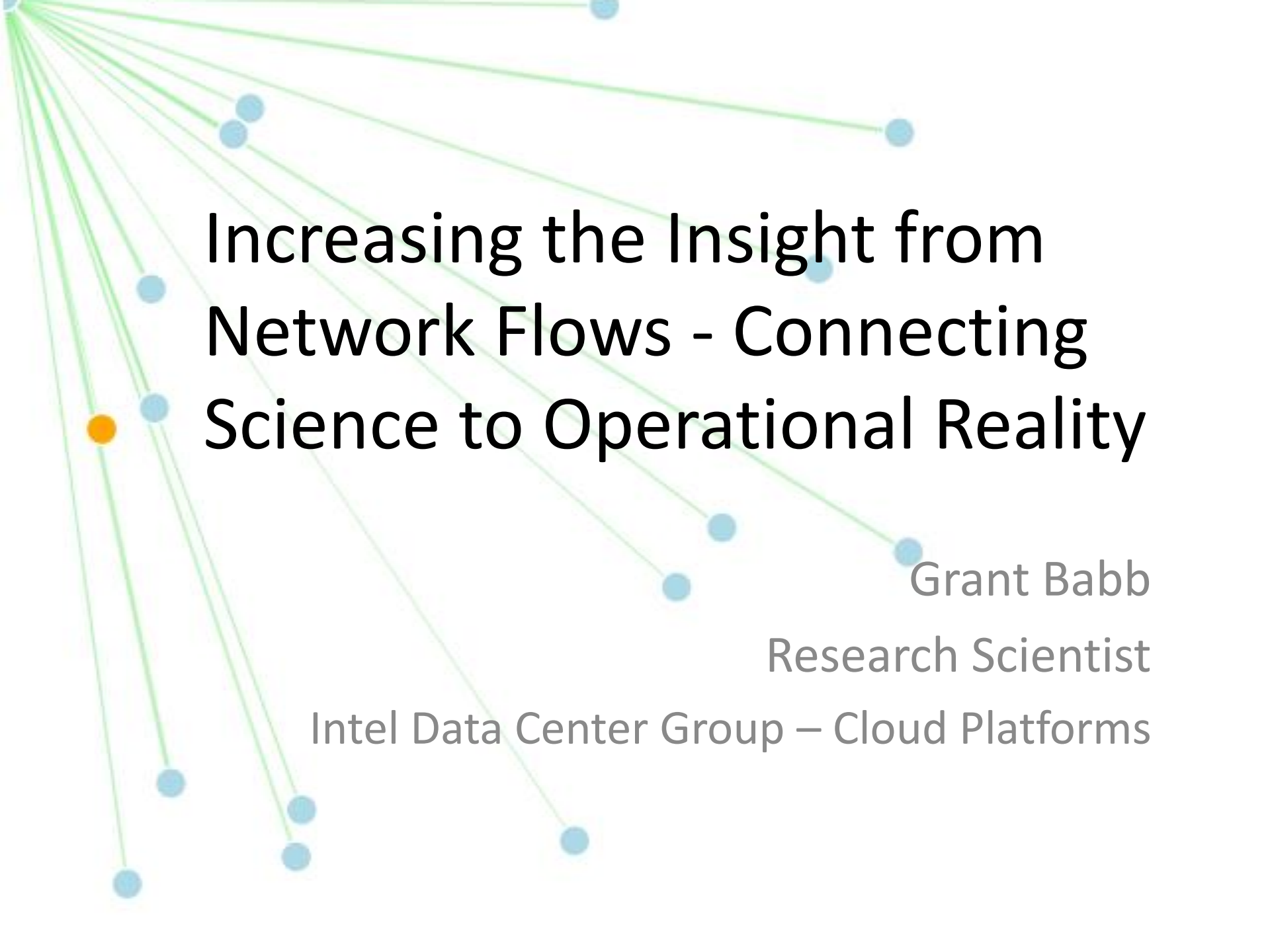


Increasing the Insight from Network Flows - Connecting Science to Operational Reality

