Very Low Frequency IP Data
Signal? Noise? or Something Else?

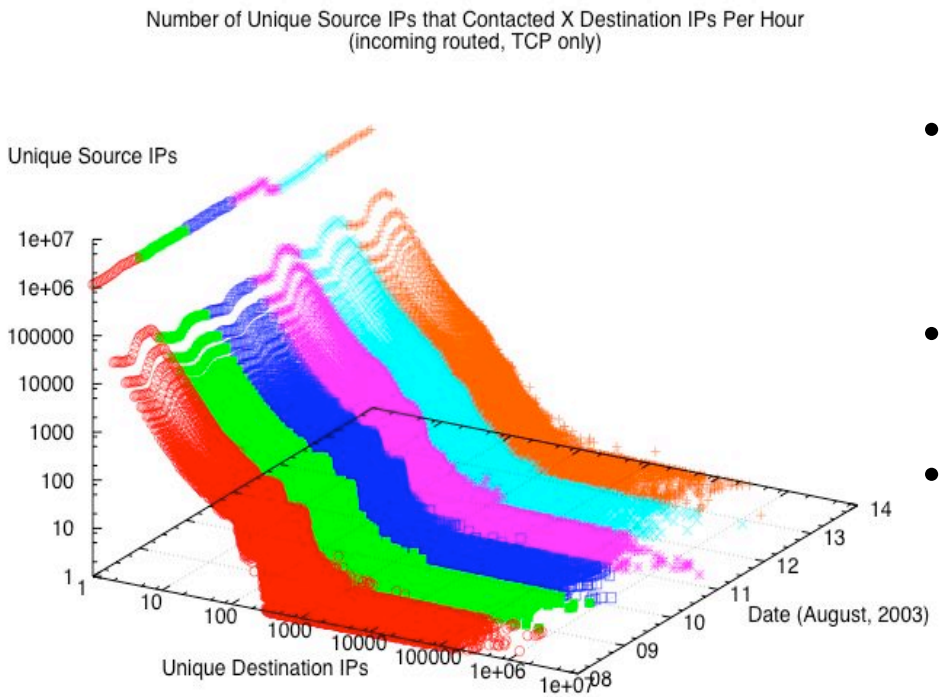
Workshop II:
Applications of Internet MRA to Cyber-Security
17 October 2008

John McHugh
mchugh at cs dot dal dot edu
(but not for long)

The internet is the quintessential MRA target

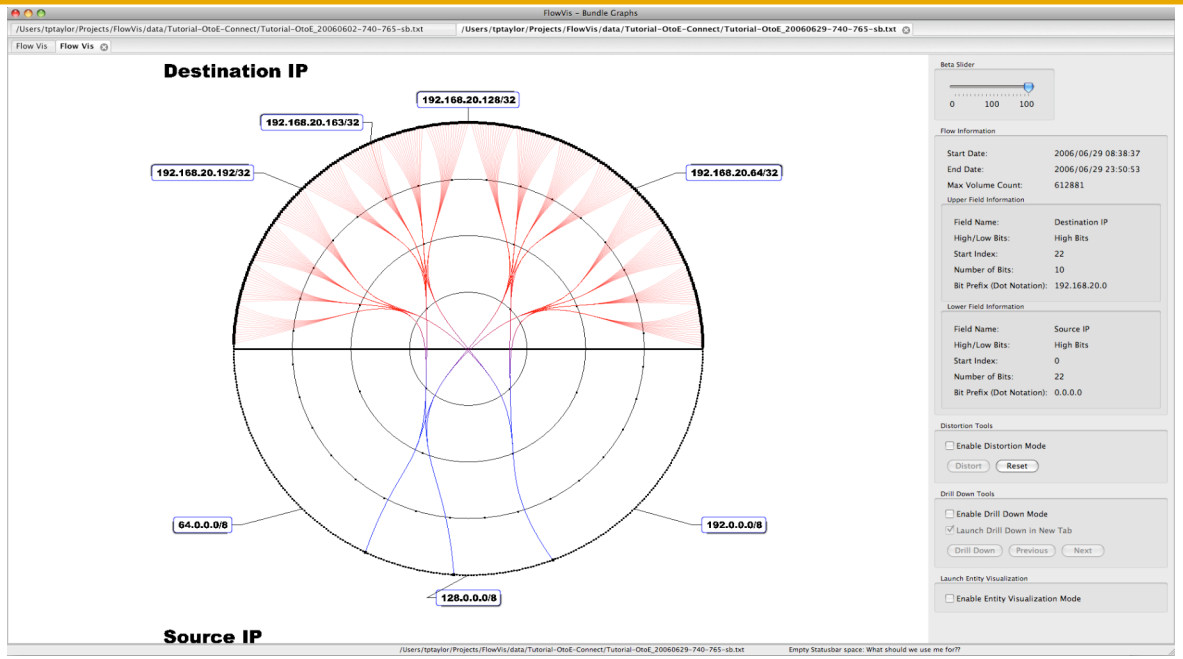
- 2^{32} IPv4 addresses, 2^{128} IPv6 addresses
- Many more complex targets
  - (src, port) X (dst, port) X protocol
- We are interested in behaviors ranging from things that affect the entire internet fabric (see next slide),
- things that reflect group interactions (and the next after that) and,
- things that represent individual host behaviors (a couple of more slides), but
- we will look at hosts that generate very infrequent traffic (for the rest of the talk).

The contact surface



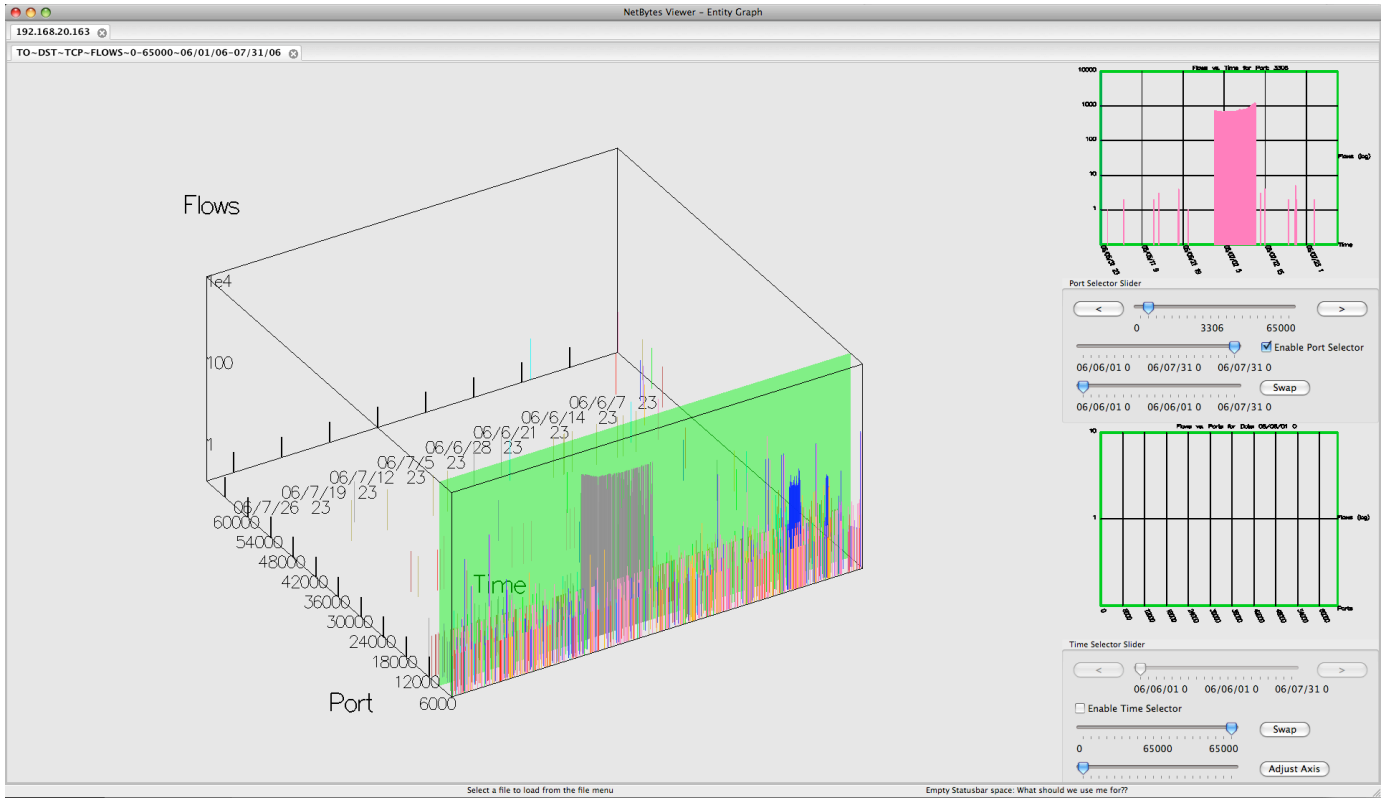
- The behavior of a few hosts disturbs the connection pattern of millions of hosts.
- This is periodic, independent, slow, scanning by a few hundred sources.
- Cause unknown, but killed by “blaster” worm on 2003/08/11.
- We can simulate the process. (see DIMVA 08 paper)

Bundle Diagrams



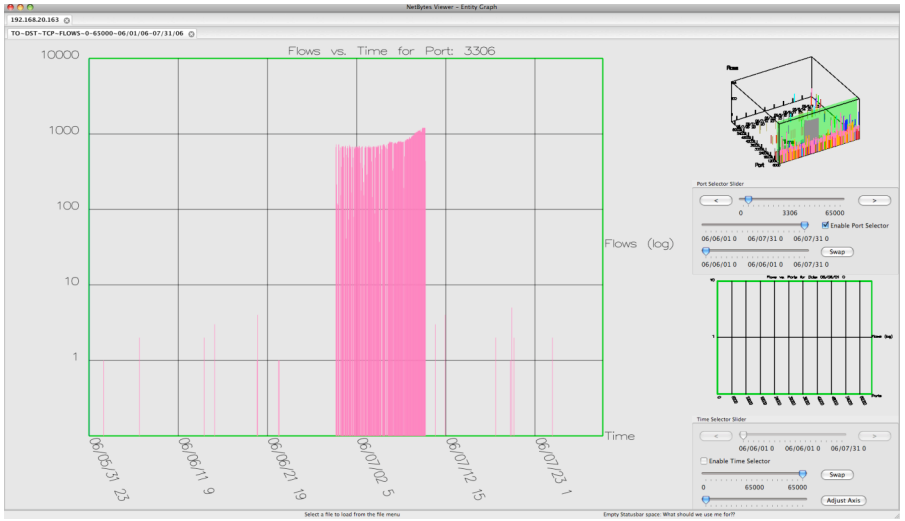
- These show connections between hosts or subnets.
- Current hack allows connection value to be up to 32 bits from any 2 scalar netflow fields
- 2^8 points on each arc
- Shows connection endpoints and relative volume levels
- Can expand portion of arc for additional detail.
- Can pivot to host view or look at NetFlow for thread.

