Network Awareness and Network Security
or
You don't have to work here to be paranoid, but it helps.

TRUST Seminar, Berkeley, 3 May 2007

John McHugh
Canada Research Chair in Privacy and Security
Director, Privacy and Security Laboratory
Dalhousie University
mchugh@cs.dal.ca

© 2005-2006 by John McHugh

Enterprise networks are complex

- Most system administrators are unsure of the configuration of their networks and the machines on it at any given time.
  - The distinction between program and data becomes ever more blurred with time
    - The blessing and curse of Von Neumann
  - You can no longer control the introduction of executable code.
    - Try turning off browser scripting and see how useful the browser is.
- Continuous, passive, observation provides insight.

© 2005,6,7 by John McHugh

A little background

- I've been in computer security for about 30 years, more if you count from the first machine I broke.
- I started doing high assurance systems in the 1980s after spending 20 years doing other things.
- I drifted from program verification to covert channel analysis, to security engineering and finally fetched up at CERT in 1999.
- Suresh Konda at CERT was looking for a source of network data. We interviewed a candidate for an analyst position who told us some interesting things ...

© 2005,6,7 by John McHugh

Differential charging and serendipity

- The candidate's organization is essentially a very large ISP. They were collecting NetFlow, inbound, with the idea that they might charge their users more for some protocols than others.
- The candidate had sorted some of the data by source IP and found an interesting hot spot.
- Someone was systematically scanning the ISP's address space, but at a rate that was on the order of 1 probe per /24 per hour.
  - There were enough /24s to make this stand out
  - We didn't get the analyst, but we sent a proposal to his former employer. The rest is history.

© 2005,6,7 by John McHugh

Let attackers do your work

- It is relatively simple to monitor the traffic that passes between your network and the internet.
- Monitoring internal networks is slightly more complex
- Even flow level data is useful in confirming expected services and finding unexpected ones.
  - You didn't know you run a music swapping service?
  - How about a world class Half Life server?
- A bit of analysis during collection can create a fairly complete inventory of hosts and their software, including applications you didn't know you ran.
- *But first a word from our ~~sponsor~~ friends.*

© 2005,6,7 by John McHugh

Advantages of Passive Mapping

- Even unused hosts are active on the network
  - Scanning is rampant on large networks
  - Between 1/3 and 2/3 of all network transactions are scan related
- Every TCP session contains information that identifies a host O/S
- Server and client applications broadcast their identity
- Spyware sends damaging information in the clear
  - Adversary perspective: hosts TRICKLER identifies already belong to an adversary!
- Why scan our networks when everyone else is doing it for us?

© 2005,6,7 by John McHugh

How do we do this?

- Leaving aside the TRICKLER component,
  - Cisco and other routers are capable of collecting NetFlow data and sending it to an archiving and analysis facility.
  - Other tools can do similar collection:
    - softflowd and YAF (next slides)
  - The SiLK tools from CERT allow analysis of flow level data.
    - Run on any UNIX (Windows is problematic using Cygwin)
    - Provide flexible analysis framework
    - Examples to follow

© 2005,6,7 by John McHugh

A little terminology

- Internet traffic consists of packets sent between hosts
- Packets contain addressing, service info, and data.
  - Sessions aggregate packets with common address and service information
- We can capture full packets, packet headers, or session information.
  - Each has advantages and disadvantages wrt space, analysis detail, privacy, etc.
- NetFlow lies between session and header in detail
  - Originally designed for accounting but useful for many other purposes.

© 2005,6,7 by John McHugh

Netflow probes: softflowd

- Softflowd semi-statefully tracks traffic flows. Upon expiry of a flow, reports to a designated collector host using the standard NetFlow protocol.
- Export using NetFlow version 1, 5 or 9 datagrams, IPv6 capable. Any standard NetFlow collector should be able to process the reports from softflowd.
- Designed to minimize host load; can read pcap files
- Normally runs as a daemon with "remote control" program (softflowctl) which allows runtime control and extraction of statistics.
- http://www.mindrot.org/projects/softflowd/
- Runs on Linux / OpenBSD / some Solaris support.

© 2005,6,7 by John McHugh

Netflow probes: YAF (from CERT)

- YAF is Yet Another Flow sensor. It processes packet data from pcap(3) dumpfiles / interface into bidirectional flows, then exports those flows to IPFIX Collecting Processes or the SiLK collector in a format that includes TCP flag extensions. Bidirectional causes time problems.
- YAF also supports partial payload capture - this feature is intended for use in "banner grabbing" for protocol verification.
- YAF is primarily intended to track developments in the IETF IPFIX working group, specifically bidirectional flow representation and archival storage formats.

