Traffic Modeling(3)

Correlation

- Correlation is a statistical measure of the relationship between two different time series

  i) Positive correlation

  : both behave similarly

  : big observation usually followed by another big, or small by small

Correlation

- Correlation is a statistical measure of the relationship between two different time series

ii) Negative correlation

: behave as opposites

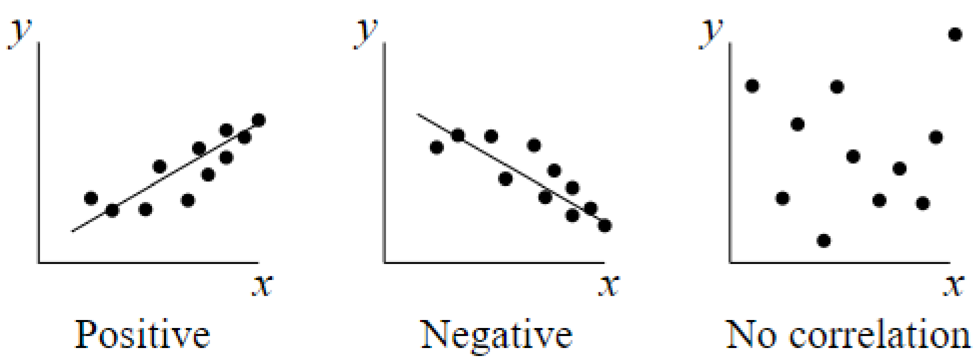
: big observation usually followed by small, or small by big

Correlation

- Correlation is a statistical measure of the relationship between two different time series

iii) No correlation

: behavior of one is unrelated to behavior of other



Autocorrelation

- Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- It's conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.

Autocorrelation

- For example, if it's rainy today, the data suggests that it's more likely to rain tomorrow than if it's clear today.

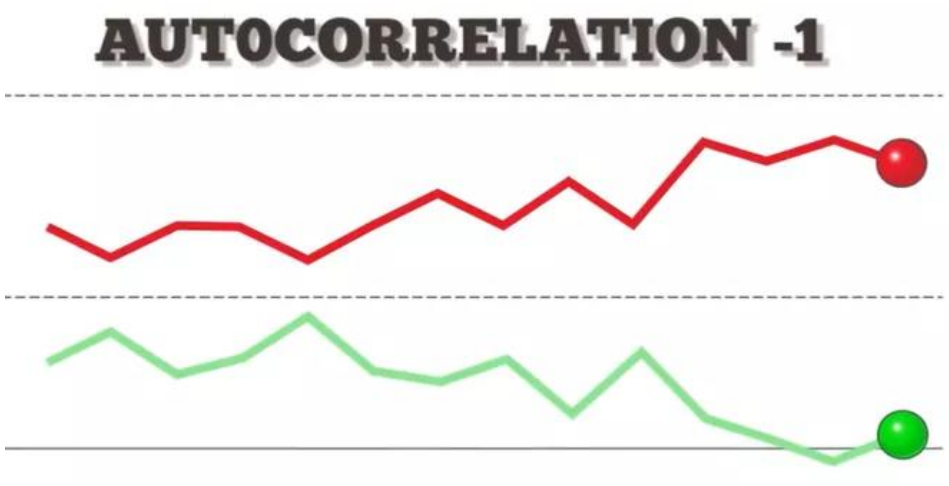
- When it comes to investing, a stock might have a strong positive autocorrelation of returns, suggesting that if it's "up" today, it's more likely to be up tomorrow, too. Technical analysts can use autocorrelation to measure how much influence past prices have on its future price.

Autocorrelation

- An autocorrelation of +1 represents a perfect positive correlation (an increase seen in one time series leads to a proportionate increase in the other time series).

Autocorrelation

- On the other hand, an autocorrelation of -1 represents a perfect negative correlation (an increase seen in one time series results in a proportionate decrease in the other time series).



Autocorrelation

- Autocorrelation coefficient can range between +1 (very high positive correlation) and -1 (very high negative correlation)
- Zero means no correlation
- Autocorrelation function shows the value of the autocorrelation coefficient for different time lags k
- Lack of independence usually results in autocorrelation

Measuring Autocorrelation

- A correlogram (also called Auto Correlation Function ACF Plot or Autocorrelation plot) is a visual way to show serial correlation in data that changes over time (i.e. time series data).

- Autocorrelation plots are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations of data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then the autocorrelations will be non-zero.

Definition of Autocorrelation

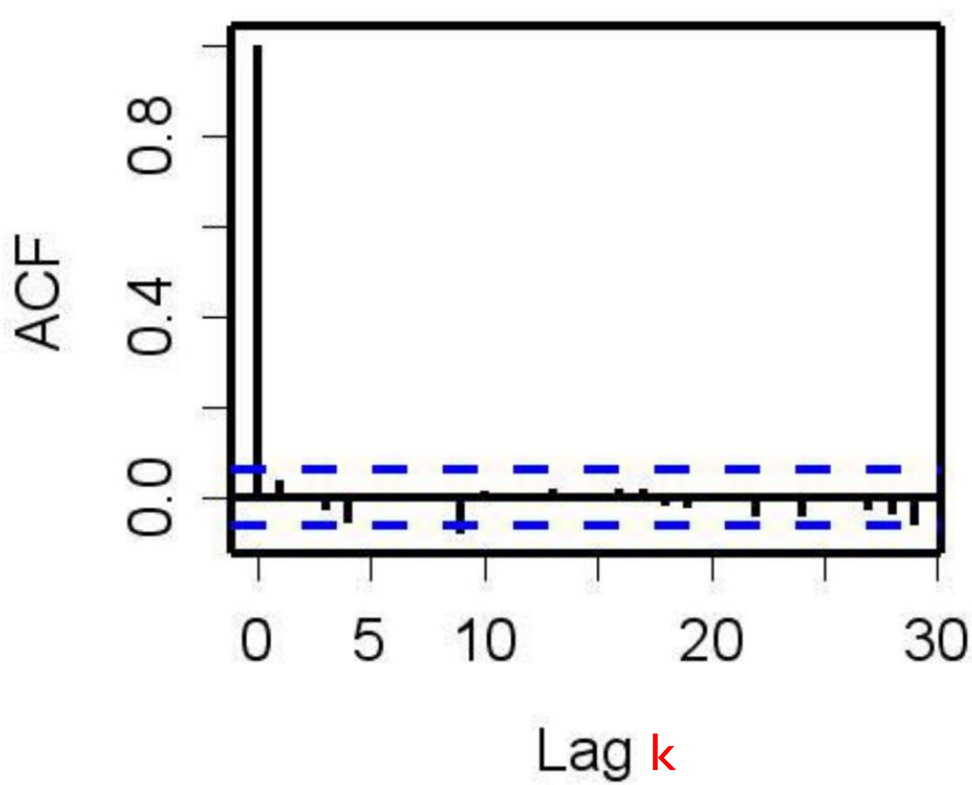
- Autocovariance Function :

$$\mathrm{r}(k) = Cov(\mathrm{X}_t, \mathrm{X}_{t+k}) = \mathrm{E}[\mathrm{X}_t \mathrm{X}_{t+k}] - \mathrm{E}[\mathrm{X}_t] \cdot \mathrm{E}[\mathrm{X}_{t+k}]$$

- Autocorrelation Function (Autocorrelation Coefficient):

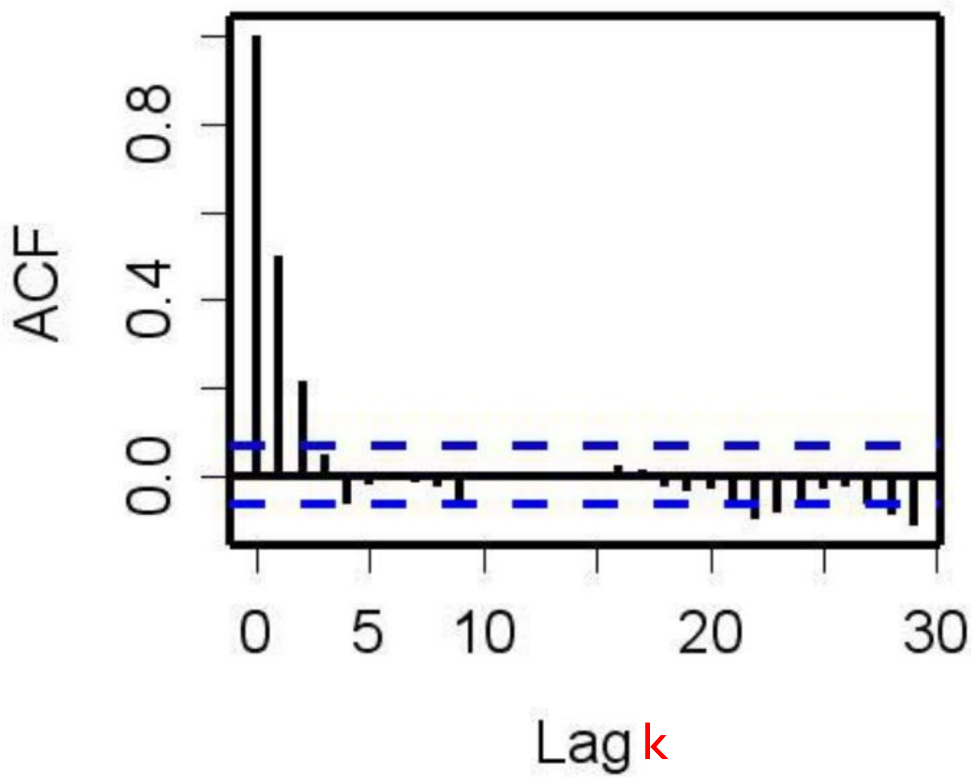
$$\rho(k) = Cor(\mathrm{X}_t, \mathrm{X}_{t+k}) = \frac{Cov(\mathrm{X}_t, \mathrm{X}_{t+k})}{\sqrt{Var(\mathrm{X}_t) Var(\mathrm{X}_{t+k})}}$$

