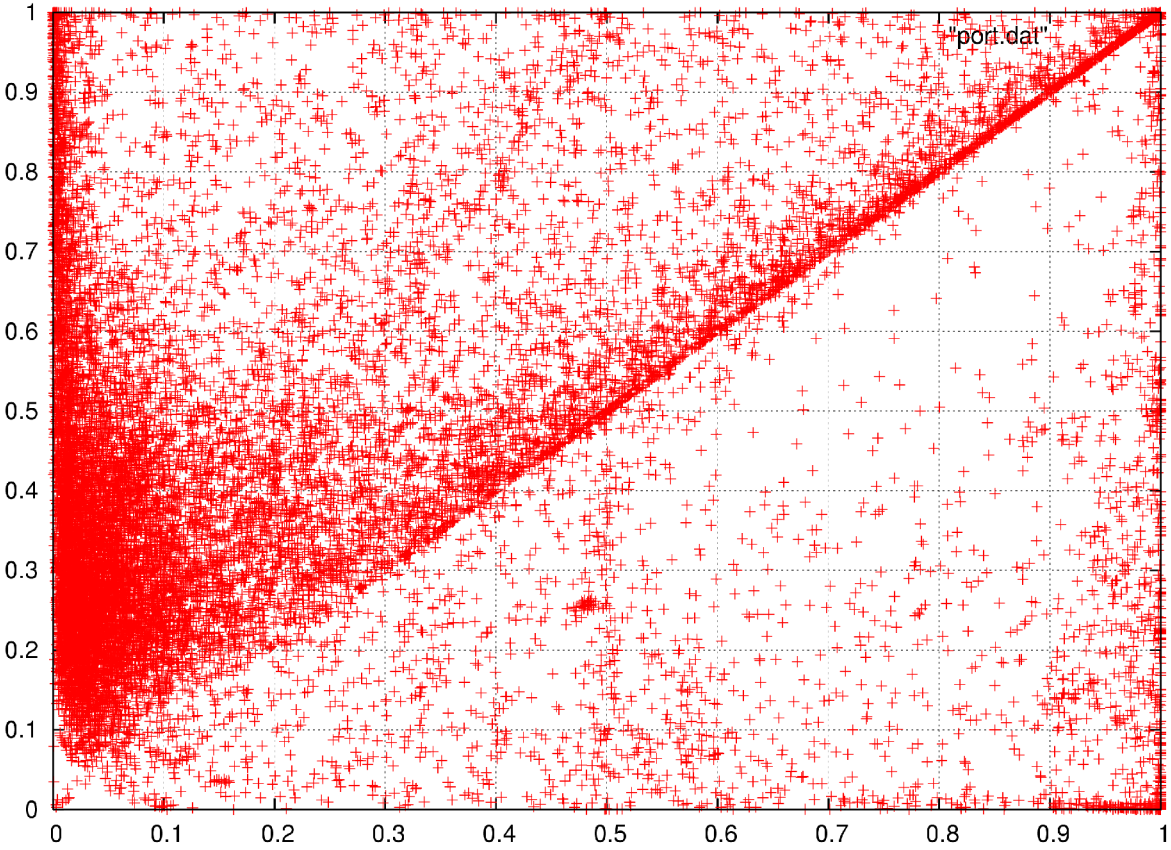
Network Host Classification Using Statistical Analysis of Flow Data

Alex Kent, Mike Fisk, Eugene Gavrilov
Los Alamos National Laboratory

Overview and Objectives

- Host/IP address profiling based on flow data over some time interval
  - 10 minutes to 7 days have been examined with 24 hours providing repetitively stationary results
  - Generate histograms of peer hosts, source ports, and destination ports over the time interval
  - Compute Shannon entropy values for the 3 dimensions
- Outcomes:
  - Provide IP behavior “snapshots” of individual hosts
  - Allow comparison of behavior through clustering
  - Build models over large host sets in real-time

Source / Destination Port Variance



Does not provide an effective representation of categorical data sets

Sample (simplified) Histogram of Flow Data

		Cumulative bytes	Cumulative packets	Cumulative sessions
Host A				
	Peer A	34958	324	54
	Peer B	3948	132	13
	Peer C	231	43	9
	Peer D	5675	123	29
	Src Port 1	2358	77	32
	Src Port 2	13246	345	67
	Src Port 3	1231	75	12
	Dst Port 1	54467	5653	199
	Dst Port 2	563	345	1
Host B	Peer X	842	347	23
	Peer Y	23879	3452	874
	Peer Z	9463	232	78
	...	...	...	...
...				