© 2005,6,7 by John McHugh

SiLK

- Developed at CERT by late Suresh L. Konda
- Allows efficient archive and analysis of large flow sets
  - Primary customer collects 50-100GB/D flow
    - Don't try to tell me your network is too big
- NOT an IDS, supports retrospective analysis
- Data organization allows parallel access
- Allows filtering on flow properties
- Sets and multisets for analysis abstractions
- Scriptable for routine analyses, reports
- Results can feed into graphical displays.
- Get SiLK at http://tools.netsa.cert.org/

© 2005,6,7 by John McHugh

SiLK - filtering and display

- `rwfilter` is a command line program that will:
  - Select appropriate files from the archive
  - Extract data based on selection criteria such as
    - address and port information - ranges, sets
    - protocol, volume parameters, time, duration
    - sensor contributing the data
  - Data can be partitioned into pass / fail outputs
  - Filter output can be further filtered or used as input to other programs
- `rwcut` produces tabular output from binary flow files

© 2005,6,7 by John McHugh

SiLK aggregating and display

- `rwset`, `rwbag` will build sets, multisets of source or destination addresses in a flow file.
- `rwbag` will also build multisets for ports and protocols
  - count can be based on flows, packets, bytes
- sets and bags can be operated on (union, intersection, addition, etc.
- sets and bags can be displayed in a variety of ways
  - listings including network structure
  - bags can be inverted and binned by count
- Sets and bags become a way of thinking about network activity.

© 2005,6,7 by John McHugh

Example - large scale scanners on a /16

```
rwfilter --type=out --proto=6 --saddress=x.y.0.0/16 ... |          \
    rwset --sip-file=actve.set --dip-file=tgt.set
```

actve approximates the network population,

tgt is the set of outside machines contacted /answered

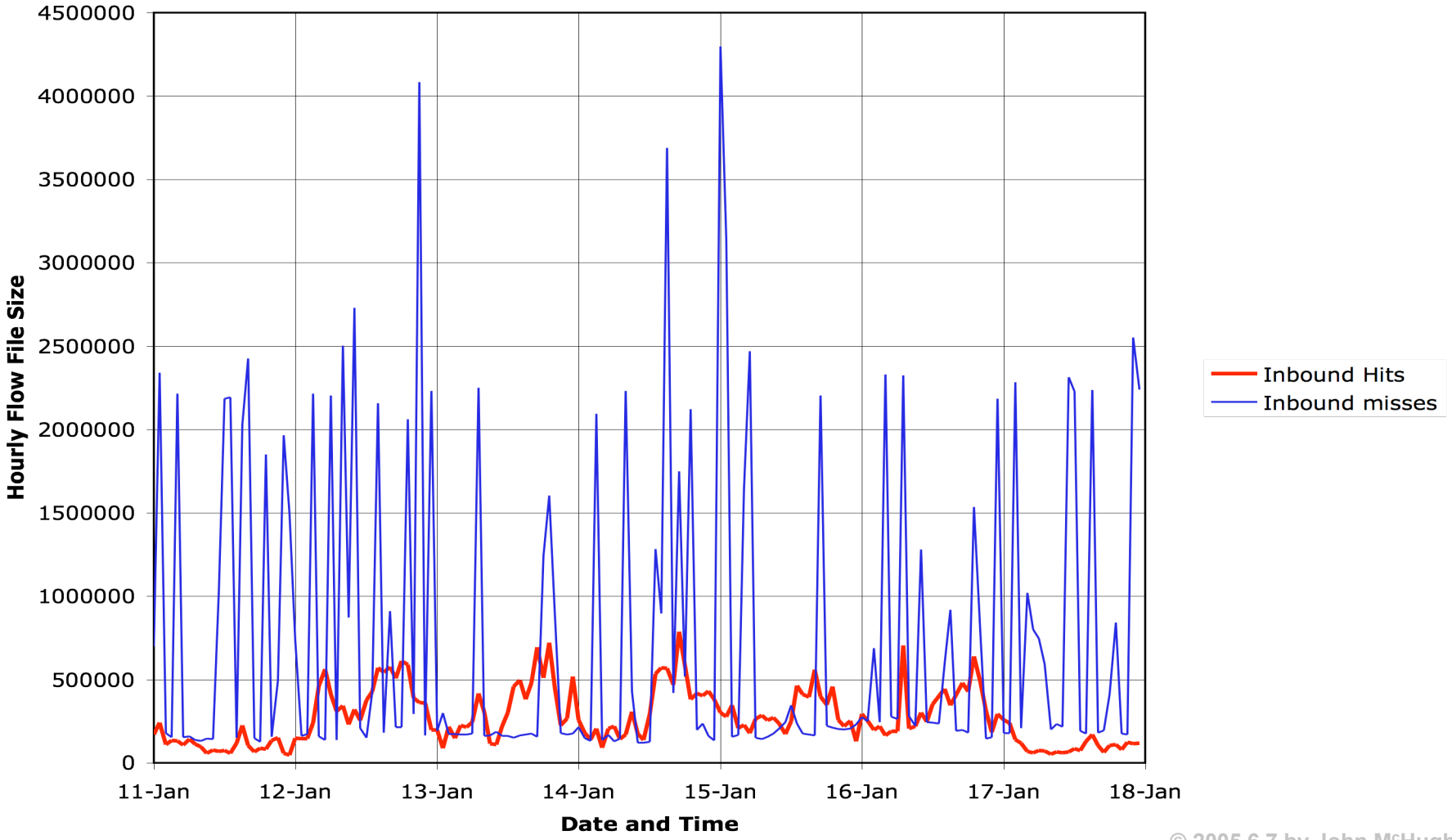
Then for each hour in the week hh do

```
rwfilter --type=in --proto=6 --dipset=actve.set \
         --pass=hit-${hh}.rwf --fail=miss-${hh}.rwf
```

See plot on next page. File size is proportional to number of flows.

