

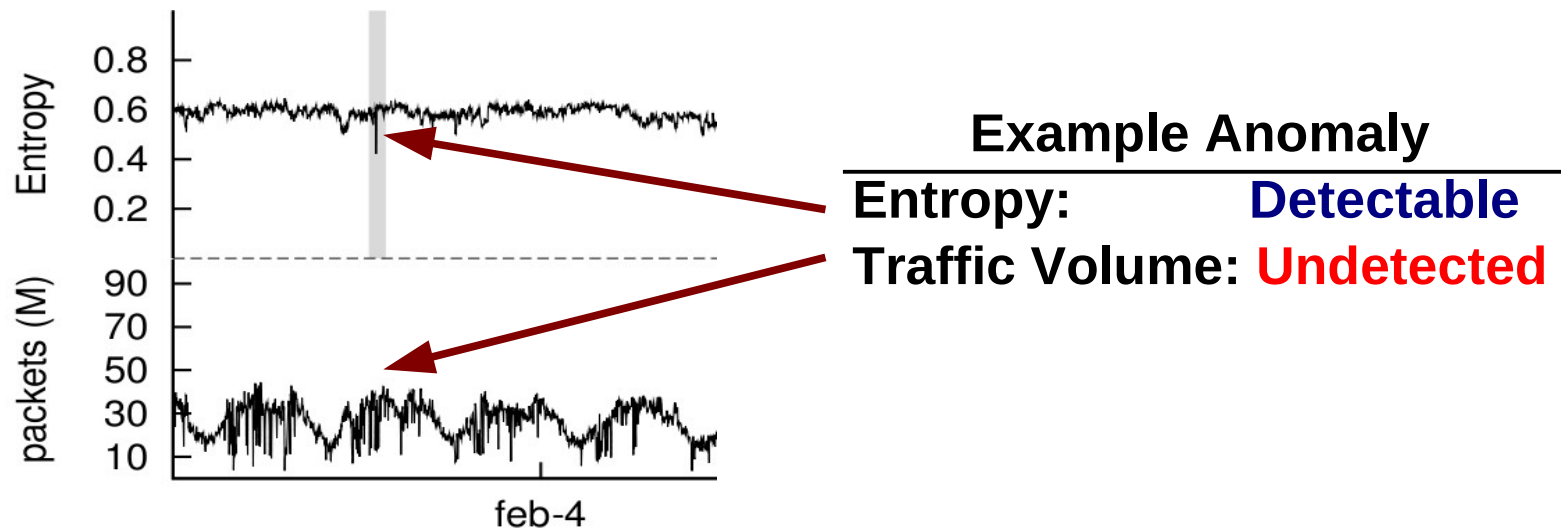
# **An Empirical Evaluation of Entropy-based Traffic Anomaly Detection**

*George Nychis, Vyas Sekar, David Andersen,  
Hyong Kim, Hui Zhang*

*Carnegie Mellon University*

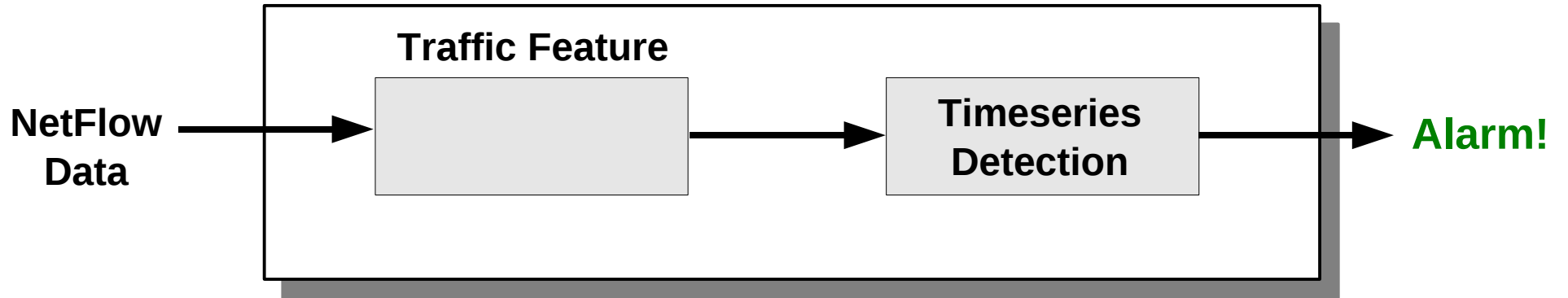
# Entropy-based Anomaly Detection

- **Goal:** detect abnormal behavior
  - scan activity, DDoS, bandwidth floods ...
- **Traditional:** raw traffic volume (*insufficient*)
  - *e.g., total number of packets in an epoch*
- **Modern:** entropy-based traffic metrics
  - *e.g., relative randomness in distribution of packets across ports*