Shannon Entropy
Using Packet Count Histograms

$$Entropy = -1 \times \sum_{i=1}^{N} p_i \times log_2(p_i)$$

$$\text{where} \quad p_i = \frac{Packets_i}{TotalPackets}$$

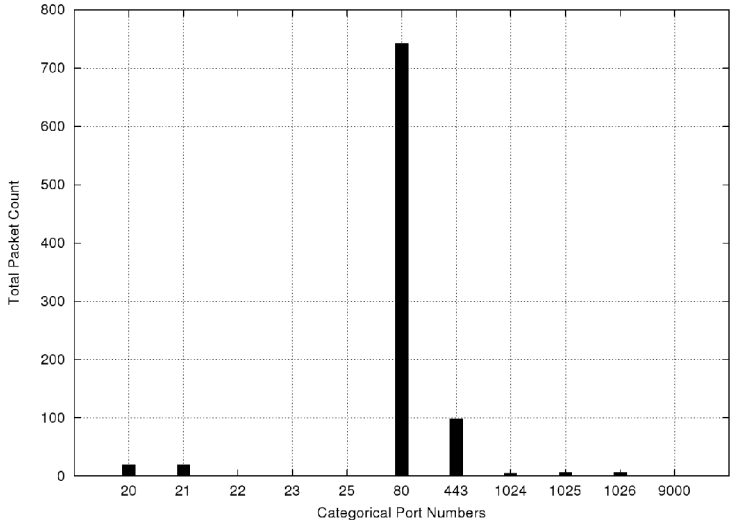
- Computed for host peers, source ports, and destination ports time-delineated histograms leveraging byte, packet, and session totals
  - Packet histograms/entropy calculation favored in final analysis
  - Since base 2: port entropy will be 0-16, peer 0-32 (IPv4)

Sample Entropy Calculation

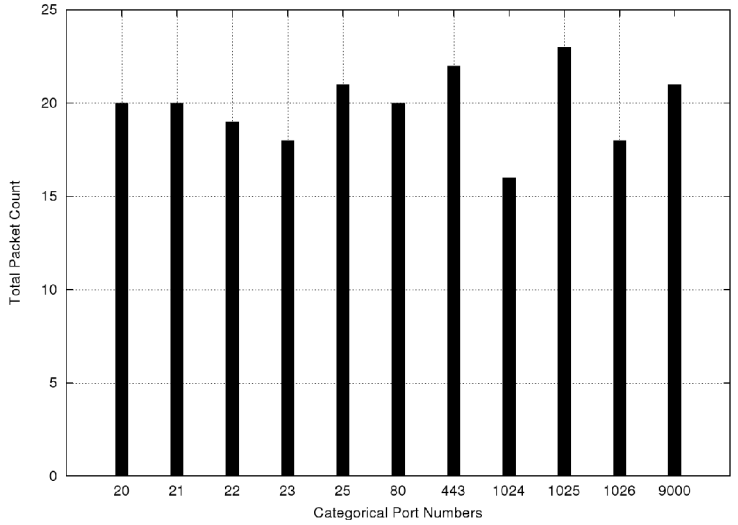
For Host A:

	Cumulative packets	*Pi* Packets	*Pi* * log2(*Pi*)	Entropy
Peer A	324	0.52	-0.49	
Peer B	132	0.21	-0.47	
Peer C	43	0.07	-0.27	
Peer D	123	0.20	-0.46	1.69
Src Port 1	77	0.15	-0.42	
Src Port 2	345	0.69	-0.37	
Src Port 3	75	0.15	-0.41	1.19
Dst Port 1	5653	0.94	-0.08	
Dst Port 2	345	0.06	-0.24	0.32