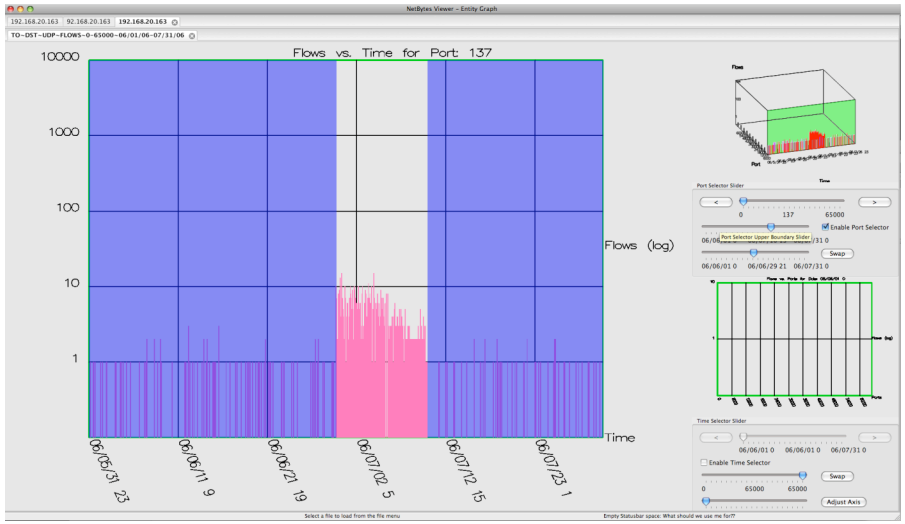
NetBytes Viewer 3D



- Pseudo 3D view of host behavior (port volumes over time)
- Rotate to highlight interestinf patterns.
- Can narrow port or time ranges.
- Can pivot to connect view or drill down to actual data (NetFlow)
- Related 2D views in time or port planes

NetBytes Viewer - 2D



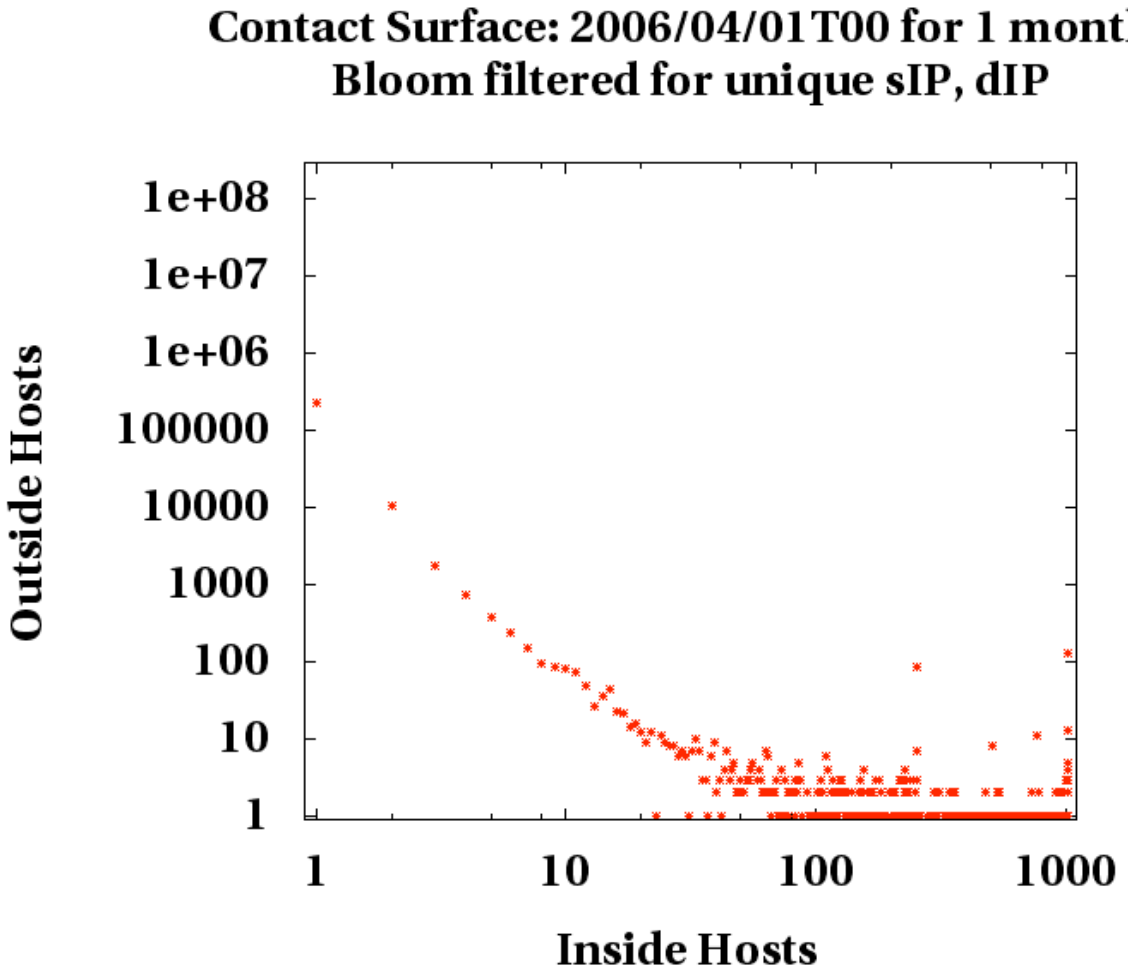


- Can see 2D view for port volumes at a given time or time volumes for a given port. Can select region for more detail (R)
- This is 3306 TCP and 137 UDP for the victim.

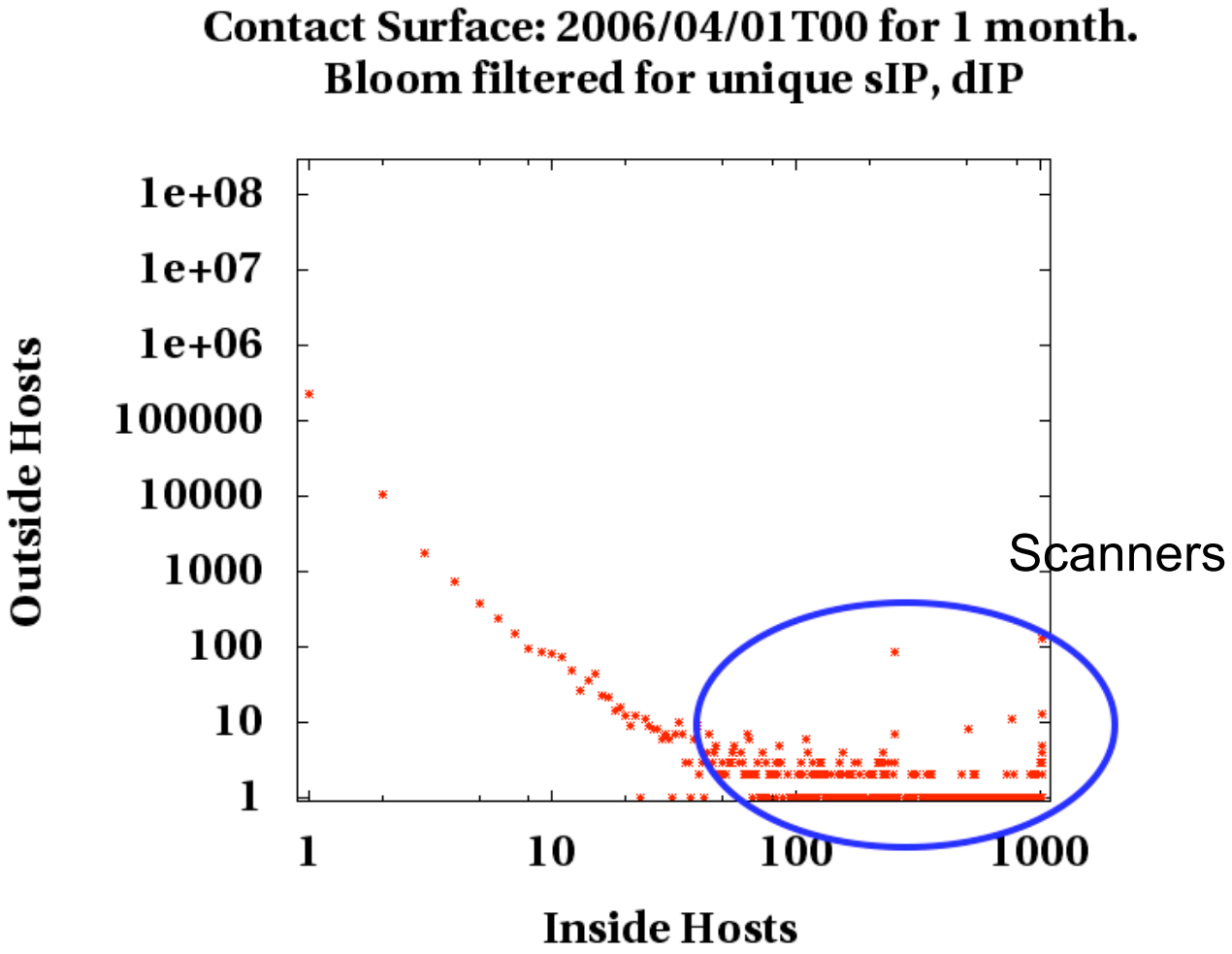
The problem at hand - VLF data

- We have been monitoring a local /22 since 2006/02
- 14 months of data till mid 2007/03 considered
  - NAT used from that point + collection failures
- 1024 addresses
  - <120 ever active, typically 80-90 during a given week
- About 13 million Outside addresses seen
  - over 90% generated 1-10 flow records
  - 93% target a single address
  - about 90% appeared in 1-5 hours
- Even at this scale, the distribution looks familiar