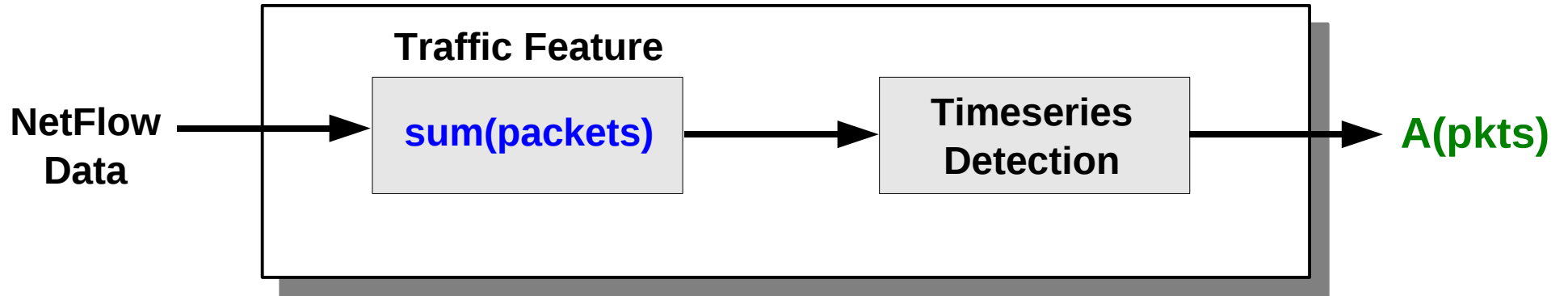
# Motivation

## Anomaly Detection



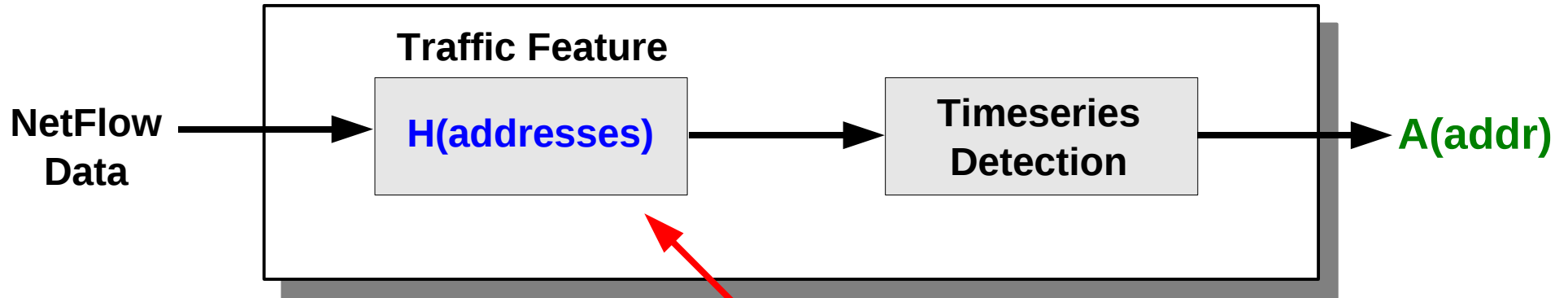
# Motivation

## Anomaly Detection



# Motivation

## Anomaly Detection

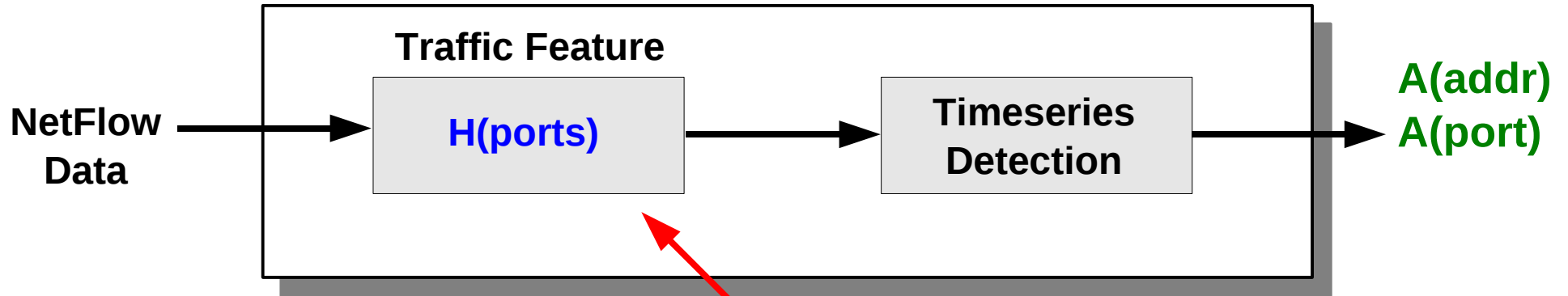


**Entropy-based Features:**

**Dist. of packets across addresses**

# Motivation

## Anomaly Detection



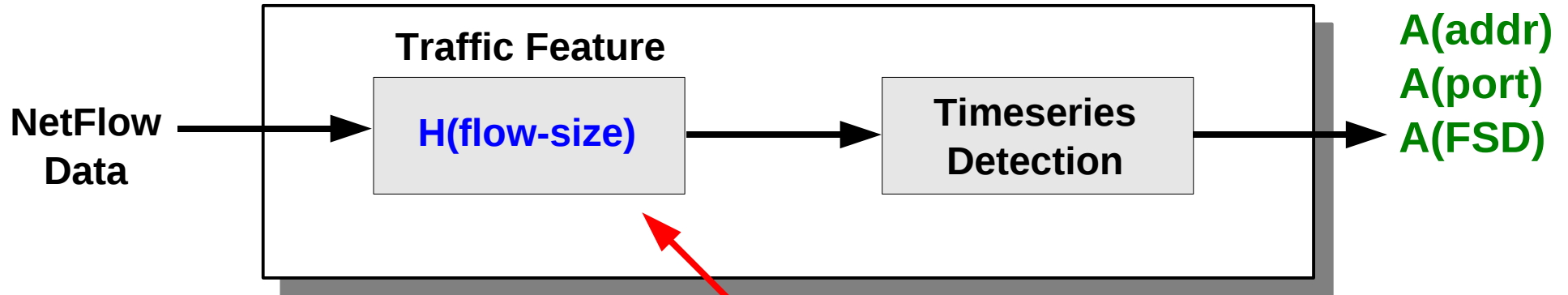
**Entropy-based Features:**

**H(addresses)**

**Distribution of packets across ports**

# Motivation

## Anomaly Detection

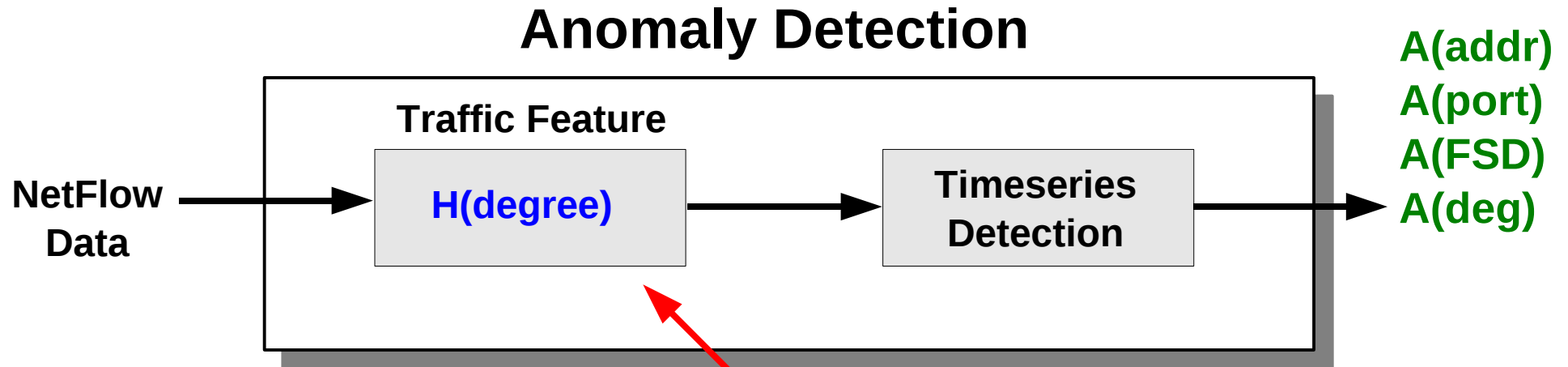


**Entropy-based Features:**

**H(addresses) H(ports)**

**Distribution of flow-sizes (in packets)**

# Motivation

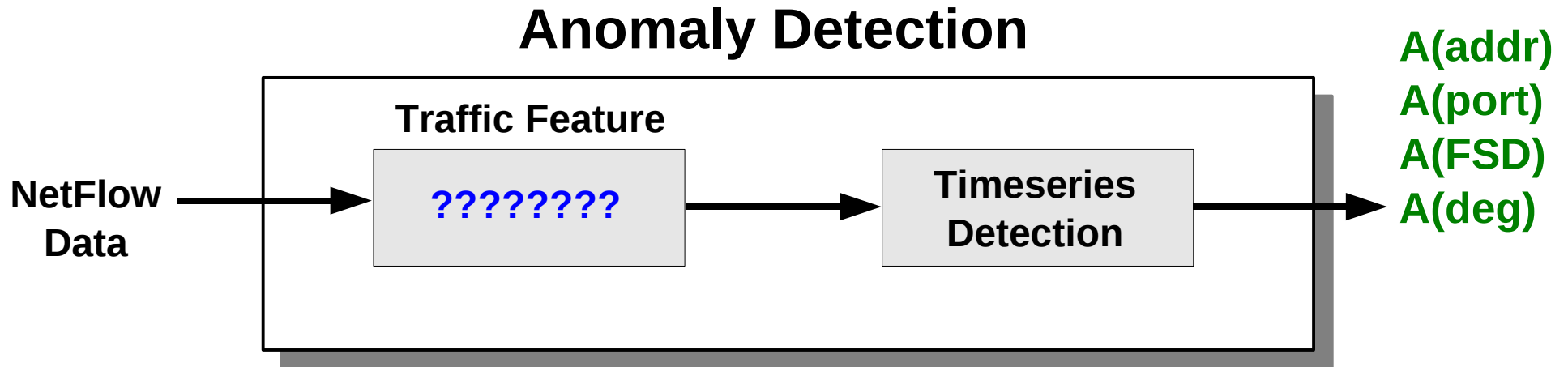


**Entropy-based Features:**

$H(\text{addresses})$     $H(\text{ports})$     $H(\text{flow-size})$

**Distribution of host communication**

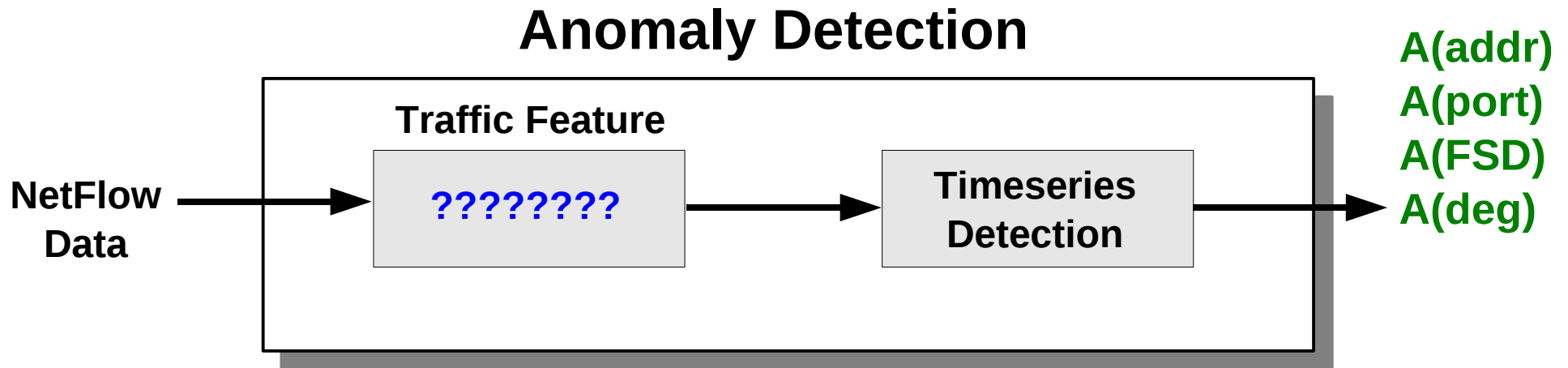
# Motivation



## Entropy-based Features:

**H(addresses) H(ports) H(flow-size) H(degree)**

# Motivation



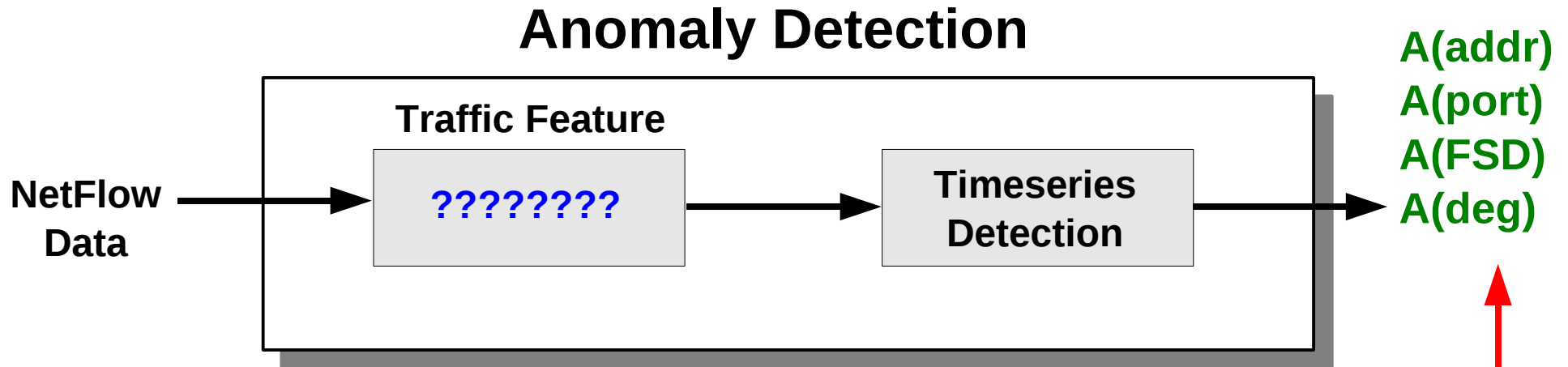
## Entropy-based Features:

$H(\text{addresses})$     $H(\text{ports})$     $H(\text{flow-size})$     $H(\text{degree})$



- **Goal:** understanding the **features**

# Motivation



## Entropy-based Features:

$H(\text{addresses})$     $H(\text{ports})$     $H(\text{flow-size})$     $H(\text{degree})$



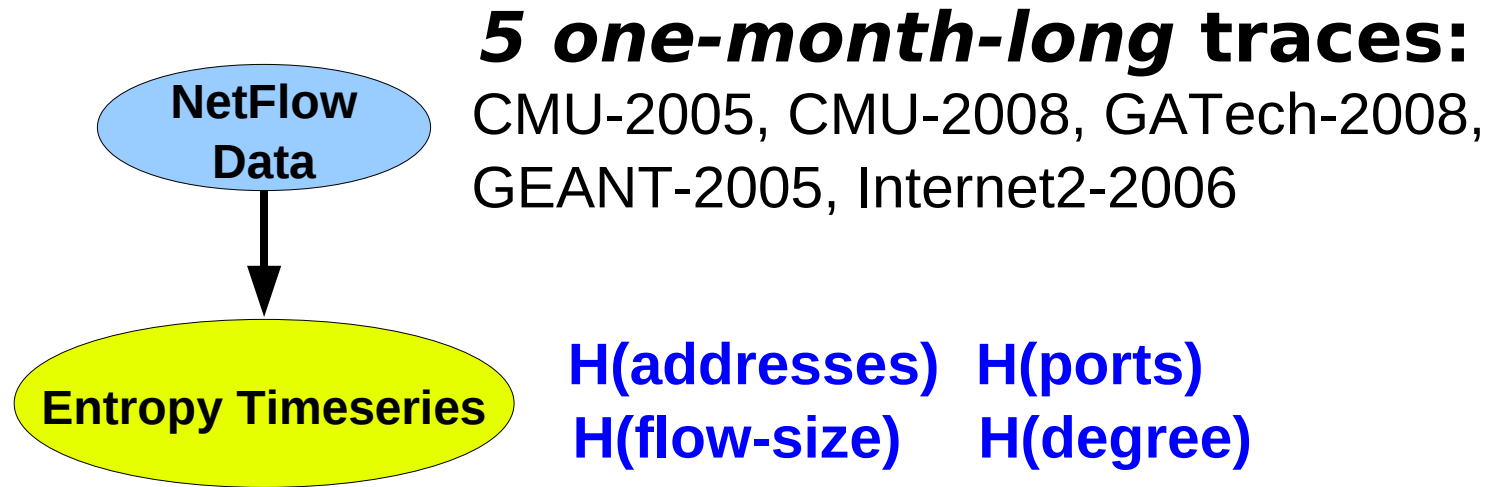
- **Goal:** understanding the **features**
  1. How **unique** are their detection capabilities? \_\_\_\_\_
  2. How **effective** are they? \_\_\_\_\_

# Analysis Method

NetFlow  
Data

***5 one-month-long traces:***  
CMU-2005, CMU-2008, GATech-2008,  
GEANT-2005, Internet2-2006

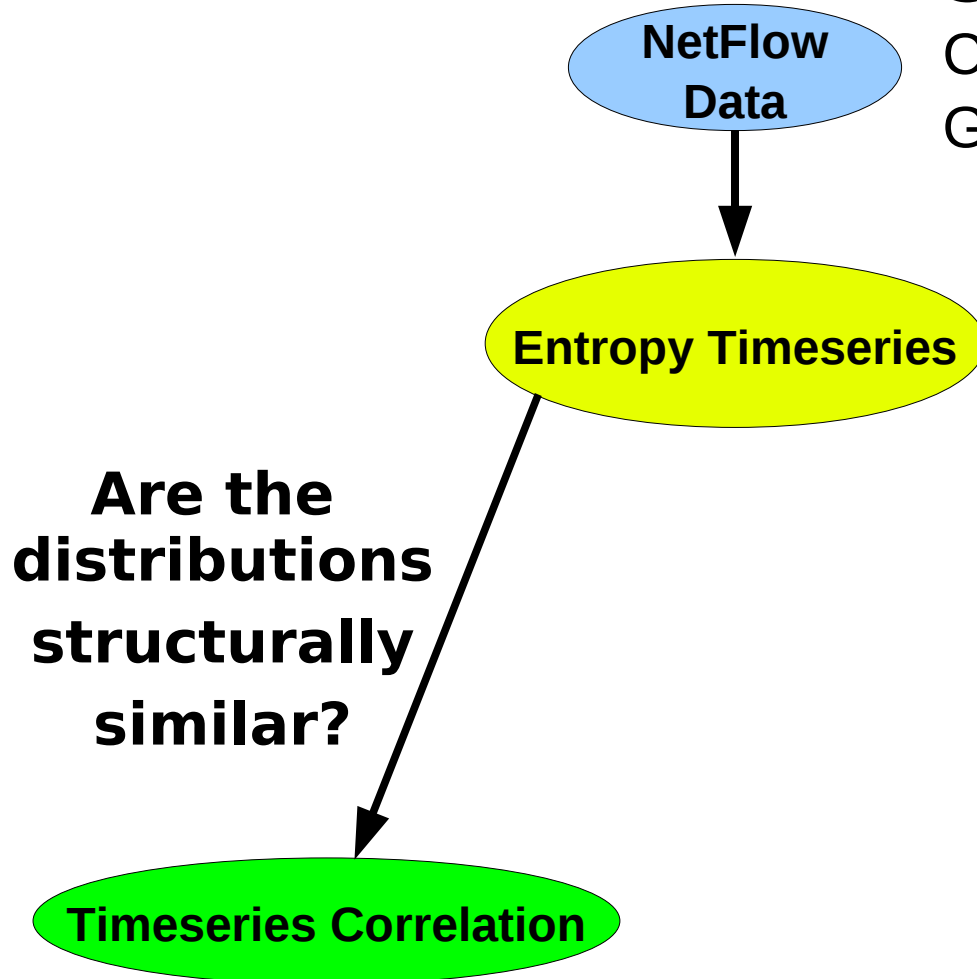
# Analysis Method



# Analysis Method

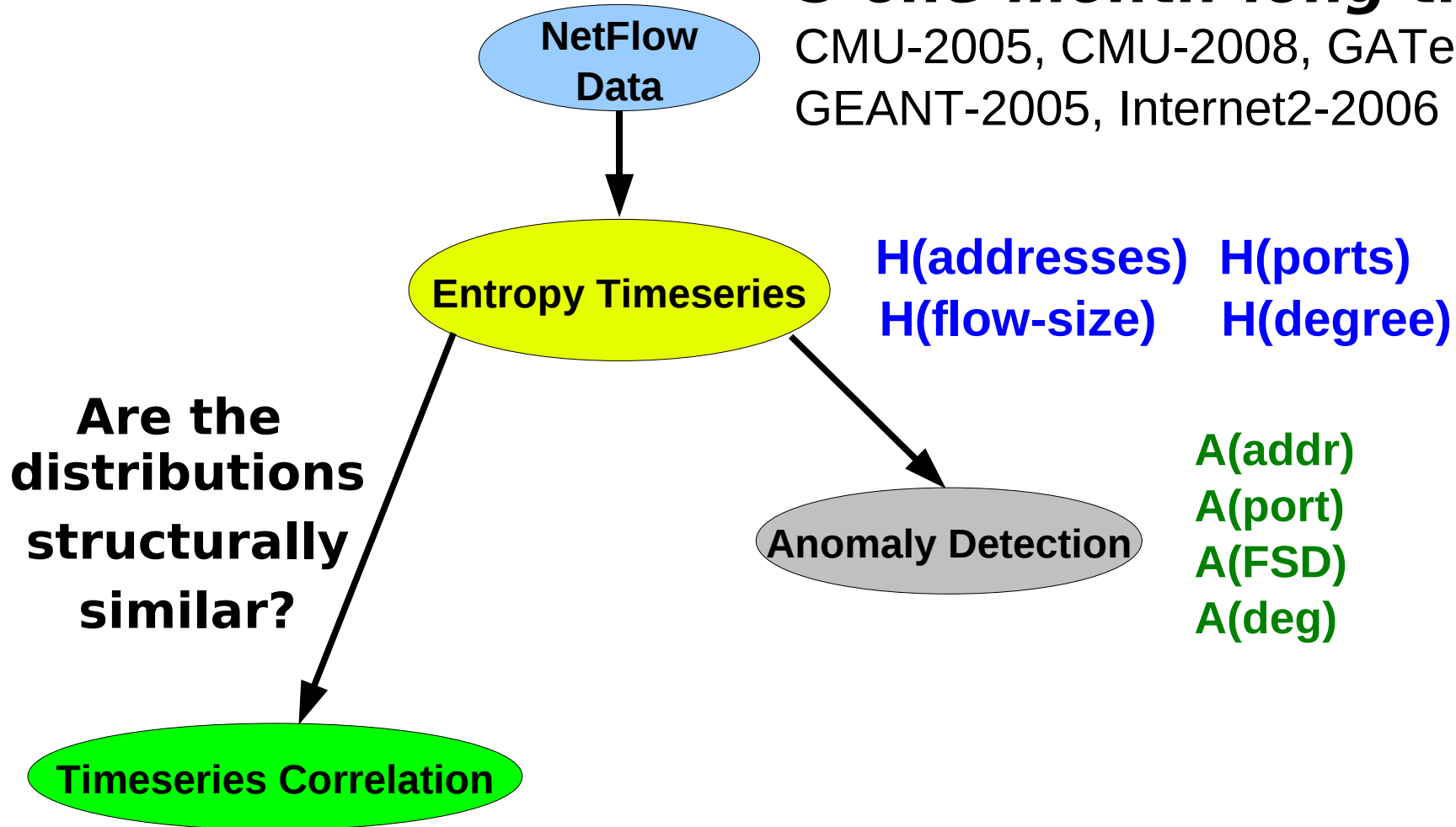
**5 one-month-long traces:**  
CMU-2005, CMU-2008, GATech-2008,  
GEANT-2005, Internet2-2006

