Pyrite or gold?
It takes more than a pick and shovel

SEI/CERT -CyLab

Carnegie Mellon University

20 August 2004

John McHugh,

and a cast of thousands

Pyrite or Gold?

Failed promises

- Data mining and machine learning techniques have been proposed as a mechanism for detecting malicious activity by examining a number of data streams representing computer and network activity.
- Although some encouraging results have been obtained, most systems do not deliver in the field what they promised in the lab.

Why is this?

- There are a number of reasons for this, but the most likely one is the failure of the developers of such systems to understand what the systems have learned and to relate it to the activity they are seeking to detect.
  - Put simply there are too many serendipitous detections, or
  - The distinguishing behavior is insufficient to establish either necessary or sufficient conditions for maliciousness.

Is this all?

- Probably
  - At this point, you should know what you need to do to fix the problem and I could go home, but, to drive home the point, I'm going to give some examples that may help to illustrate the problem.

If a tree falls in the forest …

- After some years of looking at the problem of relating abnormal behavior to malicious behavior, I am starting to realize that this is probably not fruitful.
  - Are there any necessarily normal behaviors?
  - Are there any behaviors that are necessarily both abnormal and malicious?
  - Are there any descriptors that are sufficient to identify either?

Normal Program Behavior

- A number of suggestions have been made for defining normal program behavior.
  - Execution traces under normal conditions
  - Execution traces by symbolic execution
  - Formal specifications
  - Domain and type enforcement
- We will consider the first two further

Observed Execution Traces

- If we run a program long enough, in a "typical" environment, we should collect a definitive set of "normal" traces.
- Under some measure of similarity, abnormal traces may indicate intrusive activity.
- The work begun by Forrest is based on this approach.

Possible Execution Traces

- In theory, it is possible to determine all possible execution traces for most uesful programs. This could substitute for observation, but
- What if the traces admit malicious behavior?
  - We will see an example later.
- A trace ending with “exec” may or may not be normal

Problems

- System call traces, per se, are not a good basis for detection.
- STIDE represents normal by a set of fixed length subtraces.
  - This is blind to minimal foreign subtraces longer than the length chosen
  - There are ways to modify attacks so as to be missed (Tan, Maxion, Wagner)

Example #1 -- Hiding in Normal: Modifying the `restore` attack

- The `restore` attack manifests in the following sequence:

  `write,read,write,munmap,exit`

- The `read`, `write` is due to the fact that the program used by `restore` (created by the attacker) failed to connect to the remote computer, causing `restore` to print an error and exit. Using the normal data as training data and this attack as test data,we find the following two minimal foreign sequences:

  `write,read,write,` and `write,munmap`

- Since this attack contains a MFS of size 2, stide will detect this attack with a window size of 2 or greater.

- We now need to find new ``stide-friendly'' means to achieve the same goal.

Example #1 (cont) "Stide-friendly" Method

- What if the attacker's program did not fail to connect to a remote computer, but performed the expected operations of the program normally invoked by `restore`?

- We craft such a program to mimic the normal behavior of the program normally invoked by restore (e.g. the secure shell) while still giving the attacker root privileges.

- The attacker can modify the exploit to make the system program `restore` run our ``normal-mimicking'' program instead. Now rather than printing an error and exiting, `restore` executes successfully, as shown in the manifestation of this attack:

`write,read,read,read,...`

- This manifestation contains no minimal foreign sequences, yet it still bestows root privileges upon the attacker.

Network data

- NIDS look at various representations of network data.
  - Whole packets
  - Packet headers
  - Network flows
- Ground truth (labeling) is difficult for real data, artificial data presents other problems.
- Lets look at the Lincoln data.

Context and Background
(with a hint of a research agenda)

- In the summer of 1999, shortly after I joined CERT, I became involved in the production of a report on the state of practice in intrusion detection. This led to an investigation of efforts by MIT's Lincoln Laboratory to evaluate IDSs developed under DARPA's research programs.
- Critiques of the 1998 Lincoln work were presented and published in several forums. The conclusion of these efforts is that evaluating IDSs is very difficult under the best of circumstances. The reasons range from fundamental lack of theory to complex logistics.

