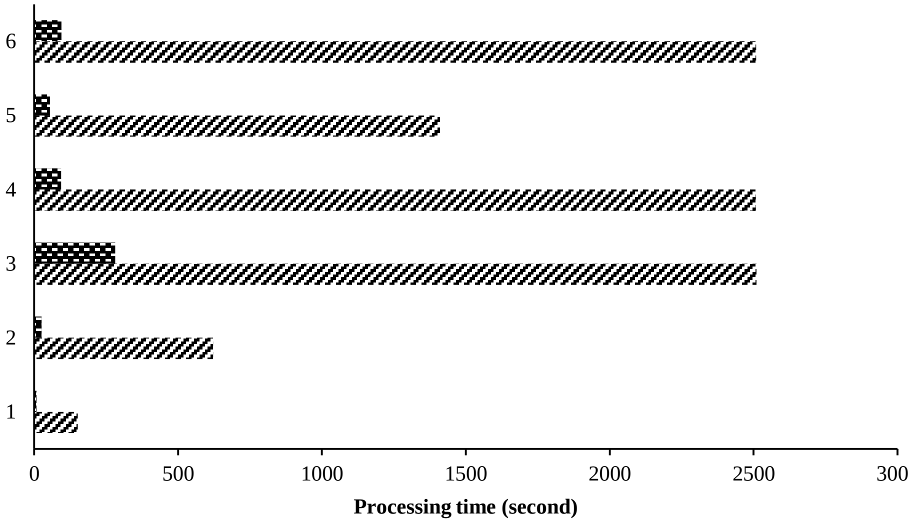
The results of Portland data tests. The unit of time measure is second. Sample sizes are rounded to convenient numbers.

Metrics	**Portland Wireless Data Sets**					
	psu_cs	***pioneer***	***cafe***	***library***	***ug***	***powells***
N	260,326	517.632	3,601,343	1,300,671	895,721	1,398,222
n	250,000	500,000	1,000,000	1,000,000	750,000	1,000,000
$\overline{sr}$	0.013131	0.027783	0.008853	0.009150	0.010156	0.012615
$\widehat{m_{critical}}$	2,500	5,000	10,000	10,000	7,500	10,000
$\widehat{m_{stable}}$	1,250	2,500	5,000	5,000	3,750	5,000
$\widehat{H}$	0.9425	0.9616	0.9174	0.9839	0.8846	0.9350
ε_H	0.01%	0.43%	−0.35%	−0.35%	0.14%	−0.39%
$t_{original}$	151.01	621.90	2510.15	2507.21	1410.22	2508.98
$t_{proposed}$	7.39	25.51	281.02	92.9590	54.34	94.07

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교



(a)



(b)

Figure 8. Processing time comparison chart. Slash shading: using original VTP method; brick shading: using proposed method. (a) Bellcore data experiments: label 1 for pAug89, 2 for pOct89, 3 for Oct89Ext, and 4 for Oct89Ext4. On average 96.22% of time is saved from original; (b) Portland data experiments: label 1 for psu_cs, 2 for pioneer, 3 for café, 4 for library, 5 for ug, and 6 for powells. On average 94.75% of time is saved from original.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