**H(addresses) H(ports)**  
**H(flow-size) H(degree)**



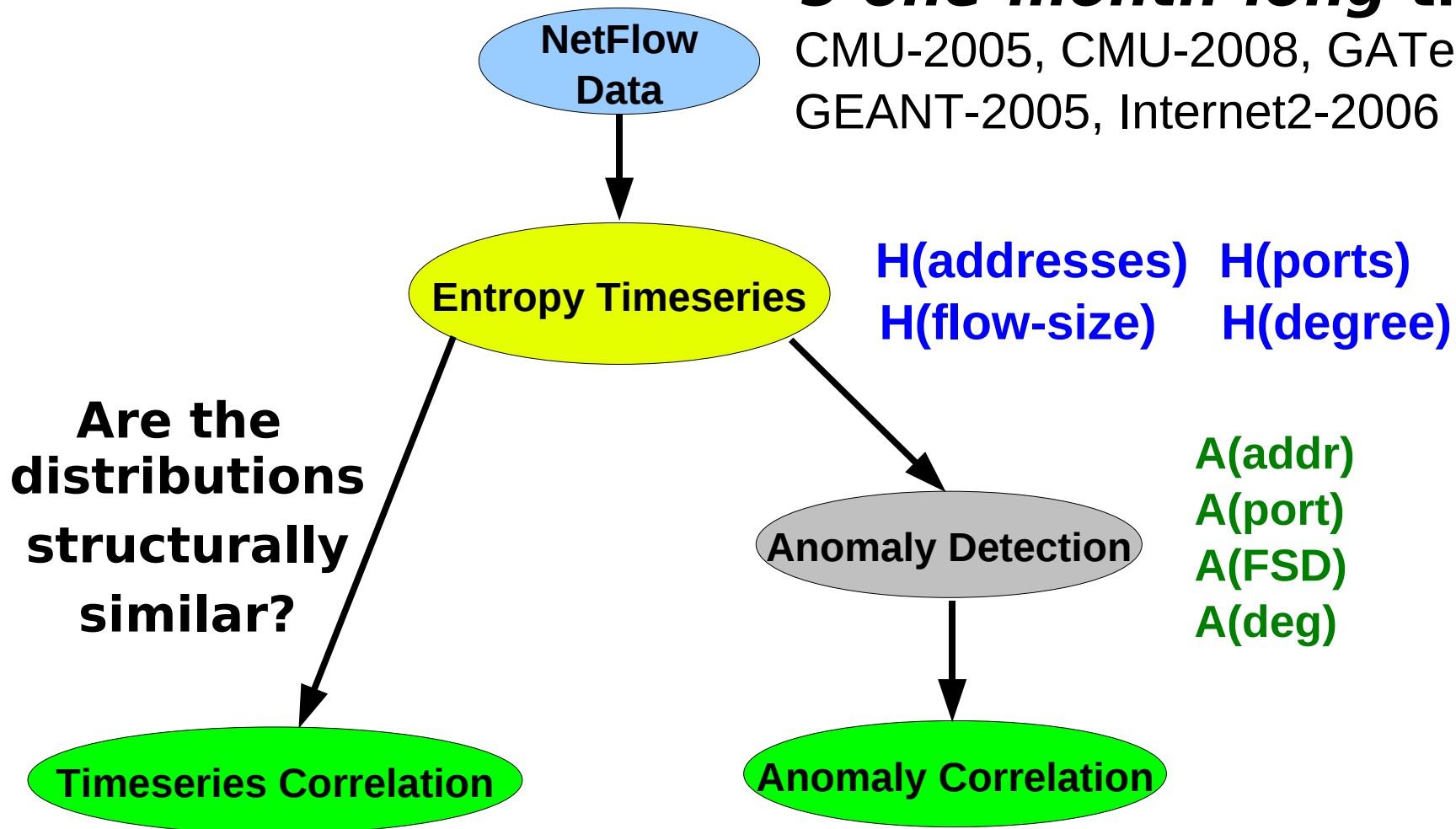
# Analysis Method

**5 one-month-long traces:**  
 CMU-2005, CMU-2008, GATech-2008,  
 GEANT-2005, Internet2-2006



# Analysis Method

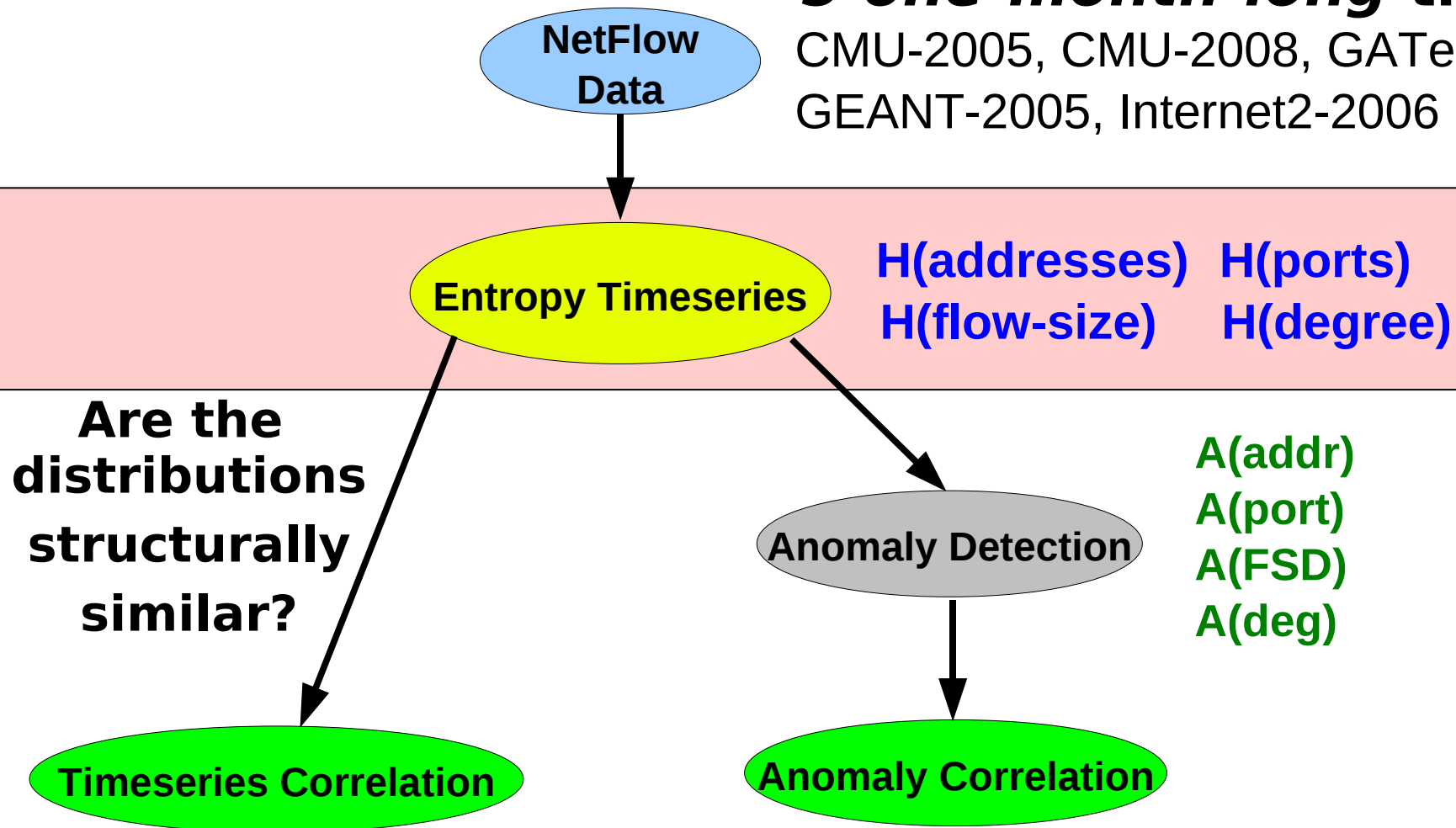
**5 one-month-long traces:**  
 CMU-2005, CMU-2008, GATech-2008,  
 GEANT-2005, Internet2-2006



