Discrete Mathematical Approaches to Traffic Graph Analysis

CLIFF JOSLYN
WENDY COWLEY, EMILIE HOGAN, BRYAN OLSEN

FLOCON 2015

JANUARY 2015

Outline

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- The challenge for analytics on cyber network data
- Multi-scale network analysis approaches
- Analysis test environment
  - Netflow traffic analysis
  - RDB and EDA tools
  - VAST challenge data set
- Basic graph statistics
- Labeled graph degree distributions
- Time interval synchrony measurement

Challenge

Asymmetric Resilient Cybersecurity Initiative (ARC), PNNL
Research effort on modeling formalisms for general cyber systems

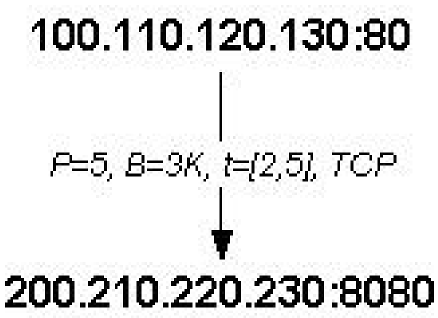
- **Cyber systems modeling needs unifying methodologies**
  - **Digital**: No space, ordinal time, no energy, no conservation laws, no natural metrics (continuity, contiguity)
  - **Engineered**: No methods from discovery-based science
- **Represent cyber systems as discrete mathematical objects interacting across hierarchically scalar levels**
  - Coarse-grained and fine-grained models
  - Each distinctly validated, but interacting
  - Similar to hybrid modeling and qualitative physics
    - Coarse grained discrete model
    - Constrains fine-grained continuous model
  - We are discrete all the way down
- **Utilize discrete mathematical foundations**
  - Labeled, directed graphs as a base representation of any discrete relation
  - But, equipped with additional constraints, complex attributes
  - And exploiting higher-order combinatorial structures and methods

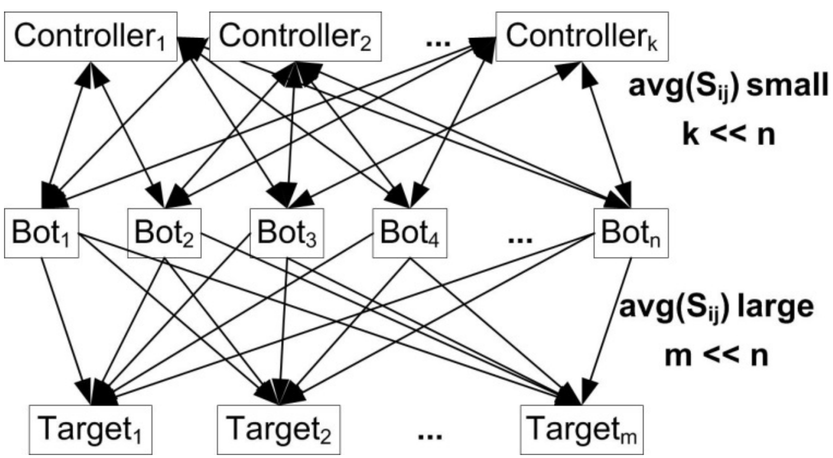
Netflow Focus

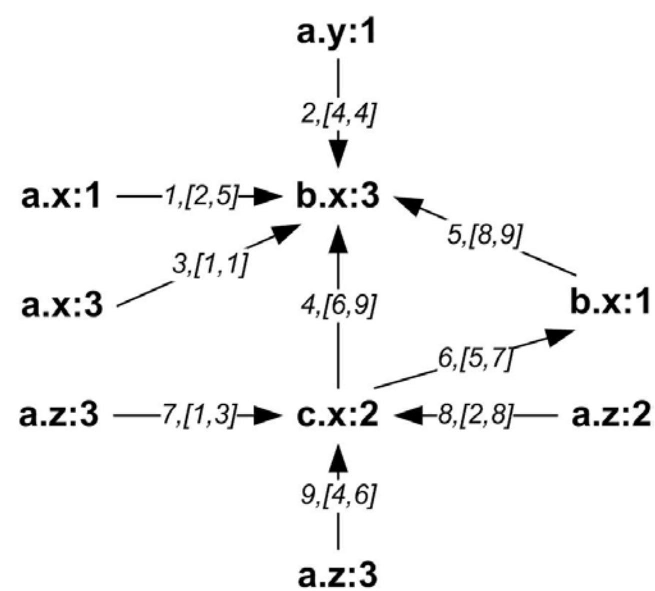
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

GOAL: ***Multi-scale network modeling***

- **Modeling assumption 1:** Netflow for first cut
  - Inherently multi-scale: drilldown to packet level, scalar "sweet spot"?
  - Broad interest beyond ARC
  - Ample use cases
  - Both public and private test databases available
- **Modeling assumption 2:** VAST Challenge fort test data
  - Open
  - Ground truth
  - Moderate size







Joslyn, CA; Choudhury, S; Haglin, D; Howe, B; Nickless, B; Olsen, B.: (2013) "Massive Scale Cyber Traffic Analysis: A Driver for Graph Database Research", *Proc. 1st Int. Wshop. on GRAph Data Management Experiences and Systems (GRADES 2013)*

Analysis Environment

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

- Test data sets

	VAST	CAIDA	Predict	NCCDC
Scope	Netflow	Packet	Packet	Netflow and Packet
# records/sample period		25M/min		65M/day
Total size	<10GB	Various	6 TB	
Payload?	Y	Y	Various	Y
Time stamps?	Y	Y	Various	Y
Total # records available	69M	Various	Various	133M
Distribution	Open	Registration	MOU	Open
Sample time period	2 weeks	Multiple	10 days	2 days
Sampling rate	Synthetically Generated	95%	?	?

- Currently scaling to O(100M) edges
  - Netezza TwinFin:
    - Parallel SQL databases appliance
    - Unique asymmetric massively parallel processing (AMPPTM) architecture
    - FPGAs for data filtering
  - Tableau 8.1 for EDA
- **Future:** Porting to PNNL's novel high-performance graph database engine GEMS, potential scaling to O(100B-1T) graph edges

Morari, A; Castellana, V; Tumeo, Antonino; Weaver, J; David Haglin, John Feo, Sutanay Choudhury, Oreste Villa: (2014) "Scaling Semantic Graph Databases in Size and Performance", *IEEE Micro*, 34:4, pp: 16-26

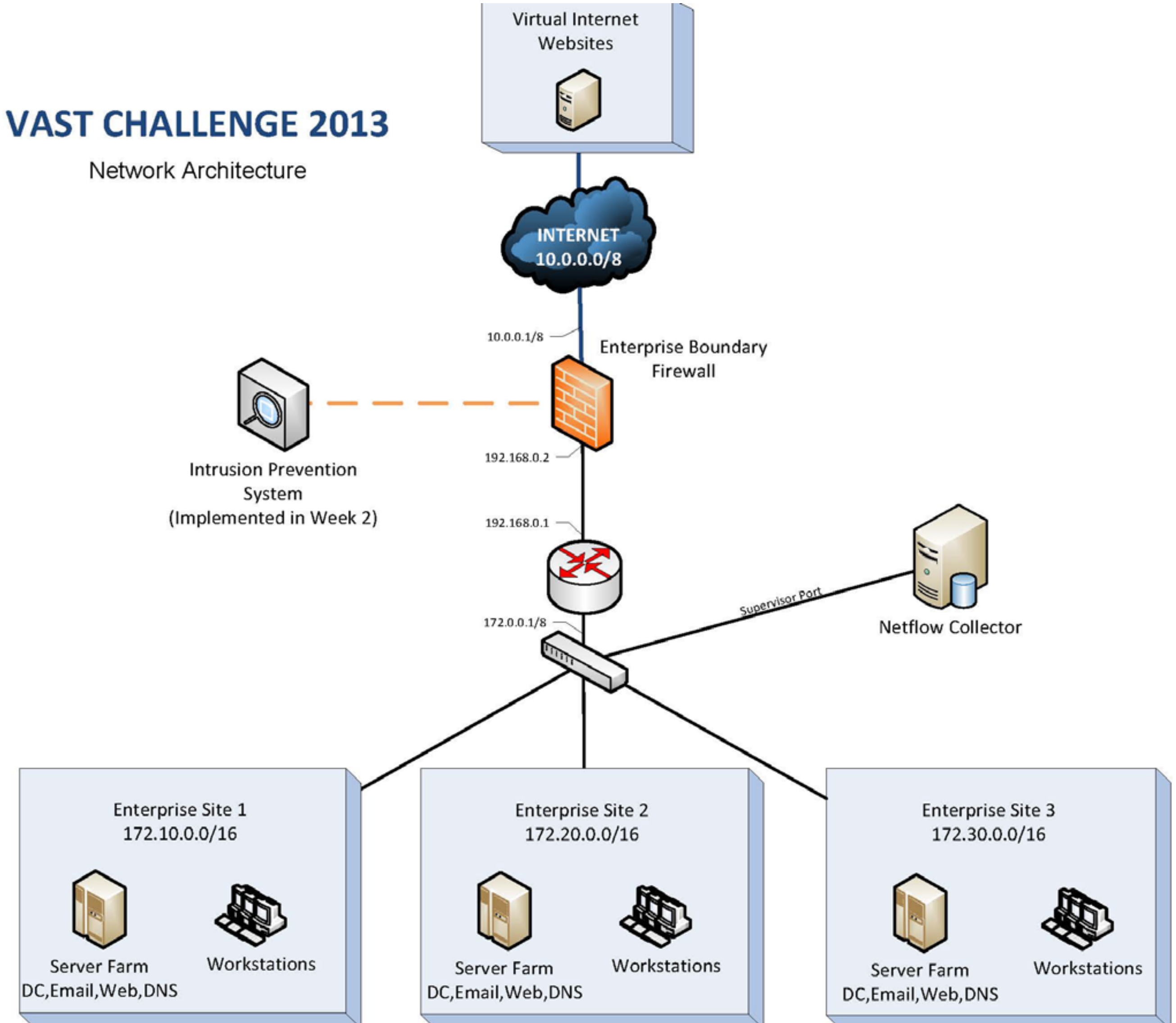
VAST Data Challenge

- Visual analytics competition co-led by PNNL since about 2005
- Co-located with Visual Analytics Science and Technology (VAST) conference
- Funded by and in the service of specific sponsors and their goals
- 2011-2013 focus on cyber challenge
- Scenario: Big Marketing Situational Awarenes
- PNNL-provided simulated netflow traffic
- Combined with IPS and BigBrother health monitoring
- Challenge
  - Provide visualizations for situational awareness
  - Report events during the timeline
- Submissions
  - About a dozen from universities, commercial partners, individuals

http://vacommunity.org/VAST+Challenge+2013

VAST Architecture

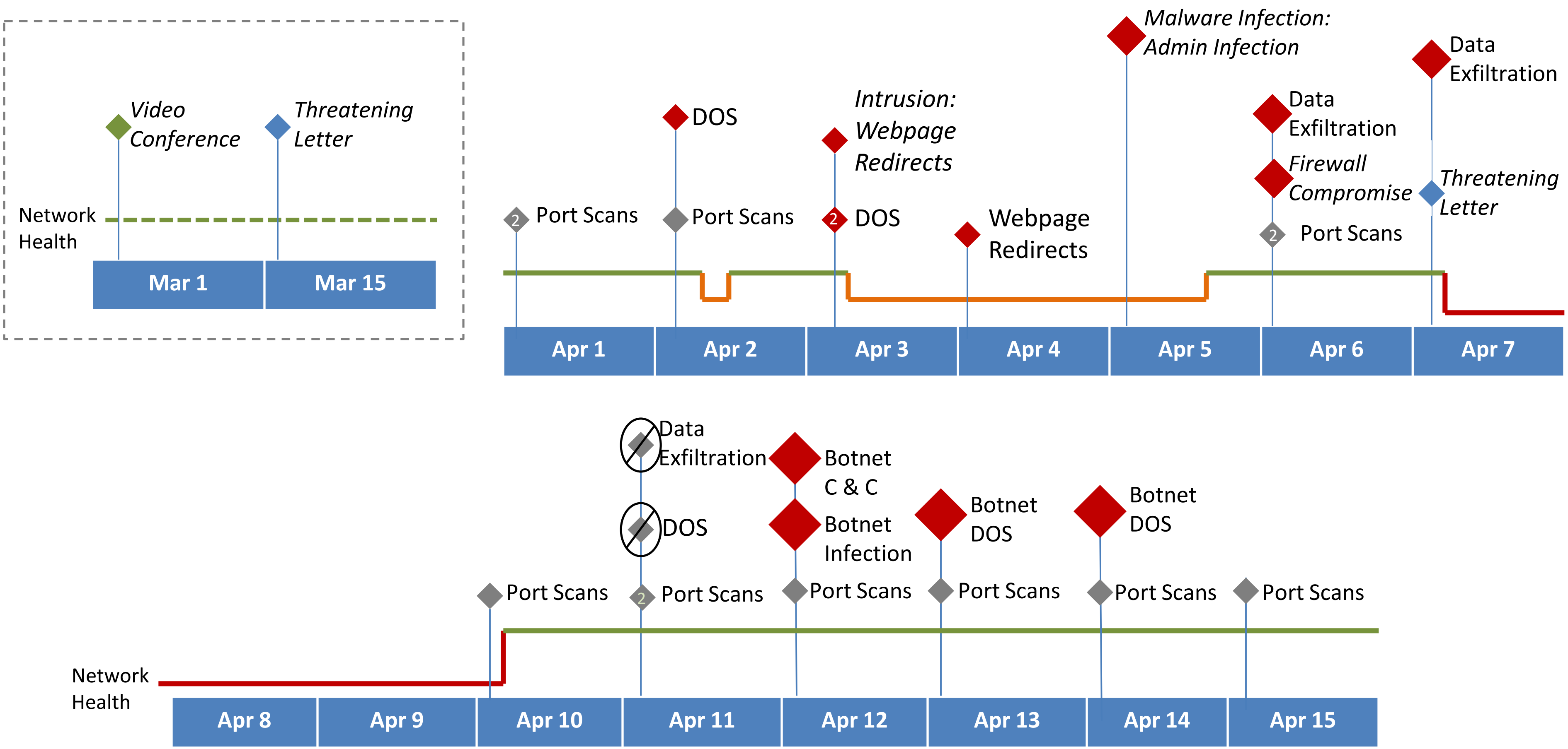
- Three BM sites
- Mostly web traffic
- Clients and servers both inside and outside
- Simulated external users hitting internal servers
- Some I/O ambiguity on bidirectional Netflow



Ground Truth

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965



Italics = Events that are not observable in supplied data

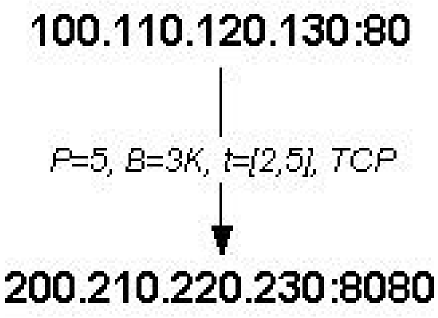
(red) = Attacks with serious consequences

