IEEE INFOCOM '19

Internet Traffic Analysis:

On the Distribution of Traffic Volumes in the Internet and its Implications

Mohammed Alasmar
Department of Informatics
University of Sussex

George Parisis
Department of Informatics
University of Sussex

Richard Clegg
School of Computer Science
Queen Mary University of London

Nickolay Zakhleniuk
School of Computer Science
University of Essex

https://ieeexplore.ieee.org/document/8737483

1. Motivations
2. Main Goals
3. Methodology
4. Datasets
5. Power-law test
- Overview
- Likelihood ratio
- Anomalous traces
- Sampling times
- Corr. coeff. test
6. Use case 1
Link Dimensioning
7. Use case 2
Traffic billing

Motivations

- Reliable traffic modelling is important for network planning, deployment and management; e.g.

  (1) network dimensioning,

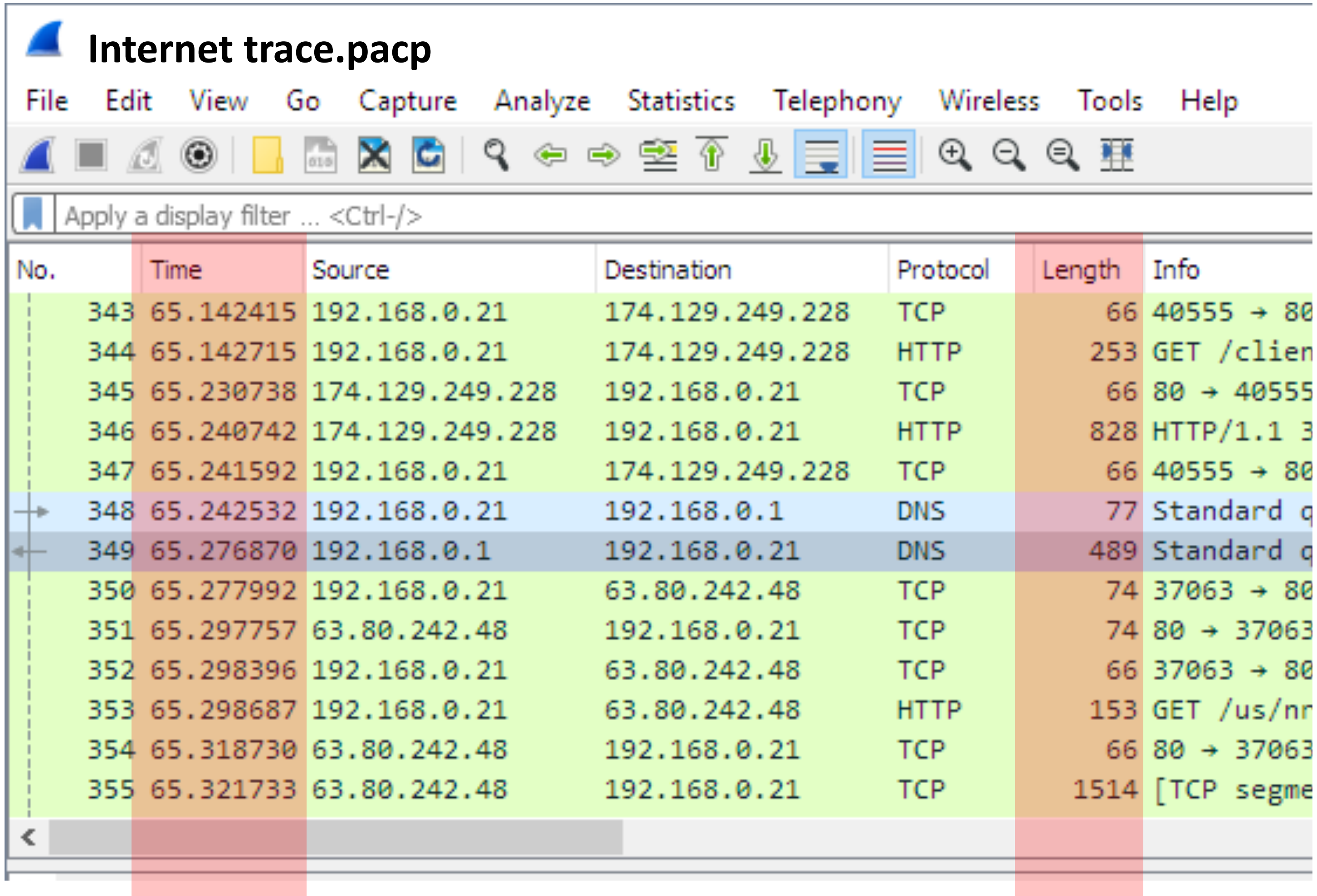
  (2) traffic billing.

- Historically, network traffic has been widely assumed to follow a **Gaussian distribution**.

- Deciding whether Internet flows could be **heavy-tailed** became important as this implies significant departures from **Gaussianity**.

Traffic volumes at different T

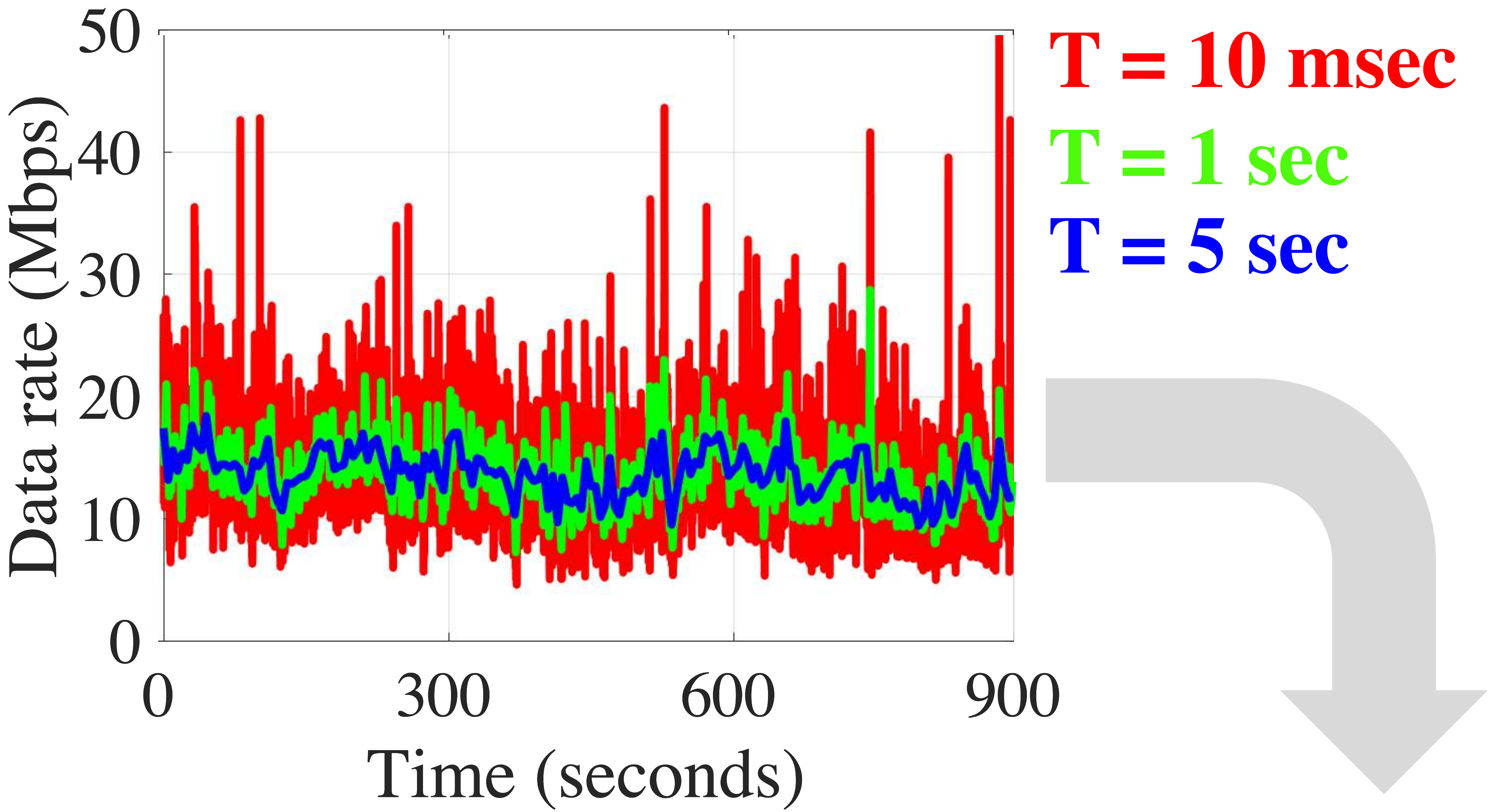
- X_i : the amount of traffic seen in the time period $[iT, (i+1)T)$



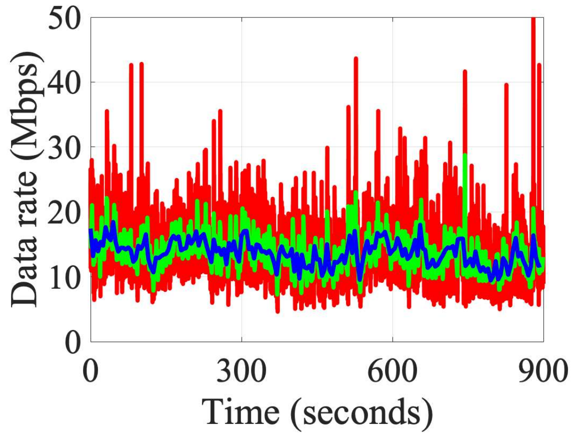
No.	Time	Source	Destination	Protocol	Length	Info
343	65.142415	192.168.0.21	174.129.249.228	TCP	66	40555 → 80
344	65.142715	192.168.0.21	174.129.249.228	HTTP	253	GET /clien
345	65.230738	174.129.249.228	192.168.0.21	TCP	66	80 → 40555
346	65.240742	174.129.249.228	192.168.0.21	HTTP	828	HTTP/1.1 3
347	65.241592	192.168.0.21	174.129.249.228	TCP	66	40555 → 80
348	65.242532	192.168.0.21	192.168.0.1	DNS	77	Standard q
349	65.276870	192.168.0.1	192.168.0.21	DNS	489	Standard q
350	65.277992	192.168.0.21	63.80.242.48	TCP	74	37063 → 80
351	65.297757	63.80.242.48	192.168.0.21	TCP	74	80 → 37063
352	65.298396	192.168.0.21	63.80.242.48	TCP	66	37063 → 80
353	65.298687	192.168.0.21	63.80.242.48	HTTP	153	GET /us/nr
354	65.318730	63.80.242.48	192.168.0.21	TCP	66	80 → 37063
355	65.321733	63.80.242.48	192.168.0.21	TCP	1514	[TCP segme