ACF of samples of i.i.d. random variables



Correlogram of samples of i.i.d random variables

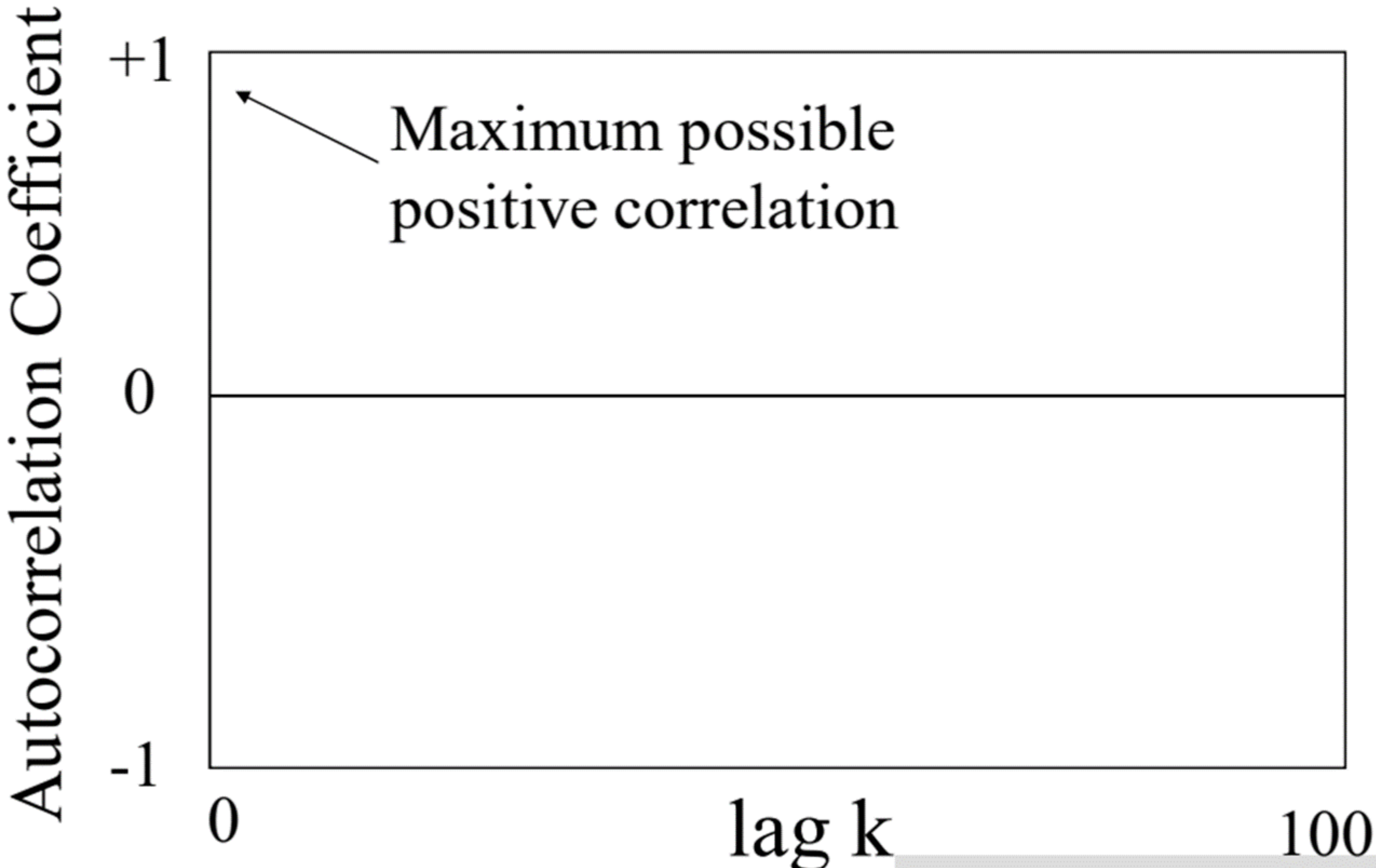
Measuring Autocorrelation



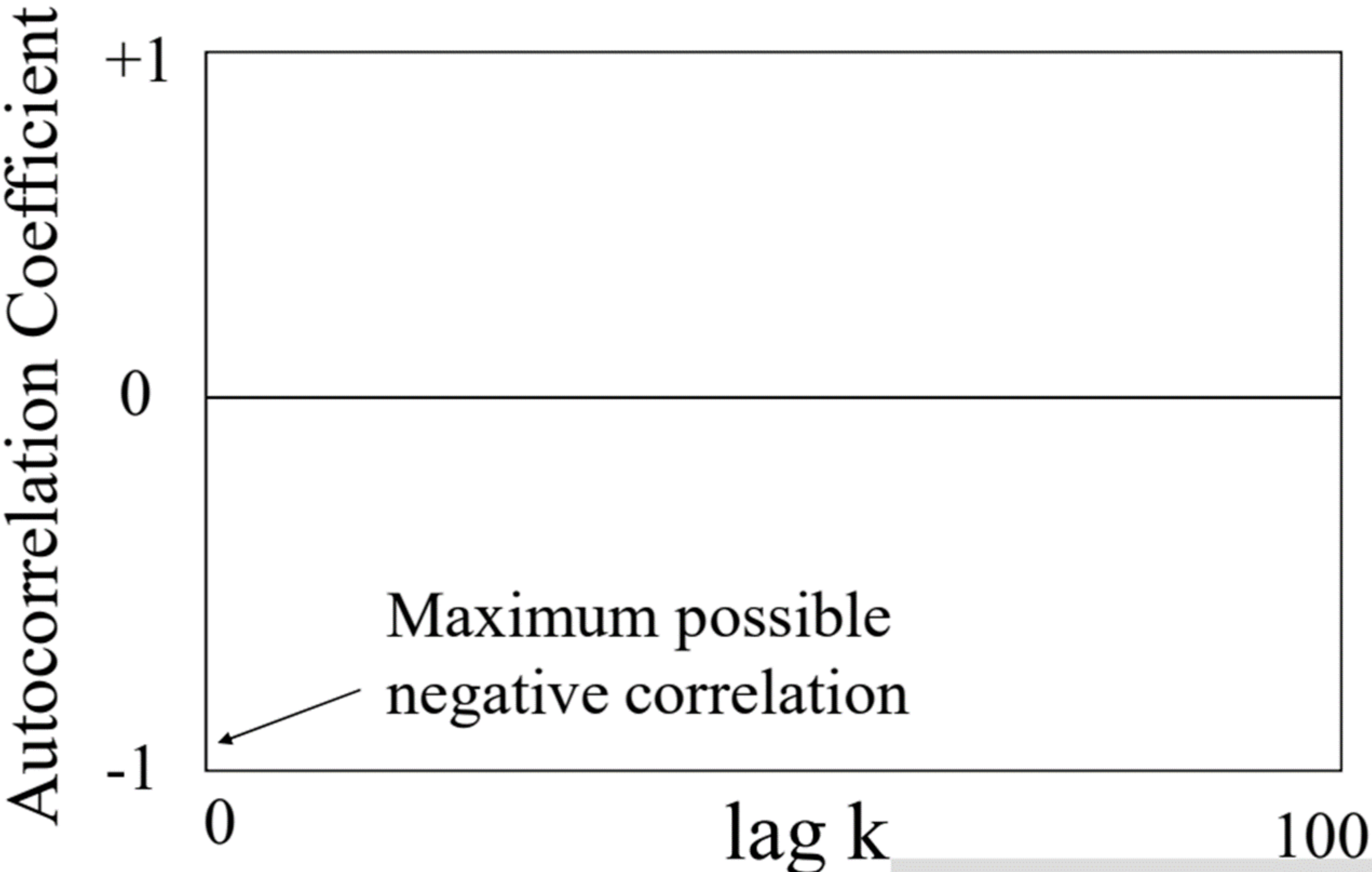
Correlogram of samples of not i.i.d random variables

Autocorrelation

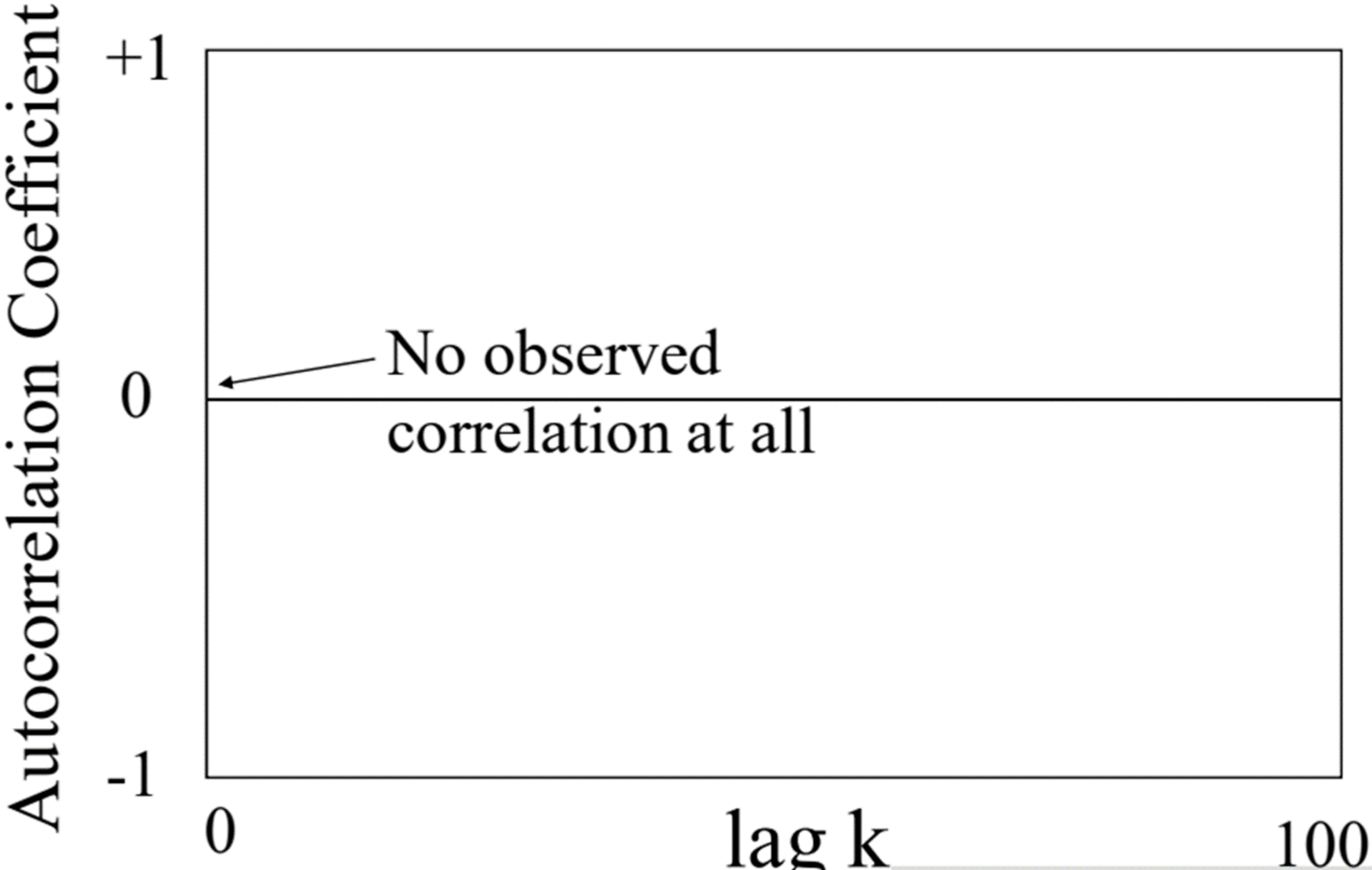
- In fact, if the process consisted of i.i.d. RVs, we would be done.
- However, most traffic has the property that its measurements are not independent.
- Lack of independence usually results in autocorrelation



