Structural Analysis of Network Traffic Flows

Mark Crovella

Anukool Lakhina, Konstantina Papagiannaki, Christophe Diot, Eric D. Kolaczyk, Nina Taft
Traditional Network Traffic Analysis

• Focus on
  – Short ‘stationary’ timescales
    • An hour or less
  – Traffic on a single link in isolation

• Principal results
  – Scaling properties
  – Packet delays and losses

What ISPs Care About

• Focus on
  – Long, nonstationary timescales
    • Hours to weeks or months
  – Traffic on all links simultaneously

• Principal goals
  – Traffic engineering
  – Anomaly detection
  – Capacity planning
Need For *Whole-Network* Traffic Analysis

- Traffic Engineering
  - How does traffic move *throughout* the network?
- Attack / Anomaly Detection
  - On *which* links is there 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.
  – Hundreds to thousands of links in a large IP backbone network
  – Even single link modeling is difficult

• High-dimensional timeseries
  – “Curse of dimensionality”
  – Significant correlation structure

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

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

• Each OD flow serves a different customer population
  – No two OD flows carry same traffic
  – Are they still correlated?

• Even more OD flows than links
  – High dimensional, multivariate timeseries
  – Still have curse of dimensionality

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

- Look for a *low-dimensional* representation preserving the most important features of data
- Often, a high-dimensional structure may explainable in terms of a small number of independent variables
- Commonly used tool: *Principal Component Analysis* (PCA)
Our Work

• Capture complete sets of OD flow timeseries from two backbone networks
• Use Principal Component Analysis to systematically understand them
  – Decompose OD flows into smaller set of features
  – Characterize the individual features
  – Reconstruct each OD flow as a sum of features
• Refer to this as “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

• 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

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 Principal Components

Each PC points in the direction of maximum (remaining) energy in the data:

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

and,

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

Set of flows mapped onto a single PC is called an eigenflow.
PCA on OD flows

$X=U\Sigma V^T$
PCA on OD flows (2)

Each eigenflow is a weighted sum of all OD flows

Eigenflows are orthonormal

Singular values indicate the energy attributable to a principal component

Each OD flow is weighted sum of all eigenflows
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

A plot of the singular values reveals how much energy is captured by each PC. Sharp elbow indicates that most of the energy captured by 5-10 singular values, for all datasets.
Approximating With Top 5 Eigenflows

\[ X' = U' \Sigma V^T \]
Approximating With Top 5 Eigenflows

Traffic in OD Flow 79

- Original
- 5 PC
Approximating With Top 5 Eigenflows

Traffic in OD Flow 96

- Original
- 5 PC
Outline

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

• Potential applications
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

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: \textit{d-eigenflows} appear to be most significant in both networks.

5-10: \textit{s-eigenflows} are next important.

12 and beyond: \textit{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

**Largest OD flows:**
Strong deterministic component.

**Smallest OD flows:**
Primarily dominated by noise.

Regardless of size, s-eigenflows account for a fairly constant portion.
Contribution to Each OD Flow

(Abilene)

Fraction of Total Energy

OD Flow (large to small)
Summary: Specific Questions

• Are there low dimensional representations for a set of OD flows?
  – 5 or 6 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?
  – High volume flows tend to be dominated by D-eigenflows
  – Low volume flows tend to be dominated by N-eigenflows
  – S-eigenflows contribute across all OD flows
Outline

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

• Potential applications

Structural Analysis
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 dim($X$) < 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.
[Papagiannaki:INFOCOM03]

Possible approach:
Build forecasting models on d-eigenflows as trends.
Allows simultaneous examination and forecasting for entire ensemble 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)
How OD flows differ

Largest OD flows contain most significant eigenflows

Smallest OD flows contain least significant eigenflows

Constituting eigenflows clustered in a tight range