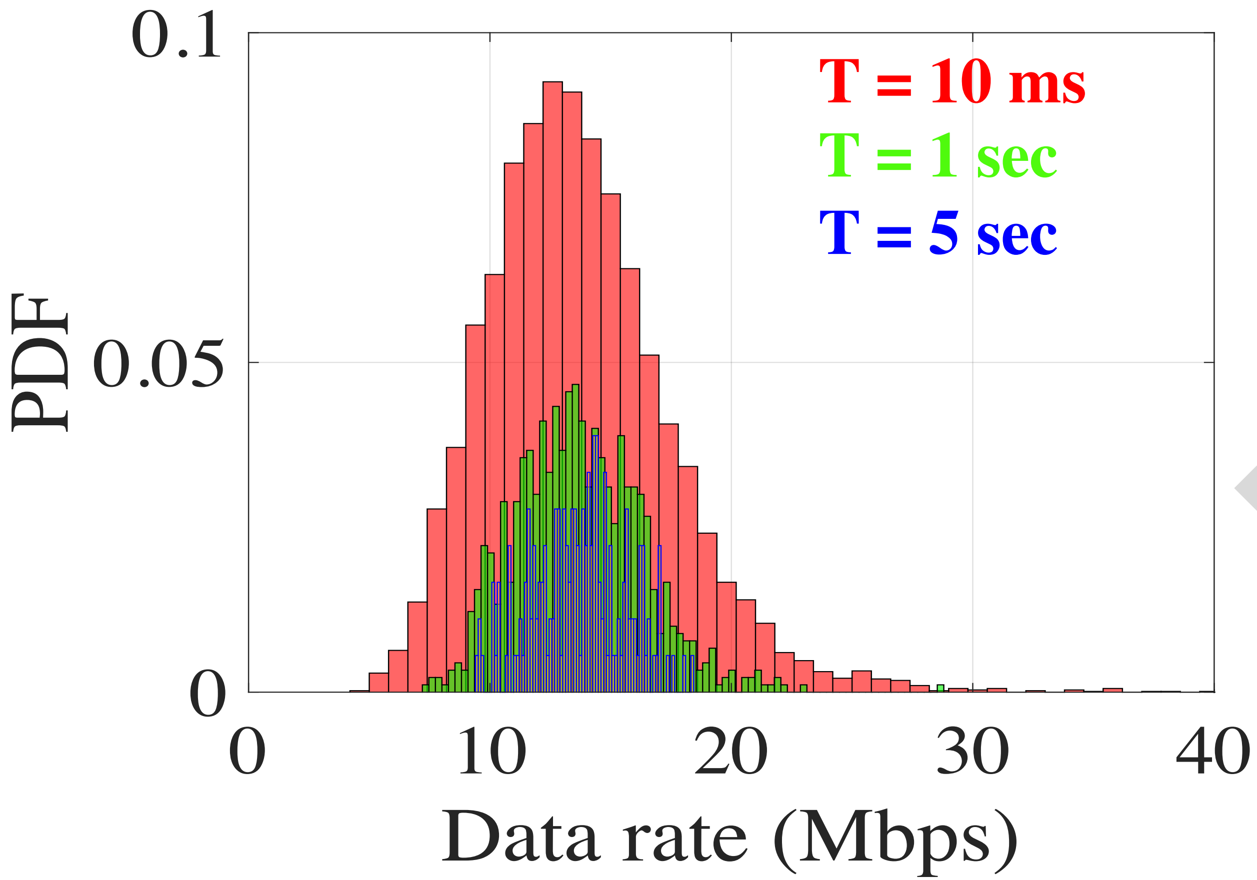
- Aggregation at different sampling times (T)

Traffic volumes at different T



Goal



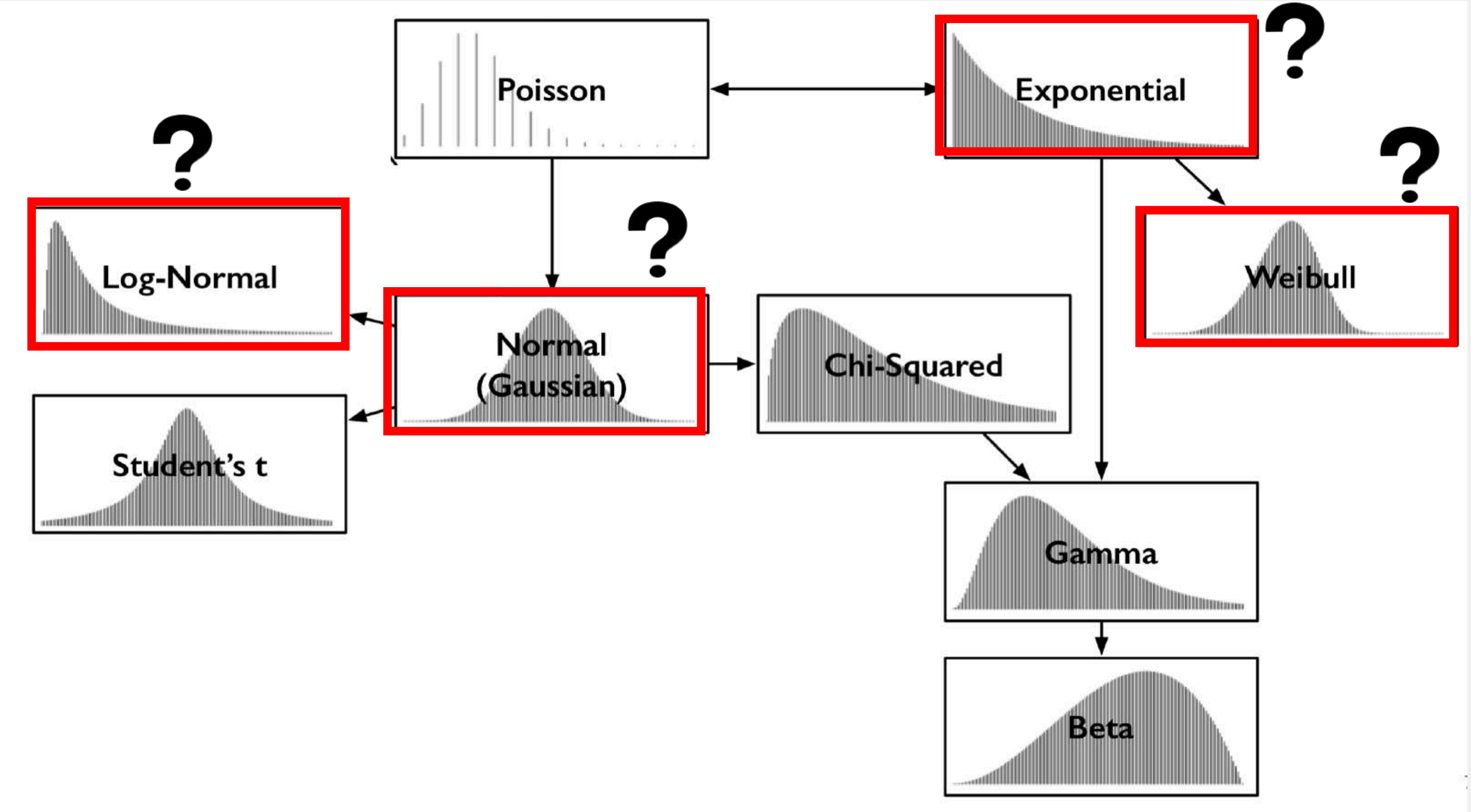


- Investigating the distribution of the amount of traffic per unit time using a robust statistical approach.

Goal

- Investigating the distribution of the amount of traffic per unit time using a robust statistical approach.

Goal



Poisson
Exponential
?
Log-Normal
?
Normal
(Gaussian)
?
Chi-Squared
Weibull
?
Student's t
Gamma
Beta

Datasets

- We study a large number of traffic traces (230) from many different networks: 2009 → 2018

Dataset	#Traces
Twente[1]	40
MAWI[2]	107
Auckland[3]	25
Waikato[4]	30
Caida[5]	27

[1] https://www.simpleweb.org/wiki/index.php/Traces , 2009.
[2] http://mawi.wide.ad.jp/mawi/ , 2016-2018.
[3] https://wand.net.nz/wits/auck/9/ , 2009.
[4] https://wand.net.nz/wits/waikato/8/ , 2010-2011.
[5] http://www.caida.org/data/overview/ , 2016.

Power-law test

- Our analysis is based on the framework proposed in:

> Power-law distributions in empirical data
> A Clauset, CR Shalizi, MEJ Newman - SIAM review, 2009 - SIAM
> Cited by 6865

- The framework combines **maximum-likelihood** fitting methods with **goodness-of-fit** tests based on the **K**olmogorov–**S**mirnov statistic and likelihood ratios.

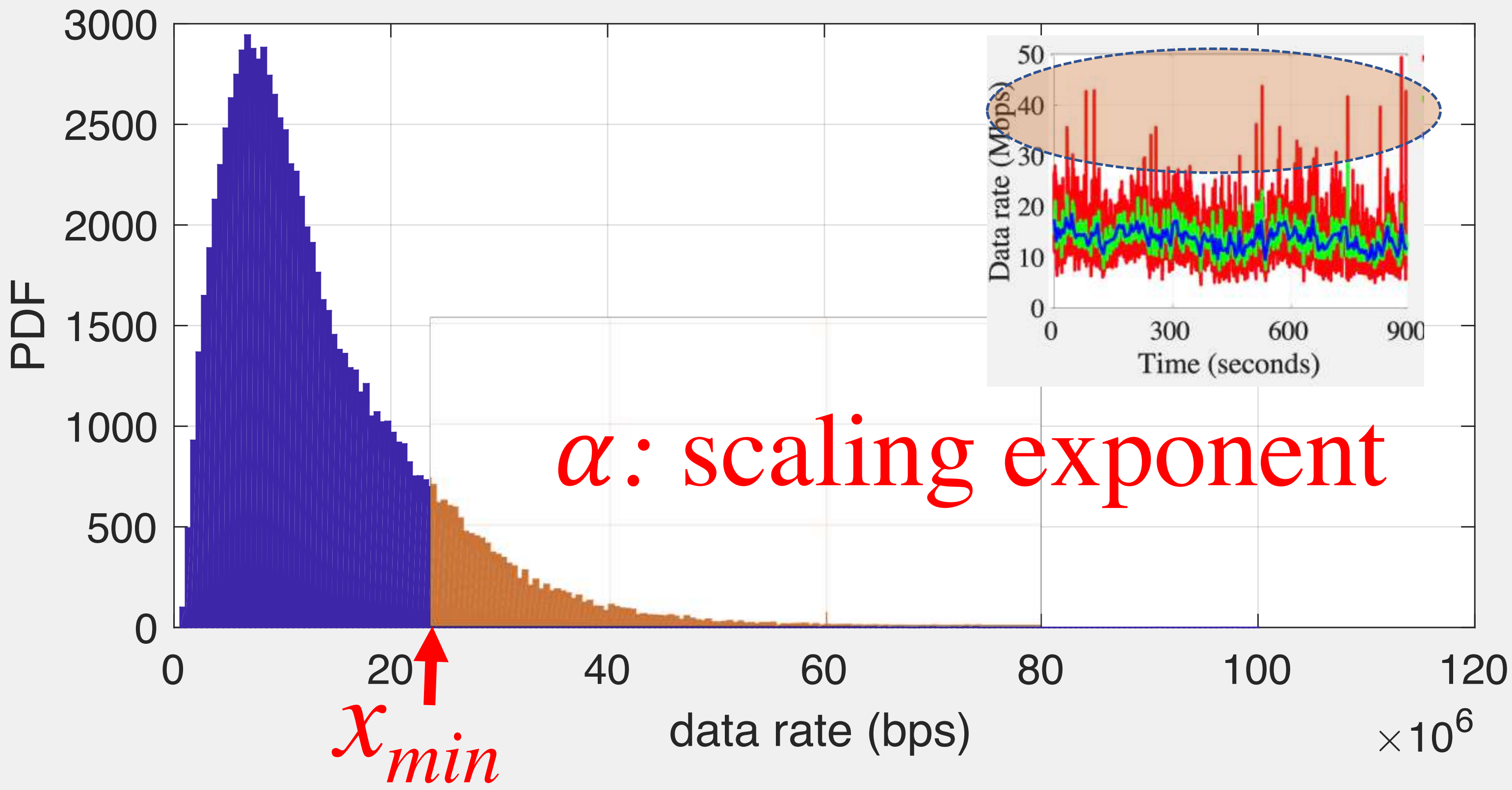
Power-law test

Power-law distribution:

