A Traffic Study to Interleaved Dark Space

Markus De Shon (mdeshon)

Agenda

- Methodology
- Results
- Discussion on Data Sharing

Google™

Methodology

Flow collection at Google

- Sampled sFlow and Netflow v{5,9} collected at network devices
- Written as annotated flow records to Google log infrastructure
- Google tools available for analysis
  - Mapreduce for batch processing
  - Near-real-time processing pipeline
  - Time series anomaly detection pipeline, with event classification and alerting

Google™

Darkspace at Google

- Some IP spaces allocated but unused (likely temporary)
- Most allocated IP space well-populated
- Some netblocks unused within larger populated blocks
- Allocated IP space identified from public IPs listed in internal network allocation database
  - Use inbound flow data instead? (messy)
- Unused space identified empirically, no outbound flows from a /24 in the last X days

→ Must keep dynamically updated list of unused IP spaces.

When traffic is observed from a /24, remove from list.
Batch runs over X days to identify new unused spaces.

Google™

Entropy timeseries

Calculate (packet count-weighted) information entropy by

- sIP
- sPort
- dIP
- dPort

cf. Zseby FloCon 2012

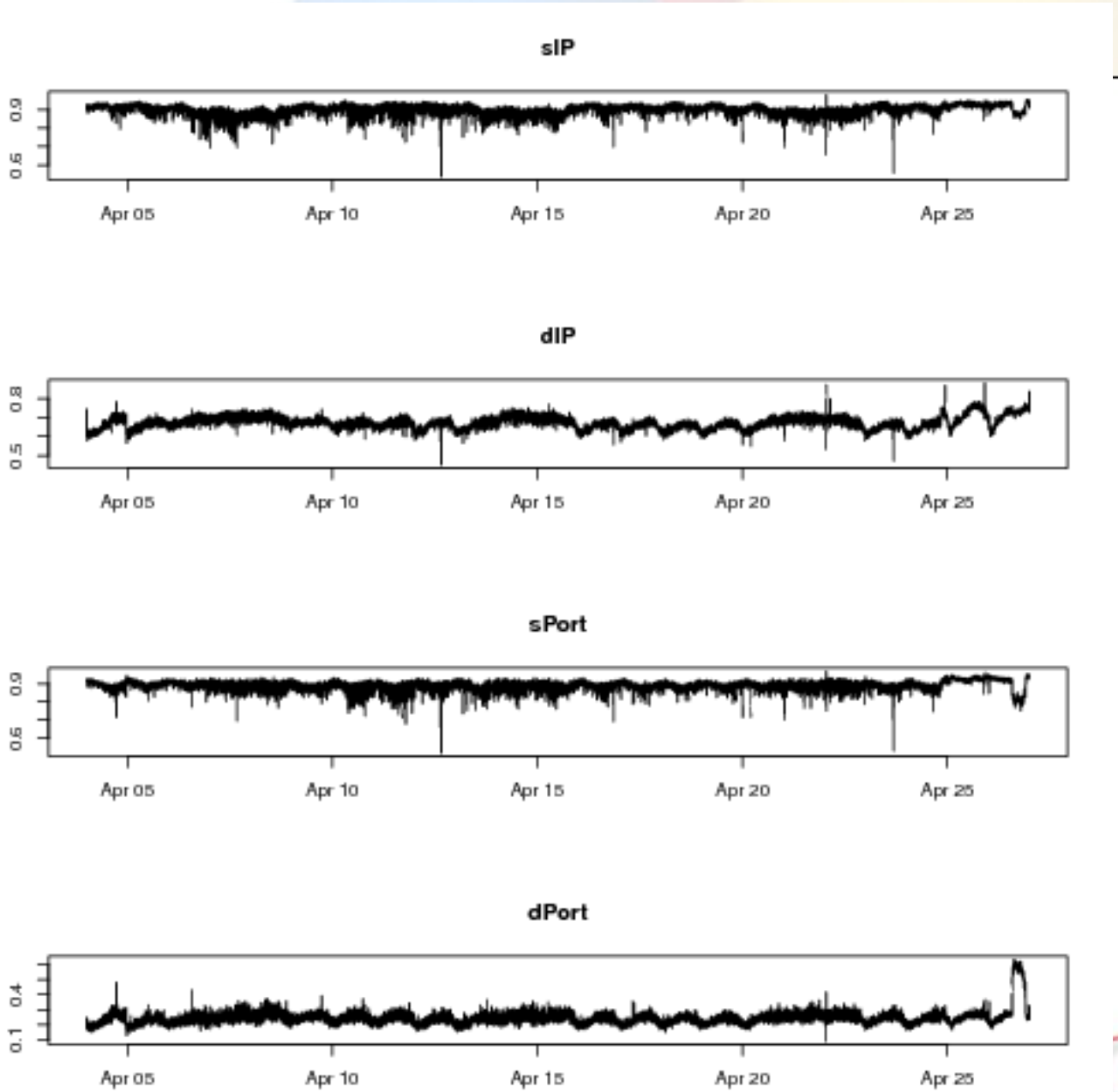
- Also calculated Bpp, not that useful so far...

