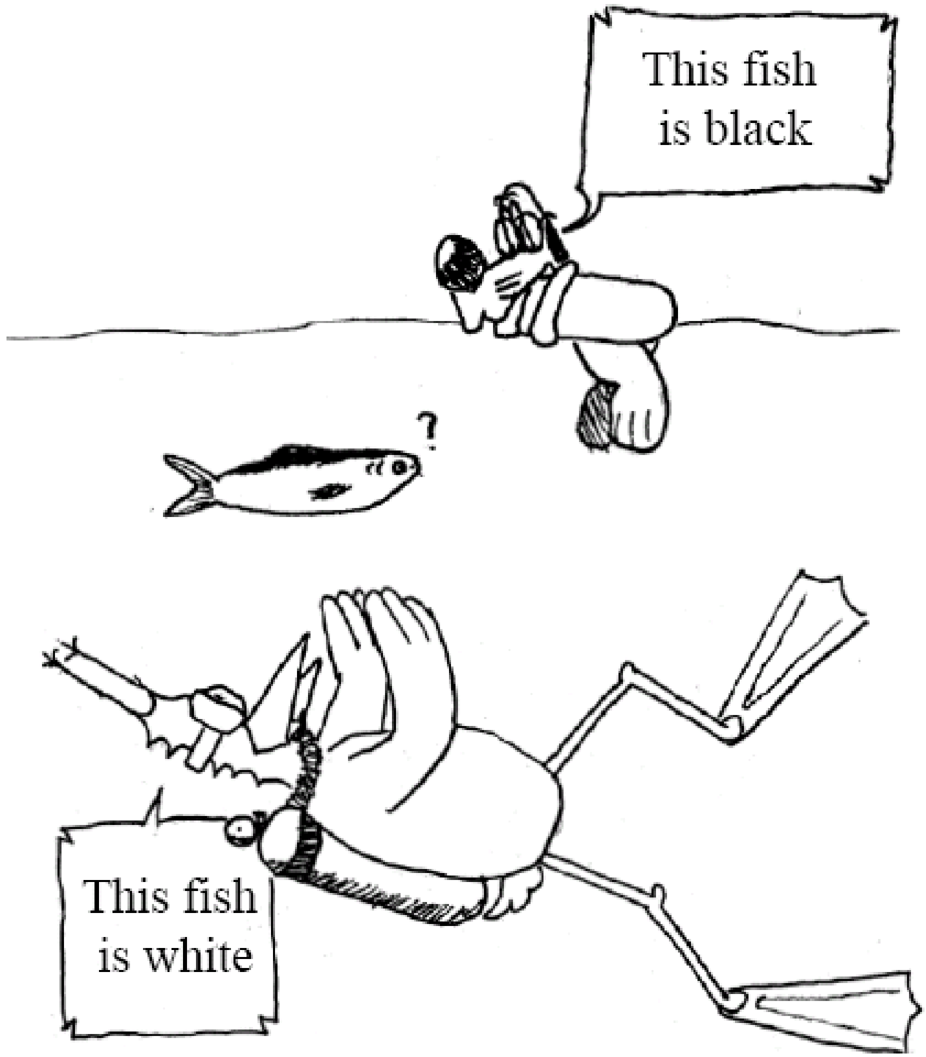
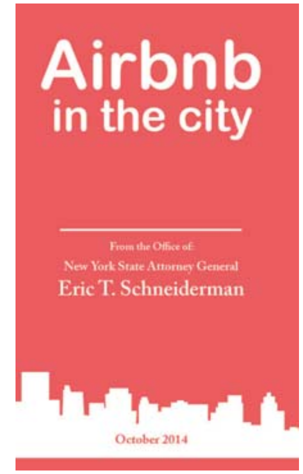
Palm Calculus
Part 1
The Importance of the Viewpoint

JY Le Boudec



Who says the truth ?



AirBnB claims:
median occupancy of rented listings is 11% (40 days a year)

Insideairbnb.com claims:
median occupancy of rented listings is 40-50% (165 days a year)

Airbnb. Data on the Airbnb community in New York City . Technical report, AirBnB corporation, Dec. 2015.

Lecuyer, M., Tucker, M. and Chaintreau, A., 2017, April. Improving the Transparency of the Sharing Economy. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 1043-1051). International World Wide Web Conferences Steering Committee.

Who says the truth ?

SovRail: according to our systematic tracking system, probability of a train being late $\leq$ 5%

BorduKonsum: according to our consumer survey, probability of being late $\approx$ 30%

1. Event versus Time Averages

Consider a simulation, state S_t

Assume simulation has a stationary regime

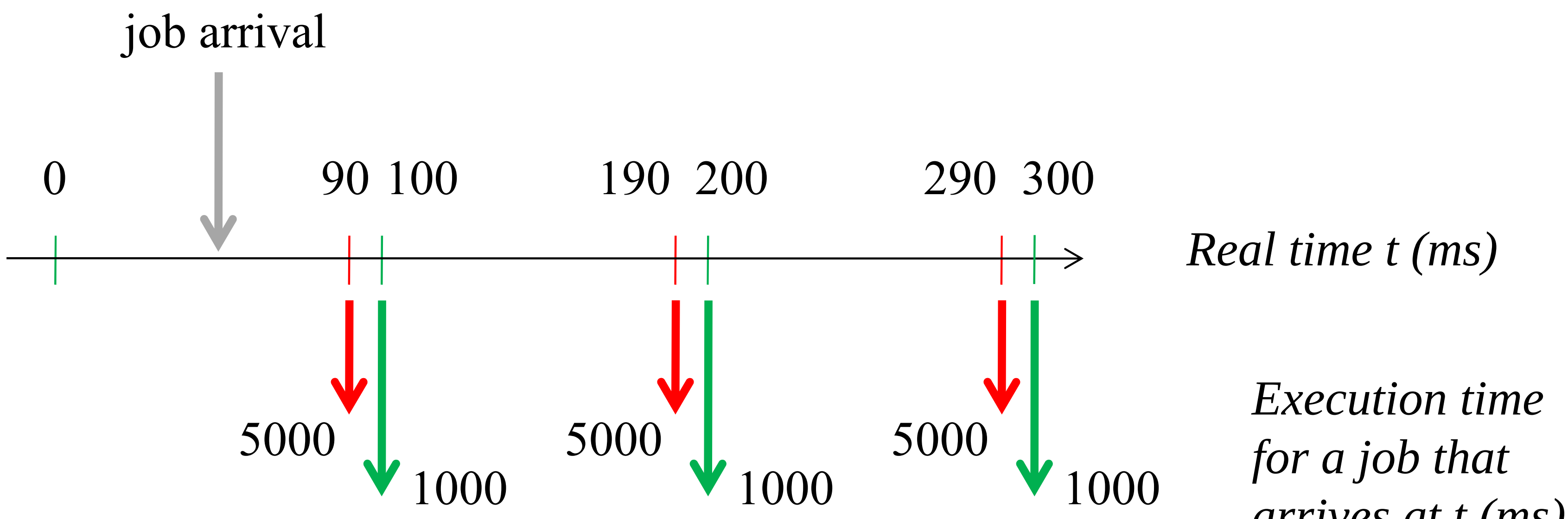
Consider an *Event Clock*: times T_n at which some specific changes of state occur

