

Mining Anomalies in Network-Wide Flow Data

Anukool Lakhina, Ph.D.
with Mark Crovella and Christophe Diot



Network Anomaly Diagnosis

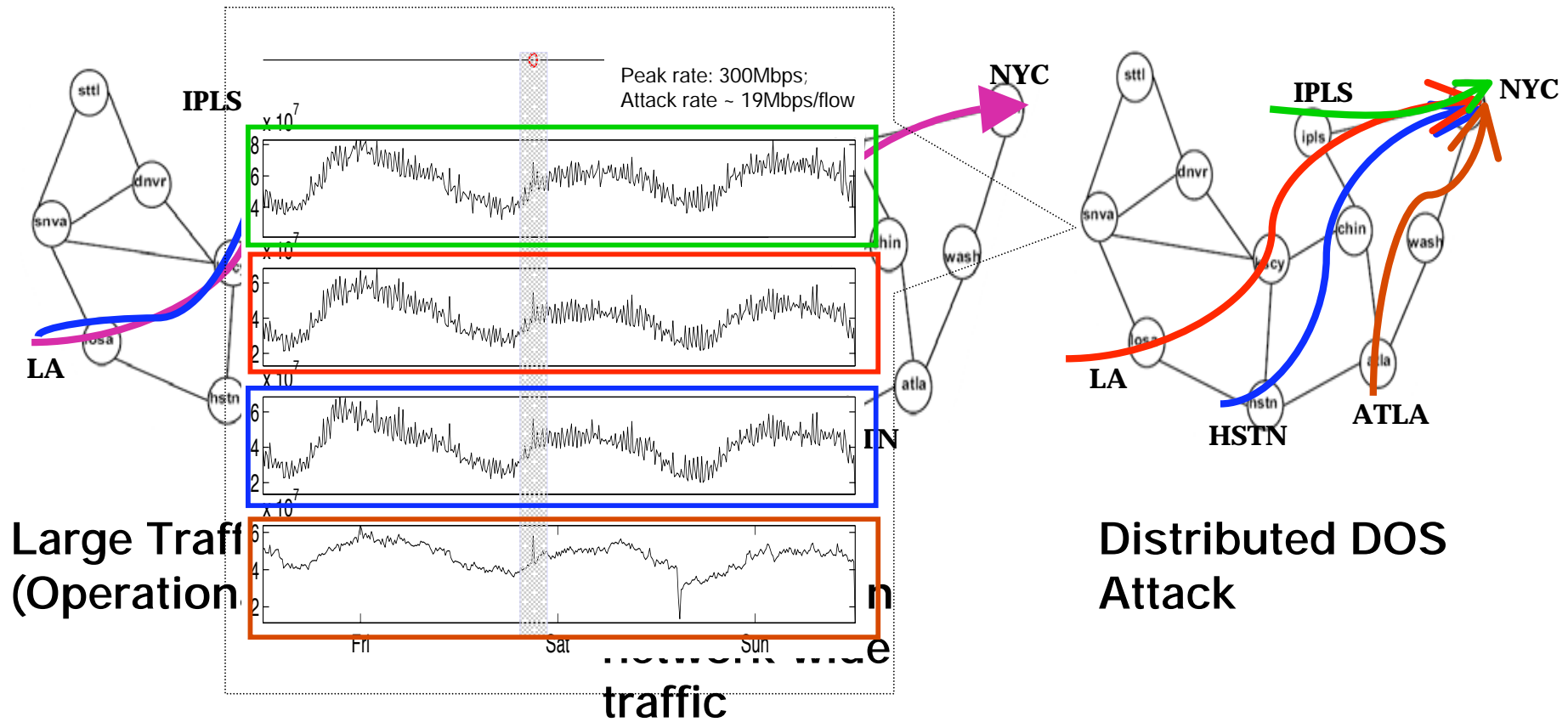
- Am I being attacked?
- Is someone scanning my network?
- Are there worms spreading?
- A sudden traffic shift?
- An equipment outage?
- Something never seen before?