April Contacts



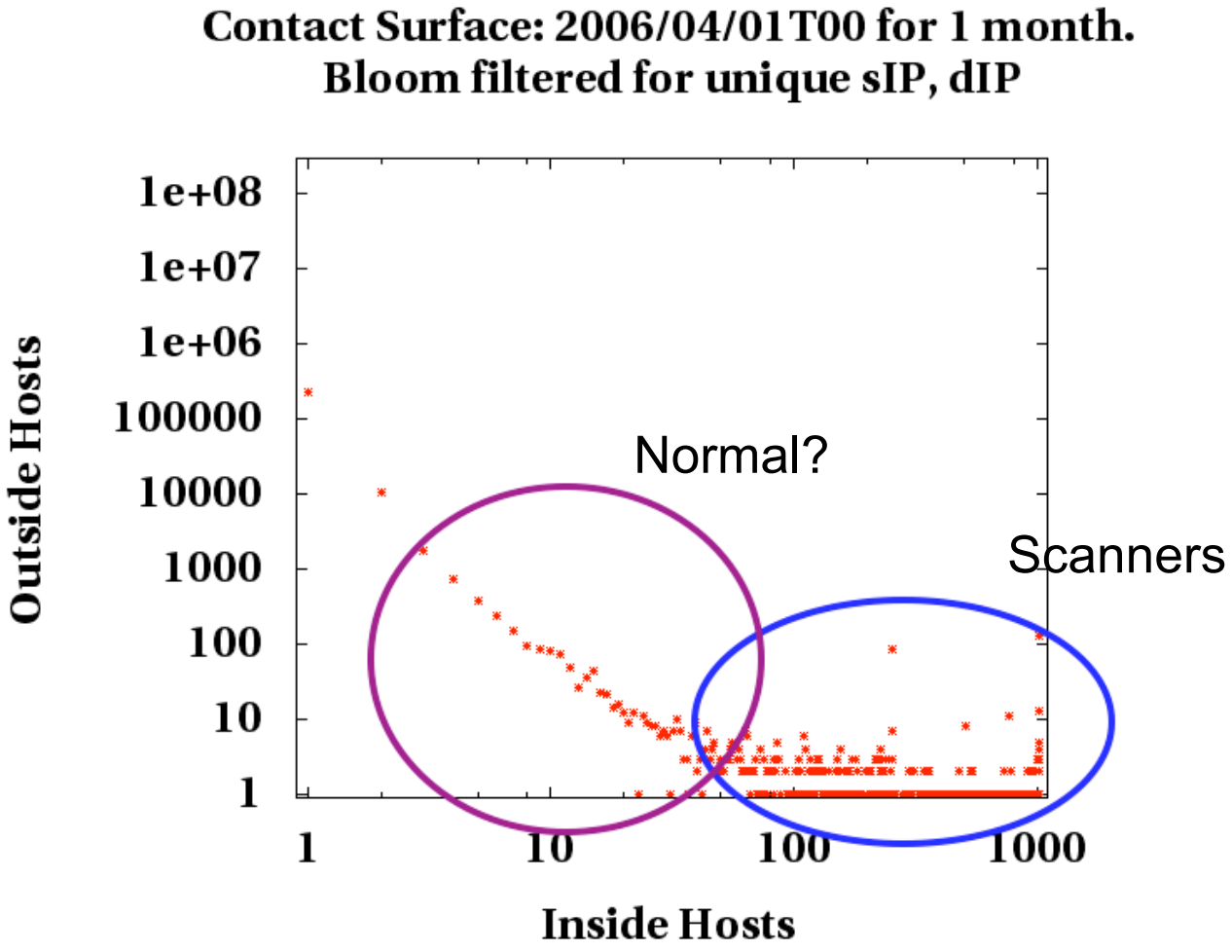
April Contacts



Scanners are ubiquitous

- We have thousands of hosts that attempt to contact all addresses in between 1 and 4 of our /24s.
- Typically scan for the vulnerability du jour
- Typically get very few responses and even fewer serious interactions.
- Account for large fraction of total flows.
- Obvious scans are trivial to identify
- Subtle ones can be detected with algorithms like TRW, etc.
- We have another project that looks at using lossy compression of scans to reduce archive volume.

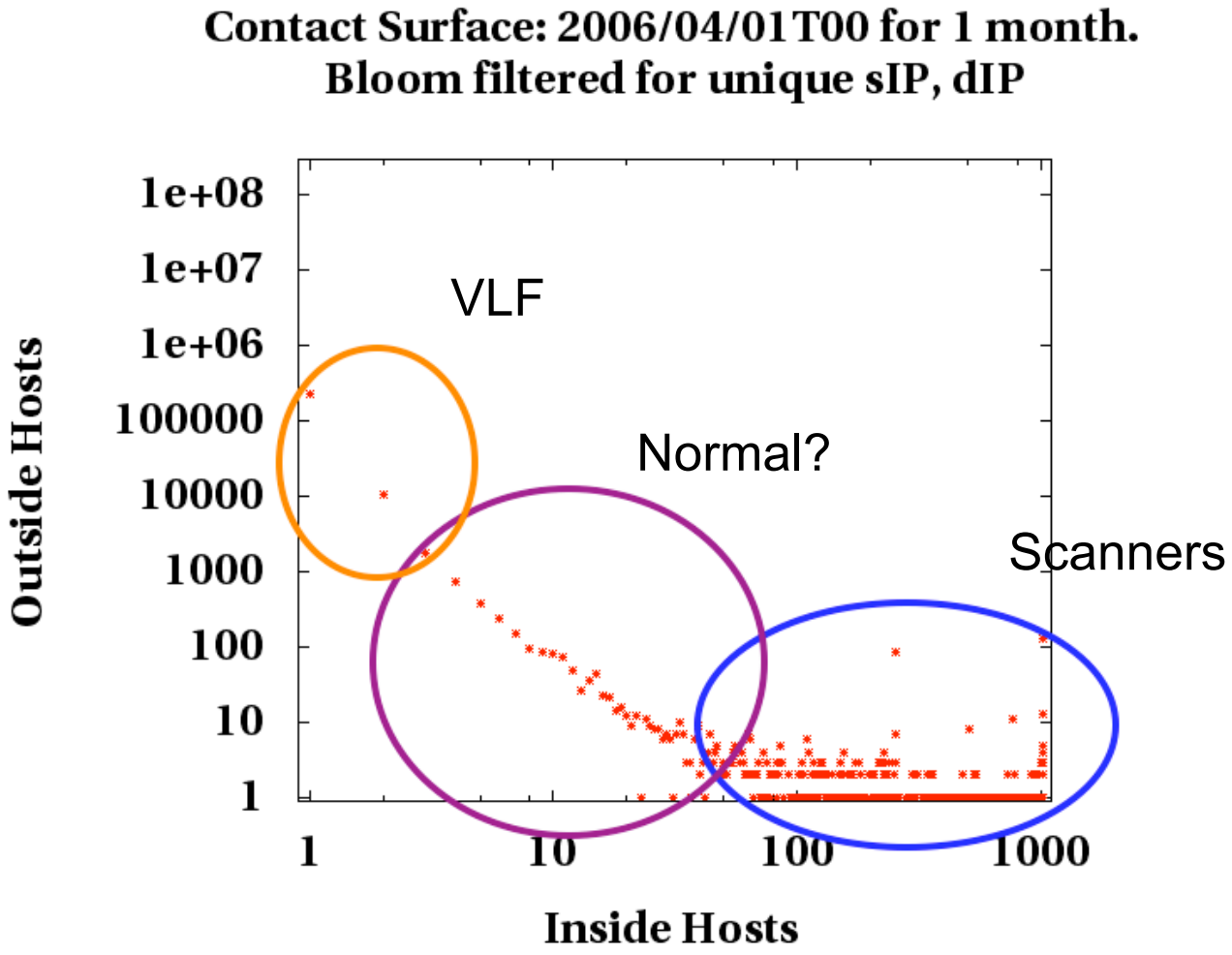
April Contacts



The middle of the distribution

- This probably includes both normal and malicious interactions.
  - Upper (fewer hosts) blends with VLF
  - Lower (more hosts) blends with scanners.
- In this network, we have seen a variety of behaviors ranging from a machine subverted to become a “half life” game server to various P2P applications in addition to the “real” work of the enterprise.
- Ron McLeod has looked at automatic host role classification on the network. Prior to the P2P invasion, it worked well. Now, it does not.