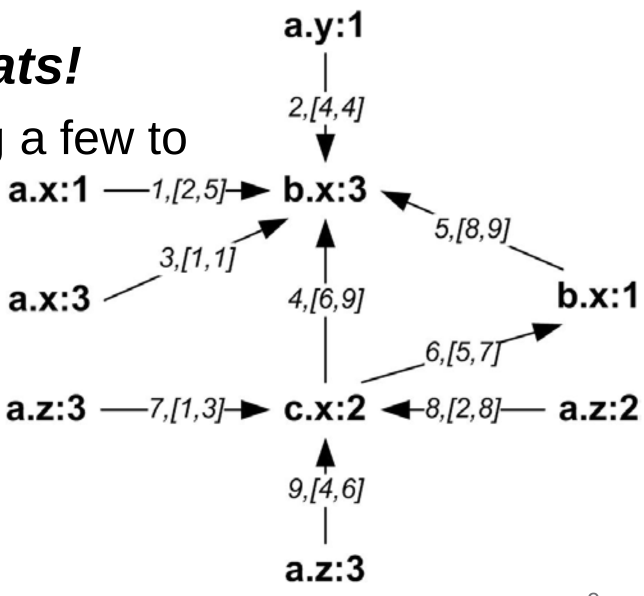
= Attack attempts blocked by IPS

Thanks to Kirsten Whitley

Netflow: Complex Data Space

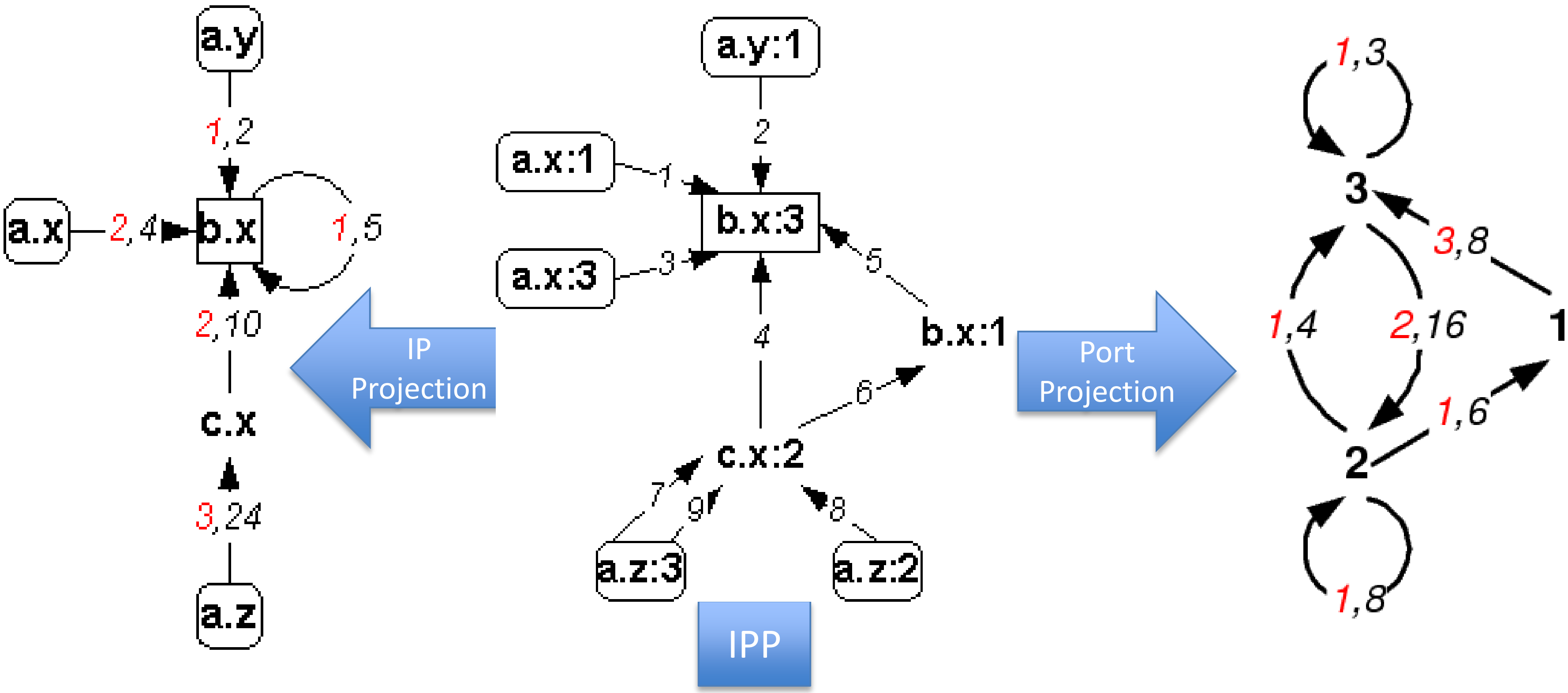
- Basic graph statistics: *all with Input X Output*
  - **Flow count**
  - **IPPs**
  - **IPs**
  - **Ports**
  - **Times:** Start, Finish, Durations
  - **Payload:** # packets, # bytes
  - **Transport protocol**
- ***Tremendous initial value just with basic stats!***
  - Many many, combinations, we're cherry-picking a few to show
- To which we bring our new measures:
  - **Degree distribution:**
    - Dispersion, Smoothness
    - Additional metrics
  - **Time intervals**





"Graph Cube" Contractions

- Projections in directed labeled graphs provide natural scalar levels
- **Netflow:** IPs and Ports



Zhao, Peixiang; Li, Xiaolei; Xin, Dong; and Han, Jiawei: (2011) "Graph Cube: On Warehousing and OLAP Multidimensional Networks", SIGMOD 2011

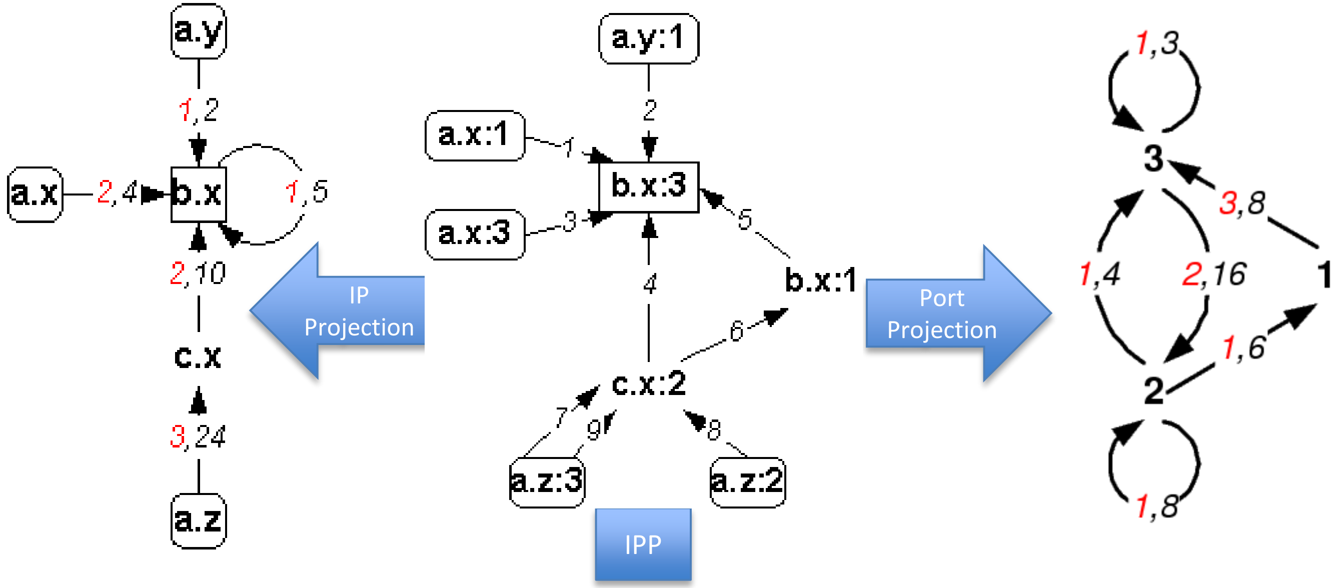
Basic Graph Statistics: VAST

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

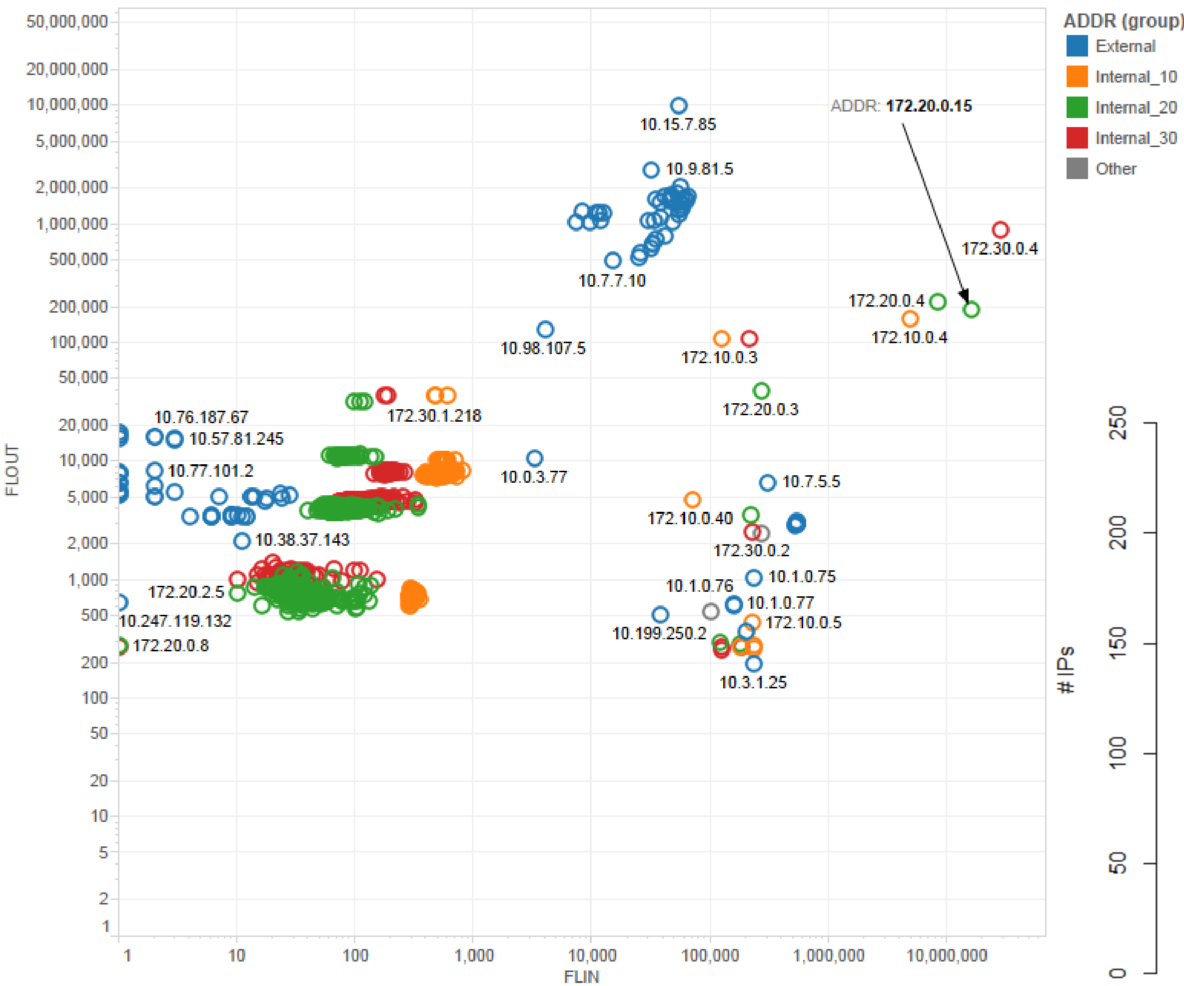
VAST IP		Mean flows per
Flows	*69,396,995*	
Nodes	*1,440*	48,192
Outs	*1,424*	48,734
Leaves	16	1.1%
Ins	*1,345*	51,596
Roots	95	6.6%
Internals	1,329	92.3%
Pairs present	*30,161*	2,301
Pairs possible	1,915,280	36
Density	1.57%	
Mean Ports/IP	6,990.41	

VAST IPP		Mean flows per
Flows	*69,396,995*	
Nodes	*10,066,187*	6.89
Outs	*8,784,807*	7.90
Leaves	1,281,380	12.7%
Ins	*2,533,742*	27.39
Roots	7,532,445	74.8%
Internals	1,252,362	12.4%
Pairs present	*14,387,421*	4.82
Pairs possible	22,258,434,457,794	0.00000312
Density	0.0000646%	

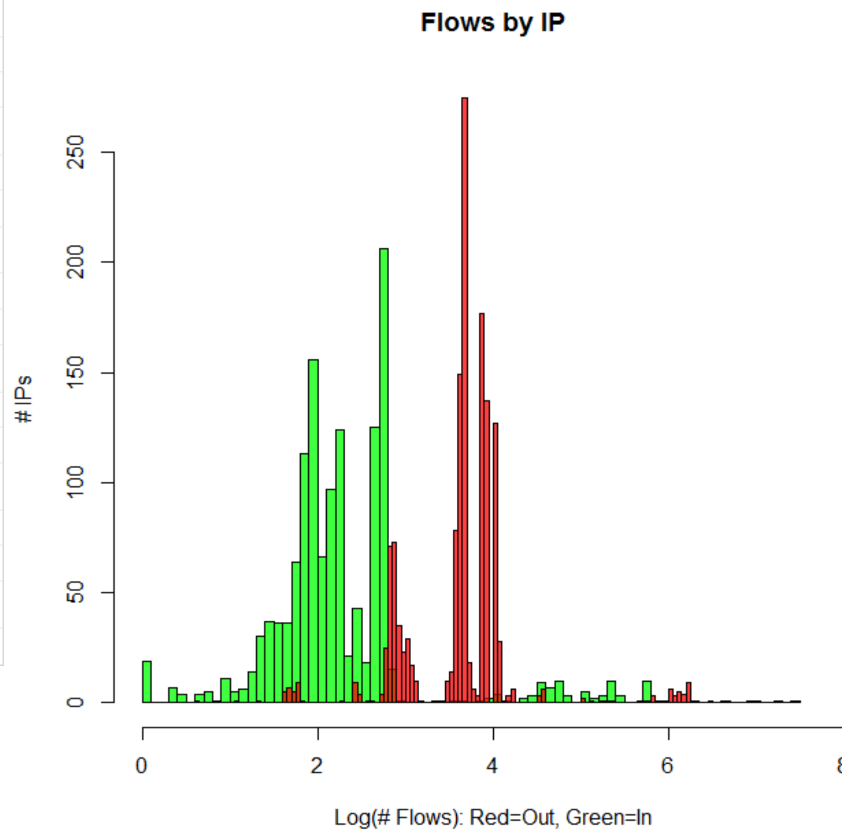
VAST Port		Mean flows per
Flows	*69,396,995*	
Nodes	*65,536*	1,058.91
Outs	*64,501*	1,075.91
Leaves	1,035	1.6%
Ins	*65,536*	1,058.91
Roots	-	0.0%
Internals	64,501	98.4%
Pairs present	*986,385*	70.35
Pairs possible	4,227,137,536	0.01641702
Density	0.023%	