© 2005,6,7 by John McHugh

One week on a /16



© 2005,6,7 by John McHugh

Who are the scanners?

Call any source that sends 100+ flows to inactive addresses in an hour a scanner

```
rwbag --sip-flows=stdout miss-${hh}.rwf |\
    rwbagtool --mincount=100 --coverset --output=scnrs.set
```

Find the addresses that interacted with a scanner

```
rwsetintersect --add-set=tgt.set --add-set=scnrs.set \          --
    set-file=intact.set
rwfilter --dipset=intact.set (as at top) ... | rwset --saddress --
    print-ips
```

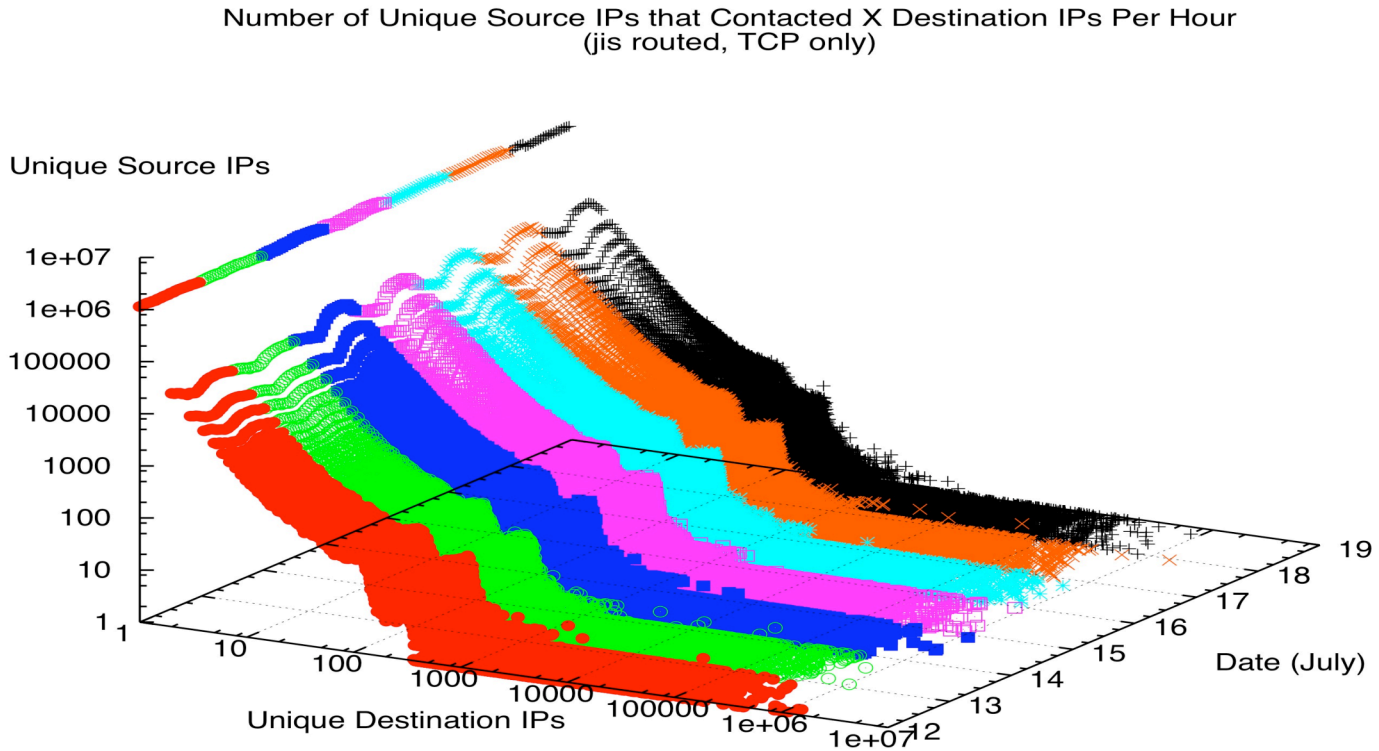
© 2005,6,7 by John McHugh

Very large scale observation

- Carrie Gates was interested in the degree of fan out from outside to inside for her scan detection work.
- How many outside hosts use exactly one inside host / service pair. (unique destination address/port)
- Bloom filters can be used to find unique sIP,dIP,dport exemplar flows
- If we make a source IP bag from the flows, the counts will be the number of different host / service pairs used.
- Invert the bag to determine how many entries have a count of 1, 2, 3, ... maxcount. Plot hourly results for a week

