The Subspace Method for Diagnosing Network-Wide Traffic Anomalies

Anukool Lakhina, Mark Crovella, Christophe Diot

What’s happening in my network?

- Is my customer being attacked? probed? infected?
- Is there a sudden traffic shift?
- An external route change?
- A routing loop?
- An equipment outage?

Automated methods for reliably and generally answering such questions are lacking

A General Framework

- We can treat all such problems as special cases of the general question:

  Is my network experiencing unusual conditions?

- Then, adopt the following framework:
  - **Detection**
    Is there an unusual event?
  - **Identification**
    Which of the possible explanations fits best?
  - **Quantification**
    How serious is the problem?

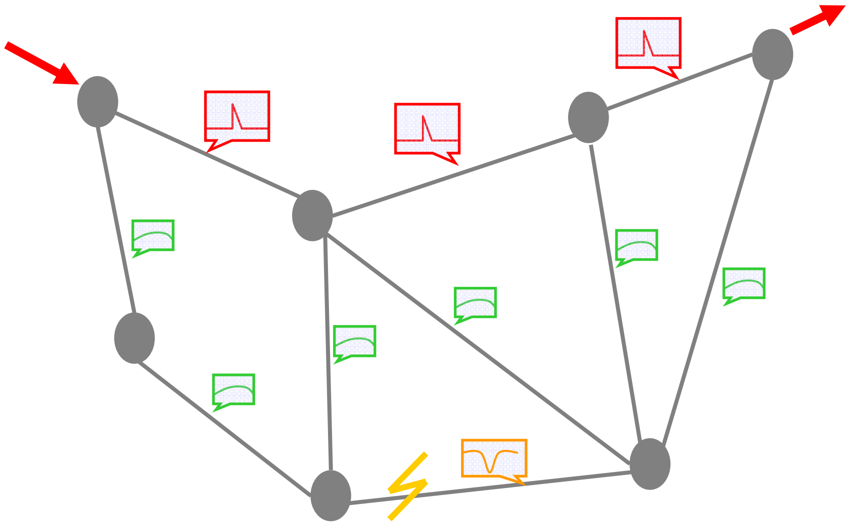
Statistical Approach

The advantage of such a framework is that it lends itself to a statistical approach:

- **Detection:** Outlier detection
- **Identification:** Hypothesis testing
- **Quantification:** Estimation

Anomaly Diagnosis

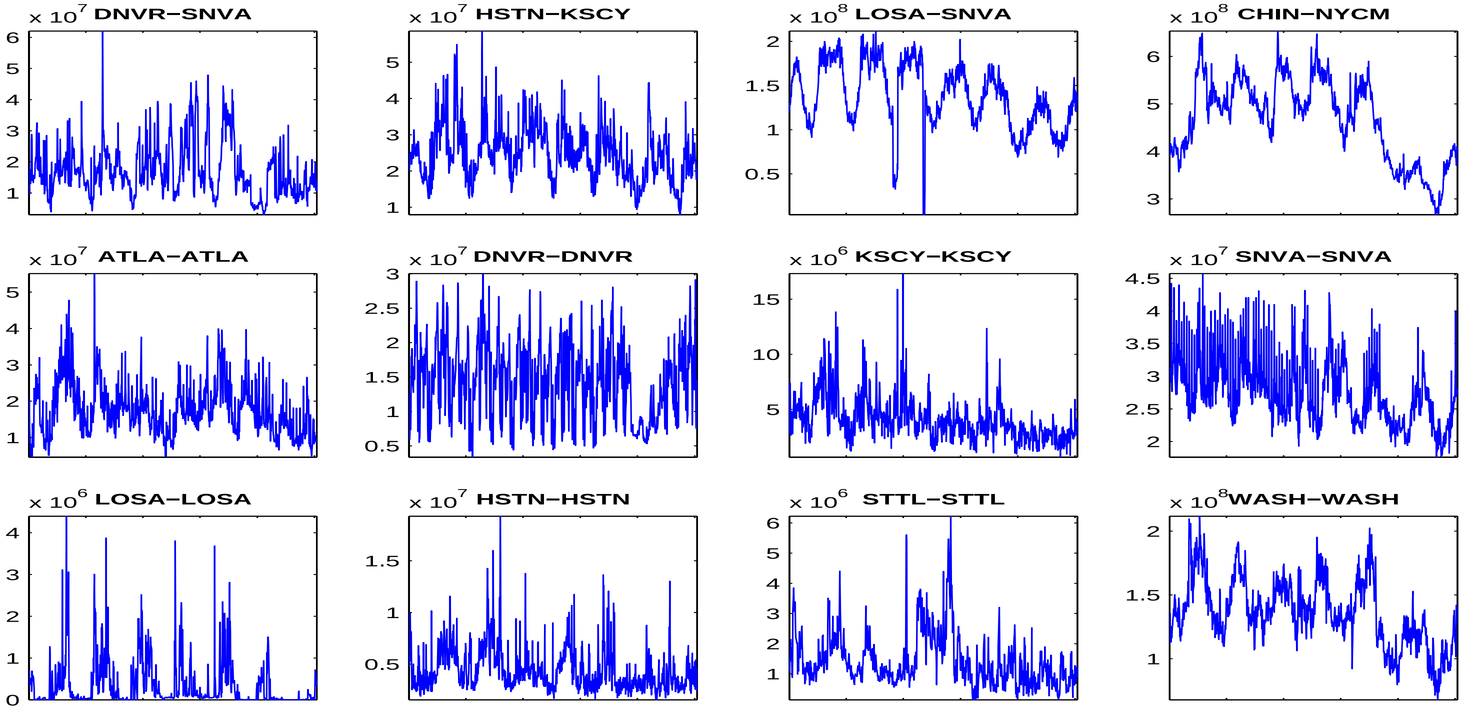
A Need for Whole-Network Diagnosis



<u>Our Thesis:</u> Effective diagnosis of network anomalies requires a **whole-network** approach

For example, diagnosing traffic anomalies requires analyzing traffic from all links

But, This Is Difficult!



