

Describing multiple aspects of cloud usage with excess entropy

100,000 servers is the simple part...

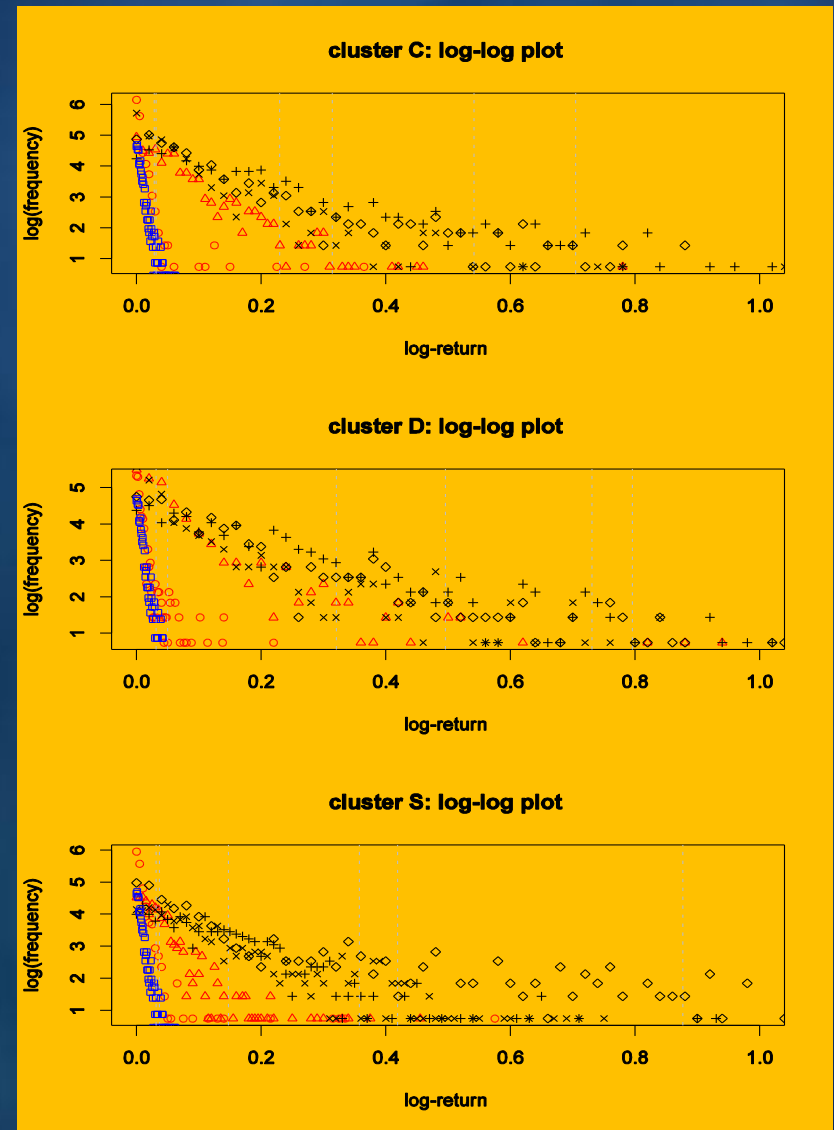
Charles Loboz
Windows Azure
cloboz@microsoft.com

Background: system size, complexity, body-of-knowledge

- Hardware: grouped in racks, clusters (of different types), cluster groups – and there is the network layer, reliability,
- Customers: different patterns – with tens of thousands of customers... you get them all – *there are no 'typical problems'*
- Very small number of people to handle large volume of information – and consider hundreds of aspects *on all levels* – daily