© 2005,6,7 by John McHugh

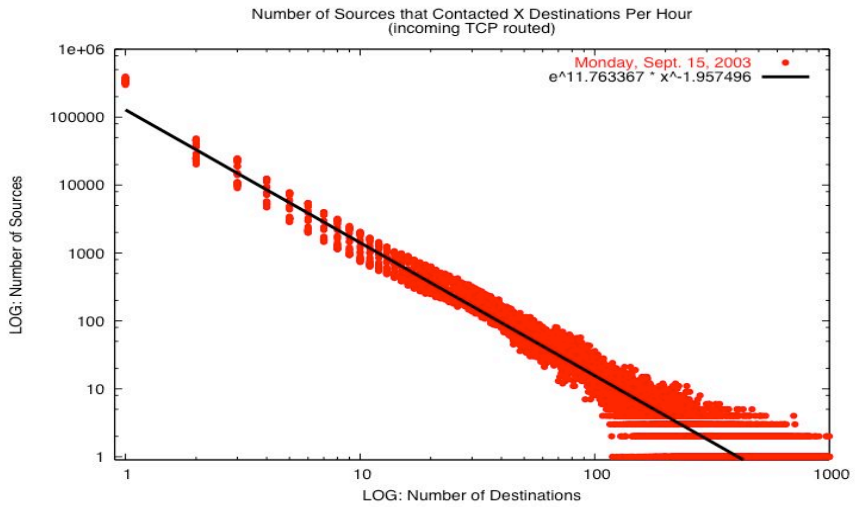
Outside to inside - July 2003



© 2005,6,7 by John McHugh

Developing the contact surface

- In the absence of the disturbance seen on the previous page, contact lines seem to follow a power law type of distribution
  - or do they[1].
  - I think this is really at least 3 separate processes
    - VLF noise
    - "normal activity"
    - Bulk scanning



[1] everything is a straight line on log/log paper, especially if you use a fat marker

© 2005,6,7 by John McHugh

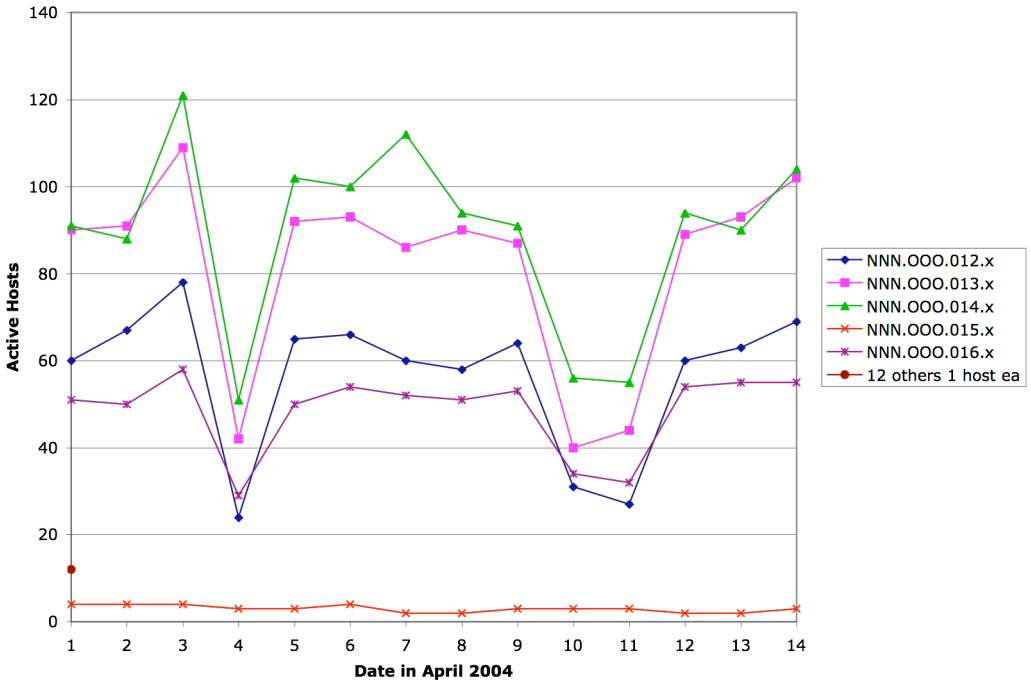
Internet wide disturbance

- The ripple in what would otherwise be a fairly straight log/log plot of connectivity was observed from at least Jan - Aug 2003.
- It went away when Blaster appeared in Aug 2003.
- A similar ripple existed from Feb 11 to May 31 2004 coinciding with the lifetime of Welchia-B
  - In this case, the ripple is due to a few hundred machines scanning at a low, fixed, rate induced by a loop with a “sleep” system call to induce a fixed scanning rate.
- In both cases, they persisted until killed, not patched.