**Goal(1): Uniqueness**

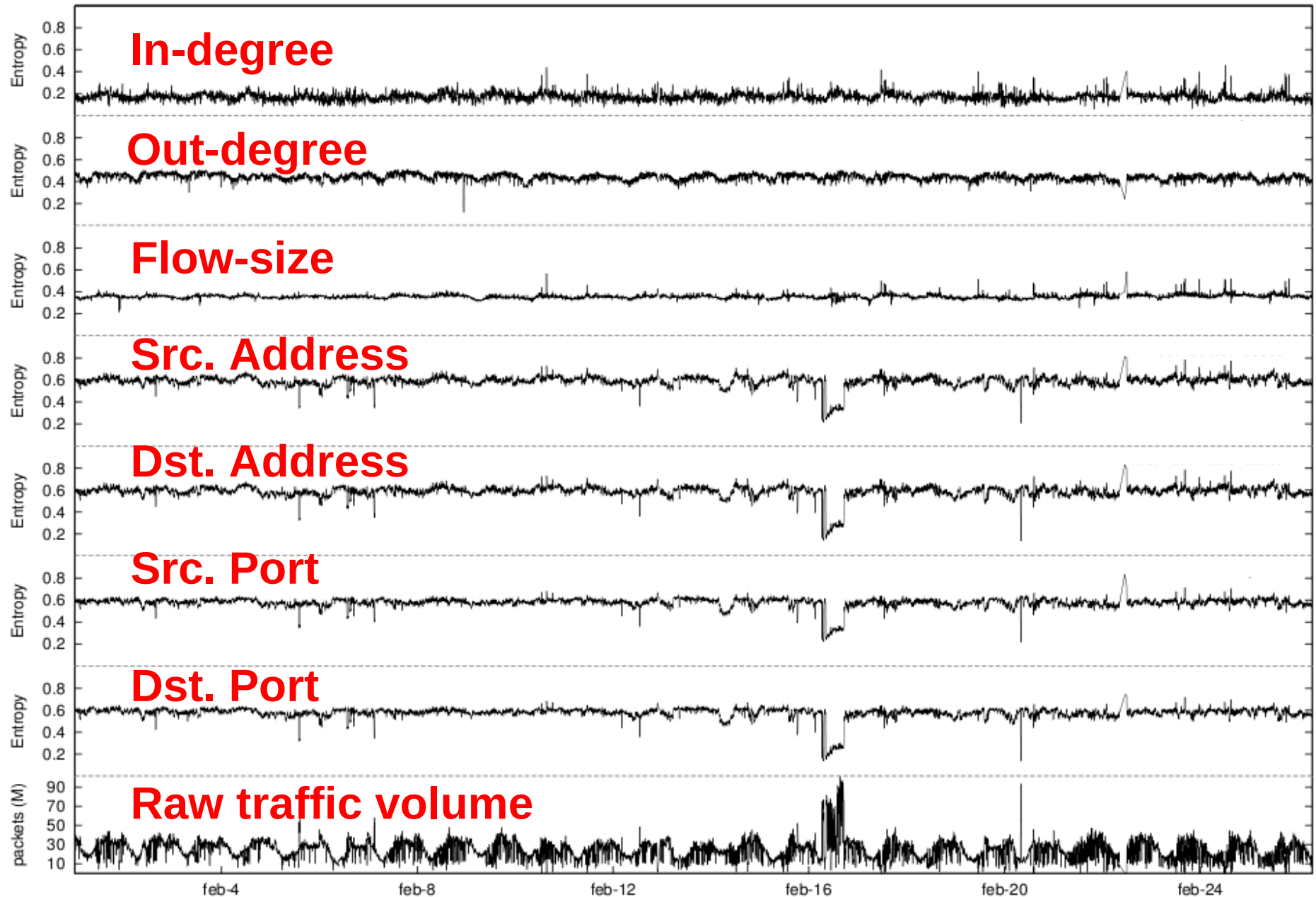
# Analysis Method

**5 one-month-long traces:**  
 CMU-2005, CMU-2008, GATech-2008,  
 GEANT-2005, Internet2-2006

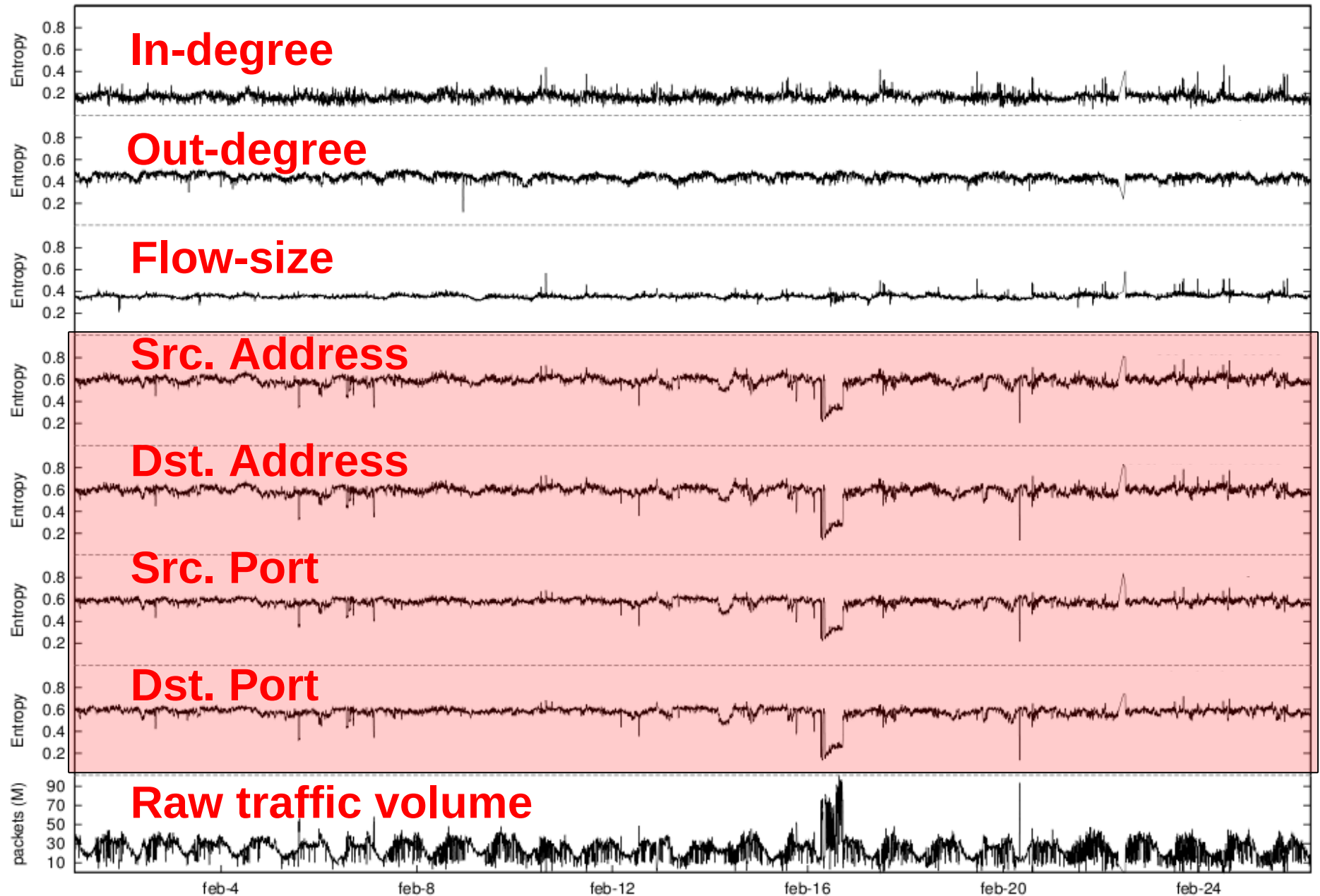


**Goal(1): Uniqueness**

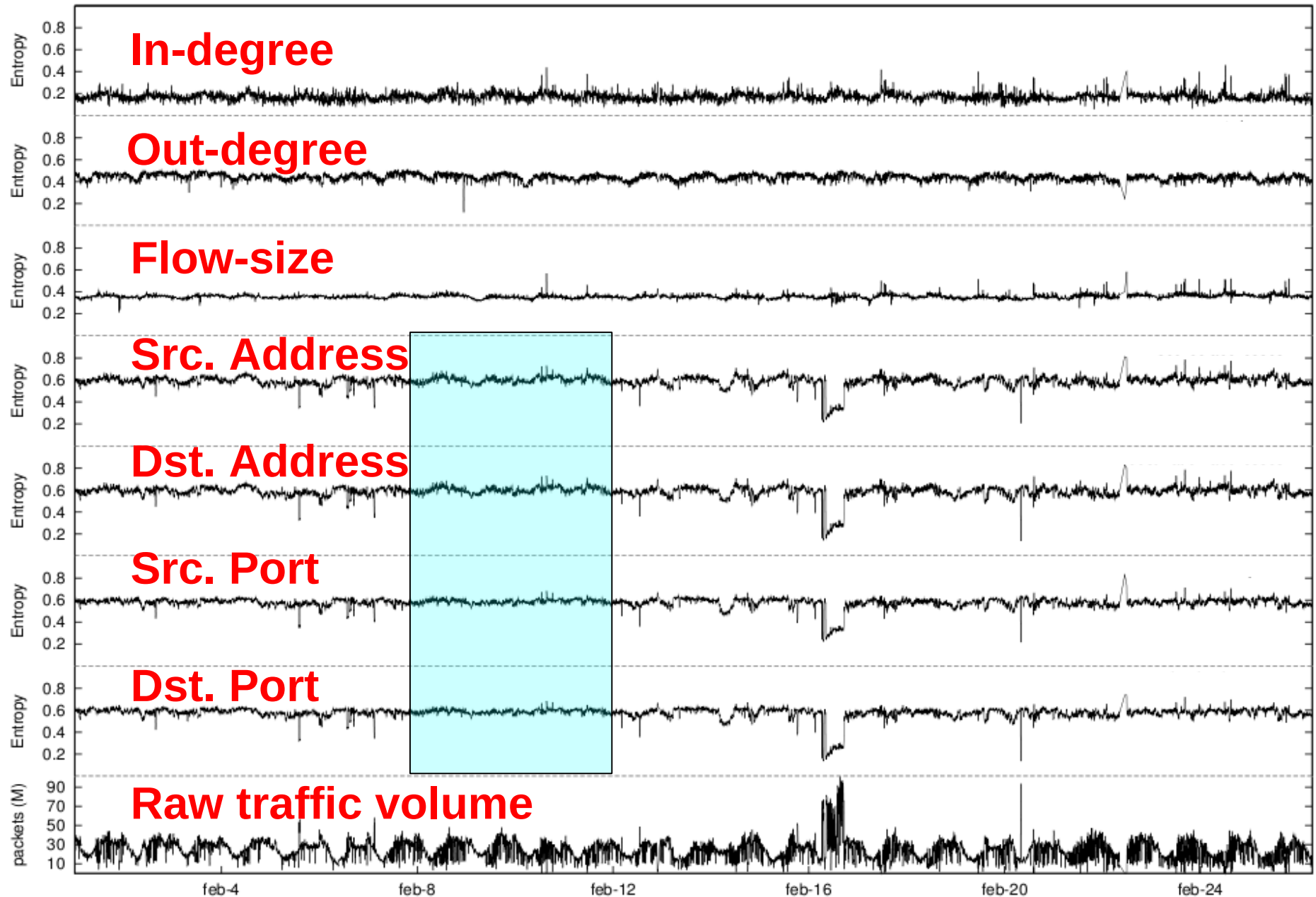
# Entropy Timeseries (February 2005)