Ex: arrival of job; Ex. queue becomes empty

Event average statistic: mean queue length seen by an arriving customer: $\bar{Q}^0 = \frac{1}{N+1} \sum_{i=0}^{N} Q(T_n^-)$

Time average statistic: mean queue length (seen by an inspector): $\bar{Q} = \frac{1}{T_N - T_0} \int_{T_0}^{T_N} Q(s) ds$

Example: Gatekeeper; Average execution time



Viewpoint 1: System Designer

Two processes, with execution times 5000 and 1000

$$W_s = \frac{5000 + 1000}{2} = 3000$$

Viewpoint 2: Customer

Inspector arrives at a random time

red processor is used with proba $\frac{90}{100}$

$$W_c = \frac{90}{100} \times 5000 + \frac{10}{100} \times 1000$$
$$= 4600$$

Sampling Bias

W_s and W_c are different, but both are *average execution times* !

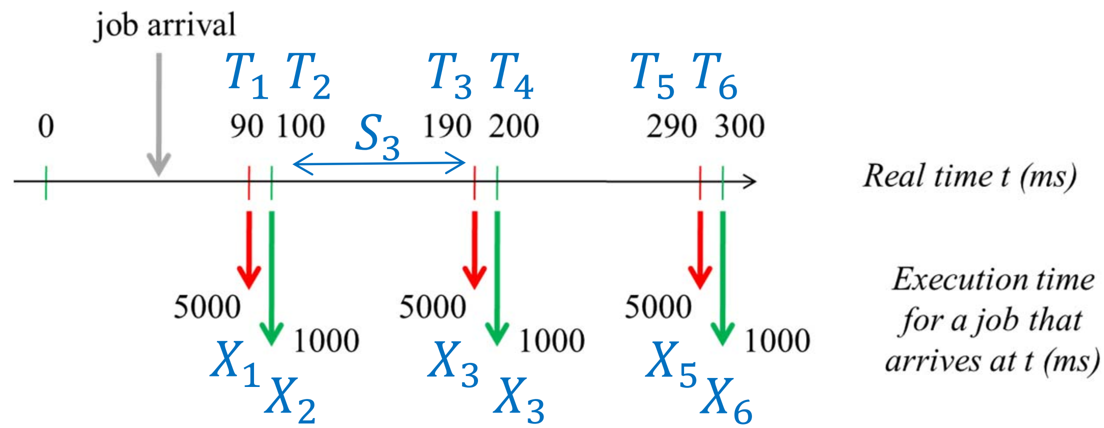
A metric definition should mention the sampling method (*viewpoint*)

Different sampling methods may provide different values: this is the *sampling bias*

Palm Calculus is a set of formulas for relating different viewpoints

Can often be obtained by means of the *Large Time Heuristic*

Large Time Heuristic Explained on an Example



We want to relate W_s and W_c

We apply the large time heuristic

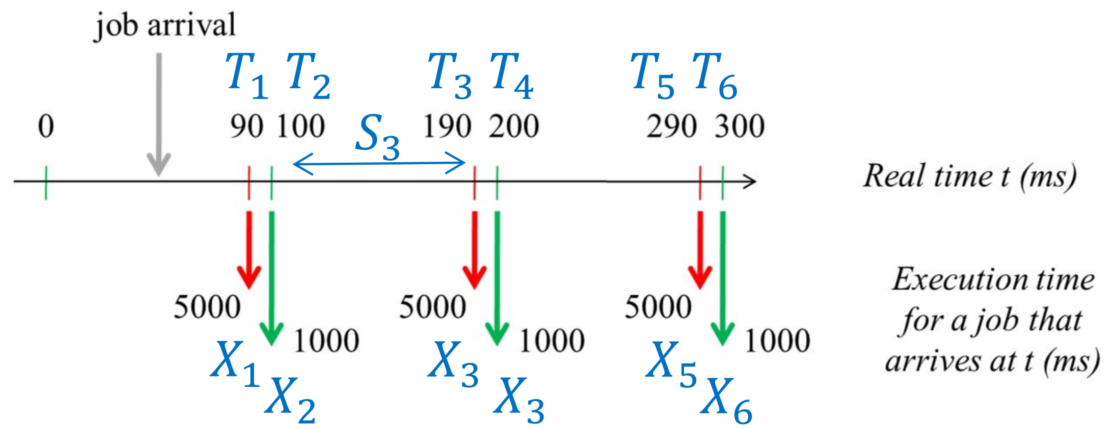
1. How do we evaluate these metrics in a simulation ?