$$p(x) = (x)^{-\alpha}$$

introduce new variables

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$$



Power-law test

name	distribution $p(x) = Cf(x)$: $f(x)$		C
power law	$x^{-\alpha}$	→	$(\alpha-1)x_{\min}^{\alpha-1}$
power law with cutoff	$x^{-\alpha}e^{-\lambda x}$	→	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$
exponential	$e^{-\lambda x}$	→	$\lambda e^{\lambda x_{\min}}$
stretched exponential	$x^{\beta-1}e^{-\lambda x^{\beta}}$	→	$\beta\lambda e^{\lambda x_{\min}^{\beta}}$
log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$		$\sqrt{\frac{2}{\pi\sigma^2}}\left[\operatorname{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$

Likelihood Ratio: $\boldsymbol{R}$

$$\boldsymbol{R}, p \ = fit.\boldsymbol{distributionCompare}(powerlaw, alternative)$$

- **Weibull**
- **Lognormal**
- **Exponential**

Likelihood ratio:

$$\boldsymbol{R} = \frac{L_1}{L_2} = \frac{\prod_{i=1}^{n} p_1(x)}{\prod_{i=1}^{n} p_2(x)}$$

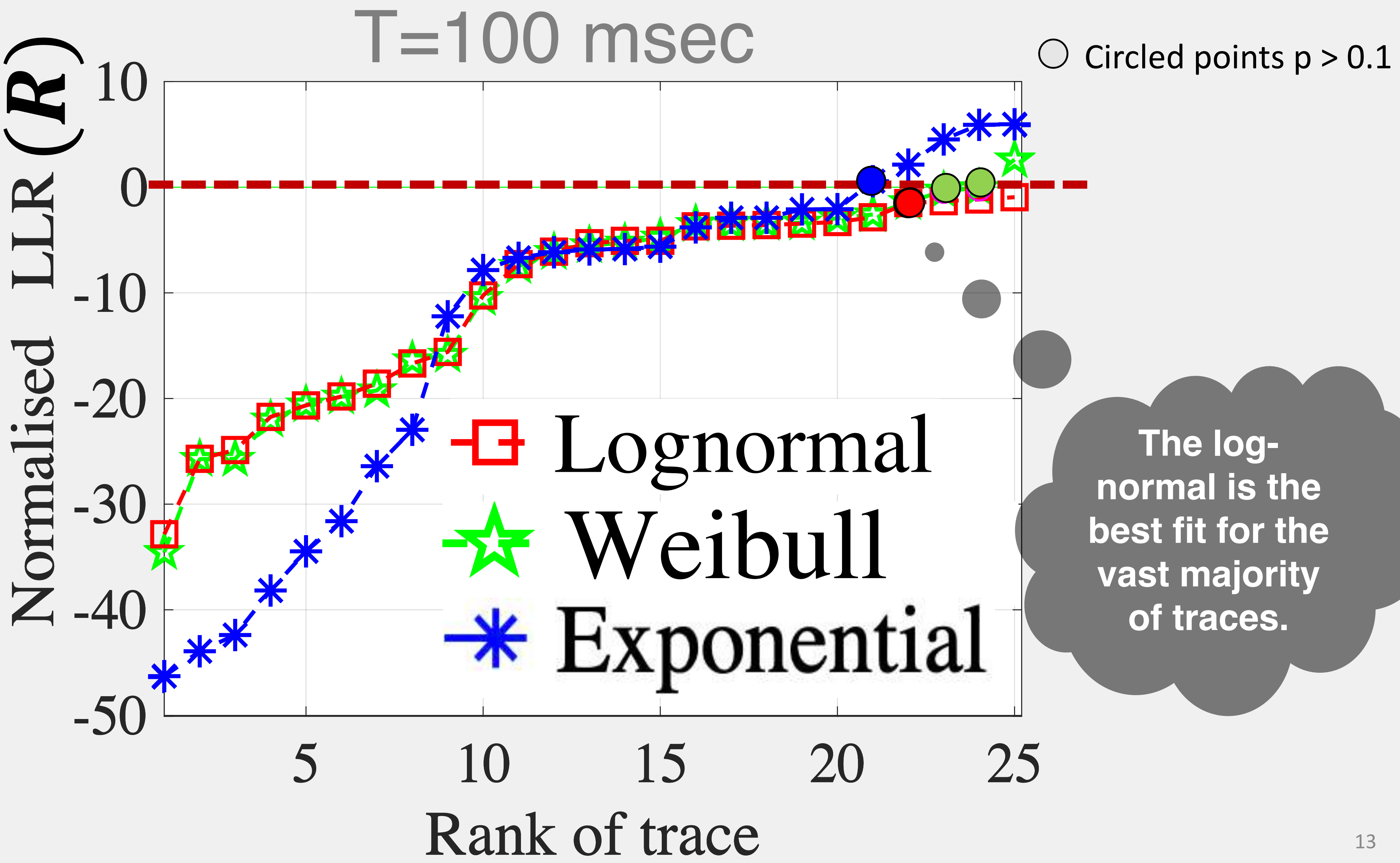
$\prod_{i=1}^{n} p_1(x)$ → power-law likelihood function

$\prod_{i=1}^{n} p_2(x)$ → alternative likelihood function

- $\boldsymbol{Log}$−Likelihood ratio: $\boldsymbol{R}$

- If $\boldsymbol{R} > 0$, then the power-law is favoured.
- If $\boldsymbol{R} < 0$, then the alternative is favoured.
- If $p < 0.1$, then the value of $\boldsymbol{R}$ can be trusted.

Normalised Log-Likelihood Ratio (LLR)



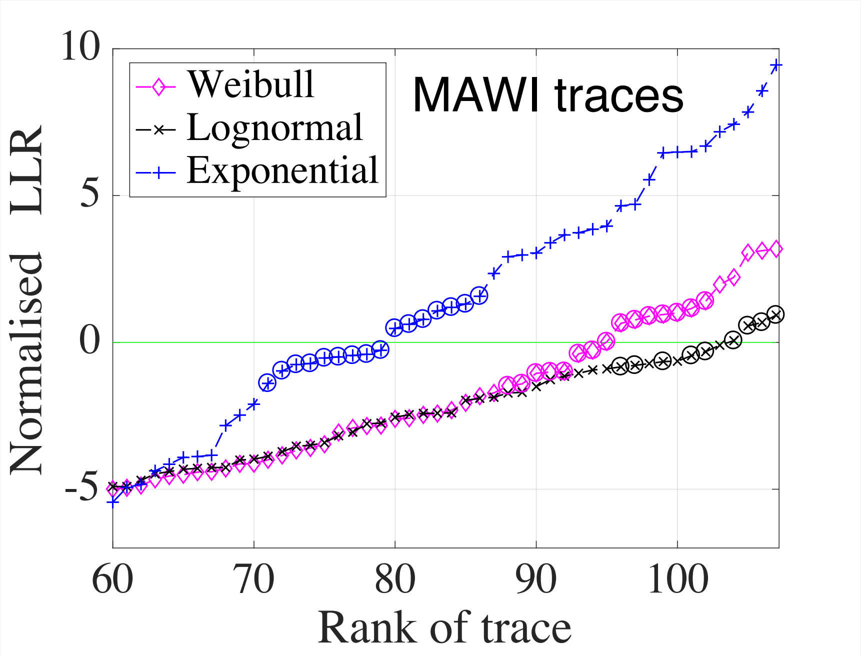
The log-normal is the best fit for the vast majority of traces.

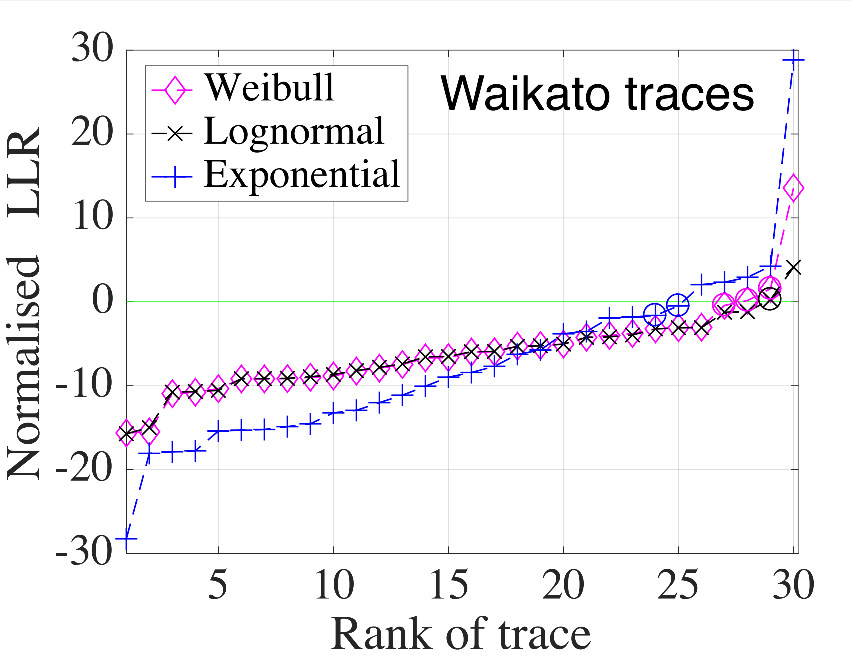
The log-normal distribution is the best fit for the vast majority of traces.