Descriptors and higher-level descriptors

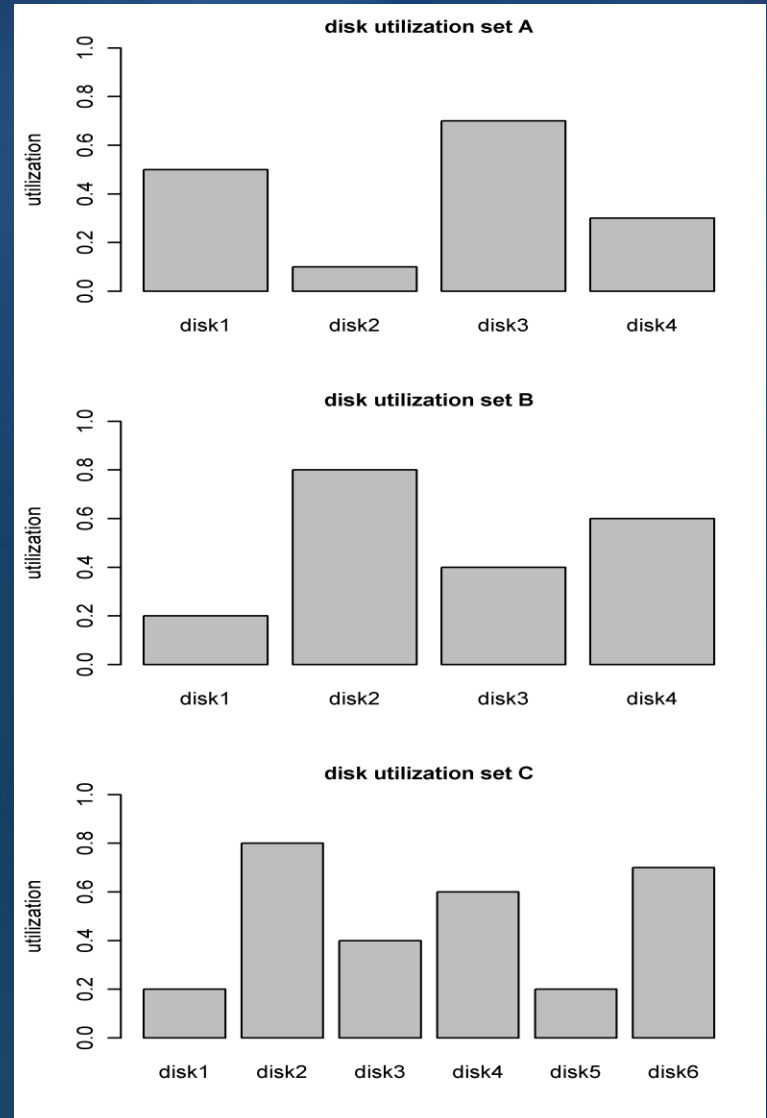
- Averages, standard deviations - we do *not* have normal distribution of customer 'size' – even the aggregated usage is more volatile than a stock market index (Stor-data here)
- The questions are *not* 'what is the (average) utilization':
 - which system components are *overloaded* (*not* just highly utilized)?
 - Is the load 'well distributed' – if not which element needs to be fixed?
 - On which days the load is different from other days?