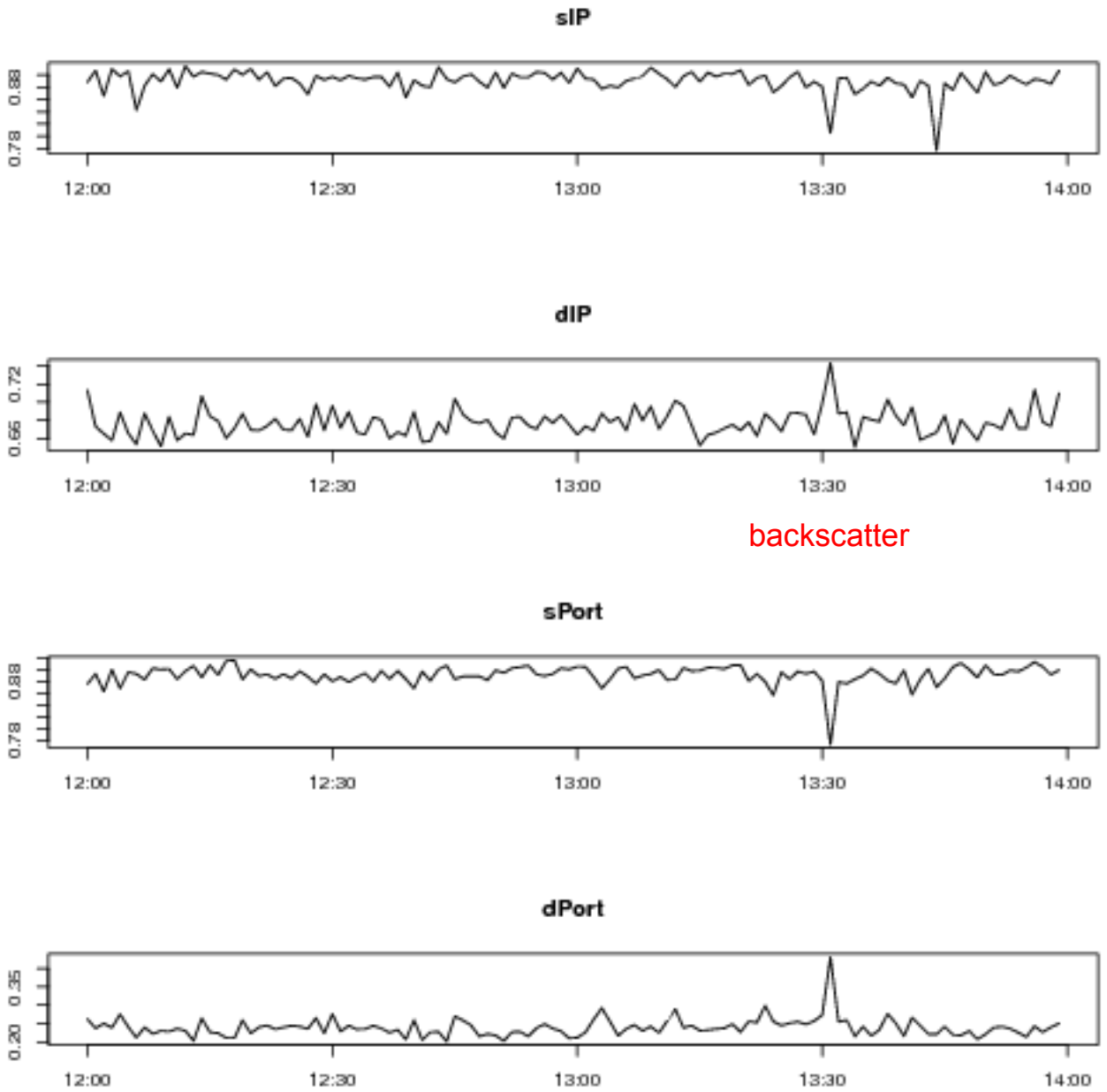
Scalable counting by unique keys in first Mapreduce
Entropy sums in second Mapreduce
All darkspace traffic aggregated, single timeseries per entropy