Grant Babb

Research Scientist

Intel Data Center Group – Cloud Platforms

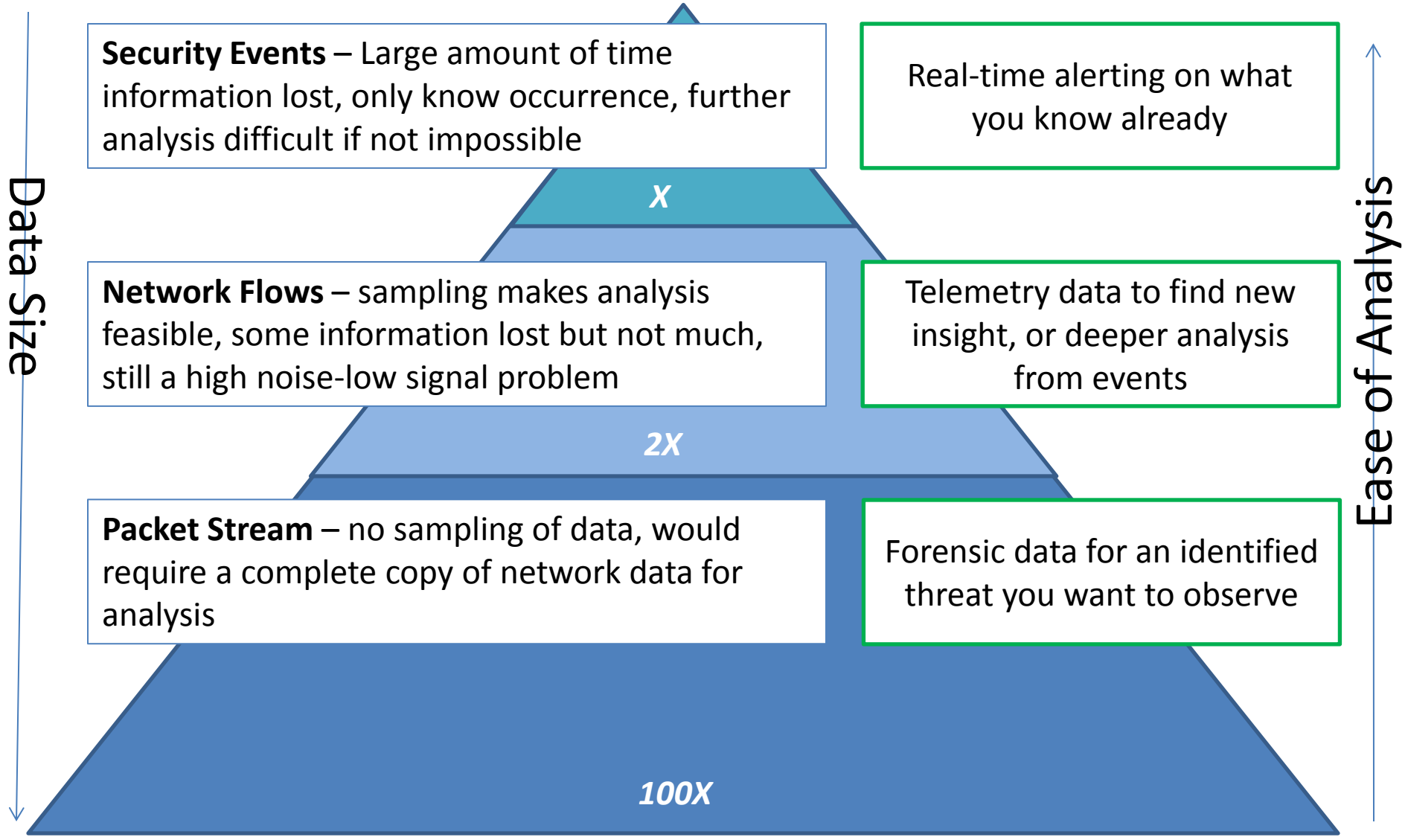
Objectives

- The BIG question
- Why netflows?
- Why transform them?
- What analytics to use?

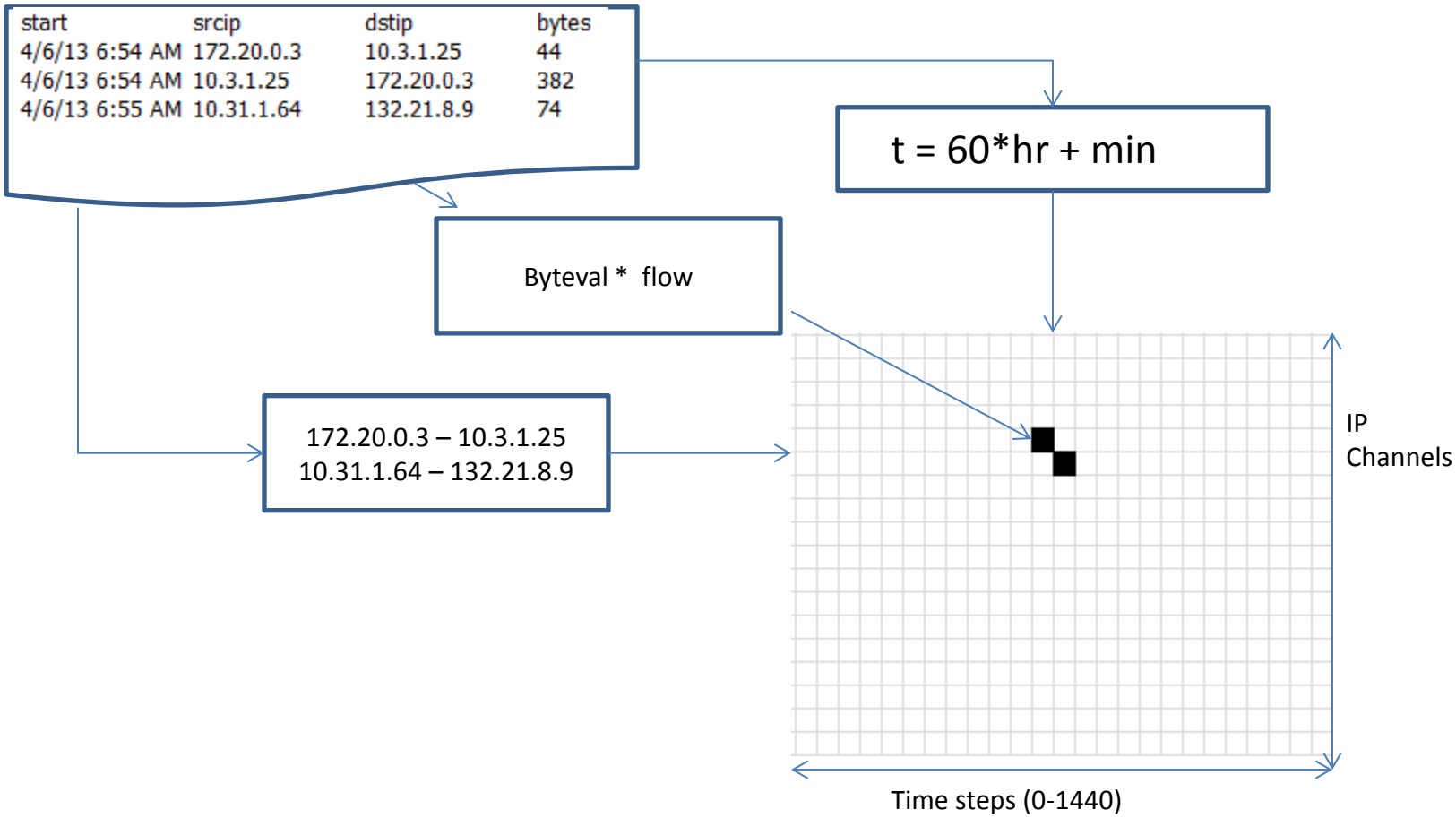
The BIG Question

What are the patterns in my network flow data that will identify a potential security threat?