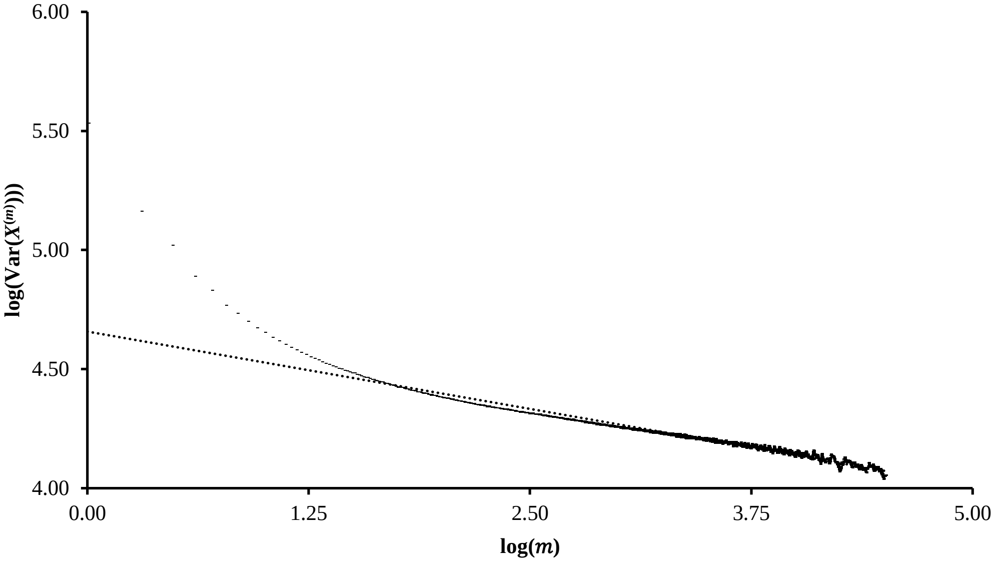


Figure 9. The Variance-Time Plot for Portland powells sub data traces. Variance based on samples of aggregated series of packet size $\mathrm{Var}\left(X^{(m)}\right)$ (y) over aggregation index m (x) with trend line of first million samples. Regression line (light dotted line) equation : $\log(y) = 4.6584 - 0.13\log(x),\ r^2 = 0.9442.$ $m \in [2,\ 32{,}000]$. The plot fluctuates heavily for $m > 10{,}000$ or $\log(m) > 4.0$.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

3.2.3 Confidence intervals

Other methods, such as MLE methods, the periodogram or Whittle, have to be combined with VTP to achieve the 95%-CI metrics [3]. We choose the maximum and minimum values of the $\frac{t_{proposed}}{t_{original}}$ and ε_H as the our "confidence intervals (CI)" metrics, because the classic 95%-CI metrics is not available for our experiment analysis of H due to the few number of data traces at hand.

The CI of ε_H for BC data is $[-1.89\%, 1.75\%]$ and for Portland data is $[-0.39\%, 0.43\%]$. While a reference 95%-CI of ε_H for BC data is at most about 14% [3]. The CI of $\frac{t_{proposed}}{t_{original}}$ is $[3.06\%, 4.05\%]$ for BC data and $[3.71\%, 11.20\%]$ for Portland data.

Chapter IV

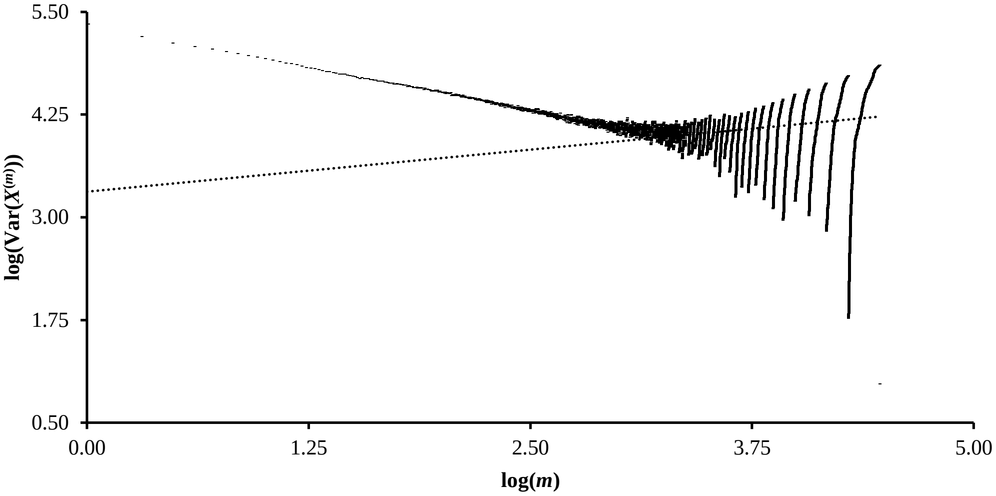
DISCUSSION

4.1 Choosing $m_{maximum}$

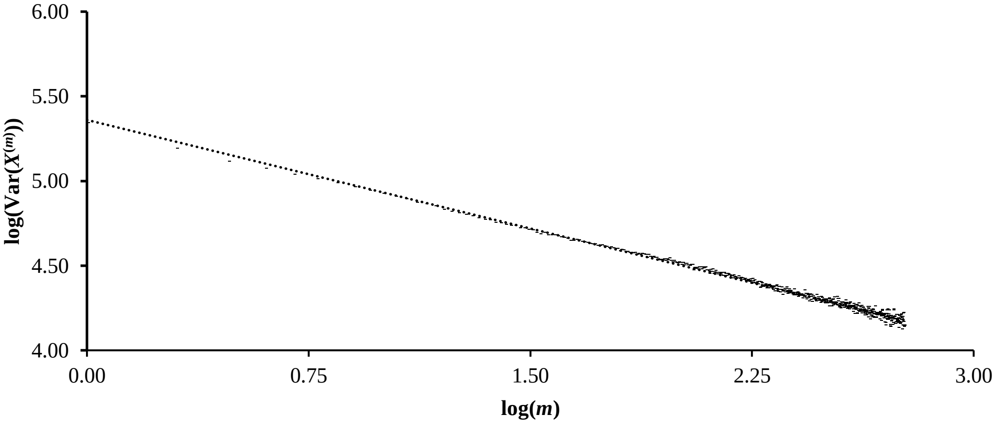
When $m_{maximum} = \left\lceil \frac{n}{m} \right\rceil$, there would be more significant fluctuation as $m \rightarrow m_{maximum}$ than using $m_{maximum} = \left\lfloor \frac{n}{m} \right\rfloor$. Because the $X^{(m)}_{\left\lceil \frac{n}{m} \right\rceil}$ usually contains less terms of X than all the preceding terms, causing the number of terms of $X^{(m)}_{\left\lceil \frac{n}{m} \right\rceil}$, periodically varies from 1 to m and consequential inconsistence in the plot (Fig. 10).