A **general, unsupervised** method for reliably **detecting** and **classifying** network anomalies is needed

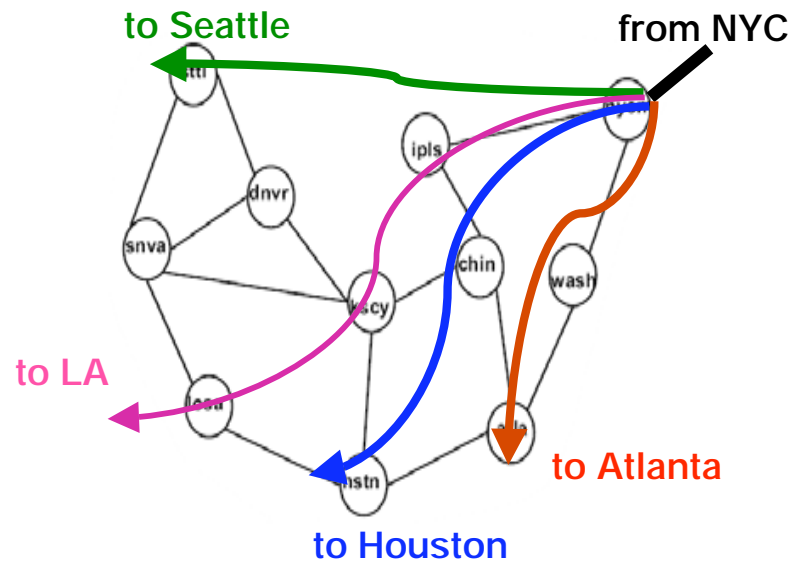
My Talk in One Slide

- *A general system to detect & classify anomalies at ISPs and large enterprises*
- **Central Message:** *Network-wide analysis of NetFlow data can expose many anomalies*
 - Detect both operational & malicious incidents
- I am here to seek **your feedback** 😊

Network-Wide Traffic Analysis



Collecting Network-Wide Traffic



- Assemble network's **traffic matrix**
- Traffic entering at the *origin* and leaving at the *destination*
- Use routing to aggregate NetFlow data into *OD flows*

Networks Evaluated

Abilene research network (Internet2)

- 11 PoPs, 121 OD flows, anonymized, 1/100 sampling, 5 min bins



Géant Europe research network (Dante)