Bridging the Gap



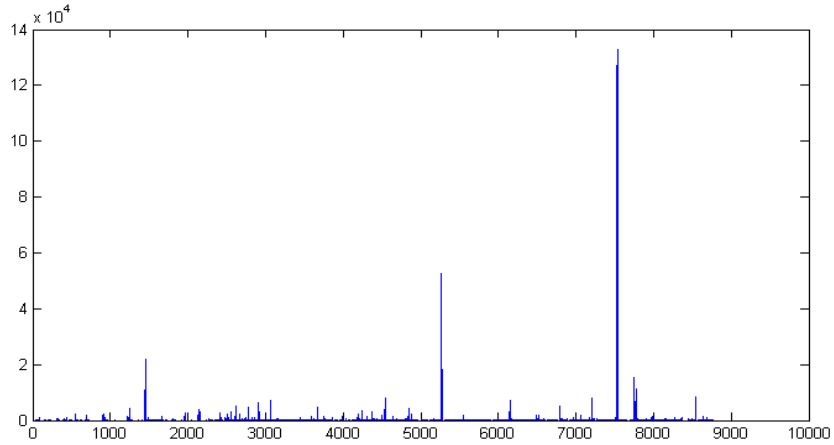
Netflows as Time Series

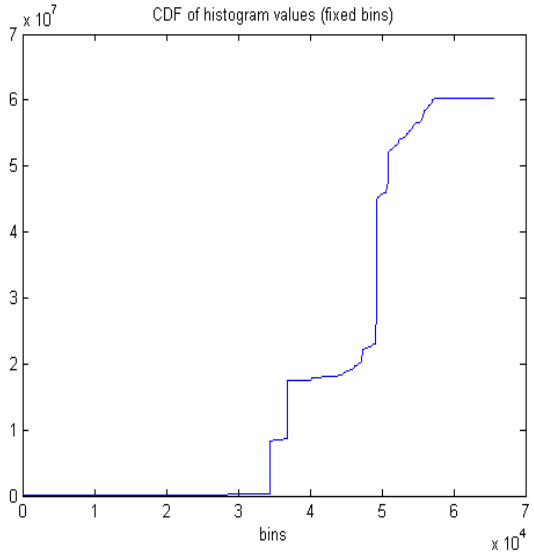


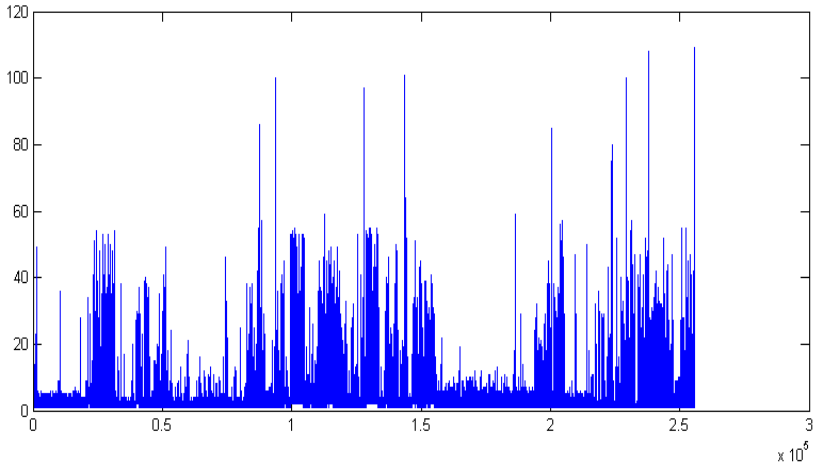
Transforming Netflows

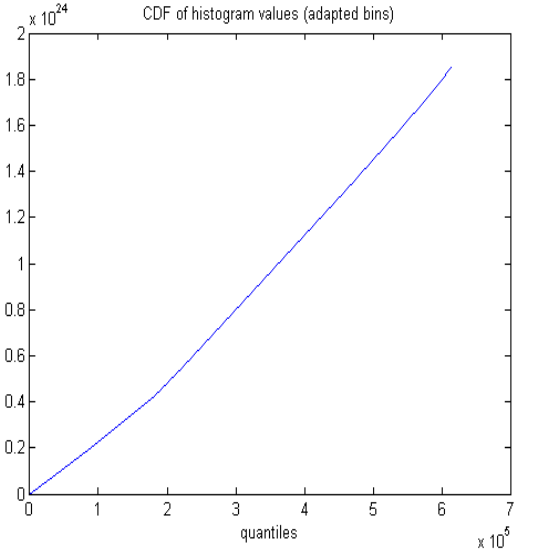
- Training – load sample of IP channels as composite 12-bit/52-bit keys
- Optimization - create the set of empirical quantiles using index keys in the training data
- Transform – use quantiles and binary search to split processing across workers, add or update values in matrix

Algorithm Results









Order of Complexity … Scalable!