Why 1998?

- At the time this effort began, the 1998 evaluation was complete, 1999 was about to begin.
- The public record associated with the 1998 effort indicated a number of problems and left many questions unanswered.
- It appeared likely that the 1999 evaluation would not address many of these issues so that many of the questions raised with respect to the 1998 evaluation would remain relevant.
- A recent analysis of the 1999 data shows it has similar problems

Data Characterization

- Artificial data was used for the evaluation.
  - Characterization of real environment incomplete, sample size unknown
  - The data is said to be based on typical traffic seen at Air Force bases. The sample may be as small as 1.
  - The abstractions used to define “typical” are not known except for word frequencies in Email traffic.

Data Characterization

- Artificial data was used for the evaluation.
  - The criteria for background data generation are not completely known - No pathologies?
  - There is a lot of “crud” on the wire. Some of it looks intrusive, but apparently isn’t. This is not seen in the test data.
  - As far as we know, there is no probe data in the background data, but some things that look like probing are normal and others are not the result of malicious intent.

Data Characterization

- Artificial data was used for the evaluation.
  - The test data set has not been validated.
    - No pilot studies or data evaluations were performed.
    - Investigators complained of having to debug test data while evaluating their systems
    - False alarm characteristics critical to evaluation, but not known to be realistic in test data

Data Characterization

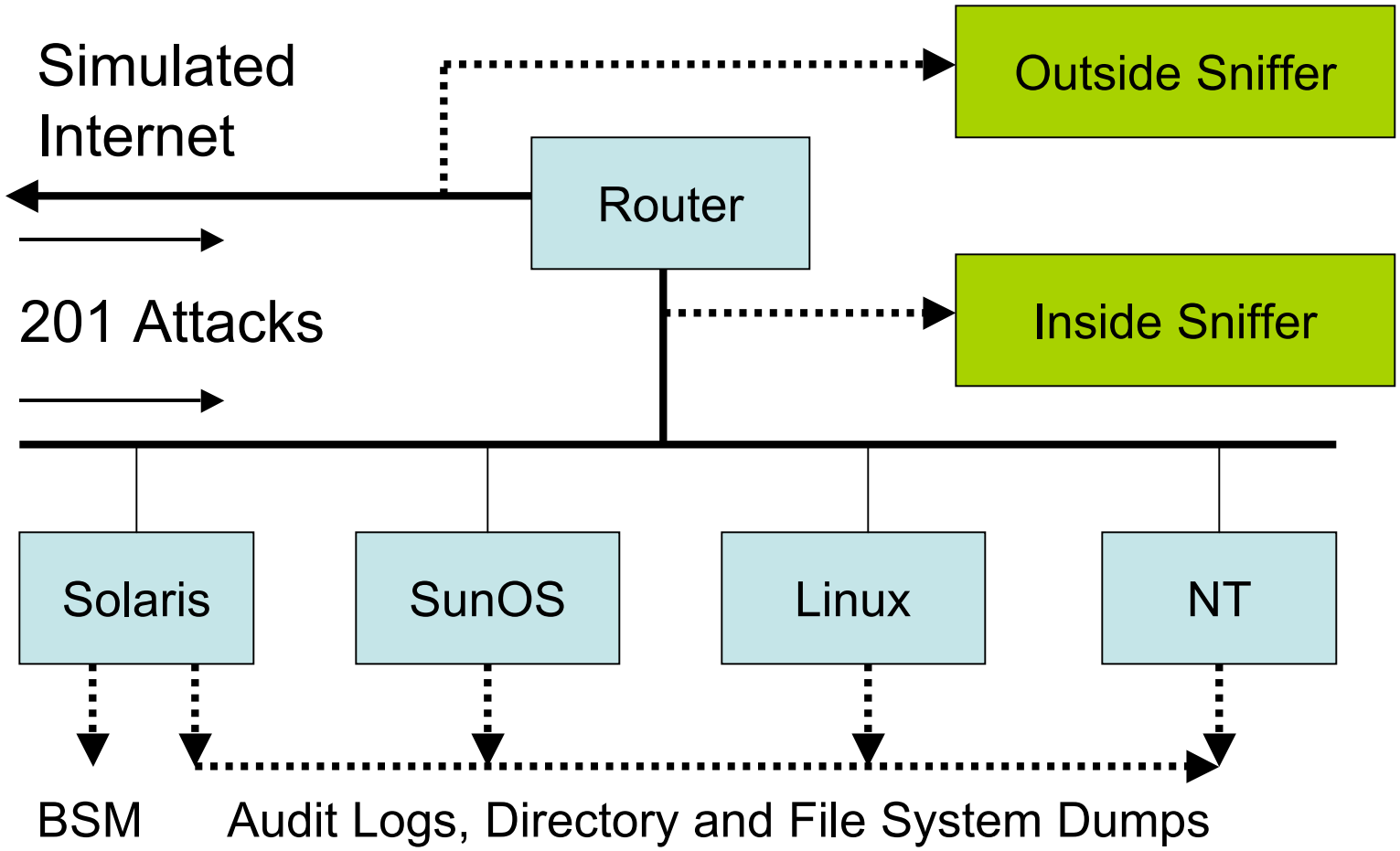
- Artificial data was used for the evaluation.
  - Data rates appear low 10-50Kb/sec at border
    - Contrast with university sites of similar size shows 10X or more difference.
  - Attack mix unrealistic - attack rate may be also
    - Minor role for probes
    - Limited variety of attacks (more in ‘99)
    - Attacks per day may be high