Excess entropy

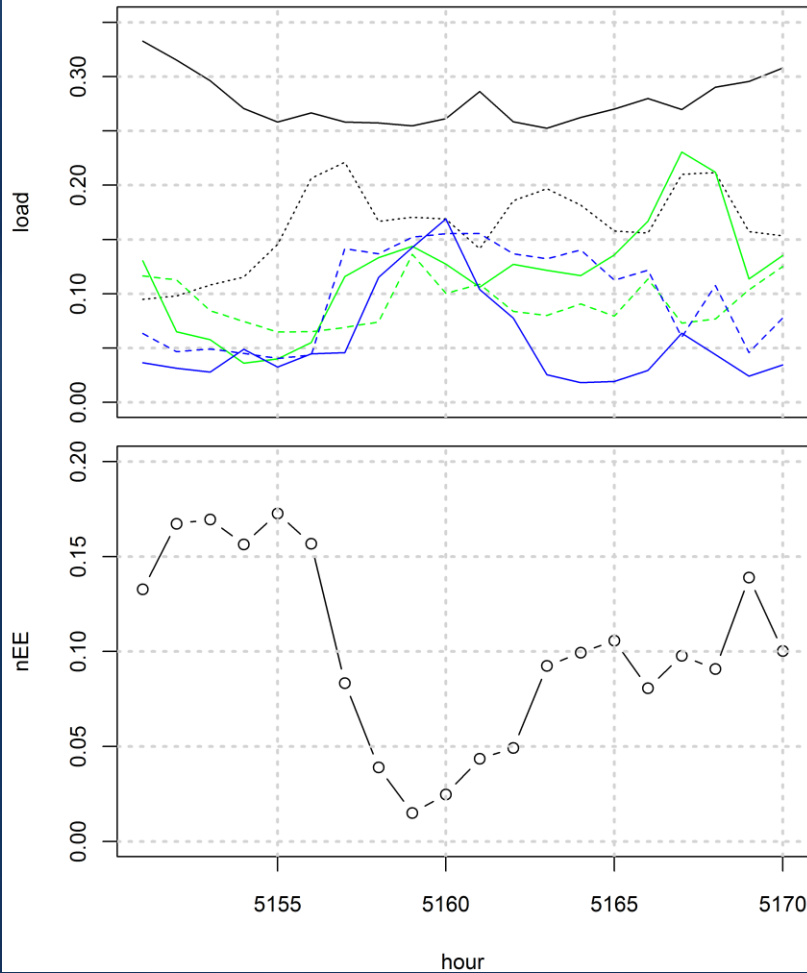
- Imbalance in computer systems – always important, never quantified
- We reach to economics and physics
- To get a formal description applicable to a range of situations

- $$EE = \frac{1}{n} \sum_{i=1}^n \frac{x_i}{x_{avg}} * \ln\left(\frac{x_i}{x_{avg}}\right)$$