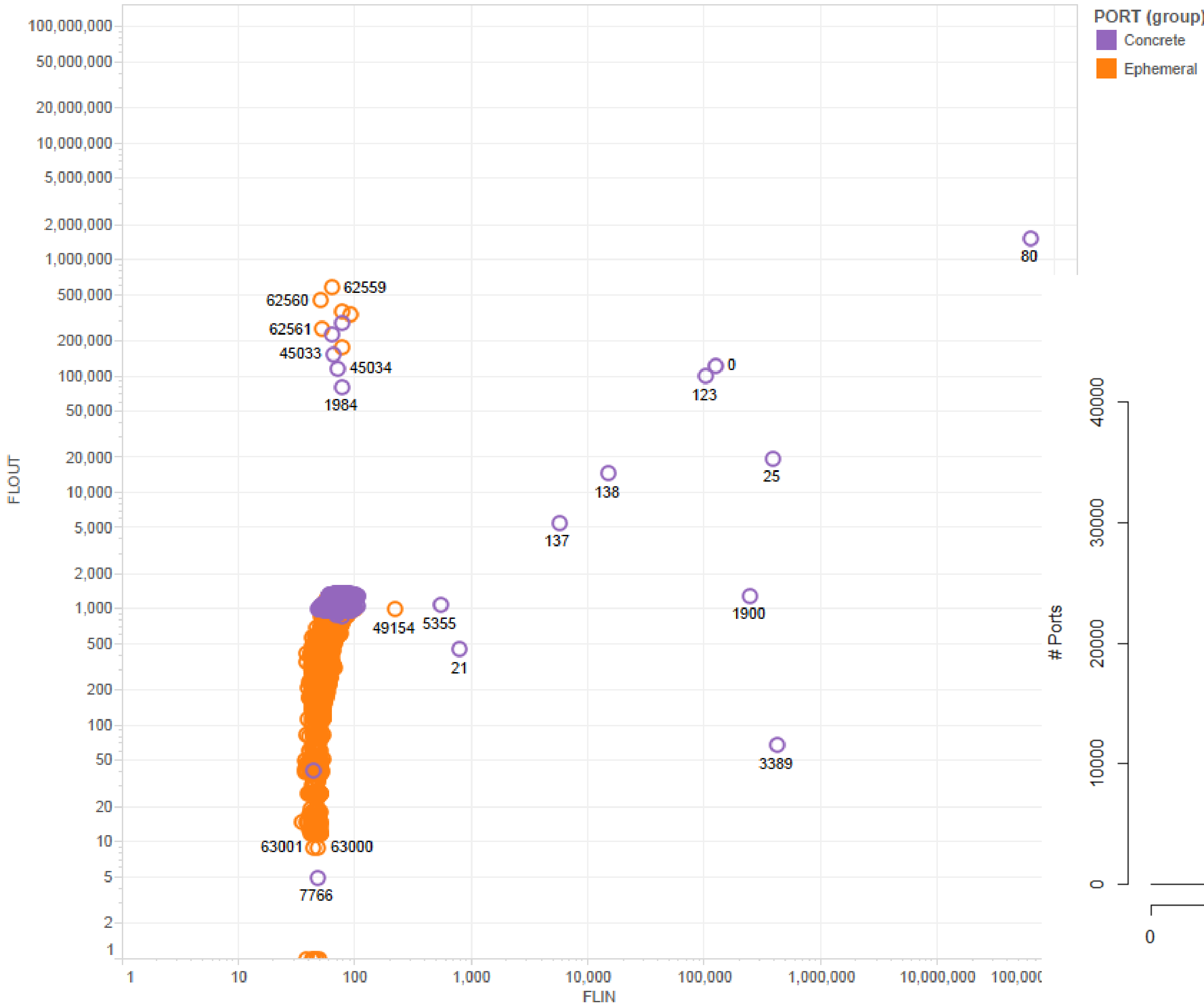
Flows by IP

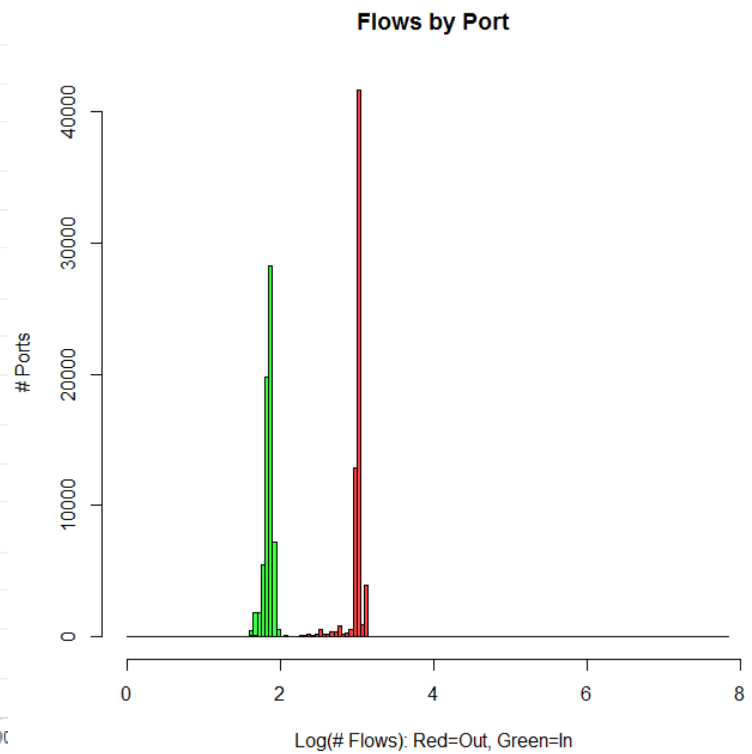


- # 0 in: 95
- # 0 out: 16
- # > 0 on both: 1328



Flows by Port



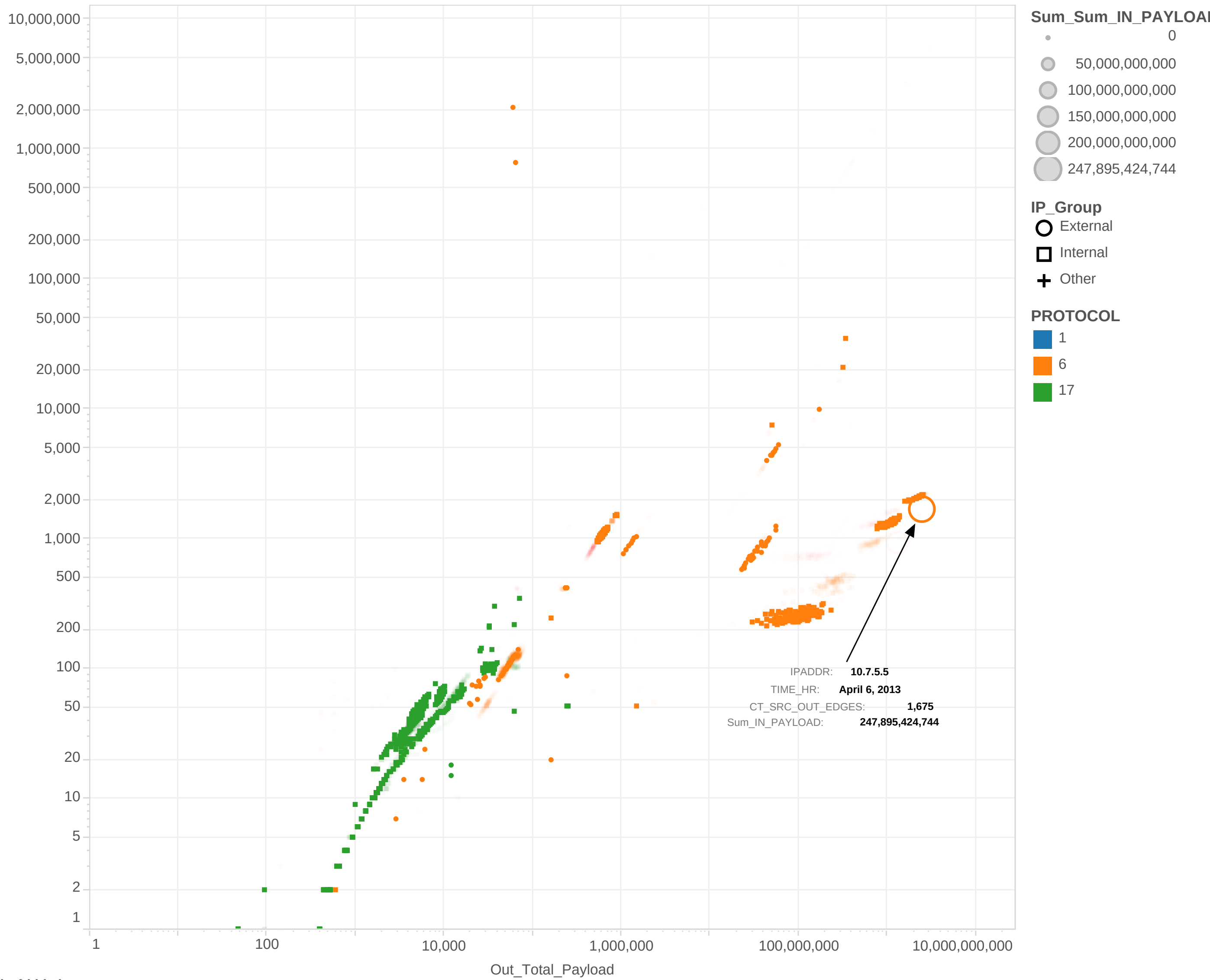


Basic Payload View: Exfiltration

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Basic Payload View: Exfiltration

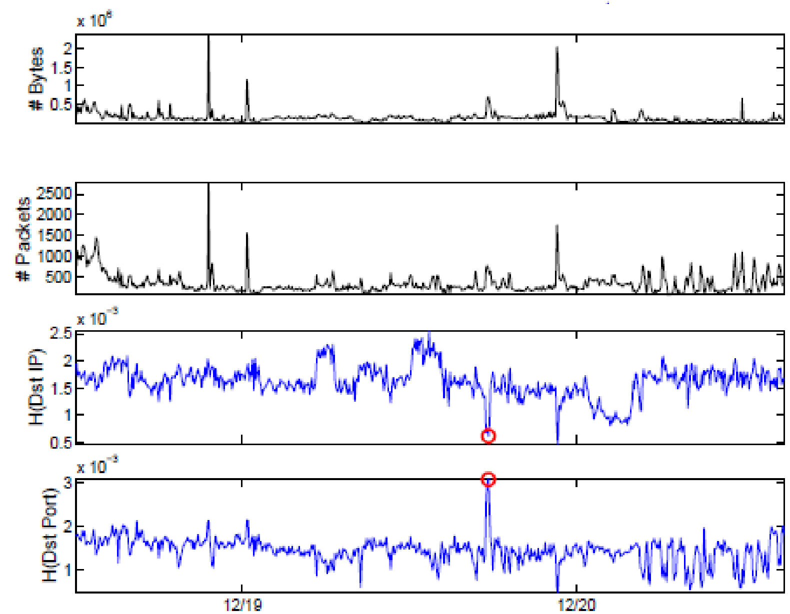
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



Beyond *Volume* for Anomaly Detection

Anomaly Label	Definition	Traffic Feature Distributions Affected
Alpha Flows	Unusually large volume point to point flow	Source address, destination address (possibly ports)
DOS	Denial of Service Attack (distributed or single-source)	Destination address, source address
Flash Crowd	Unusual burst of traffic to single destination, from a "typical" distribution of sources	Destination address, destination port
Port Scan	Probes to many destination ports on a small set of destination addresses	Destination address, destination port
Network Scan	Probes to many destination addresses on a small set of destination ports	Destination address, destination port
Outage Events	Traffic shifts due to equipment failures or maintenance	Mainly source and destination address
Point to Multipoint	Traffic from single source to many destinations, *e.g.*, content distribution	Source address, destination address
Worms	Scanning by worms for vulnerable hosts (special case of Network Scan)	Destination address and port

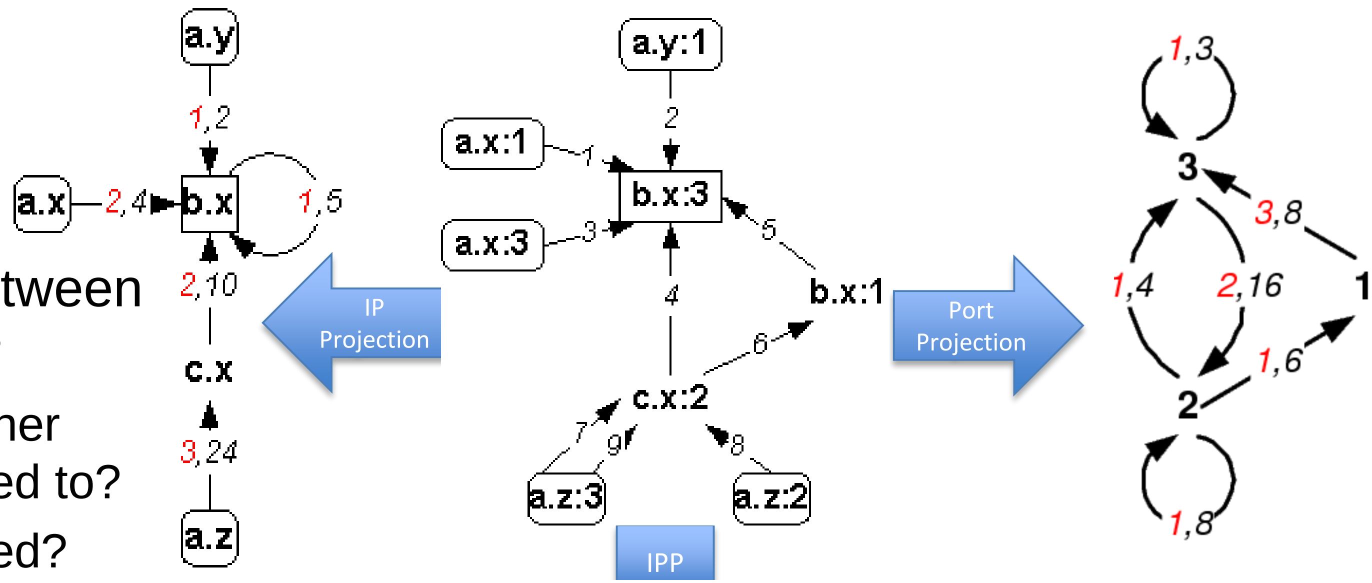
- Packets and bytes not always sufficient to identify behavioral patterns
- IP and port behavior can tell the difference
  - E.g. port scan in figure
  - Entropy of DstIP, DstPort



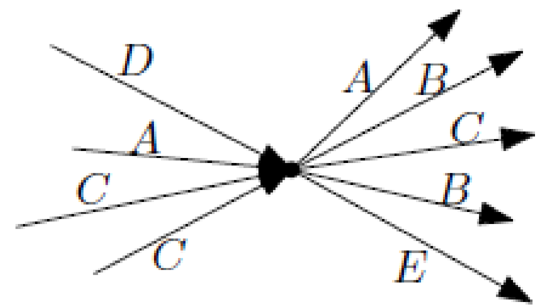
A Lakhina, M Crovella, C Diot: (2005) "Mining Anomalies Using Traffic Feature Distributions", *SIGCOMM 05*

Labeled Degree Distributions

- How can we characterize relationships between IPs, Ports, etc.?
  - How many other IPs/ports talked to?
  - How distributed?