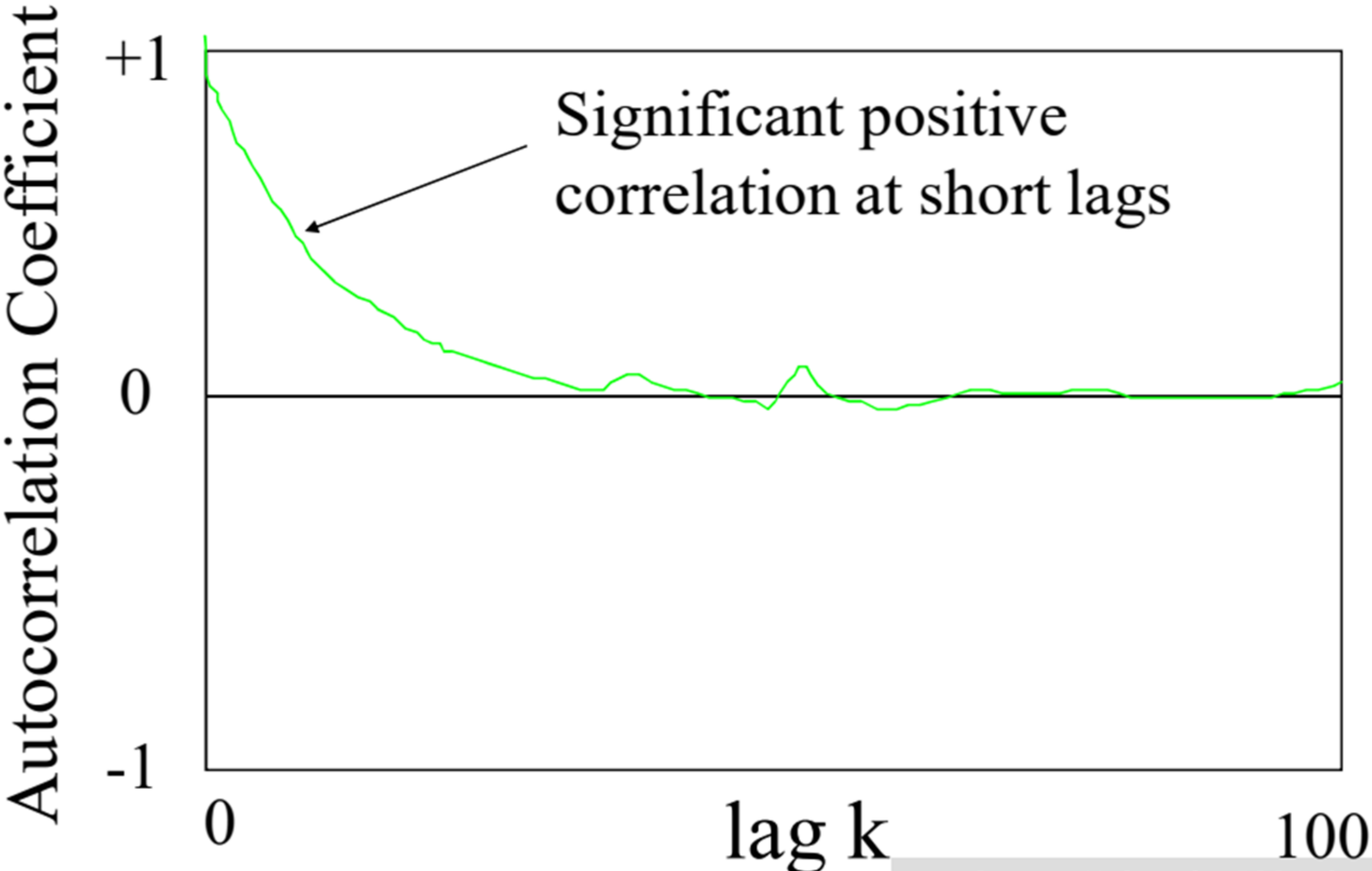
Long Range Dependence: Autocorrelation



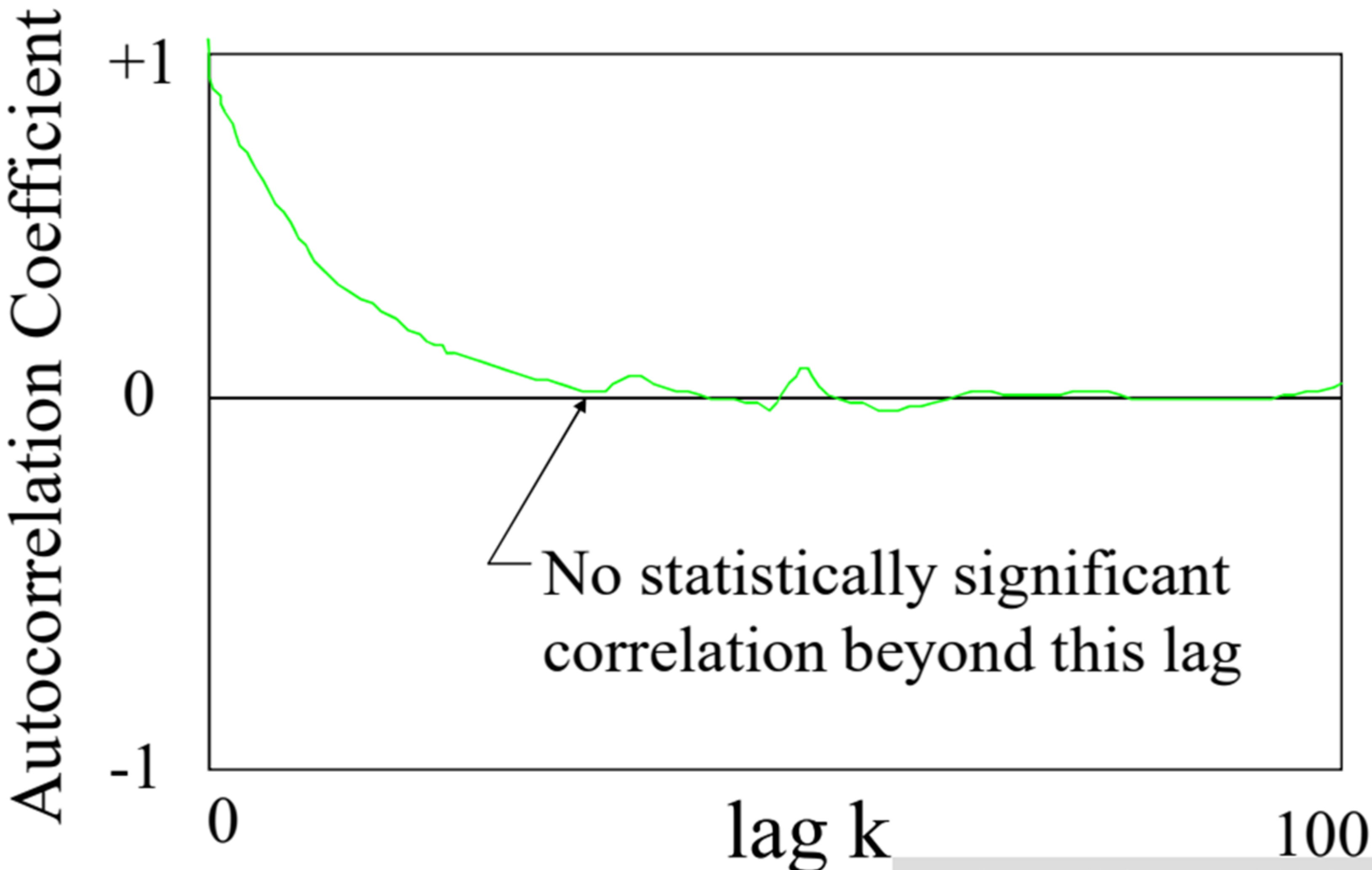
Long Range Dependence: Autocorrelation



Long Range Dependence: Autocorrelation



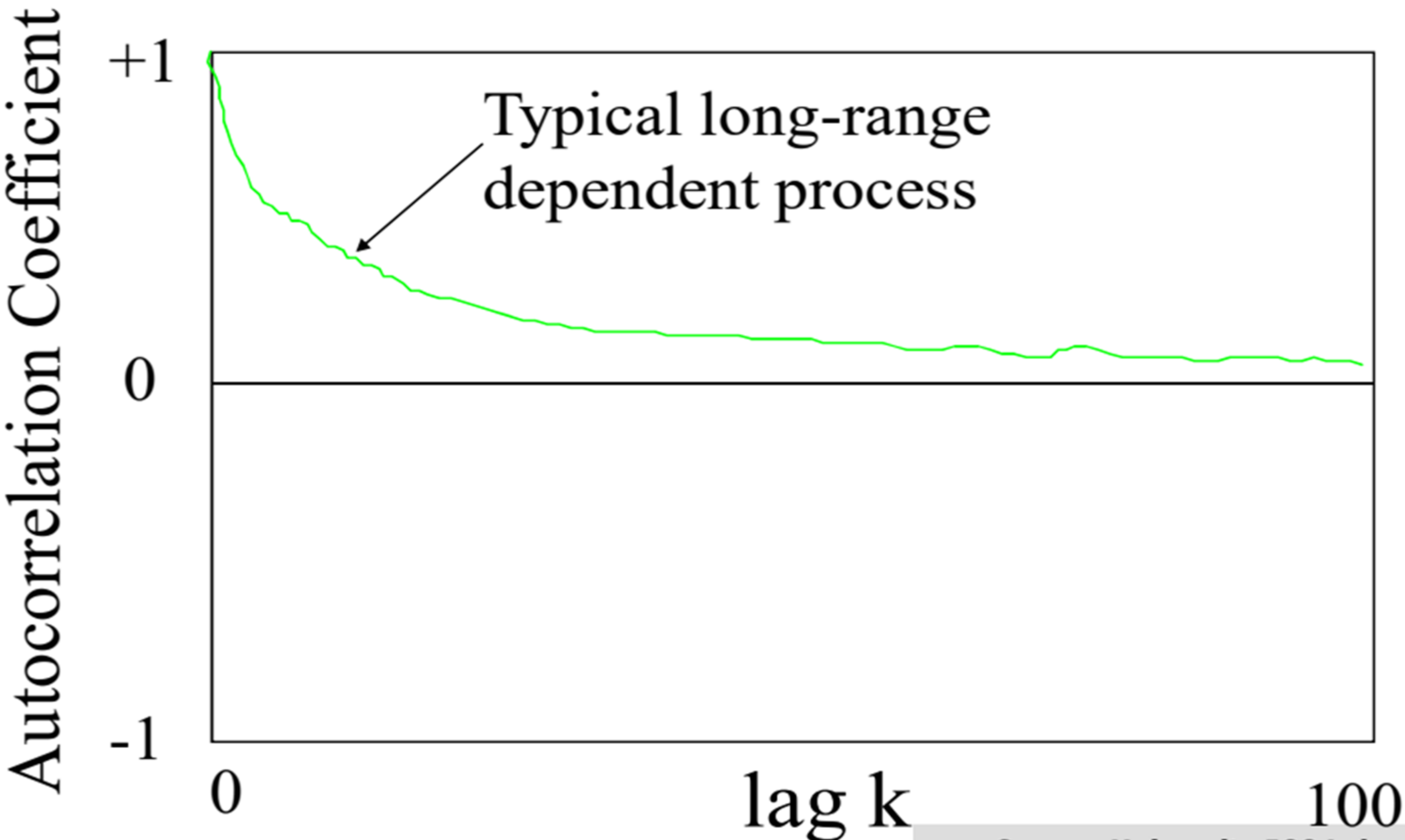
Long Range Dependence: Autocorrelation



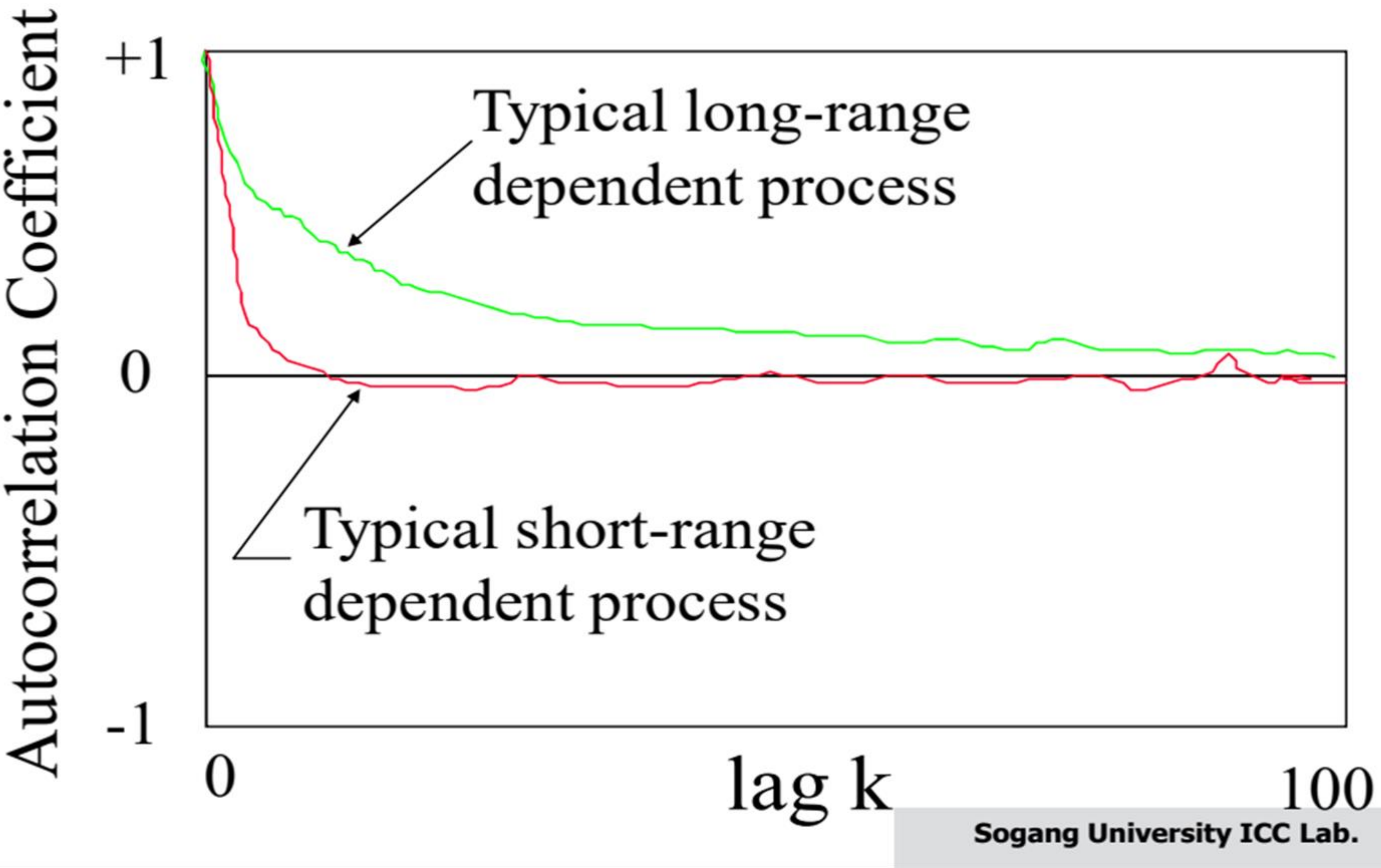
Long Range Dependence: Autocorrelation

Autocorrelation

- For most processes (e.g., Poisson, or compound Poisson), the autocorrelation function drops to zero very quickly (usually immediately, or exponentially fast)
- For self-similar processes, the autocorrelation function drops very slowly (i.e., hyperbolically) toward zero, but may never reach zero