When $m_{maximum} = \left\lfloor \frac{n}{m} \right\rfloor$, the fluctuation of plot occurs when $m \rightarrow m_{maximum}$, extreme cases being only three or two terms of the aggregated series $X^{(m)}$. But compared with the $\left\lceil \frac{n}{m} \right\rceil$, using floor is more stable and basically there is no third way to avoid the errors no matter floor or ceiling. This is one of the reason that $m_{critical}$ was introduced.

For $n = 1{,}000{,}000$, assume that $\frac{n}{100} = 10{,}000$ would place a significant error on the plot (Assumption 2). Thus $m_{maximum} \leq \frac{n}{100} = 10{,}000$ keeps the error as tolerable as we assume, which further limit the domain of m, also implying the saving of processing time.

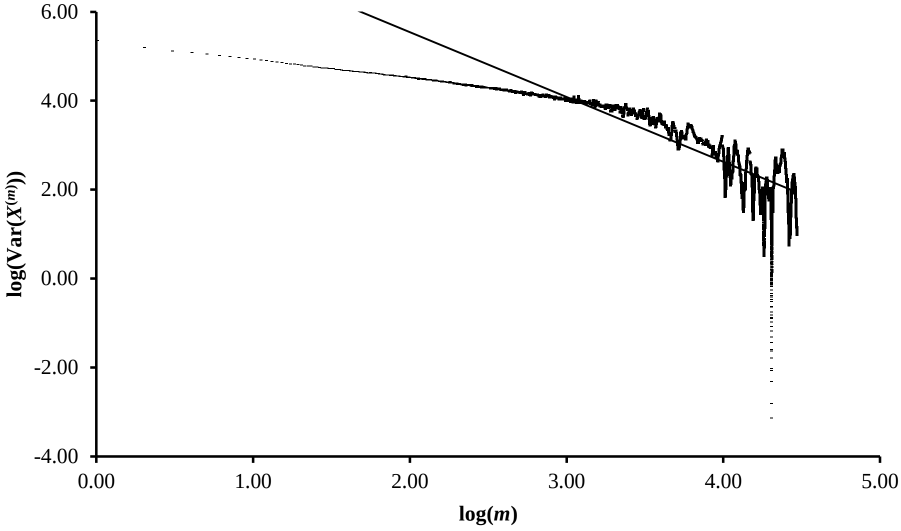


(a)

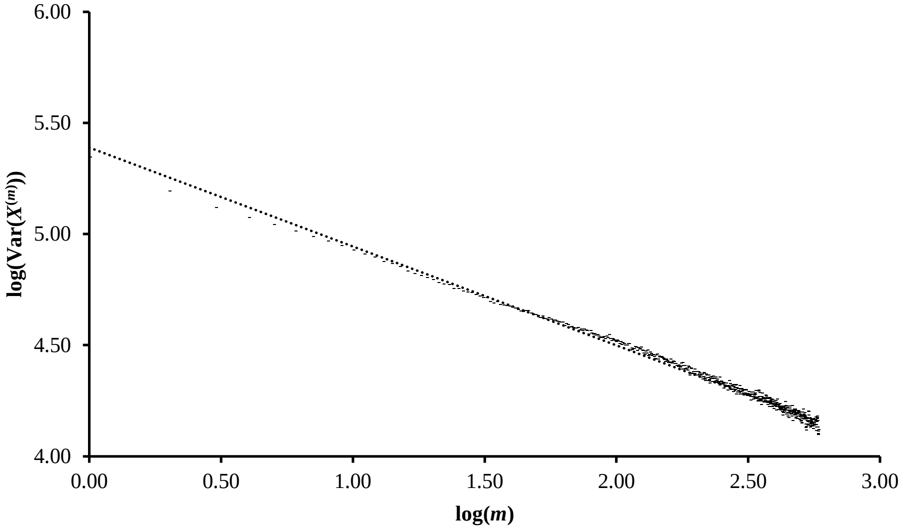


(b)