Other data issues

- Comparison of the Lincoln data, with other similar sources produces interesting results.
- BSM data shows very different conditional entropy from real data observed by Forrest, et. al. Lee S&P 2001
- Clustering attempts on tcp dump data produce “linear” clusters indicating low variability.

  Taylor NSPW 2001
- These are suspicious.
- **Lets look in detail at the 1999 network data!**

1999 IDEVAL



Thanks to Mahoney and Chan (RAID 2003)

1999 IDEVAL Results

Top 4 of 18 systems at 100 false alarms

System	Attacks detected/in spec
Expert 1	85/169 (50%)
Expert 2	81/173 (47%)
Dmine	41/102 (40%)
Forensics	15/27 (55%)

Thanks to Mahoney and Chan (RAID 2003)

Problem Statement

- Does IDEVAL have simulation artifacts?
- If so, can we "fix" IDEVAL?
- Do simulation artifacts affect the evaluation of anomaly detection algorithms?

Thanks to Mahoney and Chan (RAID 2003)

Simulation Artifacts?

- Comparing two data sets:
  - IDEVAL: Week 3
  - FIT: 623 hours of traffic from a university departmental server
- Look for features with significant differences

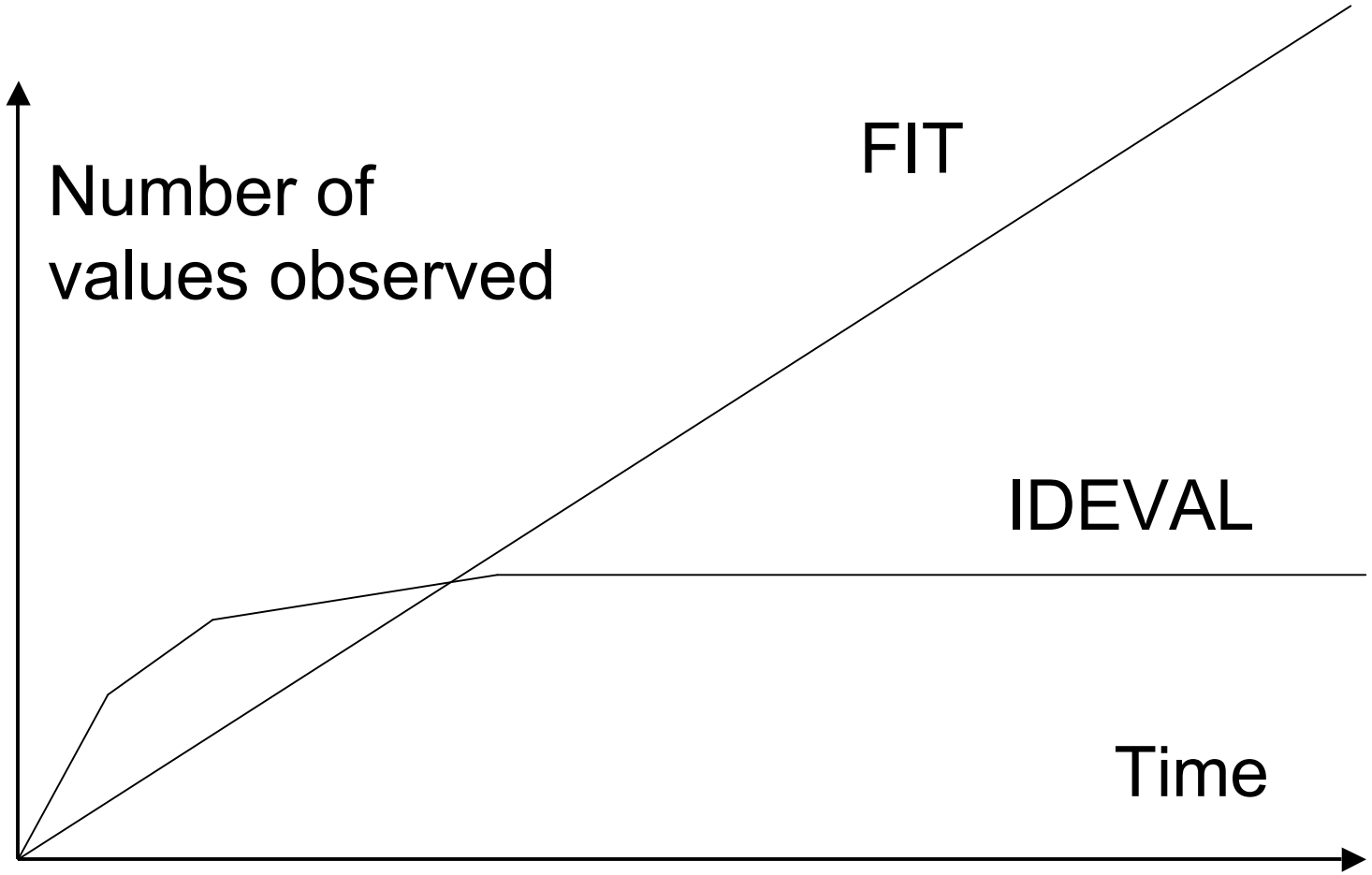
Thanks to Mahoney and Chan (RAID 2003)

of Unique Values & % of Traffic

Inbound client packets	**IDEVAL**	**FIT**