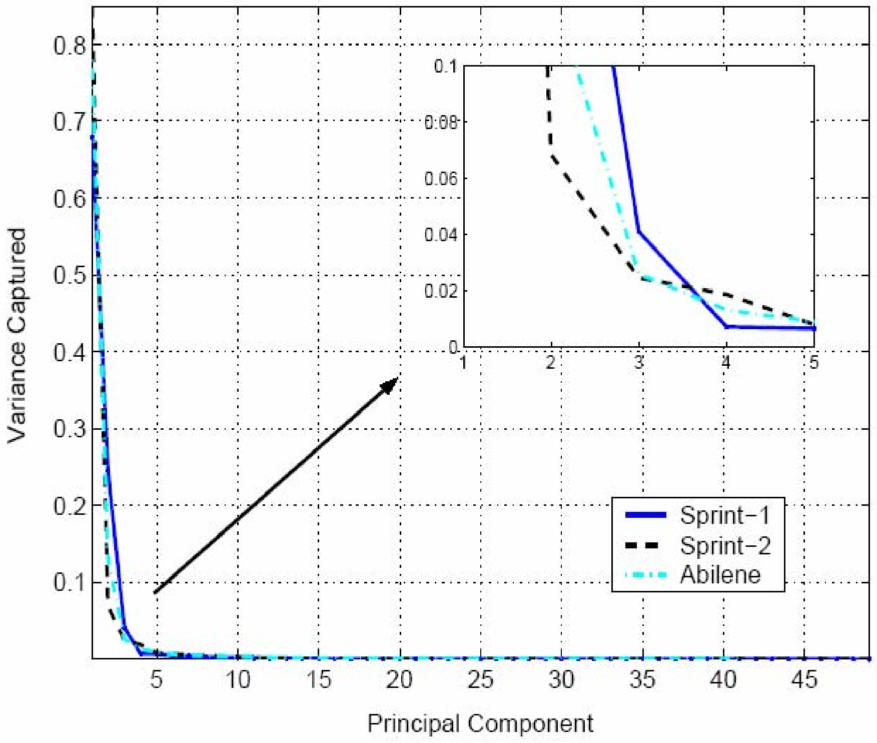
How do we extract **meaning** from such a **high-dimensional** data in a systematic manner?

Low Intrinsic Dimensionality of Link Traffic

Studied via Principal Component Analysis

Key result:
Normal traffic is well approximated by a low dimensional space

For example:
Traffic on 40+ links is well approximated in space of only 4 dimensions



Reasons for Low Dimensionality of Traffic

- Generally, traffic on different links is not independent
- Link traffic is the superposition of origin-destination flows (OD flows)
  - The same OD flow passes over multiple links, inducing correlation among links
  - All OD flows tend to vary according to common daily and weekly cycles, and so are themselves correlated
    [See SIGMETRICS 2004 paper]

The Subspace Method

- An approach to separate normal from anomalous traffic
- Define $\mathcal{S}$ as the space spanned by the first *k* principal components
- Define $\tilde{\mathcal{S}}$ as the space spanned by the remaining principal components
- Then, decompose traffic on all links by projecting onto $\mathcal{S}$ and $\tilde{\mathcal{S}}$ to obtain:

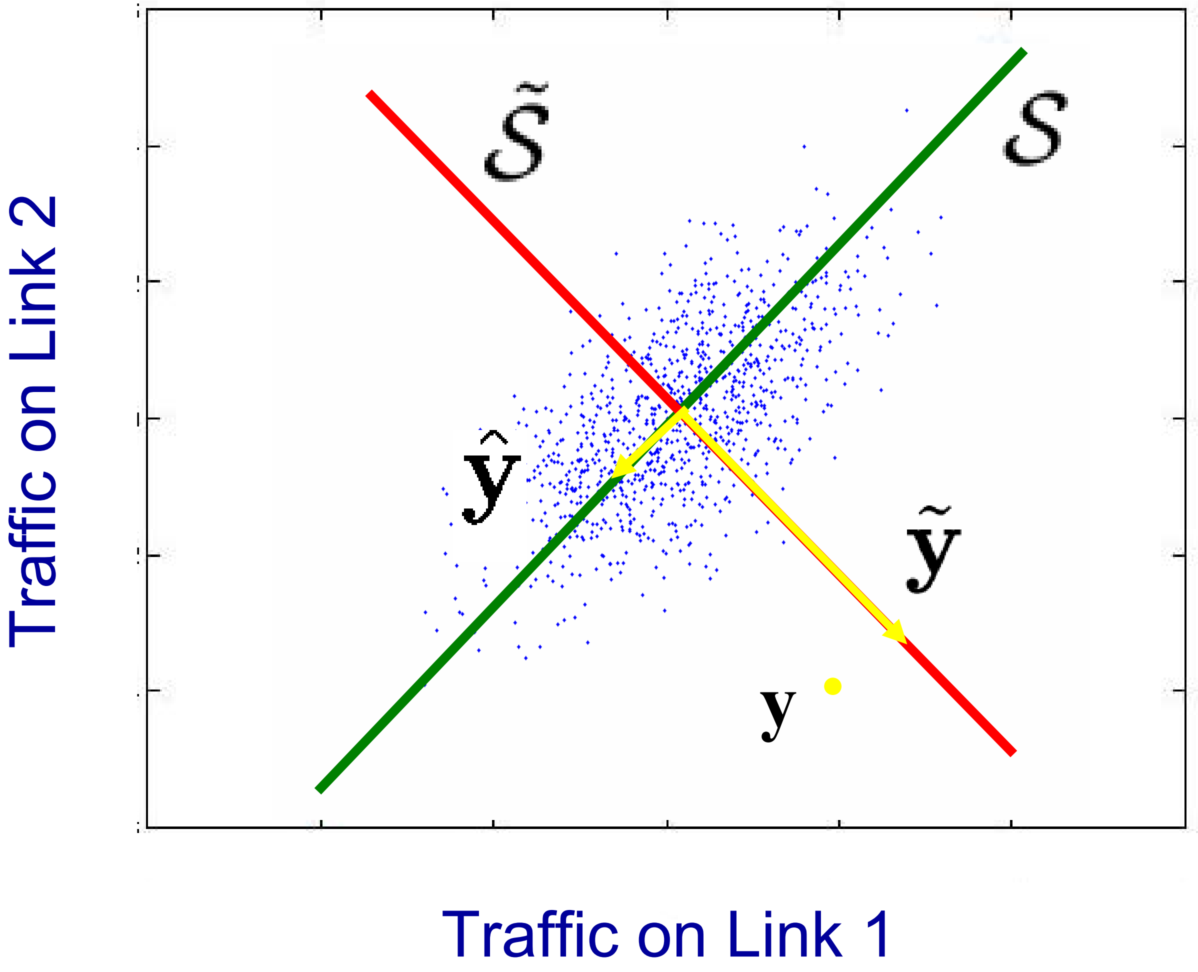
$$\mathbf{y} = \hat{\mathbf{y}} + \tilde{\mathbf{y}}$$