Visual Example of Low/High Entropy



Low Entropy Example

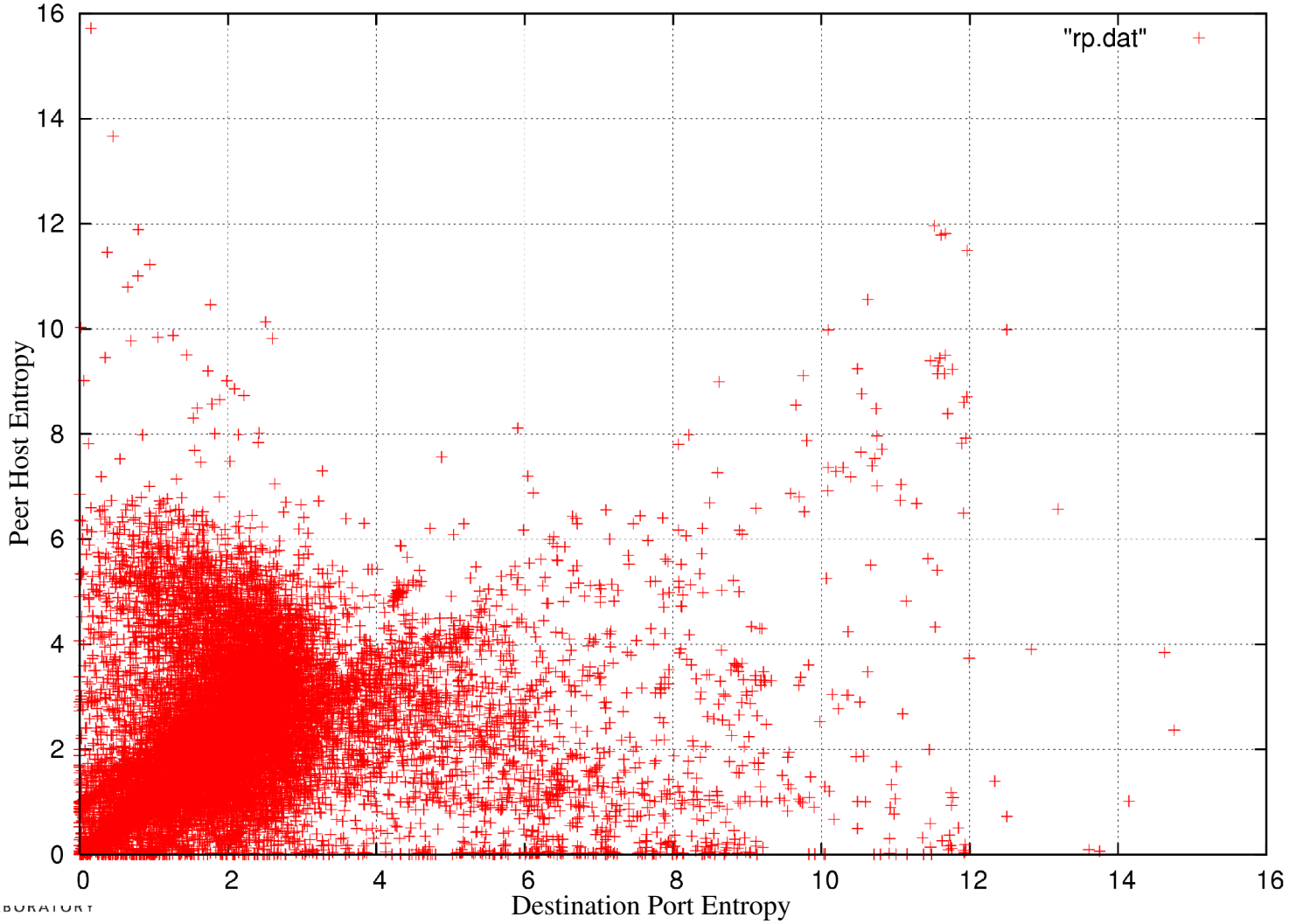


High Entropy Example

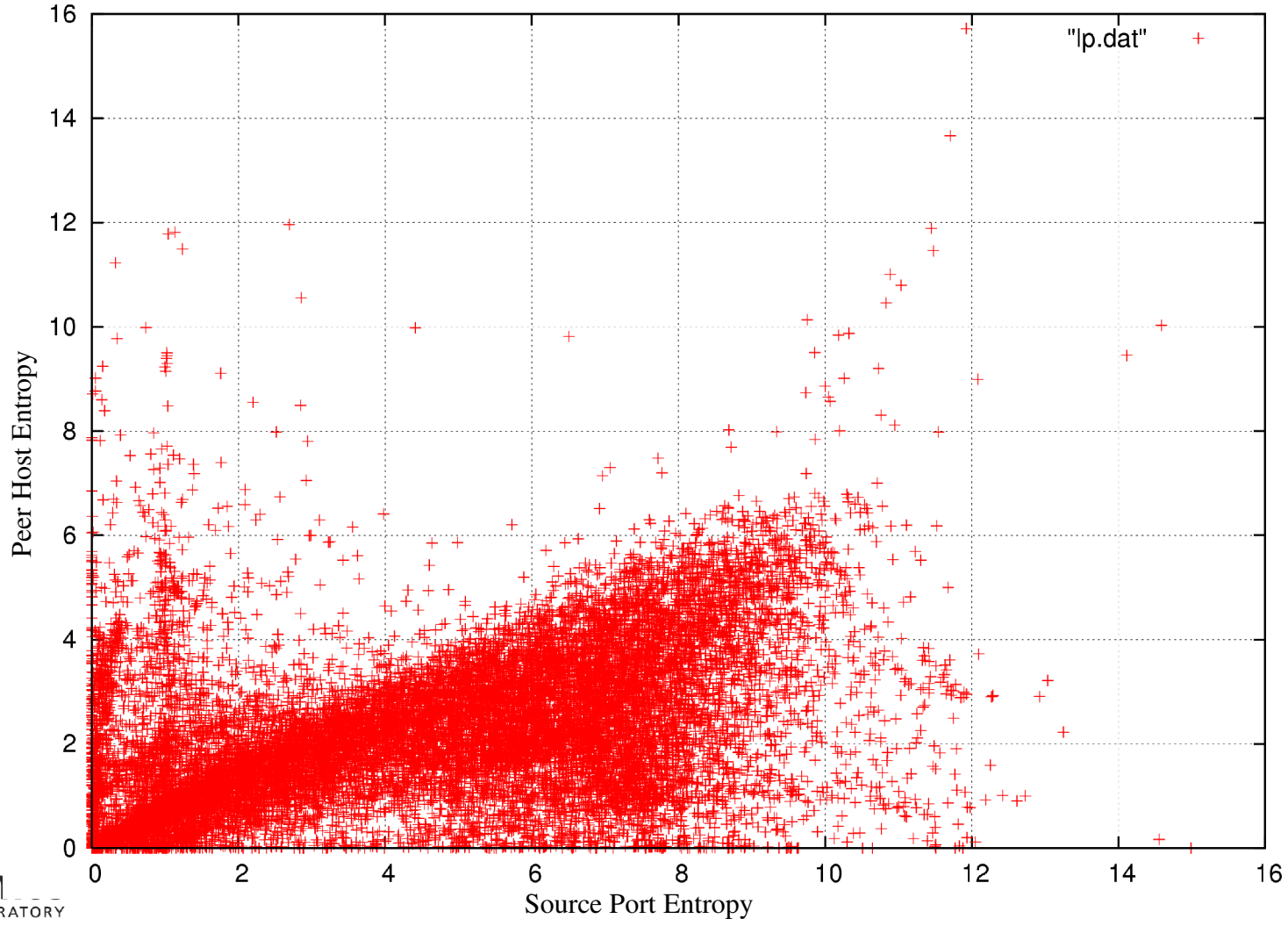
Data Overview

- Uses generic flow data
  - Required fields:
    - SRC IP, DST IP, SRC Port, DST Port, Protocol, Packets, Bytes
- Los Alamos unclassified network primarily over 24 hours (inclusive of a work day)
  - Approximately 200 million flows analyzed
  - 17,326 unique internal (Los Alamos) IP's observed
  - Day-to-day traffic very consistent