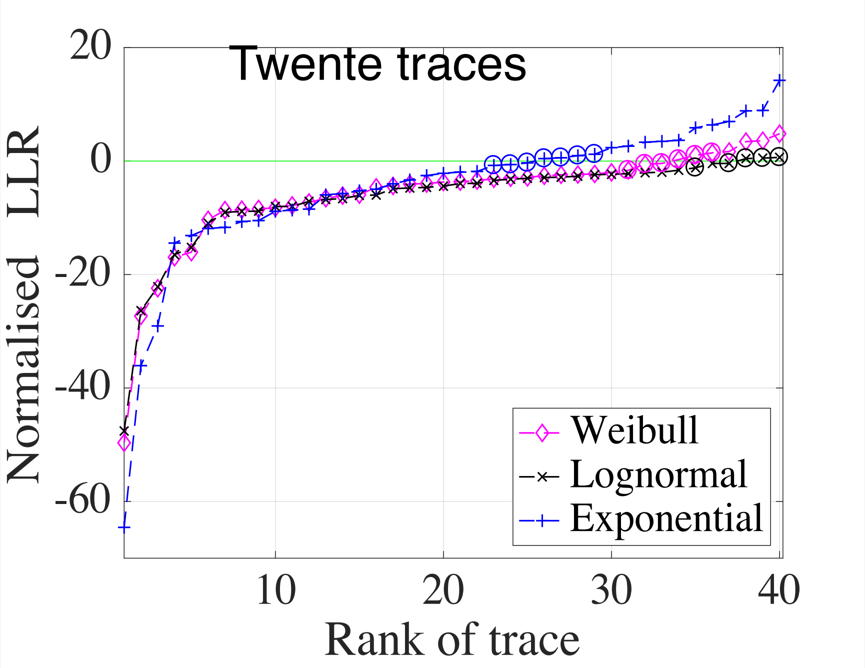
The log-normal distribution is not the best fit for …

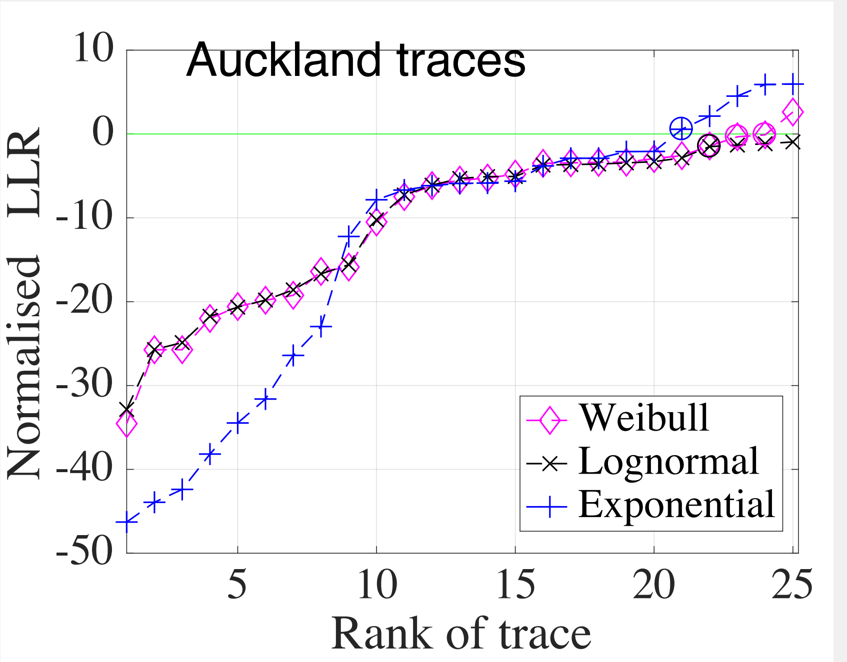
- **1 out of 27 CAIDA traces**
- **9 out of 107 MAWI traces**
- **2 out of 30 Waikato traces**
- **5 out of 40 Twente traces**
- **1 out of 25 Auckland traces**

Anomalous traces



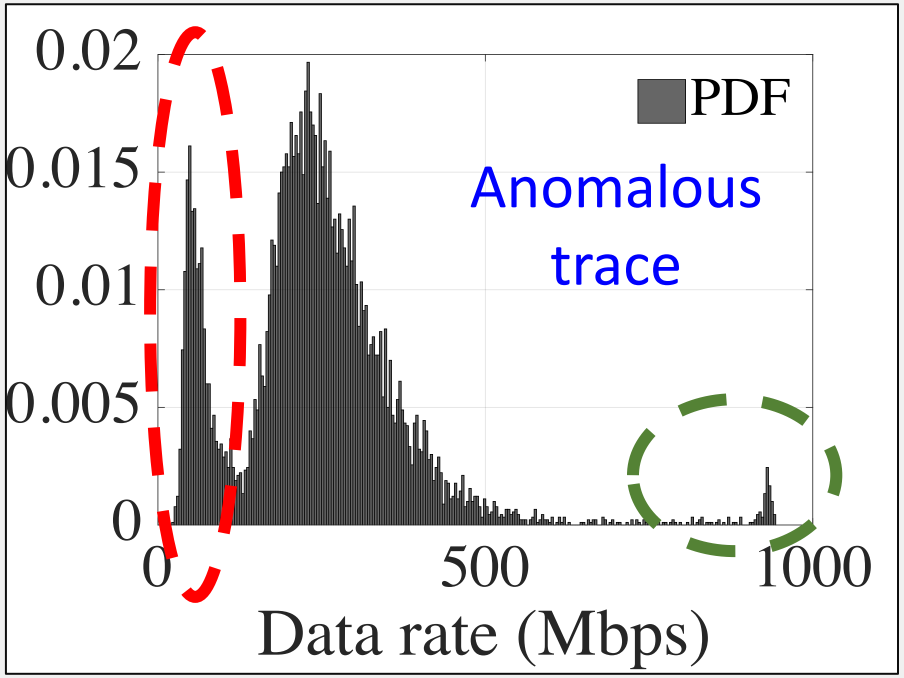


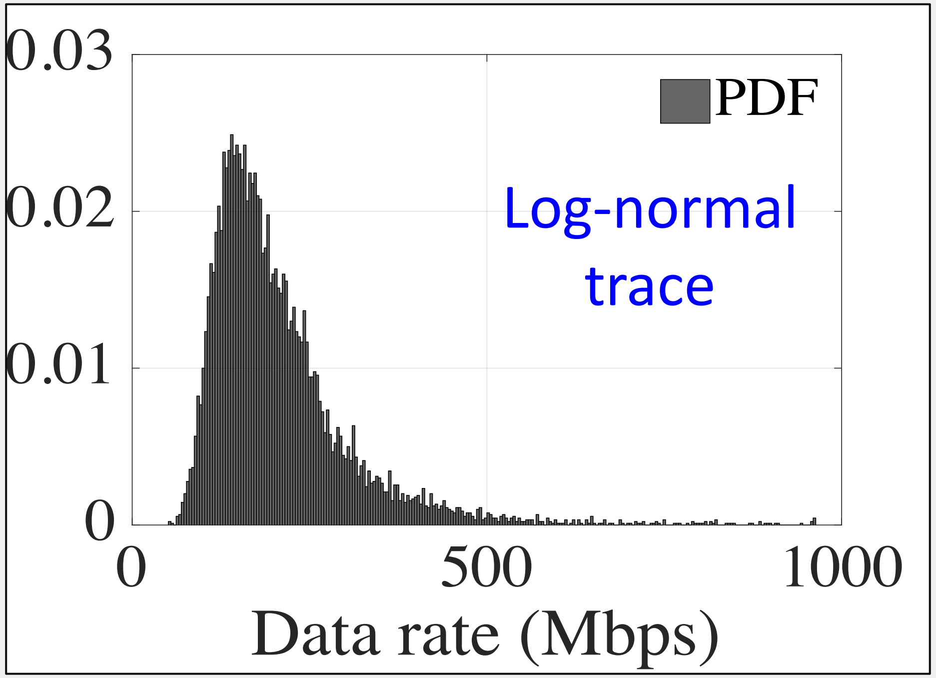




Anomalous traces

- Anomalous traces are a poor fit for all distributions tried.
- This is often due to traffic outages or links that hit maximum capacity.

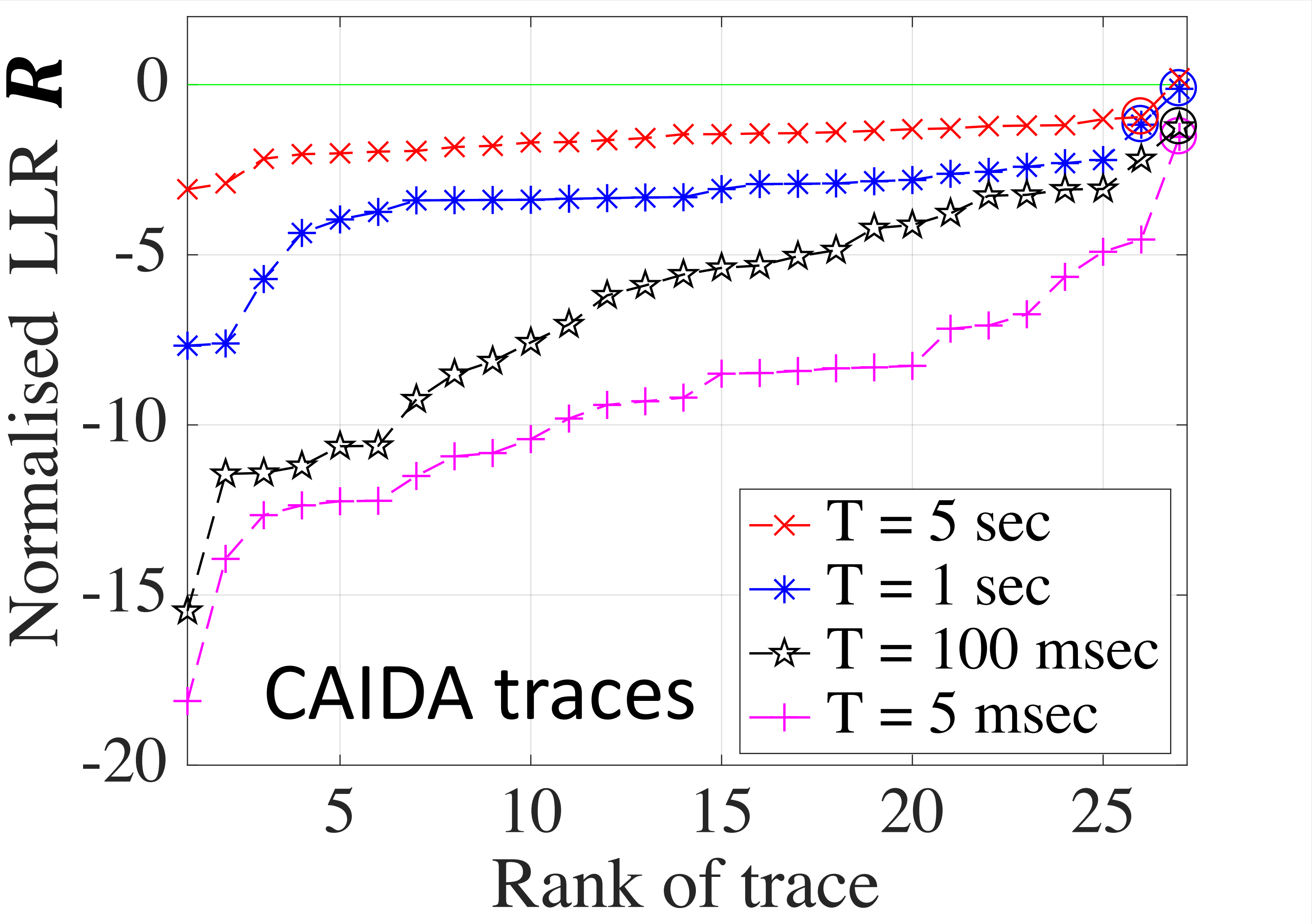




At different sampling times: T

Normalised Log-Likelihood Ratio (LLR) test results for all studied traces and log-normal distribution at different timescales