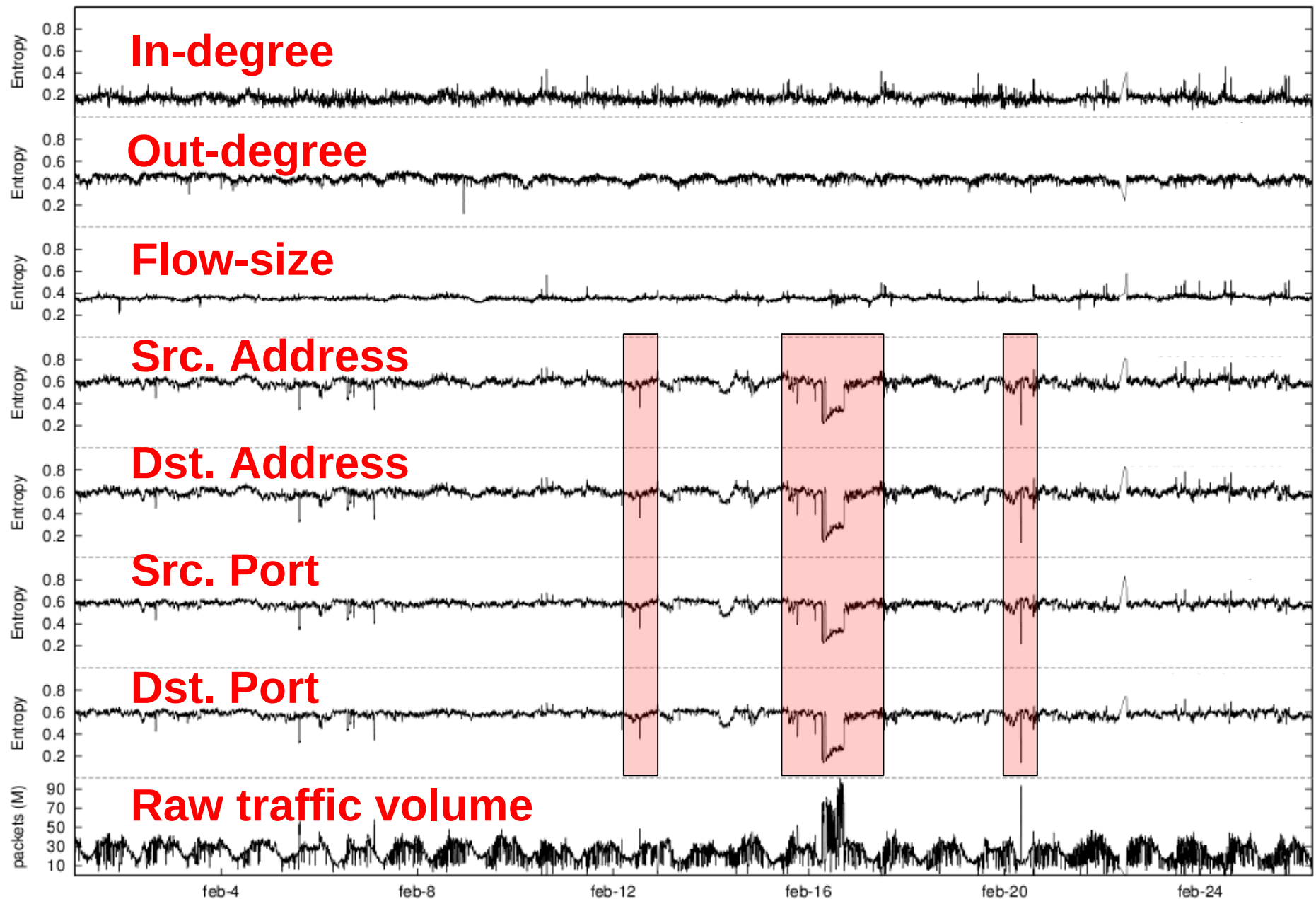
# Entropy Timeseries (February 2005)



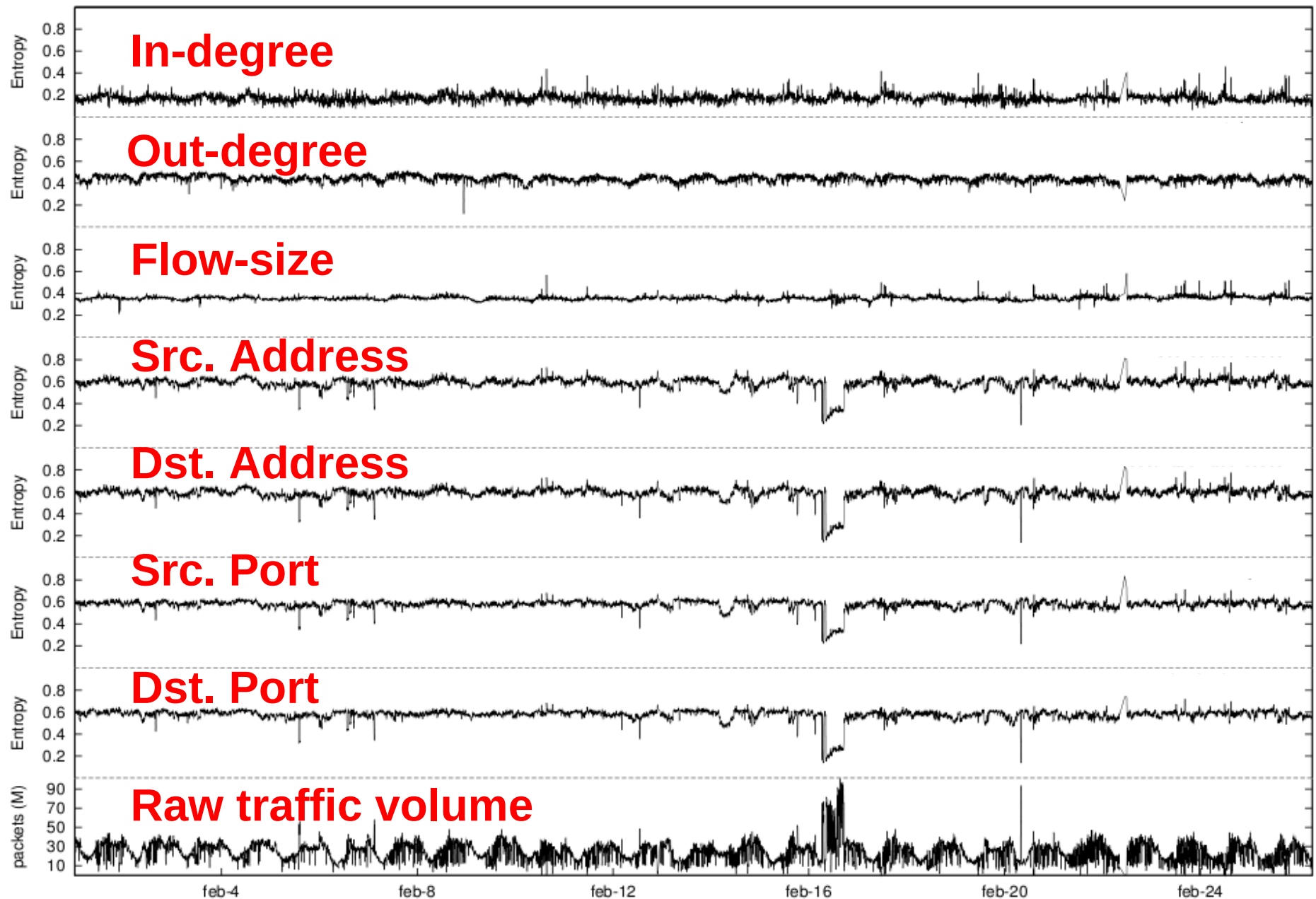
# Entropy Timeseries (February 2005)