- 22 PoPs, 484 OD flows, not anonymized, 1/1000 sampling, 10 min bins

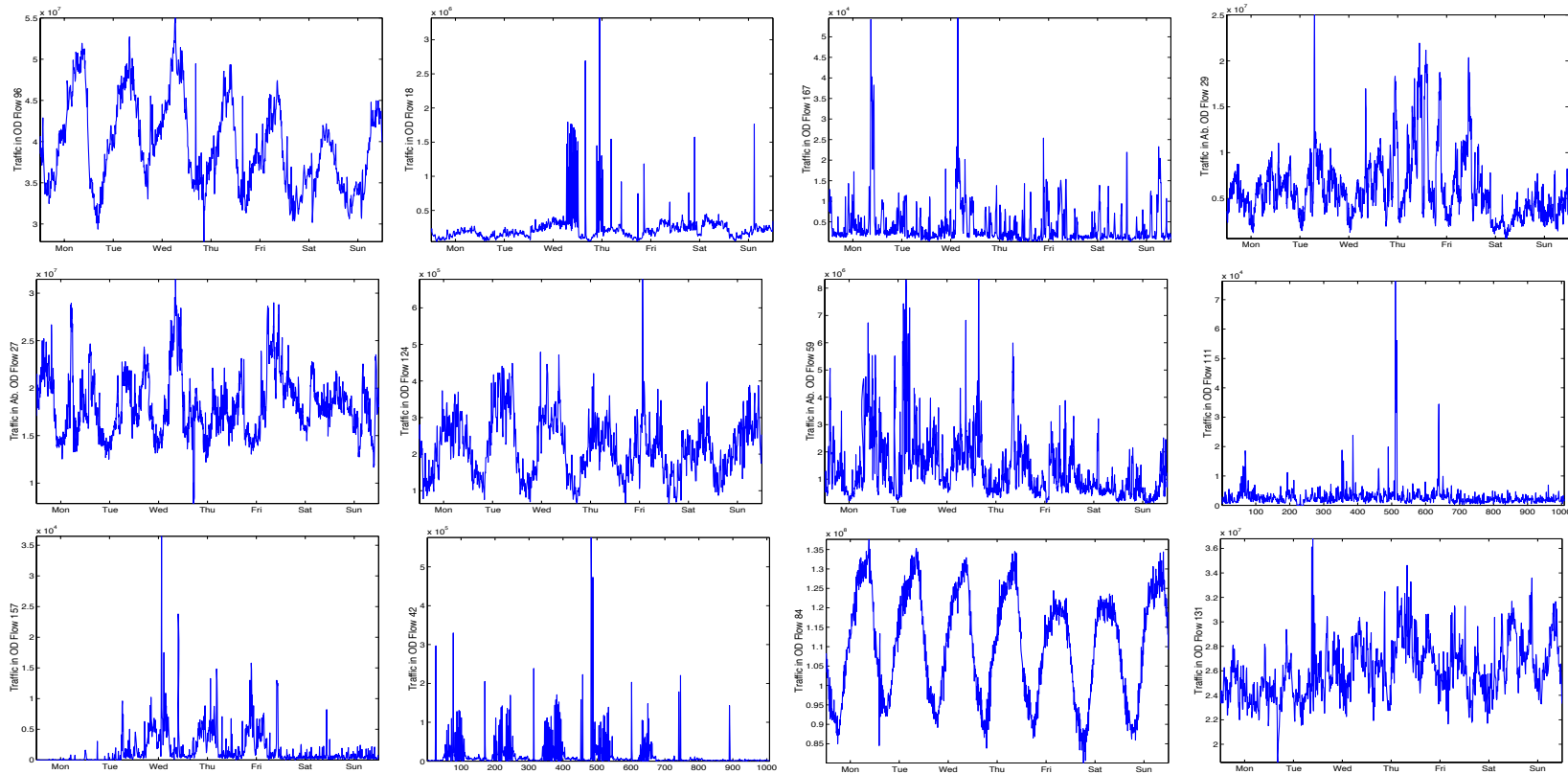


Sprint European commercial network

- 13 PoPs, 169 OD flows, not anonymized, aggregated, 1/250 sampling, 10 min bins



But, This is Difficult to Analyze!