$$W_s = \frac{1}{N} \sum_{n=1 \ldots N} X_n = \bar{X}$$

$$W_c = \frac{1}{T} \int_0^T X_{N^+(t)} dt$$

where $N^+(t) =$ index of next green or red arrow at or after T

Large Time Heuristic Explained on an Example



2. Break one integral into pieces that match the T_n's:

$$W_s = \frac{1}{N} \sum_{n=1\ldots N} X_n = \bar{X}$$

$$W_c = \frac{1}{T} \int_0^T X_{N^+(t)} dt$$

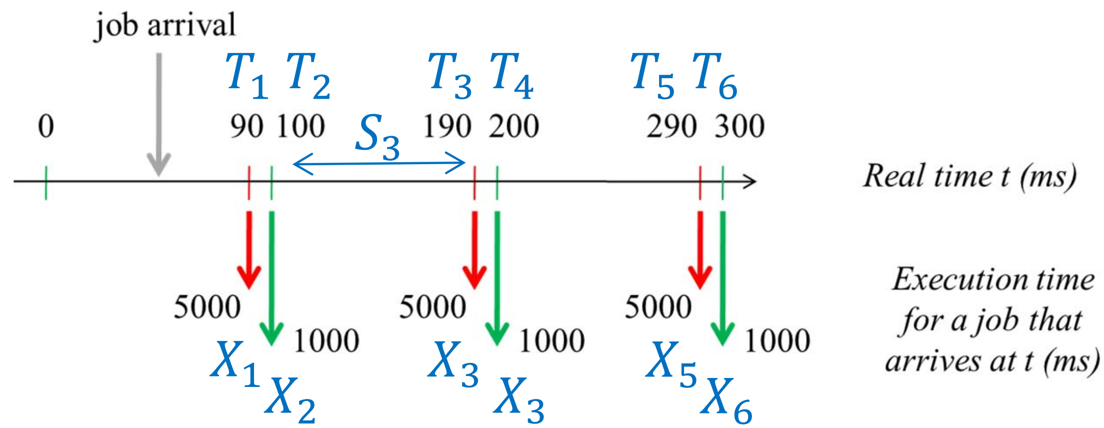
$$W_c = \frac{1}{T} \left(\int_0^{T_1} X_{N^+(t)} dt + \int_{T_1}^{T_2} X_{N^+(t)} dt + \cdots + \int_{T_{N-1}}^{T_N} X_{N^+(t)} dt \right)$$

$$= \frac{1}{T} \left(\int_0^{T_1} X_1 dt + \int_{T_1}^{T_2} X_2 dt + \cdots + \int_{T_{N-1}}^{T_N} X_N dt \right)$$

$$= \frac{1}{T} (T_1 X_1 + (T_2 - T_1)\, X_2 + \cdots + (T_N - T_{N-1})X_N)$$

$$= \frac{1}{T} (S_1 X_1 + S_2 X_2 + \cdots + S_N X_N)$$

Large Time Heuristic Explained on an Example



3. Compare

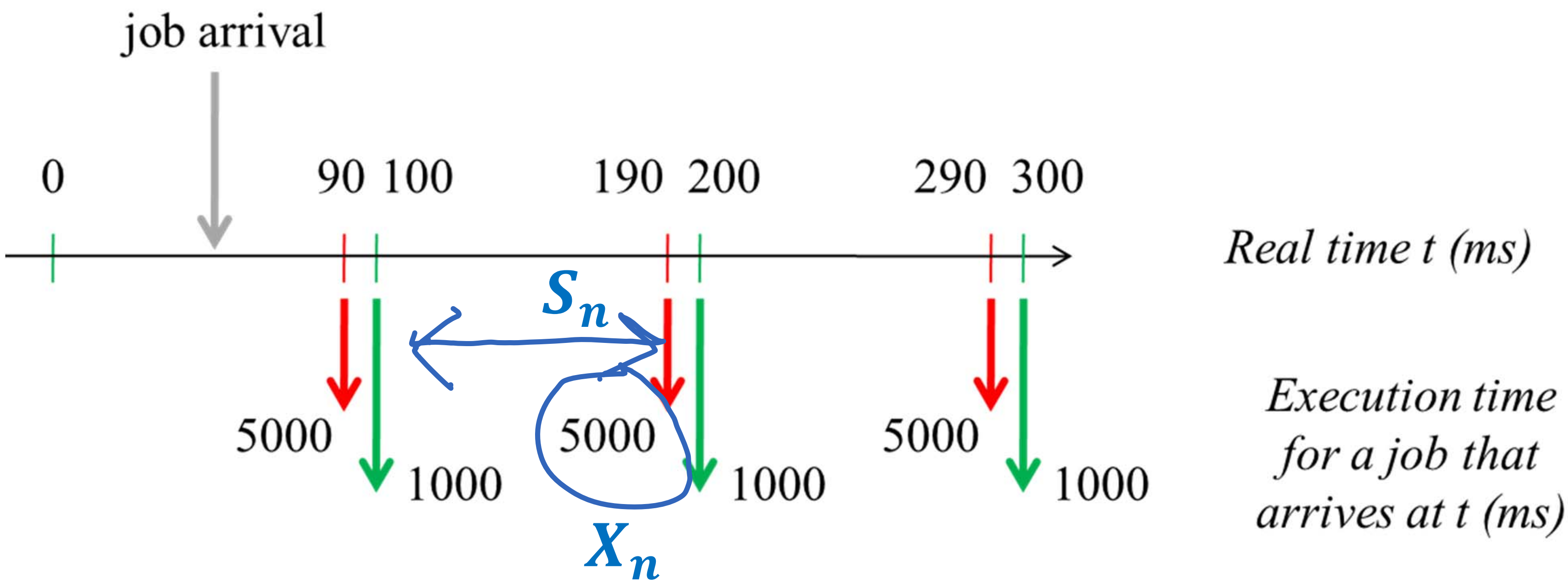
$$W_c = \frac{1}{T}(S_1X_1 + S_2X_2 + \cdots + S_NX_N)$$

$$= \frac{N}{T} \times \frac{1}{N}(S_1X_1 + S_2X_2 + \cdots + S_NX_N)$$

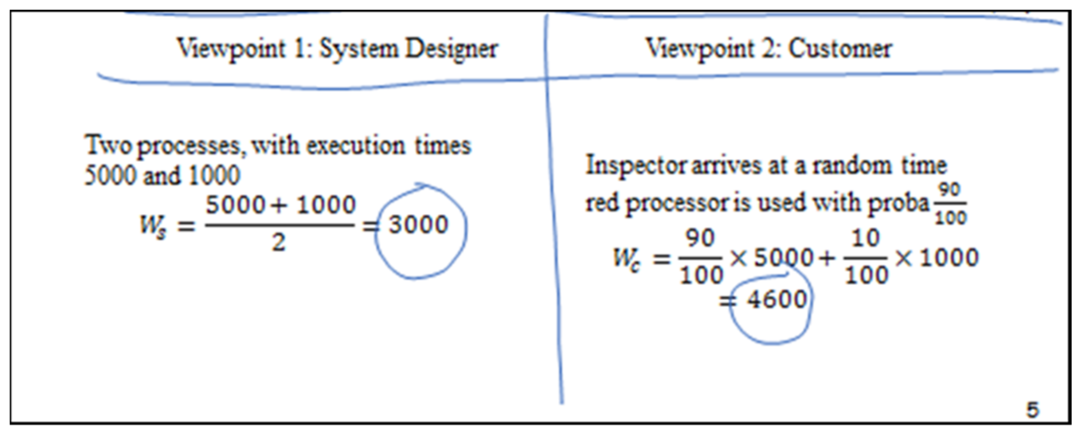
$$= \lambda \times \left(\mathrm{cov}(S, X) + \bar{S}\,\bar{X}\right) = \lambda \times \left(\mathrm{cov}(S, X) + \frac{1}{\lambda}\,\bar{X}\right)$$

$$W_c = \lambda\,\mathrm{cov}(S, X) + W_s$$

This is Palm Calculus !