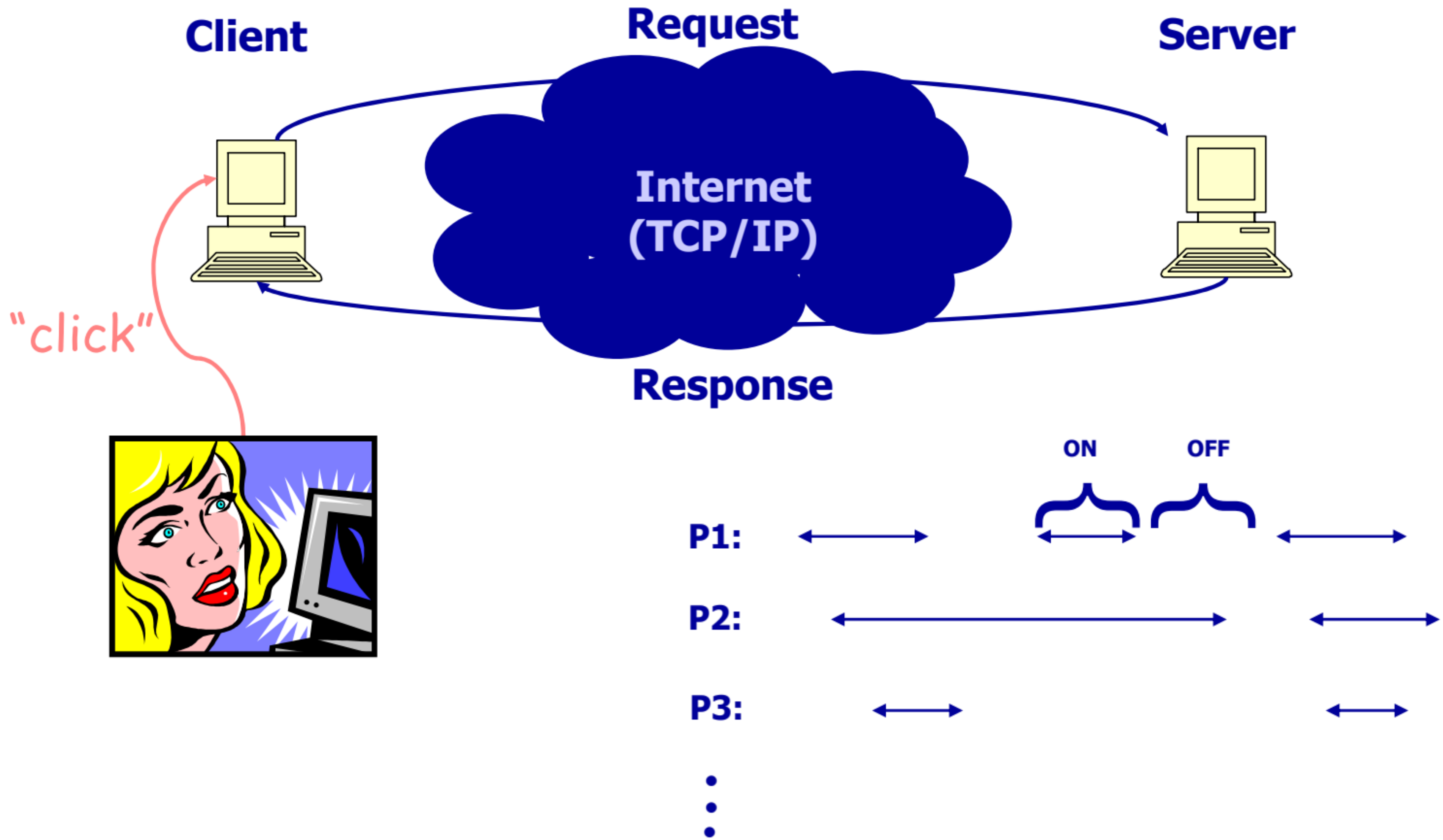
Long Range Dependence: Autocorrelation



Long Range Dependence: Autocorrelation

How Does Autocorrelation Arise?

- Traffic is the superposition of flows



Long-range-dependence

- A process with LRD has an autocorrelation function:

$$r(k) \approx k^{-\beta} \; as \; k \longrightarrow \infty \text{ where } 0 < \beta < 1, \text{and} \sum r(k) \longrightarrow \infty$$

- In other words, the autocorrelation function decays hyperbolically and is non- summable.

- For the conventional SRD (short-range dependence) process, an autocorrelation function decays exponentially.

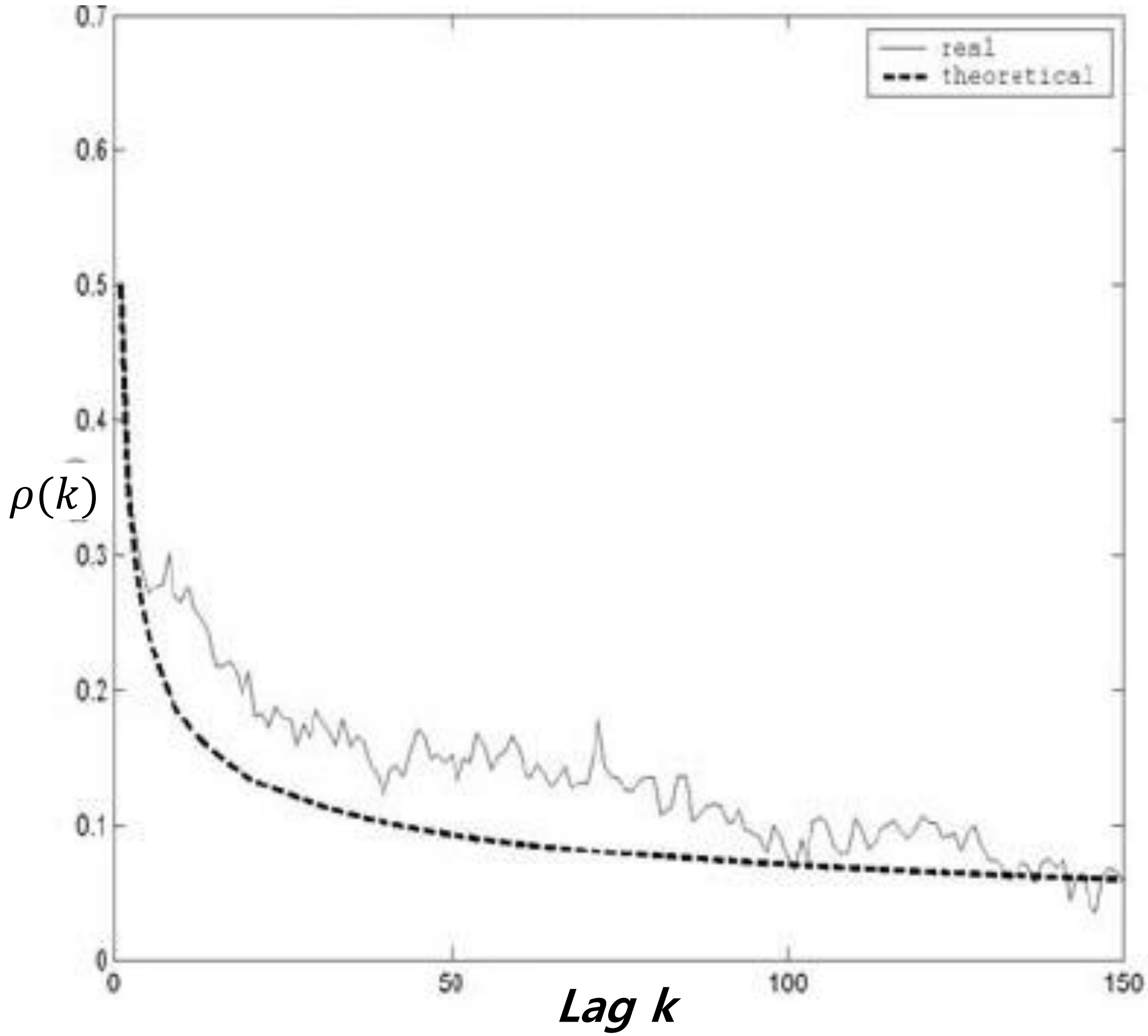
Long-range-dependence(LRD)

- It is often used to describe the tail-end behavior of the autocorrelation function of a stationary time series.
- In traffic modeling, LRD is often used to describe the aggregate traffic such as WAN (Wide Area Network), whereas self-similarity is usually used in the context of LAN (Local Area Network) or individual application traffic.

Non-Degenerate Autocorrelations

- For self-similar processes, the autocorrelation function for the aggregated process is indistinguishable from that of the original process
- If autocorrelation coefficients match for all lags k, then called "exactly self-similar"
- If autocorrelation coefficients match only for large lags k, then called "asymptotically self-similar"