Client IP addresses	29	24,924
HTTP user agents	5	807
SSH client versions	1	32
TCP SYN options	1	103
TTL values	9	177
Malformed SMTP	None	0.1%
TCP checksum errors	None	0.02%
IP fragmentation	None	0.45%

Thanks to Mahoney and Chan (RAID 2003)

Growth Rate in Feature Values



Thanks to Mahoney and Chan (RAID 2003)

Conditions for Simulation Artifacts

1. Are attributes easier to model in simulation (fewer values, distribution fixed over time)?
   - Yes (to be shown next).
2. Do simulated attacks have idiosyncratic differences in easily modeled attributes?
   - Not examined here

Thanks to Mahoney and Chan (RAID 2003)

Exploiting Simulation Artifacts

- SAD – Simple Anomaly Detector
- Examines only one byte of each inbound TCP SYN packet (e.g. TTL field)
- Training: record which of 256 possible values occur at least once
- Testing: any value never seen in training signals an attack (maximum 1 alarm per minute)

Thanks to Mahoney and Chan (RAID 2003)

SAD IDEVAL Results

- Train on inside sniffer week 3 (no attacks)
- Test on weeks 4-5 (177 in-spec attacks)
- SAD is competitive with top 1999 results

Packet Byte Examined	Attacks Detected	False Alarms
IP source third byte	79/177 (45%)	43
IP source fourth byte	71	16
TTL	24	4
TCP header size	15	2