- Analyze the distributions of labels
- Incoming and outgoing
- IPs, Ports, IPPs
- ***Labeled degree distributions***



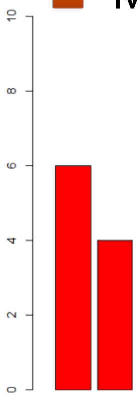
Input: C/A/D = 2/1/1

Output: B/A/C/E = 2/1/1/1

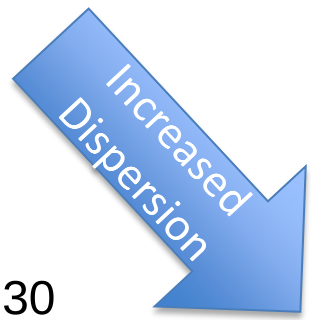
Joint: C/A/B/D/E = 3/2/2/1/1

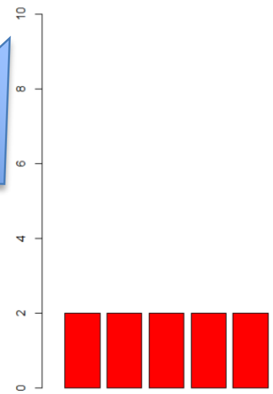
Information Measures of IP/Port Distributions

- **DISPERSION:**
  - # IPs, ports relative to # flows
  - Math: Log count ratio
- **SMOOTHNESS:**
  - Even or lumpy distribution of IPs, ports
  - Math: Normalized entropy

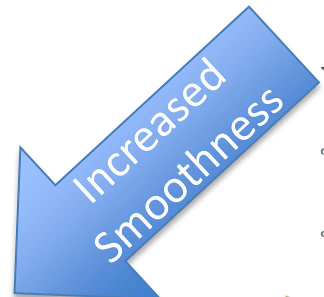


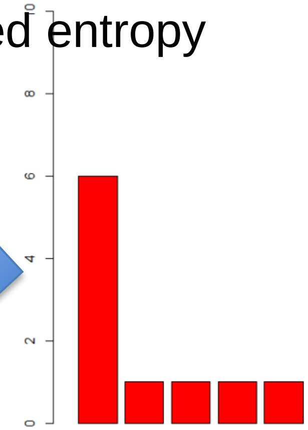
- Dispersion = 0.30
- Smoothness = 0.97





- Dispersion = 0.70
- Smoothness = 1.00





- Dispersion = 0.70
- Smoothness = 0.76

CA Joslyn, W Cowley, EA Hogan, B Olsen: (2014) “Discrete Mathematical Approaches to Graph-Based Traffic Analysis” *2014 Int. Wshop. on Engineering Cyber Security and Resilience (ECSaR14)* http://www.ase360.org/bitstream/handle/123456789/157/ecsar2014_paper4.pdf

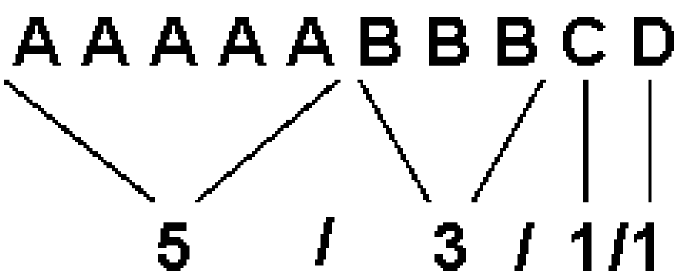
Labeled Degree Distributions

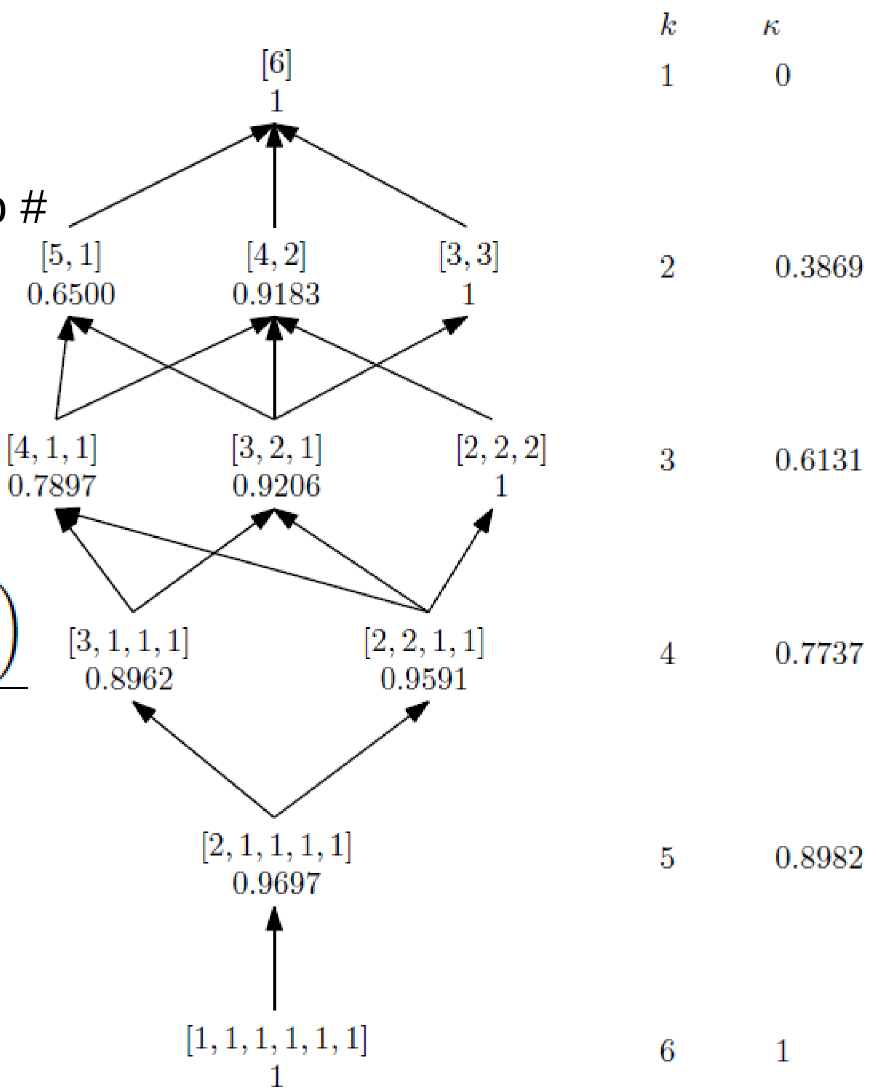
- *Information measures on integer partitions*
- *N* flows distributed into *m* <= *N* "buckets"
- **Dispersion:** How *many* buckets *m* relative to # flows *N*?

$$\kappa(\vec{C}) = \frac{\log_2(m)}{\log_2(N)}$$

- **Smoothness:** How *smoothly* are those *N* flows distributed over the *m* buckets?

$$G(\vec{C}) := \frac{\mathbf{H}(f(\vec{C}))}{\log_2(m)} = \frac{-\sum_{l=1}^{m} \frac{C_l}{N} \log_2\left(\frac{C_l}{N}\right)}{\log_2(m)}$$





Smoothness with Dispersion

- **Smoothness** is definitely significant
  - Lakhina *et al.* use IP/port smoothness (entropy) only
  - Able to identify many behavioral patterns
    - Bullet: > 1 sigma significant
    - Star: > 2 sigma significant

Anomaly	H(srcIP)	H(srcPort)	H(dstIP)	H(dstPort)
Alpha	-0.38 ± 0.32 •	-0.19 ± 0.47	-0.37 ± 0.33 •	-0.35 ± 0.35
DOS	-0.05 ± 0.57	-0.20 ± 0.51	-0.35 ± 0.20 •	-0.08 ± 0.49
Flash	0.21 ± 0.49	0.49 ± 0.26 •	-0.28 ± 0.22 •	0.13 ± 0.58
Port Scan	-0.33 ± 0.19 •	0.07 ± 0.40	-0.41 ± 0.15 *	0.70 ± 0.14 *
Net. Scan	-0.19 ± 0.22	0.84 ± 0.17 *	0.20 ± 0.21	-0.29 ± 0.16 •
Outage	0.51 ± 0.33 •	0.31 ± 0.31	0.51 ± 0.34 •	0.24 ± 0.20
Pt.-Mult.	-0.18 ± 0.16 •	-0.17 ± 0.12 •	0.66 ± 0.04 *	0.68 ± 0.06 *
Unknown	-0.28 ± 0.39	0.02 ± 0.46	-0.35 ± 0.34	0.17 ± 0.55
False	-0.01 ± 0.49	0.27 ± 0.46	-0.00 ± 0.46	-0.04 ± 0.57

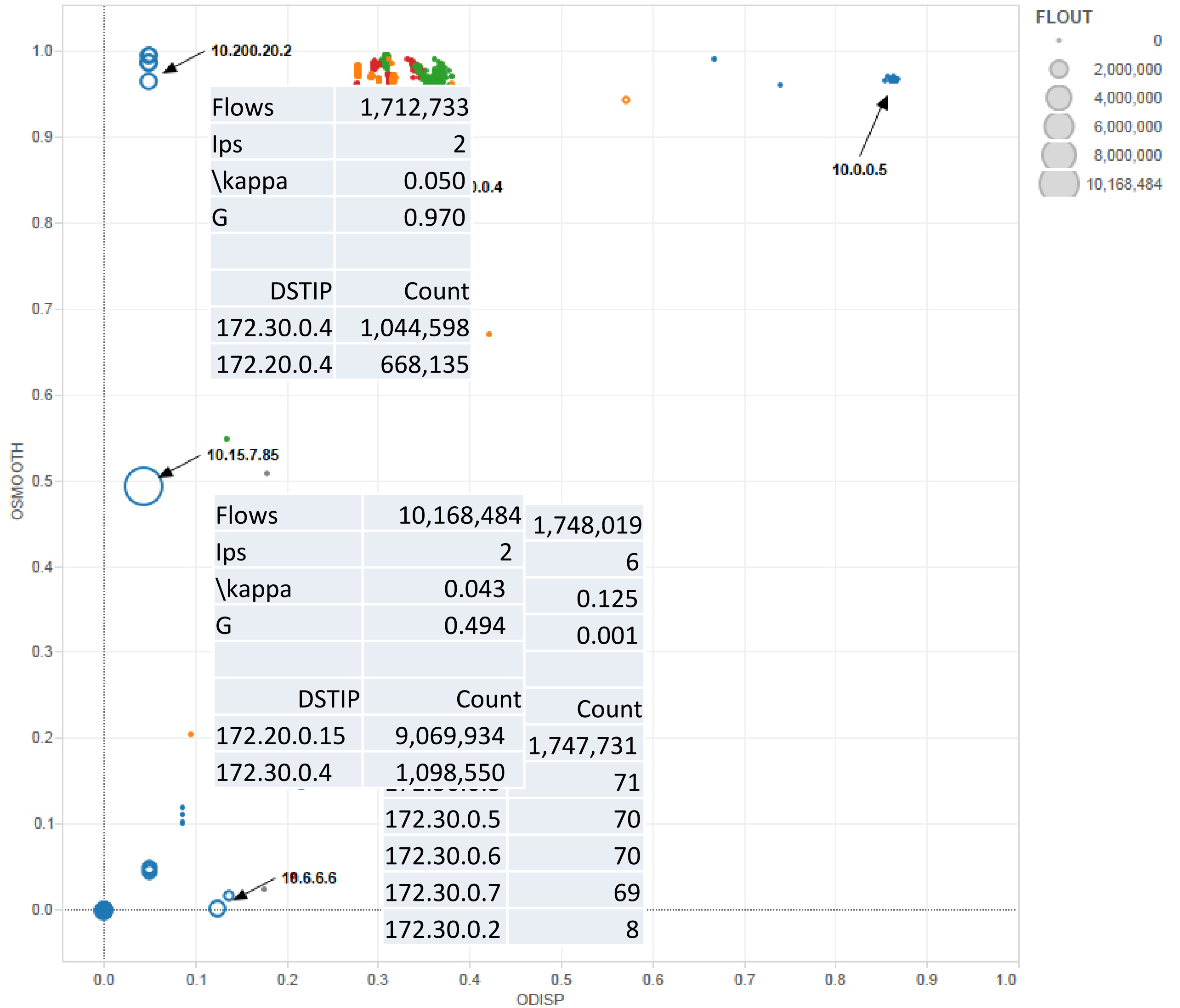
- **Dispersion** adds great value
  - Simpler computational
  - Mathematically necessary together with smoothness
  - We believe even *more* significant methodologically

A Lakhina, M Crovella, C Diot: (2005) “Mining Anomalies Using Traffic Feature Distributions”, *SIGCOMM 05*