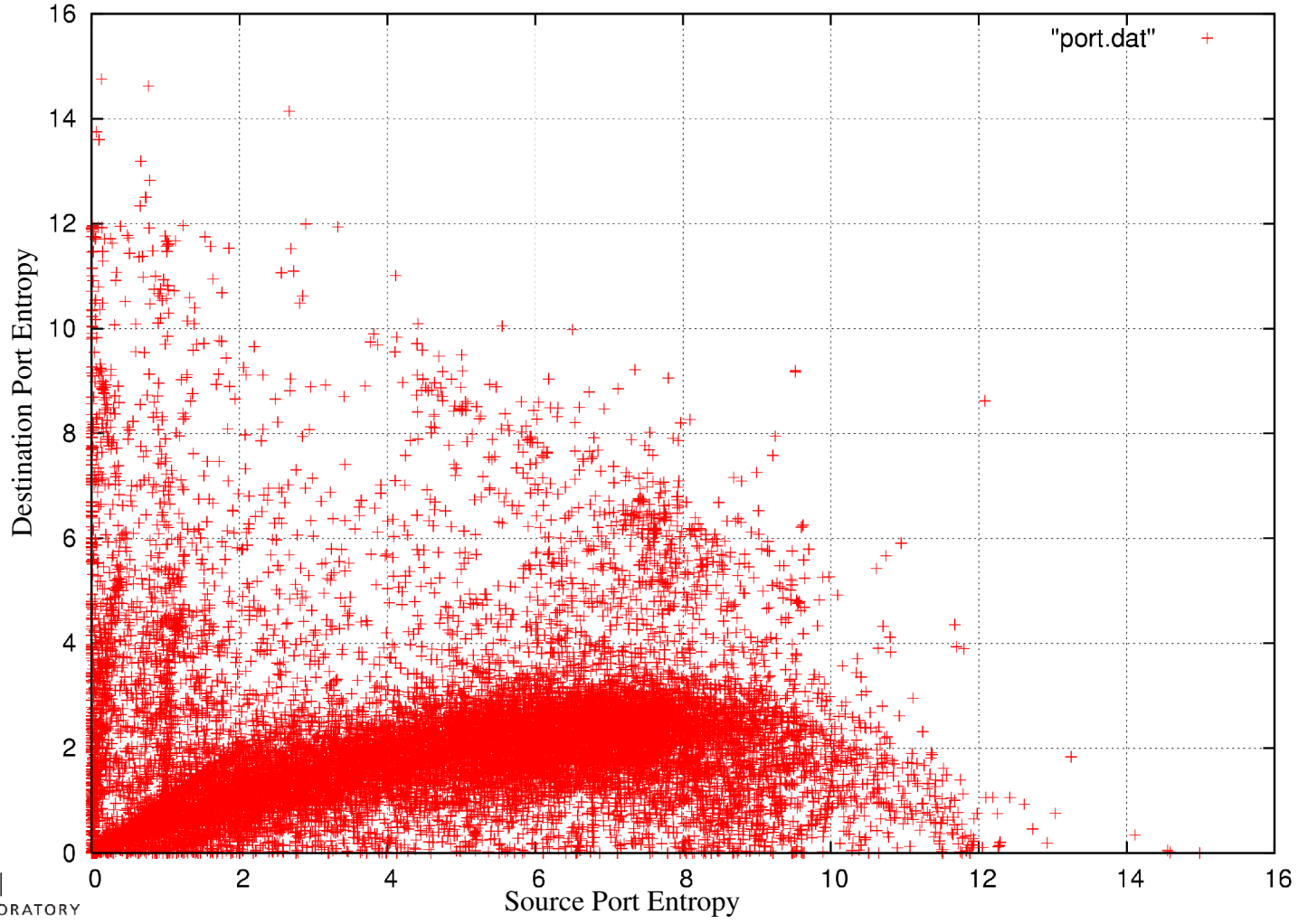
Destination Port / Peer Entropy



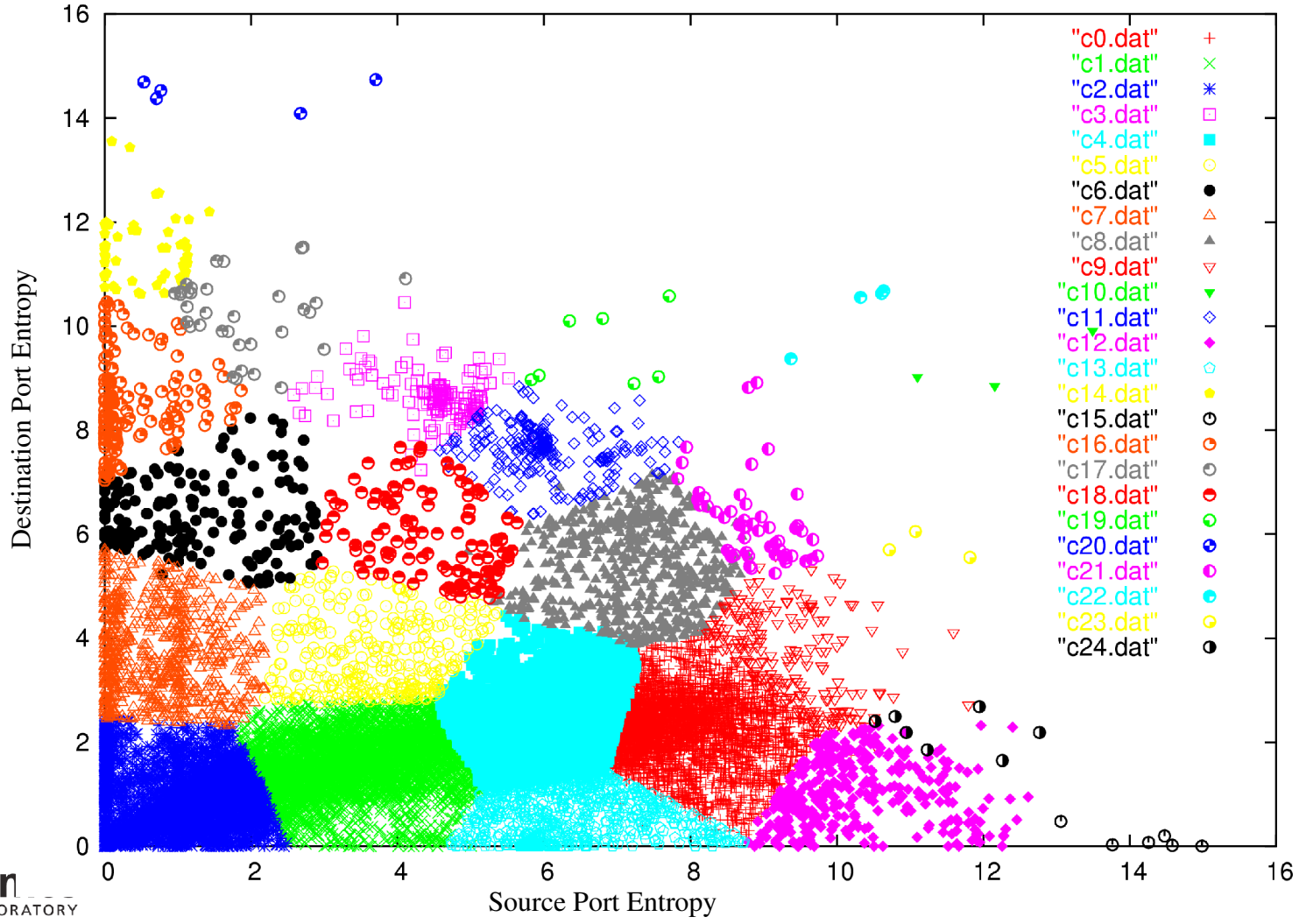
