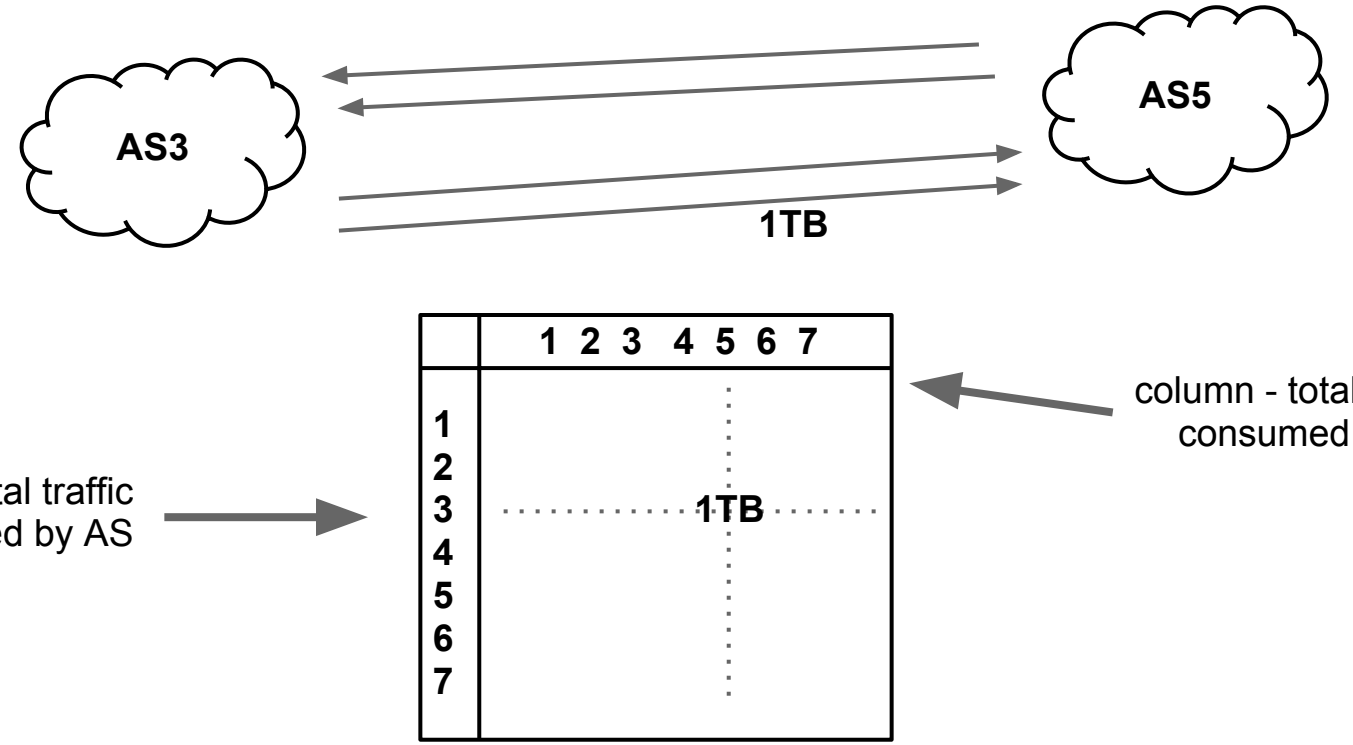
Towards a Statistical Characterization of the Interdomain Traffic Matrix

Jakub Mikians, UPC BarcelonaTech
Amogh Dhamdhere, CAIDA
Constantine Dovrolis, Georgia Tech
Pere Barlet-Ros, UPC BarcelonaTech
Josep Solé-Pareta, UPC BarcelonaTech

IFIP Networking 2012
CVUT, Prague

What is ITM?