Binary search

O(log n)

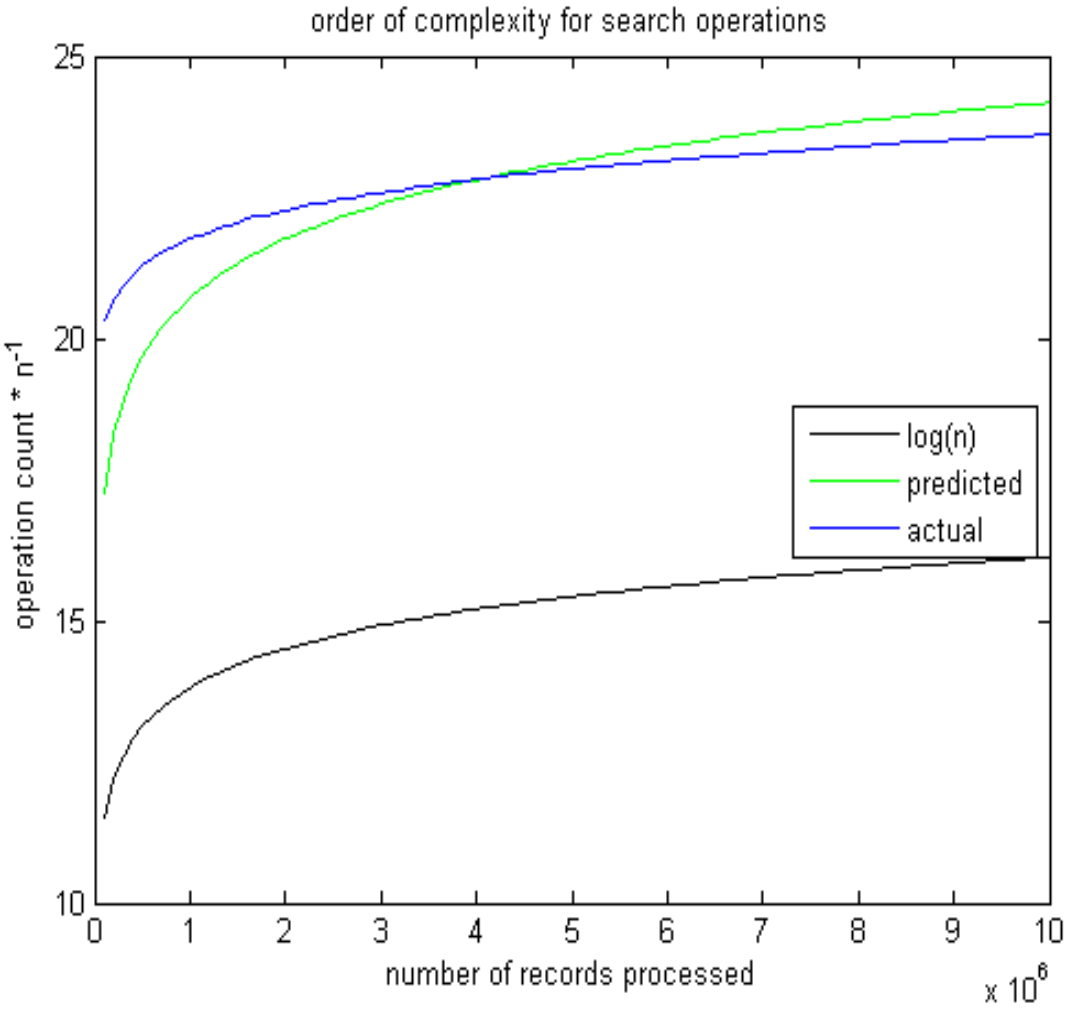
+ Direct search

O(c log n)

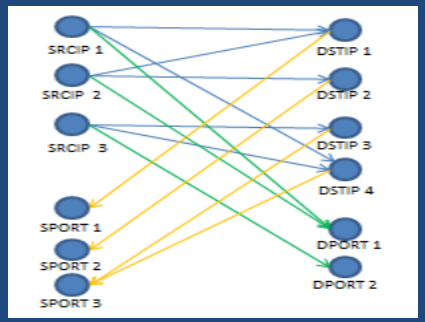
= Algorithm

O(n [1+c] log n)