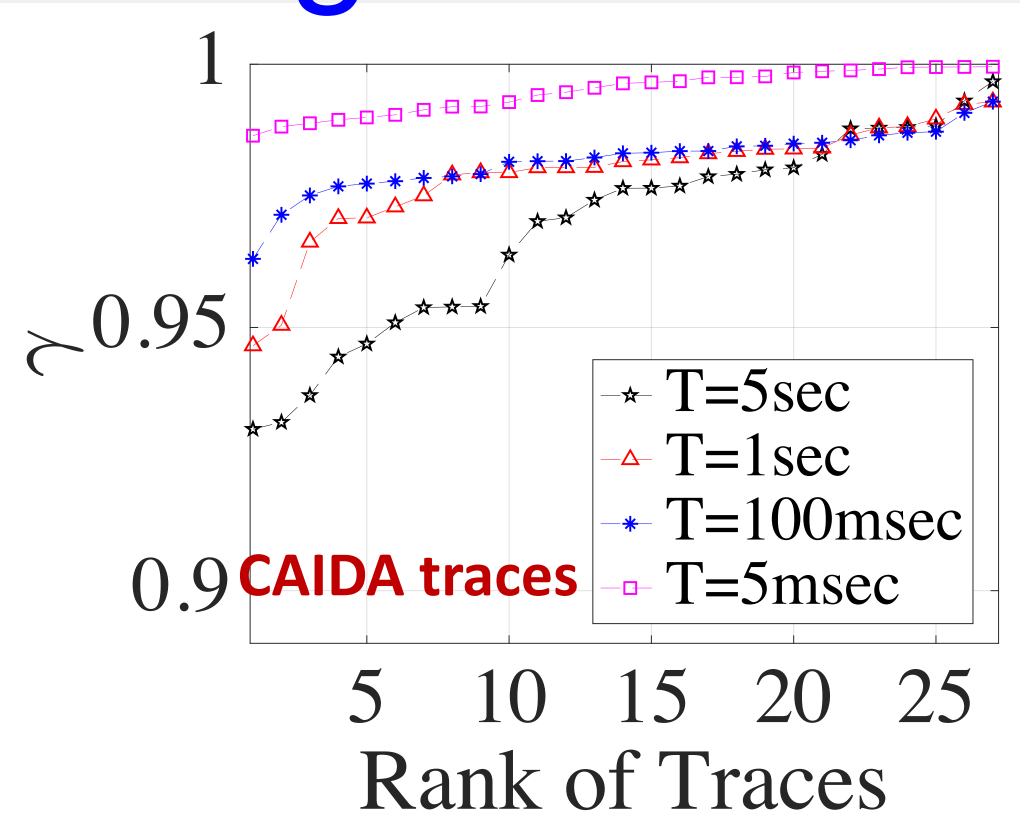
$\boldsymbol{R} < 0$, i.e., log-normal is favoured.



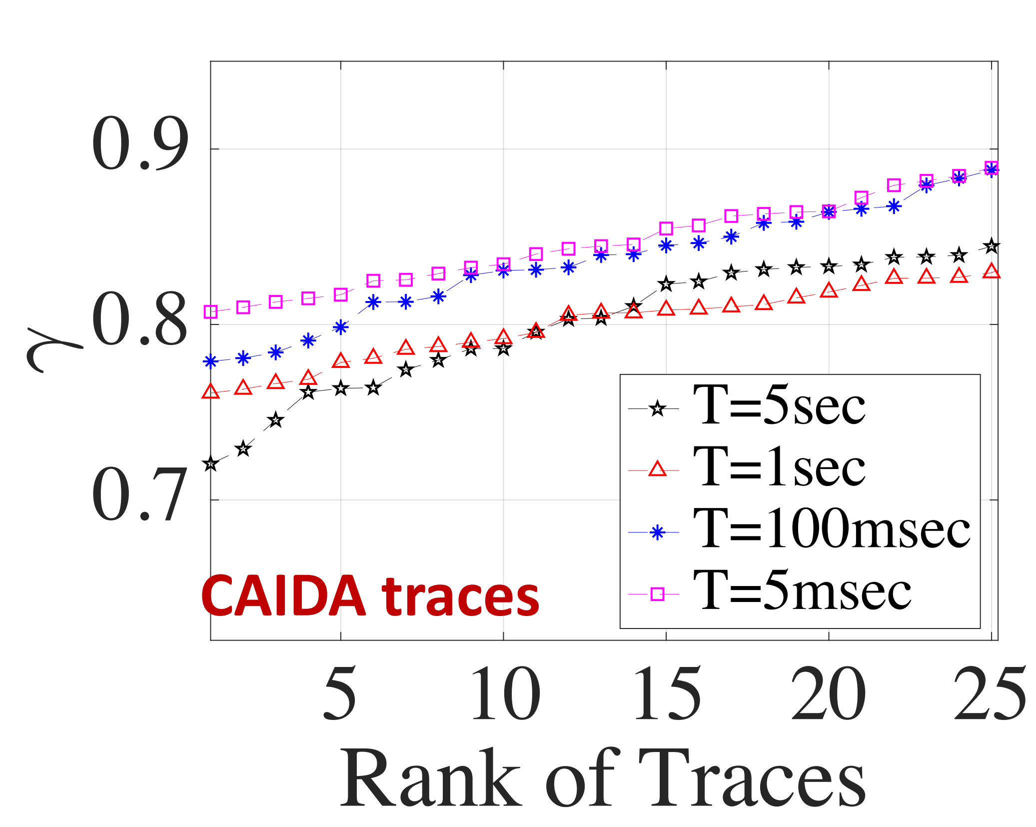
The correlation coefficient test

- Strong goodness-of-fit (GOF) is assumed to exist when the value of γ is greater than 0.95.

Log-normal

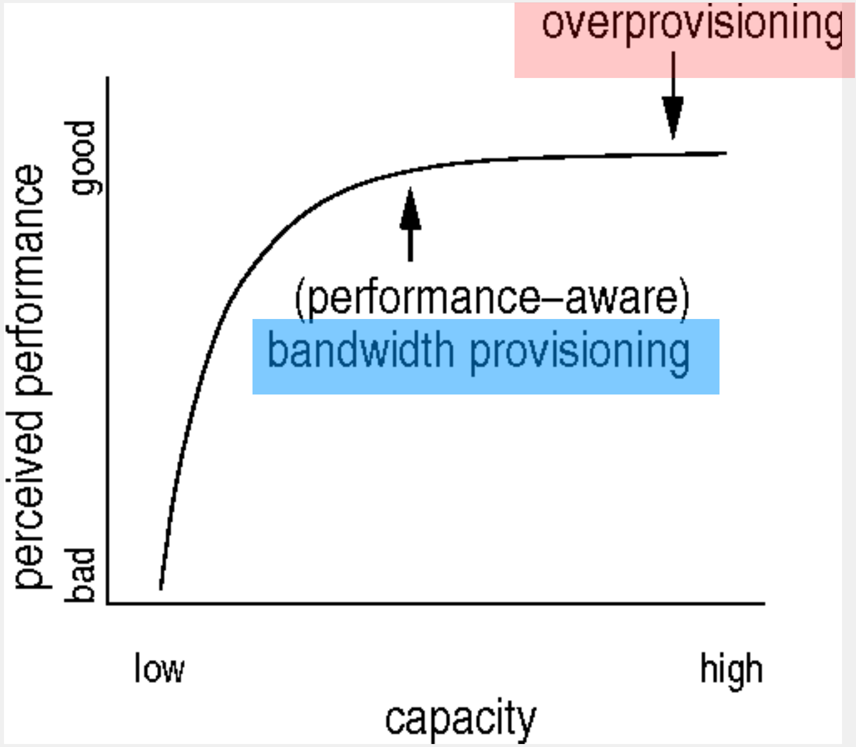


Gaussian



Use case 1: Bandwidth provisioning

- **Bandwidth provisioning** approach provides the link by the essential bandwidth that guarantees the required performance.
- **Overprovisioning**. In the conventional methods the bandwidth is allocated by up-grading the link bandwidth to 30% of the average traffic value.



Use case 1: Bandwidth provisioning

- The following inequality (the '**link transparency formula**') has been used for bandwidth provisioning:

