ITMgen - A First-principles Approach to Generating Synthetic Interdomain Traffic Matrices

Jakub Mikians  
UPC BarcelonaTech  
jmikians@ac.upc.edu

Nikolaos Laoutaris  
Telefonica Research  
nikos@tid.es

Amogh Dhamdhere  
CAIDA  
amogh@caida.org

Pere Barlet-Ros  
UPC BarcelonaTech  
pbarlet@ac.upc.edu
What is Interdomain Traffic Matrix

- ITM - matrix describing the flow of the Internet traffic at the highest level, between Autonomous Systems (AS)
Why is it useful?

- ITM - interesting for a number of reasons
- Internet economy (mainly)
- Flow of money ~ flow of traffic
- Interdomain interconnection policies
- Pricing schemes
- Routing protocols
- Peering strategies
- ...
Why is it difficult to work with ITM

- Difficulty to obtain representative traffic data - sensitive information from ISP perspective
- Impossible to have a full view of the ITM - single AS can observe a single row and column of the matrix
- +40k ASes - 40k rows x columns. Difficult to work with, in practice
Our goal...

- Be able to construct a representative, synthetic ITM
- Make it of arbitrary size - smaller (easier to work with, practical) but still useful for research purposes

- Useful in research, what-if scenarios
- Here we focus on static snapshots, and access networks
Observations

Connection-based

● We recognize that traffic comes from connections of individual users
● We model the traffic at connection level, and the traffic exchanged by an AS will depend on the number of users of the AS
Observations

Different content types

- **Different applications** (web, peer-to-peer file sharing, streaming video)
- **Traffic asymmetry** - we expect more asymmetric traffic in case of client-server web apps, and more symmetric for P2P
Observations

Regional and global popularity

- **Content popularity** shows global and regional effects
- E.g., Google and Facebook are popular worldwide, and national media are popular within a specific region
- We take into account the **global and regional popularity** associated with content objects
Traffic model

\[ T_{i,j} = \sum_{\kappa} m_{\kappa} \left( S_i p_i^{\kappa}(j) + d_{\kappa} S_j p_j^{\kappa}(i) \right) \]

Population of AS(j)

Traffic AS(i) to AS(j)

Application types

Application mix

Population of AS(i)

Popularity of AS(i) from the point of view of AS(j) with respect to application k

Popularity of AS(j) from the point of view of AS(i) with respect to application k

Forward traffic

Reverse traffic

Application k forward - to reverse ratio
Parametrization

- Used publically available data (AS sizes, content popularity, ... - *macro level*)
  - Alexa
  - marketing reports
  - ...

- Combined with packet traces (application characteristics - *micro level*)
  - packet level trace from research AS

- Focused on WEB traffic and P2P
Parametrization - number of users per AS

- Publicly available marketing data
- Rough estimation per AS, relative
- Market shares of ISPs for the top-10 countries
- ...combined with # of IP addresses in BitTorrent logs
- Empirical distribution of relative populations of ASes, for ~400 ASes (1% of the total number of ASes, ~60% of the total number of Internet subscribers)
Parametrization - WEB popularity ($p_{\text{web}}$)

- Used Alexa.com "page views" statistics
- Web pages popularity, globally and country-wise
- Statistical distribution of "popularity" of ASes
- We group ASes as globally popular, locally popular and remaining
Parametrization - P2P ($p_{p2p}$)

- P2P stats obtained from BitTorrent crawling
- Not much information about regional popularity (space for improvement...
Parametrization - application mix 
\((m_k, d_k)\)

- Characterizing forward-to-reverse ratio of applications
- 14 days of packet level trace from research AS (sampled, truncated)
- DPI to identify applications
Validating results

- Generated synthetic ITM for 1k, 2k and 3k ASes
- Compared with real 3 ISP AS statistics
Validating results

- Traffic produced and consumed (relative)

- Visible similarity, although the values differ. Distribution of the traffic in the generated matrices have "thinner tail" - is more skewed towards big ASes
- Comparing 3 real ASes and 3 selected ASes from the ITM
Validating results

- Traffic exchanged within the same region

(a) Traffic exchanged with ASes within same region; matrices of 4 different sizes are shown.

(b) Regional traffic of CPs.

- ASes in the generated ITM exchange traffic with ASes within the same region (compared with naive gravity model and CESCA measurements)
Validating results

- Application mix
- In the synthetic matrix, P2P ~27% of the traffic
- In measurements and reports P2P contribute to ~ 9% - 21%
- Model overestimates P2P with our parametrization...
Conclusions

- There is some potential in this method
- Difficult to parametrize... space for improvement here!
- No real volumes of traffic (bytes), only relative traffic flows
Acknowledgements

- J. Mikians was funded by FI Grant 2010FI B 00512 from Generalitat de Catalunya
- The research was funded by the Spanish Ministry of Science and Innovation under contract TEC2011-27474 (NOMADS project)
ITMgen - A First-principles Approach to Generating Synthetic Interdomain Traffic Matrices

Jakub Mikians  
Nikolaos Laoutaris  
Amogh Dhamdhere  
Pere Barlet-Ros

UPC BarcelonaTech  
Telefonica Research  
CAIDA  
UPC BarcelonaTech

jmikians@ac.upc.edu  
nikos@tid.es  
amoghp@caida.org  
pbarlet@ac.upc.edu