Traffic vector of all links at a particular point in time

Normal traffic vector

Residual traffic vector

The Subspace Method, Geometrically



In general, anomalous traffic results in a large value of $\tilde{\mathbf{y}}$

$$\hat{\mathbf{y}} = \mathbf{C}\mathbf{y}$$

$$\tilde{\mathbf{y}} = \tilde{\mathbf{C}}\mathbf{y}$$

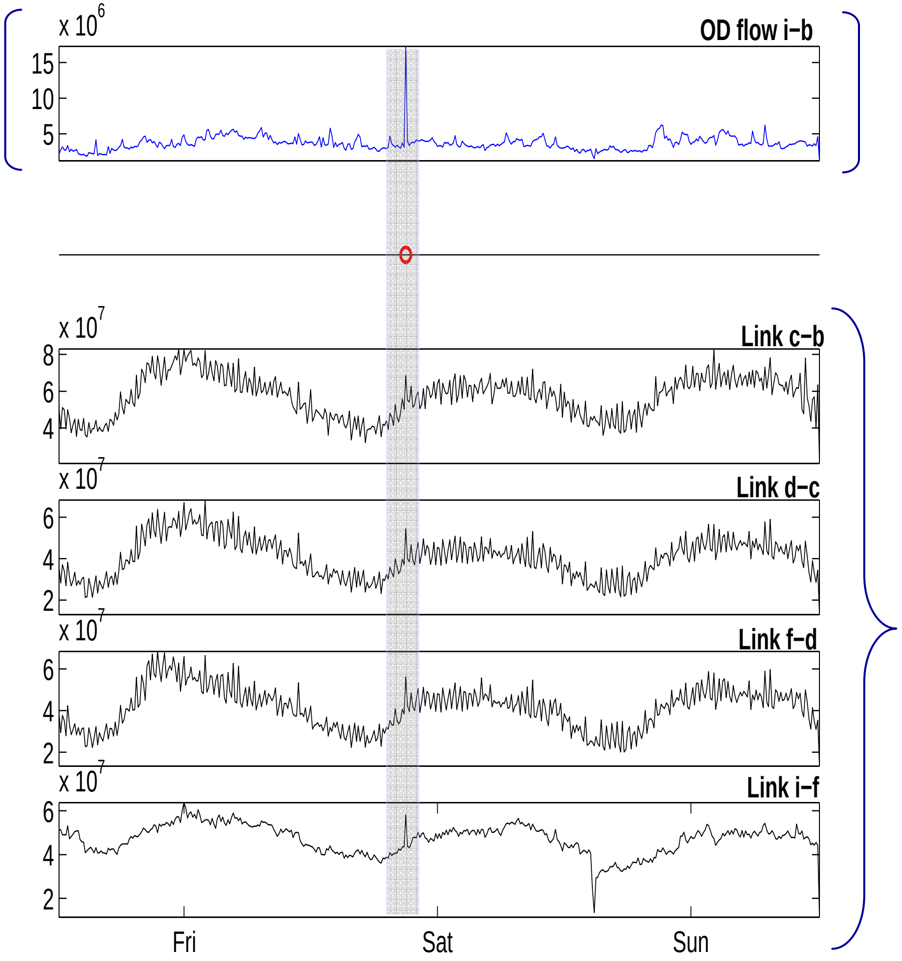
Outline

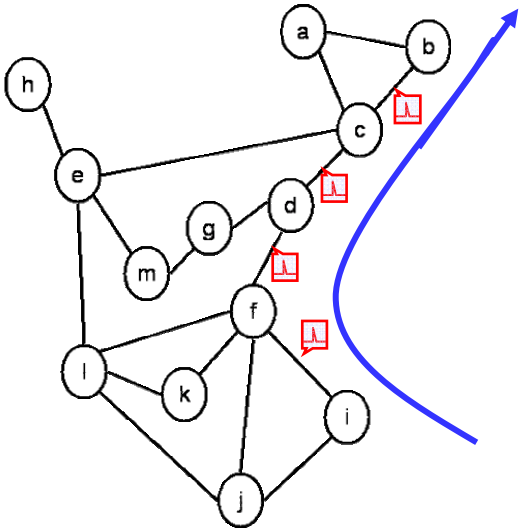
- Subspace Method applied to Link Traffic
  - Problem: Volume Anomaly Diagnosis
  - Detection, Identification, Quantification
  - Validation
- Subspace Method applied to Flow Traffic
  - Problem: General Anomaly Detection
  - Sample Results
- Conclusions

