Structural Analysis of Network Traffic Flows

Eric Kolaczyk

Anukool Lakhina, Dina Papagiannaki, Mark Crovella, Christophe Diot, and Nina Taft
Traditional Network Traffic Analysis

- Focus on
  - Short ‘stationary’ timescales
  - Traffic on a single link in isolation

- Principal results
  - Scaling properties
  - Packet delays and losses

What ISPs Care About

- Focus on
  - Long, nonstationary timescales
  - Traffic on all links simultaneously

- Principal goals
  - Capacity planning
  - Traffic engineering
  - Anomaly detection
Need for *Whole-Network* Traffic Analysis

- **Traffic Engineering:** How does traffic move *throughout* the network?

- **Anomaly Detection:** *Which* links show unusual traffic?

- **Capacity planning:** How much and *where* in network to upgrade?
This is Complicated!

• Measuring and modeling traffic on all links simultaneously is challenging.
  – Even single link modeling is difficult
  – 100s of links in large IP networks
  – High-Dimensional timeseries

• Significant correlation in link traffic

• Is there a more fundamental representation?
Origin-Destination Flows

- Link traffic arises from the superposition of Origin-Destination (OD) flows
- Modeling OD flows instead of link traffic removes a significant source of correlation
- A fundamental primitive for whole-network analysis
But, This Is Still Complicated

- Even more OD flows than links
- Still a high dimensional, multivariate timeseries

- How do we extract meaning from this high dimensional structure in a systematic manner?
High Dimensionality: A General Strategy

• Look for good low-dimensional representations

• Often a high-dimensional structure can be explained by a small number of independent variables

• A commonly used technique: *Principal Component Analysis* (PCA) (aka KL-Transform, SVD, …)
Our work

• Measure complete sets of OD flow timeseries from two backbone networks

• Use PCA to understand their structure
  – Decompose OD flows into simpler features
  – Characterize individual features
  – Reconstruct OD flows as sum of features

• Call this *structural analysis*
Datasets

• **Abilene**: 11 PoPs, 121 OD flows.
• **Sprint-Europe**: 13 PoPs, 169 OD flows.
• Collect sampled traffic from every ingress link using NetFlow
• Use BGP tables to resolve egress points
• Week-long datasets, 5- or 10-minute timesteps
Example OD Flows

Some have visible structure, some less so…
Specific Questions of Structural Analysis

• Are there low dimensional representations for a set of OD flows?
• Do OD flows share common features?
• What do the features look like?
• Can we get a high-level understanding of a set of OD flows in terms of these features?
Principal Component Analysis

Coordinate transformation method

Original Data

Transformed Data

\[
\begin{bmatrix}
x_1, x_2 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
u_1, u_2 \\
\end{bmatrix}
\]
Properties of Principle Components

• Each PC in the direction of maximum (remaining) energy in the set of OD flows
  • Ordered by amount of energy they capture

• *Eigenflow*: set of OD flows mapped onto a PC; a common trend
  • Ordered by most common to least common trend
PCA on OD flows

\[ X = UV^T \]
PCA on OD flows (2)

$$u_i = \frac{X v_i}{\sigma_i} \quad i = 1, \ldots, p$$

Each eigenflow is a weighted sum of all OD flows

$$E = \begin{bmatrix} \vdots \end{bmatrix}$$

Eigenflows are orthonormal

$$\|X v_i\| = \lambda_i \quad ; \quad \sigma_i = \sqrt{\lambda_i}$$

Singular values indicate the energy attributable to a principal component

$$\frac{X_i}{\sigma_i} = U (V^T)_i \quad i = 1, \ldots, p$$

Each OD flow is weighted sum of all eigenflows

$$E = \begin{bmatrix} \vdots \end{bmatrix} + \begin{bmatrix} \vdots \end{bmatrix} + \begin{bmatrix} \vdots \end{bmatrix}$$
An Example Eigenflow and PC
Outline For Rest of Talk

• Find intrinsic dimensionality of OD flows
  • Decompose OD flows
  • Characterize eigenflows
  • Reconstruct OD flows

• Potential applications
Low Intrinsic Dimensionality of OD Flows

Plot of (square root of) energy captured by each dimension.
Approximating With Top 5 Eigenflows

![Graph showing traffic in OD Flow 88 over a week, comparing original to 5PC approximations. The x-axis represents days of the week, and the y-axis represents traffic in millions.]
Approximating With Top 5 Eigenflows

Traffic in OD Flow 79

Original
5 PC
Approximating With Top 5 Eigenflows
Outline

• Find intrinsic dimensionality of OD flows
• Decompose OD flows
• Characterize eigenflows
• Reconstruct OD flows

• Potential applications
Structure of OD Flows

Most OD flows have less than 20 significant eigenflows

Can think of each OD flow as having only a small set of “features”
Kinds of Eigenflows

Deterministic d-eigenflows
- Predictable (periodic) trends

Spike s-eigenflows
- Sudden, isolated spikes and drops

Noise n-eigenflows
- Roughly stationary and Gaussian
D-eigenflows Have Periodicity

Power spectrum
S-eigenflows Have Spikes

Sprint–1 Eigenflow 8

Abilene Eigenflow 10

5-sigma threshold
N-eigenflows Are Gaussian

qq-plot
Hundreds of Eigenflows
But Only Three Basic Types

(a) Sprint
(b) Abilene
An OD Flow, Reconstructed

OD flow
D-components
S-components
N-components
Another OD Flow, Reconstructed

OD flow

D-components

S-components

N-components
Which Eigenflows Are Most Significant?