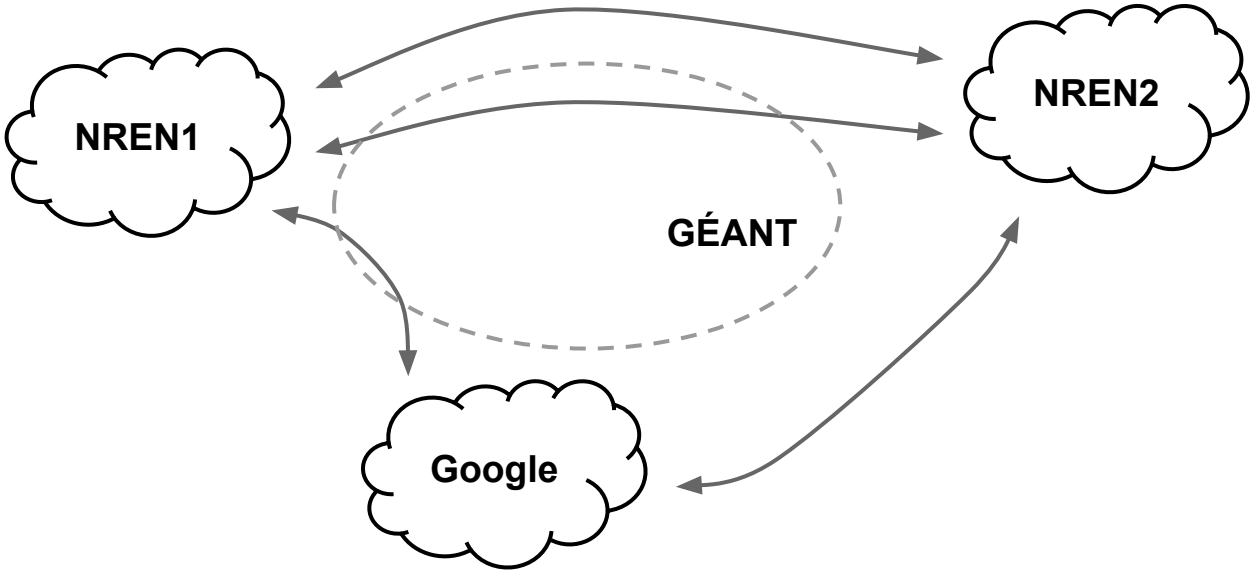


- AS or prefix granularity
- rows and columns - traffic produced and consumed
- network economics, protocols, applications -- models
- traffic -> money

Why to measure that? Why is that difficult?

- >40,000 ASes
- you can observe only a few full rows and columns (if you are lucky!)
- to work with ITM, you need to rely on models
- ...so you need to know the characteristics of the real ITM (ground truth, validation / sanity check of a generated matrix)
- Traffic reports (Arbor, Sandvine,...) are insufficient here

Dataset - GÉANT Network



- GÉANT - European-wide transit network connecting hundreds of research networks and universities - data accessible for the researchers
- Biased towards academic traffic, but also contains commercial traffic
- Not all the prefixes are routed via GÉANT - rows are incomplete
- For some of the ASes, GÉANT is the main connectivity provider
- Routing is defined for source AS and destination prefix, so later we talk about AS-to-prefix matrix

Dataset - GÉANT Network - routing stability

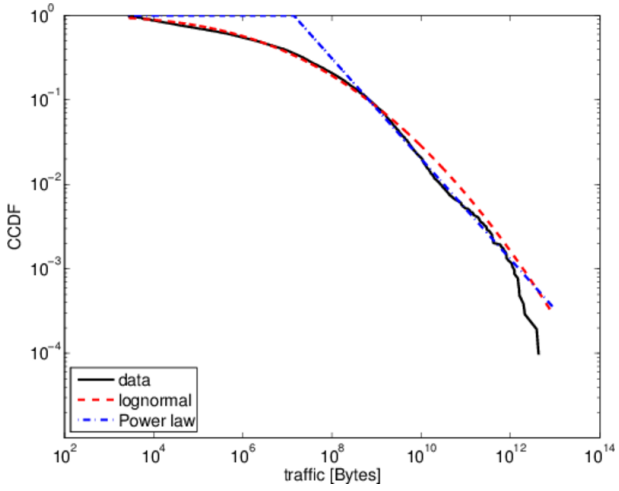
Snapshot length cannot be too short (little data) or too long (errors accumulated)