Thanks to Mahoney and Chan (RAID 2003)

Suspicious Detections

- Application-level attacks detected by low-level TCP anomalies (options, window size, header size)
- Detections by anomalous TTL (126 or 253 in hostile traffic, 127 or 254 in normal traffic)

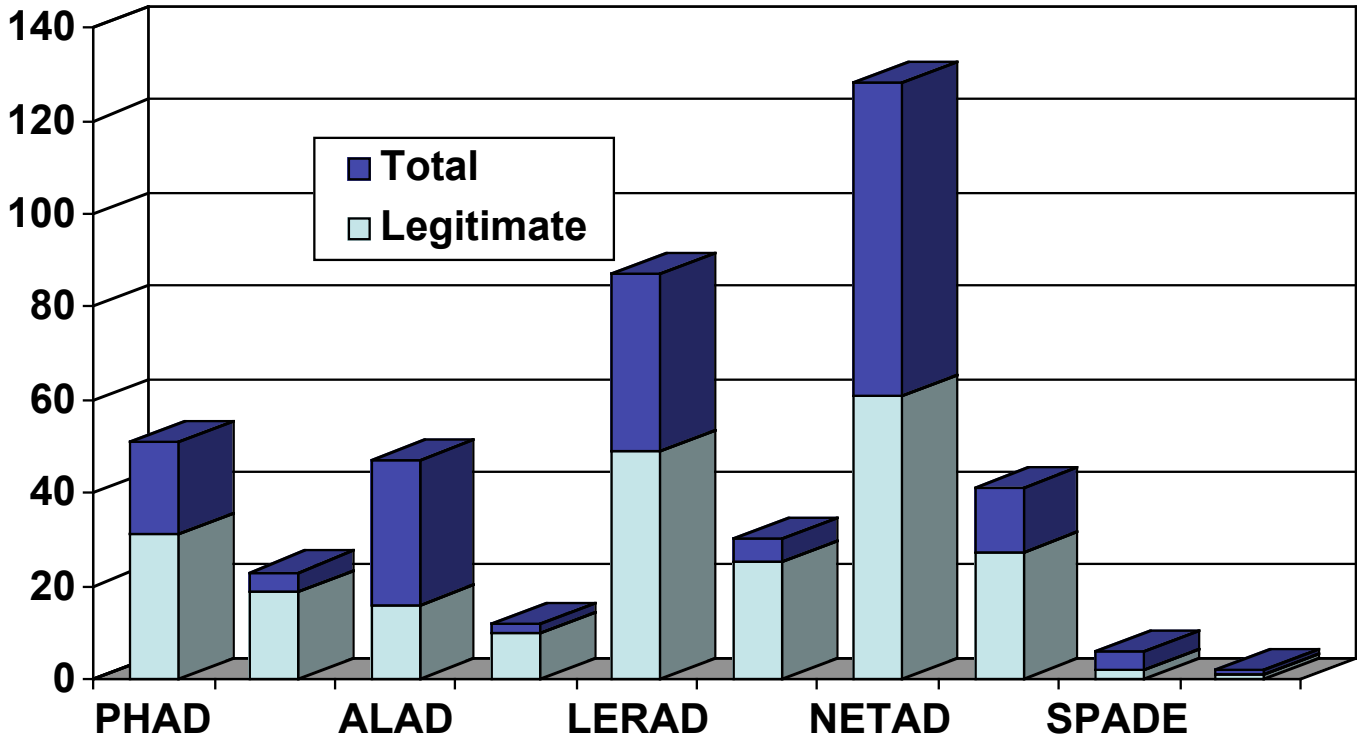
Thanks to Mahoney and Chan (RAID 2003)

Mahoney & Chan Mix

- Mahoney and Chan mixed the IDEVAL 99 data with FIT data, taking care to modify the mix (and IDS rules) to avoid developing independent models for each component.
- They then tested the mix of a number of anomaly detection algorithms.