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교



(c)



(d)

Figure 10. Errors occur when $m_{maximum} = \left\lceil \frac{n}{m} \right\rceil$. The Bellcore data subset pAug89 were

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

the testing data. Variance based on samples of aggregated series of packet size $\mathrm{Var}\left(X^{(m)}\right)$ (y) over aggregation index m (x) with trend line of first $58{,}000$ samples. (a) Trend-line (light dotted line) equation $y = 0.2039x + 3.3139$, $r^2 = 0.0315$, which is apparently not a valid trend-line because of extreme low r^2. The fluctuation occurs around $\log_{10}(m) \approx 2.90$ or $m \approx \frac{n}{100}$. (b) Plot cut by $m = \frac{n}{100} = 580$, with $y = -0.4268x + 5.3594$, $r^2 = 0.9926$, $\widehat{H} = 0.7866$. (c) Plot (floor) with $m_{maximum} = \frac{n}{2} = 29{,}000$, apparent error occurs (d) Plot (floor) cut by $m = \frac{n}{100} = 580$, $y = -0.4454x + 5.3897$, $r^2 = 0.9920$, $\widehat{H} = 0.7773$.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

4.2 Estimation of $m_{critical}$

In Step 1 of the Algorithm in Chapter II, the m_{stable} is defined from the training experiments conducted before the tests, by which the estimation time of the trend slope is shortened because plenty of repeated computation is saved.

The m_{stable} is one of the stable m that lies in the early part of the well-fitting linear trend line.

In Step 2 of the Algorithm in Chapter III, for each new value of m, the generation of aggregated series $X^{(m)}$ is one of the main cause of the computation complexity of $O(n^2)$. While some author suggested using limited memory data to shorten the processing time [17], we suggest that the effort may alternatively concentrated on how to keep m from getting larger than enough.

The one-million samples of packet size, may not be necessarily lasting $3{,}200$ seconds long for other testing data, but this one-million large samples guarantee the well observation of S-S property in the data traces, as suggested from the experiment results of $m_{critical}$ as well as the mathematical definition of LRD.

4.3 Choosing δ_{time}

An example of δ_{time} is assigned using the complexity upper bound value in Equation (2–16). Some other choices of δ_{time} can be chosen according to specific applications requirements and set in the pre-processing phase.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

4.4 Validity of result metrics

The exact direct comparison of accuracy and effectiveness between another author's work with our proposed work is not available [17], because difference testing environment and the different $m_{maximum}$ also result in the incomparability.

4.5 Validity of $\widehat{H}$ using $\widehat{m_{stable}}$

The linear regression should be based on the following Assumption 1 [30], while in practice the original VTP does not hold rigorously for these two conditions.

ASSUMPTION 1. (1) *The underlying relationship is linear*; (2) *The dots in the original scatter-plot will be dispersed equally about all segments of the regression line* [30].

Our experiment results showed that the log-log scale does not impose a dramatically different type of dot cluster nor an obvious curvilinear relationship within the interval $[m'_{critical}, m_{critical}]$, where $m'_{critical}$ is the counterpart of $m_{critical}$ near $m_{minimum}$, commonly referred to as a turning point of log-log plot. Thus the Assumption 1 is considered as not violated. Also there is no minimum requirement that how large the sample pool yields an accurate-enough regression. By the above reasoning, we claim the validity of $\widehat{H}$ using $\widehat{m_{stable}}$ somewhere in between $m'_{critical}$ and $m_{critical}$.

4.6 Reasons of curve fluctuation

It is a limitation of this paper since the conjecture is showed by the experiment instead of theoretical deduction.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Possible causes of the fluctuation when $m > m_{critical}$ are listed in following sections.

4.6.1 Defect in a Variance-Time Plot realization

ASSUMPTION 2. *If the difference between estimated parameter and the real one is within one-hundredth of the real one, the error can be tolerated.*

Assumption 2 yields the tolerance coefficient mentioned in Chapter III. When, e.g.,$n = 1{,}000{,}000$, and $m = 10{,}001$ or $\log(m) = 4.00$, there will be as many as $9{,}901$ tail samples of series *X* being discarded in computing $\log\left(\mathrm{Var}\left(X^{(10{,}001)}\right)\right)$, which is 0.99% of the population, too significant a portion to be tolerated according to Assumption 2. Then $m_{maximum}$ is reasonably confined to $\frac{n}{100}$.