# Entropy Timeseries (February 2005)



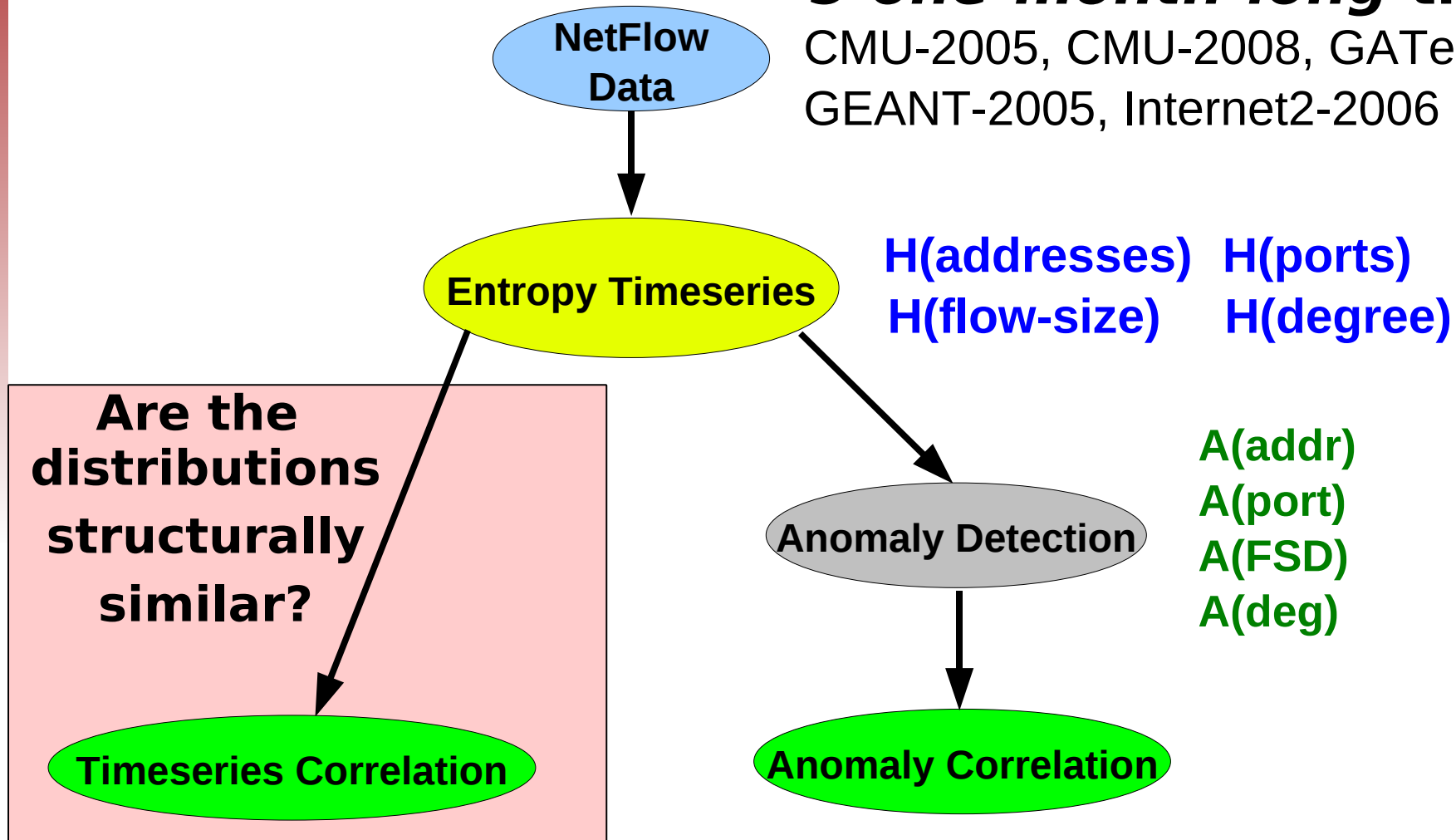
# Entropy Timeseries (February 2005)



# Analysis Method

## 5 one-month-long traces:

CMU-2005, CMU-2008, GATech-2008,  
GEANT-2005, Internet2-2006



**Goal(1): Uniqueness**

# Correlation in Entropy Timeseries

## ■ Pairwise correlation-scores for CMU-2005

	Out Deg	Src Addr	Dst Addr	Src Port	Dst Port	FSD
InDeg	0.10	0.10	0.09	0.00	0.00	0.41
OutDeg	-	-0.03	-0.03	-0.05	-0.01	-0.01
SrcAddr	-	-	0.99	0.96	0.95	0.30
DstAddr	-	-	-	0.96	0.96	0.28
SrcPort	-	-	-	-	0.98	0.17
DstPort	-	-	-	-	-	0.18

## ■ *All 4 other traces exhibit similar behavior!*

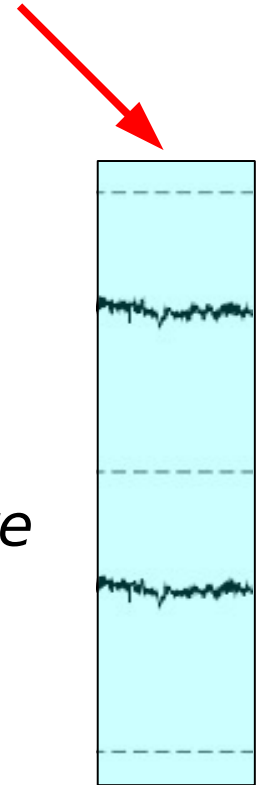
# Why Entropy is Structurally Correlated

## 1. Port / Address Correlation

- Properties of Network Traffic:

- contribute  $X$  packets to address  $A$
- contribute  $X$  packets to port  $B$

... *if hosts have few connections, and ports are uniformly random* → similar distributions



# Why Entropy is Structurally Correlated

## 1. Port / Address Correlation

- Properties of Network Traffic

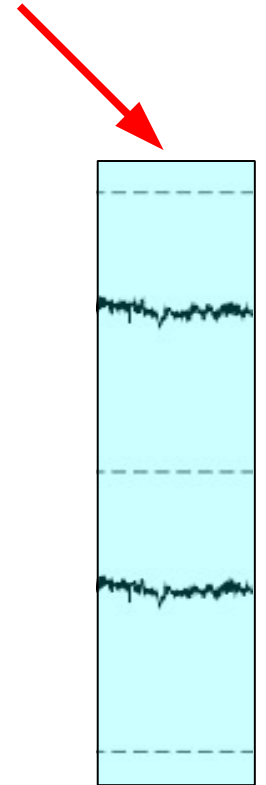
## 2. Source / Destination Correlation

- Flow accounting:
  - Bi-directional: Addr1(23) → Addr2(53)

### Bi-directional

Saddr(23)

Daddr(53)



# Why Entropy is Structurally Correlated

## 1. Port / Address Correlation

- Properties of Network Traffic

## 2. Source / Destination Correlation

- Flow accounting:

- Uni-directional: Addr1 → Addr2 (23)  
Addr2 → Addr1 (53)

### Bi-directional

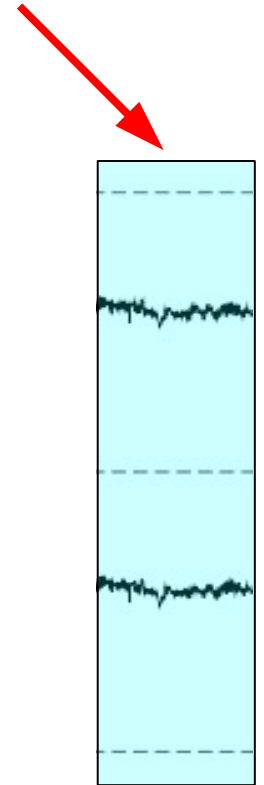
Saddr(23)

Daddr(53)

### Uni-directional

Saddr(23), Daddr(23)

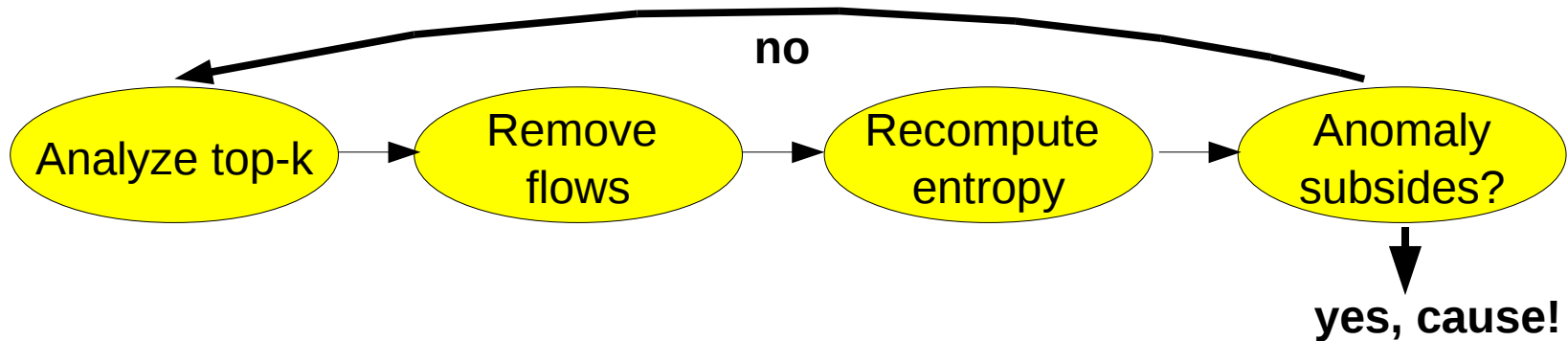
Saddr(53), Daddr(53)