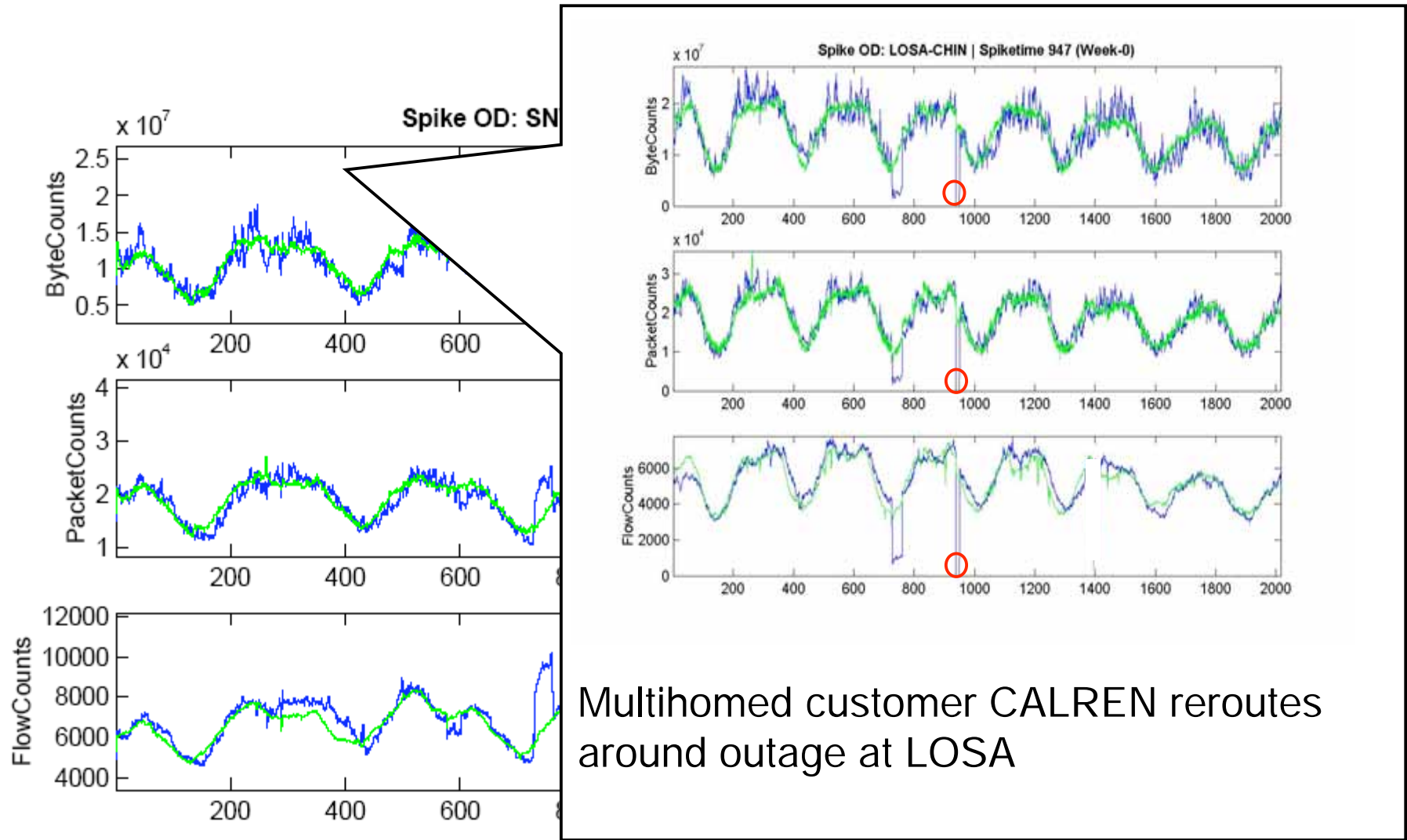


How do we extract **anomalies** and **normal behavior** from noisy, **high-dimensional** data?

The Subspace Method [LCD:SIGCOMM '04]

- An approach to separate normal & anomalous network-wide traffic
- Designate temporal patterns most common to all the traffic flows as the **normal patterns**
- Remaining temporal patterns form the **anomalous patterns**
- Detect anomalies by statistical thresholds on anomalous patterns