Google™

Results

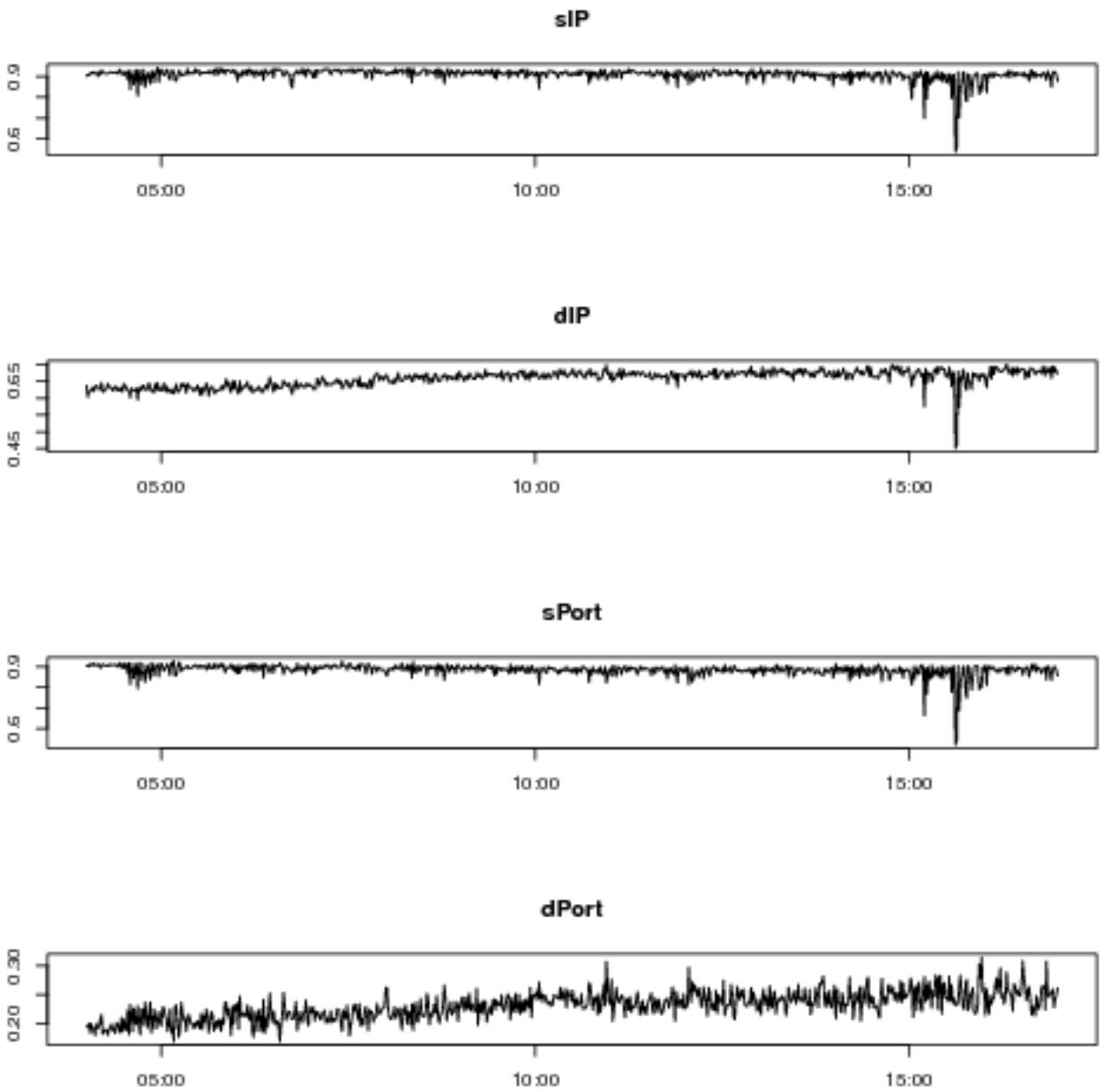
Timeseries of full time span





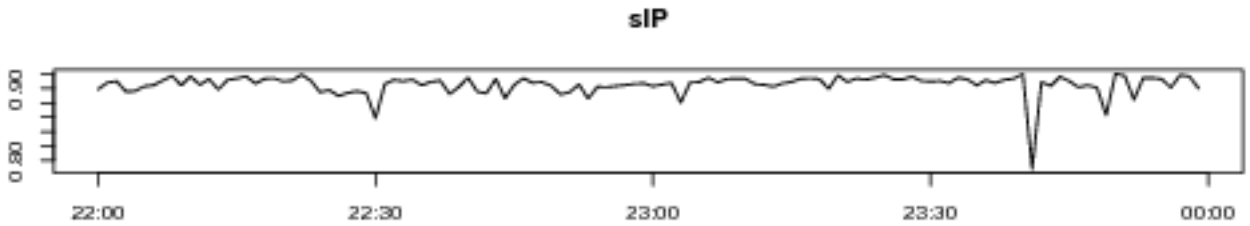
2012-04-06 12:00

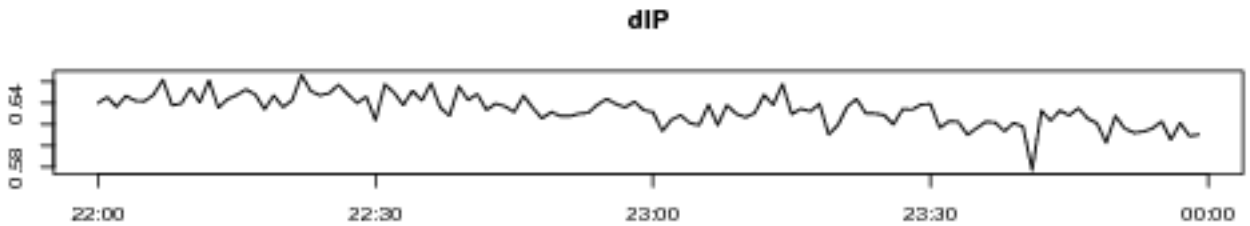
Google™

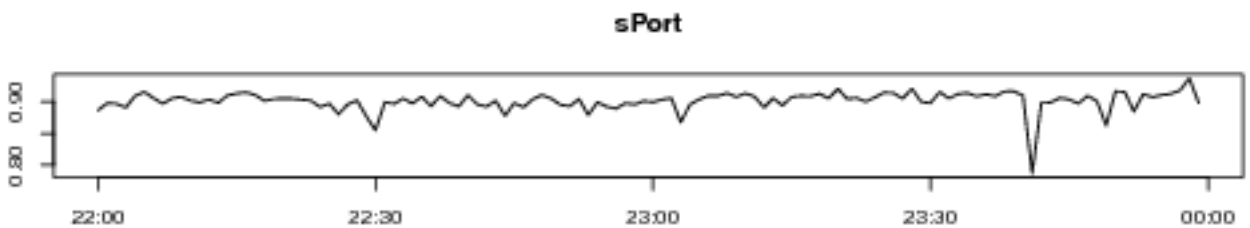


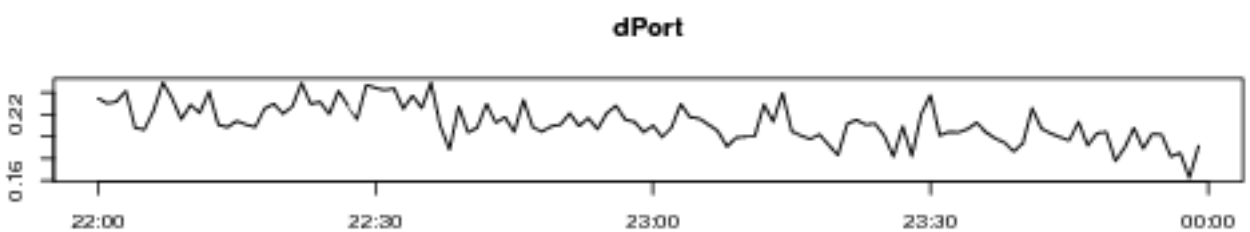
2012-04-12 04:00

Google™



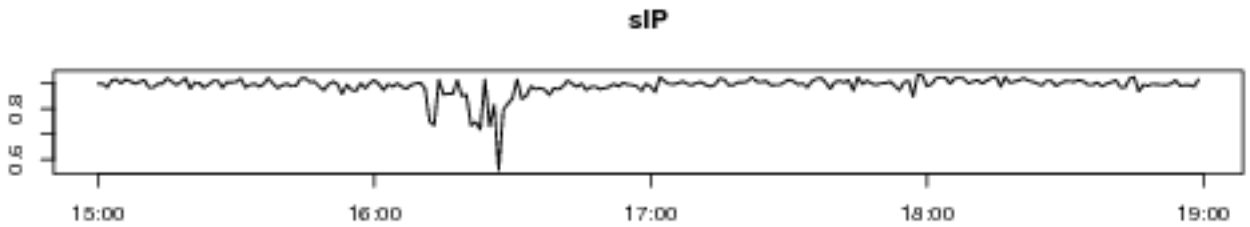


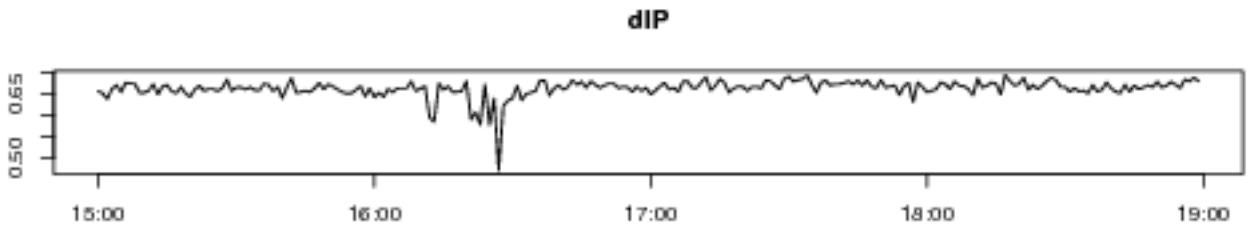


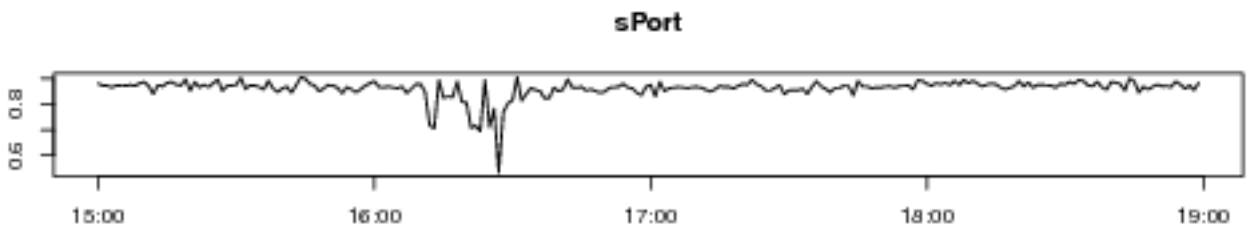


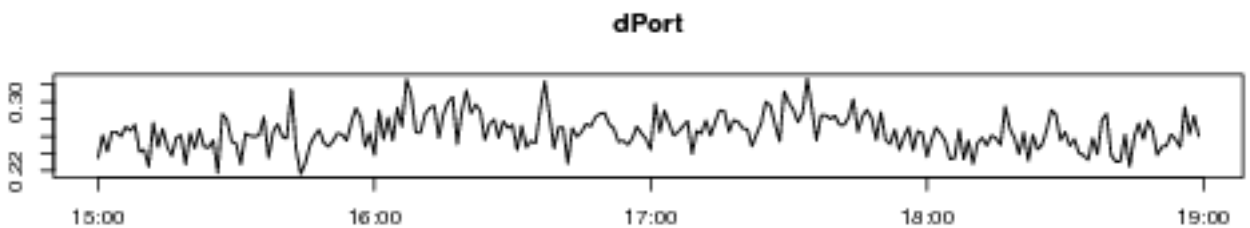
2012-04-22 22:00