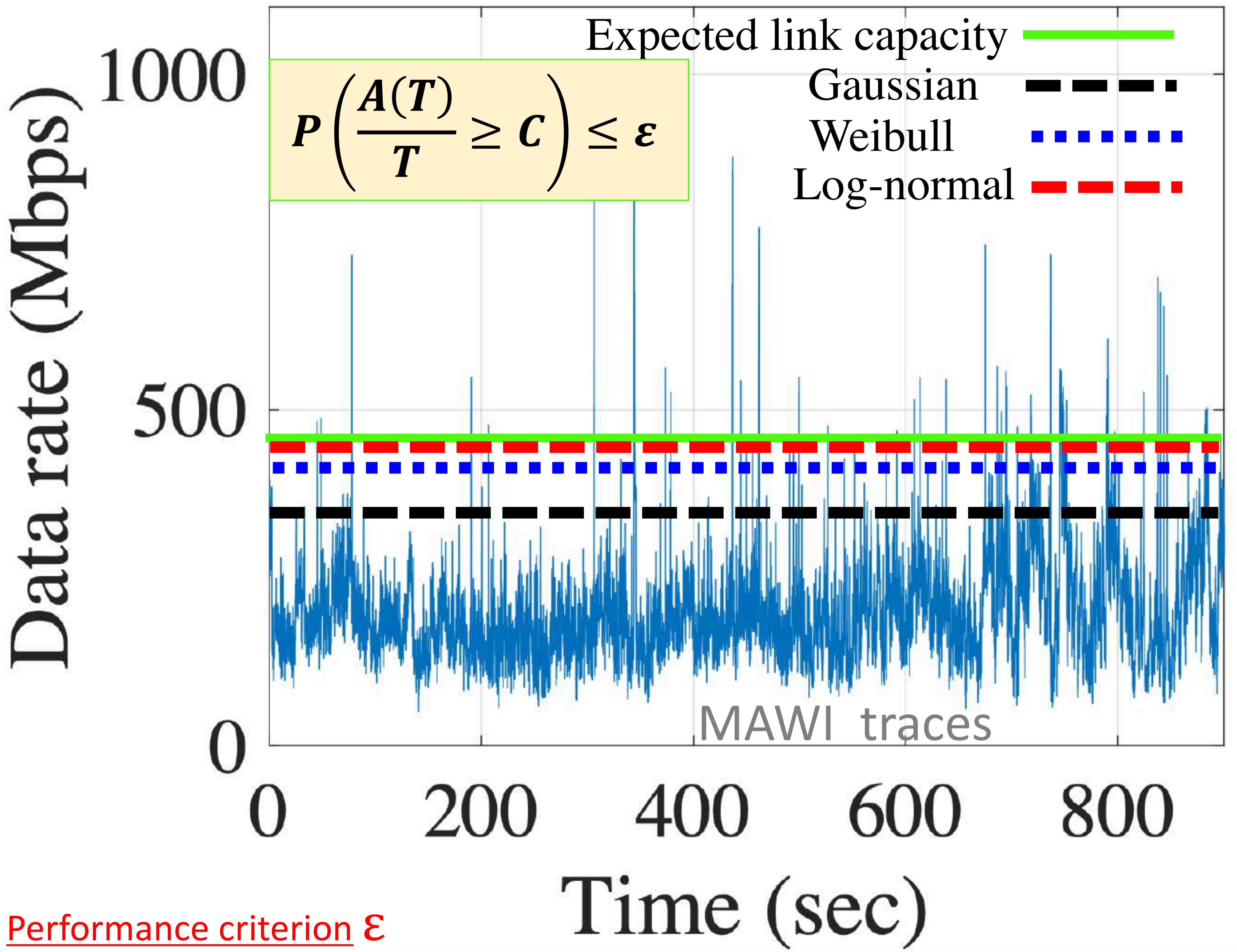
$$P\left(\frac{A(T)}{T} \geq C\right) \leq \varepsilon$$

i.e., the probability that the captured traffic $A(T)$ over a specific aggregation timescale T is larger than the link capacity C has to be smaller than the value of a performance criterion ε.

✓ **ε** has to be chosen carefully by the network provider in order to meet the specified SLA.

Use case 1: Bandwidth provisioning

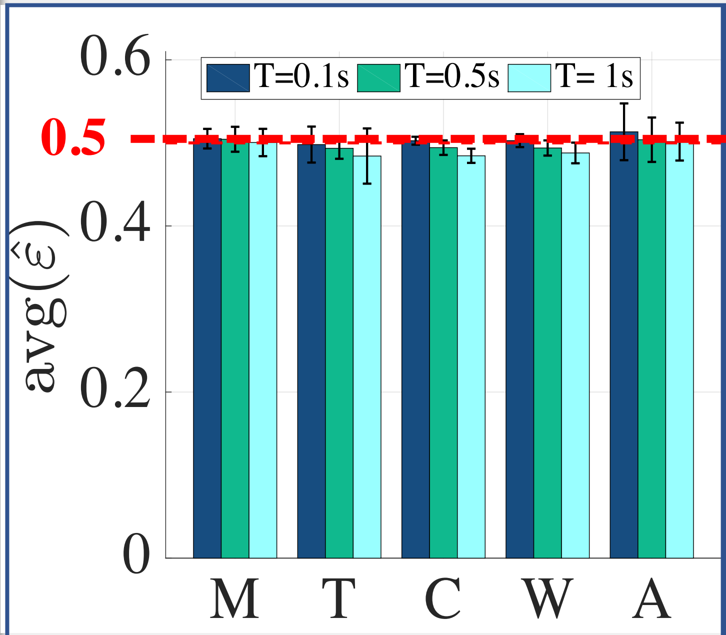
Example: $\varepsilon = 0.01$



$$P\left(\frac{A(T)}{T} \geq C\right) \leq \varepsilon$$

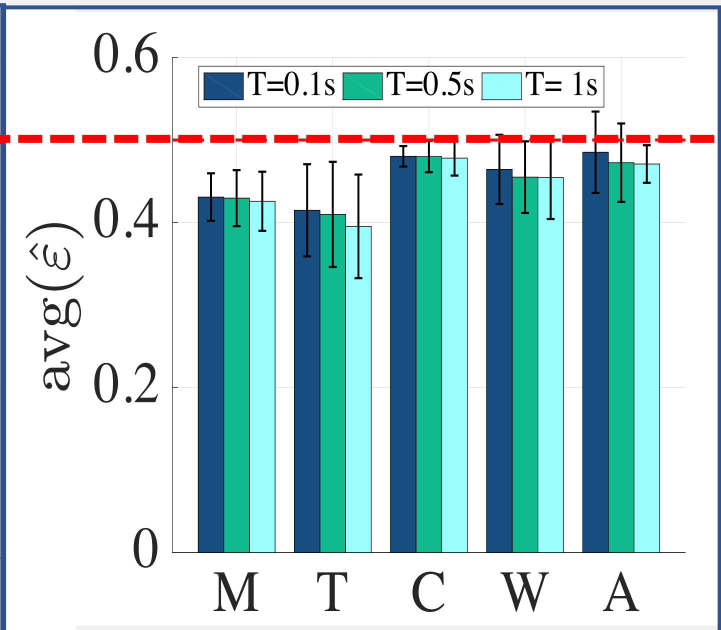
Performance criterion ε

Bandwidth provisioning: Results



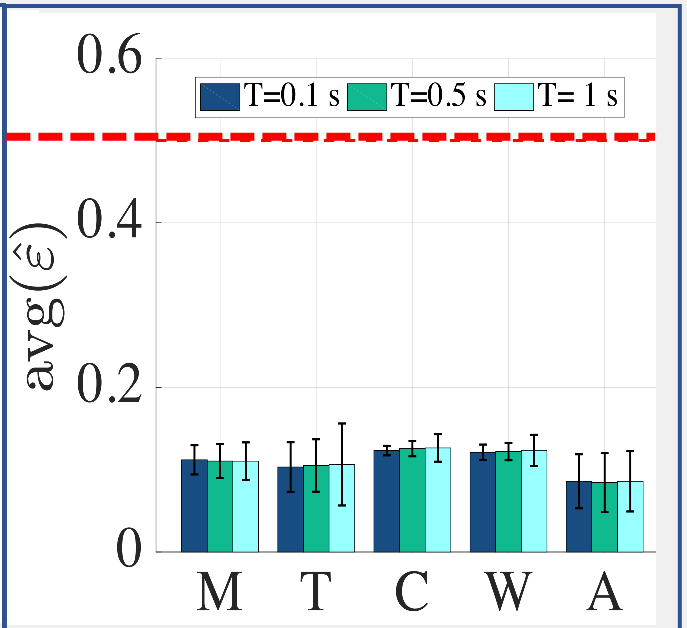
Target: ε = 0.5

Log-normal



Target: ε = 0.5

Weibull



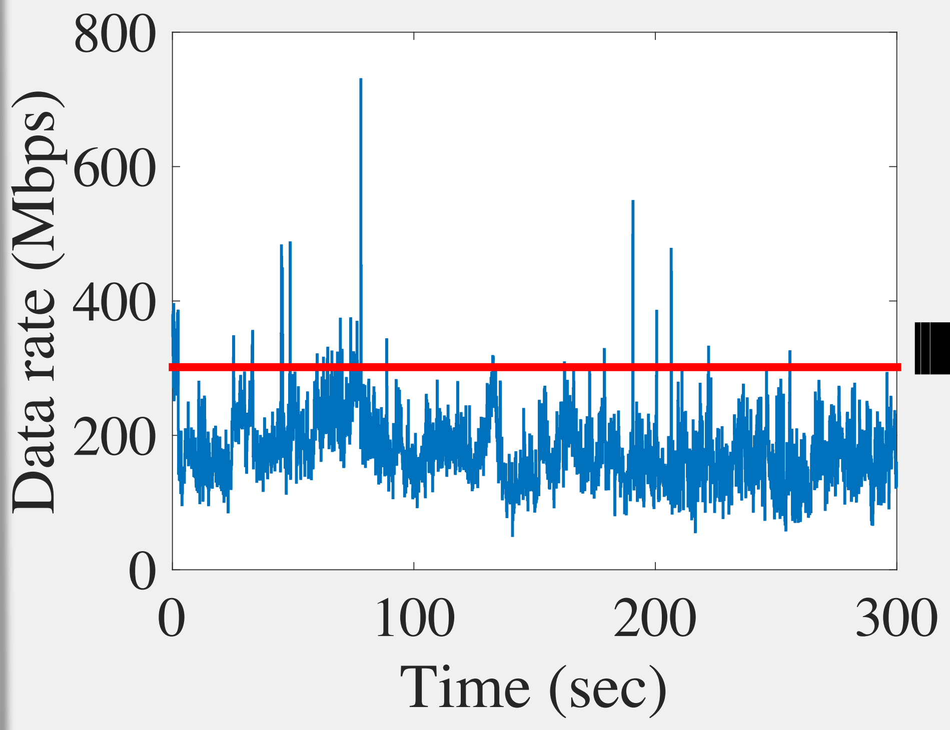
Target: ε = 0.5

Gaussian

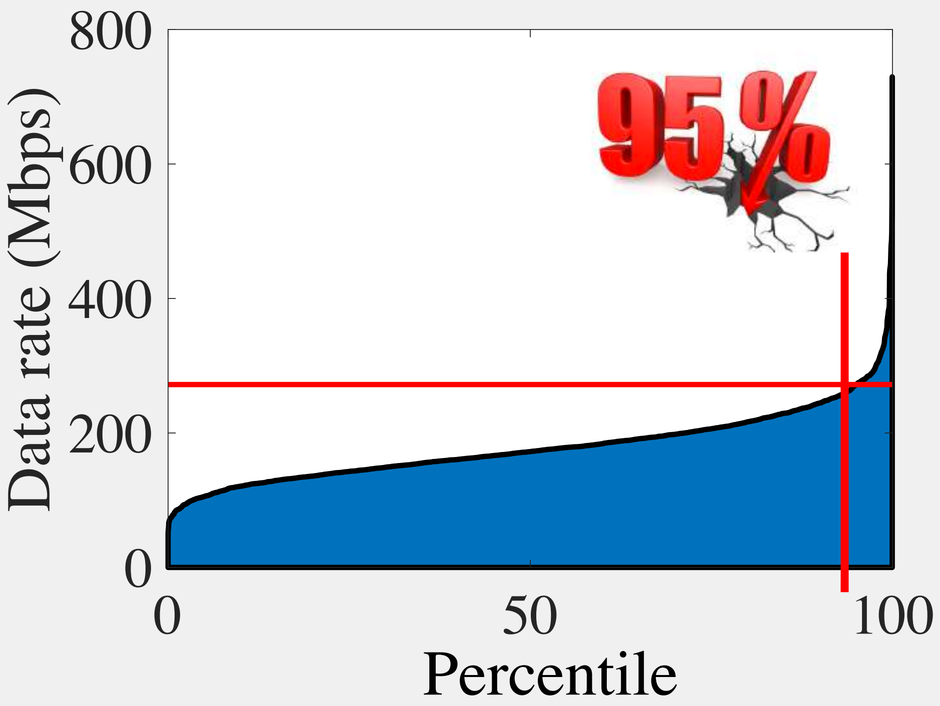
M: MAWI, T: Twente, C: CAIDA, W: Waikato, A: Auckland