Diagnosing Volume Anomalies

- A ***volume anomaly*** is a sudden change in an OD flow's traffic (*i.e.*, point to point traffic)
- Problem Statement:
  Given link traffic measurements, diagnose the volume anomalies
- A first application of the subspace method

An Illustration





Sprint-Europe Backbone Network

The *Diagnosis Problem* requires analyzing traffic on all links to:

1) **Detect** the time of the anomaly

2) **Identify** the source & destination

3) **Quantify** the size of the anomaly

Subspace Method: Detection

- Error Bounds on Squared Prediction Error:

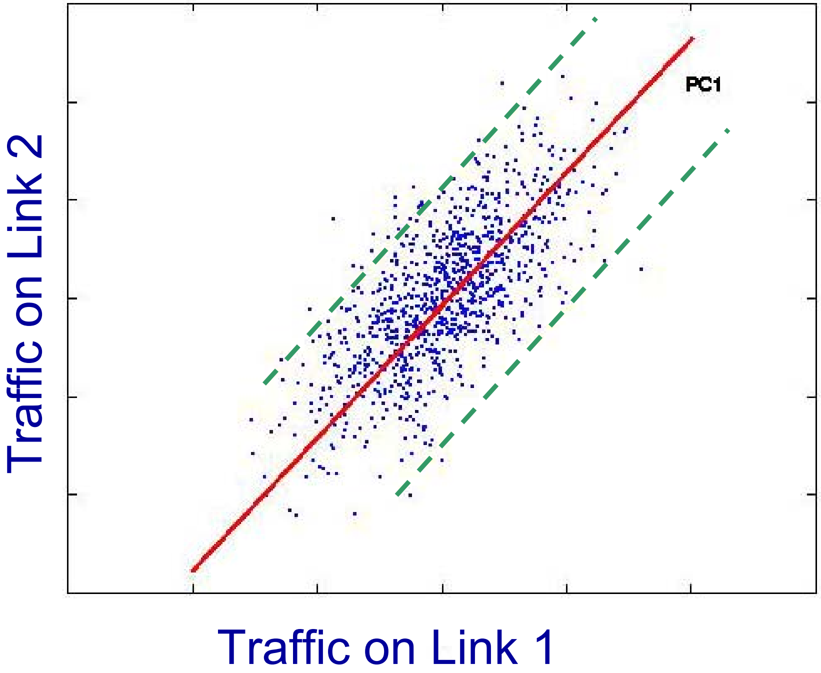
$$\mathrm{SPE} \equiv \|\tilde{\mathbf{y}}\|^2 = \|\tilde{\mathbf{C}}\mathbf{y}\|^2$$

- Assuming multivariate Gaussian data, traffic is normal when,

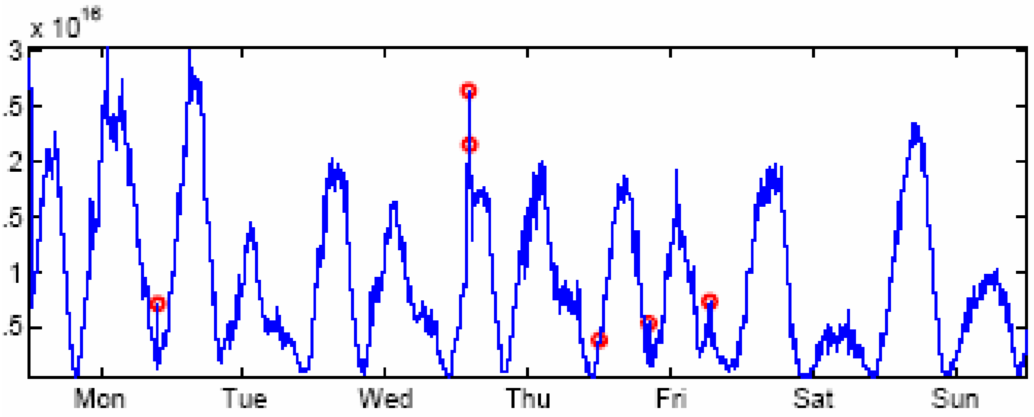
$$\mathrm{SPE} \leq \delta_\alpha^2$$

Result due to

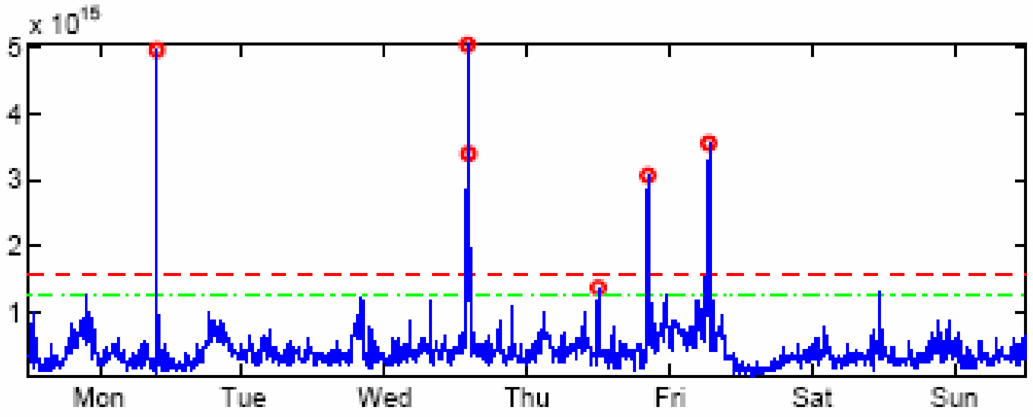
[Jackson and Mudholkar, 1979]