Google™



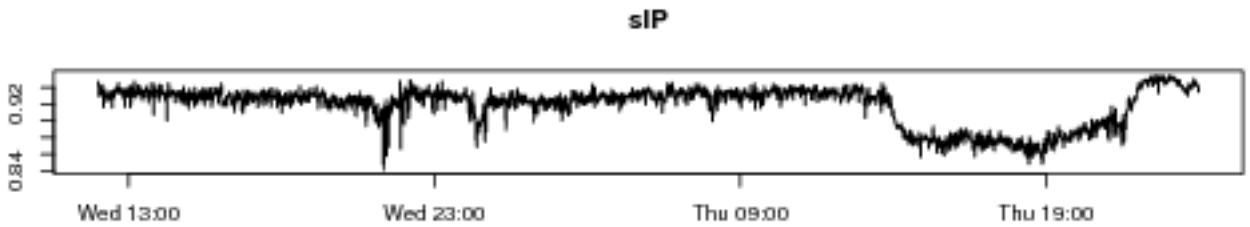




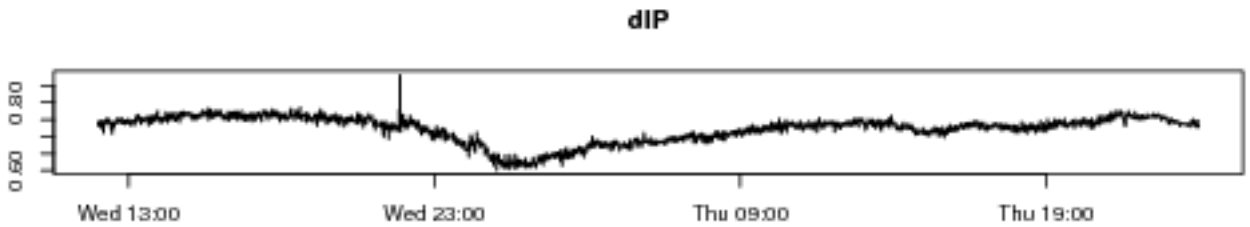


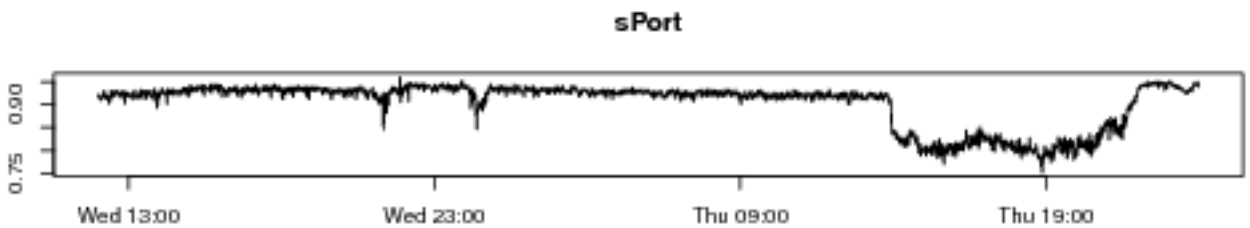
2012-04-23 15:00

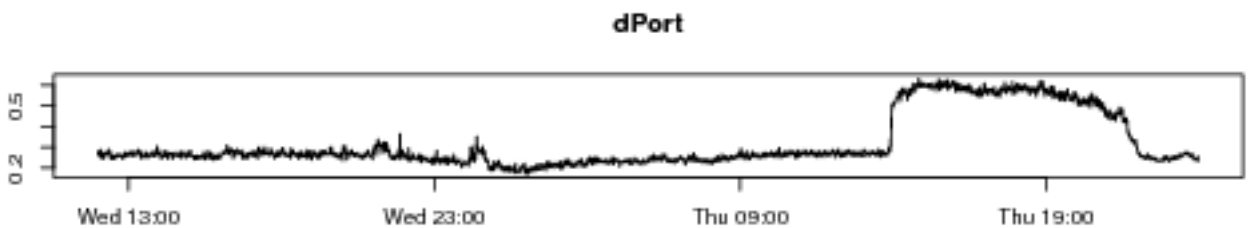
Google™



scan







2012-04-25 12:00

Google™

Future work

- Maintain a constantly updated map of active/dark network addresses
  - Darkspace telescope
  - Scan detection
- Integration of darkspace into near-real-time flow processing pipeline
- Study our IPv6 darkspace?
  - Huston NANOG 50 paper shows almost entirely misconfigured traffic, 100s of kbps across a /12
  - Will IPv6 darkspace be interesting?

Google™

Data sharing discussion

Google™

Needs for data sharing

No user data (requirement)

- Perfect identification/maintenance of dark IP space

Don't leak IP usage info (requirement)

- Nonreversible (?) map of dark IPs to reported IPs, OR
- No destination IPs reported

Google™

Needs for data sharing (2)

External source IP anonymization?

- Some kind of privacy-preserving query mechanisms... IANACrypto, but some system with features:
  - Alice delivers f(A), Bob delivers g(B)
  - Eve can perform Test(A==B) that does not reveal A or B, but permits aggregation across data sources to calculate total entropies
- Trusted sharing (e.g. SIE ISC)
- Other privacy-preserving designs (e.g. DEMONS)

Maximize aggregation (desirable)

- Share aggregate counts with one-way keys
- Perform entropy calculations in the sharing environment

Google™

Thank you!

Questions and Answers