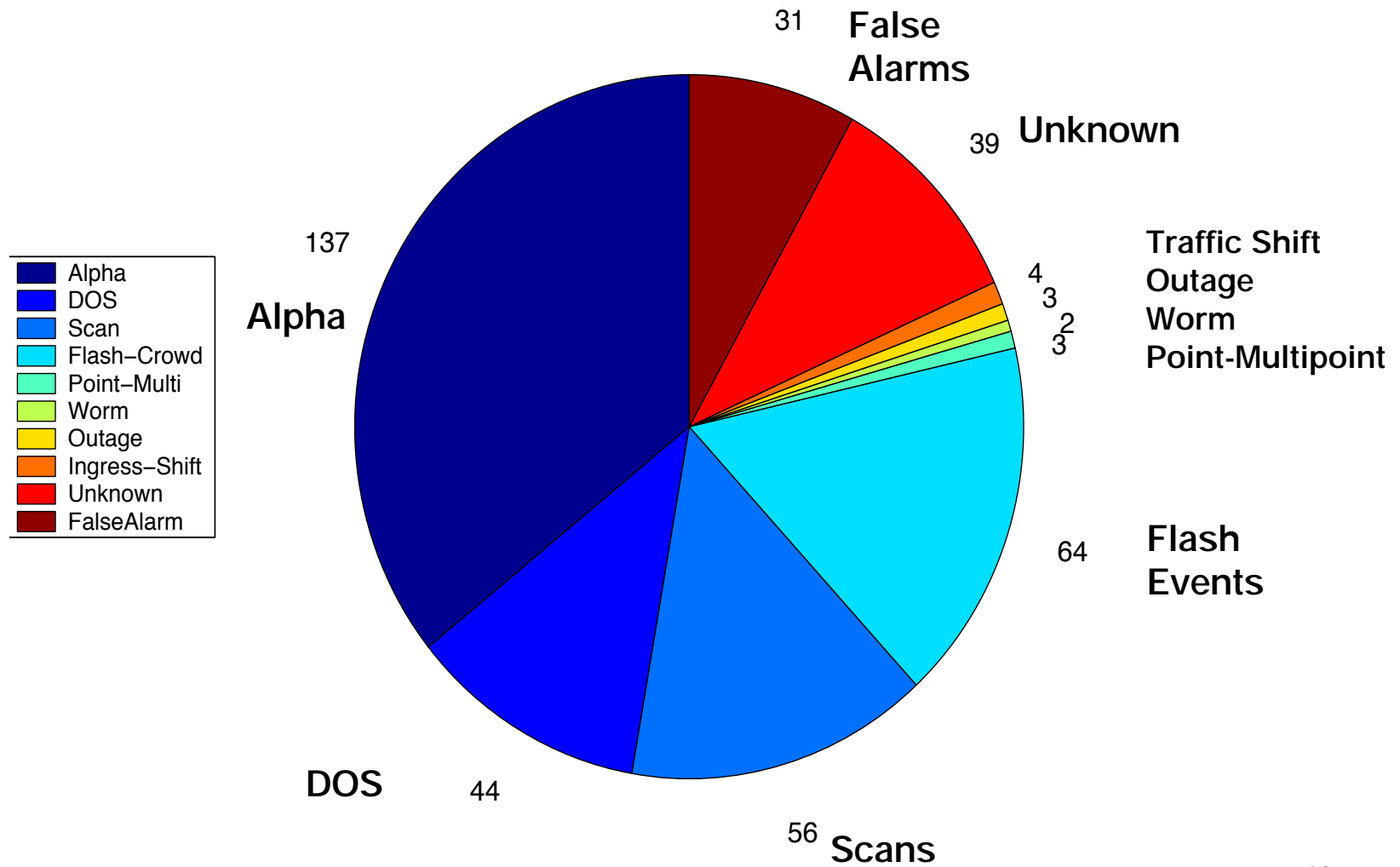
An example operational anomaly



Multihomed customer CALREN reroutes around outage at LOSA

Summary of Anomaly Types Found

[LCD:IMC04]