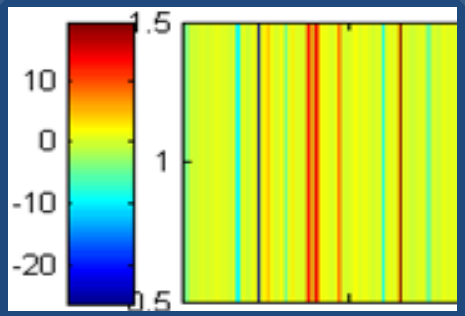
Compare to $O(n^2)$



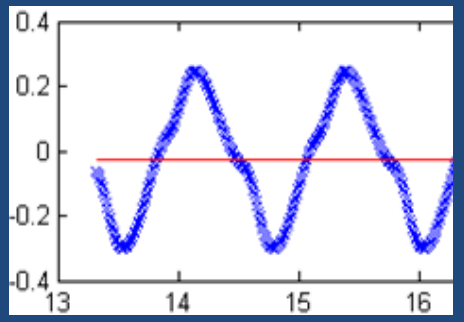
Analytic Approach



Network Analysis



Pattern Analysis

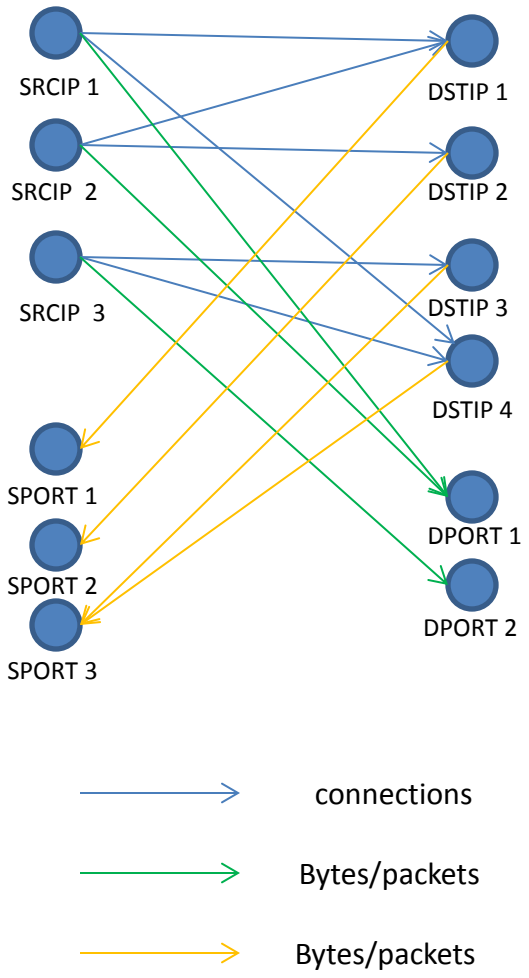


Signal Analysis

Visual Analysis

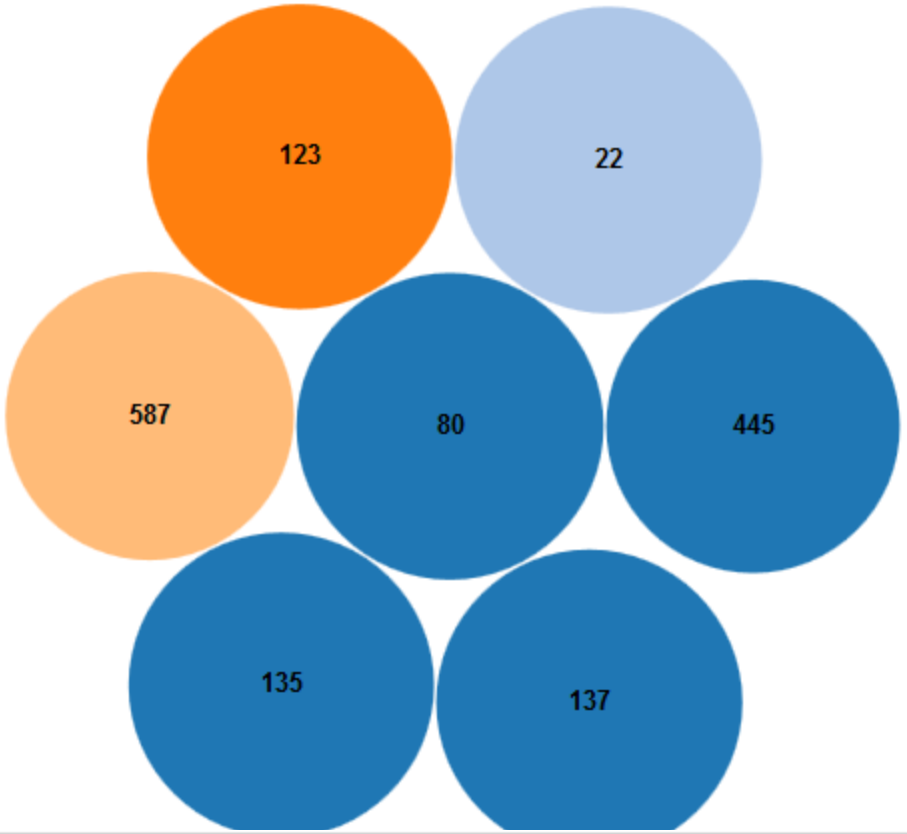
Graph Analysis: Latent Dirichlet Allocation

- Tries to put a population into sub-groups based on their similarity
- Used with documents and the words in them to suggest “topics”
- IP addresses are nodes, flow details are edges
- Use to cluster on known (profiling) or unknown (automated behavior)



