Network Awareness and Network Security

John McHugh
Canada Research Chair in Privacy and Security
Director, Privacy and Security Laboratory
12 October 2005

Overview

The internet has grown to the point where observational approaches offer one of the only approaches to its understanding.

- In this case, we have a view of the border of a large composite network with > /8 address space and several million hosts.
- Given a broad view, it is possible to look at both similarities and differences in the traffic going to/from various sites.
- In this talk, we will try to sketch a variety of analyses that can be performed when such data is available

Note: Similar data can be aggregated in a variety of ways. Opportunities to obtain similar views exist.

Network Data Collection Approach

Look at flow abstractions for a large customer network.

- No payload data just headers – Source, Destination IP and ports; protocol; times; traffic volumes (e.g., packets and bytes) - Cisco NetFlow like sources

Comprehensive coverage

- >95% coverage of the customer network
- Multiple networks, at the minimum

Collect a lot of data

- Requires a data center with large computational and storage capacity for historical analysis
- Scalable collection and analysis

Analysis Approach

Netflow Data

- Organized by hour, type, sensor (router), etc. for outside to inside and inside to outside.
- Packed format for efficient linear search
- 10s of TB and growing by 10s of GB/day

Primary operation is selection based on time, IP, flow characteristics (protocol, volume, etc.)

- Creates files of raw data satisfying selection criteria
- Statistics on files can be produced

Can create sets or bags (multisets, counted sets) of IPs with selection characteristics

- Sets can be used for further selection / filtering See ESORICS paper next week

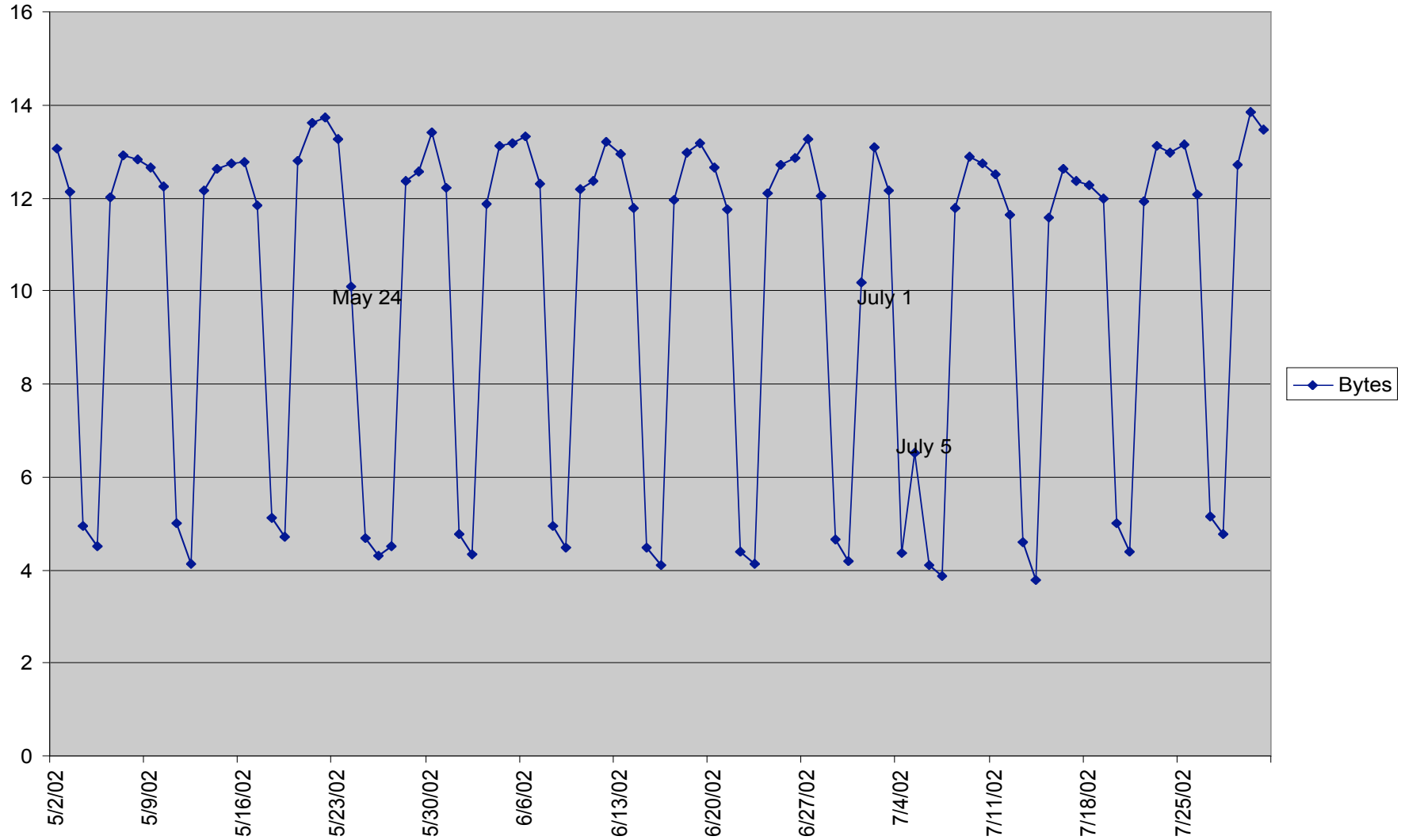
A few quick examples

The next few slides show some typical samples of the kinds of information that can be derived from the flow data.

They range from network wide analyses to examinations of the characteristics of specific subnets and even of specific hosts.

- The analysis system can be viewed as a powerful zoom lens.
- It is capable at looking at the overall traffic on a few percent of the internet at its widest angle, or at a single host at its highest magnification.

Protocol 6: TCP Routed Data (Bytes)

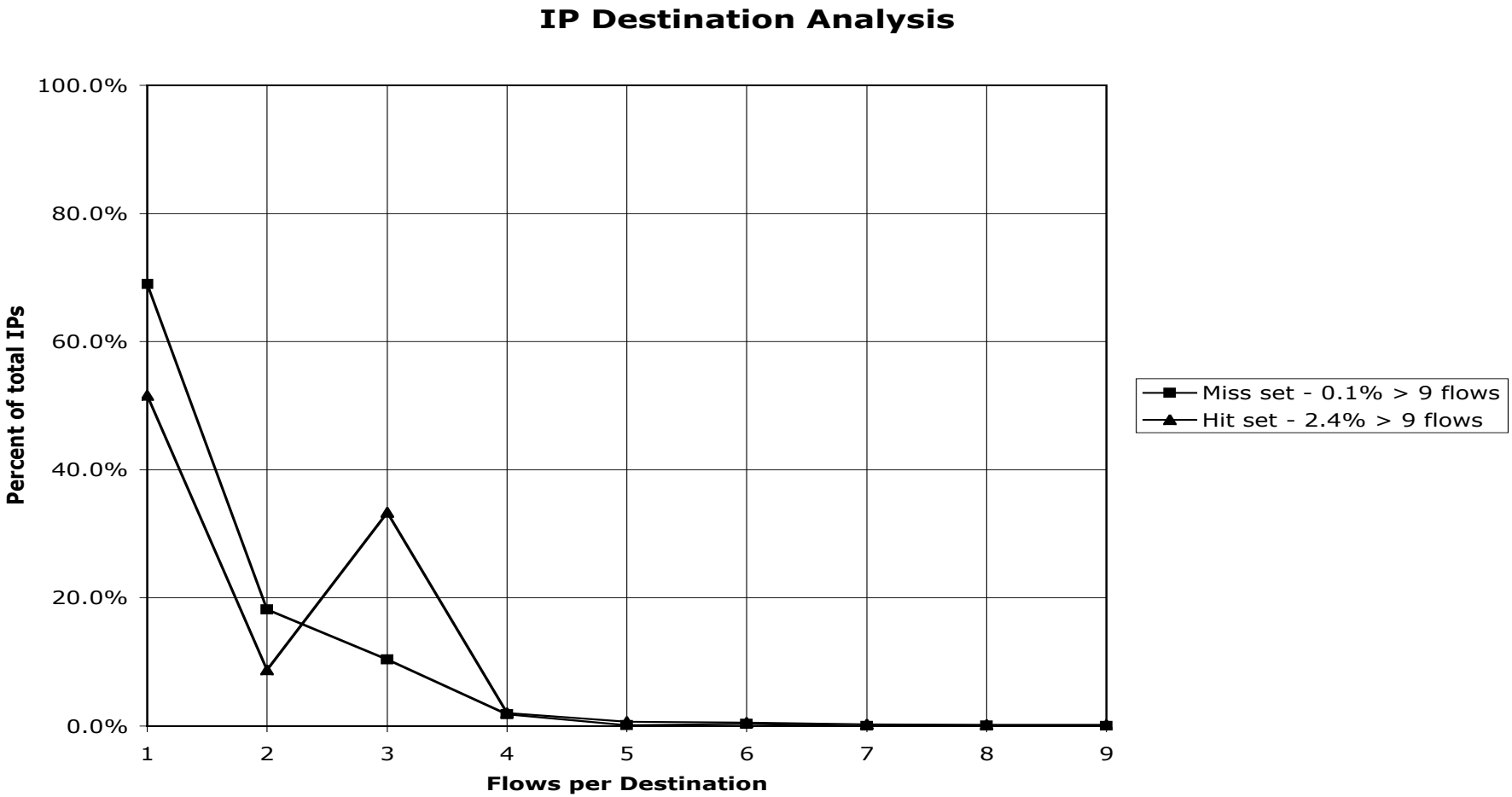


Characteristics of small sample

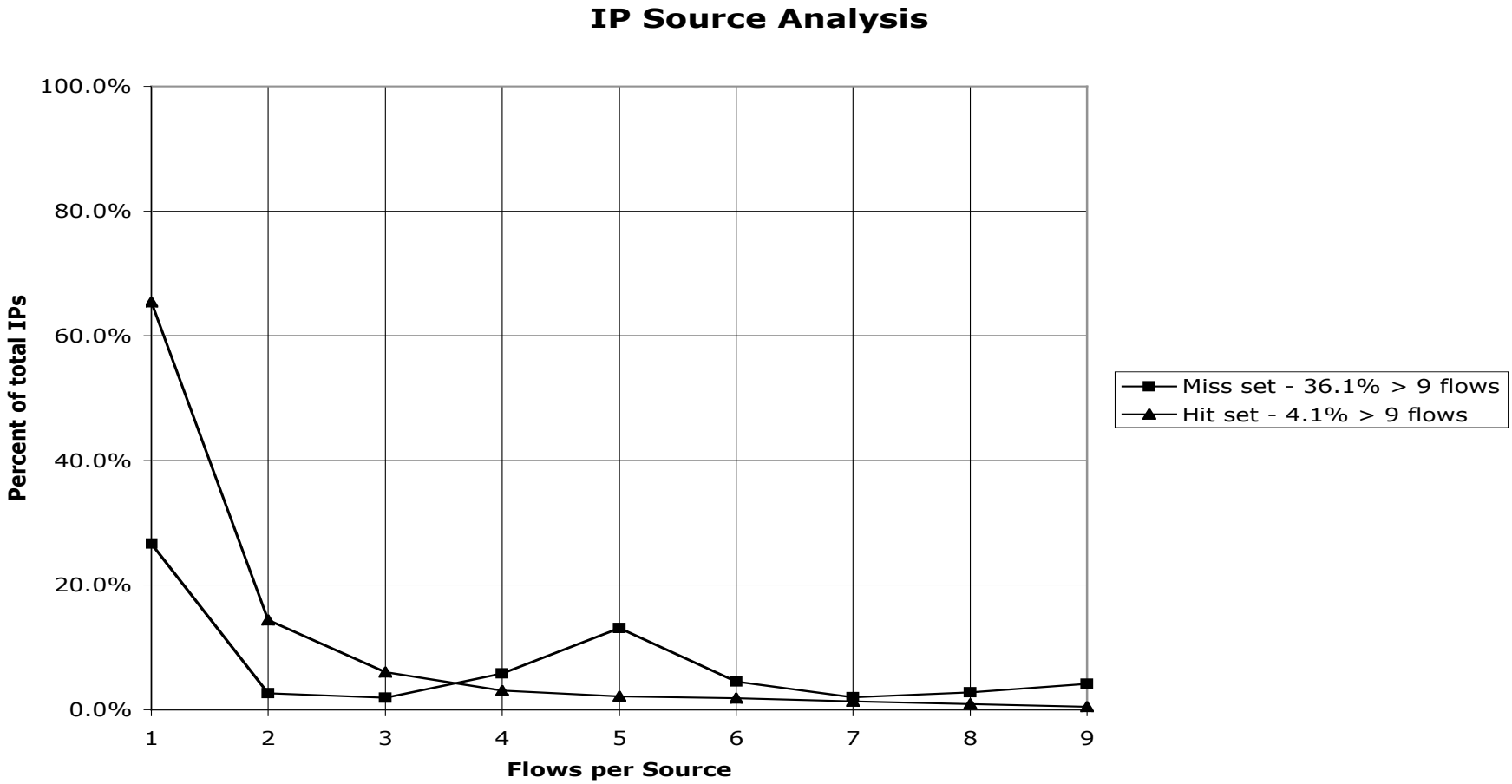
We looked at 1 minute of data from the monitored network