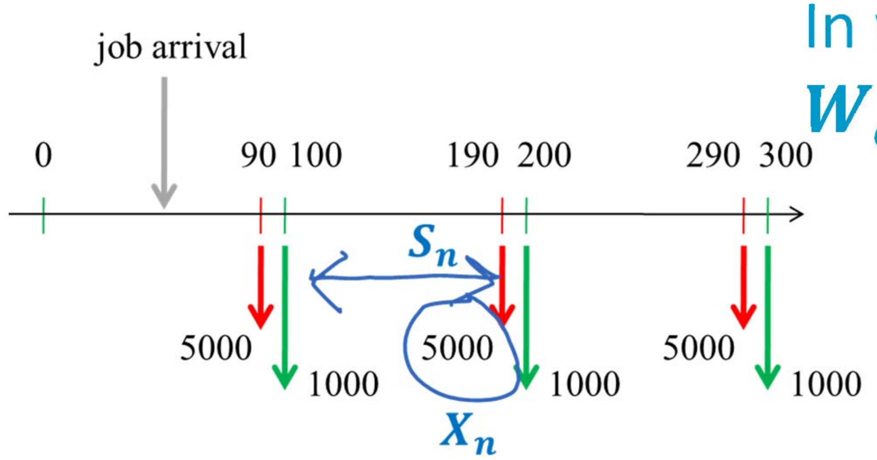
$$W_c = \lambda \operatorname{cov}(S, X) + W_s$$



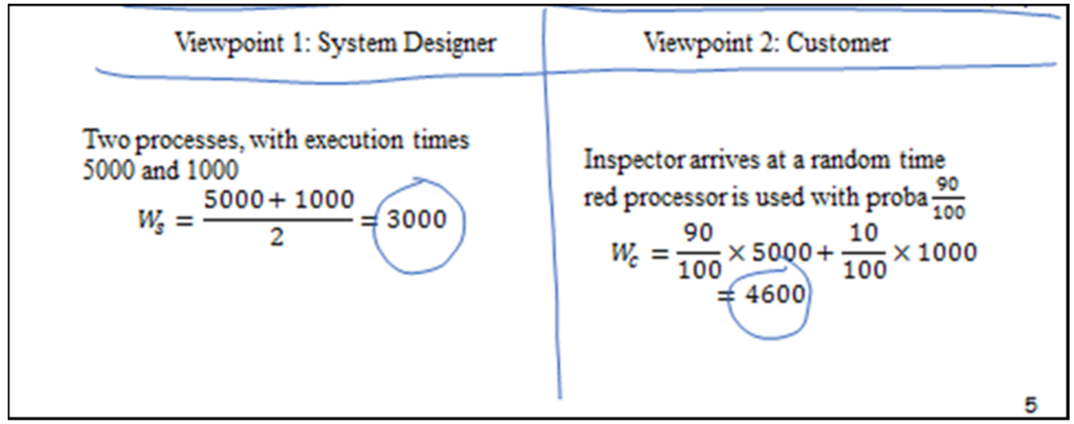
Viewpoint 1: System Designer	Viewpoint 2: Customer
Two processes, with execution times 5000 and 1000	Inspector arrives at a random time red processor is used with proba $\frac{90}{100}$
$W_s = \frac{5000+1000}{2} = 3000$	$W_c = \frac{90}{100} \times 5000 + \frac{10}{100} \times 1000 = 4600$

5

In which case do we expect to see $W_c > W_s$?



$$W_c = \lambda \operatorname{cov}(S, X) + W_s$$



Viewpoint 1: System Designer	Viewpoint 2: Customer
Two processes, with execution times 5000 and 1000	Inspector arrives at a random time red processor is used with proba $\frac{90}{100}$
$W_s = \frac{5000 + 1000}{2} = 3000$	$W_c = \frac{90}{100} \times 5000 + \frac{10}{100} \times 1000 = 4600$

5

A. S_n = 90, 10, 90, 10, 90; X_n = 5000, 1000, 5000, 1000, 5000

B. S_n = 90, 10, 90, 10, 90; X_n = 1000, 5000, 1000, 5000, 1000

C. Both

D. None

E. I don't know

Solution

In case A, S_n and X_n are positively correlated (when the interval is long, so is the processing time), i.e. $\mathbf{cov}(X, S) > 0$. By the Palm calculus formula: $W_c > W_s$

In case B, the correlation is negative, therefore $W_c < W_s$

Answer A

The Large Time Heuristic

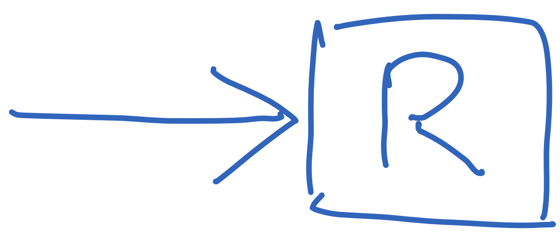
1. formulate each performance metric as a long run ratio, as you would do if you would be evaluating the metric in a discrete event simulation;
2. take the formula for the time average viewpoint and break it down into pieces, where each piece corresponds to a time interval between two selected events;
3. compare the two formulations.