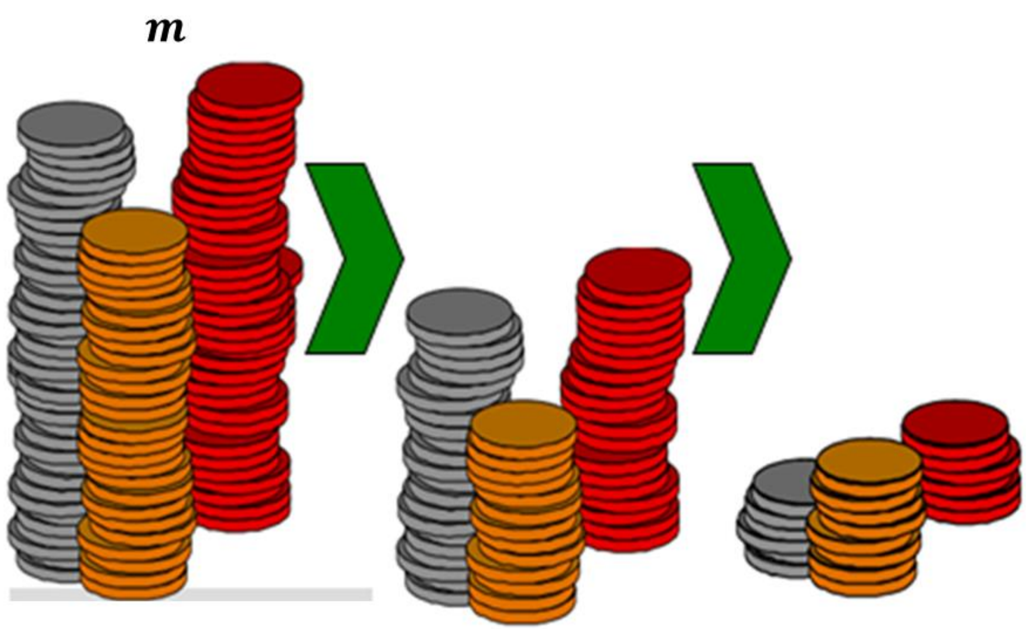
Non-Degenerate Autocorrelations



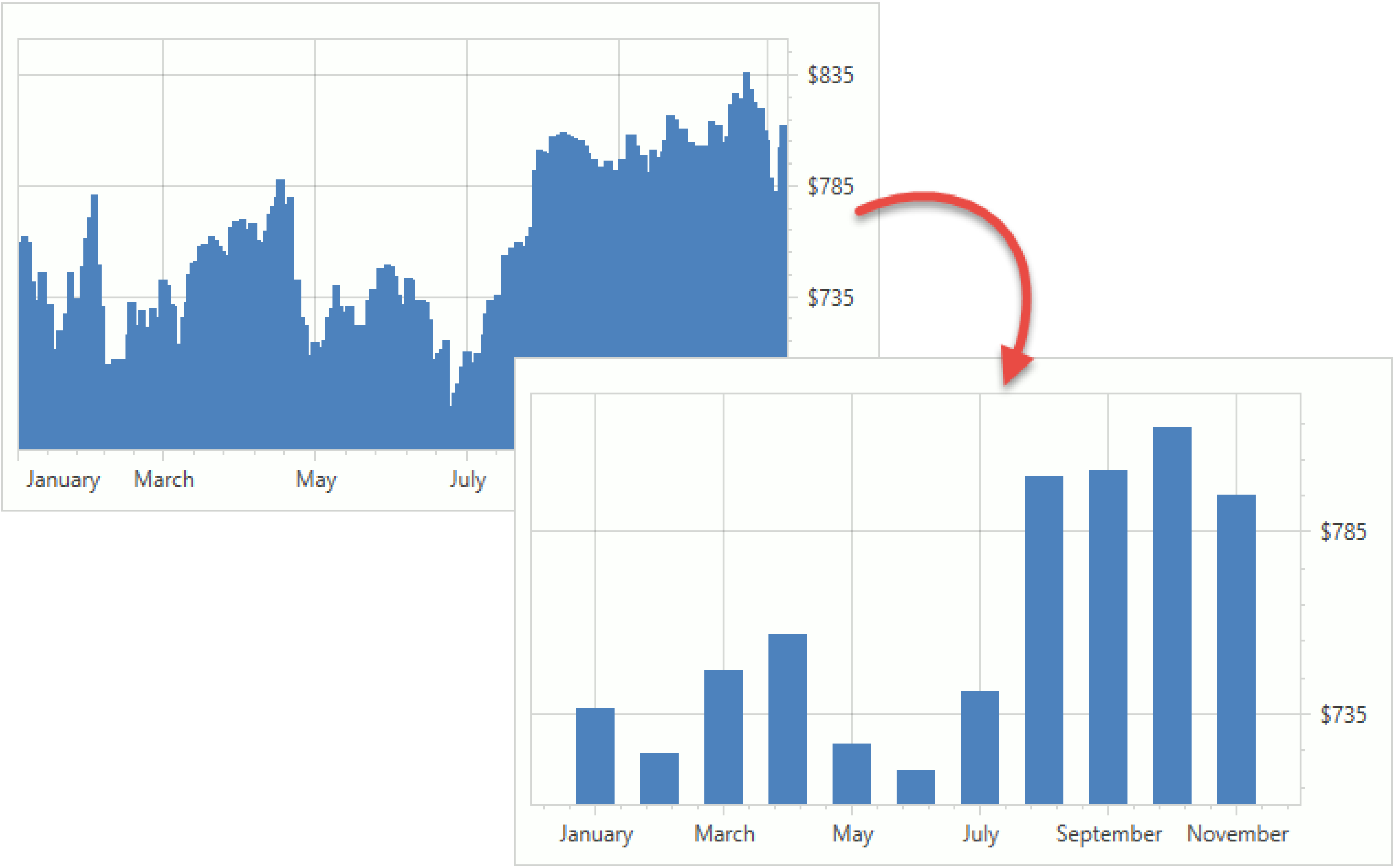
Autocorrelation for Self-Similar Traffic

Meaning of Aggregation

- Aggregation of a time series $X(t)$ means smoothing the time series by averaging the observations over non-overlapping blocks of size m to get a new time series $X'(t)$



Non-Degenerate Autocorrelations → Aggregation



Example of Data Aggregation

Examples of aggregation

- Suppose the original time series X(t) contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1...

Then the aggregated time series for $m = 5$ is:

Examples of aggregation

- Suppose the original time series X(t) contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1...

Then the aggregated time series for $m = 5$ is:

6.0

Examples of aggregation

- Suppose the original time series X(t) contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1...

Then the aggregated time series for $m = 5$ is:

6.0 4.4

Examples of aggregation

- Suppose the original time series X(t) contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1...

Then the aggregated time series for $m = 5$ is:

6.0 4.4 6.4 4.8 ...

Examples of aggregation

- Suppose the original time series X(t) contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1...

Then the aggregated time series for $m = 10$ is:

5.2

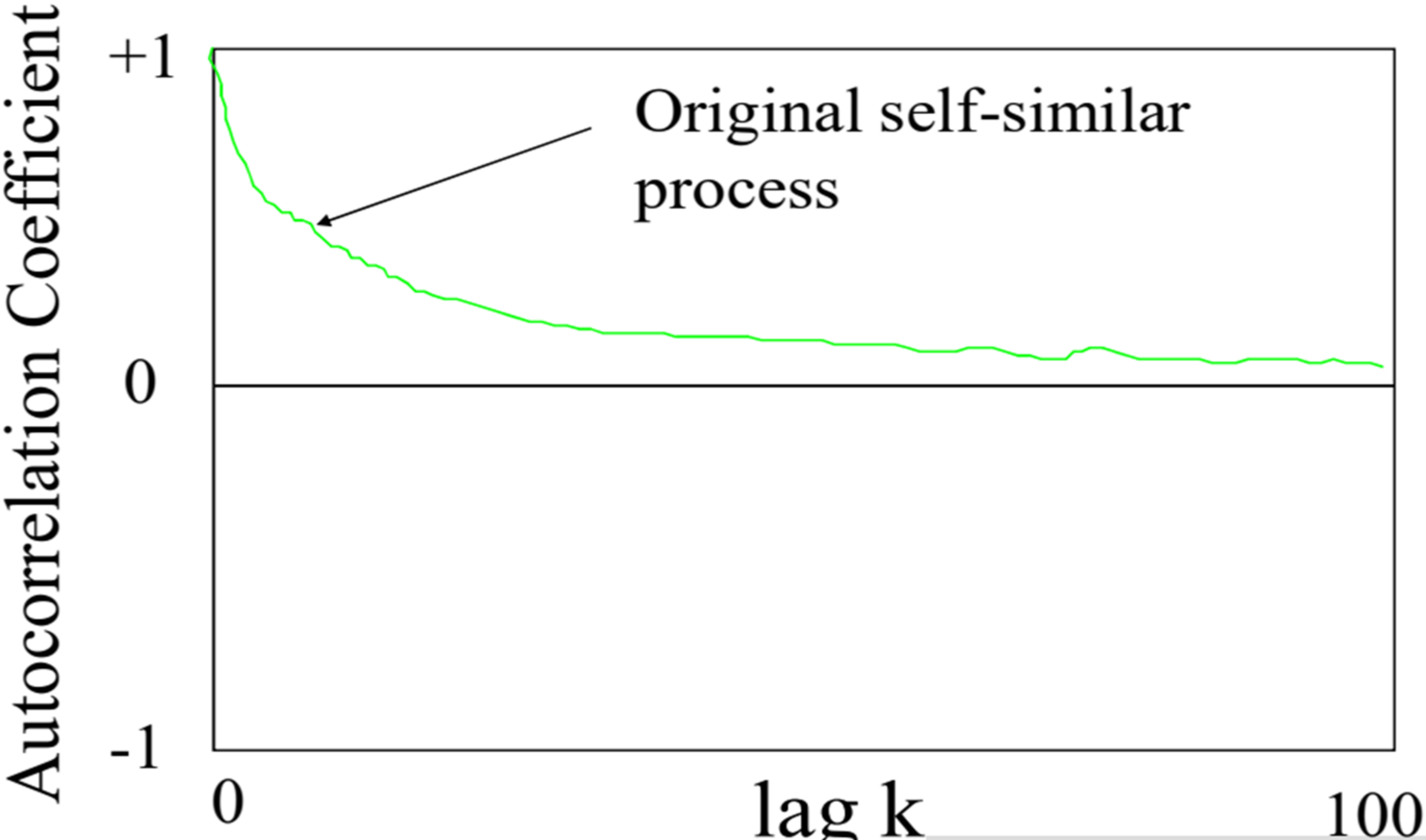
Examples of aggregation

- Suppose the original time series X(t) contains the following (made up) values:

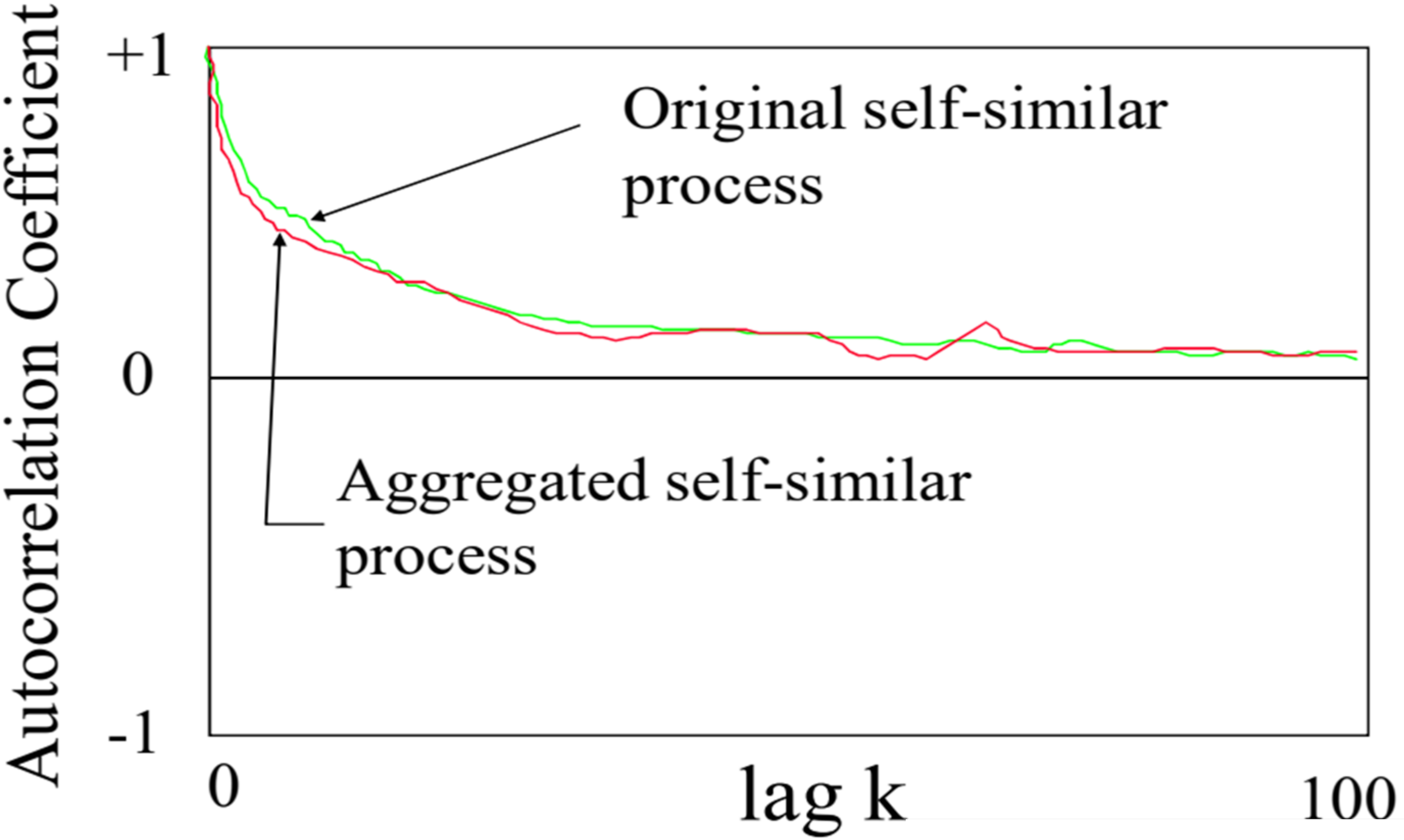
2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1...