Mixed Traffic: Fewer Detections, but More are Legitimate

Detections out of 177 at 100 false alarms



Thanks to Mahoney and Chan (RAID 2003)

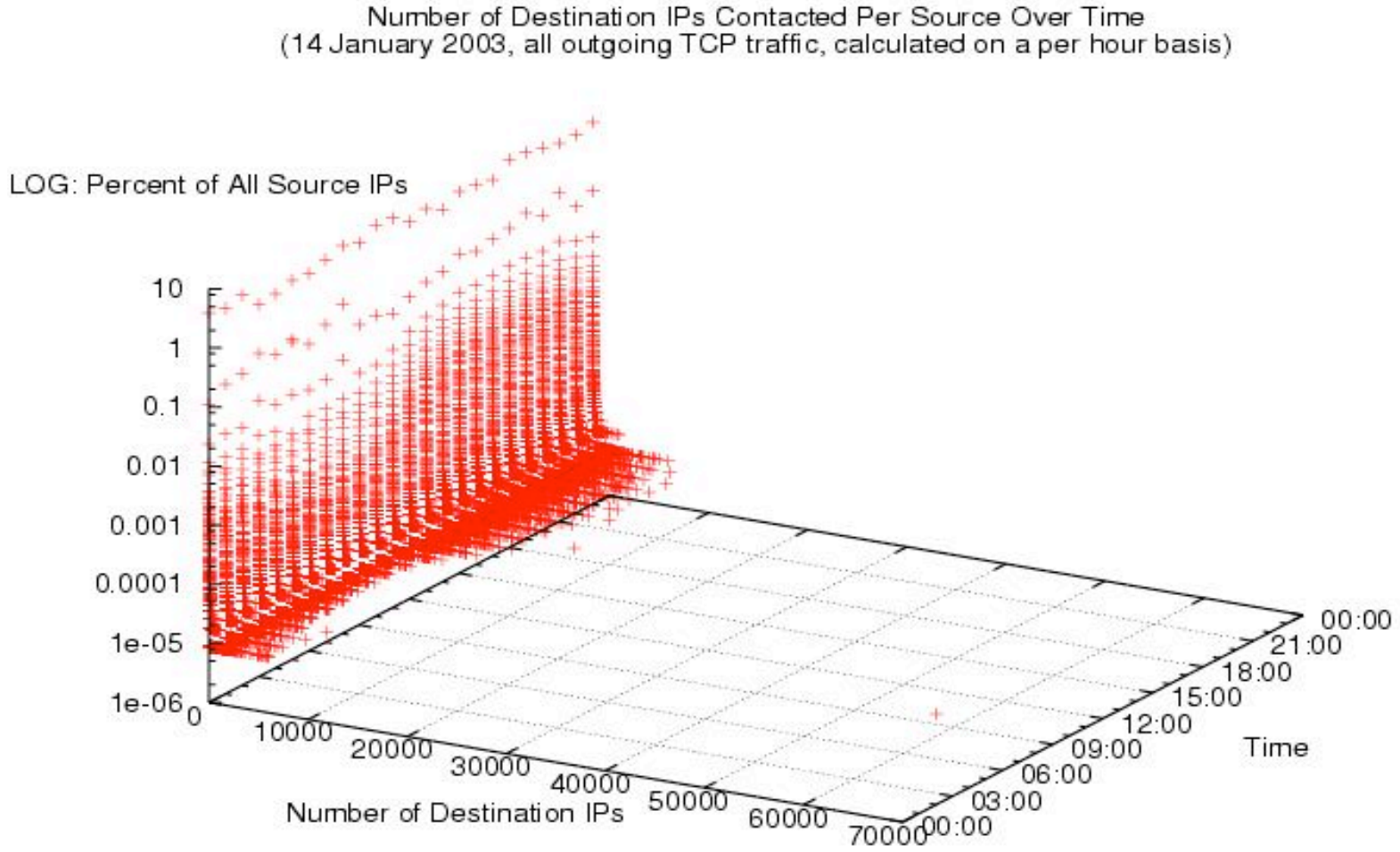
Affect on IDEVAL?