Formally correct if simulation is stationary

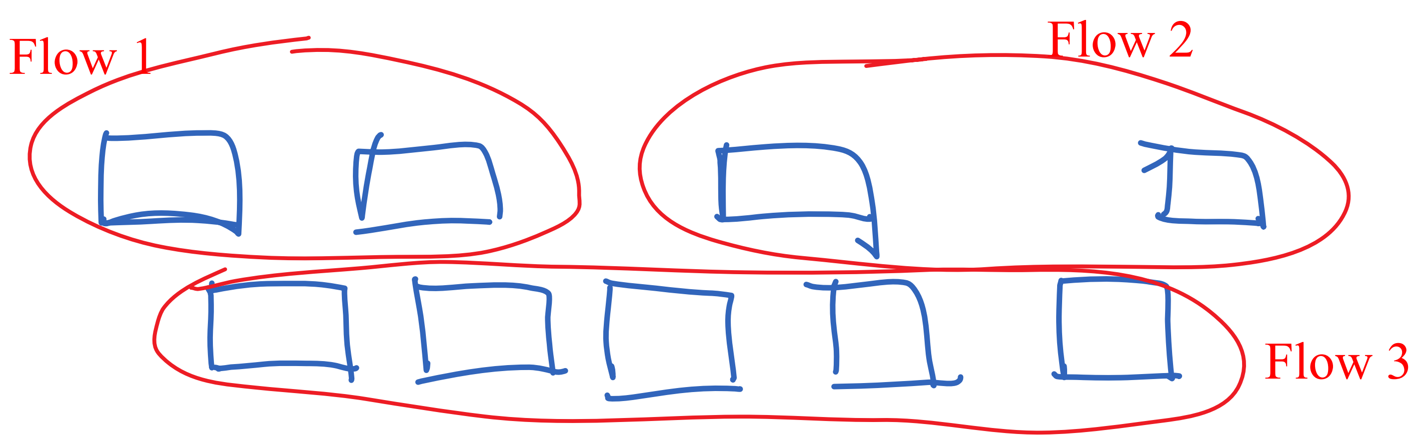
It is a *robust* method, i.e. independent of assumptions on distributions (and on independence)

Other «Clocks»

EXAMPLE 7.4: FLOW VERSUS PACKET CLOCK [96]. Packets arriving at a router are classified in "flows". We would like to plot the empirical distribution of flow sizes, counted in packets. We measure all traffic at the router for some extended period of time. Our metric of interest is the probability distribution of flow sizes. We can take a flow "clock", or viewpoint, i.e. ask: pick an arbitrary flow, what is its size ? Or we could take a packet viewpoint and ask: take an arbitrary packet, what is the magnitude of its flow ? We thus have two possible metrics (Figure 7.3):



Distribution of flow sizes
for an arbitrary flow
for an arbitrary packet



Which curves are for the per-packet viewpoint ?

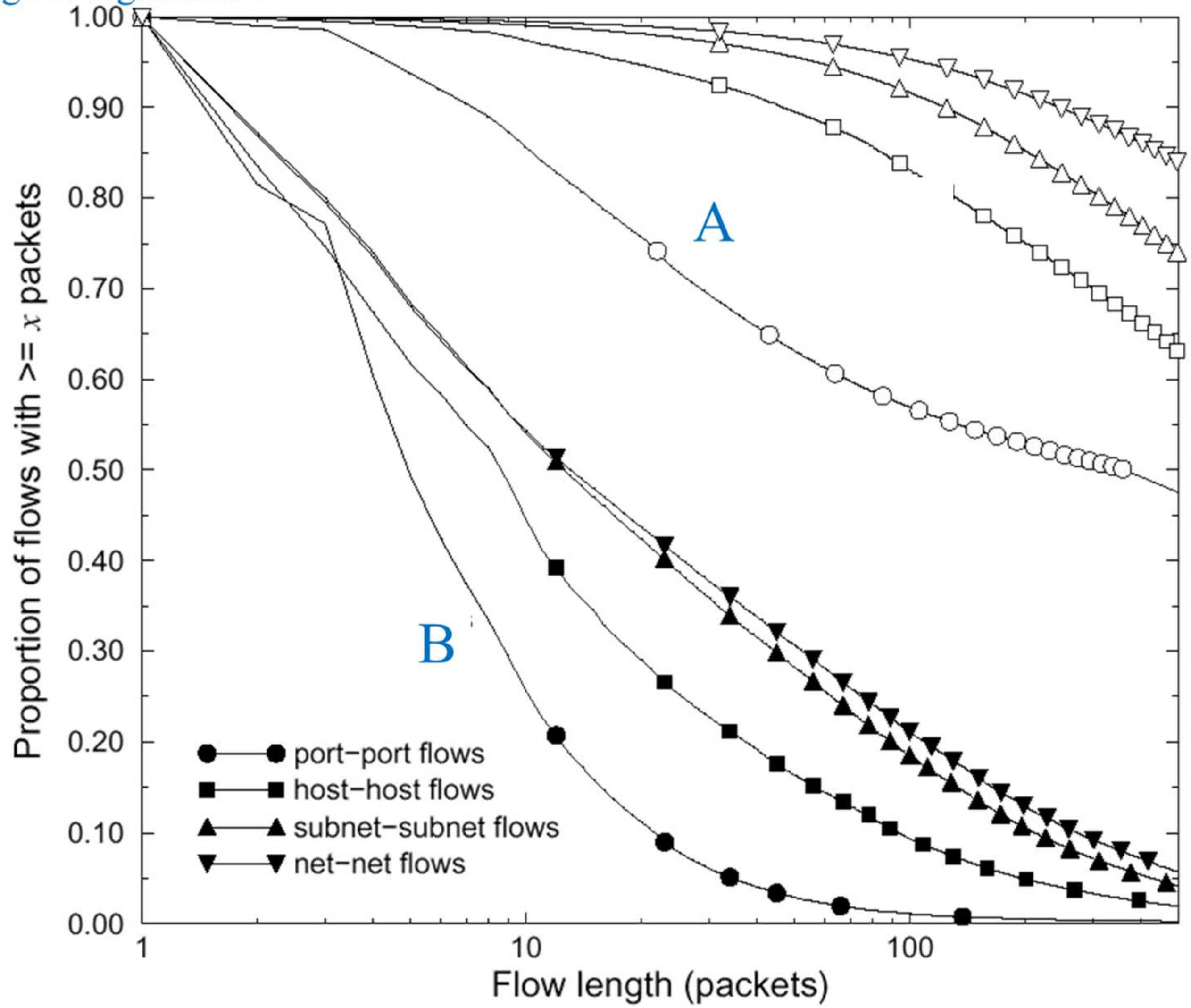
A. A

B. B

C. It depends

D. I don't know

Load Sensitive Routing of Long-Lived IP Flows
Anees Shaikh, Jennifer Rexford and Kang G. Shin
Proceedings of Sigcomm'99