- Used set of active inside IPs to partition out into
  - Hit data is addressed to “active” hosts
  - Miss data is addressed to “inactive” hosts
- Partitions have very different characteristics
  - Miss data appears to be mix of scans and noise

1 Min sample - destinations



1 Min Sample - sources



top 5 in 1 min sample

```
(39) lip $ readbag --count --print jcm-tcp-s-
   10+.bag|\                              sort -r -n | head
```

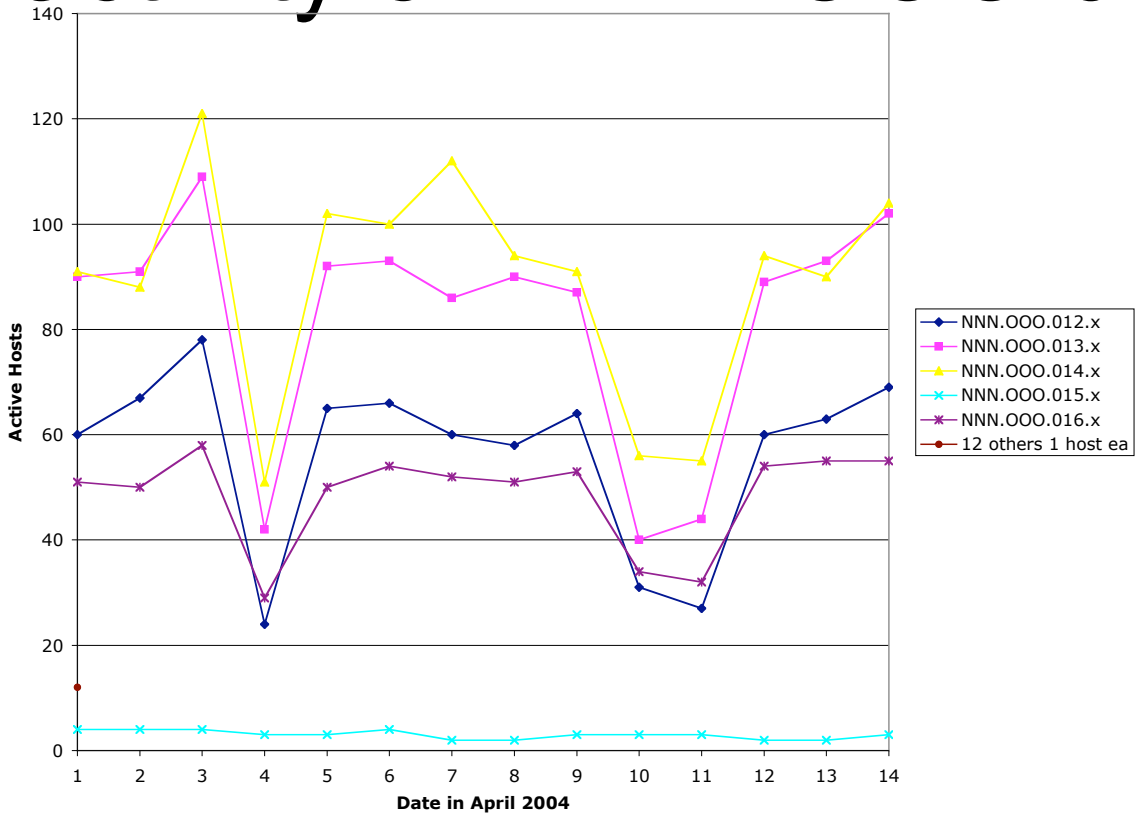
```
12994  AAA.BBB.068.218 - scan 4899 (Radmin)
 6598  CCC.DDD.209.215 - scan 7100 (X-Font)
 5944  EEE.FFF.125.117 - scan 20168 (Lovegate)
 5465  GGG.HHH.114.052 -  ditto
 5303  III.JJJ.164.126 - scan 3127 (My doom)
```

Bottom of bag in 1 min sample

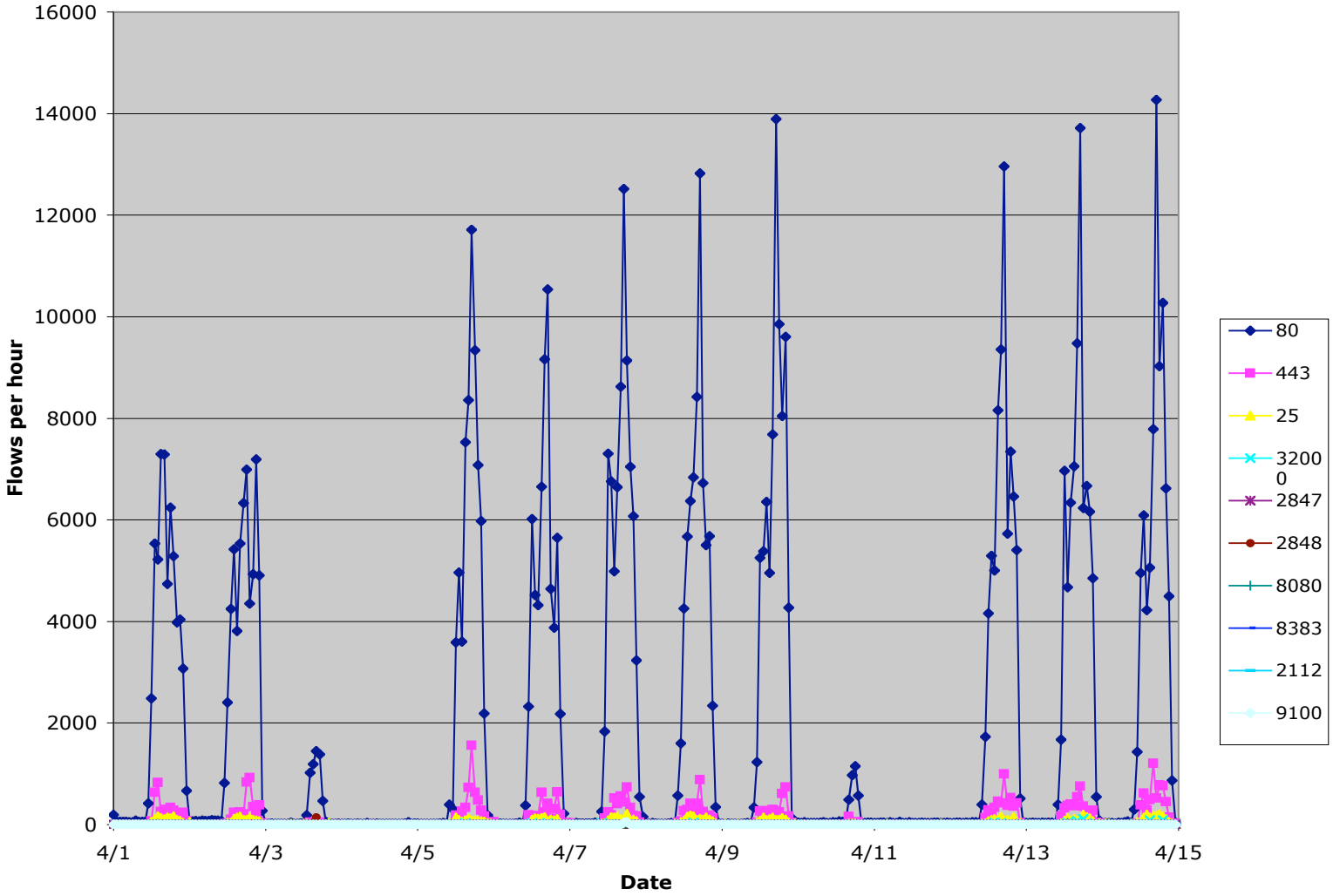
3335 external hosts sent exactly one TCP flow

- SYN probes for port 8866 449 times
  - W32.Beagle.B@mm is a mass-mailing worm-back door on TCP port 8866.
- SYN probes for port 25 are seen 271 times.
- Most remainder are SYNs to a variety of ports, mostly with high port numbers.
- There are a number of ACK/RST packets which are probably associated with responses to spoofed DDoS attacks.

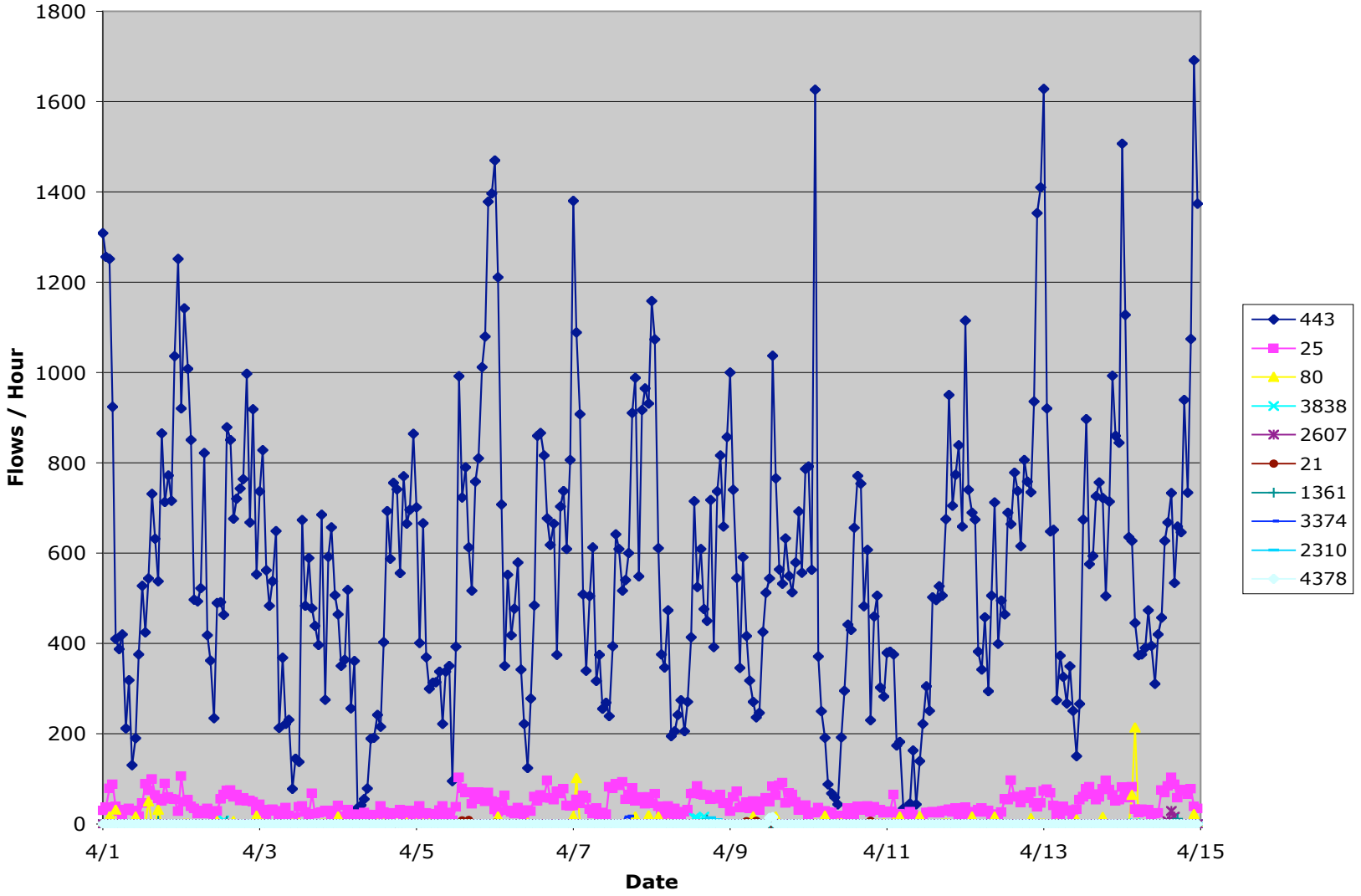
Host activity on NNN.OOO.0.0/16



NNN.OOO in to out TCP dst



NNN.OOO in to out TCP src



What does this mean?