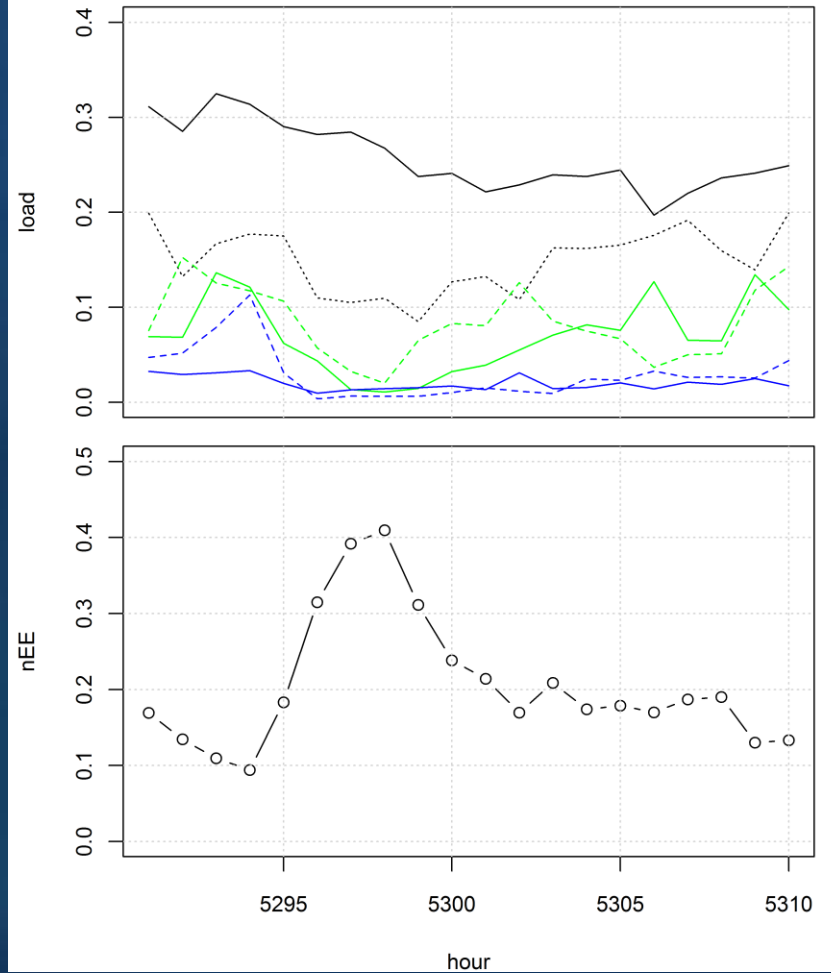


Intra-day imbalance between clusters

cluster load and excess entropy over 20 hours