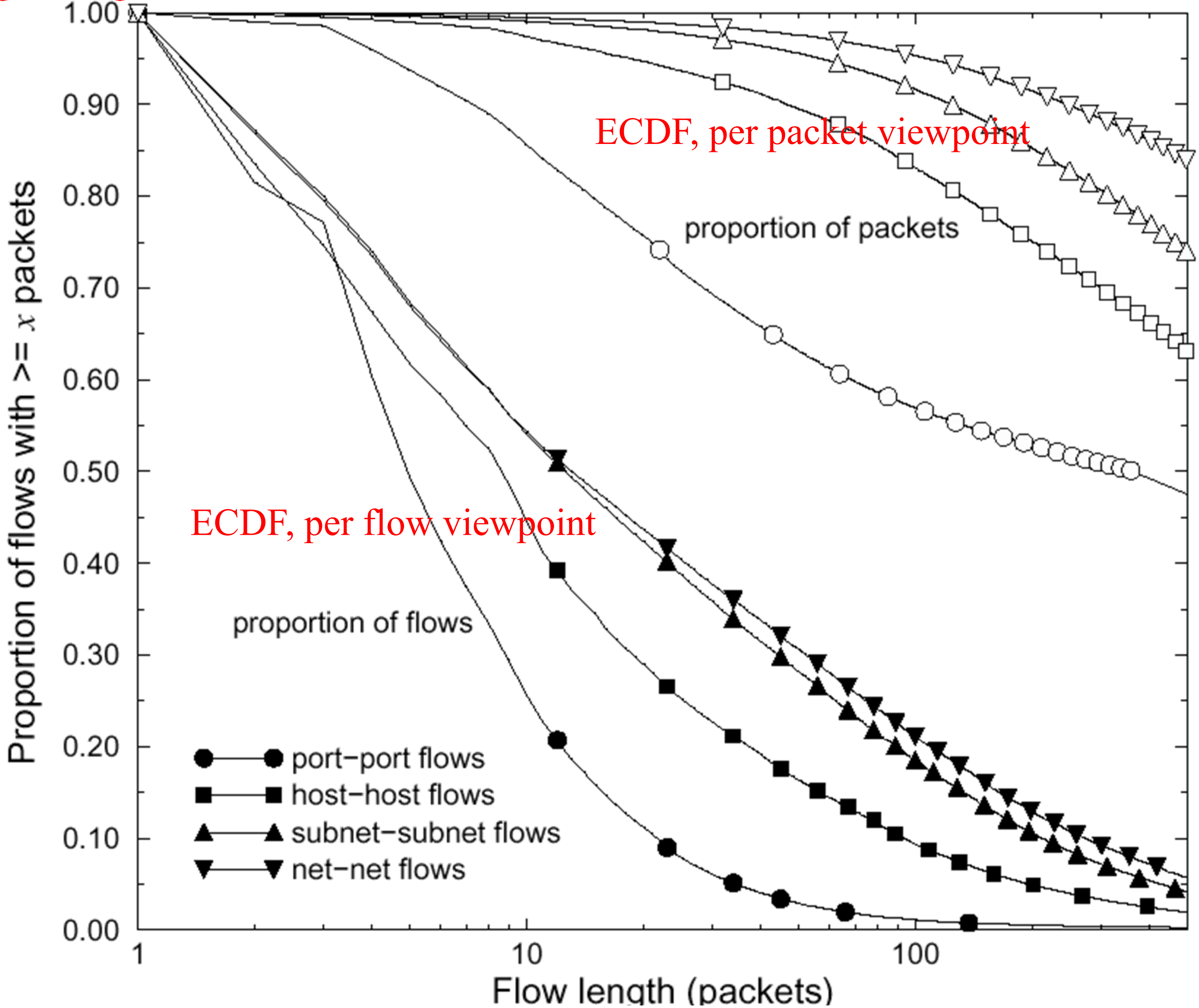


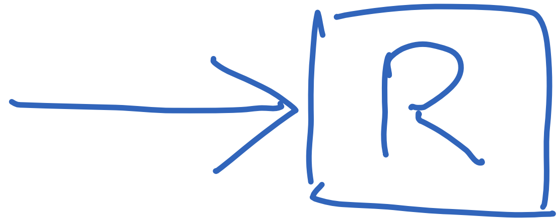
Solution

Answer A

There are more packets in the large flows. So more packets experience a large flow size.

Load Sensitive Routing of Long-Lived IP Flows
Anees Shaikh, Jennifer Rexford and Kang G. Shin
Proceedings of Sigcomm'99

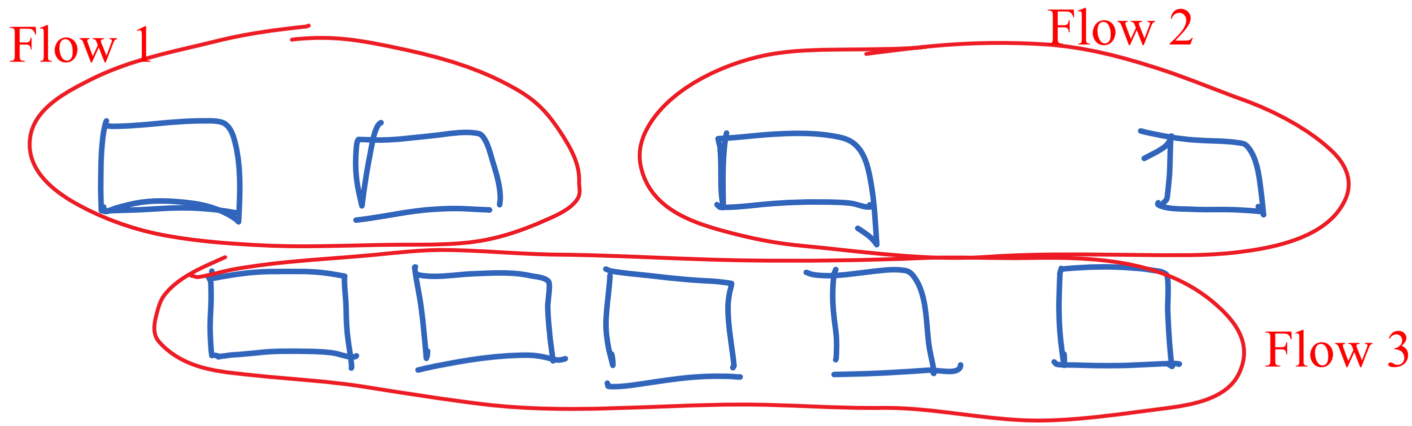




Distribution of flow sizes

for an arbitrary flow

for an arbitrary packet



Per flow $f_F(s) = 1/N \times$ number of flows with length s, where N is the number of flows in the dataset;