Then the aggregated time series for $m = 10$ is:

5.2 **5.6 ...**



Non-Degenerate Autocorrelations



Non-Degenerate Autocorrelations

R/*S* Plot

- The rescaled range is a statistical measure of the variability of a time series introduced by the British hydrologist Harold Edwin Hurst (1880–1978).
- Its purpose is to provide an assessment of how the apparent variability of a series changes with the length of the time-period being considered.
- Another way of testing for self-similarity, and estimating the Hurst parameter

R/S Plot

- Plot the R/S statistic for different values of n, with a log scale on each axis
- If time series is self-similar, the resulting plot will have a straight line shape with a slope H that is greater than 0.5
- Called an R/S plot, or R/S pox diagram

R/S Plot

- For almost all naturally occurring time series, the rescaled adjusted range statistic (also called the R/S statistic) for sample size n obeys the relationship

$$E\left[\frac{R(n)}{S(n)}\right] \cong n^H$$

where $R(n) = max(0, W_1, \ldots, W_n) - min(0,, W_1, \ldots, W_n)$,

$S^2(n)$ is the sample variance,

and $W_k = \sum_{i=1}^{k} X_i - k\overline{X_n}$ for $k = 1, 2, \ldots, n$

R/S Plot

- For almost all naturally occurring time series, the rescaled adjusted range statistic (also called the R/S statistic) for sample size n obeys the relationship

$$E\left[\frac{R(n)}{S(n)}\right] \cong n^H$$

$$\log E\left[\frac{R(n)}{S(n)}\right] = \log cn^H = H \log n + \log c$$

So, H is slope.

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

  2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1

- There are 20 data points in this example

An example of $\boldsymbol{R/S}$ statistic

- Suppose the original time series $\boldsymbol{X(t)}$ contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1

- There are 20 data points in this example

- For R/S analysis with n = 1, you get 20 samples, each of size 1:

Block 1: $\overline{\boldsymbol{X_n}} = \mathbf{2.0}$, $\boldsymbol{W_1} = \mathbf{0}$, $\boldsymbol{R(n)} = \mathbf{0}$, $\boldsymbol{S(n)} = \mathbf{0}$

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1

- There are 20 data points in this example
- For R/S analysis with n = 1, you get 20 samples, each of size 1:

Block 2: $\overline{X_n} = 7.0\ ,\ W_1 = 0,\ R(n) = 0,\ S(n) = 0$

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1

- There are 20 data points in this example

- For R/S analysis with n = 2, you get 20 samples, each of size 2:

Block 1: $\overline{X_n} = 4.5$, $W_1 = -2.5$, $W_2 = 0$,
$R(n) = 0 - (-2.5) = 2.5$, $S(n) = 2.5$,

$$R(n) / S(n) = 1.0$$

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

  2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1

- There are 20 data points in this example

- For R/S analysis with n = 2, you get 20 samples, each of size 2:

Block 2: $\overline{X_n} = 8.0$, $W_1 = -4.0$, $W_2 = 0$,
$R(n) = 0 - (-4.0) = 4.0$, $S(n) = 4.0$,

$$R(n)/S(n) = 1.0$$

An example of R/S statistic 슬라이드 추가

- Suppose the original time series $X(t)$ contains the following (made up) values:

2 7 4 12 5 0 8 2 8 4 6 9 11 3 3 5 7 2 9 1

Block 1

- For R/S analysis with n=5 , you get 4 samples, each of size 4 :

Block 1: $\overline{X_n} = 6.0$, $W_1 = -4.0$, $W_2 = -3.0$, $W_3 = -5.0$, $W_4 = 1.0$, $W_5 = 0$, $S(n) = 3.41$, $R(n) = 1.0 - (-5.0) = 6.0$, $^{R(n)}/_{S(n)} = 1.76$

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

2 7 4 12 5 | 0 8 2 8 4 | 6 9 11 3 3 5 7 2 9 1

Block 1 Block 2

- For R/S analysis with n=5 , you get 4 samples, each of size 4 :

Block 2: $\overline{X_n} = 4.4$, $W_1 = -4.4$, $W_2 = -0.8$, $W_3 = -3.2$, $W_4 = 0.4$, $W_5 = 0$, $S(n) = 3.2$, $R(n) = 0.4 - (-4.4) = 4.8$, $R(n)/_{S(n)} = 1.5$

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

2	7	4	12	5	0	8	2	8	4	6	9	11	3	3	5	7	2	9	1

Block 1 (2 7 4 12 5), Block 2 (0 8 2 8 4), Block3 (6 9 11 3 3)

- For R/S analysis with n=5 , you get 4 samples, each of size 4 :

Block 3: $\overline{X_n} = 6.4$, $W_1 = -0.4$, $W_2 = 2.2$, $W_3 = 6.8$, $W_4 = 3.4$, $W_5 = 0$, $S(n) = 3.2$, $R(n) = 6.8 - (-0.4) = 7.8$, ${}^{R(n)}/_{S(n)} = 2.4375$

An example of R/S statistic

- Suppose the original time series $X(t)$ contains the following (made up) values:

Block 1	Block 2	Block3	Block4
2 7 4 12 5	0 8 2 8 4	6 9 11 3 3	5 7 2 9 1

- For R/S analysis with n=5 , you get 4 samples, each of size 4 :

Block 4: $\overline{X_n} = 4.8$, $W_1 = 0.2$, $W_2 = 2.4$, $W_3 = -0.4$, $W_4 = 3.8$, $W_5 = 0$, $S(n) = 3.2$, $R(n) = 3.8 - (-0.4) = 4.2$, ${}^{R(n)}/_{S(n)} = 1.3125$

An example of R/S statistic

- Suppose the original time series $\boldsymbol{X(t)}$ contains the following (made up) values:

Block 1	Block 2	Block3	Block4
2 7 4 12 5	0 8 2 8 4	6 9 11 3 3	5 7 2 9 1

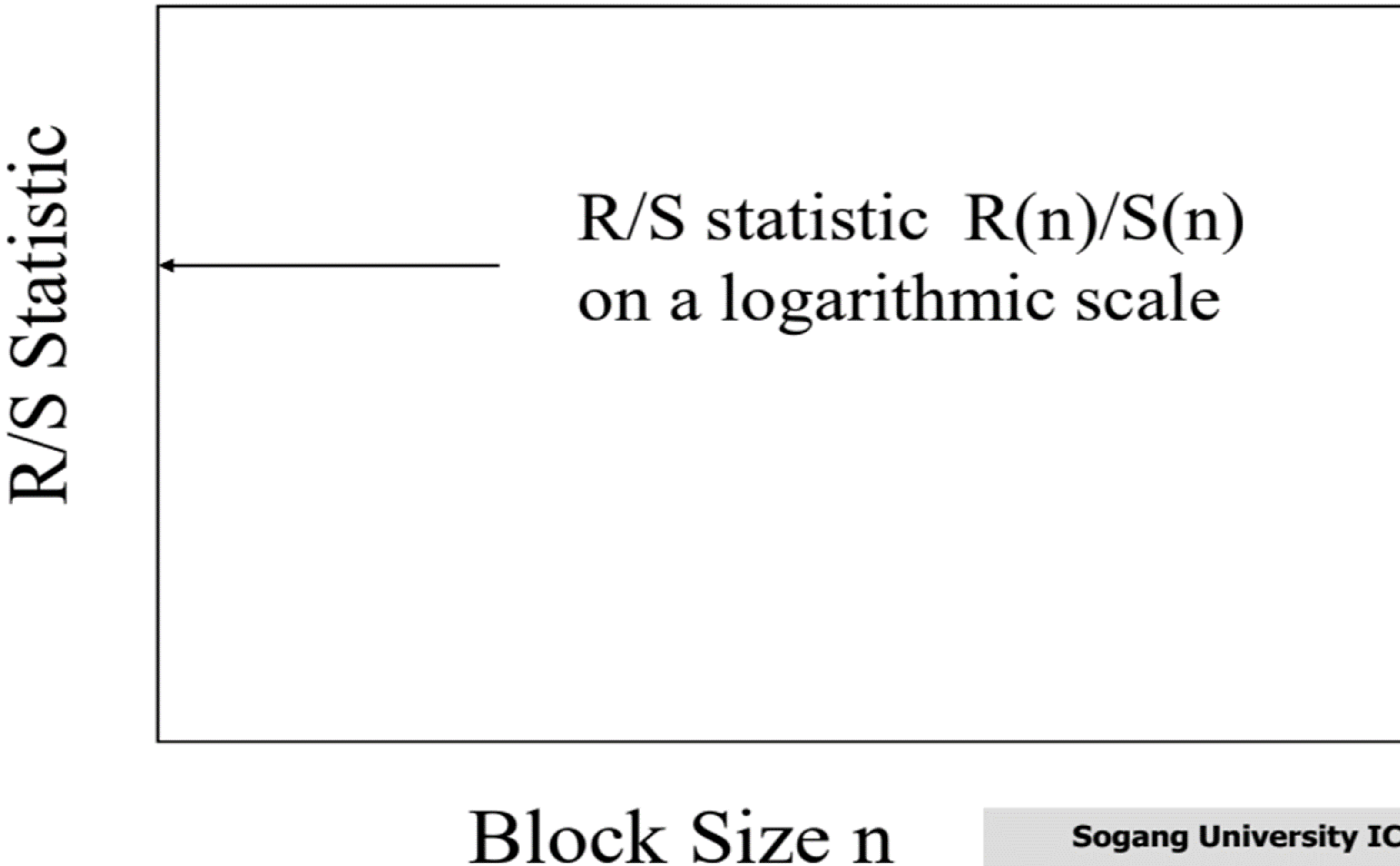
- For R/S analysis with $\boldsymbol{n}$=**5** , you get **4** samples, each of size **4** :