April Contacts



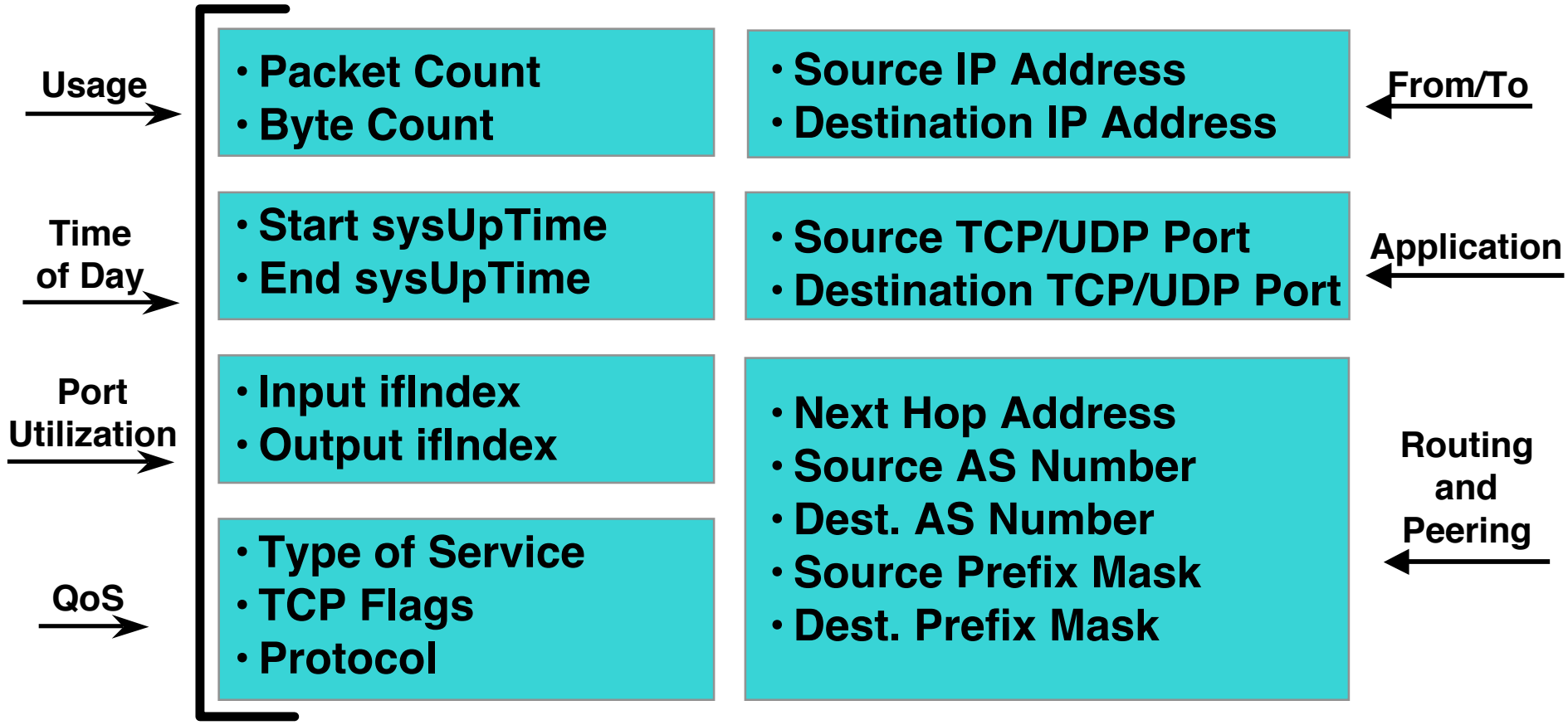
What is the VLF component?

- Intuitively, we expected to find:
  - Spoofed DDoS backscatter
  - Low frequency worm scans
  - Misconfigured hosts
  - Typos
  - Etc.
- We did find this, but we found a lot of intentional, full up connections and some strange cases.
- Most of the rest of the talk is observational and descriptive. At this point, I do not understand what I am seeing.
- But first, a digression on data and methods ...

NetFlow Origination & Innovation

- Developed by Darren Kerr and Barry Bruins at Cisco Systems in 1996
- The value of information in the cache was a secondary discovery
  - Initially designed as a switching path
- NetFlow is now the primary network accounting technology in the industry
- Sampled NetFlow a Cisco innovation
- NetFlow version 9 an emerging IETF standard
- Answers questions regarding IP traffic: *who, what, where, when, and how*

Version 5 - Flow Format



Silk Netflow

- Selected fields from the Cisco record
  - sIP, dIP, Protocol
    - sPort, dPort (UDP, TCP)
    - flags (TCP), [recently, first pkt flags, also]
    - ICMP msg/code
  - packets, bytes
  - start time, duration (1 sec [recently ms])
  - sensor
  - input, output interface IDs
  - next hop IP

File organization

- Partitioned by collection and analysis strategy
  - in, inweb, inicmp, out, ... null, int-int, etc.
- Within a partition, temporal hierarchy
  - YYYY/MM/DD/
  - file names code hirearchy and source
    - <part>-<sensor>_YYYYMMDD.HH
- Hourly files are packed to minimize time under head
  - Redundant fields removed to file header
  - variable fields reduced to minimum bits

The SiLK Tools

- Requirements:
  - Provide a historical data collection that supports retrospective analysis
  - Optimize for retrieval; don't forget this is an intrinsically I/O-bound problem
  - Scale affordably and easily
  - Provide a flexible foundation for analysis--don't build yet another intrusion/anomaly detection system
- Guiding Principles:
  - Leverage Unix, Unix file system, Unix "mind-set"
  - Keep it simple
  - Provide a toolbox for building more complex applications
  - Prototype, get feedback, then generalize

SiLK Tools

- Currently maintained by CERT NetSA group
  - GPLd ans available at http://tools.netsa.cert.org/silk/index.html
- Runs on most flavors of Unix, including Mac OS X and Solaris.
- For a good introduction, take my tutorial at ACSAC (Dec. 9 2008 in Anaheim)
- CERT Flow conference (Flocon) in Phoenix in mid January also has a tutorial.
  - Mine is better for researchers as it stresses approach

The key tools (for this analysis, anyhow)

- rwfilter - select data from flow file / archive
  - rwcut - print selected fields
- rwmatch - find corresponding flows in 2 unidirectional files using related field pairs, i.e. sIP <-> dIP, etc
  - cutmatch - print matches
- rwset - create sets of IP addresses from flow files
  - rwsettool - union, intersection, difference
  - rwsetcat - print sets [show subnet structure]
- rwbag - create multisets of IP (ports, protocols)
  - rwbagtool - operations on bags add, coverset, mask
  - rwbagbuild - make bag w default count from set, list
  - rwbagcat - print bags [subnet, bin, etc.]

rwfilter

- Partitions data into pass, fail files based on
  - addresses, ports, protocols, and other flow fields
  - set membership for src, dst, and nh IPs
  - Time ranges, etc.
  - plugins can filter on complex relationships
    - Bloom filter based plugin can extract exemplars of unique field combinations (sIP, dIP), (sIP, dIP, Protocol), etc.
  - silk.conf file specifies partitioning and sensors
    - allows data extraction from archive by start and end dates

rwbag

- Count volumes for IPs, ports, etc. in flow files
  - Volume measure is flows, pkts, bytes
  - Index by
    - sIP, dIP, nhIP (32 bit key)
    - sPort, dPort, sensor, in index, out index (16 bit key)
    - Protocol (8 bit key)
  - Can create multiple bags from 1 run
  - Implementation is /9, /9, /9 pointers /27 counters
    - Counters are 64 bit
    - will not scale to IPv6 or connection keys

rwbagtool

- Can add a group of bags
- Can extract the coverset of a bag
- Can mask a bag with a set (or its complement) giving a sub-bag
- Can restrict bag based on min and max index or min or max count
- Other operations (subtraction, division) are problematic

So how do we find VLF activity?

- minimal flows: 1 .. 10 in 14 months

  rwfilter (dates) | rwbag (source flows) | \

  rwbagtool (maxcount=10 coverset) > VLF-f.set

  - or add up hourly bags
- minimal destinations: 1 target in 14 months

  rwfilter (dates, bloomSD) | rwbag (source flows) | \

  rwbagtool (maxcount=10 coverset) > VLF-d.set
- both

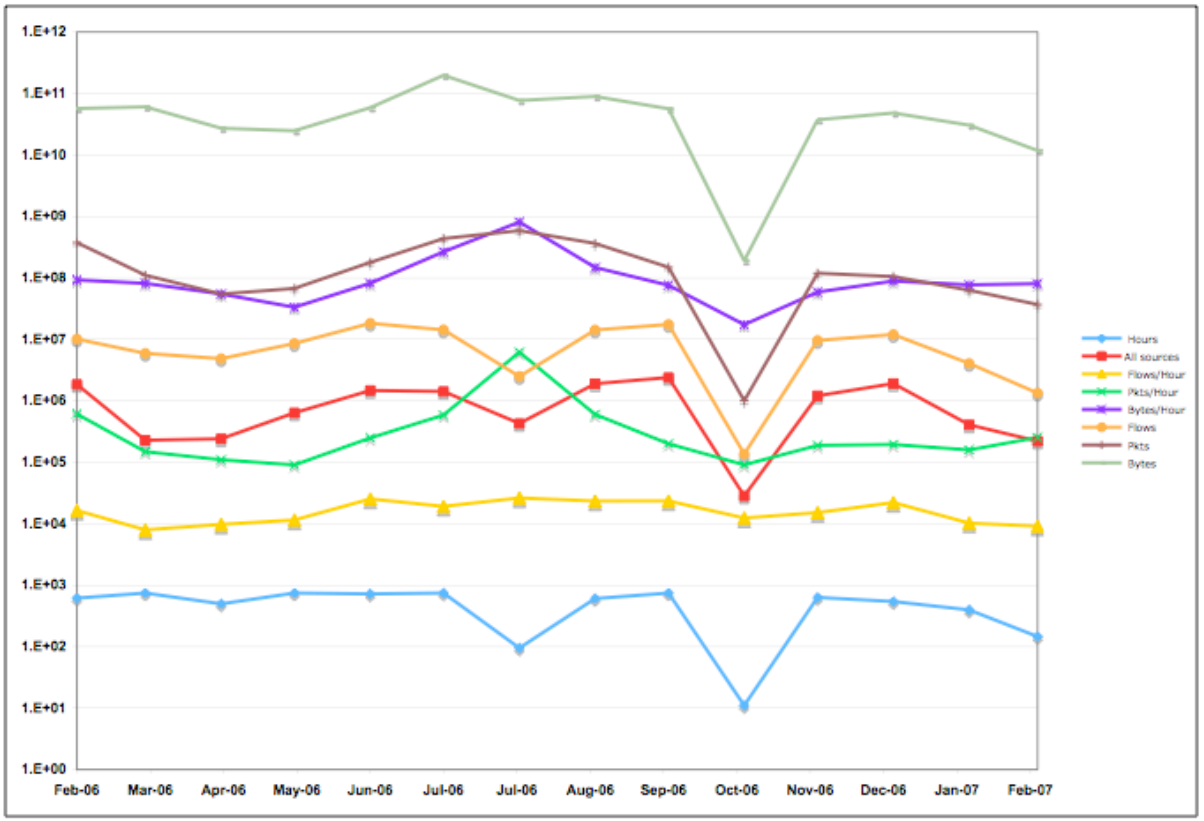
  rwsettool intersect VLF-f.set VLF-d.set > VLF-fd.set
- Can use sets to extract the flows

  rwfilter (dates, sIP in VLF-fd.set) > whatever

Counting active hours

- This is more subtle
  - Create hourly bags for sources active during hour
  - Extract cover set from bag and make new bag with count of 1 for each active IP (rwbagbuild does this)
  - Add the count 1 bags for all hours
  - Resulting bag has count of hours in which each IP in the index was active.