Per packet $f_P(s) = 1/P \times$ number of packets that belong to a flow of length s, where P is the number of packets in the dataset;

Mean flow size:

per flow S_F

per packet S_P

Large «Time» Heuristic

1. How do we evaluate these metrics in a simulation ?

per flow $S_F = \frac{1}{N} \sum_n S_n$

per packet $S_P = \frac{1}{P} \sum_p S_{F(p)}$

where $F(p) = n$ when packet p belongs to flow n

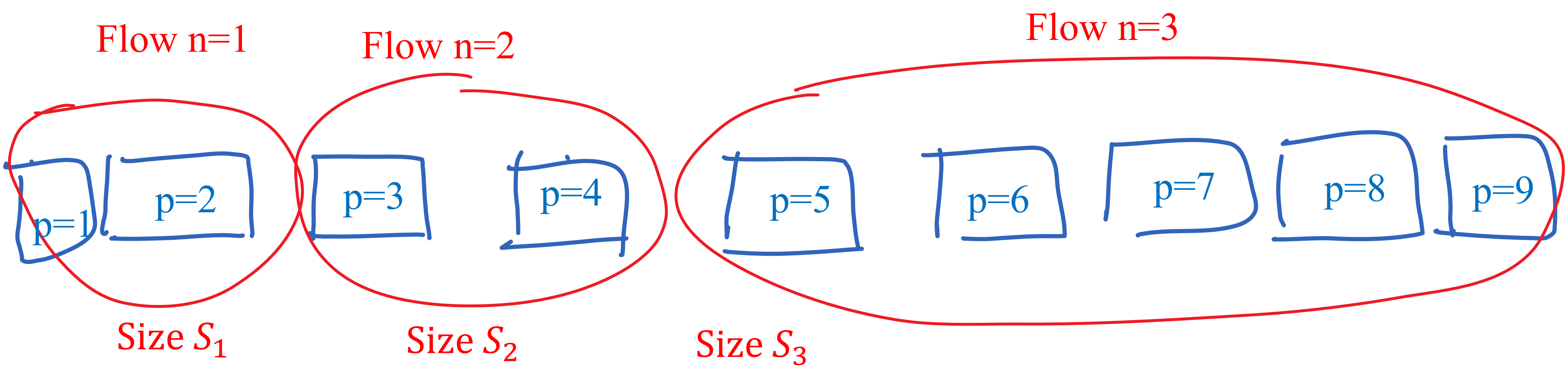
1. How do we evaluate these metrics in a simulation ?

   per flow $S_F = \frac{1}{N} \sum_n S_n$

   per packet $S_P = \frac{1}{P} \sum_p S_{F(p)}$

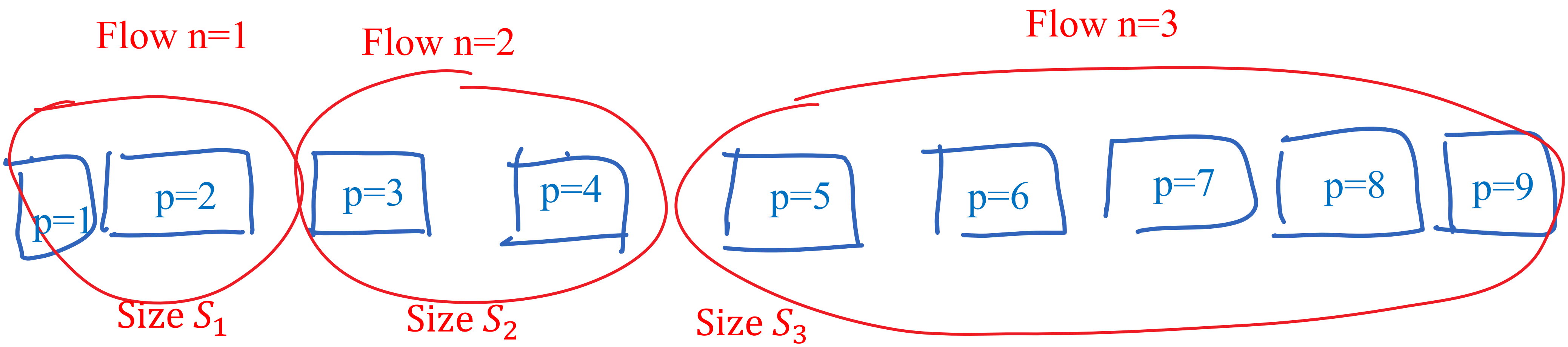
   where $F(p) = n$ when packet p belongs to flow n

2. Put the packets side by side, sorted by flow



$$S_P = \frac{1}{P}(S_1 + S_1 + S_2 + S_2 + S_3 + S_3 + S_3 + S_3 + S_3 + \cdots)$$

$$= \frac{1}{P}(S_1 \times S_1 + S_2 \times S_2 + S_3 \times S_3 + \cdots) = \frac{1}{P} \sum_n S_n^2$$