$$E\left[\frac{R(n)}{S(n)}\right] = 1.7525 \cong 5^H$$

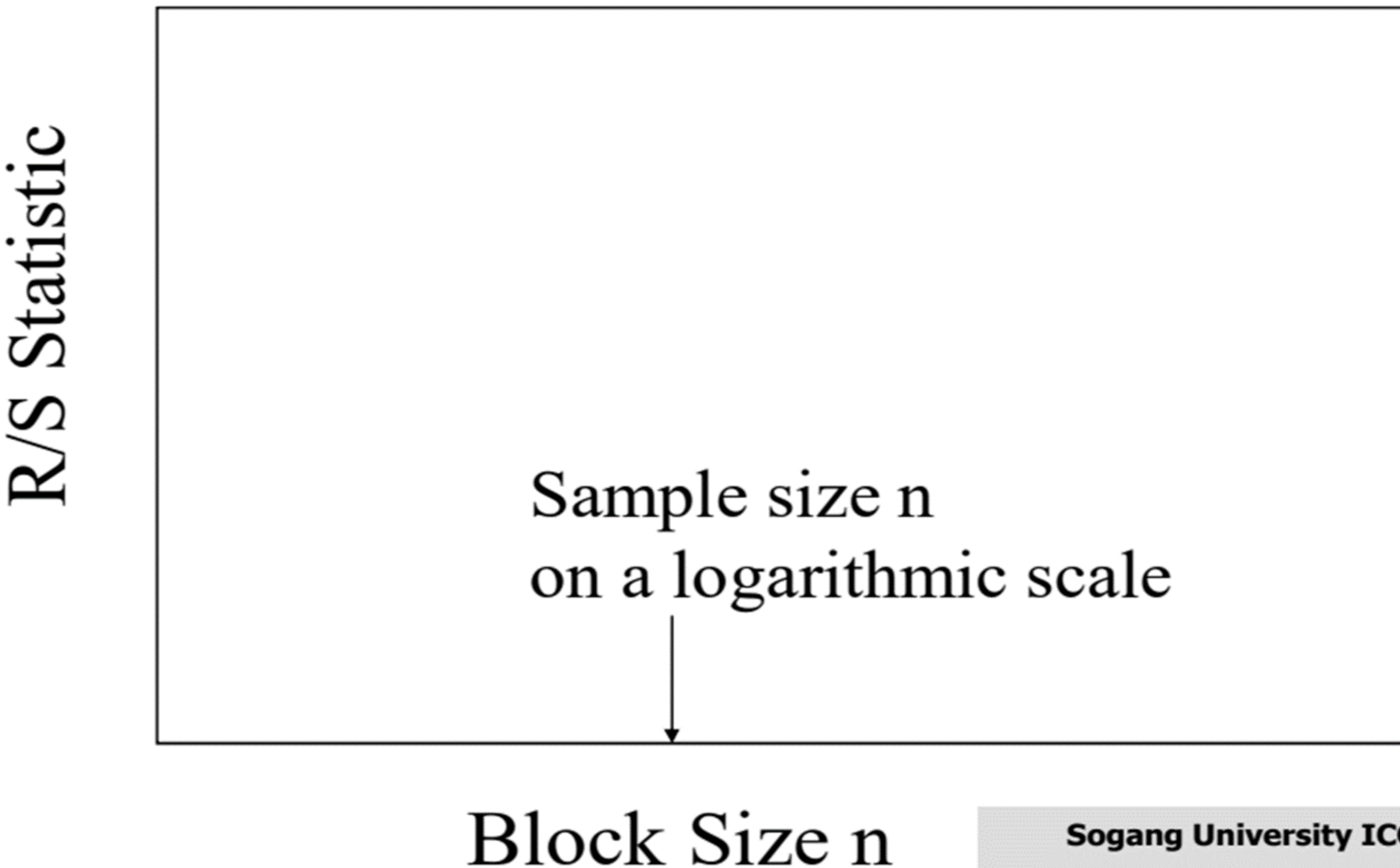
$$\log_5 1.7525 \approx 0.35 \cong H$$

The Hurst Effect

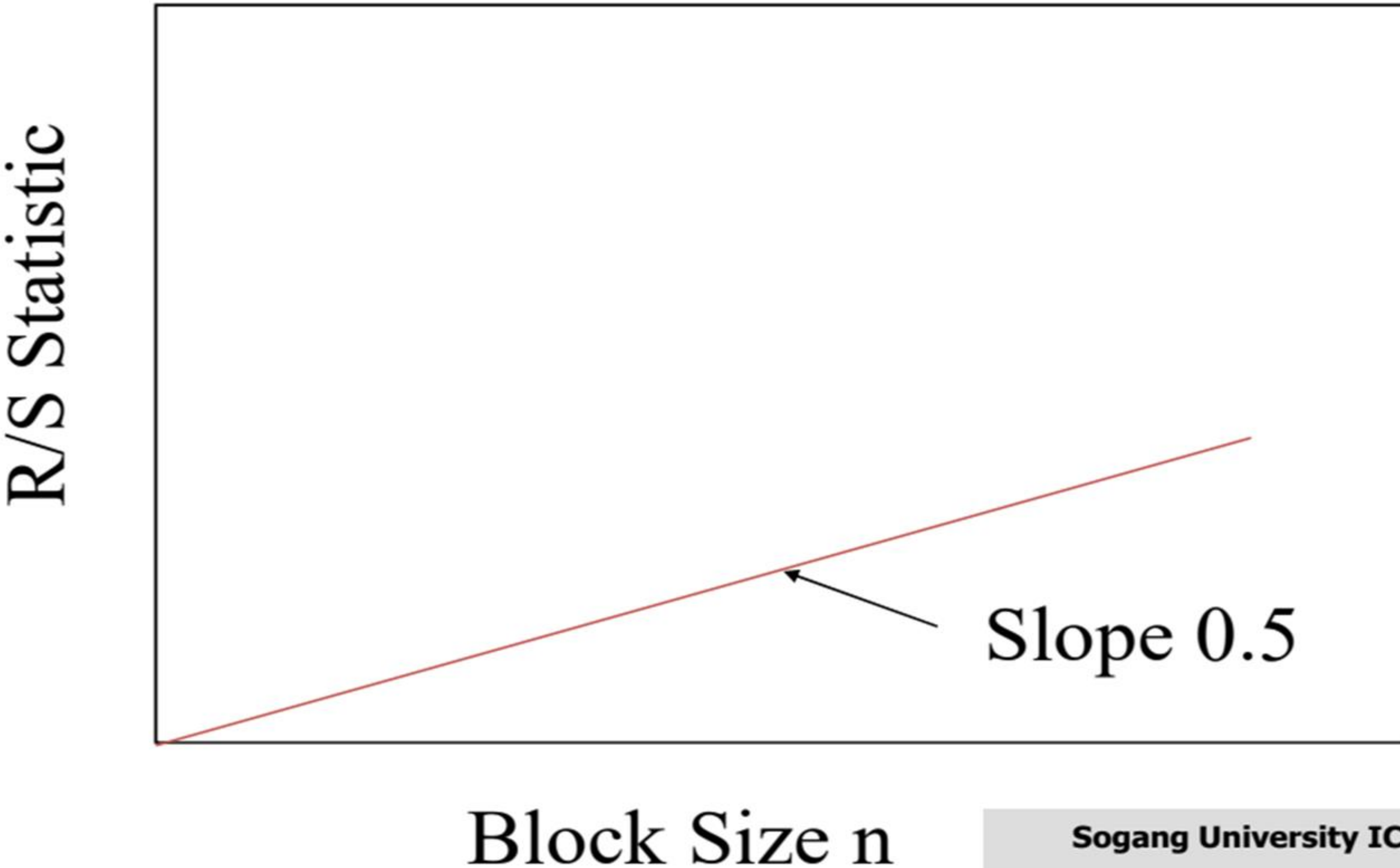
- For models with only short range dependence, H is almost always 0.5
- For self-similar processes, $0.5 < H < 1.0$
- This discrepancy is called the Hurst Effect, and H is called the Hurst parameter
- Single parameter to characterize self-similar processes



