Wow! It's sure *looks* like a power law!

John Mark Agosta, Jing Xu, Jaideep Chandrashekar, Dhiman Barman, Frédéric Giroire, & Nina Taft, with thanks to Denver Dash, Carl Livadas & Eve Schooler.

Intel Research, Santa Clara, CA

Fitting well known distributions to network traffic fails

We're studying enterprise traffic distributions of flow counts in time intervals of 4 to 512 seconds.

It is commonly known that these distributions are not fit accurately by well-known parametric models, as these chi-squared tests show:

Distribution	Fraction of Users whose traffic fit distributions (95% significance)	
	KS	Chi-Square
Gaussian	0.295	0.282
Exponential	0.605	0.395
LogNormal	0.103	0.048
Gamma	0.869	0.814

Its hard to know if bursty traffic distributions are really heavy-tailed

Strong dependencies among network flows lead to bursty traffic, which makes modeling and prediction hard.

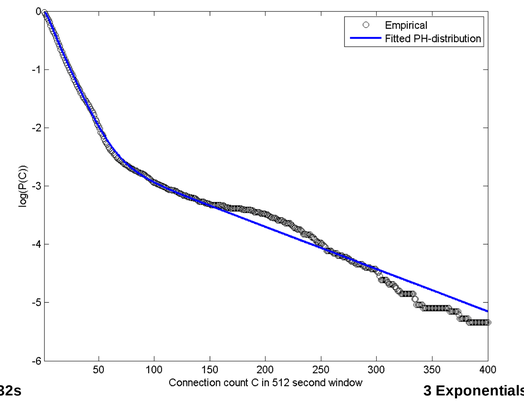
Many presumed fits to heavy-tailed distributions don't truly fit power-laws. [1] This is good news, since, if they fit more well-mannered distributions, prediction and anomaly detection are easier.

[1] arXiv:0706.1062 **Power-law distributions in empirical data.** Aaron Clauset, Cosma Rohilla Shalizi, M. E. J. Newman. physics.data-an (physics.dis-nn stat.AP stat.ME).

..but phase type model fits succeed!

In general, hierarchical models, e.g. mixture models, are both economical (fewer parameters) and offer better explanations.

$$P(Z \mid \Theta)\, P(\Theta)$$

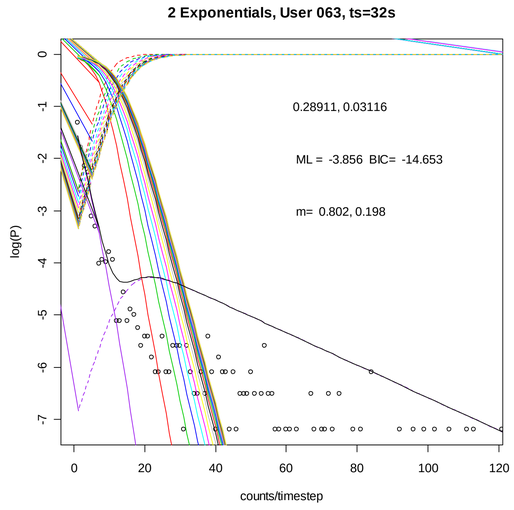


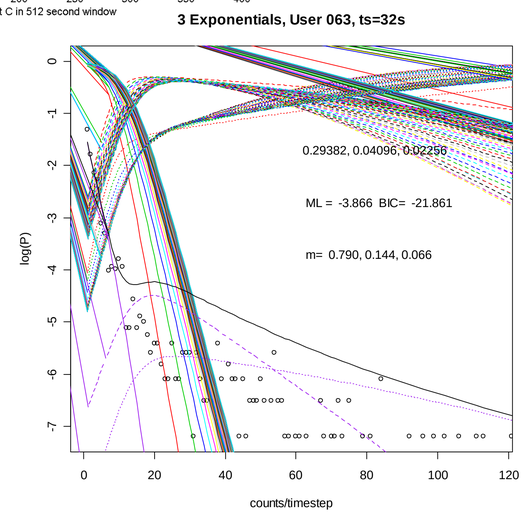
Phase-type complementary cumulative

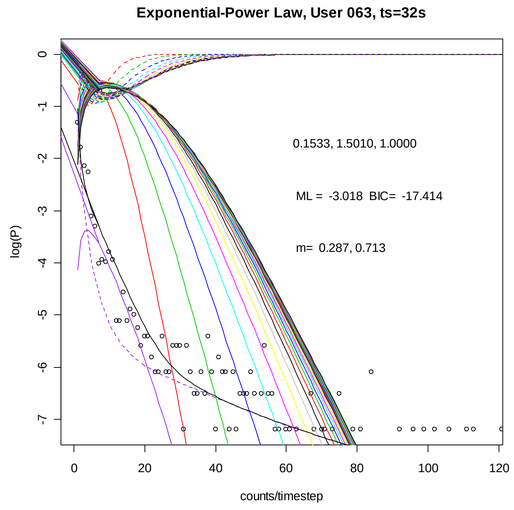
$$F(\tau) = 1 - \pi^{\top} e^{Q\tau} \mathbf{1}.\ \tau > 0$$

$$\mathbf{Q} = \begin{bmatrix} -0.1917 & 0.1888 & 0.0005 & 0.0006 & 0.0003 & 0.0002 \\ 0.0005 & -0.2206 & 0.1284 & 0.0003 & 0.0001 & 0.0002 \\ 0.0005 & 0.0002 & -0.1222 & 0.0897 & 0.0002 & 0.0003 \\ 0.0004 & 0.0001 & 0.0002 & -0.0902 & 0.0565 & 0 \\ 0.0002 & 0.0005 & 0.0003 & 0.0004 & -0.0943 & 0.0254 \\ 0.0002 & 0.0003 & 0 & 0.0004 & 0.0003 & -0.0075 \end{bmatrix}$$