SPE vs. All Traffic



Value of $||\mathbf{y}||^2$ over time

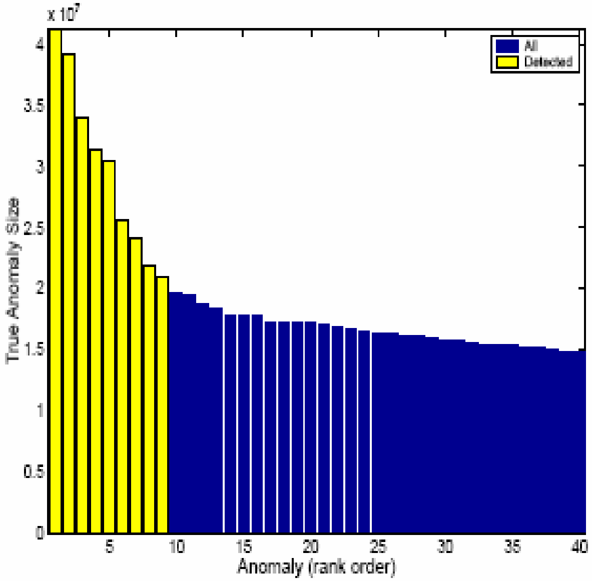


Value of $||\tilde{\mathbf{y}}||^2$ over time

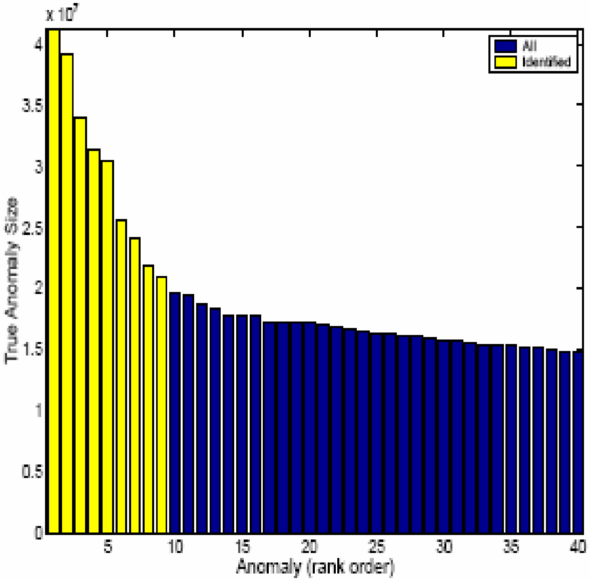
SPE $(||\tilde{\mathbf{y}}||^2)$ at anomaly time points clearly stand out

Results on True Anomalies: Sprint-1

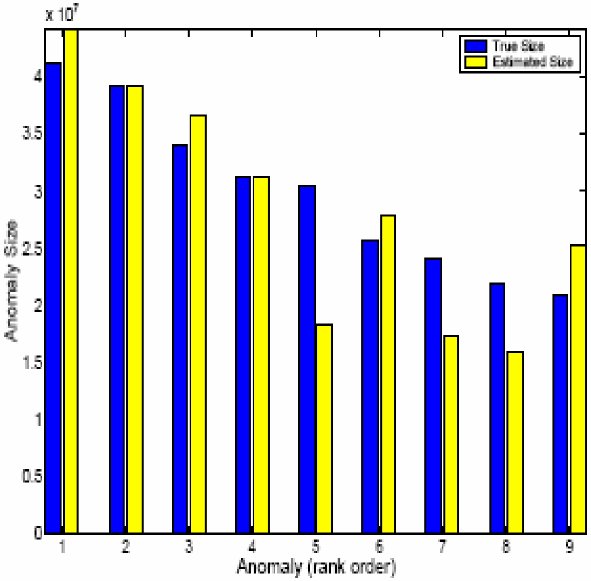
40 Largest deviations in OD flows via Fourier



Detection



Identification



Quantification

"Knee" in curve - natural cutoff for detection

Outline

- Subspace Method applied to Link Traffic
  - Problem: Volume Anomaly Diagnosis
  - Detection, Identification, Quantification
  - Validation
- Subspace Method applied to OD Flow Traffic
  - Problem: General Anomaly Detection
  - Sample Results
- Conclusions

Beyond Volume Anomalies

- Volume anomalies: important, but not the entire set of anomalies of interest to operators.

- Operators are also interested in:
  - DOS attacks, flash crowds, port scans, worm propagation, network equipment outages, changes in ingress/egress traffic patterns, ...

- Link data doesn't seem to hold enough information to accurately detect such a wide range of anomaly types.