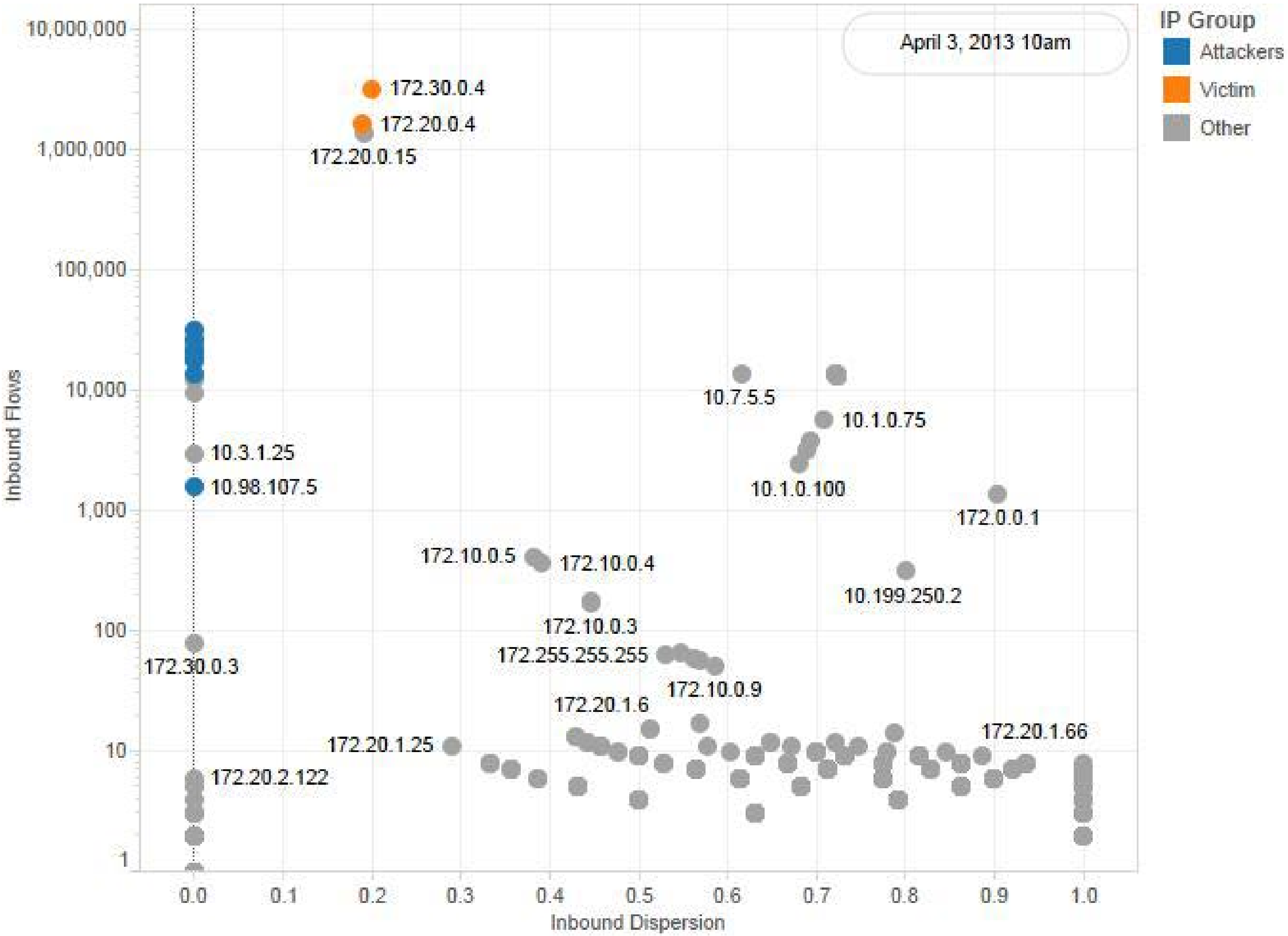
IP Distributional Statistics

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **Servers:** Unexceptional
- **Attackers:** Small dispersion, smoothness related to # victims
- **Upper right:** Outlier artifacts from simulation



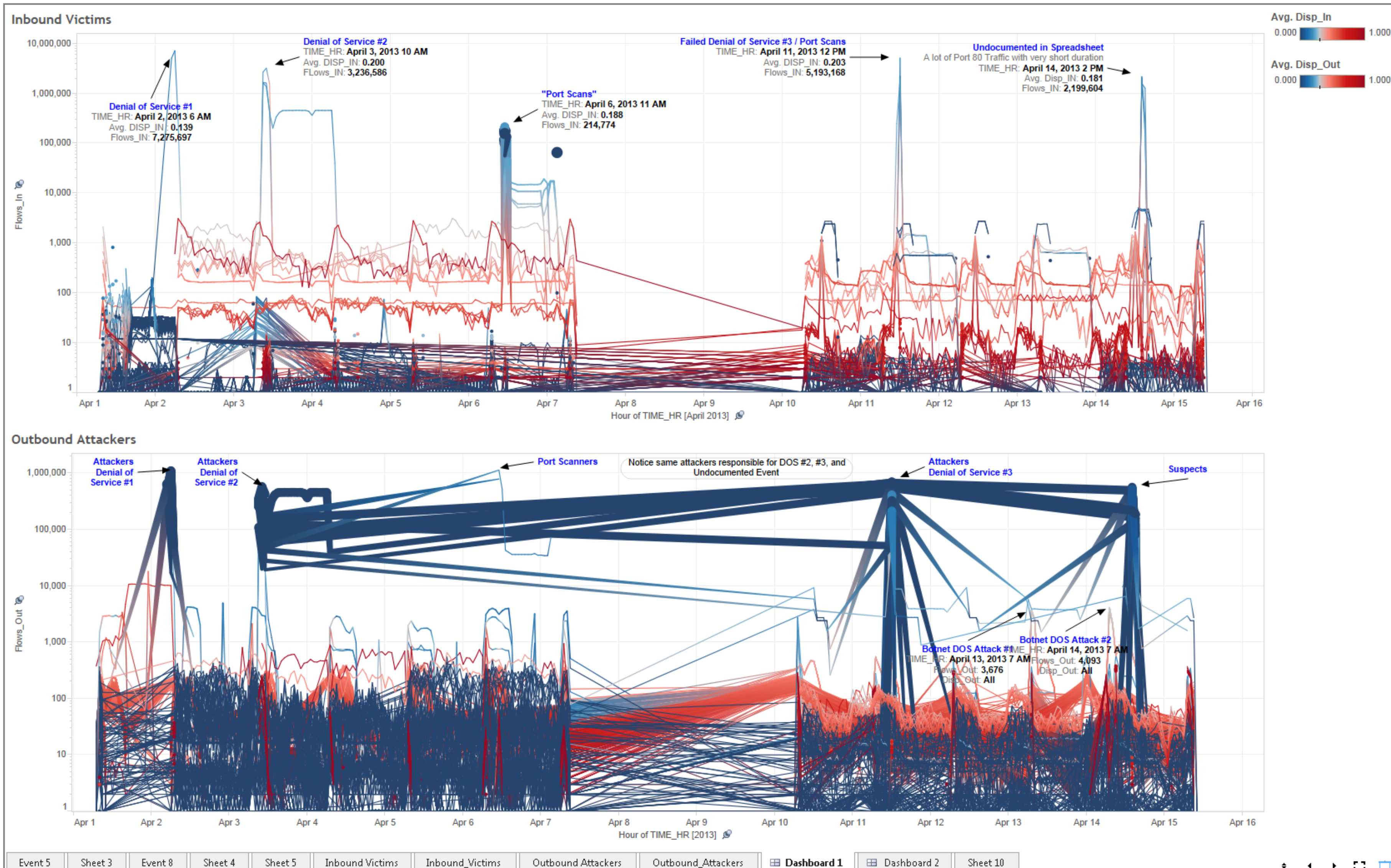
DOS Attack



April 3, 2013 10am
IP Group
Attackers
Victim
Other
Inbound Flows
Inbound Dispersion
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
172.30.0.4
172.20.0.4
172.20.0.15
10.3.1.25
10.98.107.5
172.30.0.3
172.20.2.122
10.7.5.5
10.1.0.75
10.1.0.100
172.0.0.1
172.10.0.5
172.10.0.4
172.10.0.3
10.199.250.2
172.255.255.255
172.10.0.9
172.20.1.6
172.20.1.25
172.20.1.66

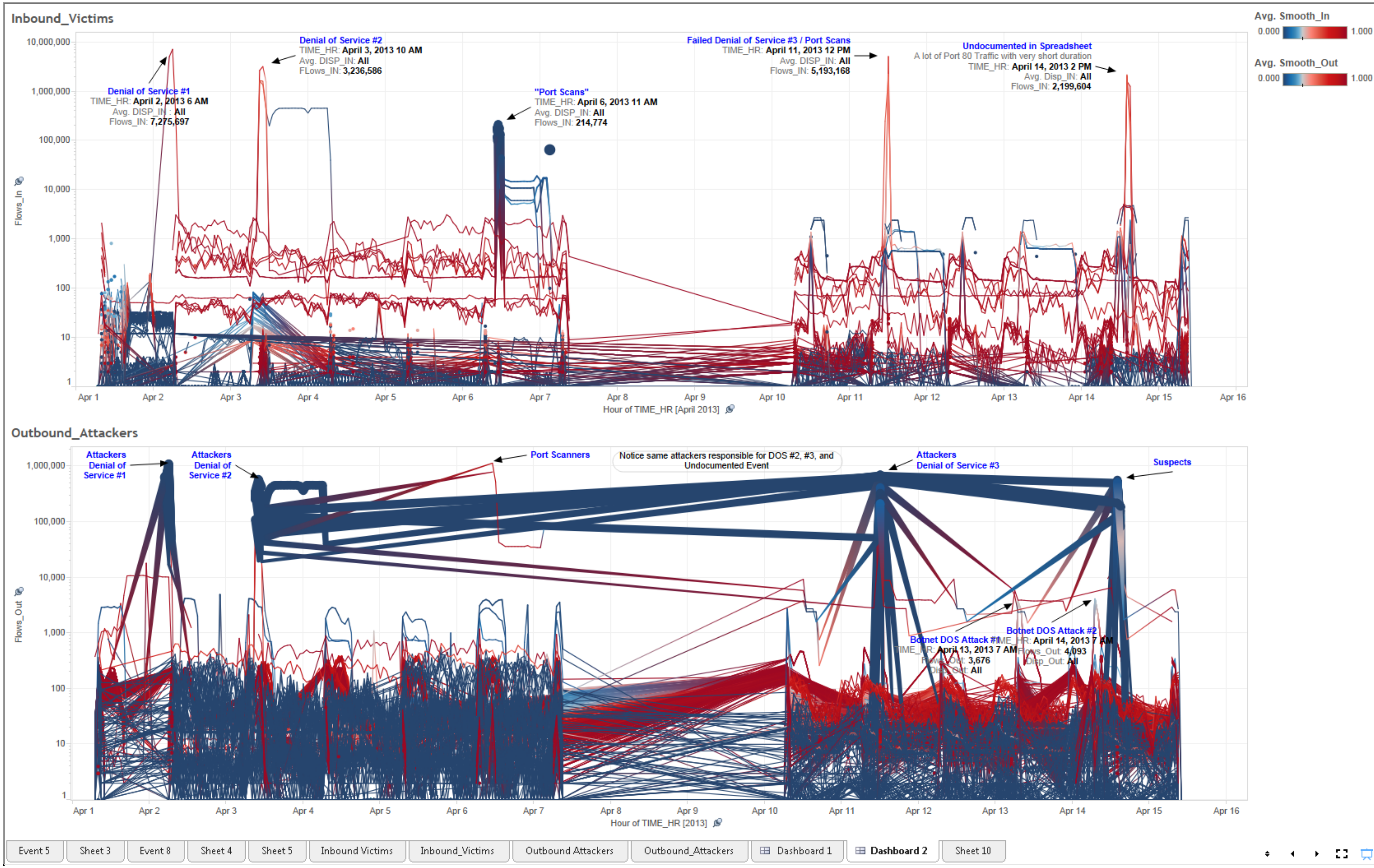
Attacks: Flows and Dispersion

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*



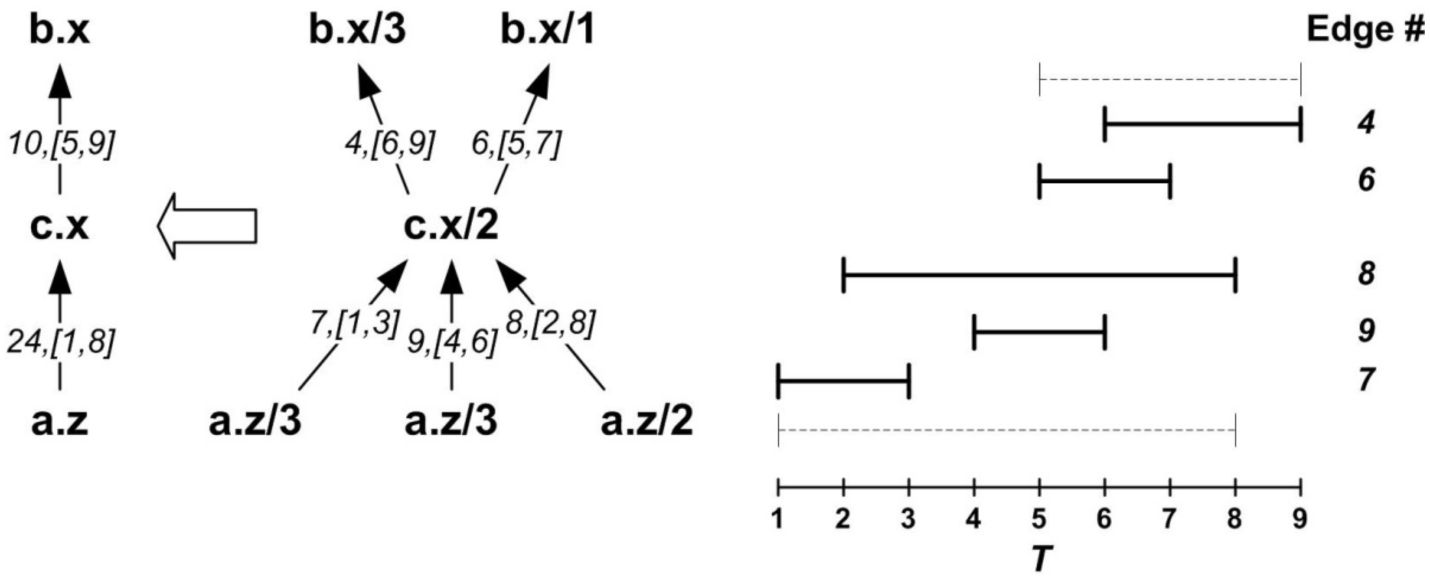
Attacks: Flows and Smoothness

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



Time Intervals

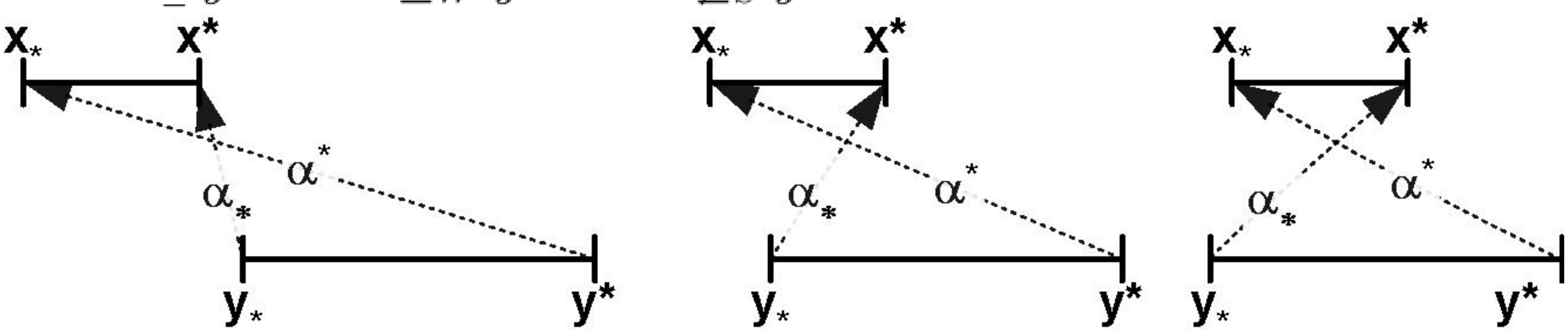
- Series and parallel relations between events
- Aggregations over graph contractions
- Measures of synchrony