Overall Traffic Volumes



Overall Port Usage

	TCP				UDP			
	Source		Destination		Source		Destination	
Rank	Port	Flows	Port	Flows	Port	Flows	Port	Flows
1	80	4246669	4662	3724509	4672	13977848	4672	24050590
2	22	3941820	139	1208545	53	7078937	7986	12934591
3	443	2531082	25	1084206	5060	1116107	53	7607168
4	1863	1585616	445	1067113	6672	492530	27015	6167979
5	139	1439111	1433	882422	32459	408733	7717	5142715
6	25	1200603	3531	863342	3531	388498	41639	980437
7	110	899930	7986	611319	6881	382606	56604	973268
8	4662	843710	56604	604490	123	365249	137	918731
9	6881	839712	80	596298	1501	157110	3531	463953
10	3101	434613	5900	430174	4803	140593	2051	413205
11	32459	254030	41639	358494	7571	120201	123	341651
12	7000	145825	22	341478	500	114677	17552	328209
13	49152	95263	7717	317584	1025	109978	17537	209013
14	143	80971	3306	270549	4671	98553	17536	184636
15	6667	65359	17306	207693	5672	97867	25383	175525
16	3002	63347	25383	168570	1026	93360	1027	168986
17	6000	63053	6881	131921	32768	92854	1026	168106
18	48602	62289	135	129084	1027	87387	43635	167296
19	48601	62266	10000	113971	7890	84221	5060	160939
20	48603	62221	4899	109668	4673	80272	1069	159056

Per Host Flow Distribution

Flows	Sources	%	Cum. %	Flows	%	Cum %
1	6196337	48.74%	48.74%	6196337	5.03%	5.03%
2	2261307	17.79%	66.53%	4522614	3.67%	8.70%
3	1034589	8.14%	74.67%	3103767	2.52%	11.22%
4	648227	5.10%	79.77%	2592908	2.11%	13.33%
5	419928	3.30%	83.07%	2099640	1.70%	15.03%
6	310678	2.44%	85.52%	1864068	1.51%	16.55%
7	224984	1.77%	87.28%	1574888	1.28%	17.83%
8	184987	1.46%	88.74%	1479896	1.20%	19.03%
9	145821	1.15%	89.89%	1312389	1.07%	20.09%
10	126003	0.99%	90.88%	1260030	1.02%	21.12%
1...100	12616333	99.24%	99.24%	53706284	43.61%	43.61%
> 100	96127	0.76%	0.76%	69457047	56.39%	56.39%
Total	12712460	100.00%	100.00%	123163331	100.00%	100.00%

A small sample of 1 flow / host data

sIP	dIP	sPort	dPort	pro	packets	bytes	flags	sTime	dur
211.192.232.210	10.1.25.225	0	2048	1	1	28		2006/04/01T00:00:13.000	0.000
141.150.143.10	10.1.24.165	3359	27015	17	1	53		2006/04/01T00:03:14.000	0.000
4.227.17.37	10.1.24.202	4137	445	6	3	144	S	2006/04/01T00:07:48.000	9.000
202.170.99.73	10.1.25.118	27318	16729	6	1	52	S	2006/04/01T00:08:45.000	0.000
212.38.199.54	10.1.25.31	0	769	1	1	56		2006/04/01T00:11:10.000	0.000
216.123.177.141	10.1.24.69	4289	25	6	15	1518	FS PA	2006/04/01T00:15:55.000	6.000
172.156.249.160	10.1.24.165	3325	2701	17	1	53		2006/04/01T00:16:42.000	0.000
4.252.250.124	10.1.24.202	3950	445	6	3	144	S	2006/04/01T00:17:28.000	9.000
68.185.242.208	10.1.27.73	0	2048	1	1	92		2006/04/01T00:20:14.000	0.000
68.127.237.162	10.1.27.28	0	771	1	1	112		2006/04/01T00:22:37.000	0.000
172.190.94.218	10.1.24.165	3403	27015	17	1	53		2006/04/01T00:22:39.000	0.000
65.92.205.153	10.1.26.178	48106	25383	17	1	63		2006/04/01T00:23:23.000	0.000
219.107.35.199	10.1.26.132	2398	25	6	17	11890	FS PA	2006/04/01T00:24:45.000	4.000

A sample of matched flows

<->	Match#	sIP	dIP	sPort	dPort	pro	packets	bytes	flags	sTime
->		4.227.17.37	10.1.24.202	4137	445	6	3	144	S	2006/04/01T00:07:48.000
->		202.170.99.73	10.1.25.118	27318	16729	6	1	52	S	2006/04/01T00:08:45.000
->	247	216.123.177.141	10.1.24.69	4289	25	6	15	1518	FS PA	2006/04/01T00:15:55.000
<-	247	10.1.24.69	216.123.177.141	25	4289	6	10	772	FS PA	2006/04/01T00:15:55.000
->		4.252.250.124	10.1.24.202	3950	445	6	3	144	S	2006/04/01T00:17:28.000
->	271	219.107.35.199	10.1.26.132	2398	25	6	17	11890	FS PA	2006/04/01T00:24:45.000
<-	271	10.1.26.132	219.107.35.199	25	2398	6	14	825	FS PA	2006/04/01T00:24:45.000

Details by type of data

- We considered flows from sources that generate between 1 and 10 flows during the observation period, 2006/02 - 2007/03
  - Unmatched TCP
  - Unmatched UDP
  - Matched TCP
  - Matched UDP
  - ICMP
  - Other (anything else)

Unmatched TCP Flag combinations

Rank	Flags	Count	Percent
1	S	871028	92.4%
2	RA	22192	2.4%
3	R	16342	1.7%
4	SA	12103	1.3%

Unmatched TCP Port Usage

	VLF TCP Destination ports									
	1 flow		2 - 4 flows		5 - 7 flows		8 - 10 flows		1 - 10 flows	
Rank	Port	Flows	Port	Flows	Port	Flows	Port	Flows	Port	Flows
1	445	71991	4662	106979	4662	45348	4662	23681	4662	239705
2	4662	63697	445	55202	445	22390	445	15000	445	164583
3	25	13353	35372	25124	35372	12955	41639	9047	35372	58994
4	35372	13302	41639	16360	41639	11892	35372	7613	41639	42492
5	17306	8474	80	14326	25	8019	25	6279	25	41785
6	9272	6601	25	14134	80	5868	1433	5302	17306	26723
7	41639	5193	17306	10241	24263	5665	7986	4693	80	26517
8	139	4960	9272	9920	1433	5586	6881	4649	24263	22113
9	24263	3758	24263	8918	6881	5559	25383	4369	9272	21110
10	7717	2509	6881	8372	7986	4814	80	3996	139	20636
11	80	2327	139	7380	17306	4769	7717	3931	6881	20051
12	40987	2126	7717	6768	139	4613	24263	3772	1433	18960
13	1433	1964	25383	6532	25383	4592	139	3683	7717	17676
14	47190	1872	40987	6486	7717	4468	17306	3239	25383	17266
15	25383	1773	1433	6108	40987	3499	34001	2119	7986	16156
16	6662	1739	7986	5251	47190	3209	47190	2083	40987	13615
17	34001	1715	47190	4775	9272	3167	40987	1504	47190	11939
18	3127	1593	113	3730	34001	2681	9272	1422	34001	9893
19	2967	1580	34001	3378	14662	1497	3531	1279	6662	6858
20	23	1538	6662	3323	5900	1390	56604	1016	14662	6662

TCP Flags for top 10 ports at 1 flow / host

	Port 445		Port 4662		Port 25		Port 35372		Port 17306	
Rank	Flags	Count	Flags	Count	Flags	Count	Flags	Count	Flags	Count
1	S	71900	S	63634	S	9584	S	13296	S	8438
2	R	49	SR	33	FSA	3065	R	4	FS	10
3	SR	34	R	20	FSRA	433	SR	2	SR	10
4	SRA	7		6	SRA	187			SRA	7
5	RA	1	RA	2	R	35			RA	6
6			SA	1	SA	18			R	3
7			SRA	1	SR	17				
8					RA	8				
9					A	4				
10					FA	2				

	Port 9272		Port 41639		Port 139		Port 24263		Port 7717	
Rank	Flags	Count	Flags	Count	Flags	Count	Flags	Count	Flags	Count
1	S	6599	S	5016	S	4955	S	3745	S	2457
2	SR	2	FS	109	SR	3	FS	11	R	18
3			R	34	R	2	SRA	2	SR	12
4			SR	22					FS	10
5			RA	8					RA	10
6			SRA	4					SRA	2

Port Usage for top 10 unmatched TCP

- 445 Used for Microsoft file sharing. This service has been associated with a long series of vulnerabilities.
- 4662 Service port for the eDonkey2000 peer to peer system.
- 25 SMTP Email.
- 35372, 17306, 9272, 41639 No information available.
- 139 Netbios session service. There is a long history of vulnerabilities and exploits associated with this port.
- 24263, 7717 No information available.

Where do they go (1 flow per host)?