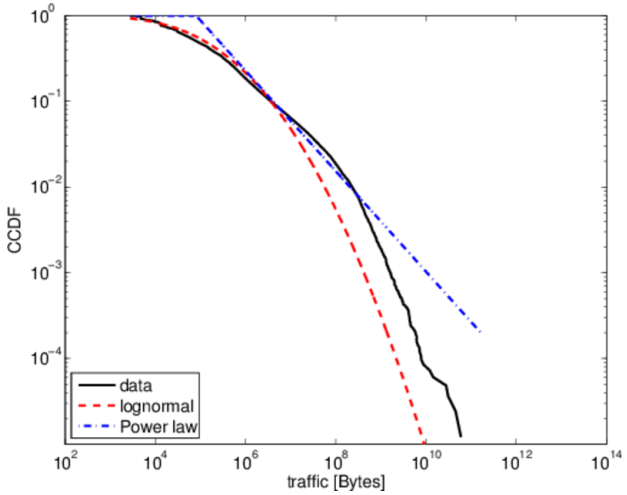
Probability that a prefix that is routed via GÉANT is routed through G. during the whole measurement interval

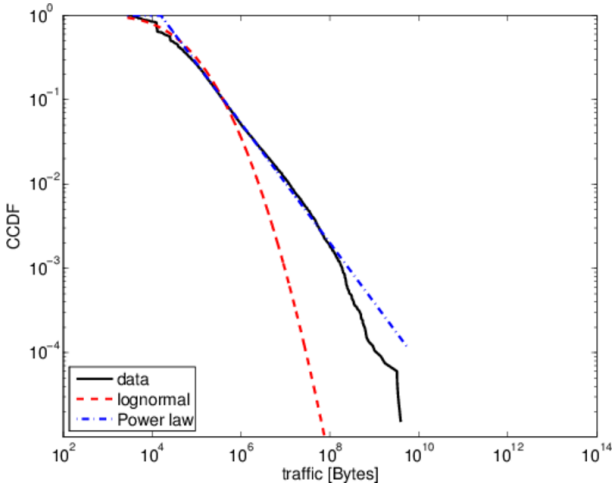
99.9%	**95.2%**	**75.0%**
1 day	1 week good tradeoff	1 month

Properties - statistical distributions

CCDF of the traffic produced by three exemplary ASes and the statistical models



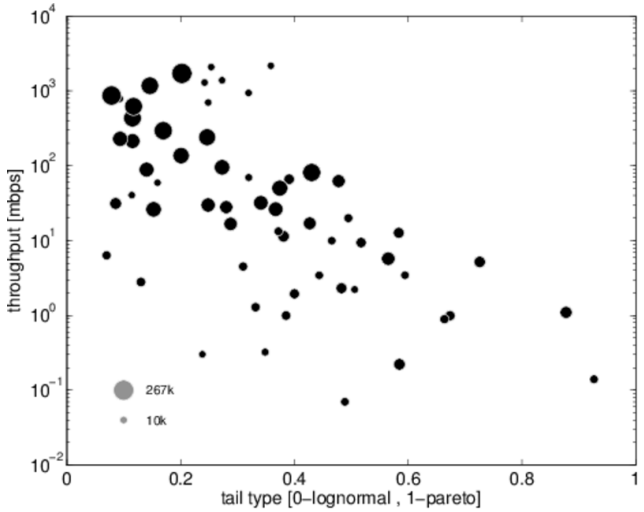




- Over 3,000 rows analysed overall in 52 week snapshots, 94% were heavy tailed
- Some distributions are Pareto-like (straight line - right), others LogNormal-like (bent - left) and others somewhere in between those two
- What is a potential cause of the difference between the distributions?

Properties - statistical distributions

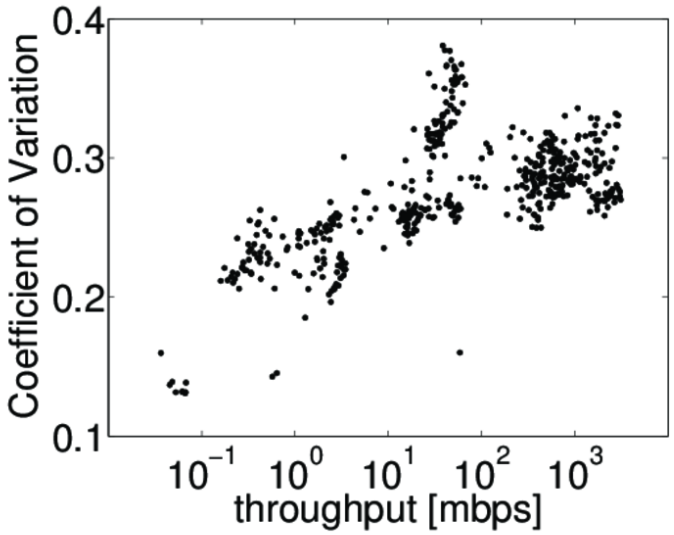
Shape of the of the distributions (per row) as a function of the observed throughput - larger throughput is connected with more "bent" characteristic