© 2005,6,7 by John McHugh

Active host counts

- Simply counting hosts that are active is interesting
  - Active means generating traffic from an address
  - These are daily, but hourly is also possible



© 2005,6,7 by John McHugh

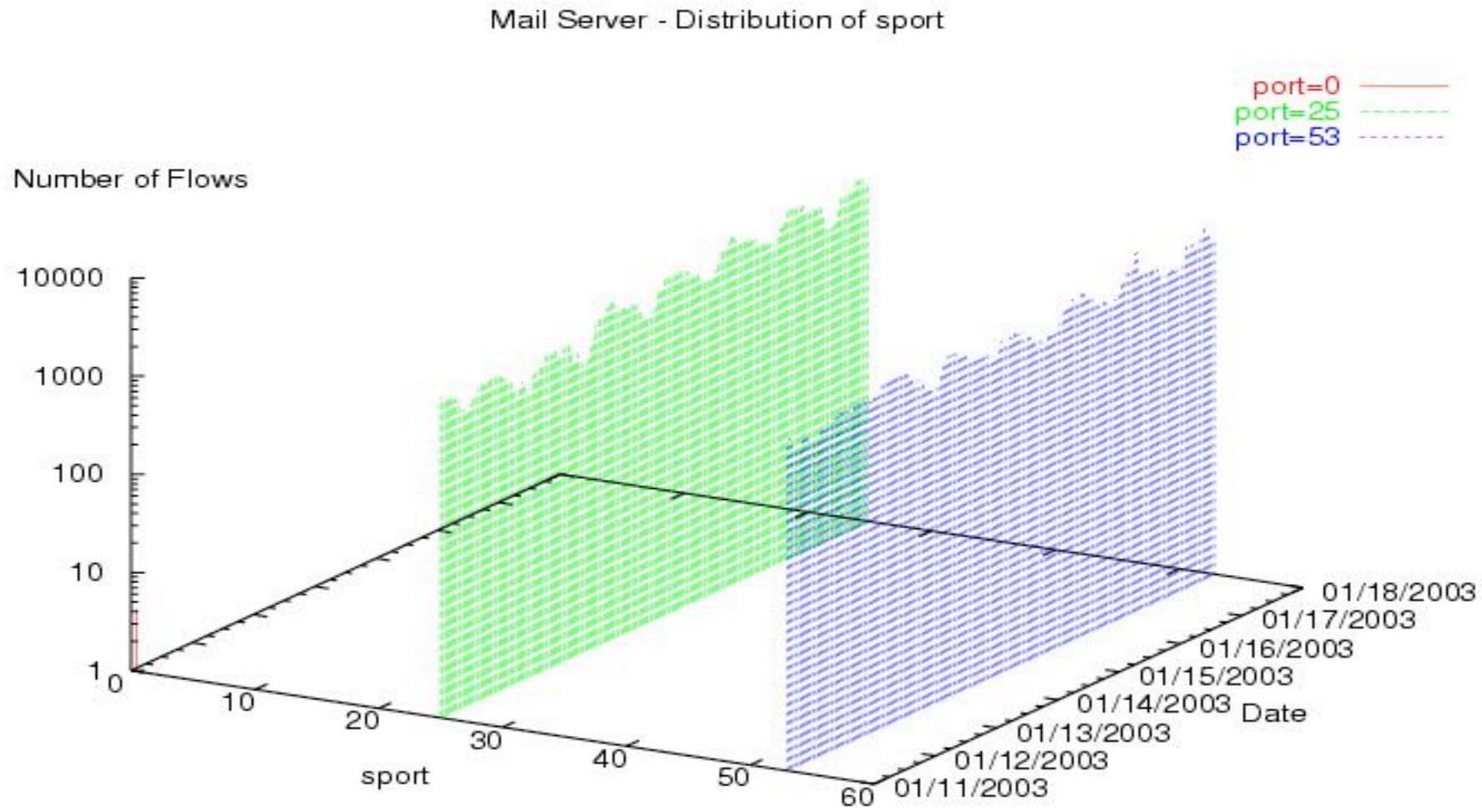
Host characterization

- We can characterize an individual host by a plot showing port activity as a function of time.
- The log scale allows small and large flows to be seen
- One might animate this to cover long time periods
- An interactive graphic would allow zooming or drill down
- Just as they are, they are informative.

- These were done by Maj. Damon Becknel while at CERT. I have a student doing an interactive version.