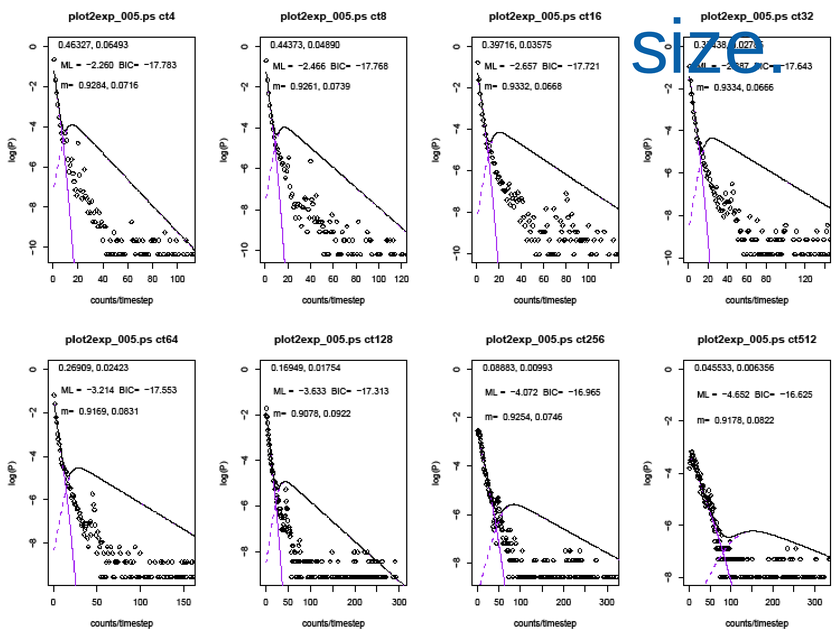
More precisely, fits to a mixture model reveal a substantial exponential component…

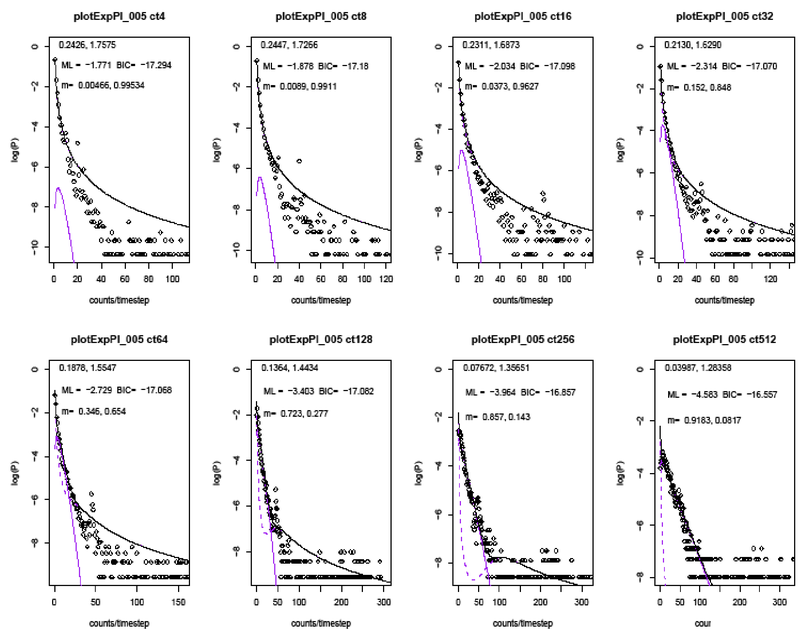






… whose parameters are stable across time-step size.



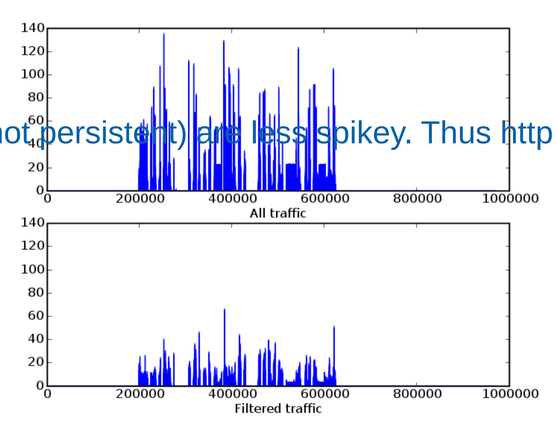


In mixtures, the fitted component parameters vary sub-linearly with the time-step with which flow counts are binned. Intuitively, the counts are due to the intensity of short duration spikes and not to the average traffic in the time-step.

Removing *the usual suspects* might explain it

Conditioning on traffic from persist destination addresses shows strong differentiation in spiky-ness. Rare connections (those not persistent) are less spikey. Thus http connections

We'd like to show that components discovered in traffic can be explained by observations like these.



This work is based on 300+ data sets of user network traffic that we collected or a five week period. These *Forbidden City* traces are available on projects in collaboration with our research.