This installation operates 5/8 or 6/8 for the most part.

It is sparsely populated

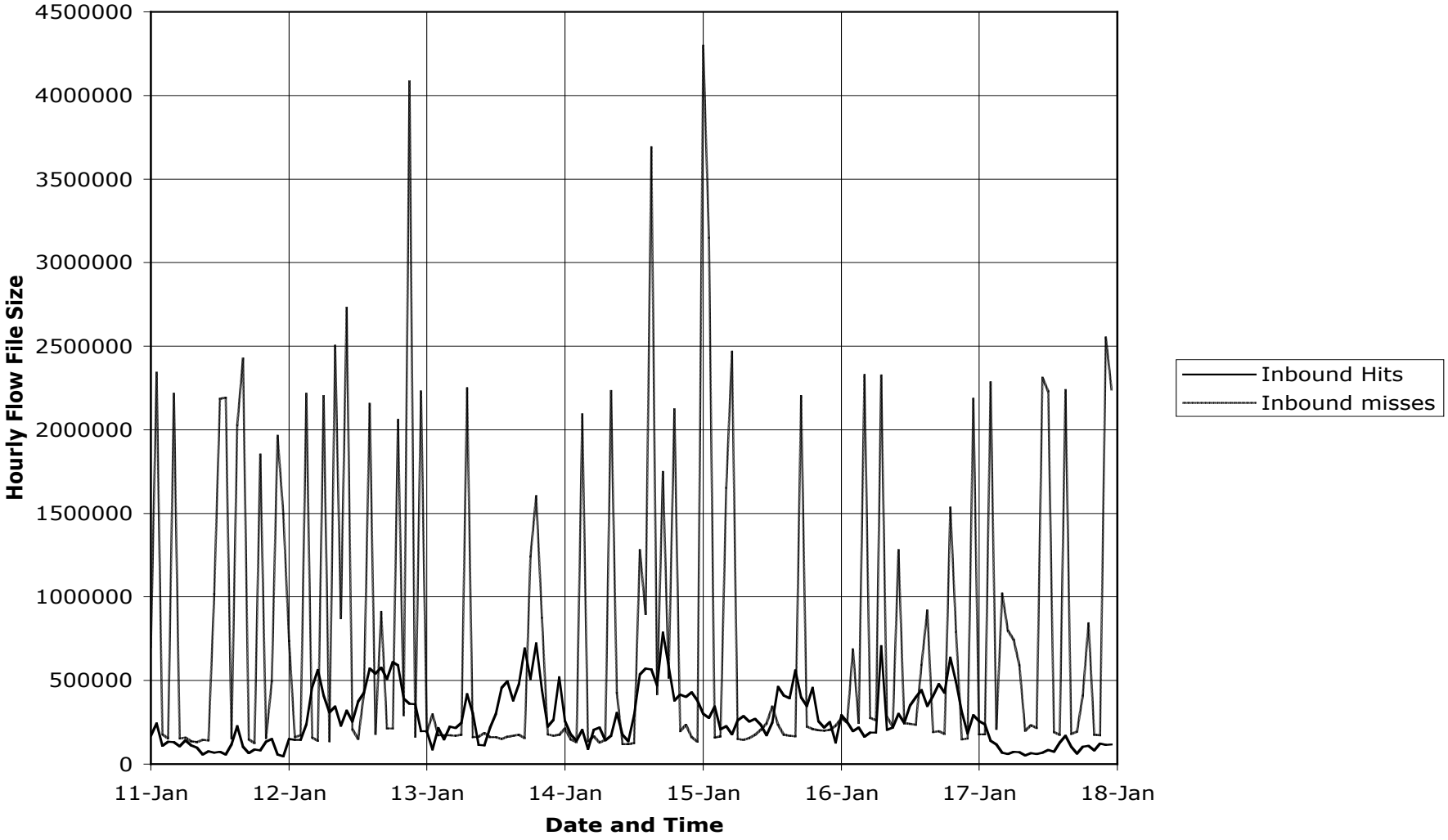
- A handfull of /24s
- 25% to 50% populated

Outgoing traffic typical of workstations in sport

Outgoing traffic typical of https server in dport

Interesting mix of minority ports, but numbers very small after first few.

One week on another /16

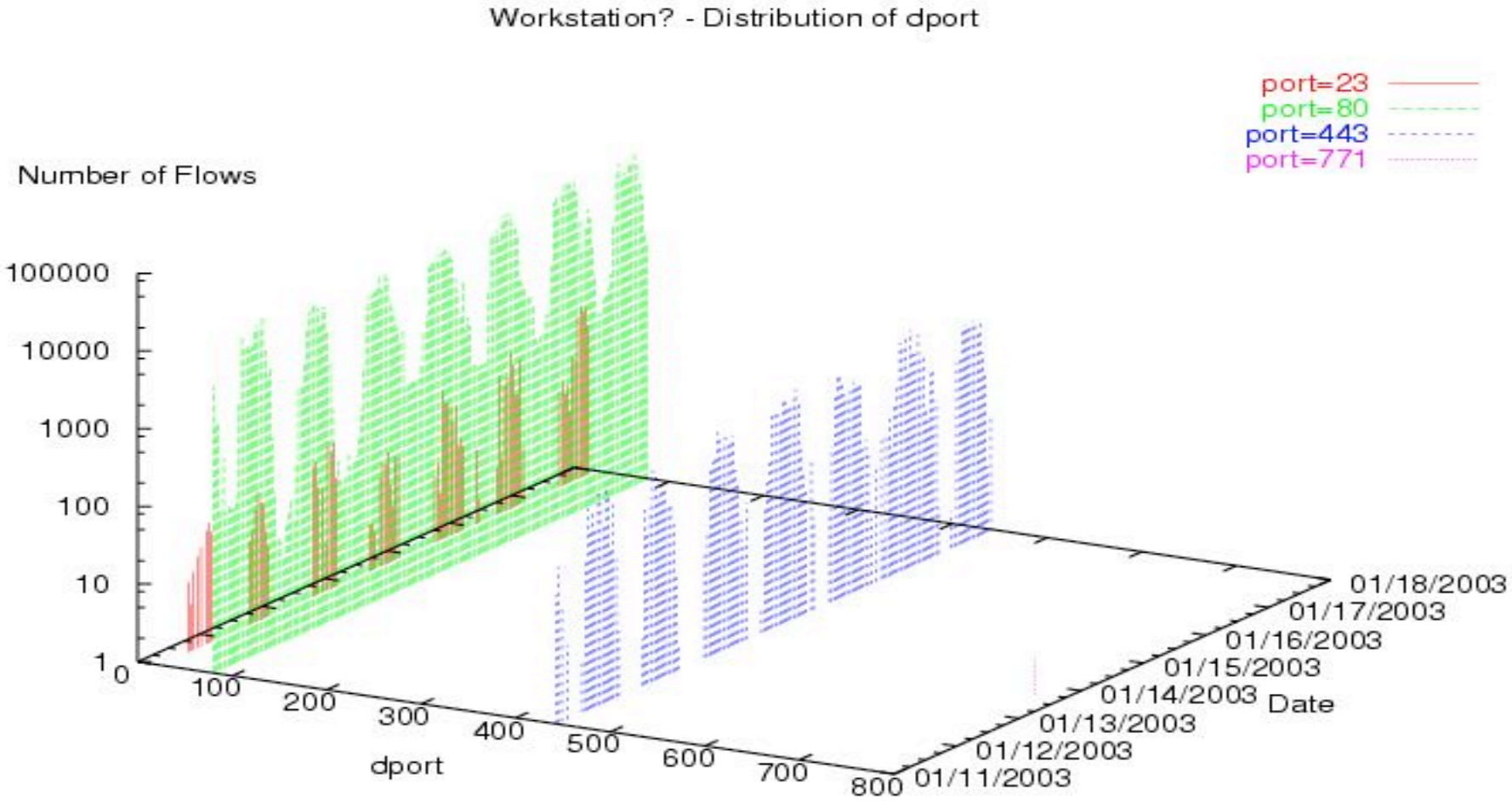


Host Characterization

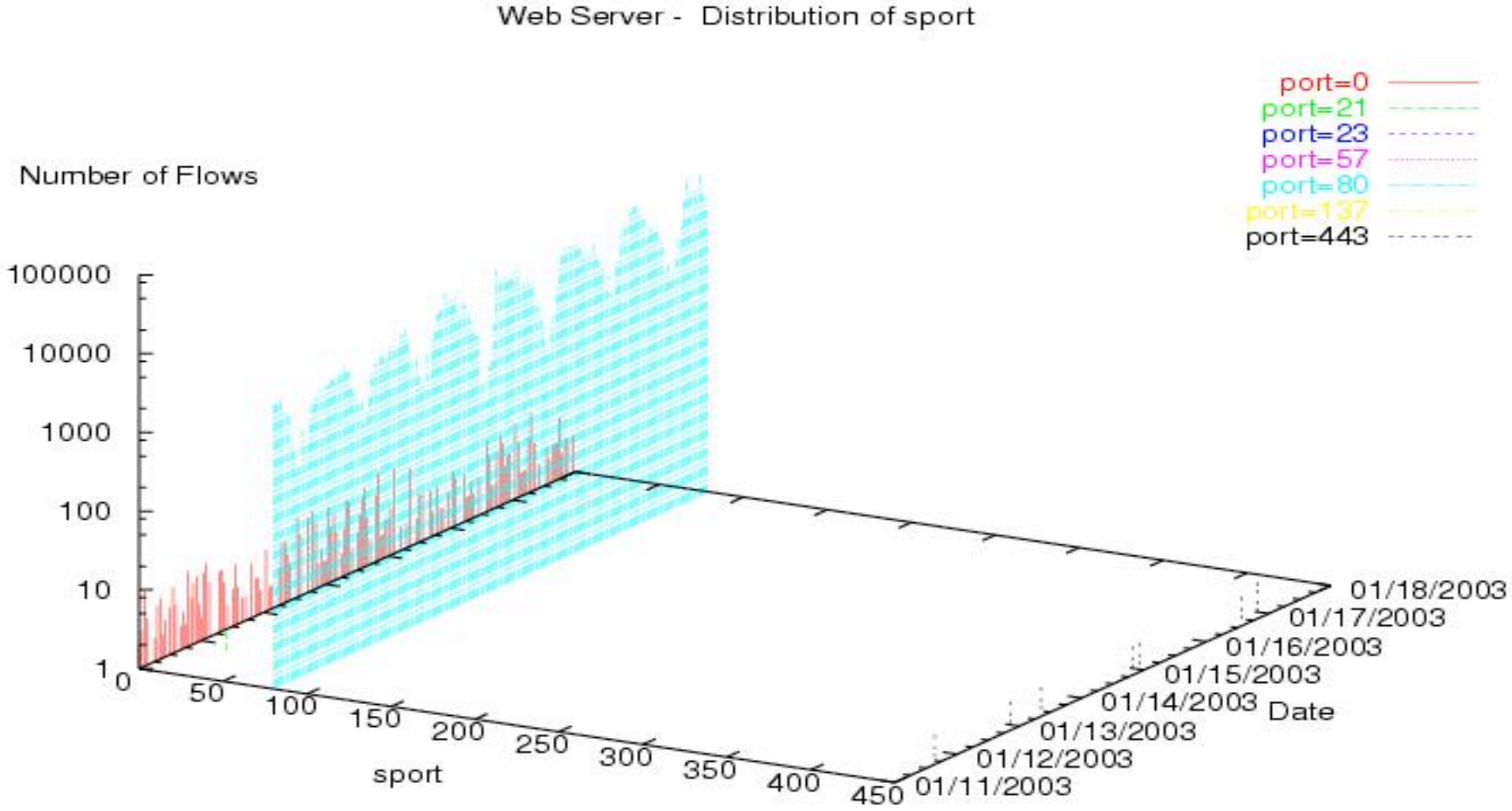
Cpt. Damon Becknel, USA looked at host characterization based on port mixes. We developed the following visualizations to help in the process.