Rank	Port	Flows	Hit	Miss	% Hit	# Dst	Dst Hit	% Dst Hit
1	445	71991	20981	51010	29.14	1006	109	10.83
2	4662	63697	49352	14345	77.47	17	5	29.41
3	25	13353	9535	3818	71.40	116	15	12.93
4	35372	13302	13302	0	100.00	1	1	100.00
5	17306	8474	8474	0	100.00	1	1	100.00
6	9272	6601	0	6601	0.00	1	0	0.00
7	41639	5193	5193	0	100.00	1	1	100.00
8	139	4960	418	4542	8.42	942	96	10.19
9	24263	3758	3758	0	100.00	1	1	100.00
10	7717	2509	2509	0	100.00	2	2	100.00
11	80	2327	306	2021	13.14	812	82	10.09
12	40987	2126	2126	0	100.00	1	1	100.00
13	1433	1964	148	1816	7.53	782	66	8.43
14	47190	1872	0	1872	0.00	1	0	0.00
15	25383	1773	1773	0	100.00	1	1	100.00
16	6662	1739	1636	103	94.07	3	1	33.33
17	34001	1715	0	1715	0.00	1	0	0.00
18	3127	1593	173	1420	10.86	624	66	10.57
19	2967	1580	127	1453	8.03	783	64	8.17
20	23	1538	149	1389	9.68	766	77	10.05

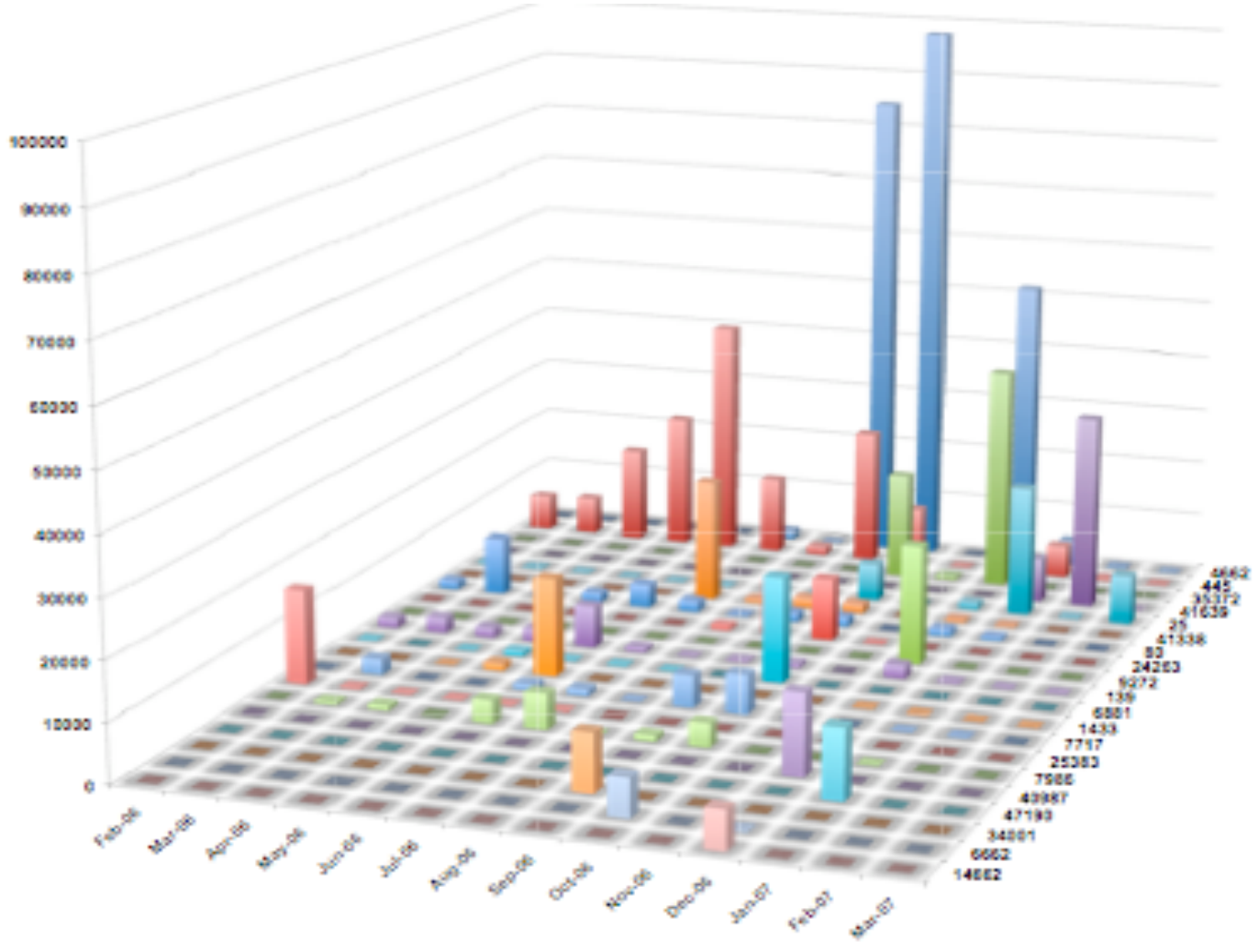
Where do they go (1-10 flows / host)?

Rank	Port	Flows	Hit	Miss	% Hit	# Dst	Dst Hit	% Dst Hit
1	4662	239705	174522	65183	72.80	18	6	33.33
2	445	164583	33487	131096	20.34	1013	110	10.85
3	35372	58994	58994	0	100.00	1	1	100.00
4	41639	42492	42492	0	100.00	1	1	100.00
5	25	41785	32185	9600	77.02	817	74	9.05
6	17306	26723	26718	5	99.98	3	1	33.33
7	80	26517	2693	23824	10.15	1012	109	10.77
8	24263	22113	22113	0	100.00	1	1	100.00
9	9272	21110	0	21110	0.00	2	0	0.00
10	139	20636	1850	18786	8.96	1013	110	10.85
11	6881	20051	19998	53	99.73	23	6	26.08
12	1433	18960	1370	17590	7.22	984	81	8.23
13	7717	17676	17676	0	100.00	2	2	100.00
14	25383	17266	17266	0	100.00	1	1	100.00
15	7986	16156	16156	0	100.00	2	2	100.00
16	40987	13615	13615	0	100.00	1	1	100.00
17	47190	11939	0	11939	0.00	1	0	0.00
18	34001	9893	0	9893	0.00	1	0	0.00
19	6662	6858	6575	283	95.87	3	1	33.33
20	14662	6662	0	6662	0.00	3	0	0.00

The destinations are a little strange

- The network is about 10% populated
  - Would expect about 10% hit rate for random addressing, but
- Many probes have a 100% hit rate on a small number of destinations
- Others have a 0% hit rate
- This is unmatched traffic, so none of these were answered. One would like to know the source of the intelligence that resulted in this targeting, especially for the Uncommon ports.

When did they happen?



Temporal distribution

- Many of the ports appear to cluster in relatively small time intervals.
- Several appear to be localized to single months
- Others grow and fall over a fairly short period.
- None exhibit a regular, stedy presence.

Geographical Distribution

- Using MaxMind, we obtained the country code distribution for these ports.
- The US is the primary source for many, but the remaining rankings vary widely.

Geographical distribution

	4662			445			35372			41639			25		
R	CC	cnt.	pct.	CC	cnt.	pct.	CC	cnt.	pct.	CC	cnt.	pct.	CC	cnt.	pct.
1	fr	84185	35.1%	us	39873	24.2%	fr	24515	41.6%	us	10679	25.1%	us	7534	18.0%
2	es	40720	17.0%	tw	31387	19.1%	es	8055	13.7%	ca	5639	13.3%	pl	2624	6.3%
3	de	17376	7.2%	fr	13379	8.1%	de	4261	7.2%	gb	3768	8.9%	it	2082	5.0%
4	it	15883	6.6%	jp	11119	6.8%	it	3684	6.2%	de	1500	3.5%	kr	2050	4.9%
5	il	12753	5.3%	de	10624	6.5%	us	2463	4.2%	se	1454	3.4%	fr	2013	4.8%
6	us	9145	3.8%	pl	8716	5.3%	il	2324	3.9%	pl	1284	3.0%	es	2006	4.8%
7	br	6637	2.8%	gb	7990	4.9%	br	1316	2.2%	au	1194	2.8%	cn	1990	4.8%
8	pl	5479	2.3%	cn	4557	2.8%	pl	1245	2.1%	nl	1142	2.7%	de	1741	4.2%
9	ar	4469	1.9%	ca	3474	2.1%	cn	1122	1.9%	my	1132	2.7%	br	1615	3.9%
10	cn	4301	1.8%	es	2314	1.4%	kr	1046	1.8%	fr	1116	2.6%	ru	1578	3.8%
T		239705			164583			58994			42492			41785	

	41338			80			24263			9272			139		
R	CC	cnt.	pct.	CC	cnt.	pct.	CC	cnt.	pct.	CC	cnt.	pct.	CC	cnt.	pct.
1	us	6043	22.6%	us	7579	28.6%	us	14390	65.1%	fr	17660	83.7%	kr	4437	21.5%
2	mx	4233	15.8%	cn	6675	25.2%	cn	2782	12.6%	us	1104	5.2%	us	3620	17.5%
3	cl	2763	10.3%	kr	2015	7.6%	gb	1111	5.0%	be	679	3.2%	tw	1180	5.7%
4	ar	1724	6.5%	ca	1438	5.4%	au	385	1.7%	ch	312	1.5%	jp	919	4.5%
5	br	1589	5.9%	hk	964	3.6%	fr	357	1.6%	es	192	0.9%	cn	916	4.4%
6	es	1275	4.8%	tw	741	2.8%	jp	310	1.4%	ca	144	0.7%	fr	905	4.4%
7	pe	1127	4.2%	br	724	2.7%	nl	258	1.2%	il	102	0.5%	gb	901	4.4%
8	ca	1105	4.1%	jp	712	2.7%	br	220	1.0%	pt	85	0.4%	de	637	3.1%
9	co	795	3.0%	ro	545	2.1%	de	210	0.9%	be	83	0.4%	ca	617	3.0%
10	tr	783	2.9%	gb	537	2.0%	–	157	0.7%	ma	75	0.4%	br	458	2.2%
T		26723			26517			22113			21110			20636	