Interval Orders

- $\overline{\mathbb{R}}$ = the set of all real intervals $\bar{x} = [x_*, x^*] \in \overline{\mathbb{R}}$, where $x_* \le x^* \in \mathbb{R}$.
  Strong Order: $\bar{x} \le_S \bar{y} \; := x^* < y_*$ or $\bar{x} = \bar{y}$.
  Weak Order: $\bar{x} \le_W \bar{y} \; := x_* \le y_*$ and $x^* \le y^*$.
  Subset Order: $\bar{x} \subseteq \bar{y} \; := x_* \ge y_*$ and $x^* \le y^*$.
- **Dual Orders:** $\ge_S, \ge_W, \supseteq$
- $\bar{x} \le_S \bar{y} \rightarrow \bar{x} \le_W \bar{y}$
- **Near Conjugacy:** $\bar{x} \le_W \bar{y}$ iff $\bar{x} \not\subseteq \bar{y}$, where no endpoints are equal
- Proper intersection (from the left) (not an order):

$$\bar{x} \circ_{\le} \bar{y} \; := \bar{x} \le_W \bar{y} \text{ and } \bar{x} \not\le_S \bar{y}.$$



Joslyn, Cliff; Hogan, Emilie; and Pogel, Alex: (2014) “Interval Valued Rank in Finite Ordered Sets”, submitted, arXiv:1409.6684

Interval Operations

Addition (interval, Minkowski sum): $\bar{x} + \bar{y} \ := [x_* + y_*, x^* + y^*]$

Subtraction (interval): $\bar{x} - \bar{y} \ := [x_* - y^*, x^* - y_*]$

Absolute Value (interval): $|\bar{x}| = [|\bar{x}|_*, |\bar{x}|^*]$, where

$$|\bar{x}|_* \ := \begin{cases} 0, & x_* x^* \leq 0 \\ \min(|x_*|, |x^*|), & x_* x^* > 0 \end{cases}$$

$$|\bar{x}|^* \ := \ \max(|x_*|, |x^*|).$$

Separation (interval): $\|\bar{x}, \bar{y}\| \ := |\bar{x} - \bar{y}|$

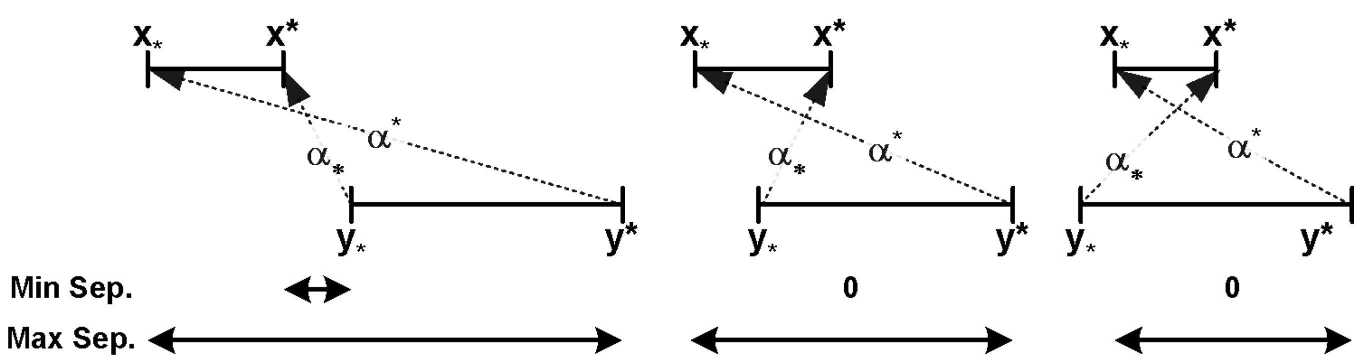
Midpoint (scalar): $\hat{\bar{x}} = \frac{x_* + x^*}{2} \in \mathbb{R}$

Width (scalar): Scalar values: $W(\bar{x}) \ := |x^* - x_*| \in \mathbb{R}$.

Mean (interval): For $X = \{\bar{x}_i\}_{i=1}^N$, $\text{mean}(X) \ := \frac{\sum_{i=1}^N \bar{x}_i}{N}$

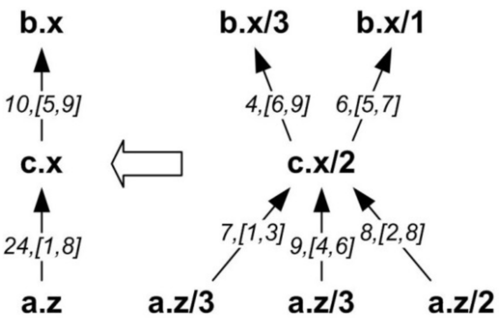
Union Over Gaps (interval):

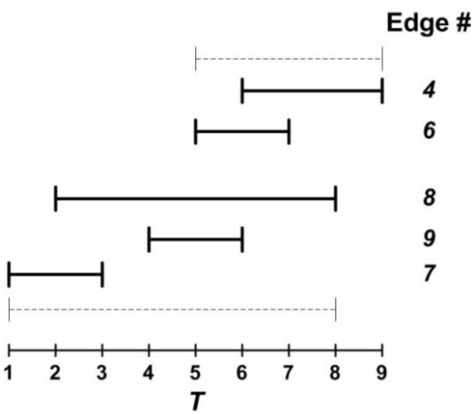
$$\bar{x} \cup \bar{y} \ := [\min(x_*, y_*), \max(x^*, y^*)]$$

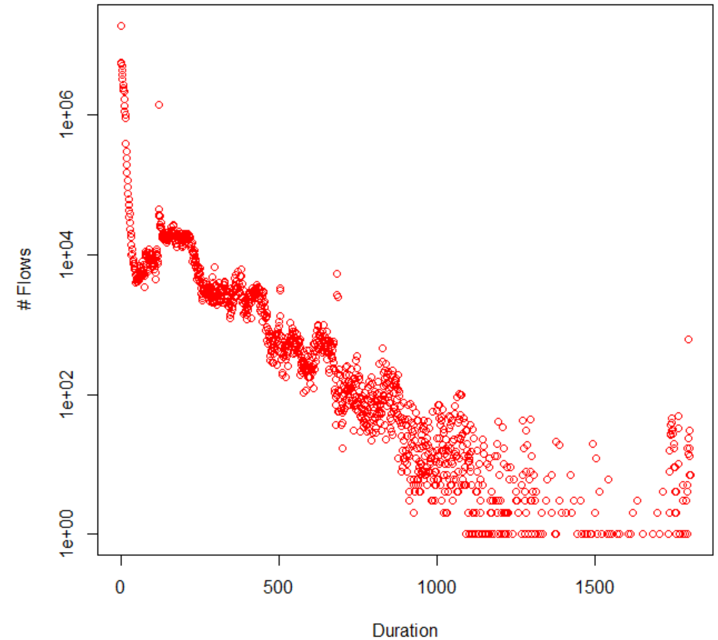


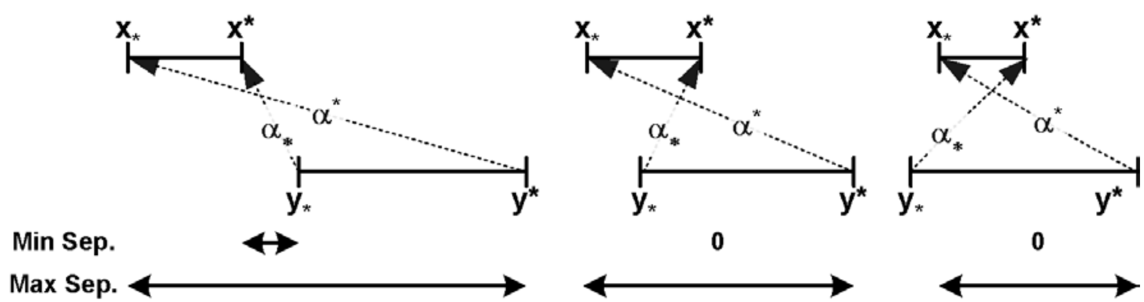
Interval Analyses

- **First effort:** Overall statistical analysis
  - Average widths
  - Counts for three overlap categories
  - Amount of overlap
- Problem in VAST: Too many short flows



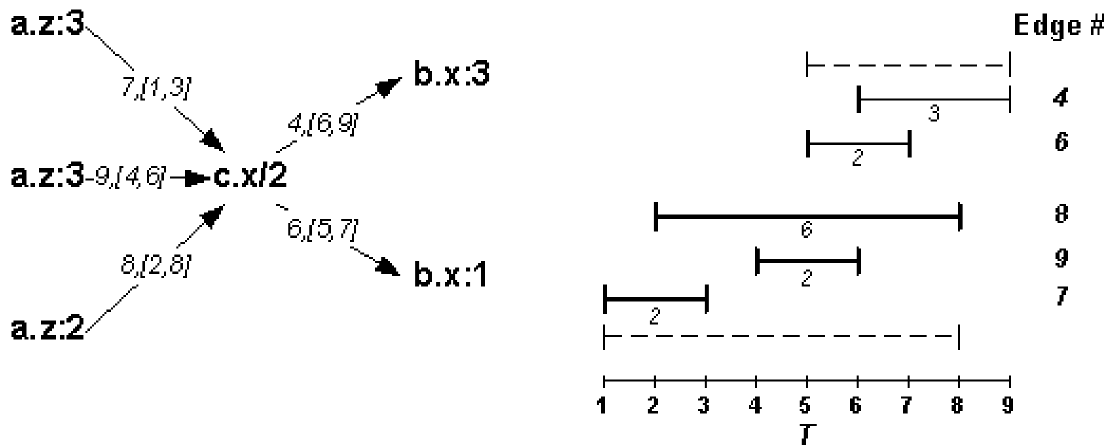


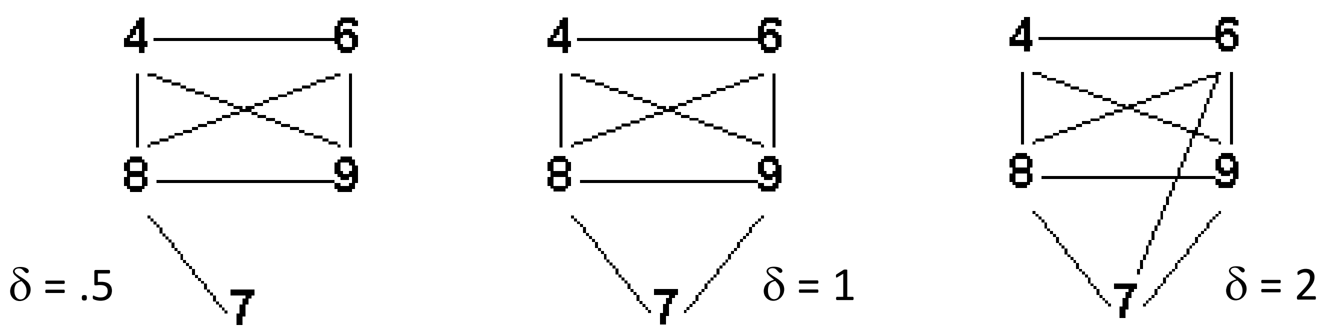




Metcalf's "Encounter Graphs"

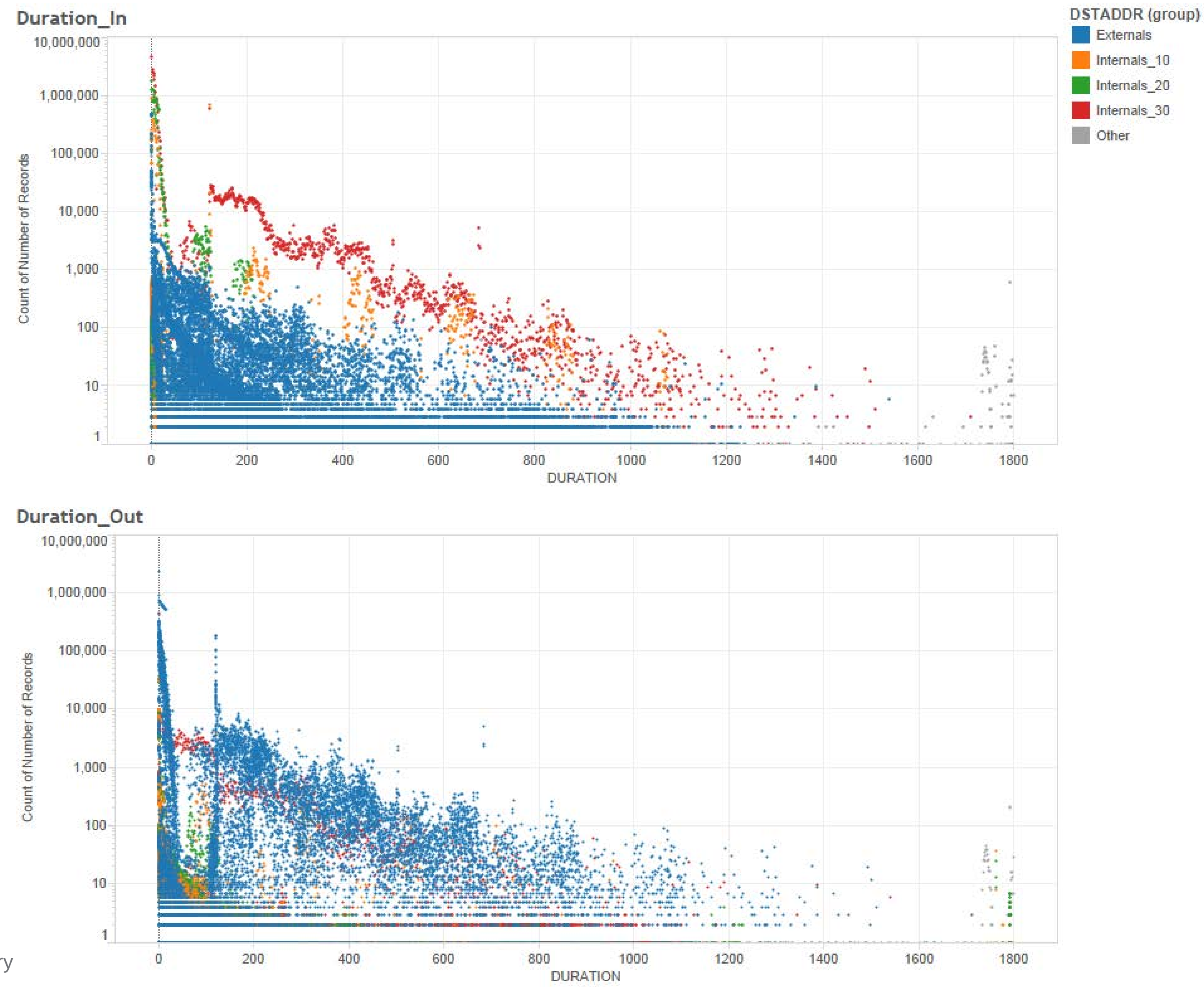
- ***Undirected*** links between ***edges***
- Link if intervals overlap or are separated by no more than δ