- The log scale on flows helps when there are large differences in flow volumes among ports
- The data was NOT filtered by protocol and there is “noise” in the port fields for some protocols, especially ICMP and possibly others.

Workstation?



Web Server

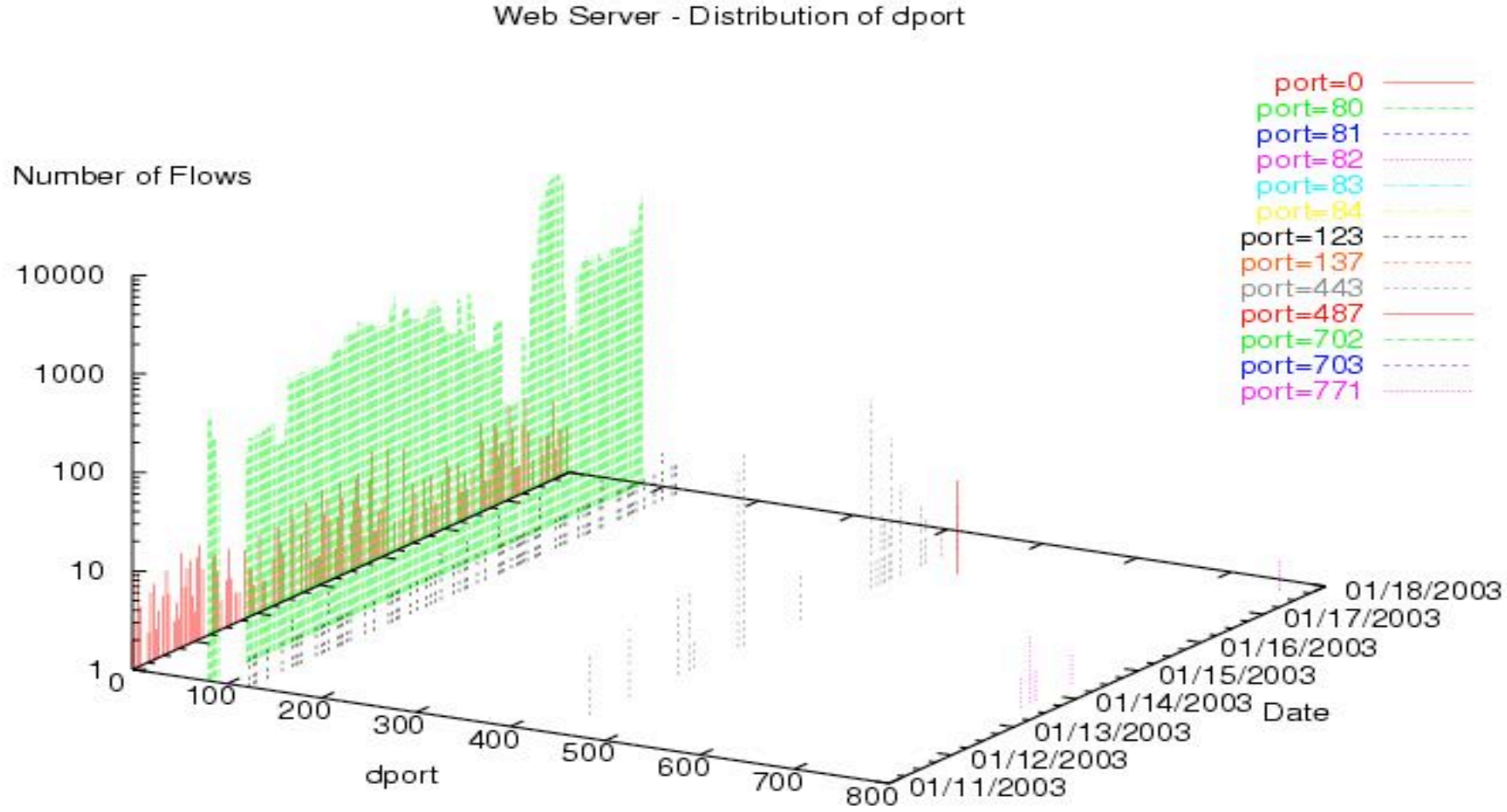

Web Server - Distribution of sport
port=0
port=21
port=23
port=57
port=80
port=137
port=443
Number of Flows
100000
10000
1000
100
10
1
0
50
100
150
200
250
300
350
400
450
sport
01/18/2003
01/17/2003
01/16/2003
01/15/2003
01/14/2003
01/13/2003
01/12/2003
01/11/2003
Date

Analysis of Misconfigurations and Malicious Activity

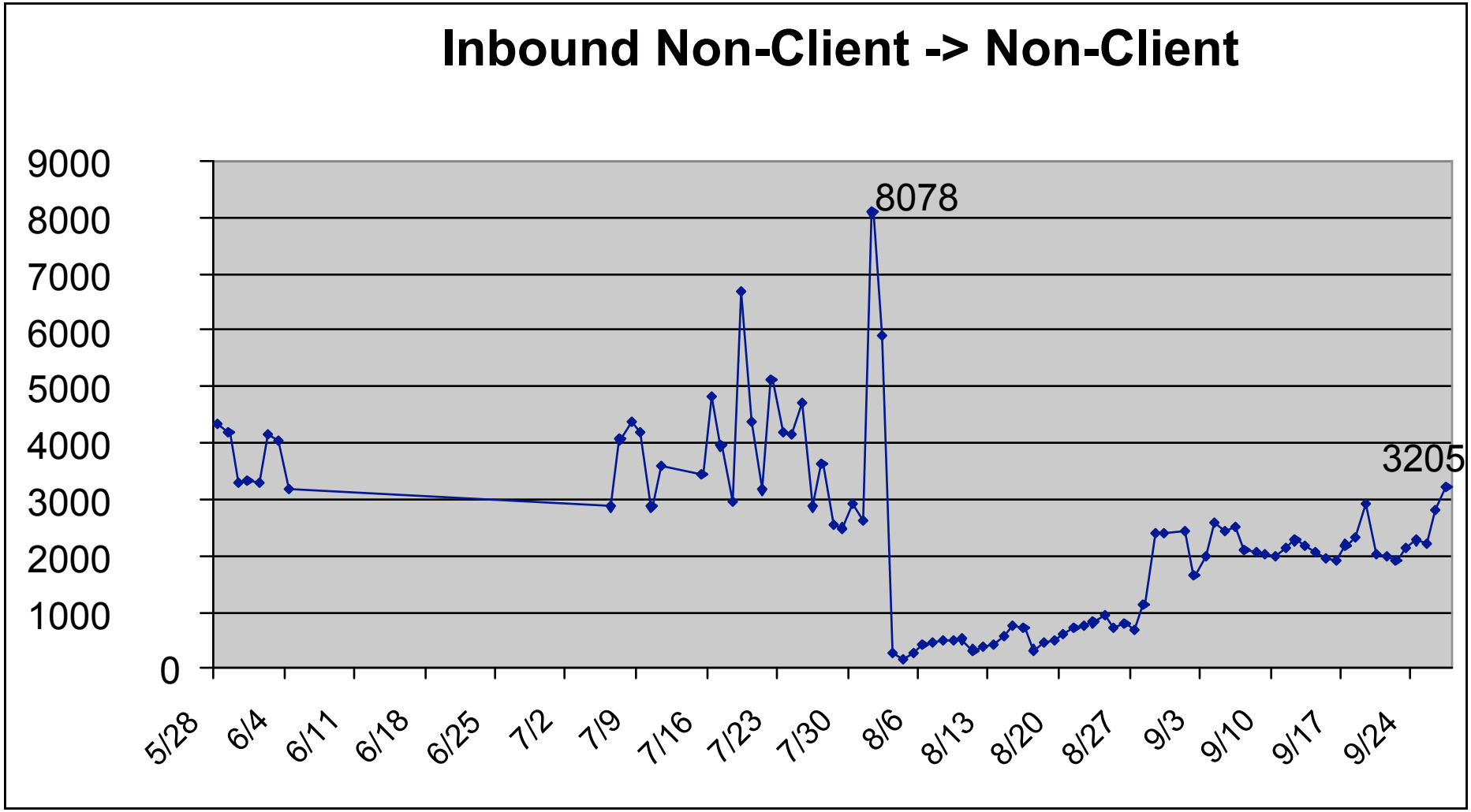
Identify at different levels of detail:

- Web server revisited
- Non-client traffic routed through client networks
- Client traffic inbound, Non-client traffic outbound
- DDoS attack traffic
- Worms of various kinds
- Precursor activities
- Scan Detection
- Contact surface anomaly

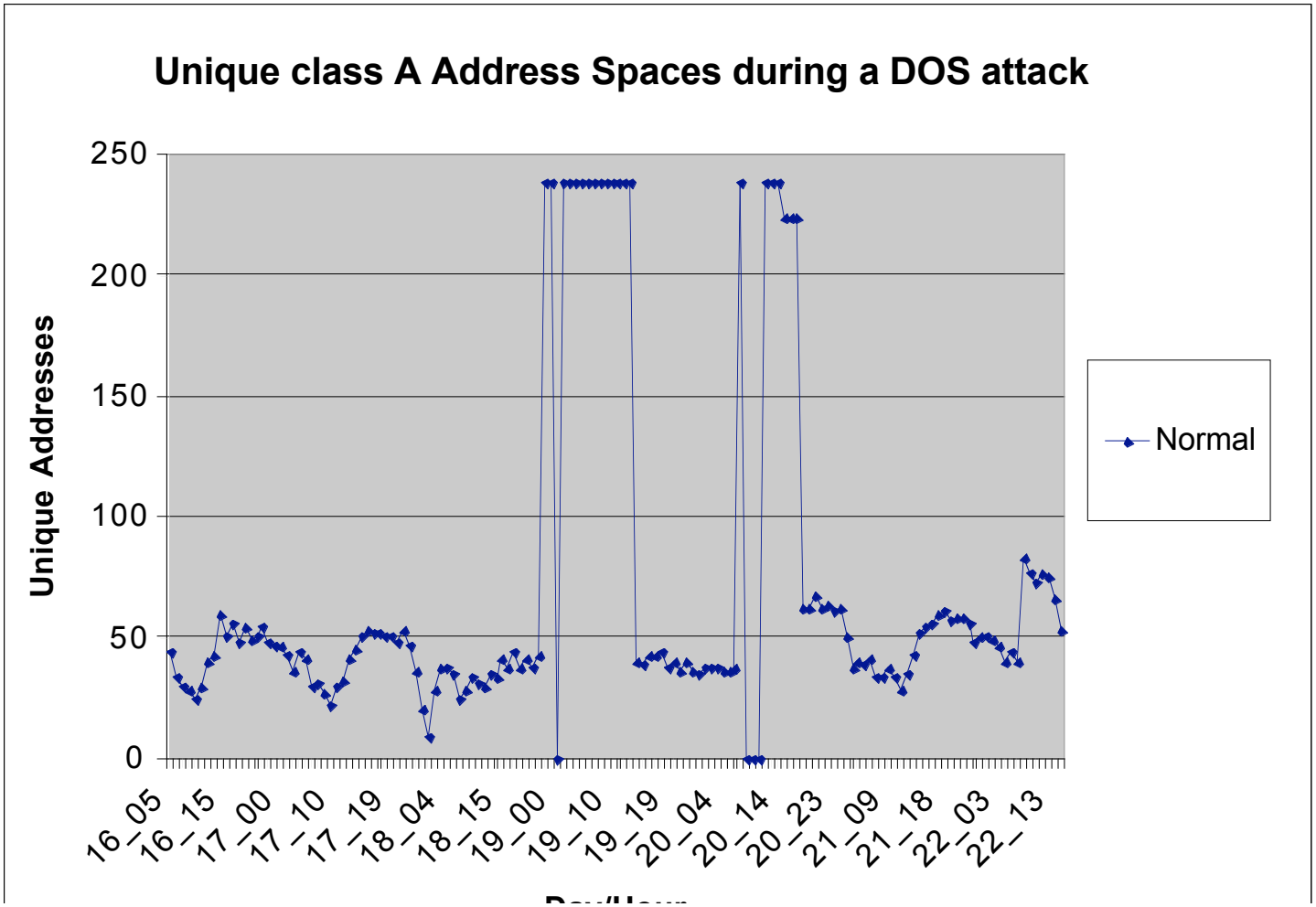
Web Server


Web Server - Distribution of dport
Number of Flows
10000
1000
100
10
1
0
100
200
300
400
500
600
700
800
dport
01/11/2003
01/12/2003
01/13/2003
01/14/2003
01/15/2003
01/16/2003
01/17/2003
01/18/2003
Date
port=0
port=80
port=81
port=82
port=83
port=84
port=123
port=137
port=443
port=487
port=702
port=703
port=771