Use case 2: 95th percentile pricing

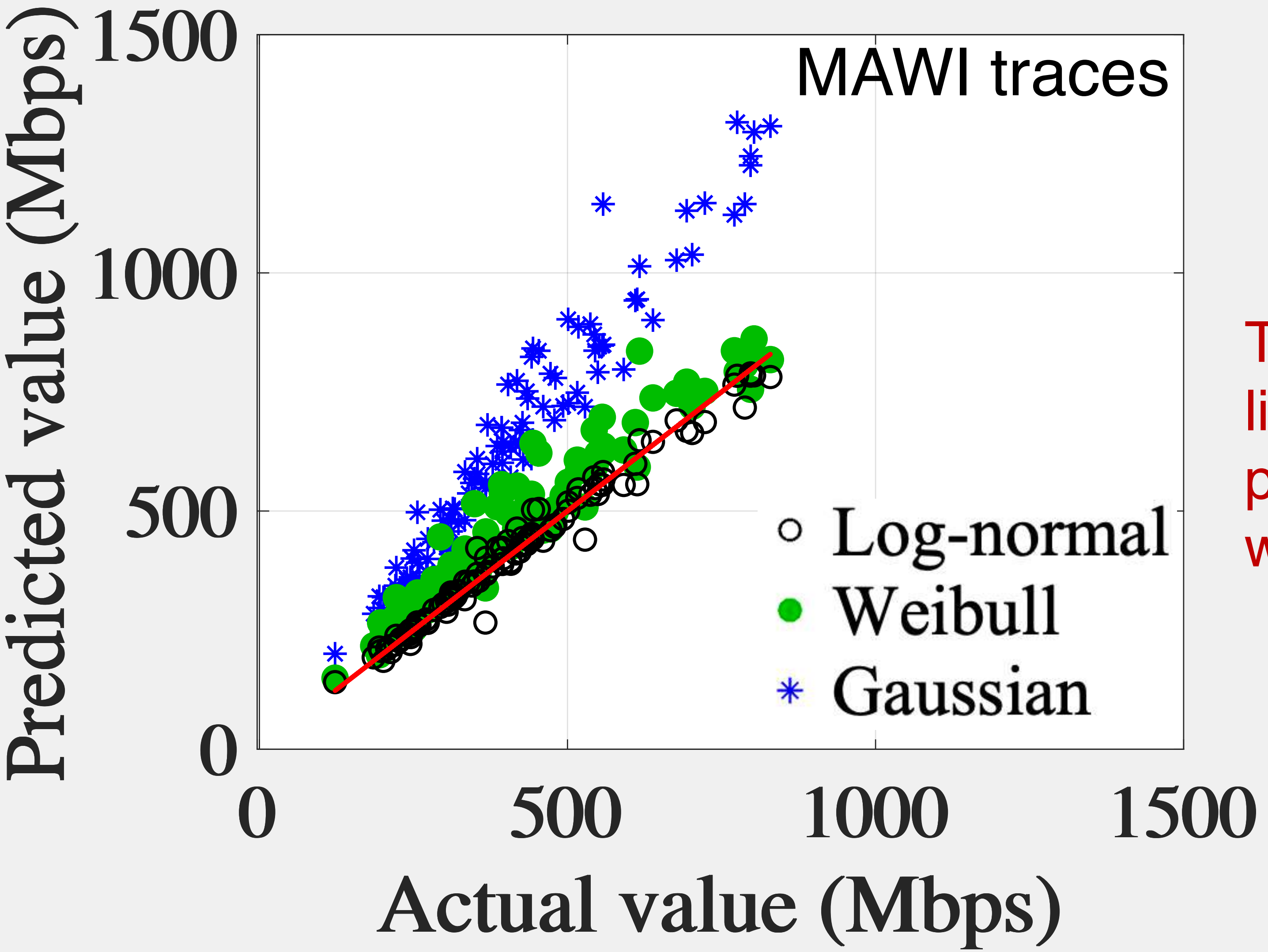
- Customers are not billed for brief spikes in network traffic.



[5 minutes]



95th percentile pricing: Results

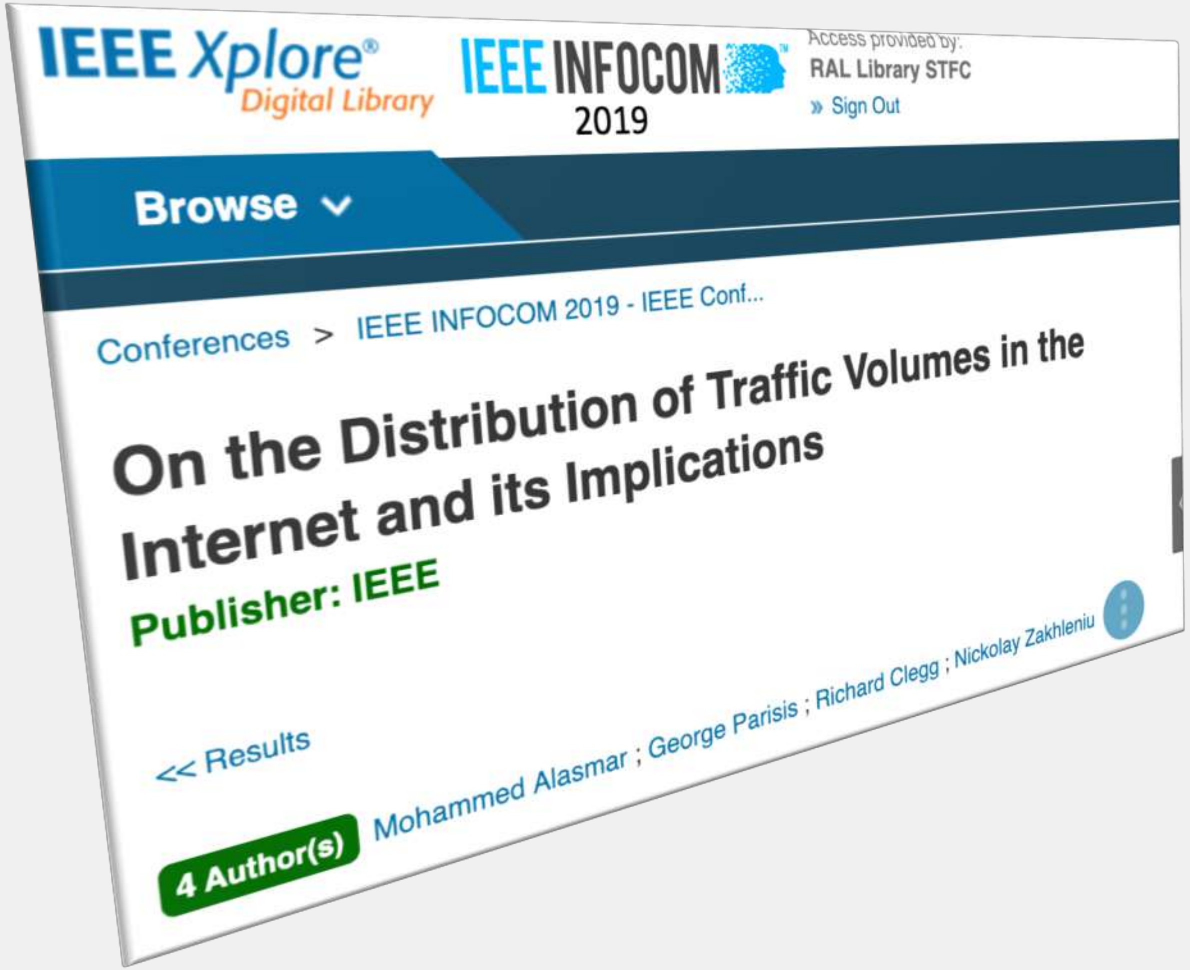


The red reference line to show where perfect predictions would be located.

- Log-normal model provides much more accurate predictions of the 95th percentile.

1. Motivations
2. Main Goals
3. Methodology
4. Datasets
5. Power-law test
- Overview
- Likelihood ratio
- Anomalous traces
- Sampling times
- Corr. coeff. test
6. Use case 1
Link Dimensioning
7. Use case 2
Traffic billing

More details ….



Thanks! Questions?

Summary

- The distribution of traffic on Internet links is an important problem that has received relatively little attention.

- We use a well-known, state-of-the-art statistical framework to investigate the problem using a large corpus of traces.

- We investigated the distribution of the amount of traffic observed on a link in a given (small) aggregation period which we varied from 5 msec to 5 sec.

- The vast majority of traces fitted the lognormal assumption best and this remained true all timescales tried.

- We investigate the impact of the distribution on two sample traffic engineering problems.

    1. Firstly, we looked at predicting the proportion of time a link will exceed a given capacity.

    2. Secondly, we looked at predicting the 95th percentile transit bill that ISP might be given.

- For both of these problems the log-normal distribution gave a more accurate result than heavy-tailed distribution or a Gaussian distribution.