4.6.2 Mixed properties for the data traces

Internet traffic may have both the Poissonity (SRD-like) and S-S (LRD-like) properties [26]. In particular, the trend of log-log plot of experiments for $m \in (m_{critical}, m_{maximum})$ is apparently reaching the horizontal axis faster than the segment for $m \in (1, m_{critical})$, regardless of errors from the inherent defect of the aggregation realization. Another $m'_{critical}$ that is closer to $m_{minimum}$ can be found out using a modified algorithm in Chapter III and define a well-fitting segment of trend line on the interval of $[m'_{critical}, m_{critical}]$.

Together with the curvilinear segment for $m \in (1, m'_{critical})$, a non-linear fitting curve (e.g. a skewed third-order polynomial or exponential curve) may fit more accurately

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

instead of a linear one. Then the parameters of the curve, or of VTP, may be related to $\hat{k}$, $\widehat{m_{stable}}$ and $\widehat{m_{critical}}$ in some way. This alternative model of the plot is left to future work.

4.7 Non-linear representation of variance-time plot

We examined polynomial (second and third degree) and exponential trend-lines on top of the linear trend-line. Results concentrated on accuracy instead of processing time, because the linear trend-line plotting can be the most efficient one while keeping the accuracy level.

4.7.1 Polynomial trend-line

Polynomial trend-lines of second and third degree have been examined to possibly provide alternative methods of estimating H (Fig. 11, 12). The counterpart for the slope in linear trend-line method is now a derivative of the polynomial curve somewhere stable. We adopted the m_{stable} to represent that stable position of $\log(m)$ axis, using the derivative at the point of m_{stable} to estimate H.

Results showed that both the second and third degree polynomial trend-line could match the variance-time plot well enough, though the third degree one did even more accurate (Fig. 11, 12). Forth and above degrees are not adopted in the experiments because it is appropriate to remain a concise set of parameters, that is, the coefficients in the polynomial equations.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

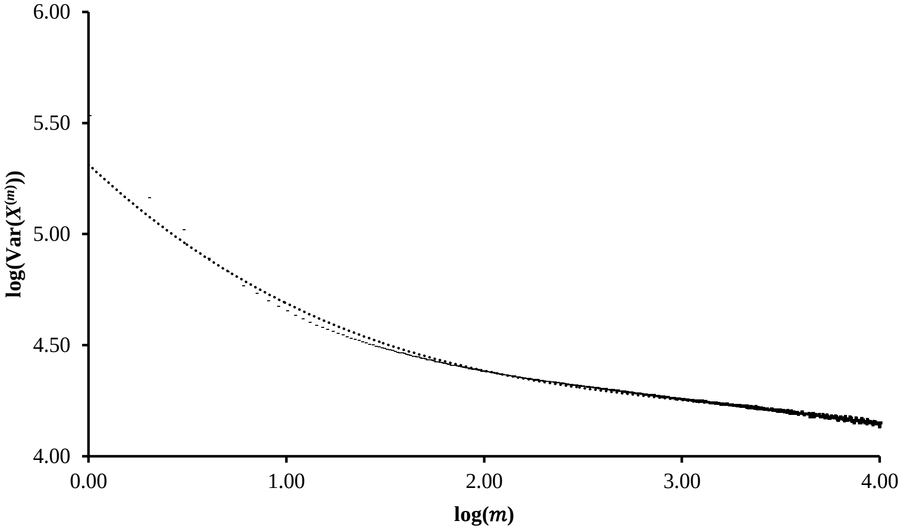


Figure 11. Polynomial third degree trend-line test (Portland powells sub data trace). Variance based on samples of aggregated series of packet size $\mathrm{Var}\left(X^{(m)}\right)$ (y) over aggregation index m (x) with trend line of first million samples. Trend-line (light dotted line) equation $y = -0.0251x^3 + 0.236x^2 - 0.8353x + 5.313,\ r^2 = 0.986,\ y' = -0.0753x^2 + 0.4720x - 0.8353.$ $\widehat{H} = 0.9402$ by using the m_{stable} in the derivative curve of the trend-line.