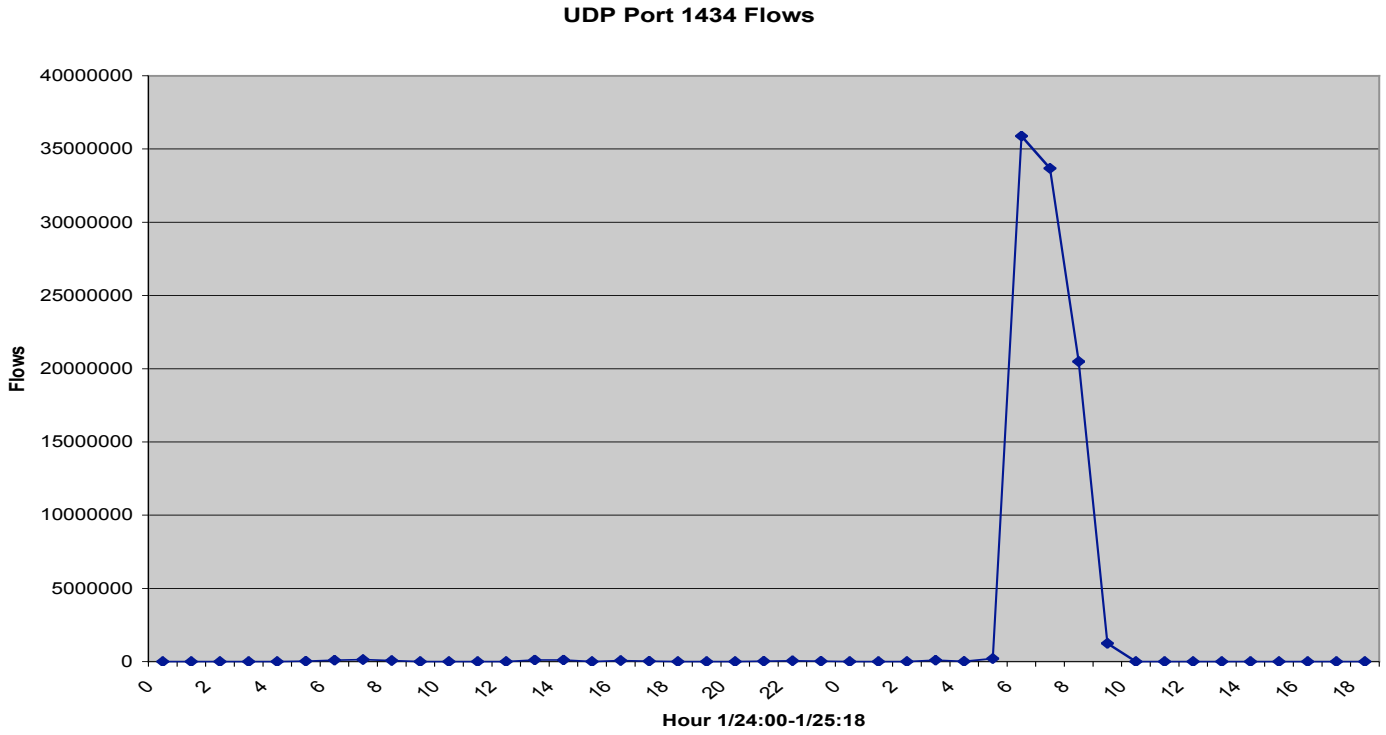
Routing Anomalies and Backdoors



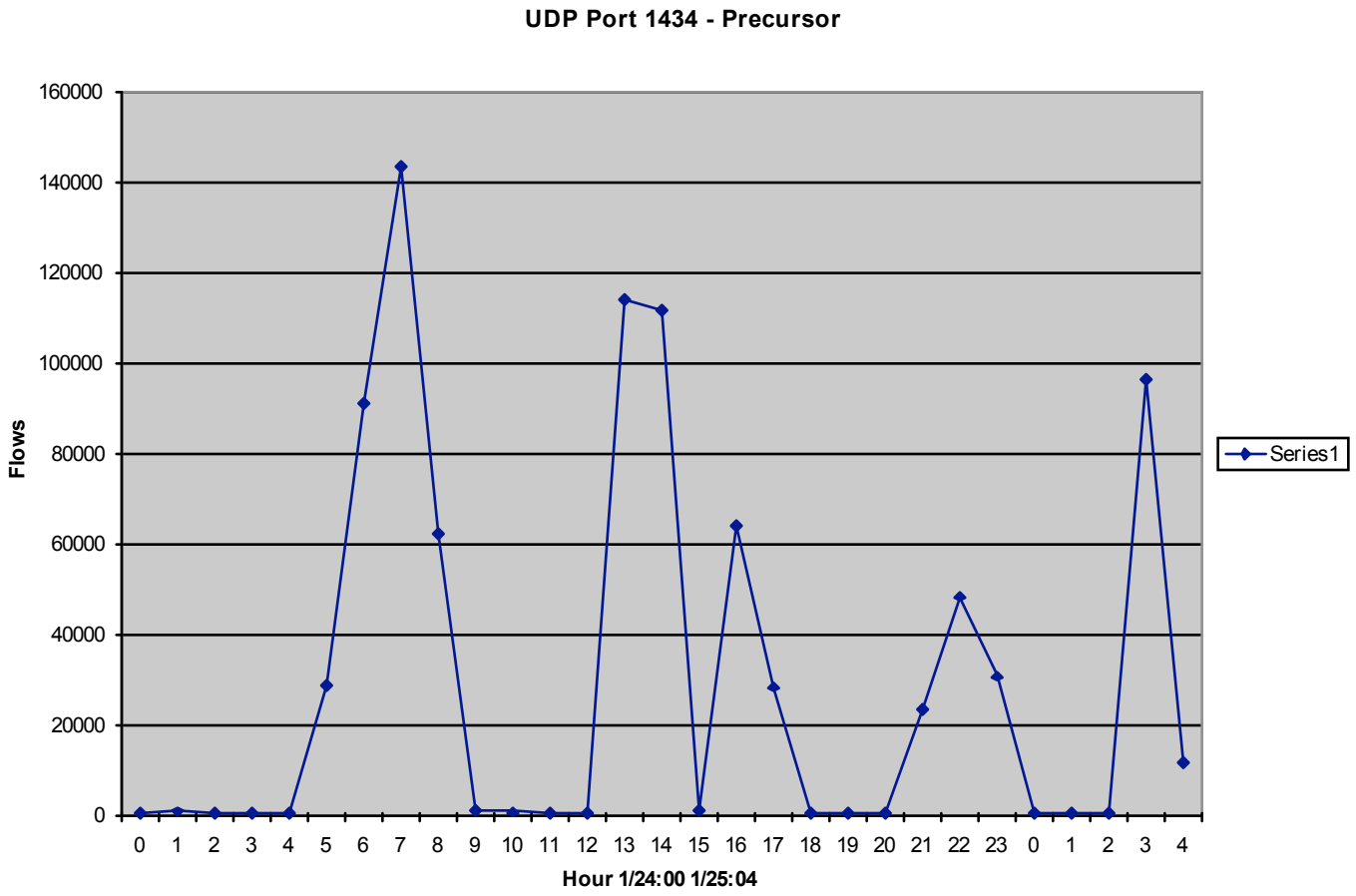
Examining Denial Of Service



Inbound Slammer Traffic



Slammer: Precursor Detection



Slammer: Precursor Analysis

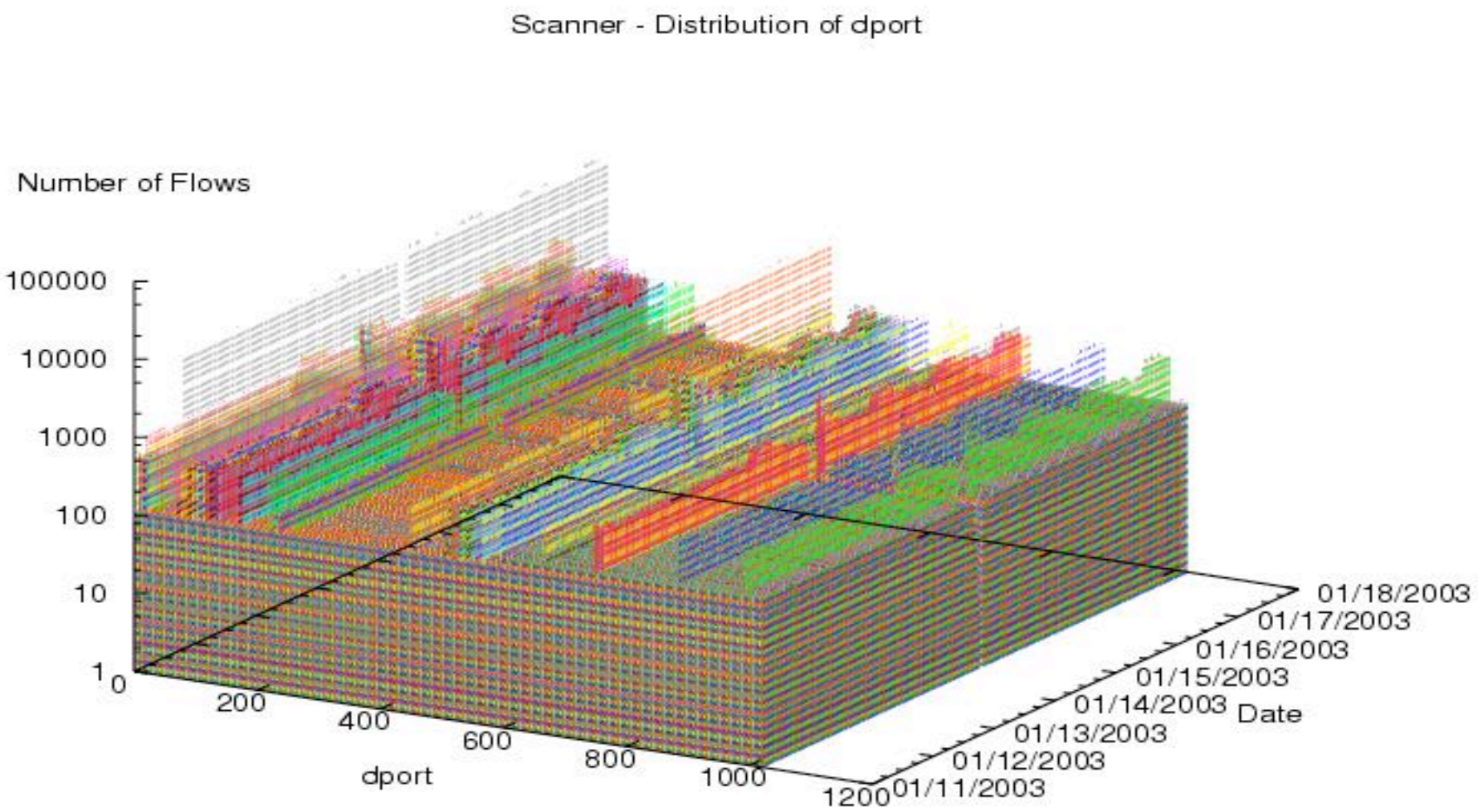
Focused on hours 6, 7, 8, 13, 14

Identified 3 primary sources, all from a known adversary

All 3 used a fixed pattern

Identified responders: 2 out of 4 subsequently compromised.