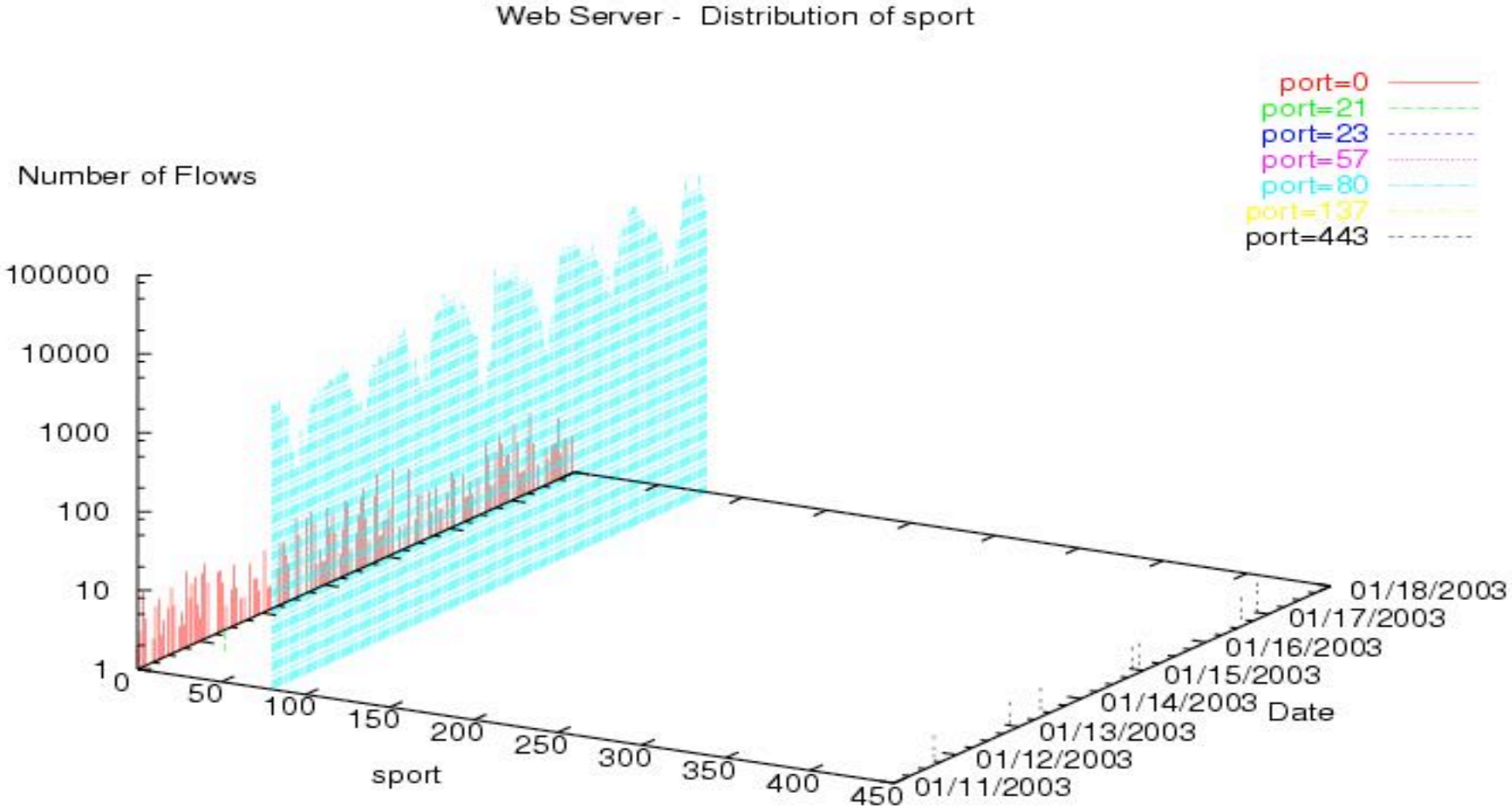
© 2005,6,7 by John McHugh

Mail Server?