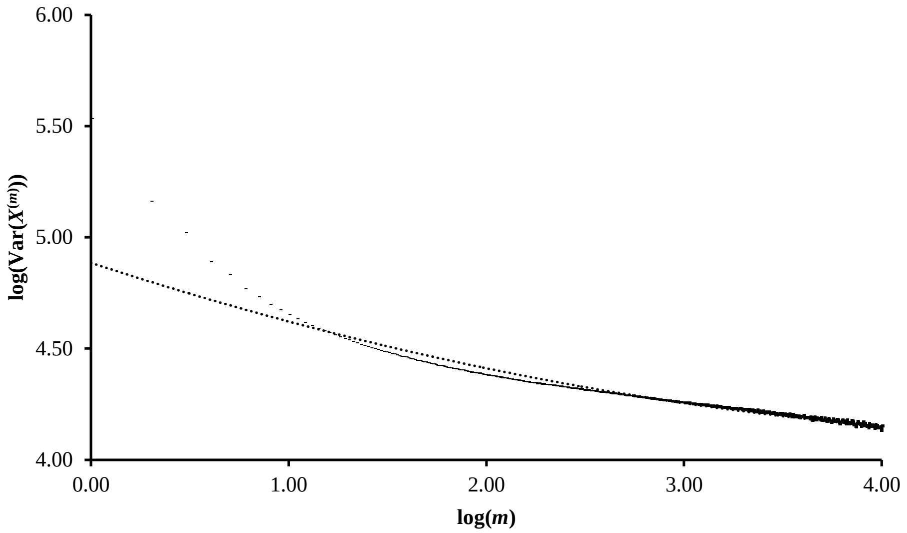


Figure 12. Polynomial second degree trend-line test (Portland powells sub data trace). Variance based on samples of aggregated series of packet size $\mathrm{Var}\left(X^{(m)}\right)$ (y) over aggregation index m (x) with trend line of first million samples. Trend-line (line dotted line) equation: $y = 0.0267x^2 - 0.2902x + 4.8852,\ r^2 = 0.956.$ $y' = 0.0534x - 0.2902.$ $\widehat{H} = 0.9537$ by using the m_{stable} in the derivative curve of the trend-line.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

4.7.2 Exponential trend-line

Exponential trend-line method uses the form of:

$$y = ae^{bx}$$

where y is the variance in logarithmic scale, x is the aggregation index m in logarithmic scale, a is a positive constant and b is a negative constant.

Results showed that the trend-line did not fit as tight as the polynomial ones, though the accuracy was still acceptable (Fig. 11).

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

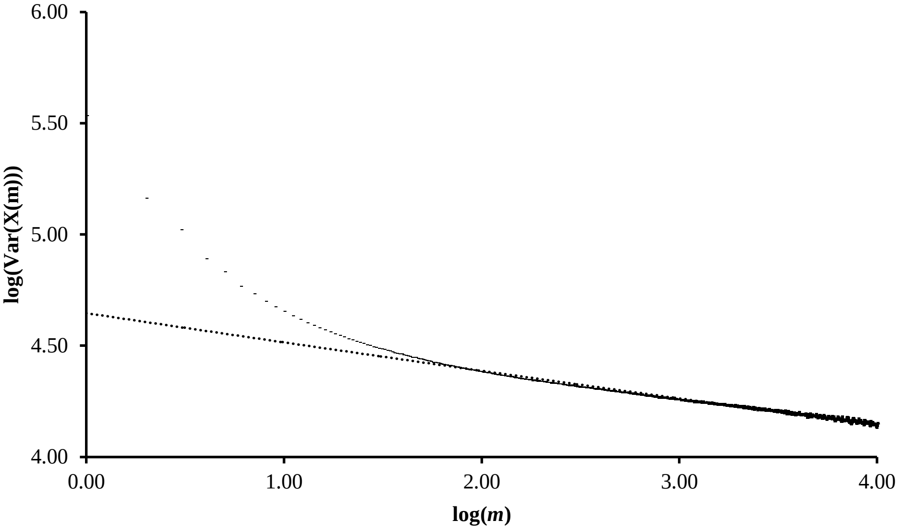


Figure 13. Exponential trend-line for test (Portland powells sub data trace). Variance based on samples of aggregated series of packet size $\mathrm{Var}\left(X^{(m)}\right)$ (y) over aggregation index m (x) with trend line of first million samples. Trend-line (line dotted line) equation: $y = 4.6463e^{-0.029x}$, $r^2 = 0.9413$. $y' = -0.1347e^{-0.029x}$. $\widehat{H} = 0.9395$ by using the m_{stable} in the derivative curve of the trend-line.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Chapter V

CONCLUSION

For some recent 802.11 wireless network traffic traces in specific environment, our VTP-based real-time Hurst parameter estimation method presents a much faster and accurate enough estimation, according to experimental results. The S-S property of testing 802.11 wireless data traces has been properly proven to be existing. It can be represented by the three aggregation level values ($m'_{critical}$, $m_{critical}$ and m_{stable}) with the slope of the fitting trend line of the middle segment for the VTP log-log plot. This proposed representation is an alternative of the one-parameter (H) representation. The S-S property may be also represented by polynomial trend-line coefficients with m_{stable}. Experimental results showed that the processing time can be reduced by more than 90% in relative comparison from the original VTP method while the accuracy remained in accepted range. The proposed method could generate results quick enough for some applications that require real-time estimation of the S-S property parameter.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

REFERENCES

[1] B. B. Mandelbrot and J. W. Van Ness, “Fractional Brownian motions, fractional noises and applications”, *SIAM Review*, vol. 10, no. 4, 1968, pp. 422–437.