- Small throughput - flat power-law, large throughput - bent LogNormal
- We excluded possibility that it is a measurement artifact (i.e., of the number of observed prefixes per row)
- Size of the dot - number of prefixes per row observed in our dataset

Properties - statistical distributions

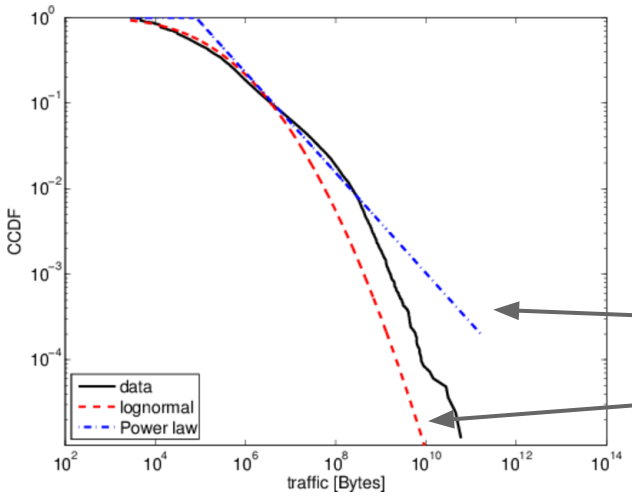
Coefficient of Variation (CoV) also depends with the throughput



- CoV = std. dev. / mean -> measure of dispersion
- CoV is also a function of the average throughput
- Again, changes in the *shape* and not only the *scale* of the distributions

Properties - congestion suspected?

An *information bottleneck* can cause changes in the statistical distributions (tail truncation)



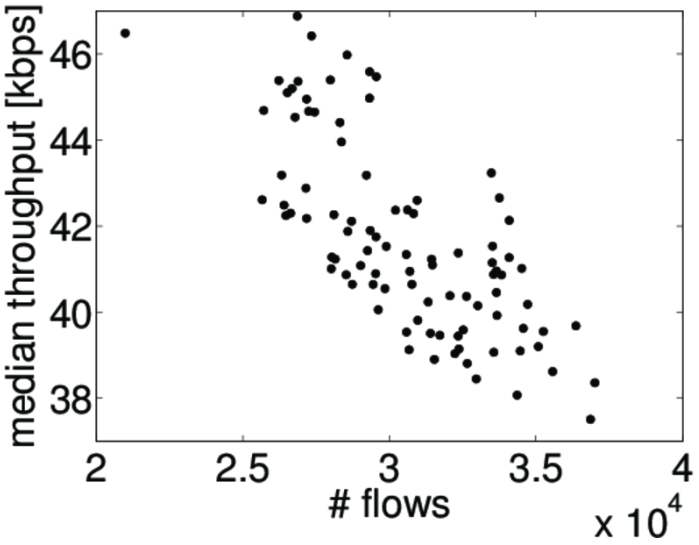
In this case, the bottleneck would be caused by congestion - "big guys" do not get that much traffic as they would get if the congestion was not be there...