Source Port / Peer Entropy



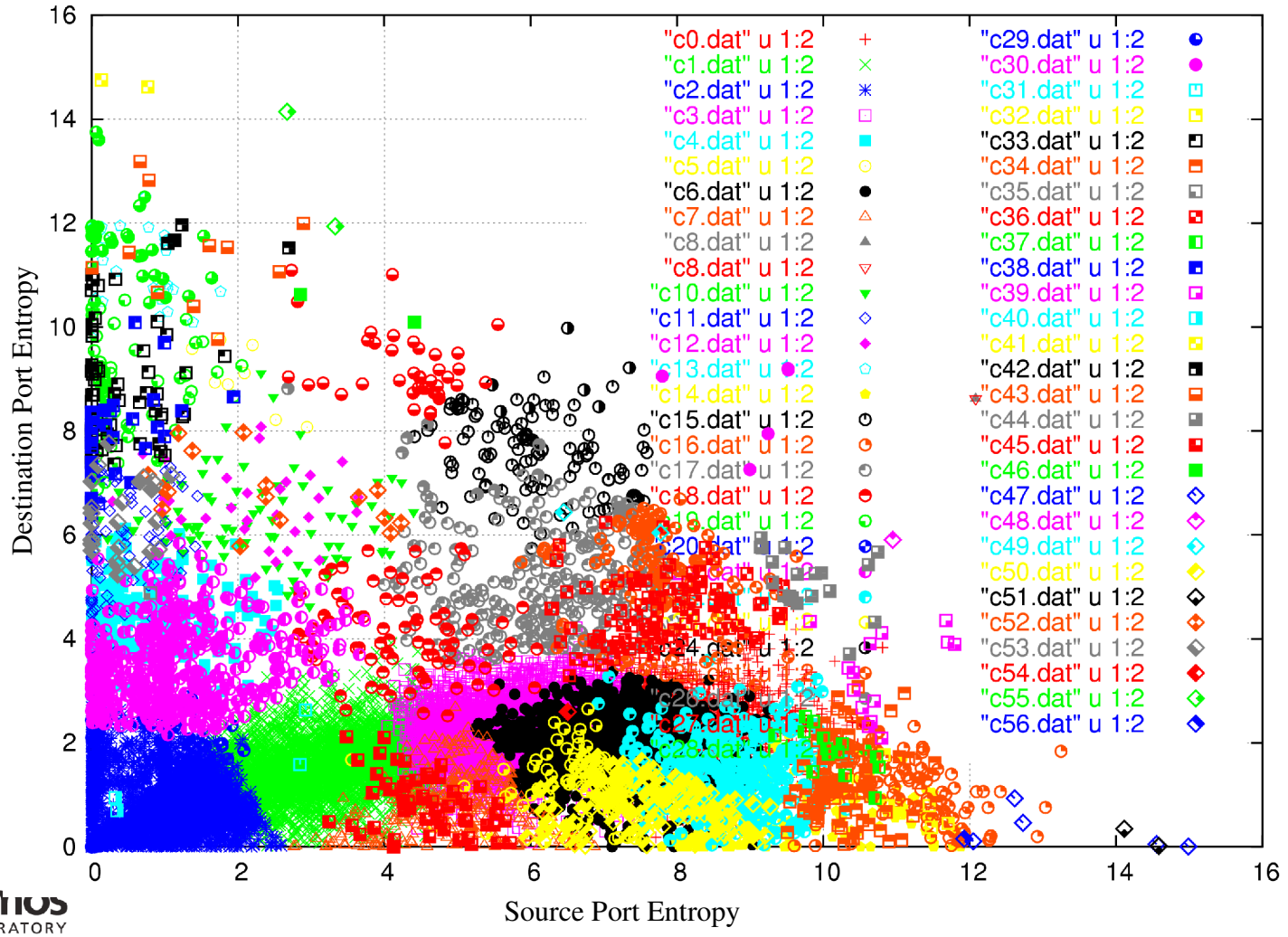
Source Port / Destination