1-6: $d$-eigenflows appear to be most significant in both networks.

5-10: $s$-eigenflows are next important.

12 and beyond: $n$-eigenflows account for rest.
Contribution of Eigenflow Types

Fraction of total OD flow energy captured by each type of eigenflow

<table>
<thead>
<tr>
<th>Eigenflow type</th>
<th>Sprint-1</th>
<th>Abilene</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-eigenflow</td>
<td>92.17%</td>
<td>69.79%</td>
</tr>
<tr>
<td>s-eigenflow</td>
<td>5.59%</td>
<td>18.60%</td>
</tr>
<tr>
<td>n-eigenflow</td>
<td>2.24%</td>
<td>11.61%</td>
</tr>
</tbody>
</table>
Contribution to Each OD Flow (Sprint)

Largest OD flows: Strong deterministic component.

Smallest OD flows: Primarily dominated by spikes.

Regardless of size, n-eigenflows account for a fairly constant portion.
Contribution to Each OD Flow (Abilene)

Largest OD flows: Strong deterministic component.

Smallest OD flows: Dominated by noise, but have diurnal trends also.

Regardless of size, spikes account for a fairly constant portion.
Summary: Specific Questions

• Are there low dimensional representations for a set of OD flows?
  – 5-10 eigenflows is sufficient for good approximation of a set of 100+ OD flows

• Do OD flows share common features?
  – The common features across OD flows are eigenflows

• What do the features look like?
  – Each eigenflow can be categorized as D, S, or N

• Can we get a high-level understanding of a set of OD flows in terms of these features?
  – Both networks: Large flows are primarily diurnal
  – Sprint: Small flows are primarily spikes; noise constant.
  – Abilene: Small flows have N and D; spikes constant.
Outline

• Find intrinsic dimensionality of OD flows
• Decompose OD flows
• Characterize eigenflows
• Reconstruct OD flows

• Potential applications
Traffic Matrix Estimation

Problem Statement:
Infer OD flows \((X)\) given link measurements \((Y)\) and routing matrix \((A)\): \(Y^T = AX^T\)

State of the Art:
dim\((X)\) > dim\((Y)\), so treat as ill-posed linear inverse problem. Infer \(Y\) on stationary (short) timescales.

Possible Approach:
On longer timescales, intrinsic dimensionality of OD flows is small, so effective \(\text{dim}(X) < \text{dim}(Y)\)
TM estimation of largest eigenflows now becomes a “well-posed” problem.
Anomaly Detection

State of the art:
Use wavelets to detrend each flow in isolation.  
[Barford:IMW02]

Possible approach:
Detrend all OD flows simultaneously by subtracting d-eigenflows.
Traffic Forecasting

State of the art:
Treat each flow timeseries independently.
Use wavelets to extract trends.
Build timeseries forecasting models on trends.
[PTZC:INFOCOM03]

Possible approach:
Build forecasting models on d-eigenflows as trends.
Allows simultaneous examination and forecasting for entire ensemble of OD flows.
Problem Statement:
How does one identify important traffic flows, so that they can be treated differently?

State of the art:
Measure all flows on a single link
Find “heavy-hitters” or “elephant” flows based on preset thresholds [PTC:INFOCOM04, PTBTSC:IMW02]

Possible approach:
Look across all flows and extract common features
Taxonomize each flow into D, S, or N
Final thoughts

OD flows a useful primitive to engineer networks

Set of OD flows have low dimensional representations

A Structural Analysis approach can provide useful insight into nature of OD flows
Thanks!

• Help with Abilene Data  
  • Rick Summerhill, Mark Fullmer (Internet2)  
  • Matthew Davy (Indiana University)

• Help with Sprint-Europe Data  
  • Bjorn Carlsson, Jeff Loughridge (SprintLink),  
  • Richard Gass (Sprint ATL)
Backup slides
Principal Component Analysis

For any given dataset, PCA finds a new coordinate system that maps maximum variability in the data to a minimum number of coordinates.

New axes are called Principal Axes or Components.
Properties of Principle Components

• Each PC in the direction of maximum (remaining) energy in the set of OD flows
  • Ordered by amount of energy they capture

\[ v_1 = \arg \max_{||v||=1} \| Xv \| \]

and,

\[ v_k = \arg \max_{||v||=1} \| (X - \sum_{i=1}^{k-1} Xv_i v_i^T)dv \| . \]

• **Eigenflow**: set of OD flows mapped onto a PC; a common trend
  • Ordered by most common to least common trend
Energy captured by each PC

![Energy captured by each PC](image)