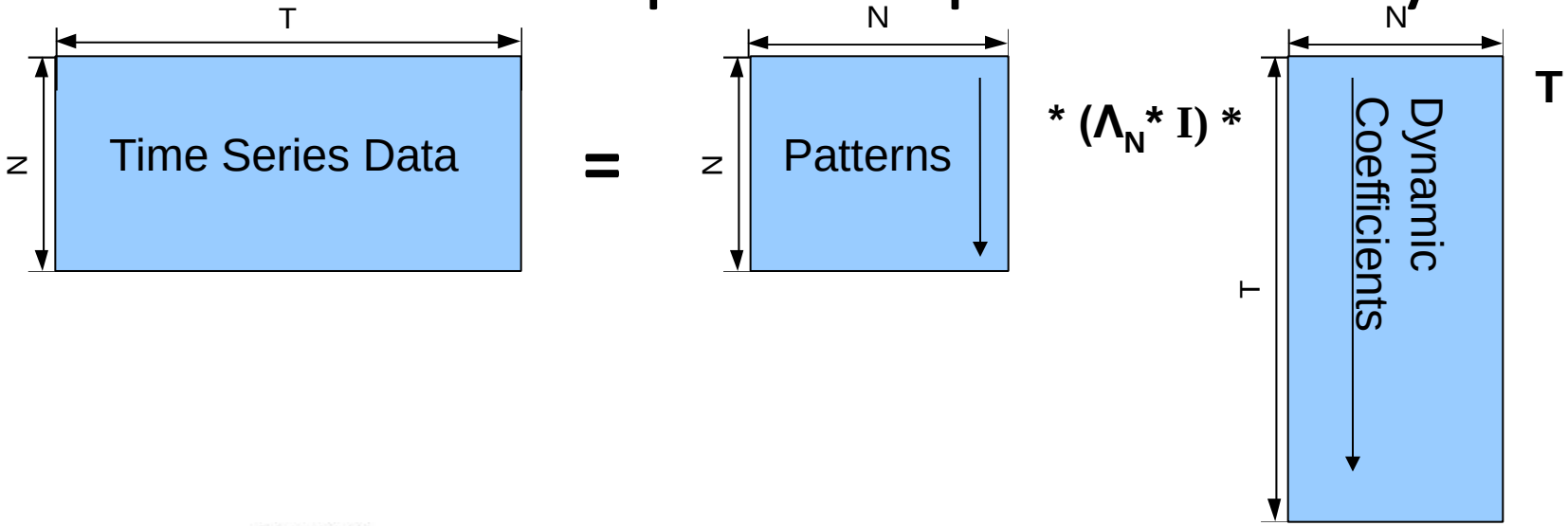
LDA results

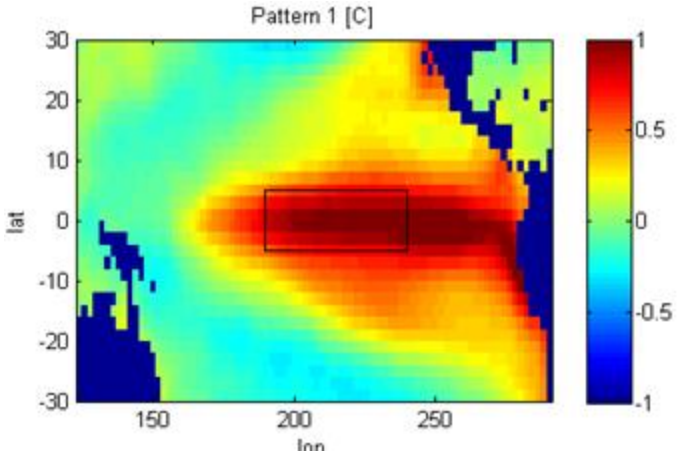
- Question: What are the strongest matches for groups based on automated communication to well-known ports ?
- Answer: Seven ports in four different groups are the strongest matches

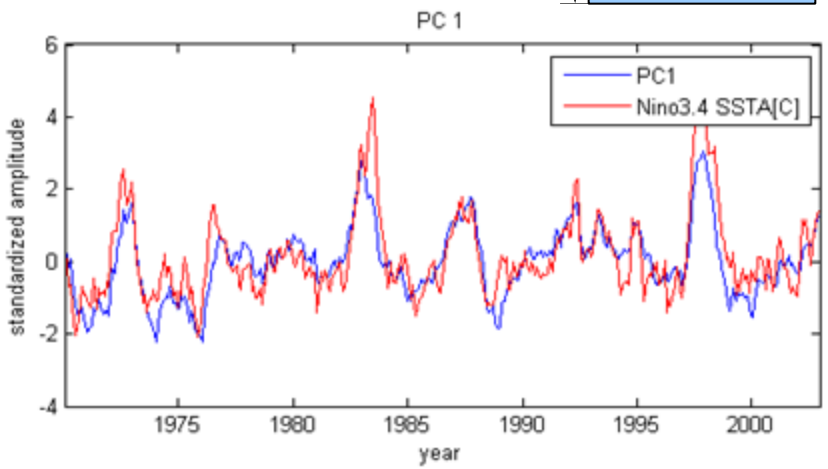



```
In [ ]: gname = 'netflow_topic'
        g = get_graph(gname)
        graph_result1 = g.query.gremlin("g.V.has('dport').has('lda_result',T.gte,0.9f).has('dport',T.lte,1024)")
        print 'results retrieved'
```