Port Entropy



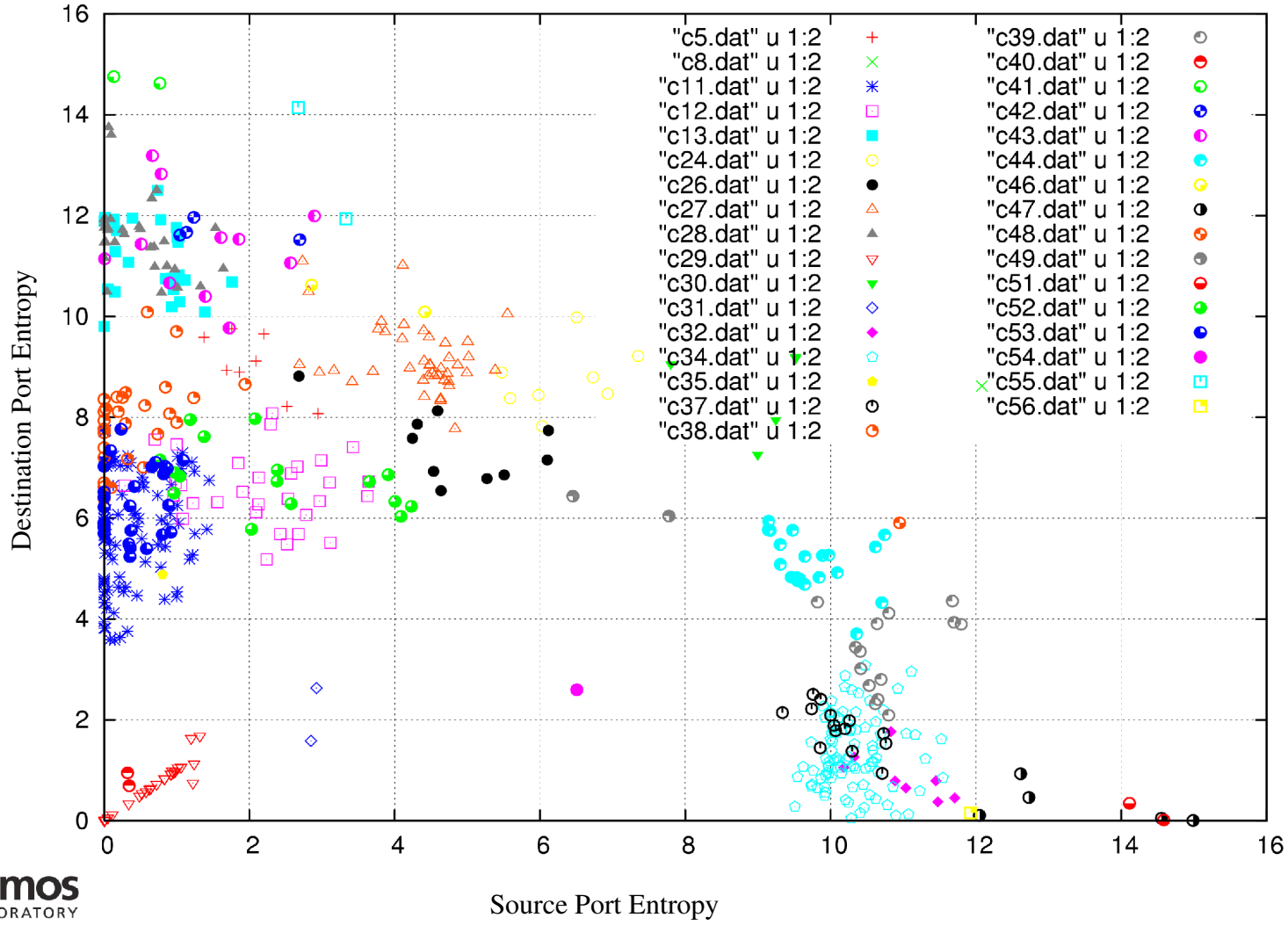
Source/Destination Port Entropy Clustering