Scanner

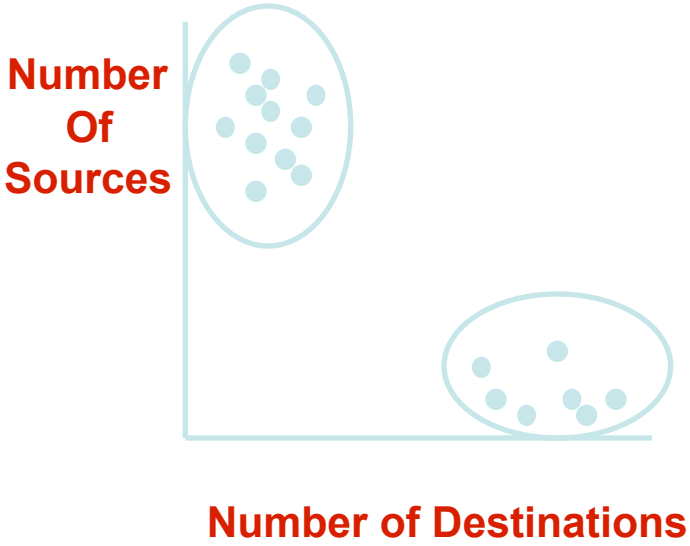

Scanner - Distribution of dport
Number of Flows
100000
10000
1000
100
10
1
0
200
400
600
800
1000
1200
dport
01/11/2003
01/12/2003
01/13/2003
01/14/2003
01/15/2003
01/16/2003
01/17/2003
01/18/2003
Date

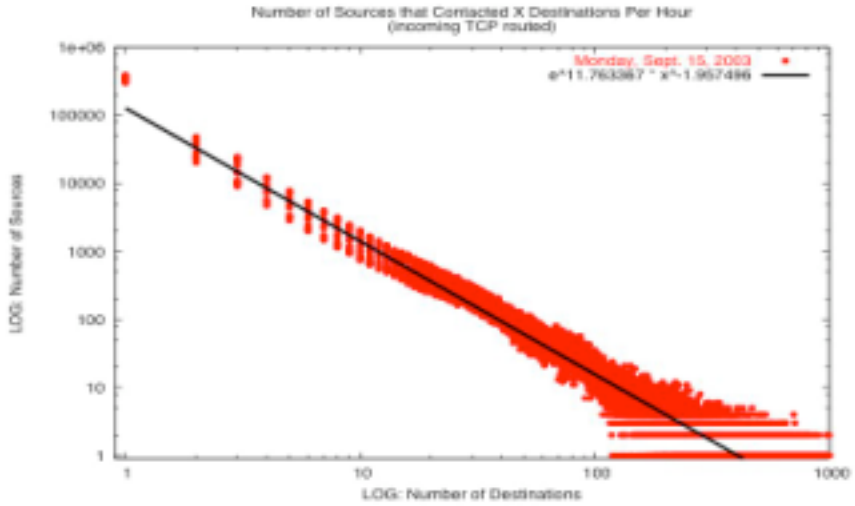
Developing the contact surface

In looking for scanners, Carrie Gates asked if there is a normal contact pattern?

- i.e. how many external hosts contact how many internal hosts per unit time ?

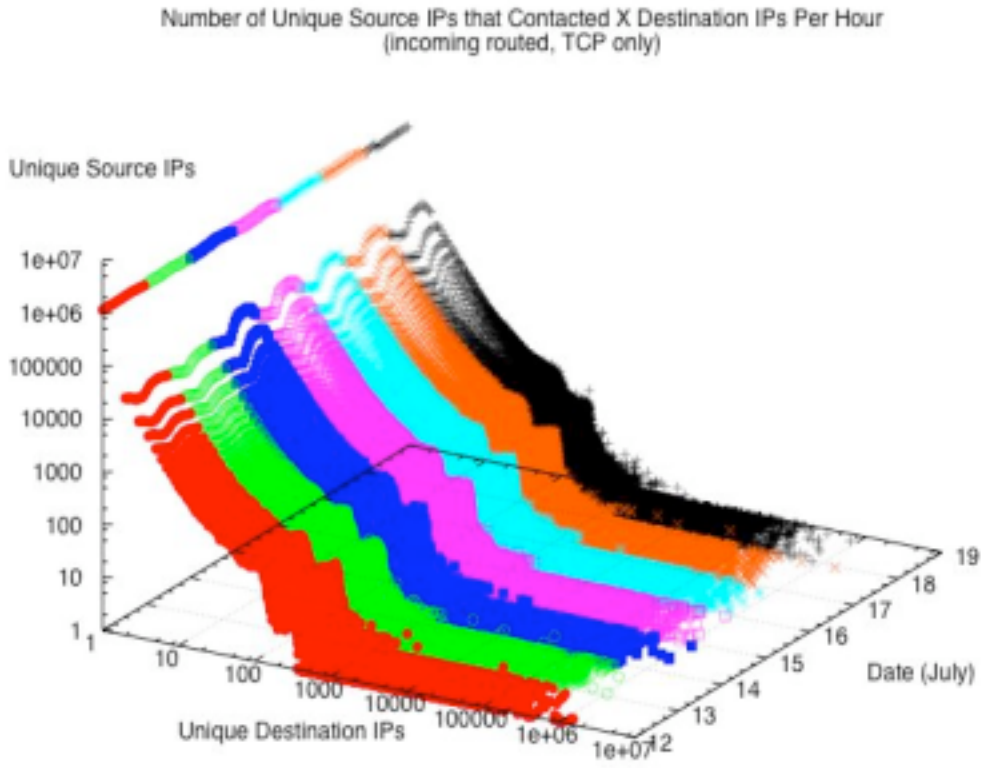
One might expect clustering or a power law, but

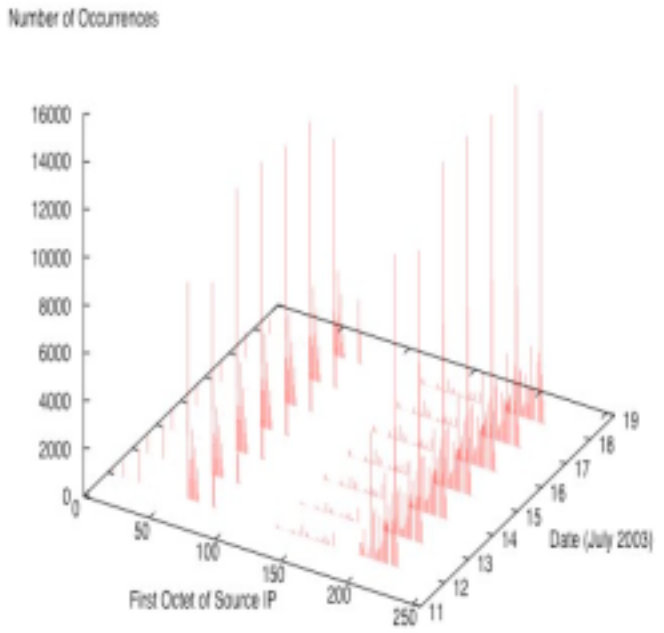




Structure in the contact surface

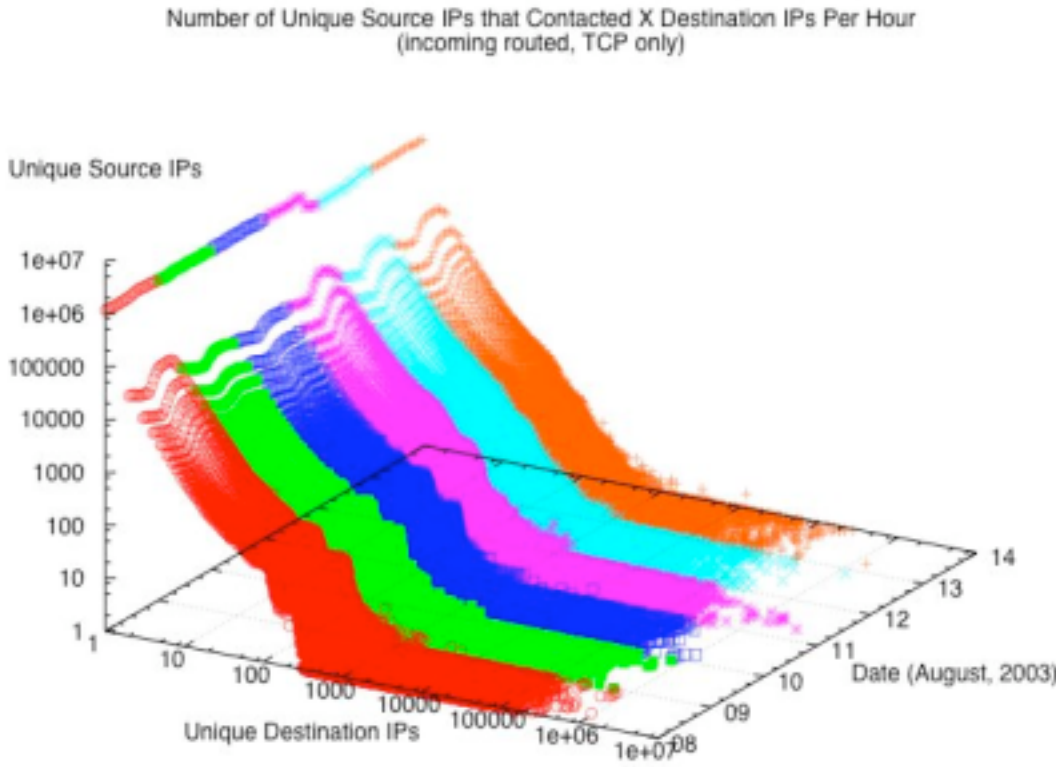
Now you see it (July 2003)