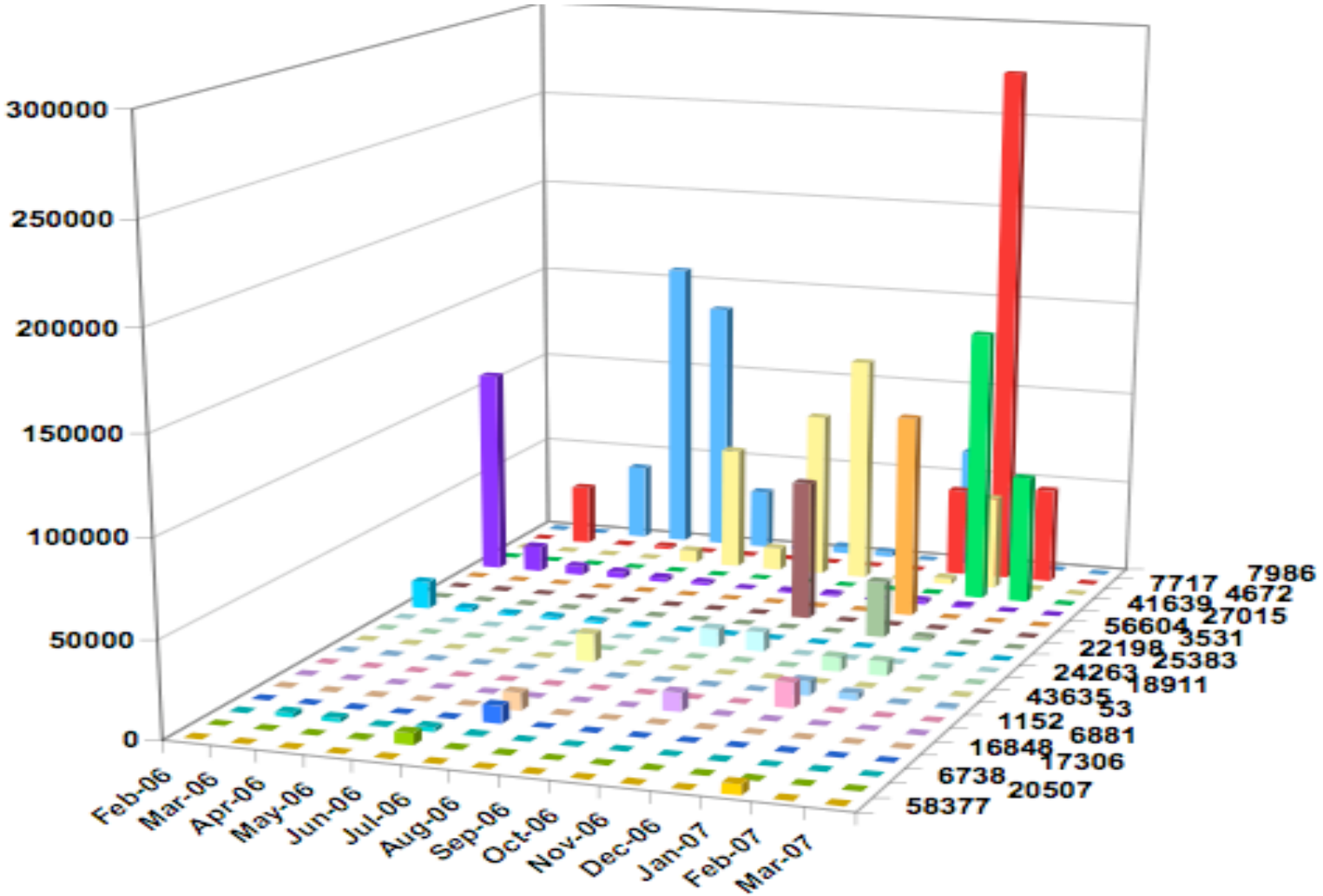
UDP per port volumes (1 - 10 flows / ports)

	VLF UDP Destination ports									
	1 flow		2 - 4 flows		5 - 7 flows		8 - 10 flows		1 - 10 flows	
Rank	Port	Flows	Port	Flows	Port	Flows	Port	Flows	Port	Flows
1	7986	127091	7986	158455	7986	105951	7986	88392	7986	479889
2	4672	111760	7717	141481	7717	88210	7717	68927	7717	410036
3	7717	111418	4672	115523	4672	68624	4672	51642	4672	347549
4	41639	89806	41639	72769	41639	31192	3531	43706	41639	213221
5	27015	50098	27015	53700	27015	23031	41639	19454	27015	140354
6	56604	46000	56604	33964	3531	19169	27015	13525	56604	107250
7	1152	10951	22198	11536	56604	17057	56604	10229	3531	73635
8	25383	7100	3531	10708	22198	7589	22198	5590	22198	31315
9	22198	6600	25383	7685	24263	4930	53	3537	25383	22255
10	43635	5170	24263	7628	53	4698	24263	3353	24263	20815
11	24263	4904	43635	5749	25383	4307	25383	3163	18911	15621
12	18911	4837	18911	5558	18911	3230	6881	3012	43635	14932
13	6738	3917	53	4418	6881	3154	18911	1996	53	13964
14	10413	3631	6881	3930	43635	2437	43635	1576	1152	13252
15	34718	3561	17306	3428	16848	1611	16848	1296	6881	10100
16	20507	3381	1026	2989	17306	1511	137	980	16848	9173
17	16848	3348	16848	2918	137	1193	17306	955	17306	8965
18	17306	3071	1028	2775	6738	986	21715	837	6738	7744
19	1027	2055	10484	2531	58377	891	6738	618	20507	7070
20	58377	2014	20507	2424	10484	809	58377	609	58377	5513

Top 10 UDP ports

- 7986, 7717 No information
- 4672 eMule / eDonkey P2P software. (4662 TCP)
- 41639 (4) No information
- 27015 HalfLife game server. During the first month of this capture, a compromised host supported a HalfLife server (several million flows per hour). Substantial connection attempts still made to the address. Server removed over two years ago.
- 56604 No information
- 3531Joltid PeerEnabler, a P2P download system
- 22198, 25383 (14), 24263 (8) No information

UDP Temporal distribution



Hit Miss for unmatched UDP

Rank	Port	Flows	Hit	Miss	% Hit	# Dst	Dst Hit	% Dst Hit
1	7986	479889	479889	0	100.00	2	2	100.00
2	7717	410036	410036	0	100.00	2	2	100.00
3	4672	347549	331845	15704	95.48	1016	113	11.12
4	41639	213221	213221	0	100.00	1	1	100.00
5	27015	140354	139317	1037	99.26	4	1	25.00
6	56604	107250	107250	0	100.00	1	1	100.00
7	3531	73635	73635	0	100.00	1	1	100.00
8	22198	31315	31315	0	100.00	1	1	100.00
9	25383	22255	22255	0	100.00	1	1	100.00
10	24263	20815	20815	0	100.00	1	1	100.00
11	18911	15621	15621	0	100.00	2	2	100.00
12	43635	14932	14932	0	100.00	2	2	100.00
13	53	13964	12248	1716	87.71	631	76	12.04
14	1152	13252	13252	0	100.00	1	1	100.00
15	6881	10100	10099	1	99.99	3	2	66.66
16	16848	9173	9173	0	100.00	1	1	100.00
17	17306	8965	8965	0	100.00	1	1	100.00
18	6738	7744	7744	0	100.00	3	3	100.00
19	20507	7070	7070	0	100.00	1	1	100.00
20	58377	5513	5513	0	100.00	1	1	100.00

Hit Miss results

- Most of the traffic is directed to a single host that does exist.
- The UDP traffic is mostly highly targeted and based on solid information.
- Only ports 4672 and 53 appear to involve scanning

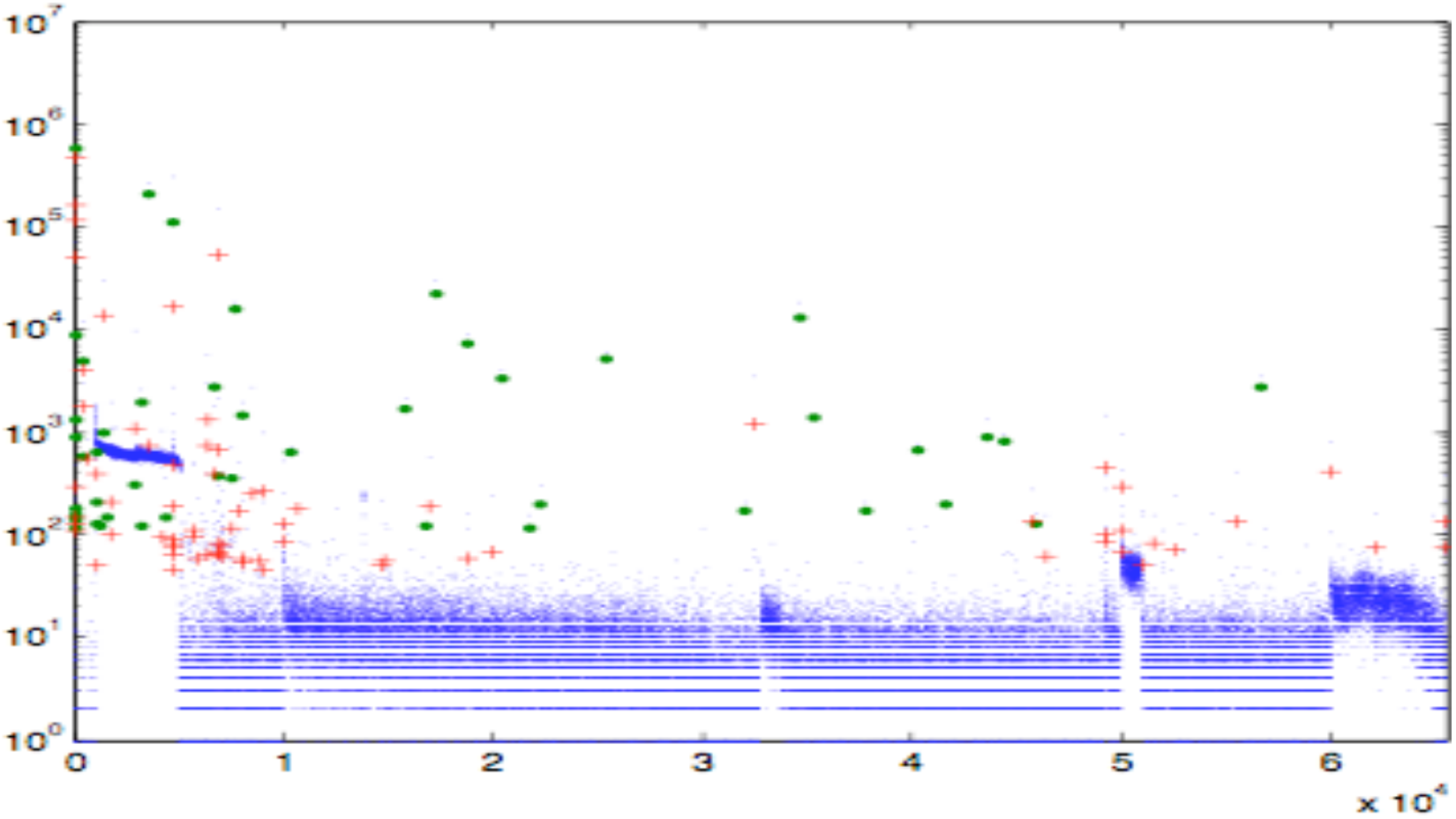
Matched TCP and UDP

- By definition, there are no misses in the match data
- Since we are unable to reliably determine the origin of the flows, we look at both the source and the destination ports in determining port frequencies.
- This means that we are looking at some connections that originate within the enterprise. Ports with high E counts may indicate scanning from inside.
- We plot overall port usage for low volume flows and note interesting differences between TCP and UDP
- Anecdotally, we see some interesting examples of “Port Numerology”