- Therefore, we turn to IP flow data

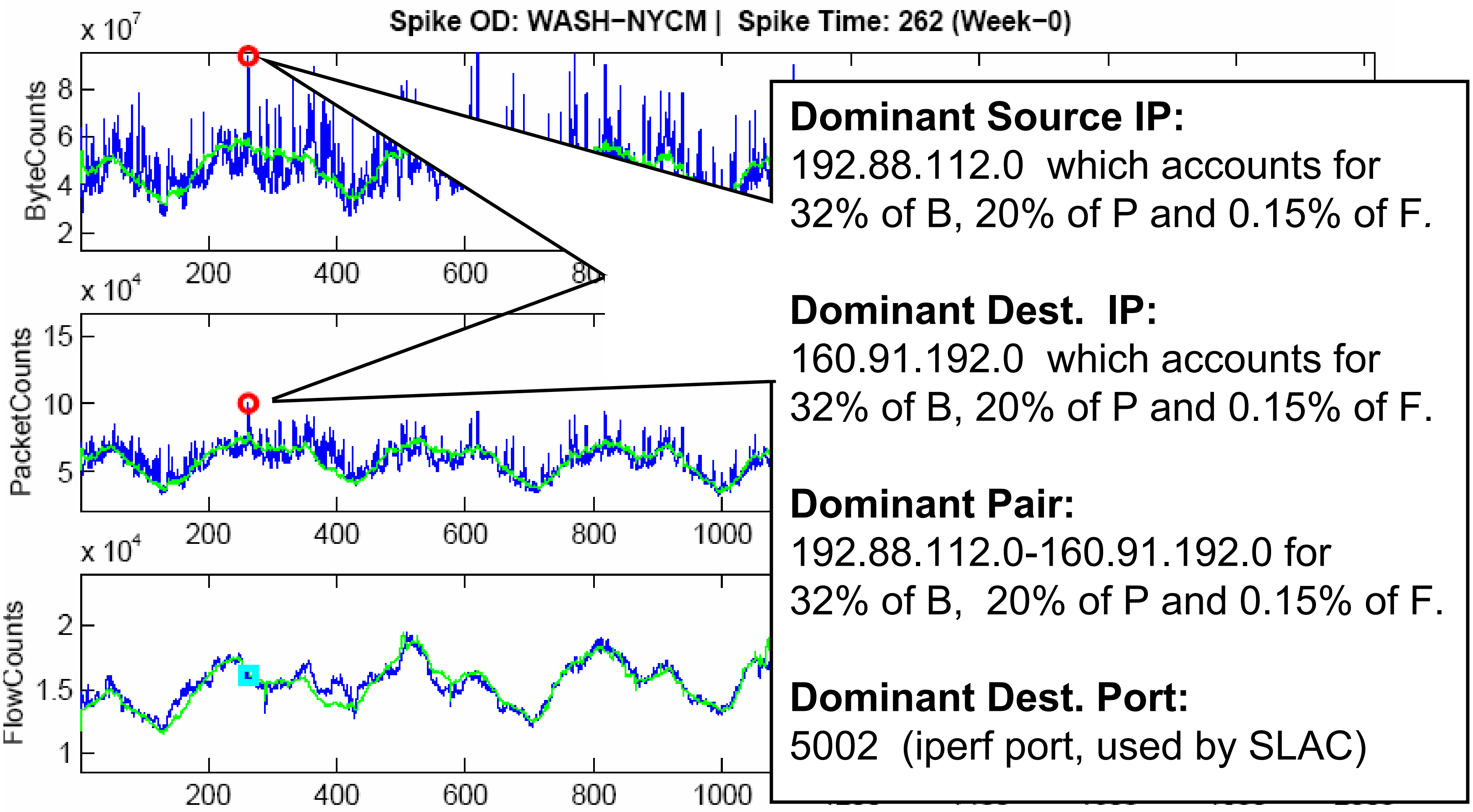
Characterization Methodology

- Extend subspace method to diagnose anomalies directly in OD flow traffic timeseries
  - Detection in both $\mathcal{S}$ and $\tilde{\mathcal{S}}$ subspaces

- Examine OD flow traffic as three separate views: **# Bytes, # Packets, # IP-flows**

- Manually inspect each anomaly found over 4 week period in Abilene network
  - Using 5-tuple headers of sampled flow data

An example BP anomaly (heavy flow)



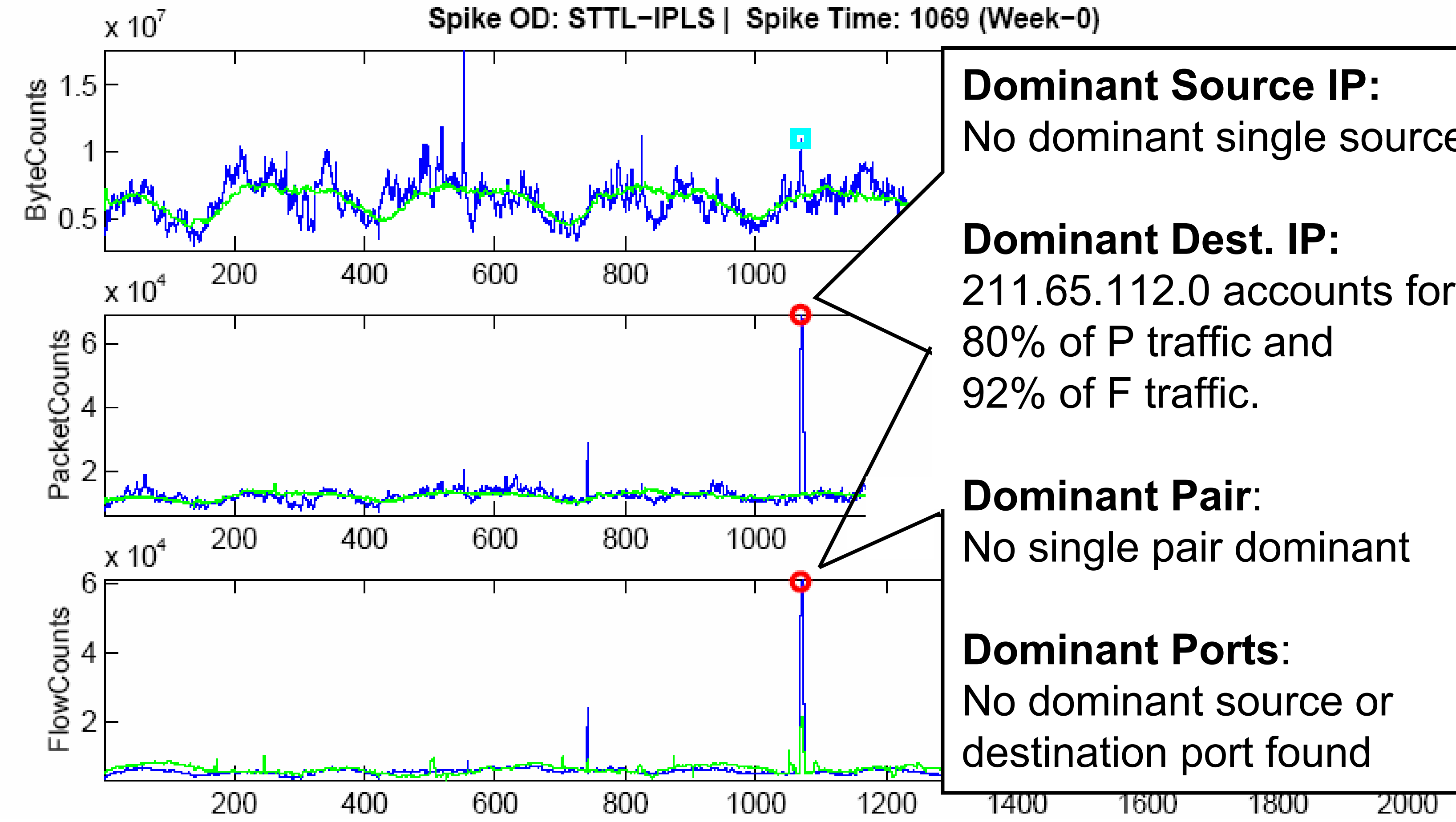
Dominant Source IP:
192.88.112.0 which accounts for
32% of B, 20% of P and 0.15% of F.