Patterns : Principal Component Analysis



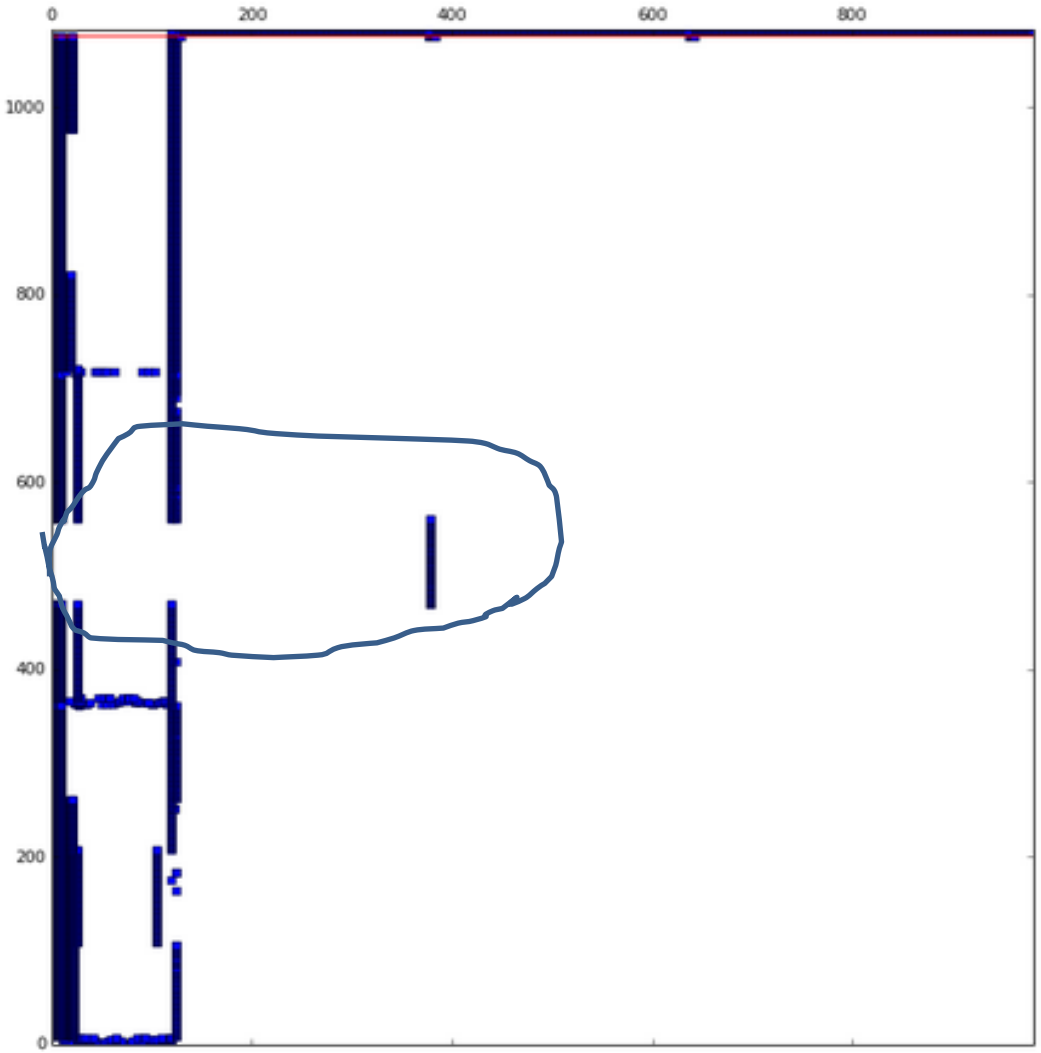




The Use of PCs to summarize … climatological fields has been found to be so valuable that is almost routine – *Joliffe, Principal Component Analysis*

PCA Results

- Question: Are there any anomalous patterns in this data?
- Answer: One source IP is talking to several destination IP's that do not exist (horizontal scan)



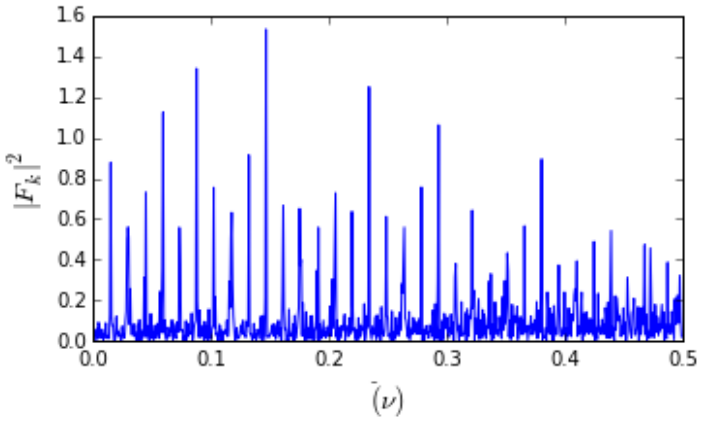
Signal Analysis: Fast Fourier Transform

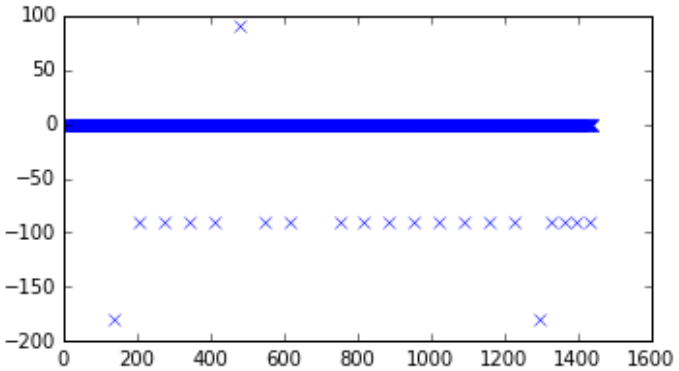
- Represent flow data as a function of sines and cosines (waves)
- Jump from time domain to frequency domain (and back)
- Easily filter noise from signal, or remove other frequencies

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos nx + b_n \sin nx \right) \qquad x \in (-\pi, \pi]$$