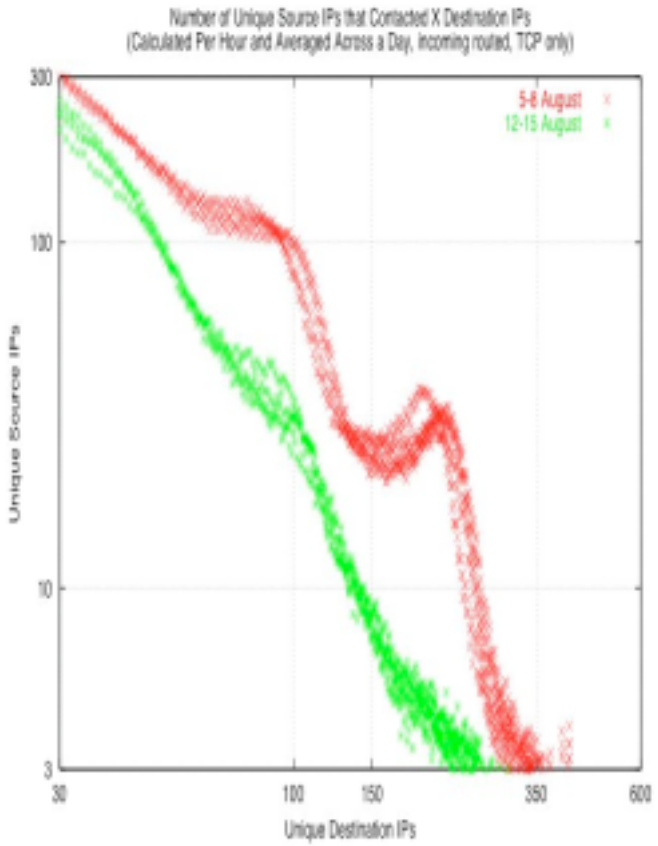




Structure in the contact surface

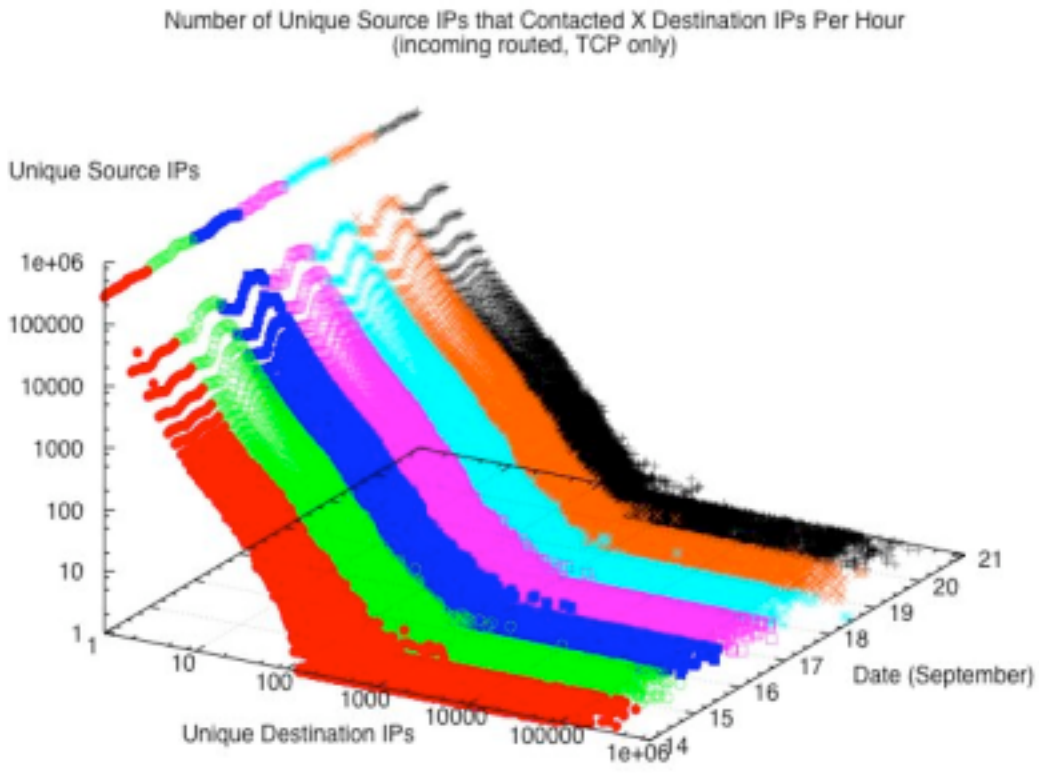
Away it goes (August 2003)

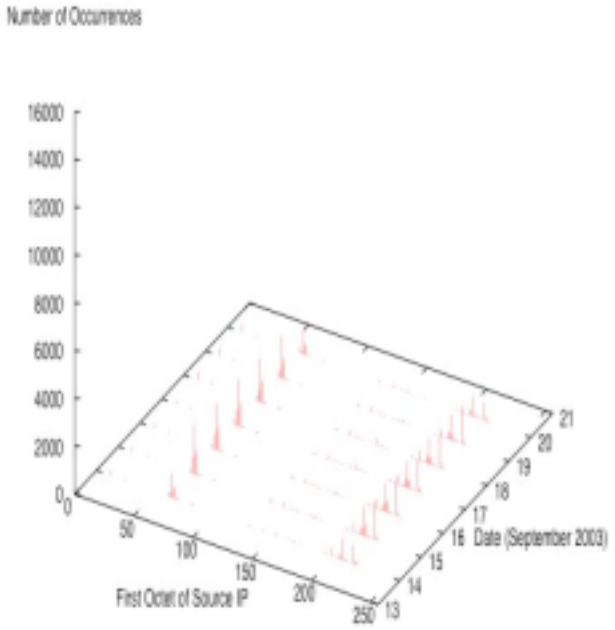




Structure in the contact surface

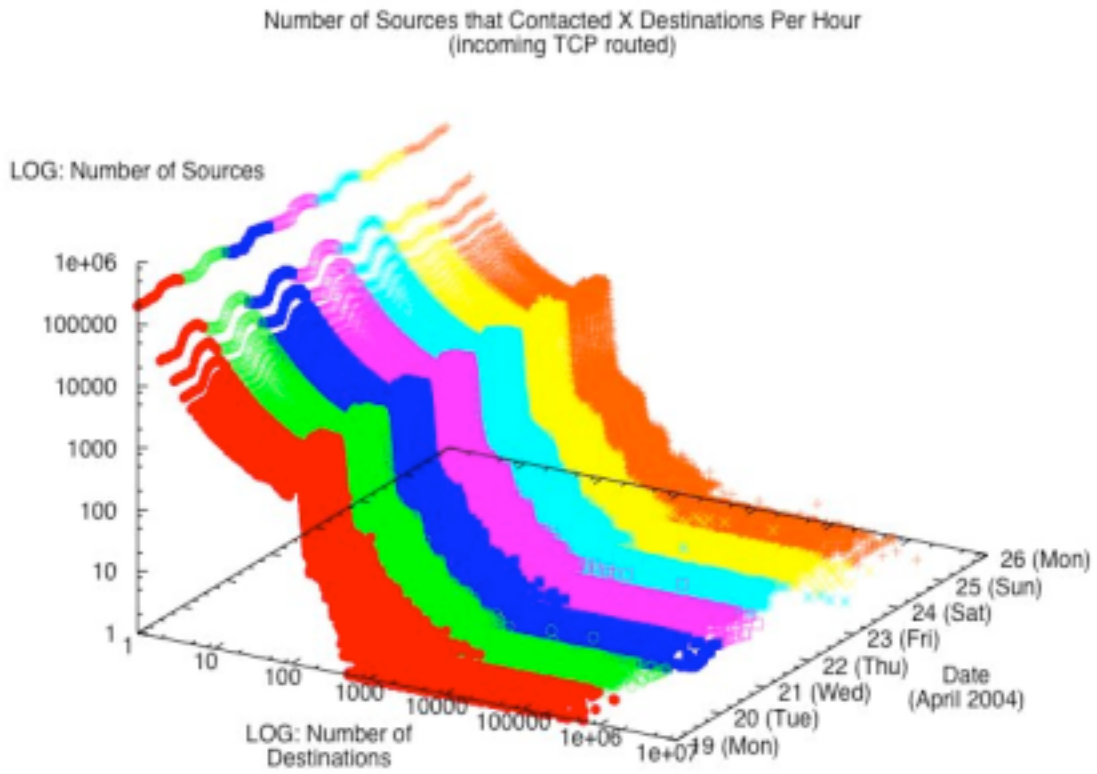
Now you don't (September 2003)

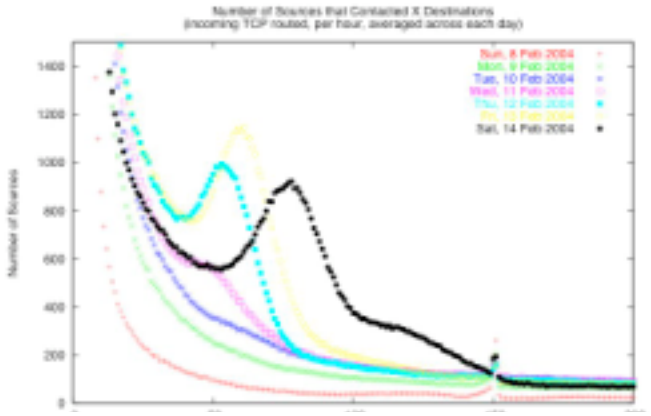


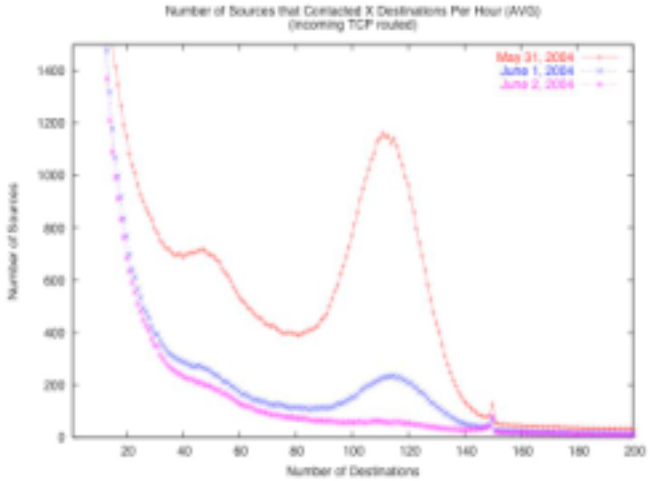


Structure in the contact surface

Here it is again (Feb / April /June 2004)







We are still studying this

Phase 1

- Persisted from Jan - Aug
- details for 1 week in July
- 91% port 80, 87% SYN, 96% unique sIP/dIP pairs
- 49% to 60 /16 in 1 /8
  - 50% of 49% to 5 /16
  - 14% to 1 /16
    - 48 /24s <3.2%ea no target
  - 14% to another /8
- 3 /8 srcs 46%, 34%, 20%
  - 2 AP, 1 SA
- Too slow to be DoS
- Too persistent for scan?

Blaster released on Aug 11

- Could this be the related?

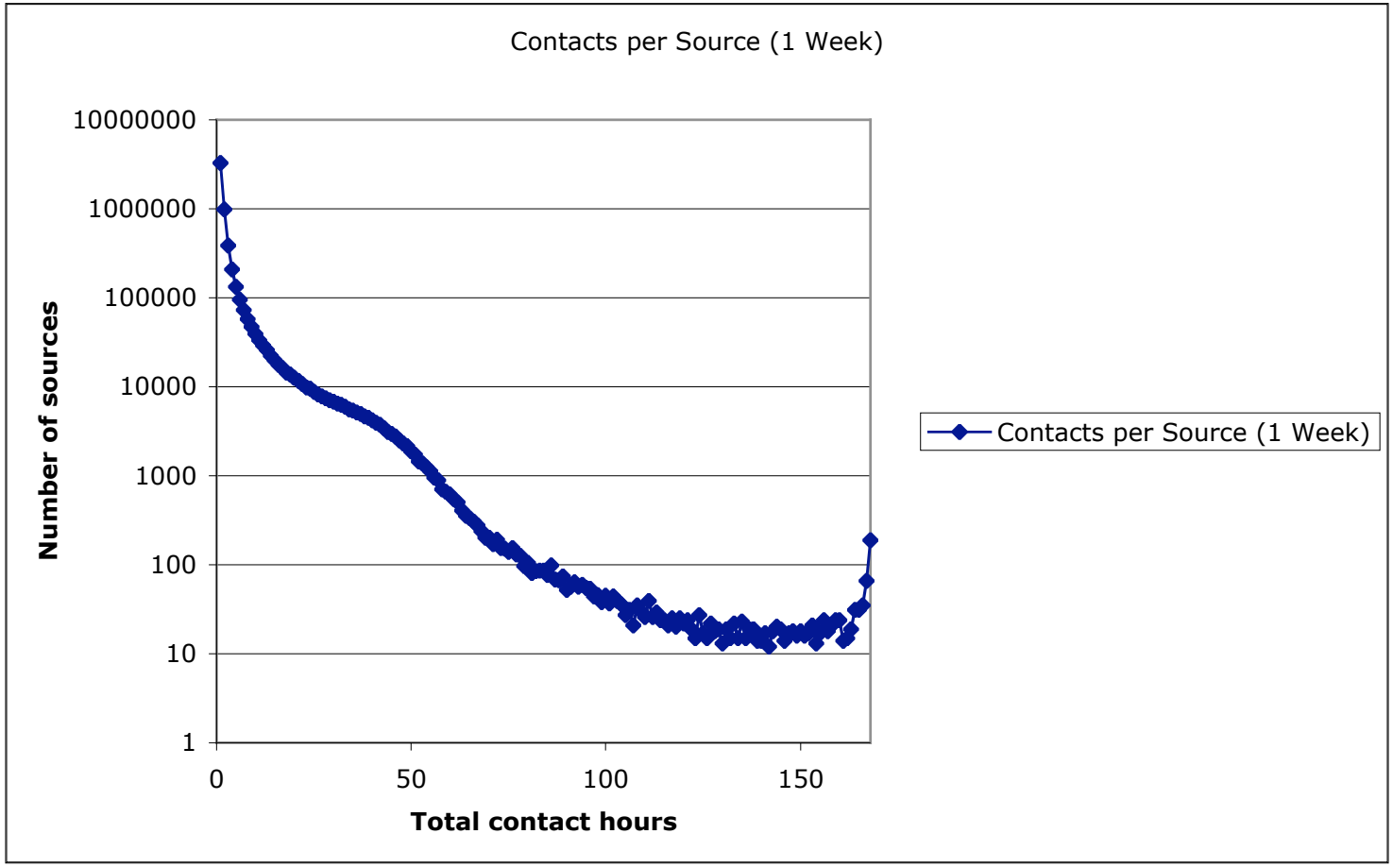
Phase 2

- From mid Feb - early June
- Again SYN to Port 80
- 2 of 3 /8 sources same
  - 3rd is there but not as strong
- 23% to a new /8

Conclusion

- Appears to be well coordinated activity, but what or why?