3. Compare

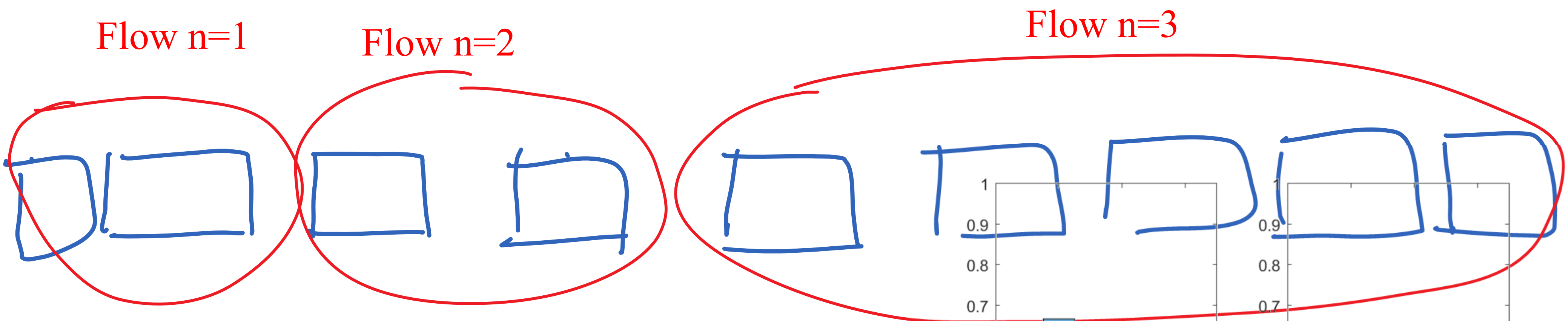
$$S_P = \frac{1}{P}\sum_n S_n^2$$

$$S_F = \frac{1}{N}\sum_n S_n = \frac{1}{N}P$$

$$S_P = \frac{N}{P} \times \frac{1}{N}\sum_n S_n^2 = \frac{1}{S_F} \times \frac{1}{N}\sum_n S_n^2 = \frac{1}{S_F} \times \left(S_F^2 + \mathrm{var_F}(S)\right)$$

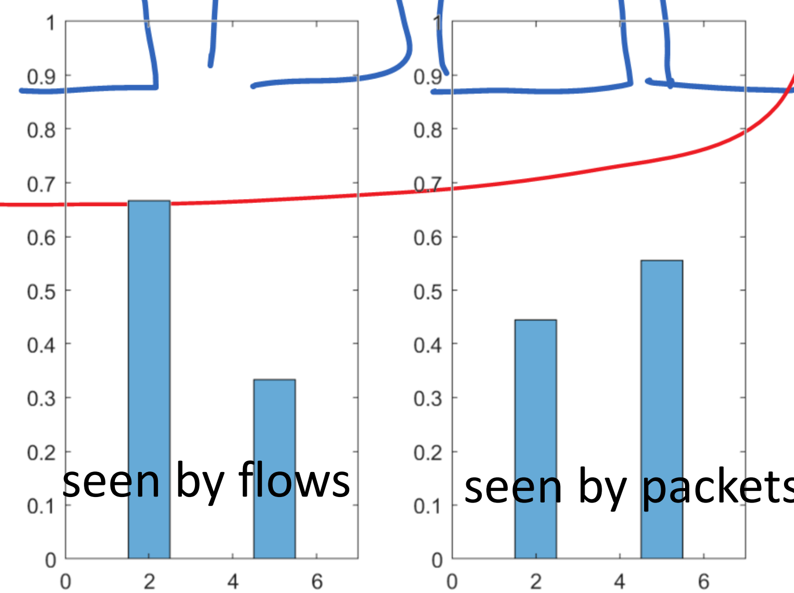
$$S_P = S_F + \frac{1}{S_F}\mathrm{var_F}(S)$$

PDFs of flow sizes



$f_F(s)$: PDF of flow size, seen by flows

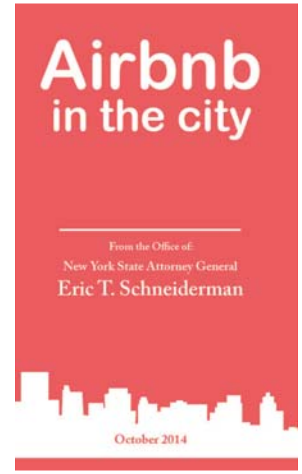
$f_P(s)$: PDF of flow size, seen by packets



Using the same approach we obtain

$f_P(s) = \eta s \, f_F(s)$ where η is a normalization constant

AirBnB’s Paradox



Occupancy PDF seen by an arbitrary object:
$f(s)$ = proportion of objects that are booked s nights per year

Occupancy PDF seen by an arbitrary traveller (estimated by insideairbnb.com) $f_T(s)$ = proportion of bookings that occur in a object booked s nights per year

A. $f_T(s) = \eta\, s\, f(s)$ where η is a normalizing constant

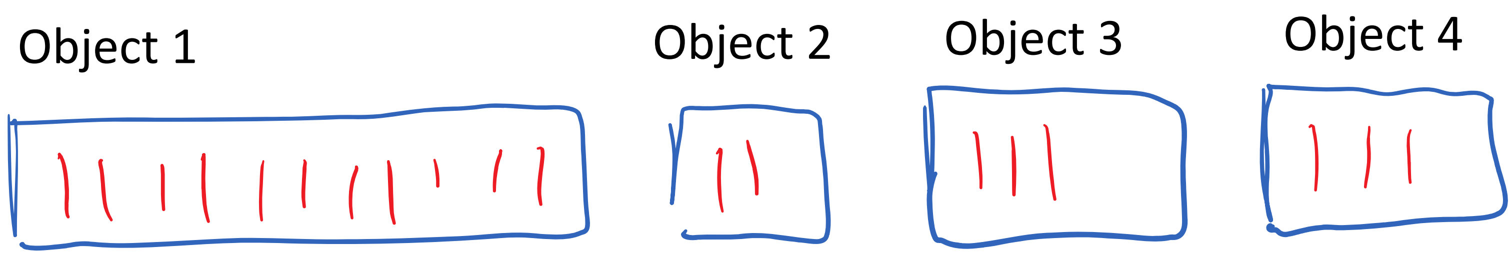
B. $f_T(s) \approx f(s)$

C. $f(s) = \eta\, s\, f_T(s)$ where η is a normalizing constant

D. I don’t know

Solution

This is the same case as with files (listings) and packets (bookings).



Therefore, with the same arguments $f_T(s) = \eta\ s\ f(s)$

The median of the distribution with PDF $f()$ is 40 days (reported by airbnb)

The median of the distribution with PDF $f_T()$ is 165 days (reported by insideairbnb.com)

An arbitrary booking is more likely to fall in a listing that is often booked.

Take-Home Message

How we sample data to compute a metric (the viewpoint) should be screened carefully.

Apparent paradoxes come from confusions in viewpoints.

Metrics may be misleading if sampling method is not appropriate.

Next we will see a formal theory (Palm Calculus) and its use in simulations.