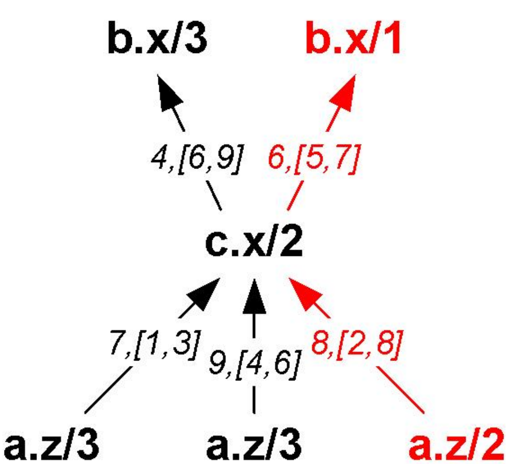


Metcalf, Leigh: (2014) "Analyzing Flow Using Encounter Complexes", Flocon 2014

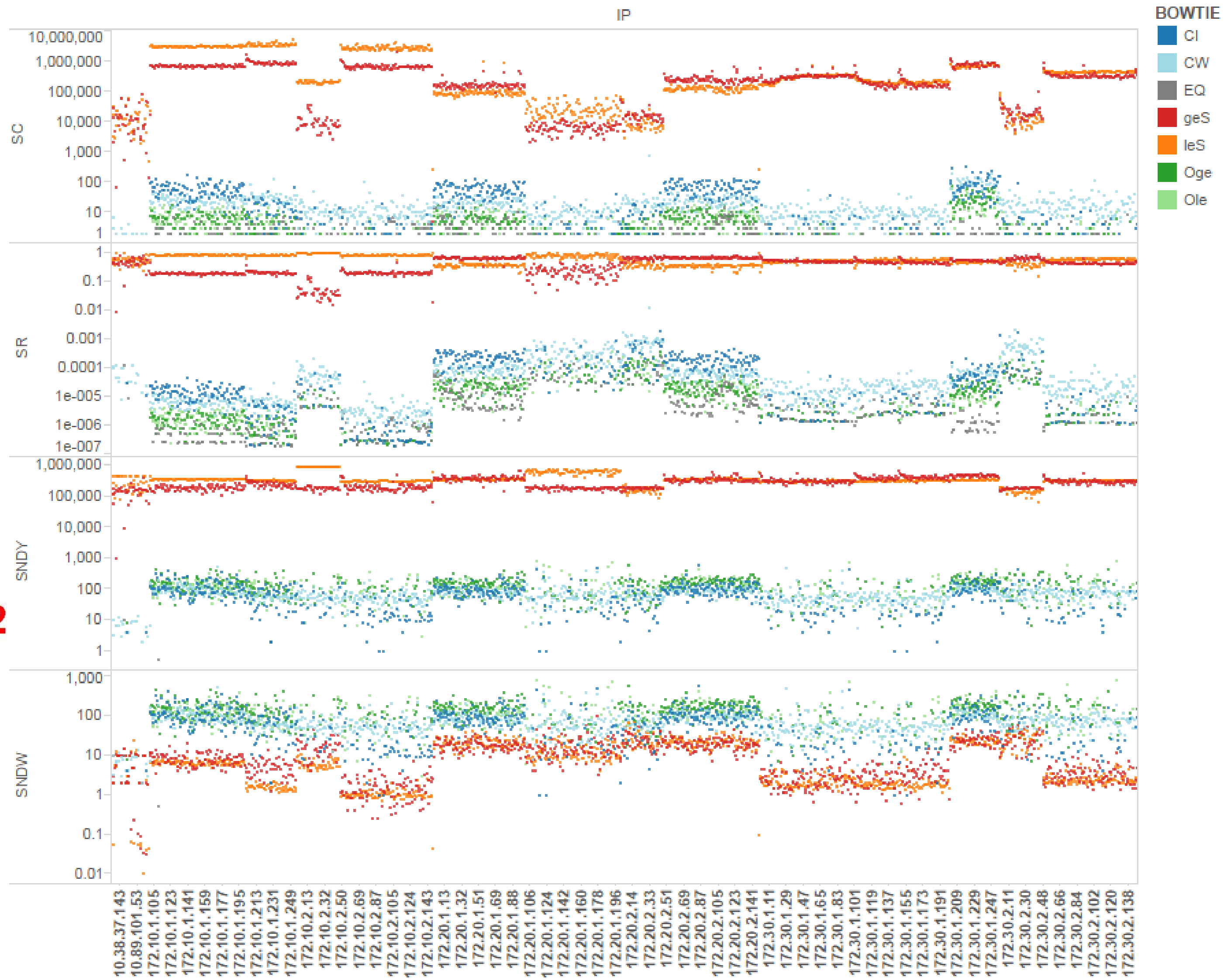
Durations by IP Group



IPs by Order Relation: Series Motifs

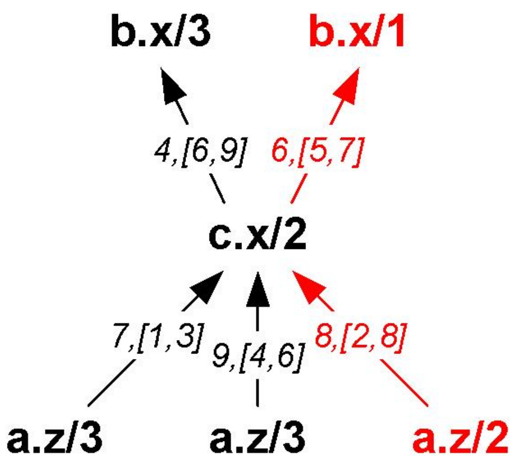


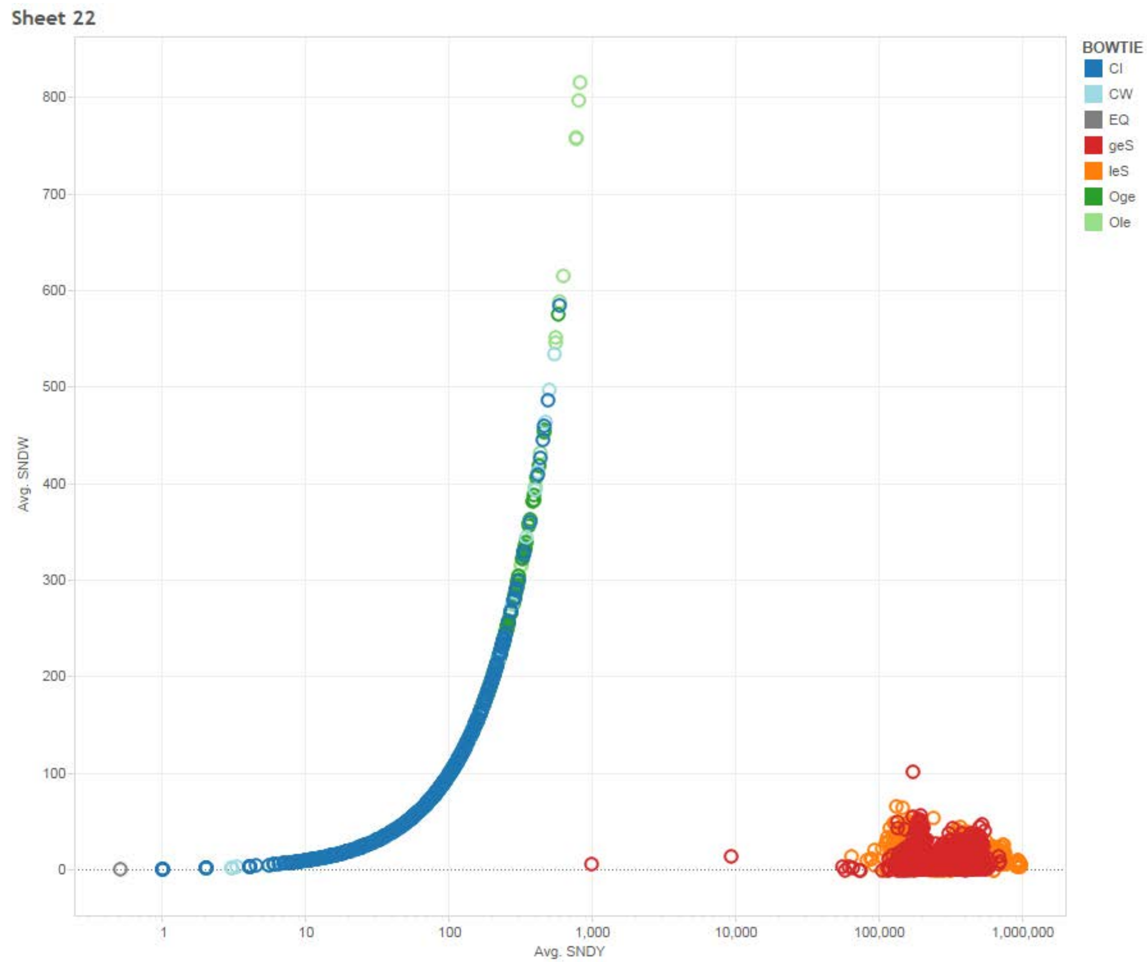
Sheet 20



Sum of SC, sum of SR, sum of SNDY and sum of SNDW for each IP. Color shows details about BOWTIE.

Max Separation and Width by Order Relation: Series Motifs





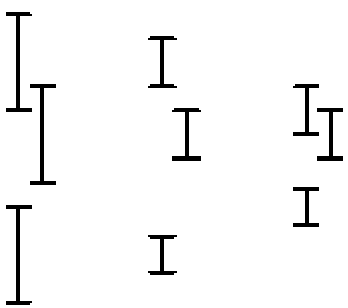
Average of SNDY vs. average of SNDW. Color shows details about BOWTIE. Details are shown for IP.

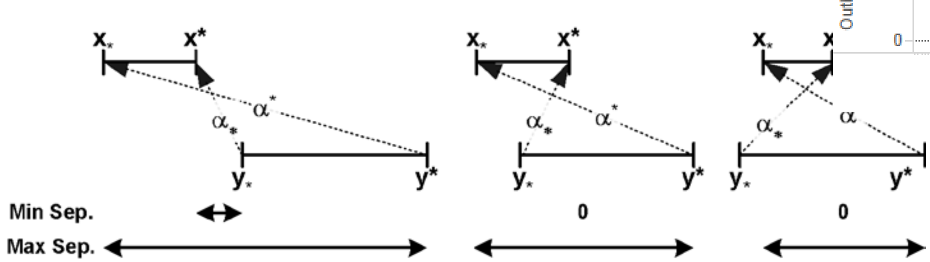
Interval Attack Analysis

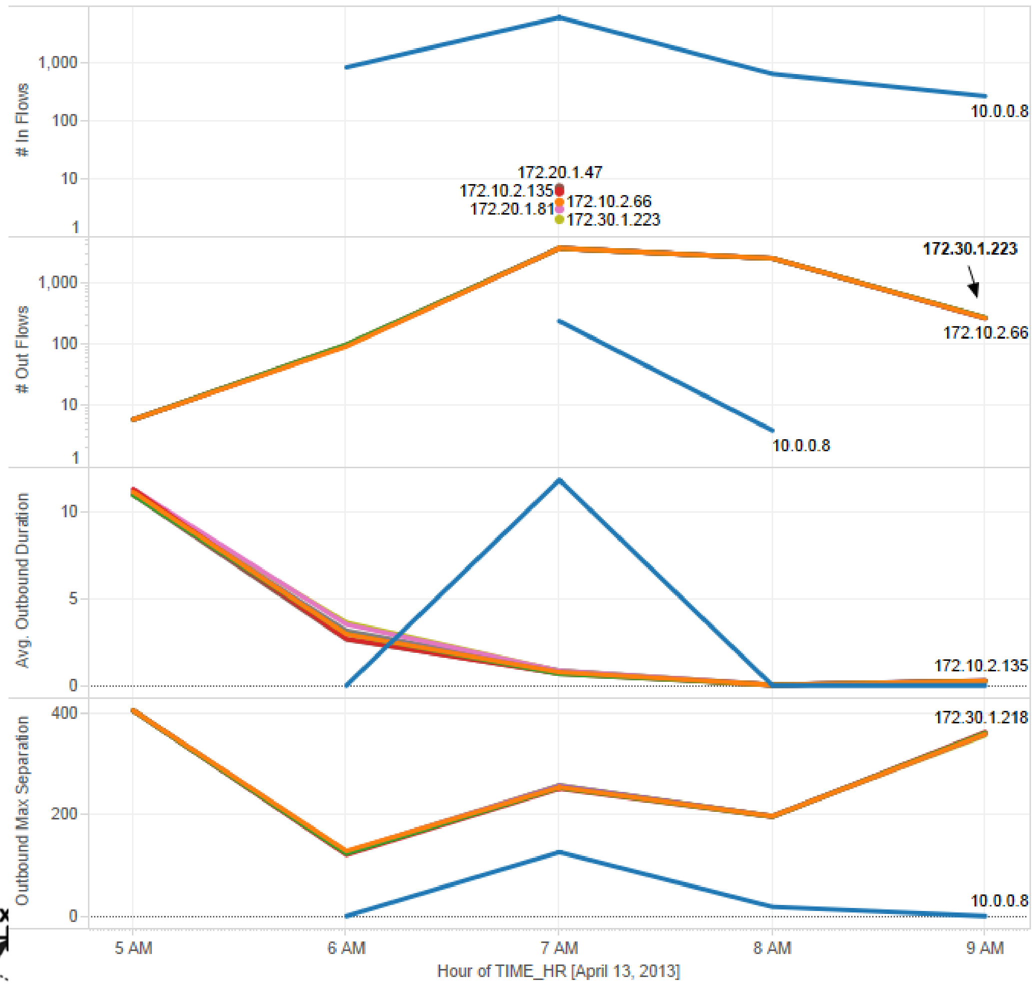
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

- **Attack:** Botnet DOS, workstations to external server
- Attacker synchrony
- Durations decrease in attack
- Separations also decrease
- Overall increase in synchrony







Thank you!