[2] W.-B. Gong, Y. Liu, V. Misra, and D. Towsley, “Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications”, *Computer Networks*, vol. 48, no. 3, 2005, pp. 377–399.

[3] 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*, vol. 2, no. 1, 1994, pp. 1−15.

[4] M. E. Crovella and A. Bestavros, “Explaining World Wide Web traffic self-similarity”, Boston University, Boston, MA, 1995.

[5] M.E. Crovella and A. Bestavros, “Self-similarity in World Wide Web traffic evidence and possible causes”, *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, 1997, pp. 835−846.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

[6] T. F. Bento and P. S. Barreto, “Evaluating the impact of management traffic and self-similarity on wireless network performance”, *Proceedings of 2011 10th IEEE International Symposium on Network Computing and Applications*, 2011, pp. 187–194.

[7] Q. Yu, Y. Mao, T. Wang, and F. Wu, “Hurst parameter estimation and characteristics analysis of aggregate wireless LAN traffic”, *Proceedings of 2005 International Conference on Communications, Circuits and Systems*, 2005, pp. 339–345.

[8] C. Phillips and S. Singh, “Analysis of WLAN traffic in the wild”, *Proceedings of the 6th international IFIP-TC6 conference on Ad Hoc and sensor networks, wireless networks, next generation internet*, 2007, 1173–1178.

[9] L. Xiang, X. Ge, K. Zhang, and C. Liu, “A self-similarity frame traffic model based on the frame components in 802.11 networks”, *Proceedings of 2009 International Conference on Computational Science and Engineering*, 2009, pp. 955–960.

[10] H. Toral-Cruz, A.-S. K. Pathan, and J. C. R. Pacheco, “Accurate modeling of VoIP traffic QoS parameters in current and future networks with multifractal and Markov models”, *Mathematical and Computer Modelling*, vol. 55, 2012, in press.

[11] R. B. Ali and S. Pierre, “UMTS-to-IP QoS mapping for voice and video telephony services”, *IEEE Network*, vol. 19, no. 2, 2005, pp.26–32.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

[12] J. Kang, Y.-Z. Song, and J.-Y. Zhang, “Accurate detection of peer-to-peer Botnet using multi-stream fused scheme”, *Journal of Networks*, vol. 6, no. 5, 2011, pp. 807–814.

[13] G. Liu, M. Hu, B. Fang, and H. Zhang, “Explaining BitTorrent traffic self-similarity”, *Proceedings of 5th International Conference of Parallel and Distributed Computing: Applications and Technologies*, 2004, pp. 839–843.

[14] J. Wang and G. Yang, “An intelligent method for real-time detection of DDoS attack based on fuzzy logic”, *Journal of Electronics (China)*, vol. 25, no. 4, 2008, pp. 511–518.

[15] M. Roughan, D. Veitch, and P. Abry, “Real-time estimation of the parameters of Long-Range Dependence”, *IEEE/ACM Transactions on Networking*, vol. 8, no. 4, 2000, pp. 467–478.

[16] E. Hernández-Orallo and J. Vila-Carbó, “Network queue and loss analysis using histogram-based traffic models”, *Computer Communications*, vol. 33, no. 2, 2010, pp. 190–201.

[17] T. Hagiwara, H. Doi, H. Tode, and H. Ikeda, “High-speed calculation method of the Hurst parameter based on real traffic”, *Proceedings of 25th Annual IEEE Conference*, 2000, pp. 662–669.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

[18] M. Hajduczenia, H. J. A. da Silva, and P. P. Monteiro, “Real-time Hurst parameter estimation based on modified embedded branching process”, *Proceedings of 2007 6th Conference of Telecommunications*, 2007.

[19] R. G. Clegg, “A practical guide to measuring the Hurst parameter”, *CS-TR No 916 21st UK Performance Engineering Workshop*, Newcastle: Newcastle University Press, 2005, pp. 43–55.

[20] Y. Xiang, W. L. Lei, and S. J. Huang, “Detecting DDoS attack based on network self-similarity”, *IEEE Proceedings of Communications*, vol. 151, no. 3, 2004, pp. 292–295.

[21] D. Guo, X. Wang, and J. Zhang, “Fast real-time Hurst parameter estimation via adaptive wavelet lifting”, *IEEE Transaction on Vehicular Technology*, vol. 53, no. 4, 2004, pp. 1266–1273.