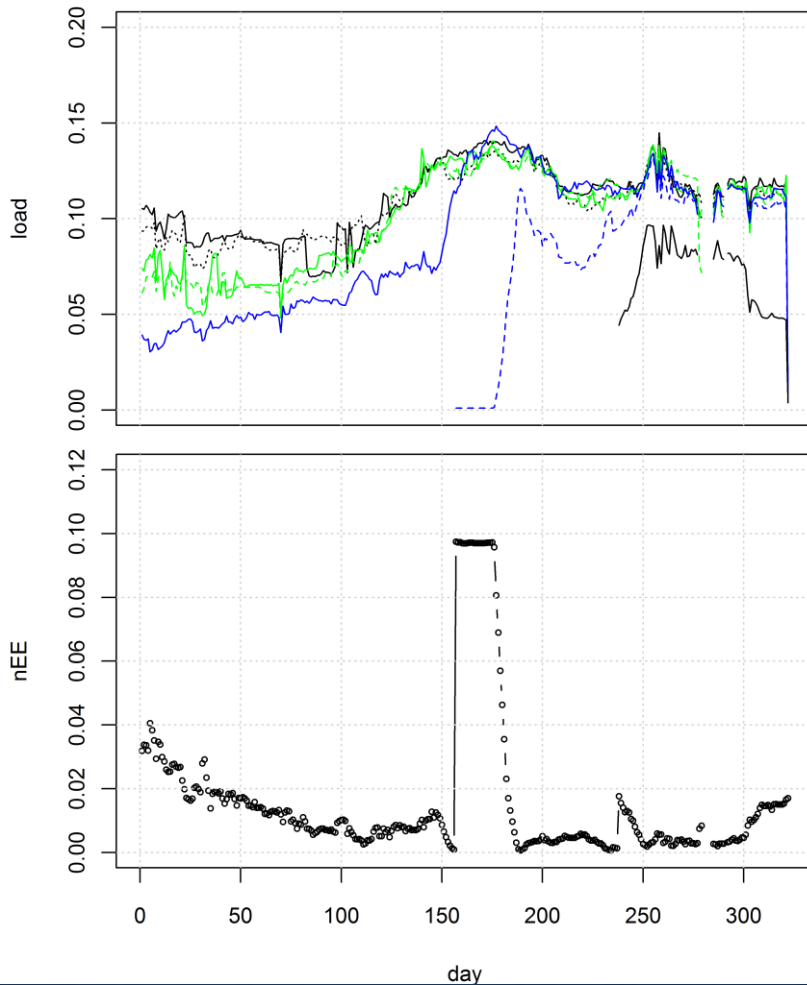


cluster load and excess entropy over 20 hours

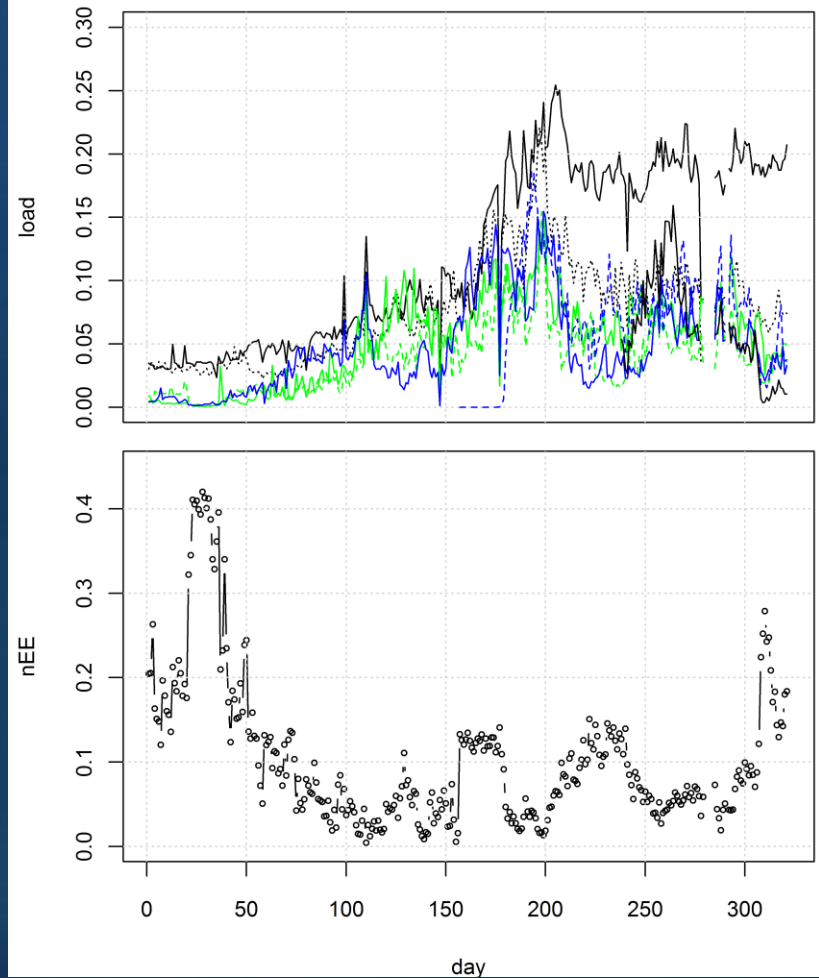


Long-term imbalance of load between clusters