**Uni-directionality destroys 2 unique distributions**

# Why Anomalies are Correlated

- **Root-cause analysis approach:**

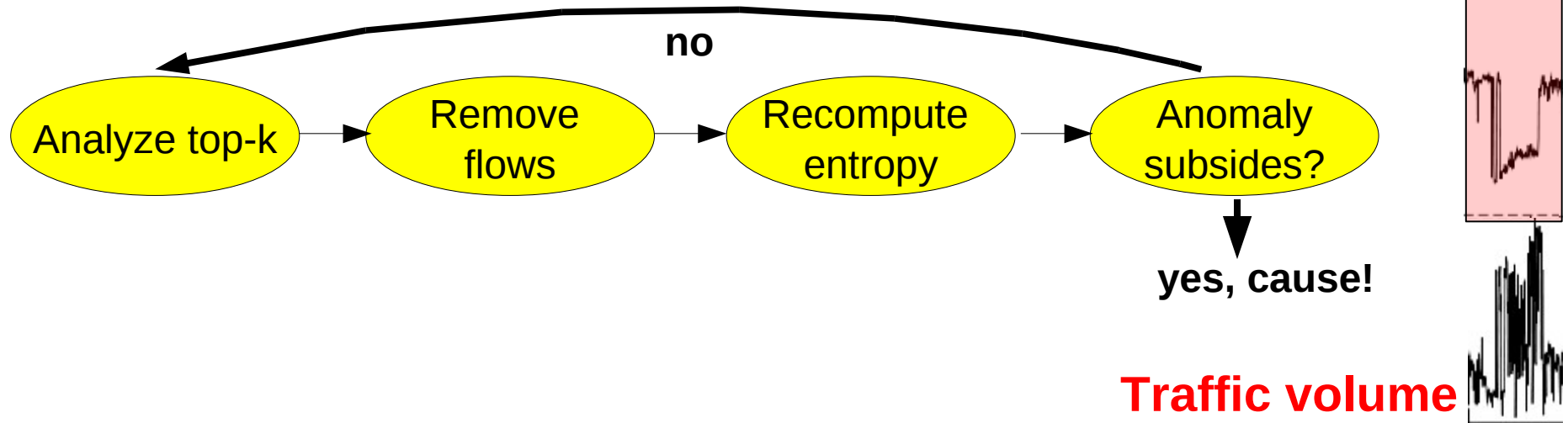


- **Our results:**

- **Ports & addresses:** *only detect alpha flows (correlation)*
- **FSD:** *detects scans*, **Degree:** SYN flood
- FSD & Degree are unique (*no correlation*)

# Why Anomalies are Correlated

- **Root-cause analysis approach:**



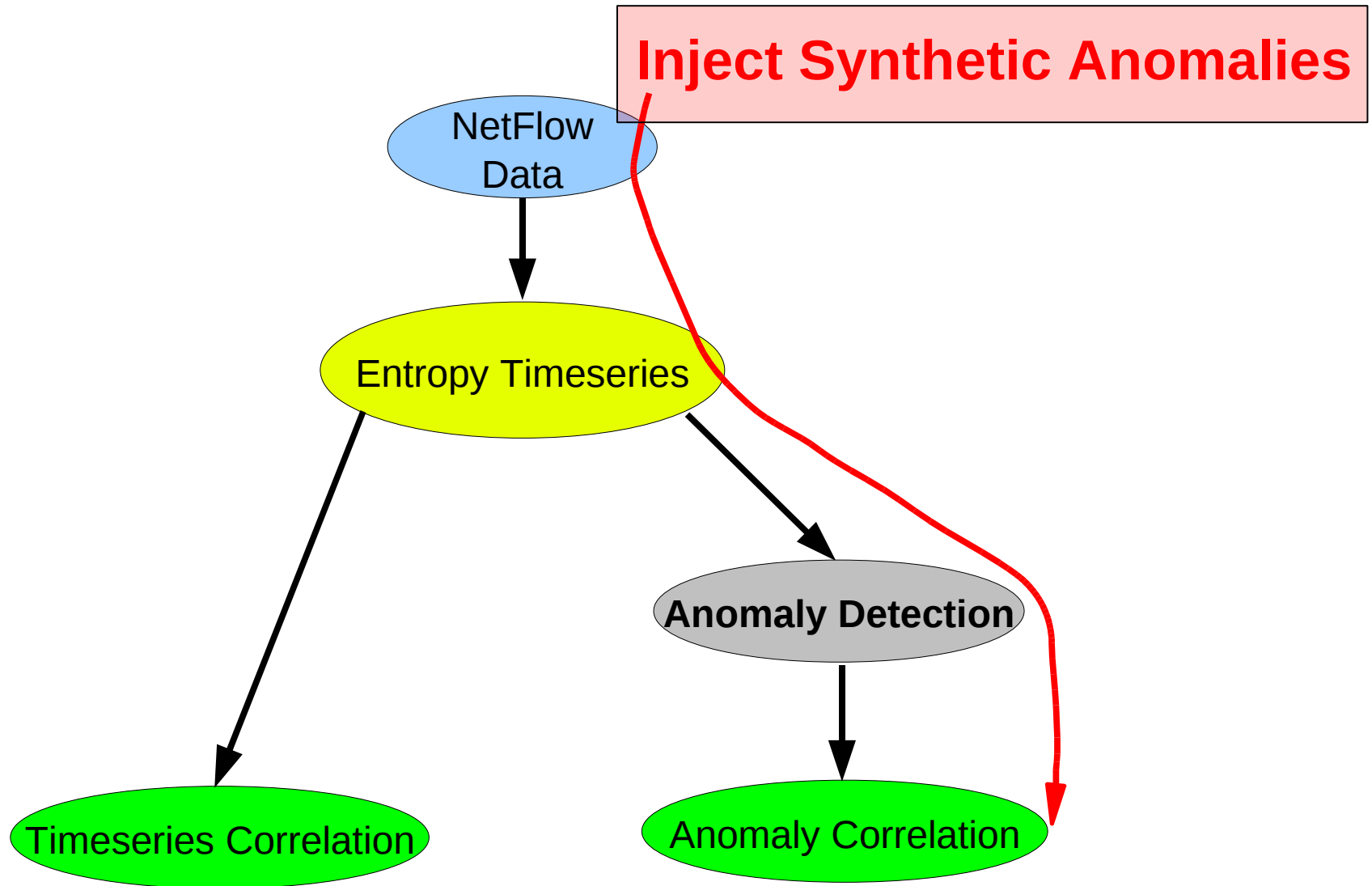
- **Our results:**

- **Ports & addresses:** *only detect alpha flows (correlation)*
- **FSD:** *detects scans*, **Degree:** SYN flood
- FSD & Degree are unique (*no correlation*)

# Summary of Goal(1): Uniqueness

- **Strong correlation** in ports and addresses
- Flow-size and degree: **unique**
- **Structural correlation**: properties of traffic
- **Anomaly correlation**: types of anomalies seen

# Understanding Effectiveness



# Best Distribution for an Anomaly?

- **Anomalies:** BW Flood, Scanner, Multiple Scanners, Port Scan, and SYN Flood

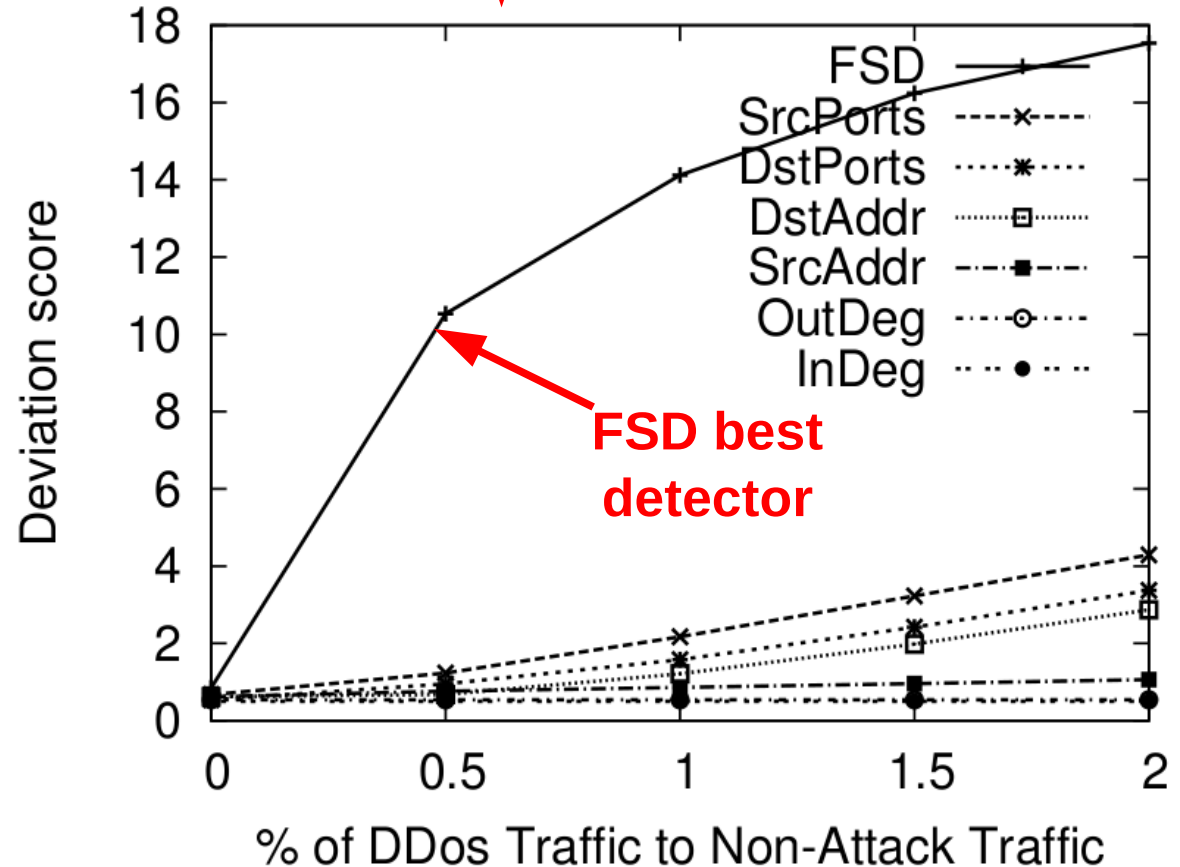
- **Other Results:**

- **BW Flood:**

- *ports & addresses*
    - already detectable by traffic volume

- **Scans:**

- difficult to detect
    - ... *FSD and degree*



# Implications and Conclusions

- **Look beyond ports and addresses**
- Select **complementary** traffic distributions
- **Uni-directional accounting** introduces biases in traffic distributions
- **Future Work:** Can correlations be leveraged?
  - during anomalies found in flow-size & degree, *correlation drops* between ports & addresses

**Questions?**