[22] O. D. Jones, Y. Shen, “Estimating the Hurst index of a self-similar process via the crossing tree”, *IEEE Signal Processing Letters*, vol.11, no.4, 2004, pp. 416–419.

[23] K. S. Trivedi, Probability and Statistics with Reliability, Queuing and Computer Science Applications, 2nd ed., New York: John Wiley & Sons, 2002.

[24] R. D. Yates, D. J. Goodman, Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers, 2nd ed., New Jersey: John Wiley & Sons, 2005.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

[25] M. Gospodinov and E. Gospodinova, “The graphical methods for estimating Hurst parameter of self-similar network traffic”, *International Conference on Computer Systems and Technologies*, Bulgaria: Varna Technical University Press, 2005.

[26] H. Gupta, A. Mahanti, and V. J. Ribeiro, “Revisiting coexistence of Poissonity and self-similarity in Internet traffic”, *Proceedings of 17th IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems*, 2009, pp. 1−10.

[27] M. Krunz, “On the limitations of the variance-time test for inference of long-range dependence”, *Proceedings of 20th Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM 2001*, vol. 3, 2001, pp. 1254−1260.

[28] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, 3rd ed., Massachusetts: The MIT Press, 2009.

[29] C. Phillips and S. Singh, “CRAWDAD data set pdx/vwave (v. 2009-07-04)”, http://www.crawdad.org/pdx/vwave, 2009.

[30] R. S. Witte and J. S. Witte, Statistics, 9th ed., New Jersey: John Wiley & Sons, 2010.

[31] J.-m. Li and R.-c. Wang, “Real-time detection of DDoS attack based on VTP”, *Acta Electronica Sinica*, vol. 35, no. 4, 2007, pp. 791−796.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

[32] J. L. Wang and A. Erramilli, “A connection admission control algorithm for self-similar traffic”, *Proceedings of 1999 Global Telecommunications Conference*, vol. 2, 1999, pp. 1623−1628.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Acknowledgments

李素碧 and 劉長江, thank you for bringing me to this world and giving without much complaint.

서영주 교수님, 유찬수 교수님 and 홍원기 교수님, thank you for giving me the opportunity to learn with you. Professor Raouf Boutaba, thank you for your teaching on this topic. 이근배 교수님, thank you for accepting me as POSTECH student.

Here goes a personal list; trying to keep it concise and chronic reverse. Some may be missed for the limitation of my memory, but you know I own you one.

青青, thank you for listening. 태열, 홍석, 승호, 승걸, 주환, and many others, thank you for taking care of all those stuff. 東昱 先輩, 楨潤 先輩, 敬學 先輩, 在弼 先輩, 영덕 형, 재국, 우중, 효련, 상욱, 석성, thank you for being tolerant. Estella, Chaysie, Taeyoung, 하연, 비진, 장혜자 선생님, 인태, 민지, and many other teachers, workers, staff members, thank you for help. 미훈 누나, 獸獸, 老趙, 大仙, 健日, 震震,, 春根大

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

叔, 碩碩, and many others, thank you for being friendly. Vinh, Do, Duc, Hien, Huong, Nguyen, Nam, and many other Vienamese friends, thank you for invitation. All my Pakistani and Indian friends, thank you for sharing happiness. Adam, Matt, Alex, Mohsen, James, Indranil, Arun, Sandip, and many other foreign friends, thank you for your time. 규송, thank you for sharing your laugher. 진식 형, thank you for your optimistic encouragement. 현숙 이모, thank you for taking care of the food. 기숙 누나, thank you for the meal. 선희 누나, thank you for showing the way. 성진, 청재, 경덕 형, thank you for being nice. 종훈 선배, thank you for insights. Zöer, thank you for patience. 加林, thank you for honesty. 輝輝, thank you for forgiving. Mickey, thank you for greeting. Jack, thank you for introducing AmE. BoBo, thank you for keeping that daydream alive. くるまだ さん, thank you for introducing drawing. 胖婆婆, thank you for teaching me how to use restroom. 爺爺, thank you for save my life. 婆婆, thank you for enduring me so long. 爺爺, thank you for your love.

*　　　*　　　*

「ユリア、これでいいのだろう」

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Curriculum Vitae

Name	Cong Liu

Education

Sept. 2005—Jul. 2009	School of Automation Engineering, University of Electronic Science and Technology of China (B.S.)
Sept. 2010—Aug. 2012	Computer Science and Engineering, Pohang University of Science and Technology (M.S.)

Experience

Mar. 2012—Jun. 2012	Teaching Assistant, CSED101 Programming Language and Problem Solving, Computer Science and Engineering, Pohang University of Science and Technology

1986