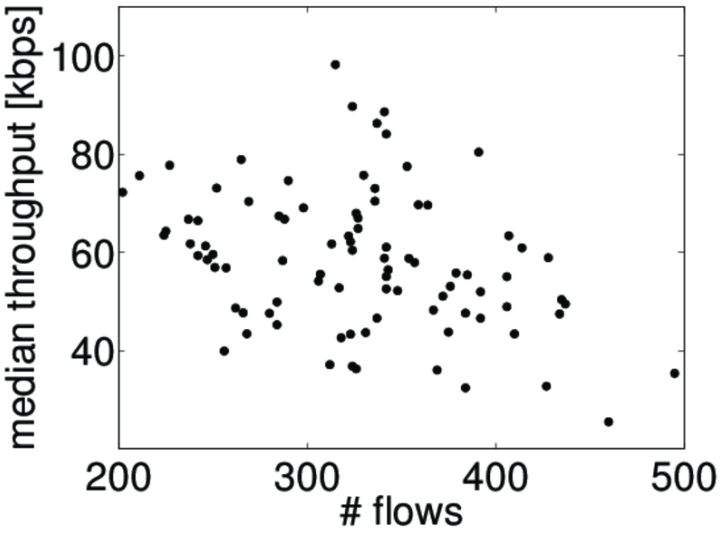
...and the tail of the distribution is "pushed" . Large players are affected more than the small ones.

- The potential congestion could be inside a network, not at the GÉANT link - hard to verify
- Bottlenecks mentioned in: Cha, M. et al. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system."

Properties - congestion suspected?

Avg. throughput as a function of number of flows for the potentially congested (left) and uncongested (right) network





- We cannot measure the congestion directly
- During the congestion - additional flows will compete for the bandwidth -> shrink the avg. throughput
- Analysed NetFlow for networks for 10:00-20:00 periods, when congestion is the most probable
- Negative correlation between the throughput and # of flows (-0.82, Spearman)

Correlations

Are the rows of the ITM dependent? Spearman correlation between the rows

0.85

highest corr.

0.28

average corr.

99%

of corr.
are positive

- We retain top 15% of the elements in the rows (rest is noise)
- Measured Spearman correlation
- Measured ITM rows can observe different set of prefixes - compared only overlapping prefixes
- Compared over 15,000 pairwise combinations of the rows

Correlations - popular prefixes

The correlations are caused by the prefixes that are "popular" among multiple ASes.

0.75%

of the prefixes have the *significance* over 0.8

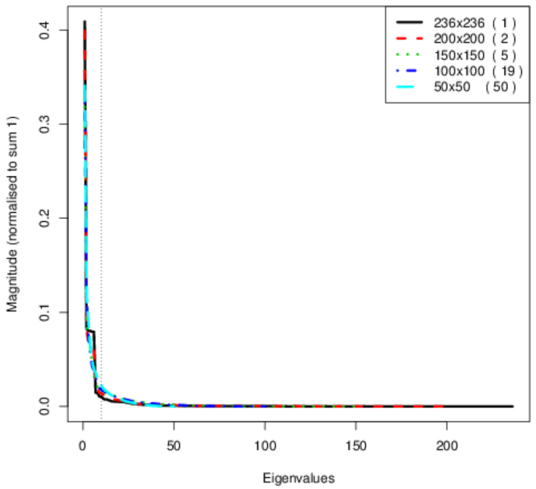
32%

...but they receive that big fraction of the overall traffic

- A prefix is *significant* if is in top-q of the AS traffic (here - top 15%)
- Significance metric for prefix- from 0 (not significant to any AS) to 1 - significant to all the measured ASes
- Small group of prefixes are *significant* for most of the ASes (globally popular)

Correlations - low effective rank

Low effective rank - matrix can be approximated by linear combination of small number of rows / columns



Only a small number of eigenvalues are significant

- Matrix completion techniques rely on the low rank (inferring invisible elements of the matrix)
- Extracted sub-squares from the matrix and calculated eigenvectors

Sparsity

How many elements are 0 (that is, how many prefixes are never used)

47%

for 1 day long snapshots

26%

for 1 week long snapshots

- Confirmed also with the university access link traffic and indirectly in other works

Conclusions + Further work

- Measuring the entire ITM is impossible
- We perform first steps towards the statistical characterization of the ITM
- FW: techniques to build a synthetic, realistic traffic matrix
- More vantage points needed

Towards a Statistical Characterization of the Interdomain Traffic Matrix

Jakub Mikians, UPC BarcelonaTech
Amogh Dhamdhere, CAIDA
Constantine Dovrolis, Georgia Tech
Pere Barlet-Ros, UPC BarcelonaTech
Josep Solé-Pareta, UPC BarcelonaTech

IFIP Networking 2012
CVUT, Prague