Dominant Dest. IP:
160.91.192.0 which accounts for
32% of B, 20% of P and 0.15% of F.

Dominant Pair:
192.88.112.0-160.91.192.0 for
32% of B, 20% of P and 0.15% of F.

Dominant Dest. Port:
5002 (iperf port, used by SLAC)

An example PF anomaly (DOS attack)



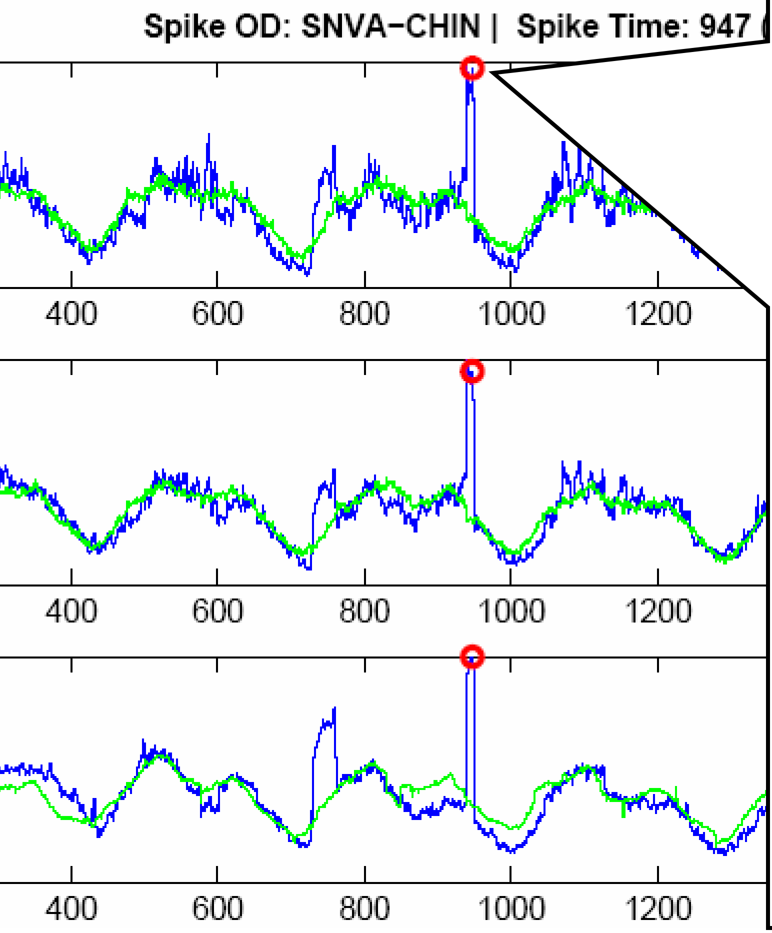
Dominant Source IP:
No dominant single source

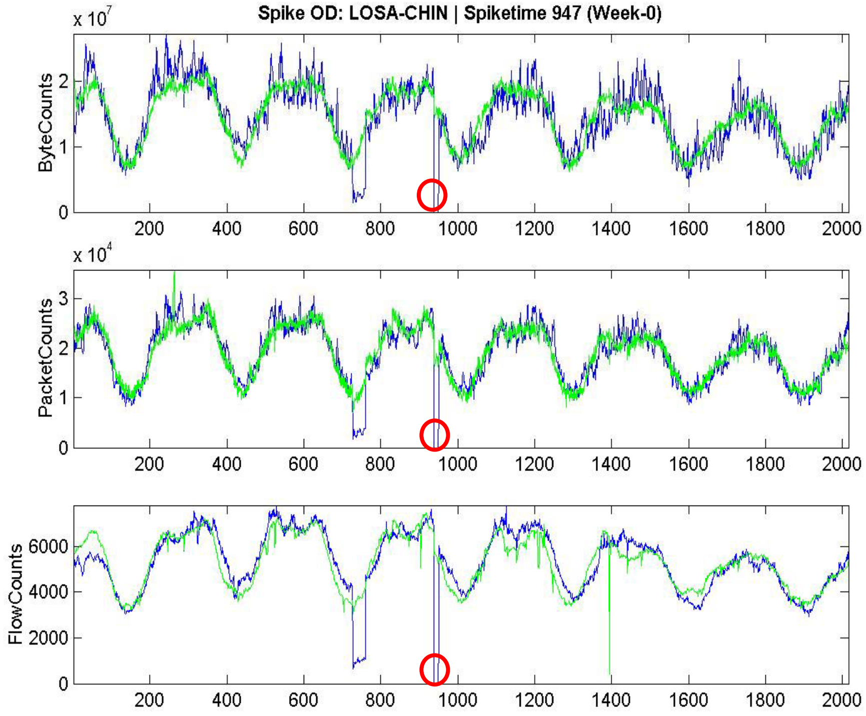
Dominant Dest. IP:
211.65.112.0 accounts for
80% of P traffic and
92% of F traffic.