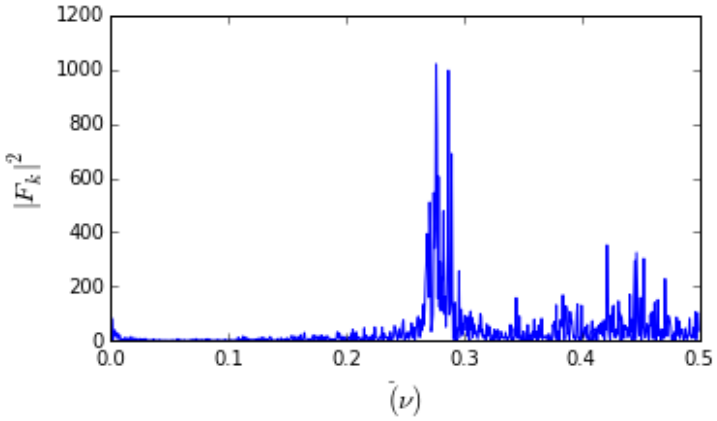
Signal Analysis - FFT

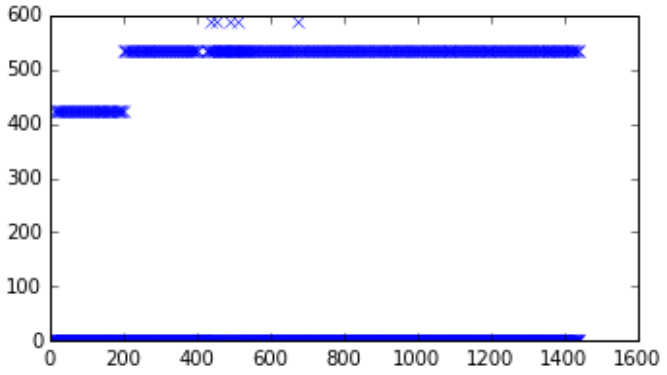
172.0.0.1 -> 172.30.0.3





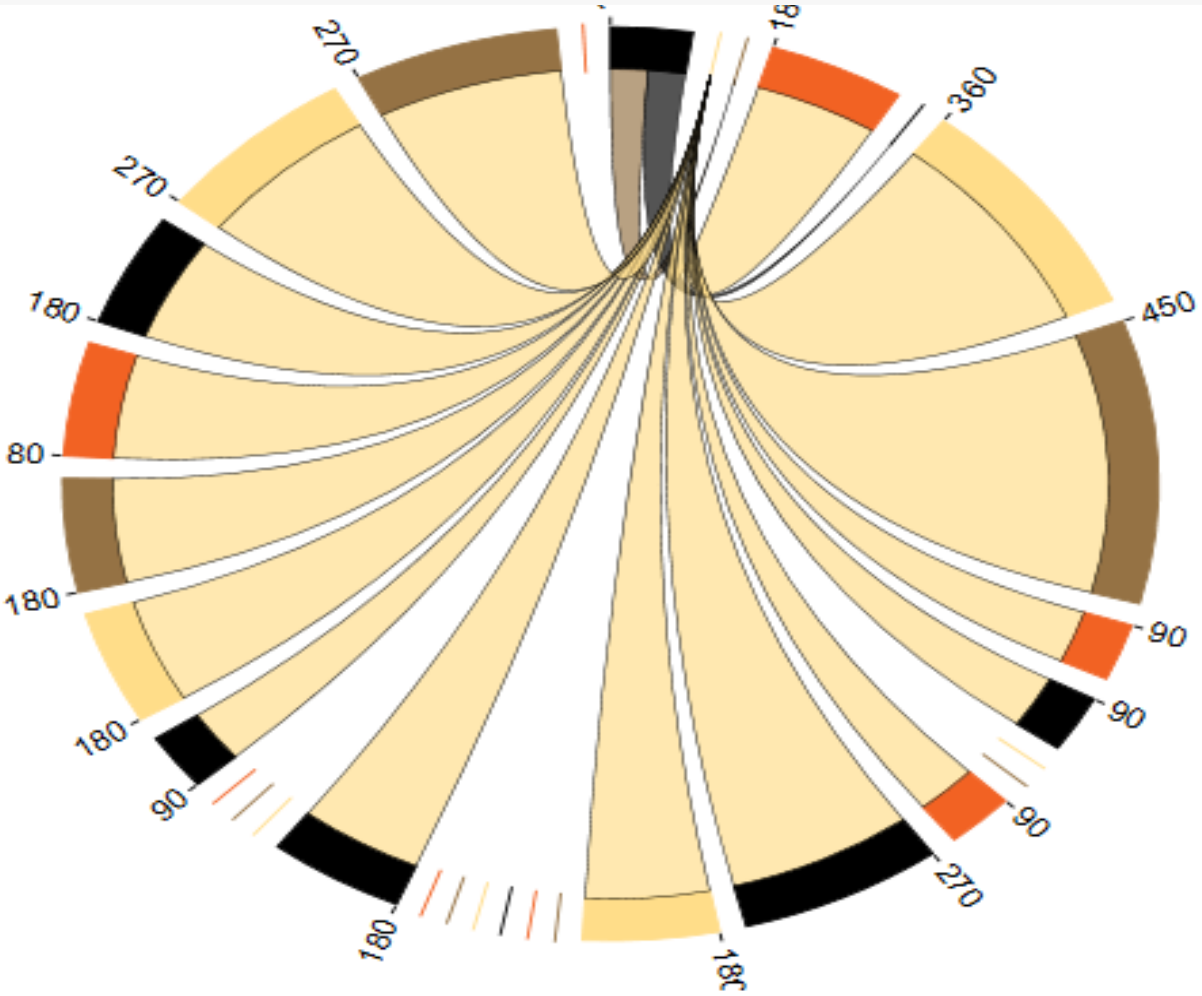
10.170.148.64 -> 172.20.0.3





Visual Analytics: IPython and D3

```
In [46]: %%javascript
         var viz = 'files/ipython/network.html'
         parent.document.getElementById('vizView').contentWindow.location.re
```



References

- Babb, Grant; Ross, Alan: *Increasing the Insight from Network Flows - Connecting Science to Operational Reality*, Draft Publication
- Kutz, J. Nathan: Data-Driven Modeling & Scientific Computation
- Joliffe, I. T.: Principal Component Analysis
- Blei, David M.: Introduction to Probabilistic Topic Models
- Chakravarty, Sambuddho et al: On the Effectiveness of Traffic Analysis Against Anonymity Networks Using Flow Records
- Cloudera Hadoop: http://cloudera.com
- Intel Analytics Toolkit: http://www.intel.com/content/www/us/en/software/intel-graph-solutions.html
- IPython, NumPy, Matplotlib: http://ipython.org
- SciPy: http://scipy.org
- D3: http://d3js.org

Questions?

Thanks