ComputeHours: load and excess entropy

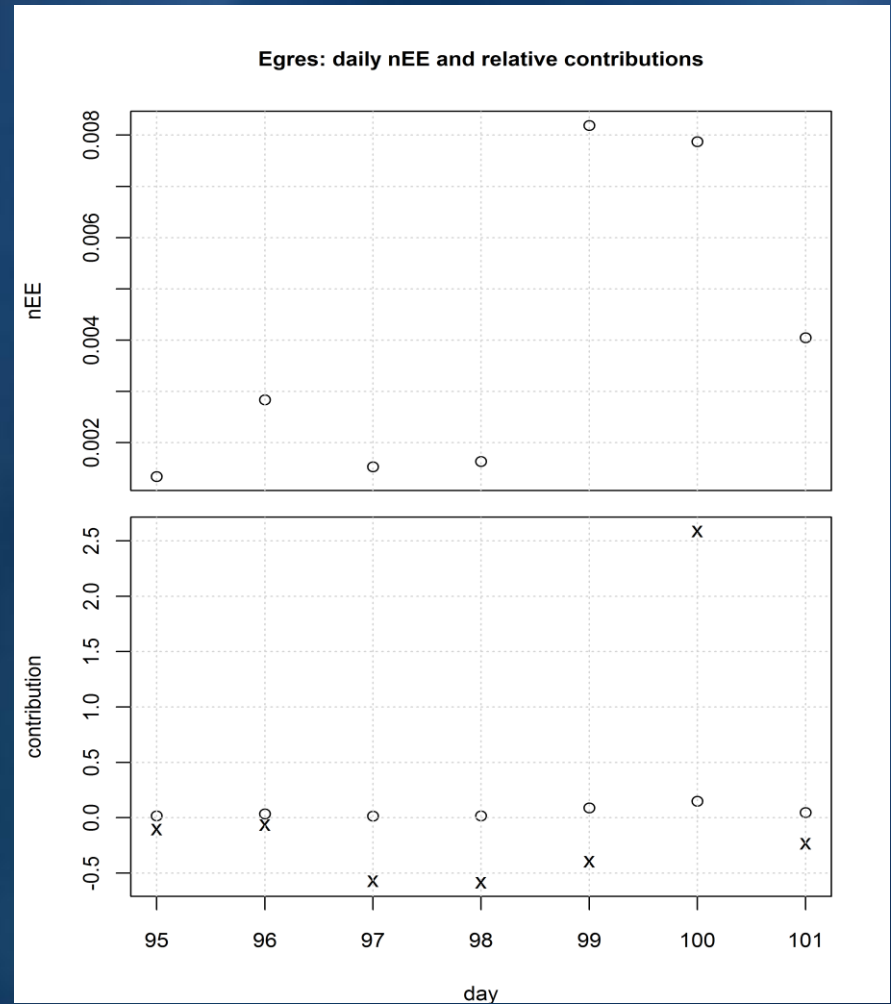
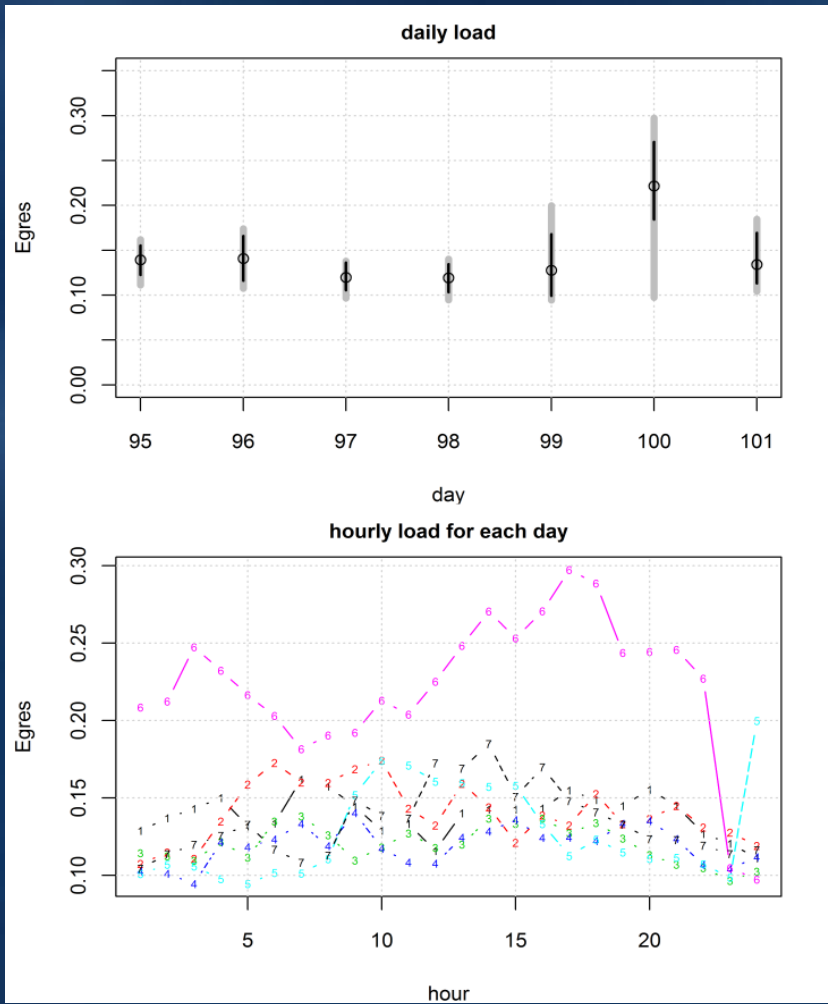