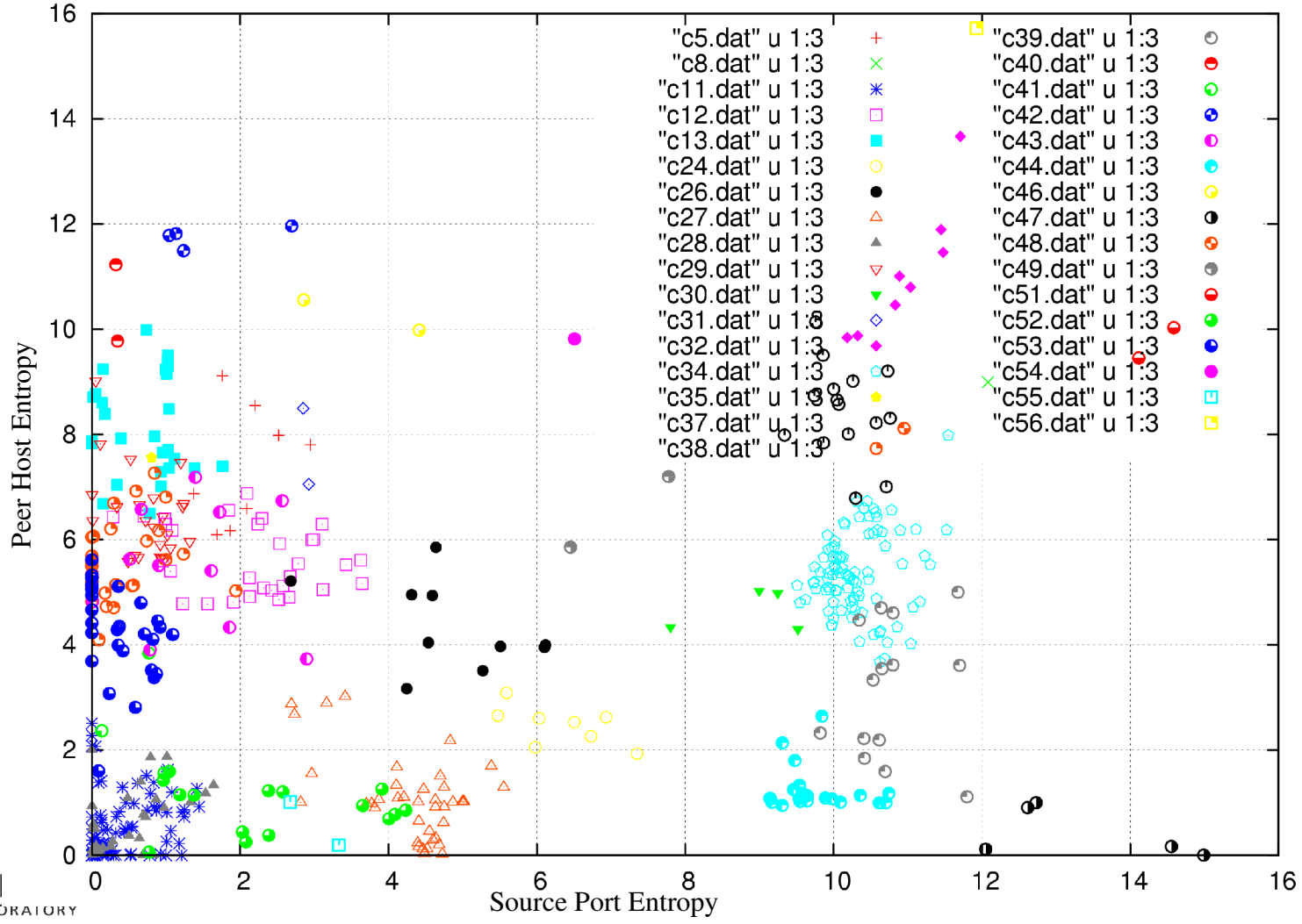
Entropy Clustering w/ 3 Dimensions



Interesting Clusters (<50)
749 Hosts (4.3% of total)



Interesting Clusters (b)
Source port versus Peers

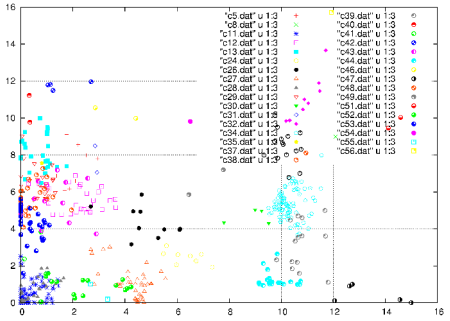


Major Servers

Host	Local Port Ent	Remote Port Ent	Peer Ent	Cluster
SMS 1	1.04	11.61	11.78	42
SMS 2	1.13	11.67	11.82	42
Int WWW	1.24	11.97	11.49	42
ActiveDir 1	2.7	11.52	11.96	42
ActiveDir 2	4.41	10.09	9.98	46
ActiveDir 3	2.86	10.62	10.56	46
DNS 1	1.76	9.76	9.11	5
DNS 2	0.74	12.5	9.99	13
VulnScanner	12.08	8.62	8.99	8
MailRelay 1	7.4	5.12	4.31	36
MailRelay 2	7.42	5.05	4.29	36
MailRelay 3	7.58	5.13	4.45	36

Clusters C32 & C56
Bad Behavior (worm variants)

Host IP	Local Port Ent	Remote Port Ent	Peer Ent
Remote Host A	11.45	0.79	11.89
Remote Host B	10.18	1.06	9.84
Internal Host A	10.32	1.26	9.87
Internal Host B	10.89	0.79	11.01
Remote Host C	10.83	1.77	10.46
Remote Host D	11.71	0.45	13.67
Remote Host E	11.04	0.65	10.8
Remote Host F	11.93	0.15	15.72



Current, On-going Work

- Demonstrated 1 million+ flows/minute processing on single system
  - Redesigning, porting system to map/reduce architecture for improved scaling and distributed processing
- Integrating additional network flow data types (e.g. custom perimeter collected flows)
- Static centroids for comparing host movements between k-means clusters
  - Enable predefined clusters, cluster definitions, and host movement between clusters
- Histogram merging that allows graceful data aging for continuous data feed and anomaly detection
- Application of novel change detection and machine learning across time series output