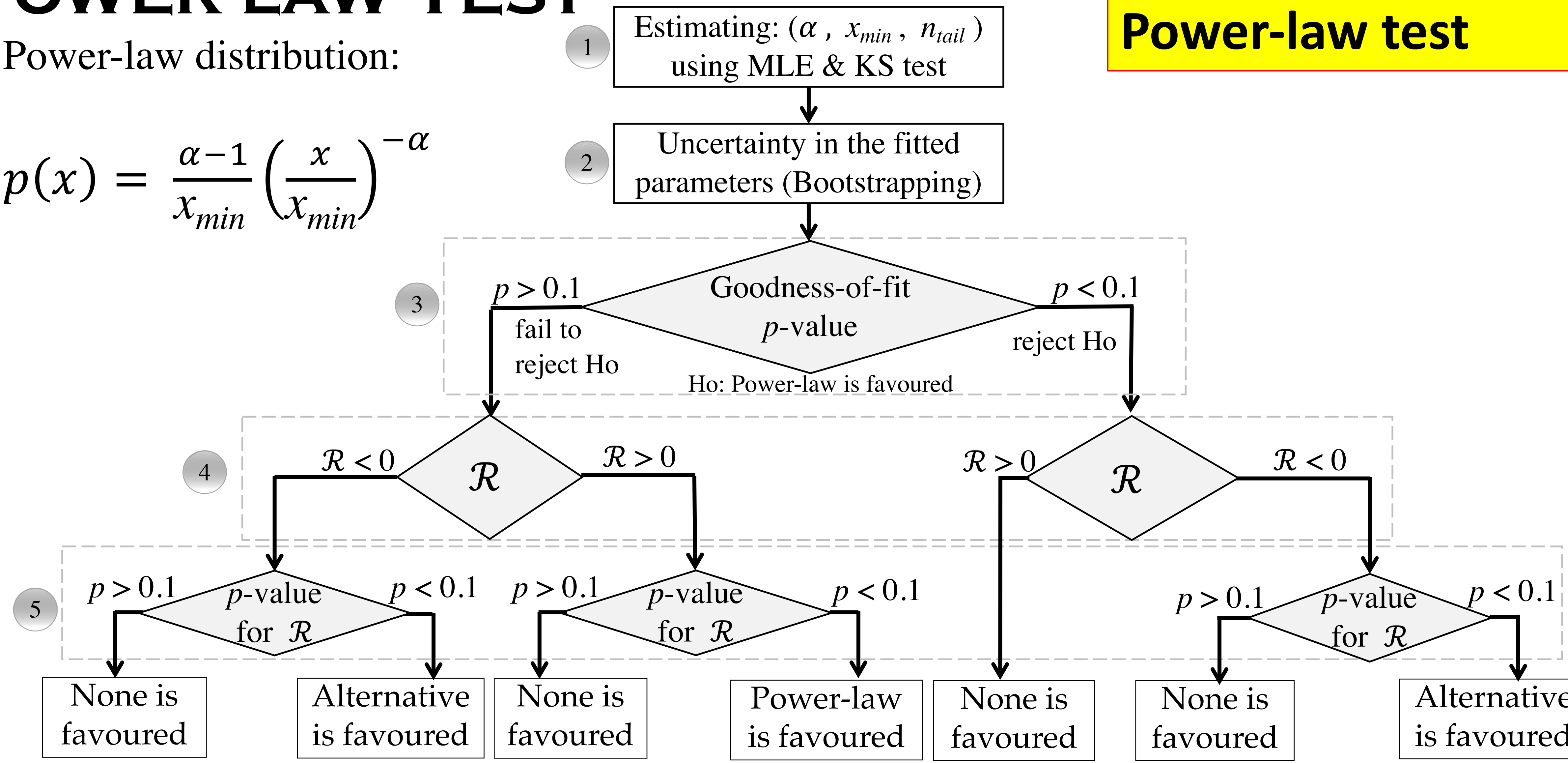
Backup ……

Power-law Test

Power-law distribution:

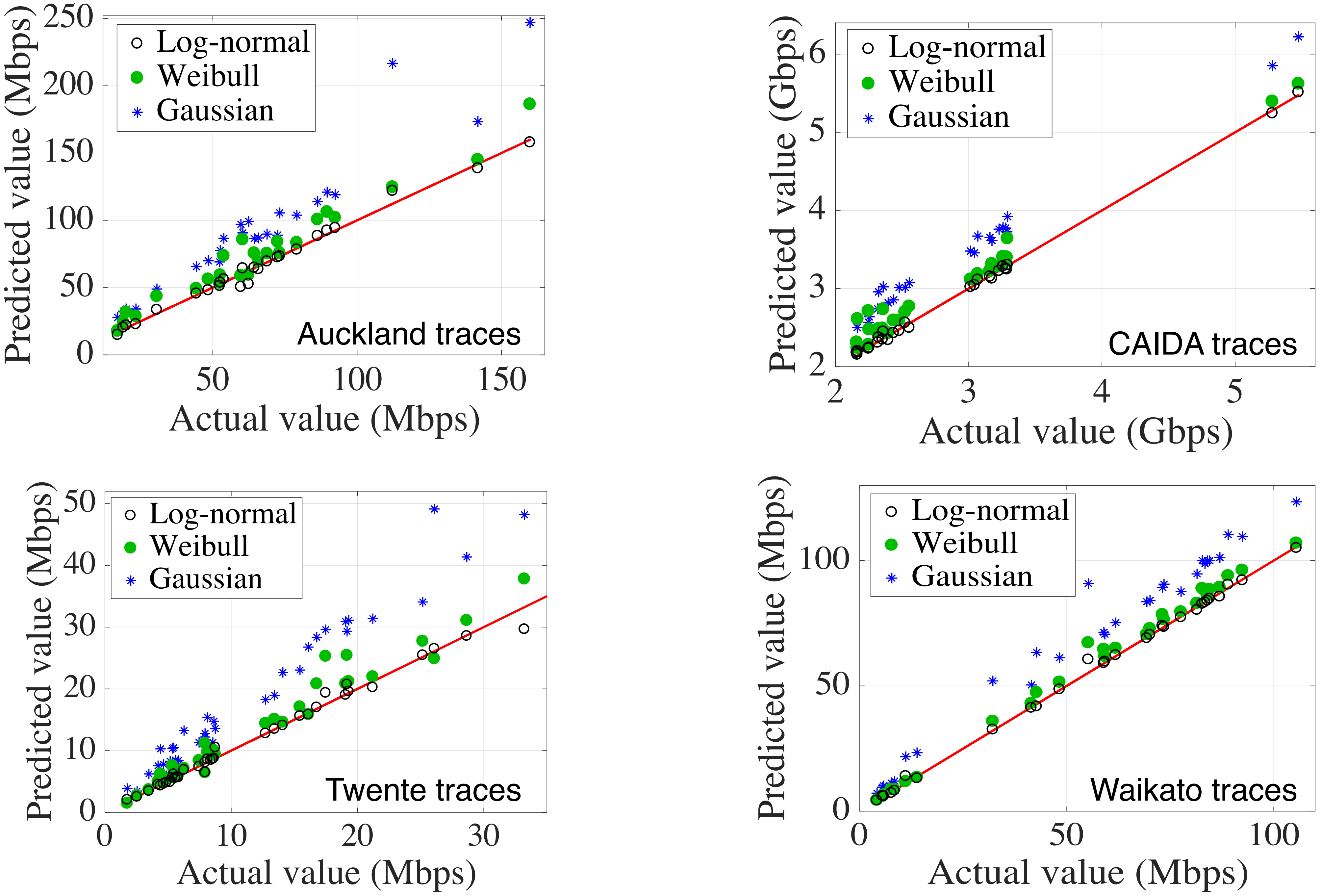
$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}$$

Power-law test

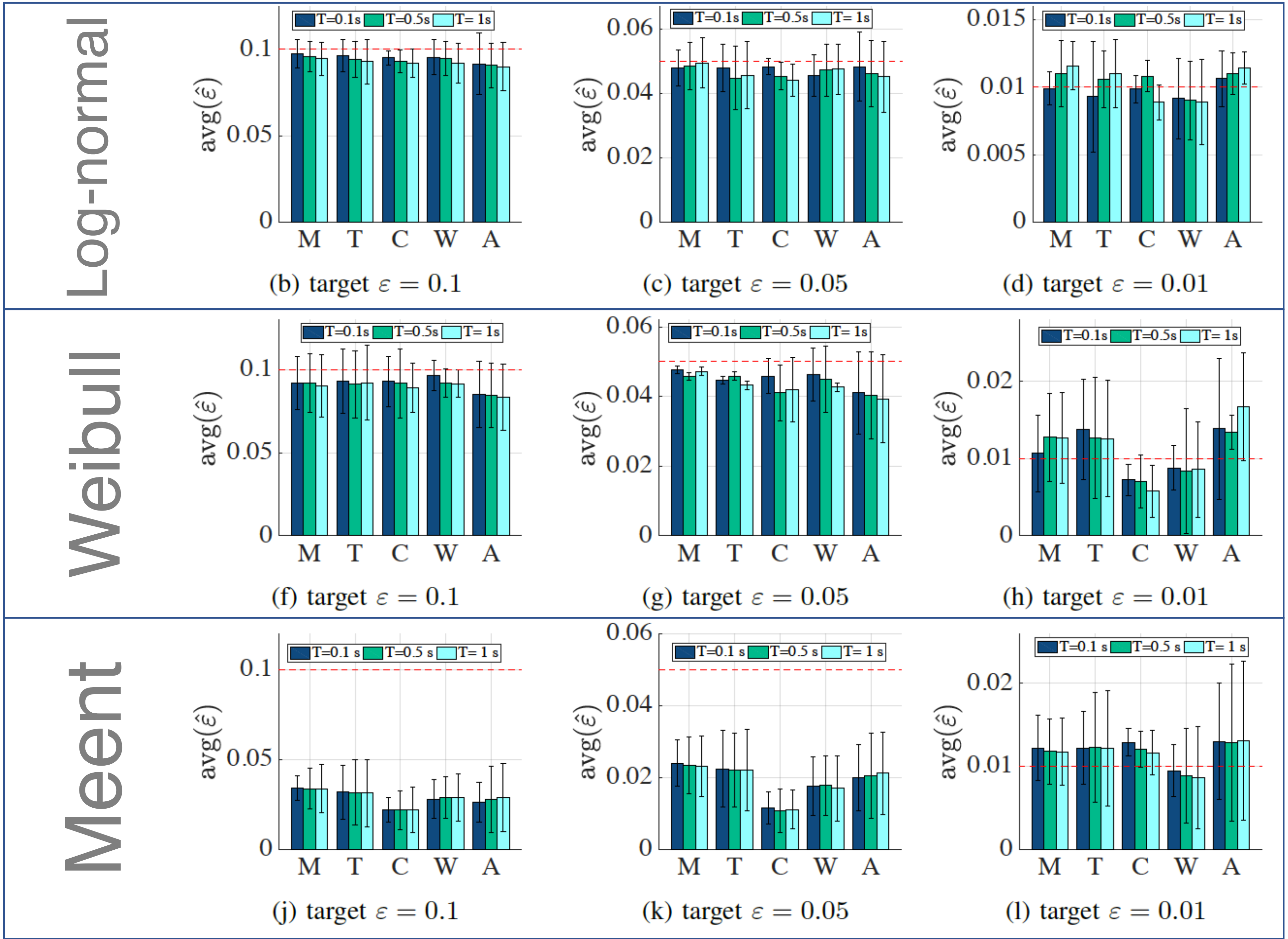


Log-Likelihood ratio ($\mathcal{R}$)

[Ref] A. Clauset, C. S. Rohilla, and M. Newman, "Power-law distributions in empirical data," arXiv:0706.1062v2, 2009.



Predicted value (Mbps)
Log-normal
Weibull
Gaussian
Auckland traces
Actual value (Mbps)
Predicted value (Gbps)
Log-normal
Weibull
Gaussian
CAIDA traces
Actual value (Gbps)
Predicted value (Mbps)
Log-normal
Weibull
Gaussian
Twente traces
Actual value (Mbps)
Predicted value (Mbps)
Log-normal
Weibull
Gaussian
Waikato traces
Actual value (Mbps)



(b) target $\varepsilon = 0.1$ (c) target $\varepsilon = 0.05$ (d) target $\varepsilon = 0.01$

(f) target $\varepsilon = 0.1$ (g) target $\varepsilon = 0.05$ (h) target $\varepsilon = 0.01$

(j) target $\varepsilon = 0.1$ (k) target $\varepsilon = 0.05$ (l) target $\varepsilon = 0.01$