Dominant Pair:
No single pair dominant

Dominant Ports:
No dominant source or
destination port found

An example BPF Anomaly (ingress-shift)



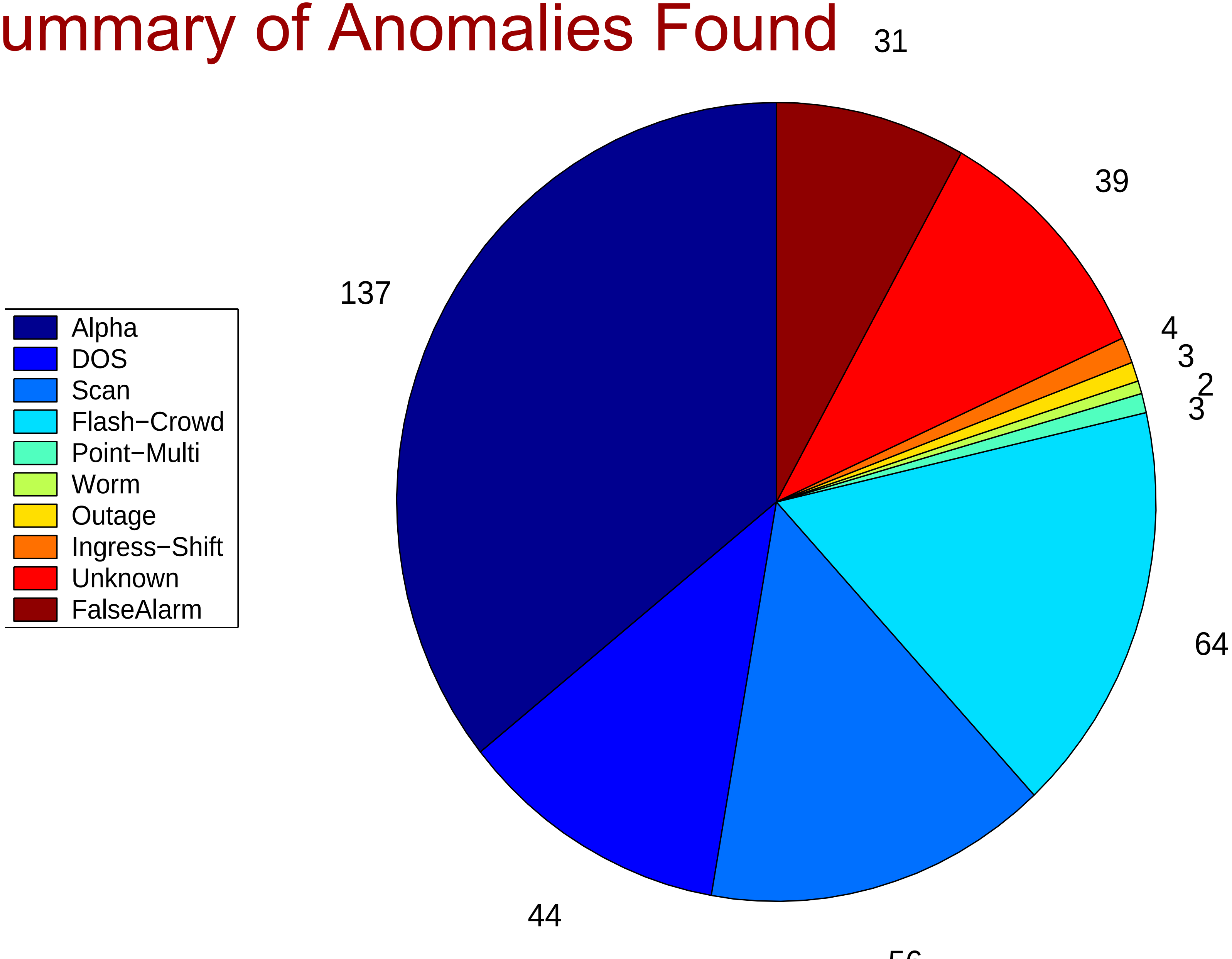


Multihomed customer CALREN reroutes around the LOSA-CHIN (scheduled) outage

Species of anomalies found

Anomaly	Definition
ALPHA	Unusually high rate point to point byte transfer
DOS, DDOS	(Distributed) Denial of service attack against a single victim
FLASH CROWD	Unusually large demand for a resource/service emerging from common set of sources
SCAN	Scanning a host for a vulnerable port (port scan) or scanning the network for a target port (network scan)
WORM	Self-propagating code that spreads across a network by exploiting security flaws
POINT to MULTIPOINT	Distribution of content from one server to many servers
OUTAGE	Equipment related events that decrease traffic exchanged by an OD pair
INGRESS-SHIFT	Customer shifts traffic from one ingress point to another

Summary of Anomalies Found



Conclusions

- Subspace method for anomaly diagnosis allows whole-network approach
  - Significant benefit accrues from whole-network analysis
- Diagnosing Volume Anomalies from Link Traffic:
  - High detection rate, low false alarm rate
  - Hypothesis-based identification is easily formalized and extended
- Detecting General Anomalies from Flow Traffic:
  - Anomalies detected span remarkable breadth
  - Almost all of the anomalies found are operationally relevant
- Whole-Network Anomaly Diagnosis with the Subspace Method is promising
  - ... more to come!

Thanks!

Help with Abilene Data

- Rick Summerhill, Mark Fullmer (Internet2)
- Matthew Davy (Indiana University)

Help with Sprint-Europe Data

- Bjorn Carlsson, Jeff Loughridge (SprintLink),
- Supratik Bhattacharyya, Richard Gass (ATL)