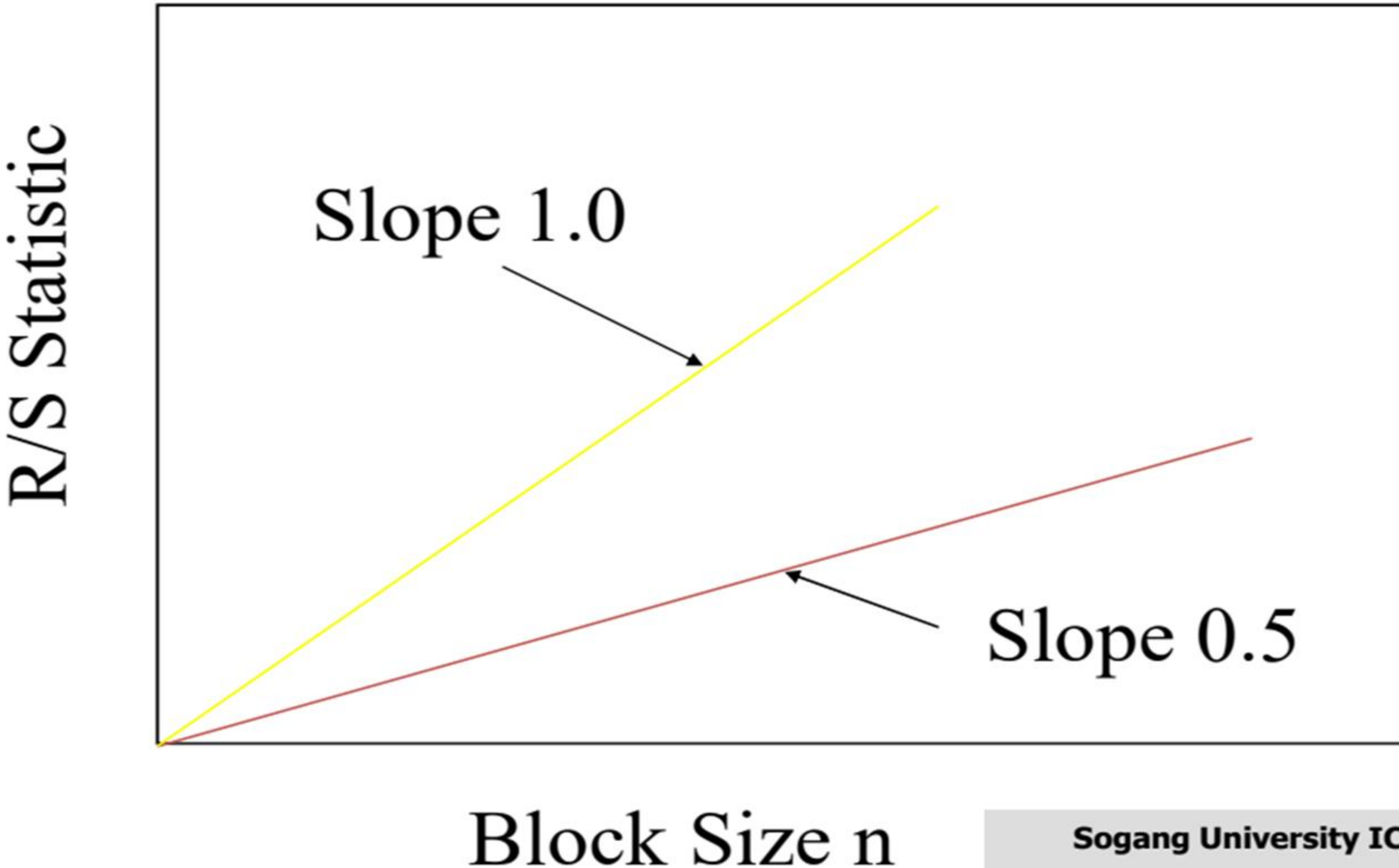
R/S Pox Diagram



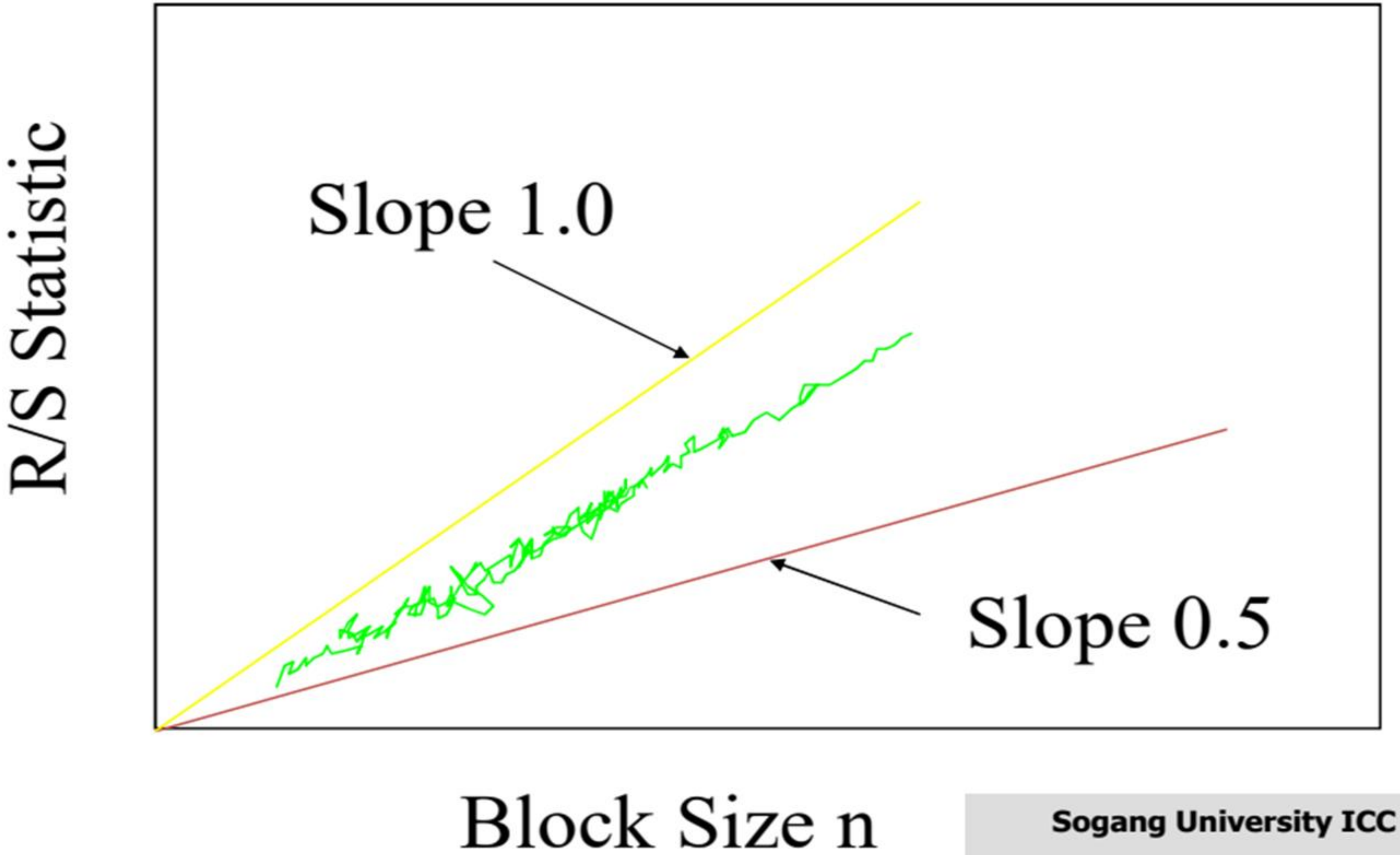
R/S Pox Diagram



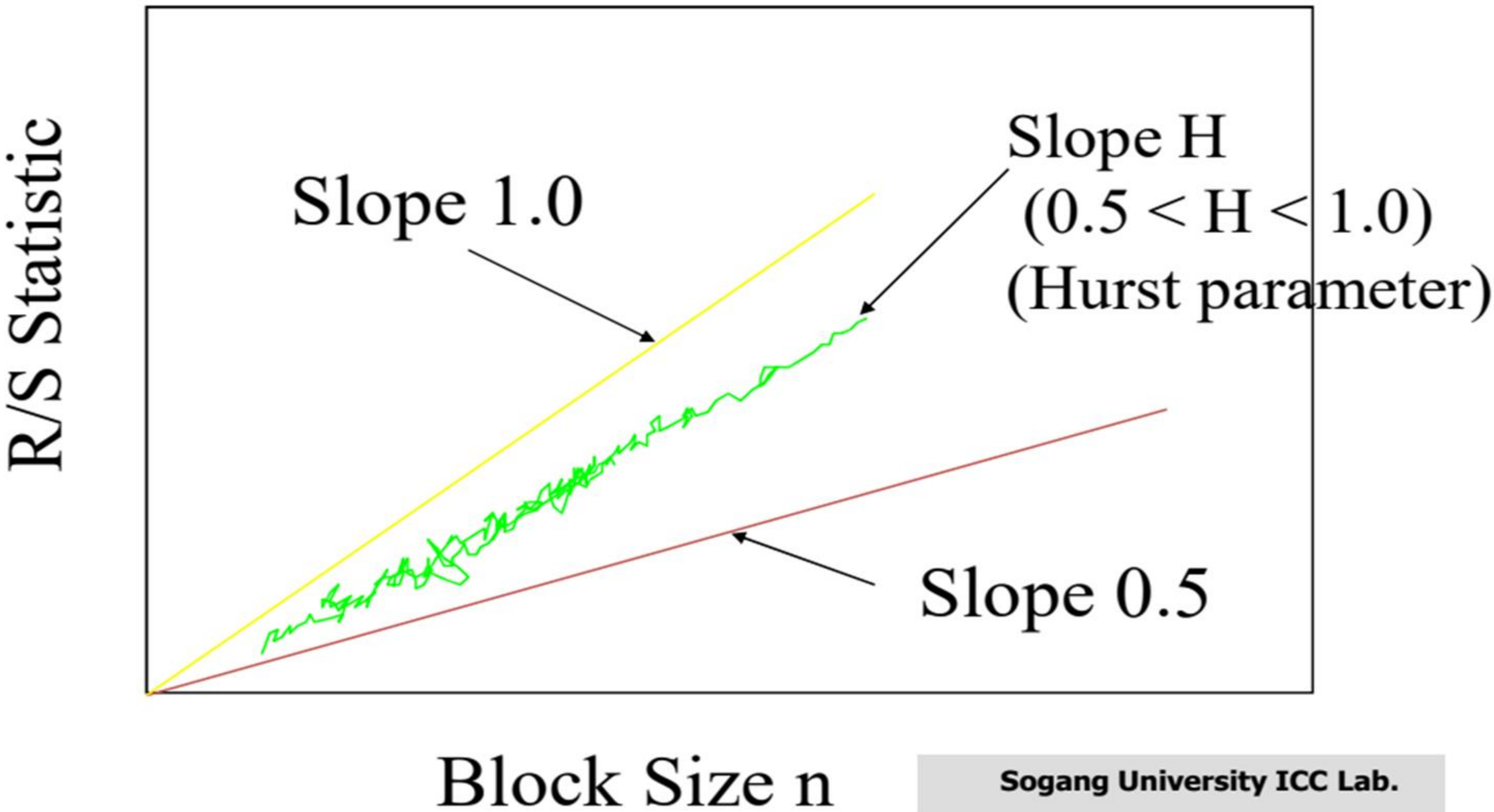
***R/S* Pox Diagram**



***R/S* Pox Diagram**

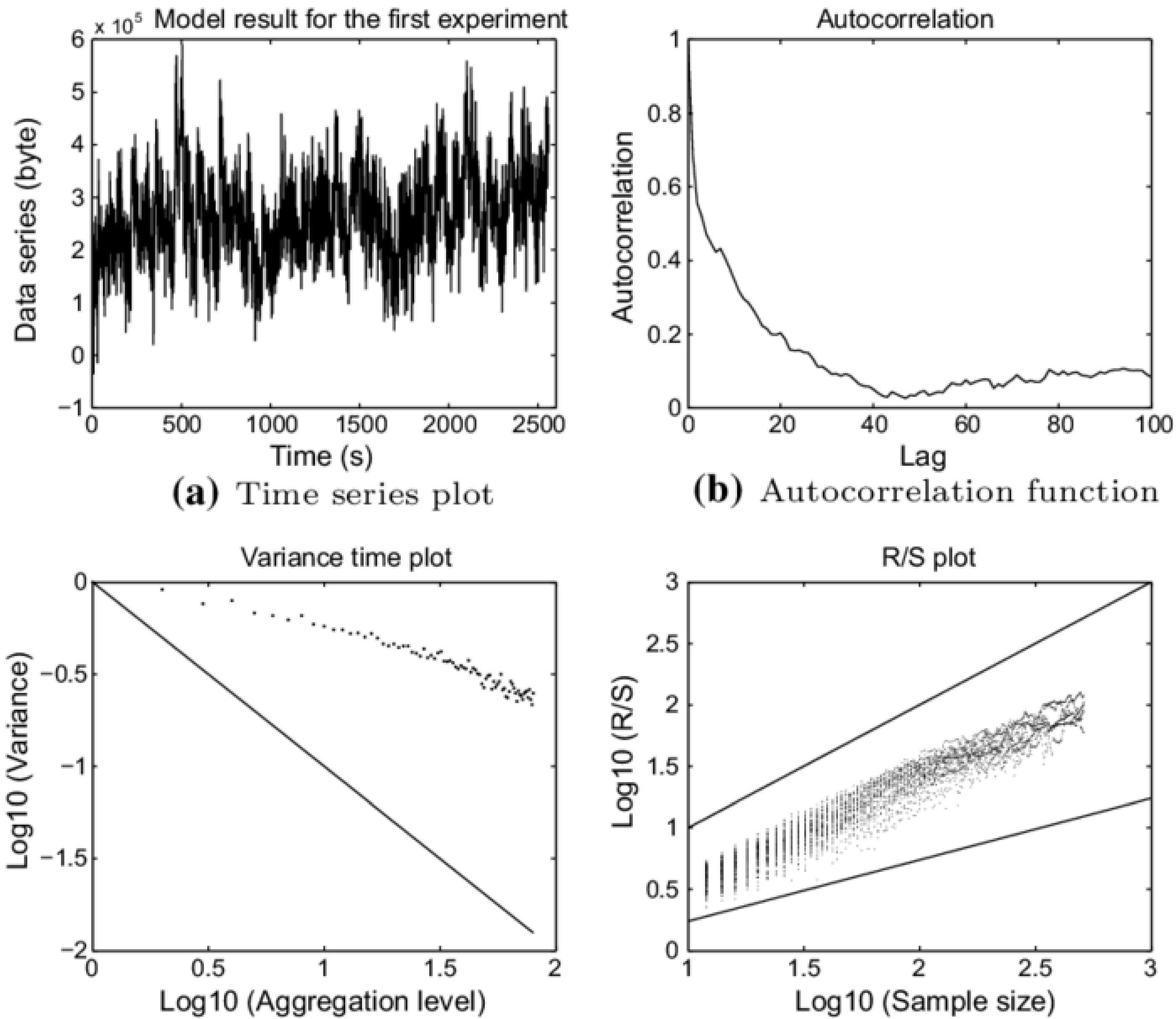


R/S Pox Diagram



Sogang University ICC Lab.

R/S Pox Diagram

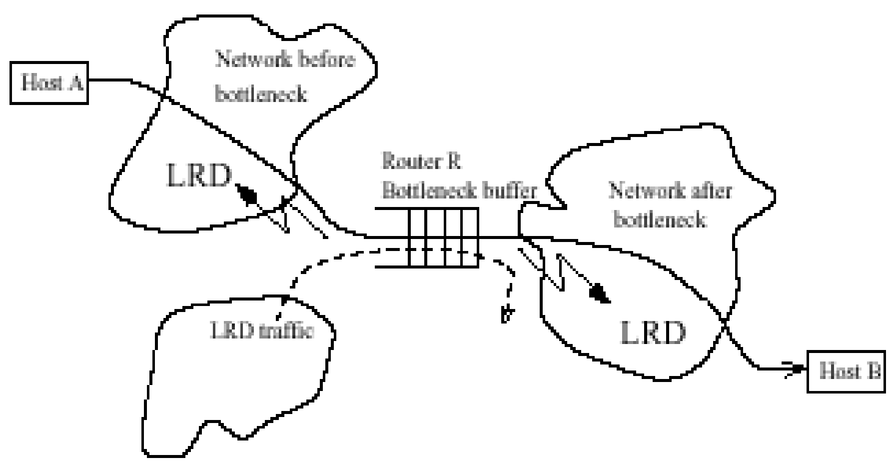


Evidence of self-similarity of video traffic in Scenario I (model data)

Analysis of traffic characteristics(Self-similar)

<Results of Willinger and Paxon’s study>

- Studies have shown that the "long-term dependence (LRD)" phenomenon, in which the correlation between traffic decreases slowly, spreads to Internet traffic through bottleneck sharing



Analysis of traffic characteristics(Self-similar)

- Relation of background traffic flowing through the bottleneck section and TCP traffic passing through the section:

$$TCP(t) = C - B(t), \qquad s.t.\, B(t) = \sum_{i=1}^{N} B_i(t)$$

  - $\mathrm{TCP}(t)$ is TCP connection traffic that shares bottleneck sections.
  - C is The total bandwidth of the bottleneck section
  - $B(t)$ is background traffic consisting of traffic generated by a large number of connections

Analysis of traffic characteristics(Self-similar)

- TCP traffic follows the characteristics of background traffic.
- If background traffic $B(t)$ shows self-similarity, TCP-connected traffic also shows self-similarity and moves self-similarity to other traffic

$\rightarrow$ The entire Internet traffic has self-similarity