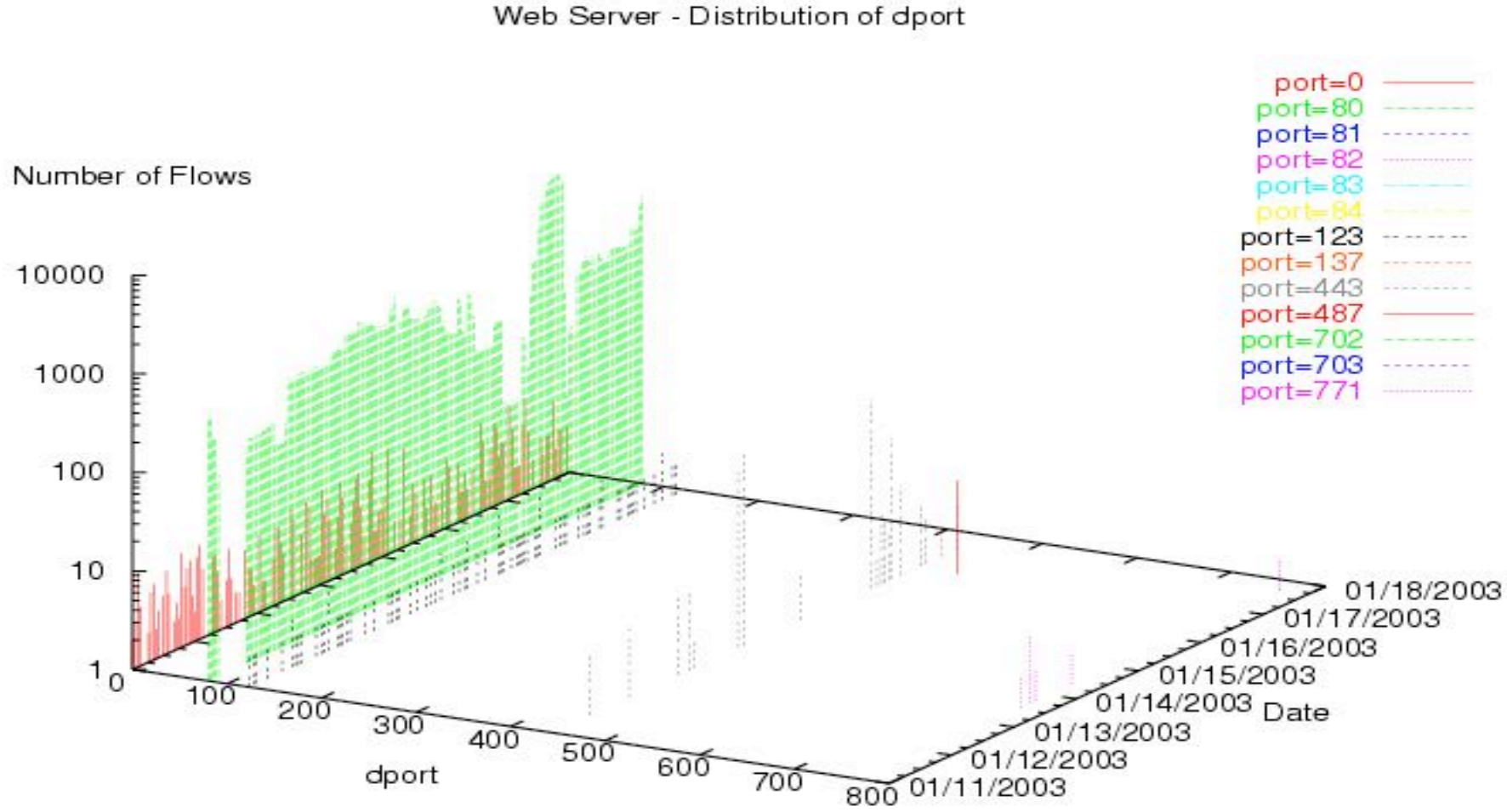
© 2005,6,7 by John McHugh

Web Server



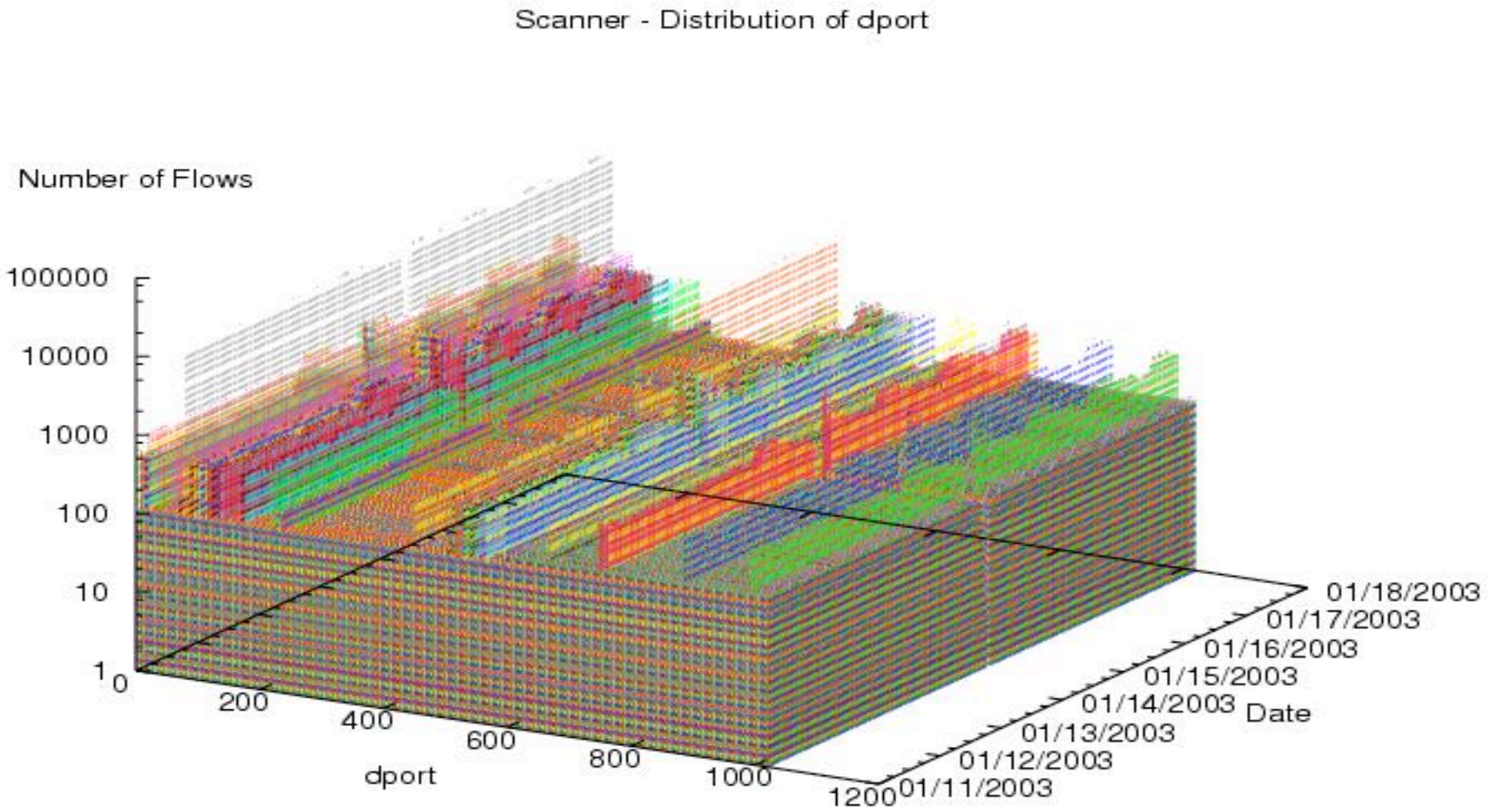
© 2005,6,7 by John McHugh

Web Server



© 2005,6,7 by John McHugh

Scanner



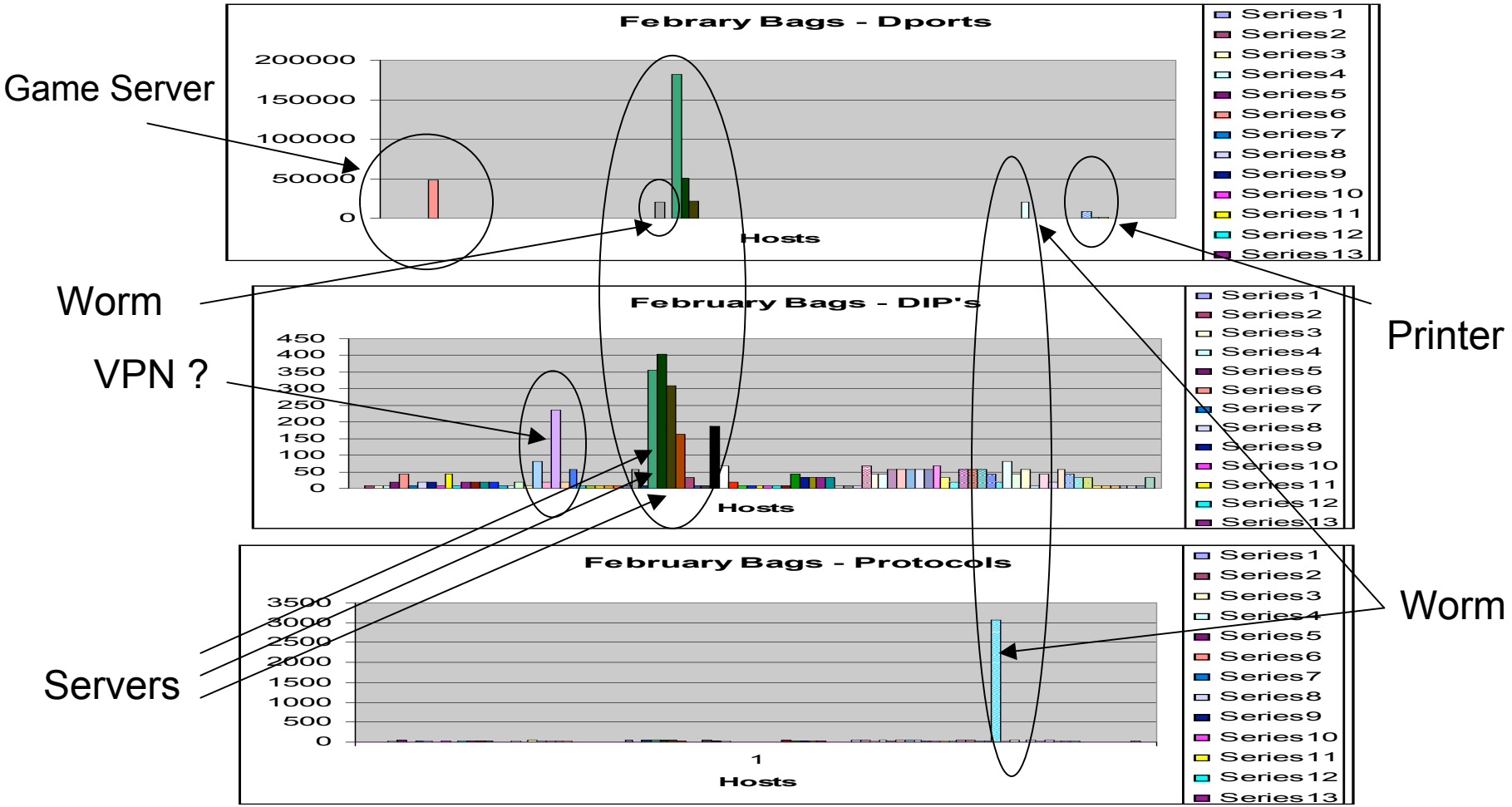
© 2005,6,7 by John McHugh

How well do you know yourself?

- One of my students monitored his work network for about 4 months. He performed a blind analysis of the outgoing traffic and characterized the hosts based on the results.
- The network was being used in surprising ways.
- The IP addresses are largely irrelevant, but, given some external knowledge, one could easily name them.
- The network contained about 80 hosts, many owned by associated companies and opaque to the hosting site.

© 2005,6,7 by John McHugh

Volumetric measures for the enterprise



© 2005,6,7 by John McHugh

Thank You

- I'll be around for the rest of the day.
- You can reach me as `mchugh` at `cs.dal.ca`
- If you want to perform similar capture and analysis of your network, we would be happy to help.

© 2005,6,7 by John McHugh