Egres: load and excess entropy



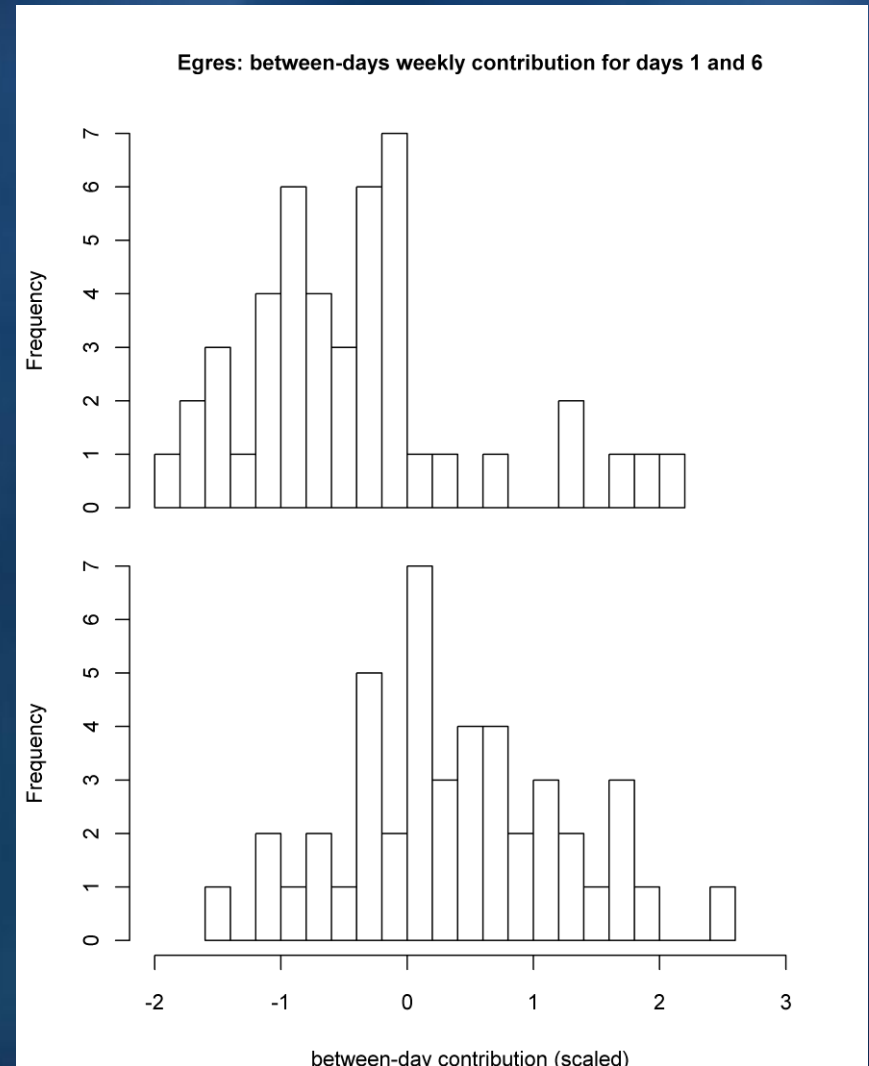
Components of system's excess

$$CEE = \sum_{m=1}^M f_m * EE_m \text{ entropy} * \ln\left(\frac{avg_m}{avg_{tot}}\right)$$