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

- Initial research effort with test data
- Transitioning certain capabilities to operational data
- Engaging multi-scale graph (logins)
- Porting to high performance graph database capability
- Eager to collaborate with community
  - Traffic analysis (Netflow)
  - Cyber graph analytics
  - Semantic graph databases

- `cliff.joslyn@pnnl.gov`

Joslyn, Cliff; Cowley, Wendy; Hogan, Emilie; and Olsen, Bryan: (2014) "Discrete Mathematical Approaches to Graph-Based Traffic Analysis", 2014 Int. Wshop. On Engineering Cyber Security and Resilience (ECSaR14), http://www.ase360.org/bitstream/handle/123456789/157/ecsar2014_paper4.pdf

Cliff Joslyn, Wendy Cowley, Emilie Hogan, Bryan Olsen: (2015) "Discrete Mathematical Approaches to Traffic Graph Analysis", Flocon 2015

Joslyn, CA; Choudhury, S; Haglin, D; Howe, B; Nickless, B; Olsen, B.: (2013) "Massive Scale Cyber Traffic Analysis: A Driver for Graph Database Research", *Proc. 1st Int. Wshop. on GRAph Data Management Experiences and Systems (GRADES 2013)*

BACKUP

Netflow Data Sizing

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

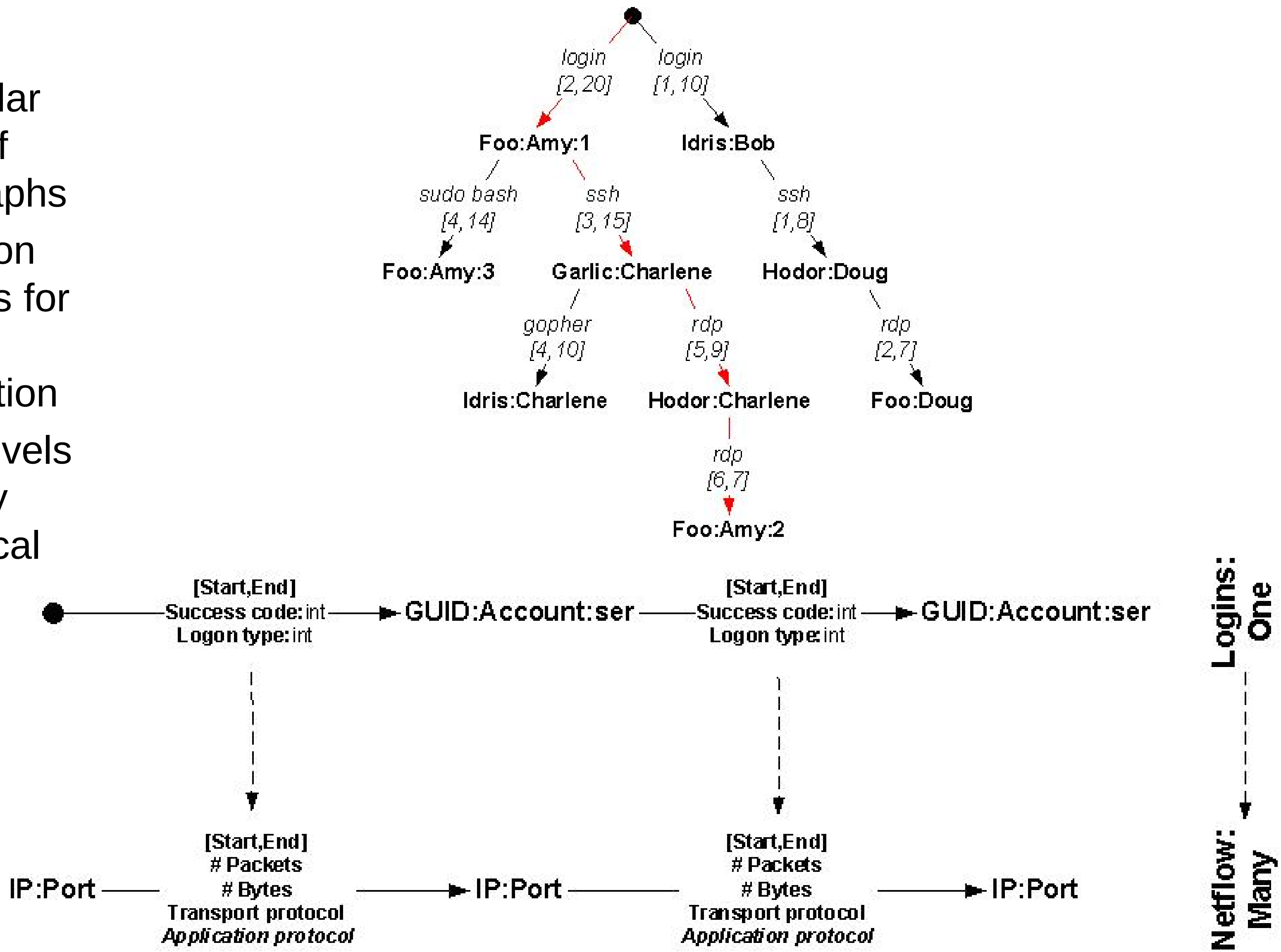
- Traffic analysis an essential big data problem
  - Direct acquisition from routers or reuse of publicly databases
  - Direct IPFLOW measurement or aggregation of packet capture
- Typical data rates from *one* typical PNNL network monitor:

	Average	**Stdev**
Flows/day (M)	613.2	242.5
Packets/day (B)	27.6	11.9
Bytes/day (T)	24.1	11.1
Packets/flow	178.7	702.6
Bytes/flow (K)	153.1	596.4

Multi-Scale
With Login Graphs from Event Logs

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Multi-scalar linkage of cyber graphs
- Information measures for feature identification
- Across levels to identify hierarchical scaling structure
- *Scale to massive graphs*

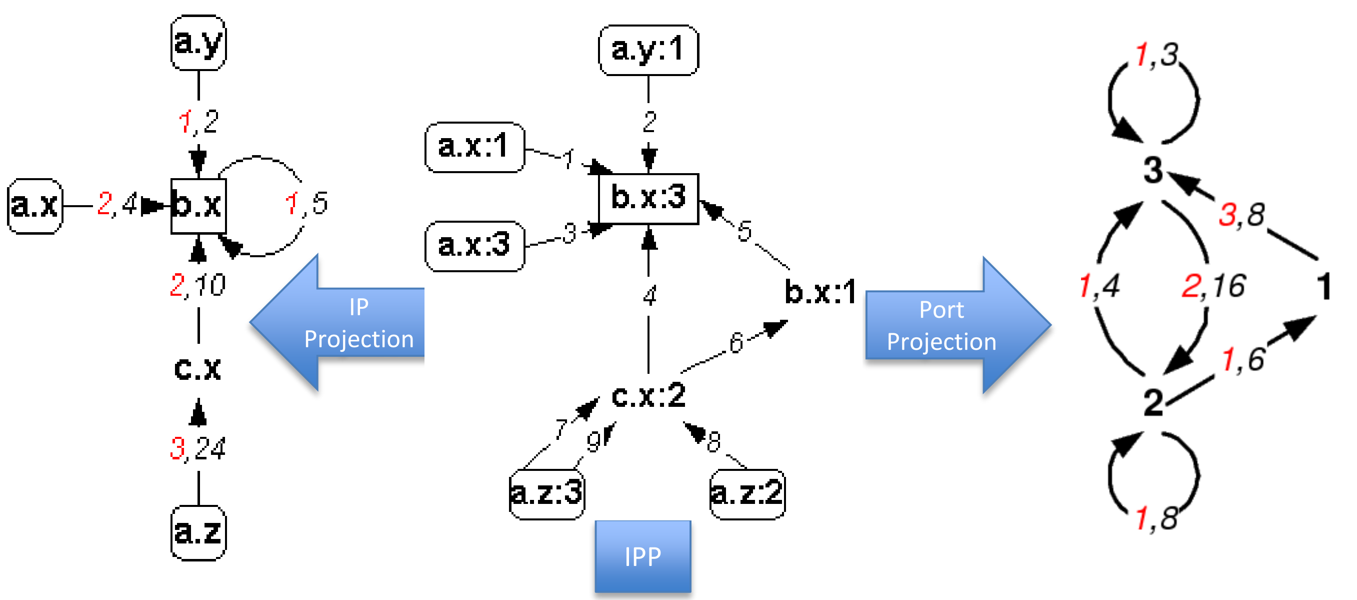


Basic Graph Statistics: Test

Test IP		Mean flows per
Flows	9	
Nodes	5	1.80
Outs	4	2.25
Leaves	1	20.0%
Ins	2	4.50
Roots	3	60.0%
Internals	1	20.0%
Pairs present	5	1.80
Pairs possible	8	1.13
Density	62.50%	
Mean Ports/IP	1.80	

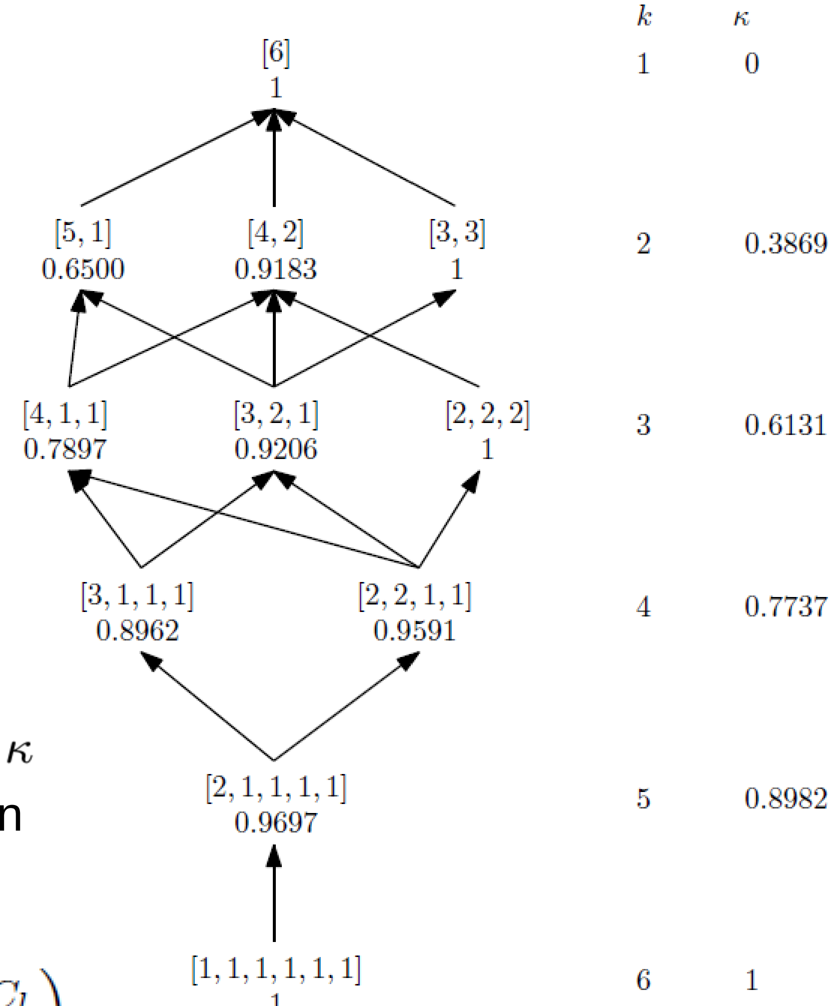
Test IPP		Mean flows per
Flows	9	
Nodes	8	1.13
Outs	7	1.29
Leaves	1	12.5%
Ins	3	3.00
Roots	5	62.5%
Internals	2	25.0%
Pairs present	8	1.13
Pairs possible	21	0.43
Density	38.10%	

Test Port		Mean flows per
Flows	9	
Nodes	3	3.00
Outs	3	3.00
Leaves	-	0.0%
Ins	3	3.00
Roots	-	0.0%
Internals	3	100.0%
Pairs present	6	1.50
Pairs possible	9	1.00
Density	66.67%	
Mean IPs/Port	2.67	



Measure Behavior

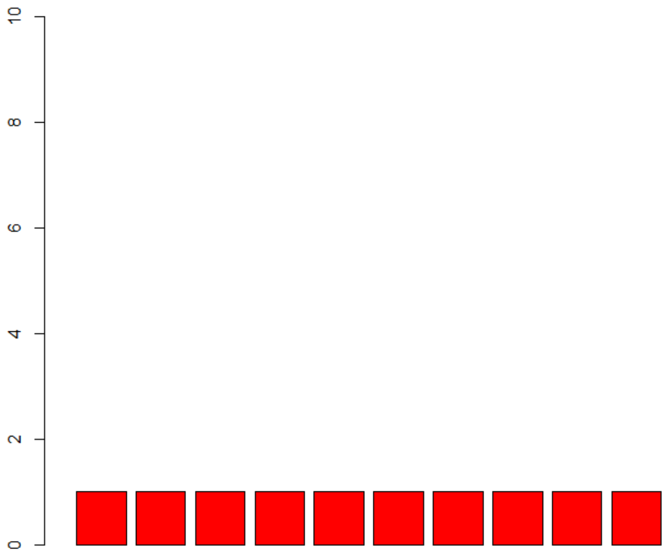
- Combinatorial measures on count distributions = integer partitions
- **Dispersion**
  - Normalized cardinality of support
  - In [0,1], varies with rank
- **Smoothness**
  - Entropy normalized over a *variable* support
  - In [0,1], increases within ranks
- Relatively independent "coordinates"
  - Consider $I = G \times \kappa = \frac{\mathbf{H}(f(\vec{C}))}{\log_2(N)} \leq G, \kappa$
  - For N >= 8, ranges of I of each rank can overlap



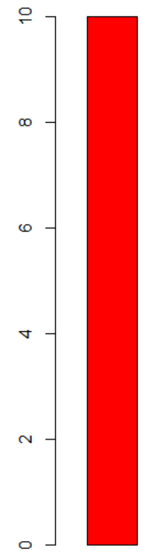
$$G(\vec{C}) := \frac{\mathbf{H}(f(\vec{C}))}{\log_2(m)} = \frac{-\sum_{l=1}^{m} \frac{C_l}{N} \log_2\left(\frac{C_l}{N}\right)}{\log_2(m)}$$

$$\kappa(\vec{C}) = \frac{\log_2(m)}{\log_2(N)}$$

Measure Behavior

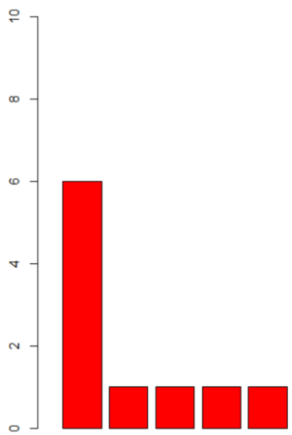


- *C=<1,1,1,1,1,1,1,1,1,1>, m = 10*
- Maximal dispersion: \kappa = 1
- Maximal smoothness: G = 1

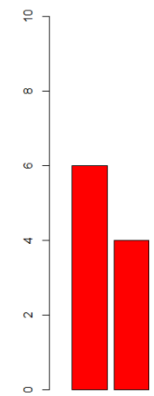


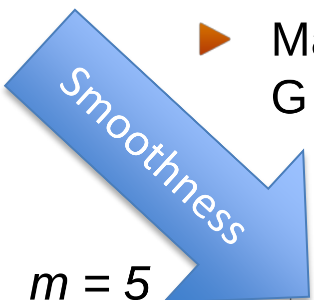
- *C=<10>, m = 1*
- Minimal dispersion: \kappa = 0
- Minimal smoothness: G = 0

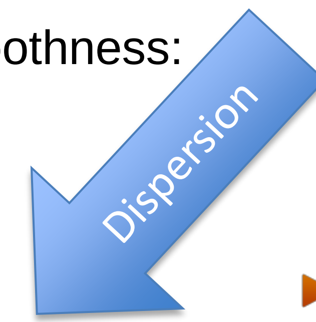
Measure Behavior



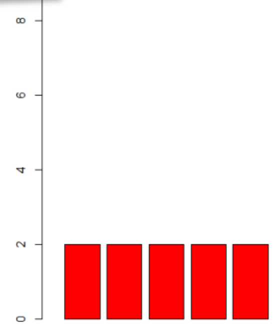
- *C=<2,2,2,2,2>, m = 5*
- Moderate dispersion: \kappa = 0.70
- Maximal smoothness: G = 1.00







- *C=<6,1,1,1,1>, m = 5*
- Moderate dispersion: \kappa = 0.70
- “Low” smoothness: G = 0.76



- *C=<6,4>, m = 2*
- Low dispersion: \kappa = 0.30
- High smoothness: G = 0.97