Weekly Contact Hours



Lack of stealthy activity

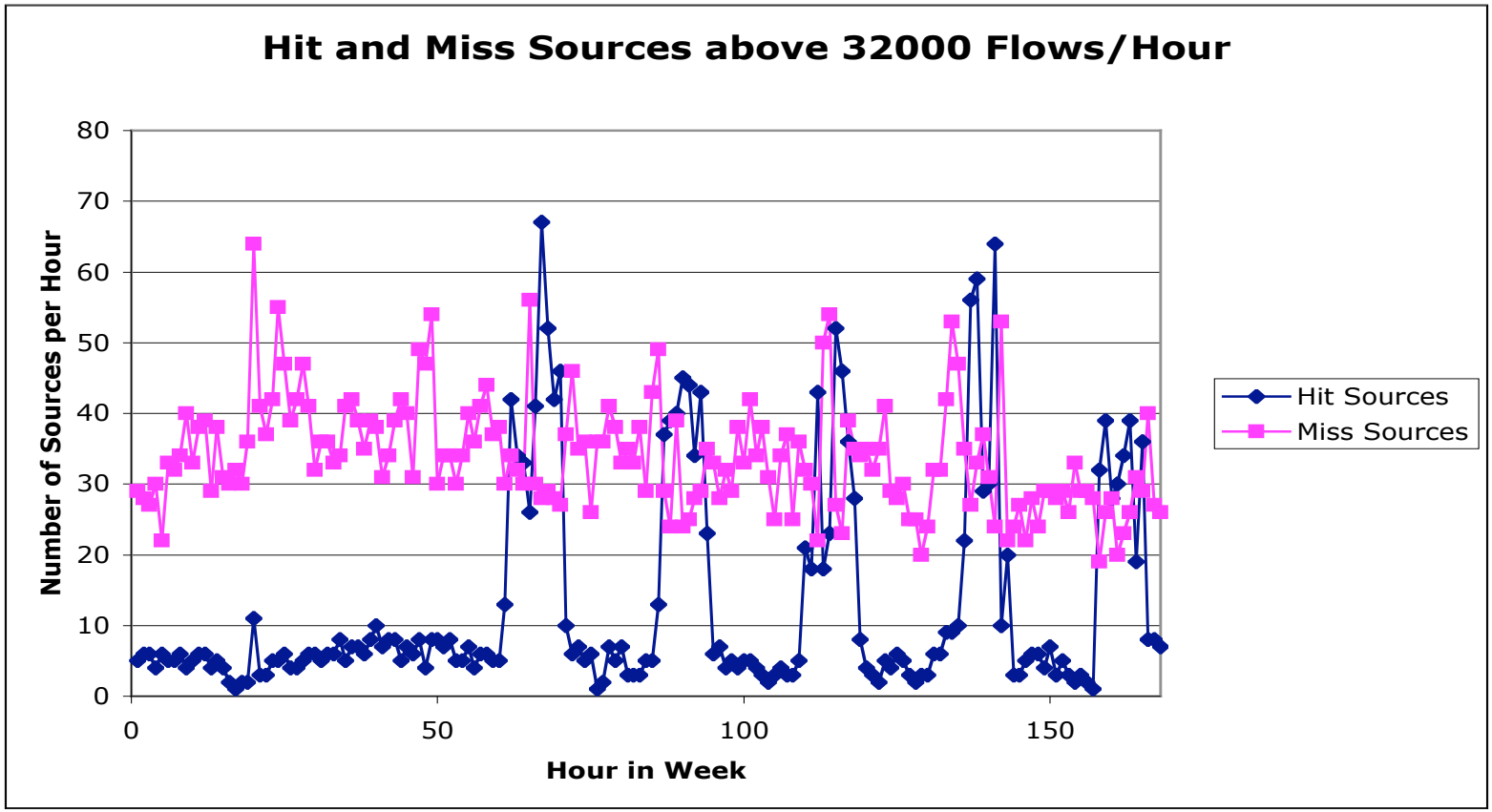
- During the week there are about 3.3 million sources that show up exactly once.
- The distribution tapers off rapidly from there.
- In a linear plot it looks like a straight line at approximately zero, but the log plot shows an interesting upturn at the end. This contains:
  - One connection that persists for the entire week with flows in each hour.
  - A number of apparently scripted transfers that occur every few minutes. At least two involve access to a weather site.
  - At first glance, no evidence of hourly, stealthy activity.

Partitioning Data

A set can be used with rwfilter to partition data into portions that have a destination (or source) IP that are in the set and those that are not.

- The set of hosts in a network that have been observed to emit traffic during some time interval approximates the active population of the network.
- Traffic sent to addresses that are not associated with hosts is arguably malicious
- The active host set can be used to separate this traffic from traffic addressed to active hosts
  - Additional refinements are possible

Hit and Miss Sources above 32000



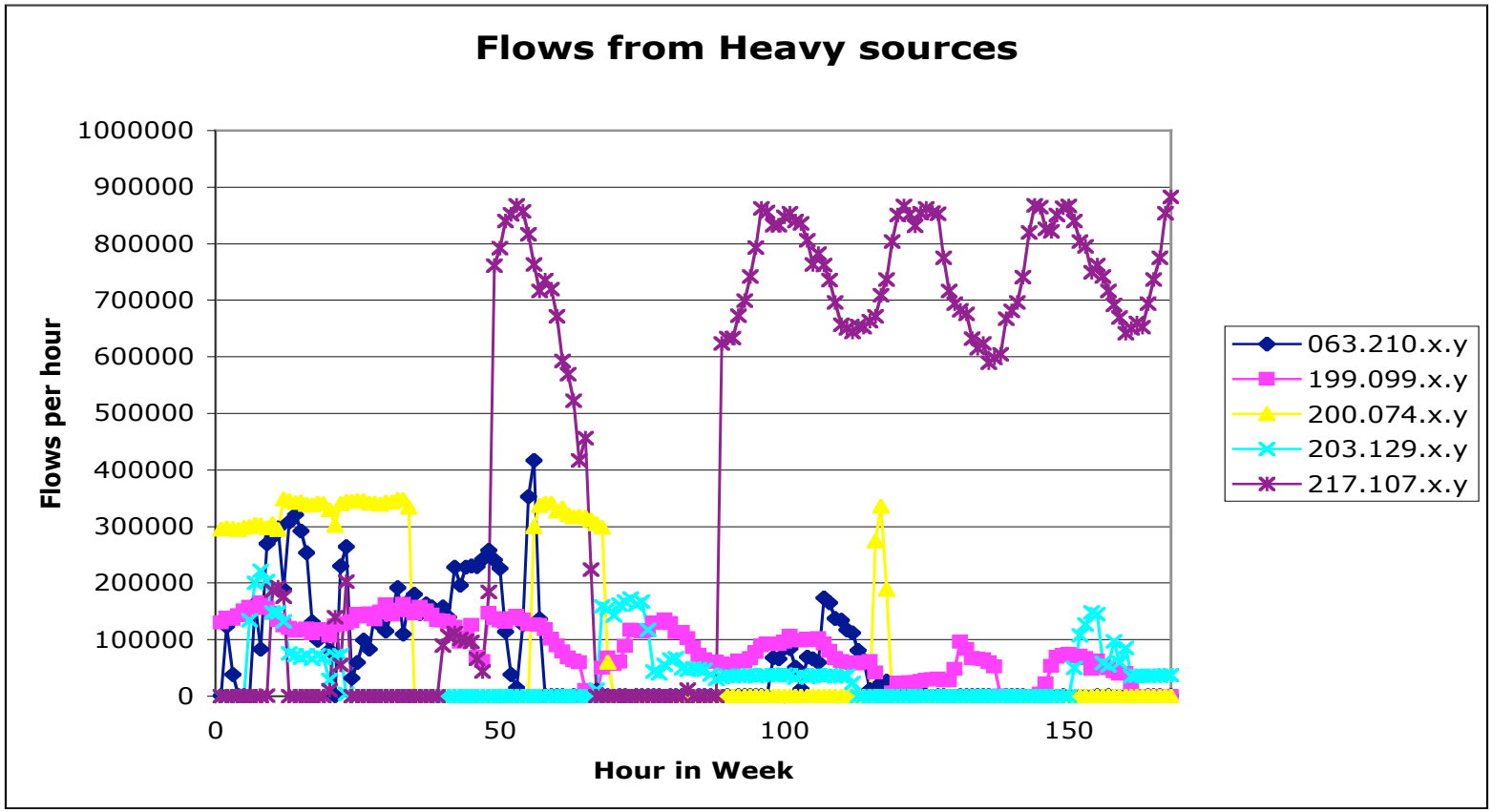
So who are these sources

The following are in the heavy miss list for more than 50 Hours in the week analyzed.

063.210.x.y	66	Level 3 Communications, Inc., CO
199.099.x.y	135	Performance Systems International, DC
200.074.x.y	51	Metropolis Intercom, Santiago, CL
203.129.x.y	77	Software Technology Parks of India, Pune, IN
217.107.x.y	113	RTComm.ru

Two are registered in ARIN, one in the Asian Registry, one in the Latin American Registry and one with RIPE.

And what are they doing?



199.99.x.y is scanning

This analysis is based on 2003/01/14:00

- rwcut -fields=1,4,5,6,7,8 finds the flow signature
- Clustering the lines shows a SYN scan

```
Clusters, counts, and order: 117075 records 2 clusters processsed.
          sIP|dPort|pro|packets|   bytes|   flags|
 199.99.x.y|    80|  6 |       2|      88|  S     |  96842     1
 199.99.x.y|    80|  6 |       1|      44|  S     |  20233     2
End of input after 117075 records forming 2 clusters.
```

- Forming the destination set shows that a single /16

```
readset --print-net=AB# 199.099.x.y-d.set
  AAA.BBB.0.0/16 : 51744 hosts in 228 /24s and 1785 /27s.
```

- This is a fairly common pattern for high volume scanners

200.74.x.y is a more complex case

Clusters, counts, and order: 376572 records 53 clusters processsed.

sIP	dPort	pro	packets	bytes	flags		
200.74.x.y	80	6	1	40	S	92354	1
200.74.x.y	8080	6	1	40	S	91678	2
200.74.x.y	3128	6	1	40	S	90959	3
200.74.x.y	1080	6	1	40	S	88865	4

- Futher down are clusters indicating responses

200.74.x.y	1080	6	2	80	SR	1246	8
200.74.x.y	3128	6	2	80	SR	136	10
200.74.x.y	8080	6	2	80	SR	64	15
200.74.x.y	8080	6	7	386	SRPA	5	21
200.74.x.y	8080	6	4	254	FS PA	4	22
200.74.x.y	80	6	7	338	SRPA	4	23

Proactive use - Spyware

Outgoing traffic from spyware creates hot spots.

- Aggregation of destinations hints at targets
- Similarity of content provides additional evidence
- The wider the spread of the spyware, the easier it is to detect this way
- The more data is aggregated the easier it is to see this.

Zombie command and control networks might be detectable this way, also.

Summary / Conclusion / Future

We have an unprecedented ability to examine network traffic

- Long periods of time, large volumes

Empirical Analyses

- How to identify scans, Denials Of Service
- What defenses work?
- Evidence of compromise

Acquire Additional Data Sources

- Different network topologies - Would like to look at interior nodes and subnets as well as border.
- Different *types* of networks (e.g., Wireless)

Credit where credit is due

To the entire Situational Awareness team, especially:

The late Suresh L. Konda

- Suresh was, and remains, the inspiration for this program

Mike Collins, Andrew Kompanek, and the SiLKtools developers

- Mike Duggan, Mark Thomas

CERT analysts and users of the data

- Carrie Gates, Marc Kellner, Capt. Damon Becknel, USA
  - Carrie and Damon are responsible many of the graphics

Tom Longstaff for managing the unmanagable