Top 20 Matched TCP ports (1 - 10 Flows)

Rank	Port	O/E	Sessions	O count	E count
1	25	O	567934	436562	33
2	139	E	483108	324742	5
3	3531	O	208094	54573	9
4	22	E	165579	141768	27
5	25	E	113514	67999	10
6	4662	O	112027	66693	11
7	6881	E	54278	31818	14
8	80	E	51386	17423	89
9	17306	O	22255	11802	1
10	4662	E	16401	13750	7
11	7717	O	16050	7203	2
12	1433	E	13378	13373	4
13	34718	O	12958	11468	3
14	113	O	8691	6822	7
15	18837	O	7228	6392	1
16	25383	O	5118	2897	1
17	445	O	4939	3772	61
18	445	E	4125	3960	2
19	20507	O	3352	2294	1
20	6738	O	2762	1797	2

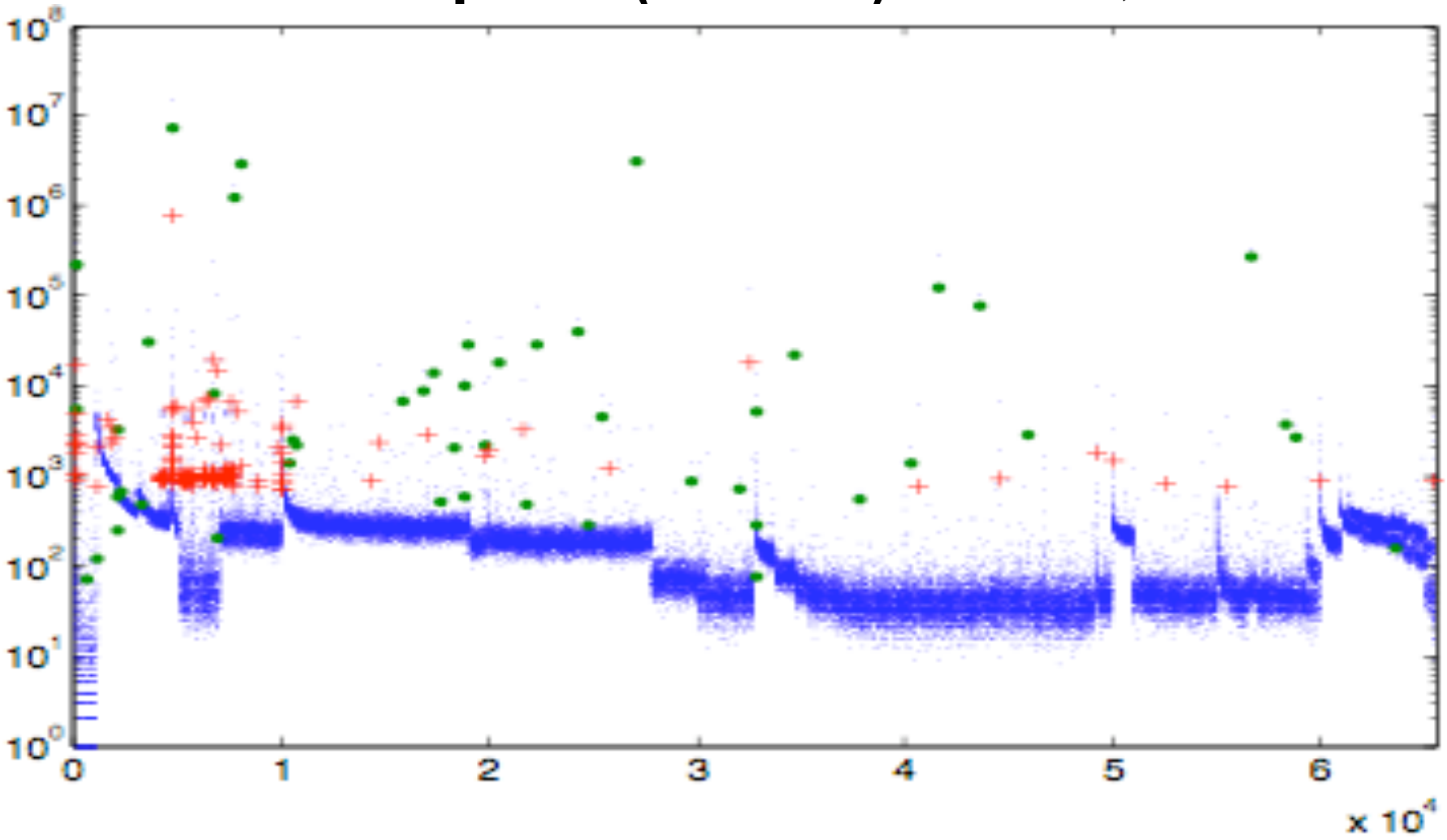
Overall TCP ports (1 - 10 F) • OtoE, + EtoO



Top 20 Matched UDP ports (1 - 10 Flows)

Rank	Port	O/E	Sessions	O count	E count
1	4672	O	7314744	3948005	2
2	27015	O	3194004	1566271	1
3	7986	O	2858847	1261078	2
4	7717	O	1206950	617366	2
5	4672	E	778674	588597	7
6	56604	O	263568	170647	1
7	53	O	227486	97813	3
8	41639	O	120483	96298	1
9	43635	O	75282	46791	2
10	24263	O	38837	37762	1
11	3531	O	31570	17967	1
12	22198	O	29022	19260	1
13	18911	O	28881	19383	2
14	34718	O	22515	21732	2
15	6672	E	18816	15002	4
16	20507	O	18681	16806	1
17	32459	E	17950	9655	13
18	53	E	17212	13558	12
19	6881	E	15339	10636	13
20	17306	O	14236	7703	1

Overall UDP ports (1 - 10 F) • OtoE, + EtoO



Port Numerology

- If we look at the relative frequencies of ports in the low frequency matched data, we see a noticeable non-uniformity in the port distribution. This was first noticed in scanning the list of ports, in order, and noticing that port numbers with usage counts that were slightly higher than their neighbors were interesting, in a sense. Sorting by count makes this somewhat clearer.
- For example, in the range from 4700 to 4850 counts, we find mostly ephemeral ports from the 1000, 4000, 5000, and 6000 region, along with some others.

Interesting port numbers (UDP)

15001	4840
52525	4710
65535	4623
10006	4596
55555	4507

- These are mixed in with a fairly typical run of ephemeral port numbers
- Counts are an order of magnitude higher than adjacent ports in some cases.
- The pattern seems to repeat at a smaller scale with local maxima being generally more “interesting” than their neighbors - 0s, palindromes, patterns, etc.

ICMP

Type	Code	Count	Rank	Meaning
3	3	419276	1	Port Unreachable
3	1	77024	3	Host Unreachable
3	13	67053	4	Communication Administratively Prohibited
3	0	16835	6	Net Unreachable
3	10	1377	8	Communication with Destination Host is Administratively Prohibited
3	4	803	10	Fragmentation Needed and Don't Fragment was Set
3	2	322	13	Protocol Unreachable
3	9	163	14	Communication with Destination Network is Administratively Prohibited
3	7	4	42	Destination Host Unknown
3	44	2	100	(Invalid code)
3	6	1	171	Destination Network Unknown
3	255	1	222	(Invalid code)
8	0	86170	2	Echo
11	0	53593	5	Time to Live exceeded in Transit
11	1	1231	9	Fragment Reassembly Time Exceeded
5	1	1618	7	Redirect Datagram for the Host
5	0	65	16	Redirect Datagram for the Network (or subnet)
4	0	532	11	Source Quench
0	0	486	12	Echo Reply
14	0	15	24	Timestamp Reply
12	0	15	25	Parameter Problem - Pointer indicates the error

ICMP

- In addition to the ICMP messages seen in the table, there are quite a few with malformed message and code fields.
- One deficiency of NetFlow with respect to ICMP is its inability to capture the offending IP address for many types of messages.
- This makes it impossible to accurately match inbound ICMP with the outbound message that provoked it, although several of my students have had some success in this area.

Other Protocols

P'col	CC	Flows	Packets	Bytes	Description
0	–	52	112332112	3454856064	IPv6 Hop-by-Hop Option
12	–	2	1074960	2752300710	PUP (Xerox)
25	–	2	284672	91103936	LEAF-1
41	us	34	34	3656	IPv6
46	us	10	10	2228	RSVP Reservation Protocol
50	ca	156	116382	29048394	ESP Encap Security Payload
54	–	2	16456	13693964	NBMA Address Resolution Protocol
64	–	14	108053086	11113219624	SATNET and Backroom EXPAK
89	–	2	98461696	5746163040	OSPFIGP
96	–	16	2854160	866707052	Semaphore Communications Sec. Pro.
128	–	6	75688960	9031922944	SSCOPMCE
132	cl ru	4	8	24	Stream Control Transmission Protocol
140	–	2	67117056	4302832512	Unassigned
153	–	2	90466	28960428	Unassigned
160	–	2	149248	57315896	Unassigned
192	–	32	177557758	13644593568	Unassigned
255	jp	6	6	360	Reserved
TOTAL		344	643797070	51132724400	

Other protocols

- There are a handfull of other protocols. The ESP traffic is probably legitimate. Many of the others are probably bogus, due to collector malfunctions or the like as the packet and byte counts are implausible.
- We know of at least one internal “protocol scan” in which every protocol from 0 to 255 was seen, despite the fact that most are undefined or reserved.

Malformed traffic

- The match program found several near matches similar to this example.

<->	Match#	sIP	dIP	sPort	dPort	pro	packets	bytes	flags
->		207.237.229.243	10.1.26.178	31185	25383	6	1	60	S
<-		10.1.26.178	207.237.229.243	25383	1965	6	1	48	S A

Future work-Credits

- The current work is exploratory. we would like to see if we can predict IPs that will not be seen again to keep them out of the “active” state that is being tracked and analyzed.
- We would also like to understand the details of what we have found, especially the transient peaks of port specific activity.
- The work reported here was partially funded by the CSE.