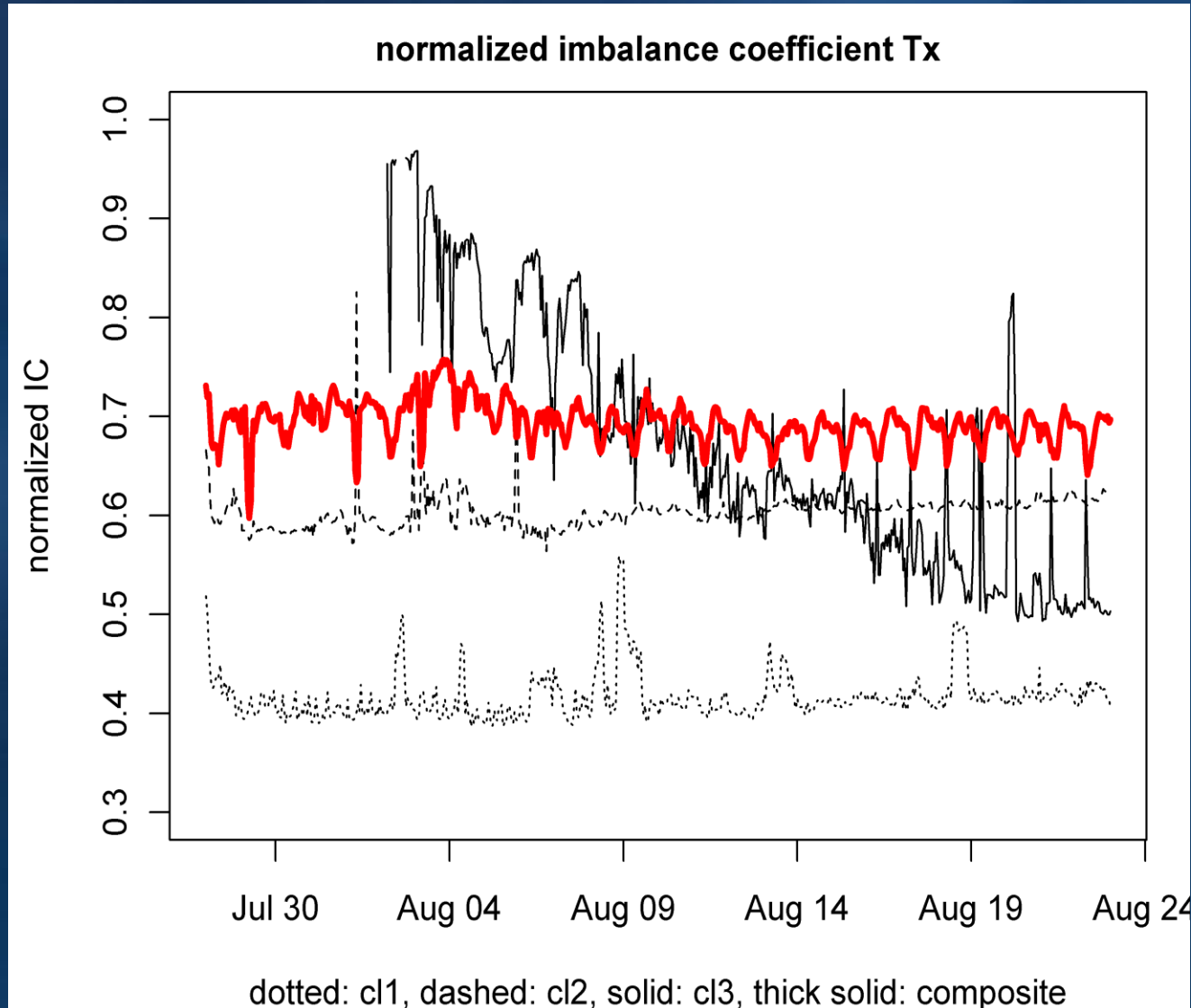


Comparing contribution to volatility

- Are Mondays different from Saturdays?



System and component imbalance



Excess entropy – why

- Dimensionless, normalizable, composable
- Applicable across wide range of problems
- Can compare apples to chocolate – because it compares *inequality* of apples with *inequality* of chocolate
- One formula – multitude of interpretations (all correct and numerically equivalent – like quantum mechanics 😊)
- Inequality, imbalance, concentration...
 - disproportion, asymmetry, flocking, variation(?), distribution(?)

Some other interpretations...

- *[The Theil index can be interpreted] as the expected information content of the indirect message which transforms the population shares as prior probabilities into the income shares as posterior probabilities.*
Henri Theil (1967:125-126) [4] [close to Nobel 1969]
- *But the fact remains that [the Theil index] is an arbitrary formula, and the average of the logarithms of the reciprocals of income shares weighted by income is not a measure that is exactly overflowing with intuitive sense.*
Amartya Sen (1997:36)[6] [Nobel 1998]
- *Or we can just think about it as excess entropy ☺*

Excess entropy – gotchas

- Context dependent
 - consider carefully what is the underlying measure: example of customers and deployments
- Direction-insensitive
 - (almost) – sensitive to difference from average
- Requires underlying data > 0
- Logarithmic, not linear
- Order-independent – important for time series

Summary – excess entropy application

- versatile
- multi-level
- handles systems and subsystems
- intuitive in most contexts
- operationally scalable
- can be fine-tuned for other inequality tests
- allows new aggregate descriptions (views) of complex systems