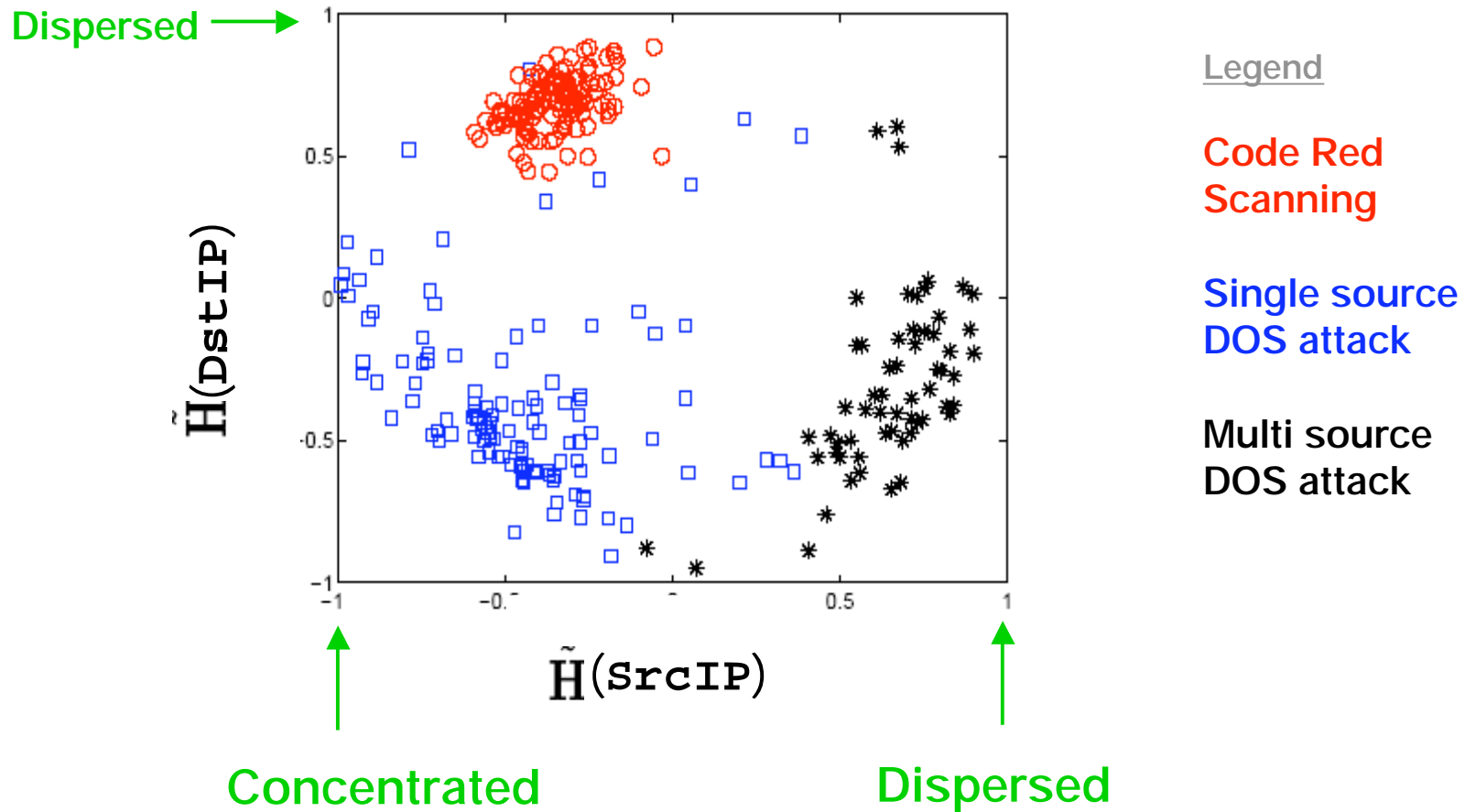


Automatically Classifying Anomalies

[LCD:SIGCOMM05]

- Goal: Classify anomalies without restricting yourself to a predefined set of anomalies
- Approach: Leverage 4-tuple header fields:
SrcIP, SrcPort, DstIP, DstPort
 - In particular, measure *dispersion* in fields
- Then, apply off-the-shelf clustering methods

Example of Anomaly Clusters



Summary

- **Network-Wide Detection:**
 - Broad range of anomalies with low false alarms
 - In papers: Highly sensitive detection, even when anomaly is 1% of background traffic
- **Anomaly Classification:**
 - Feature clusters automatically classify anomalies
 - In papers: clusters expose new anomalies
- Network-wide data and header analysis are promising for general anomaly diagnosis

Next steps

- **Ongoing Work:** implementing algorithms in a prototype system
- For more information, see papers & slides at:
<http://cs-people.bu.edu/anukool/pubs.html>
- Your feedback much needed & appreciated!
 - Data, deployment, ...