- If the results shown on the previous slides hold up for the IDEVAL 99 systems, the results of the DARPA program would be only about half as good as previously reported.
- It is known that a number of the systems evaluated during the 99 trials had difficulties when field deployments were attempted

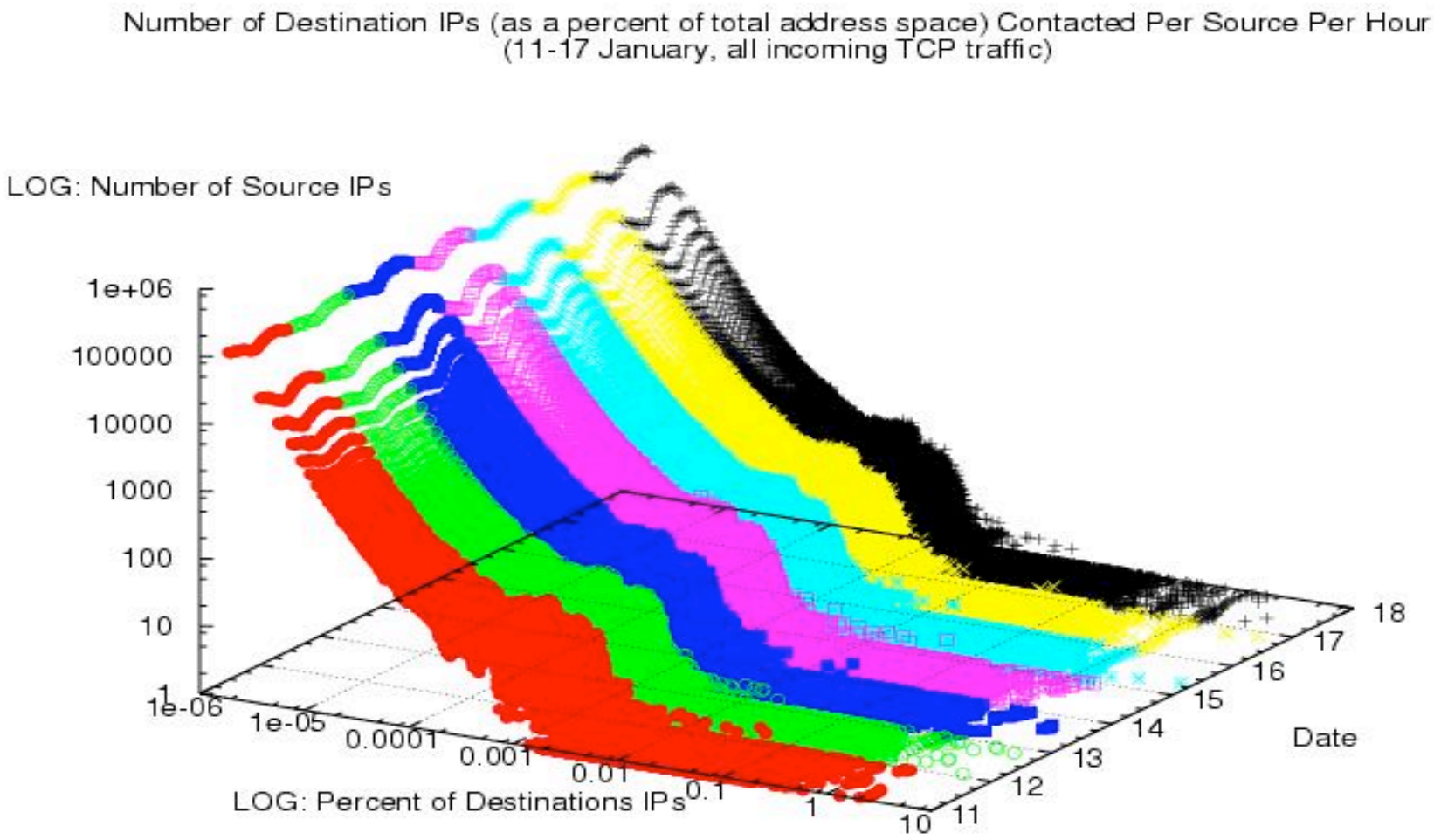
Real data is complex

- We are working with data from a customer network to characterize complexity and normal behaviors.
- The following are a few illustrations:

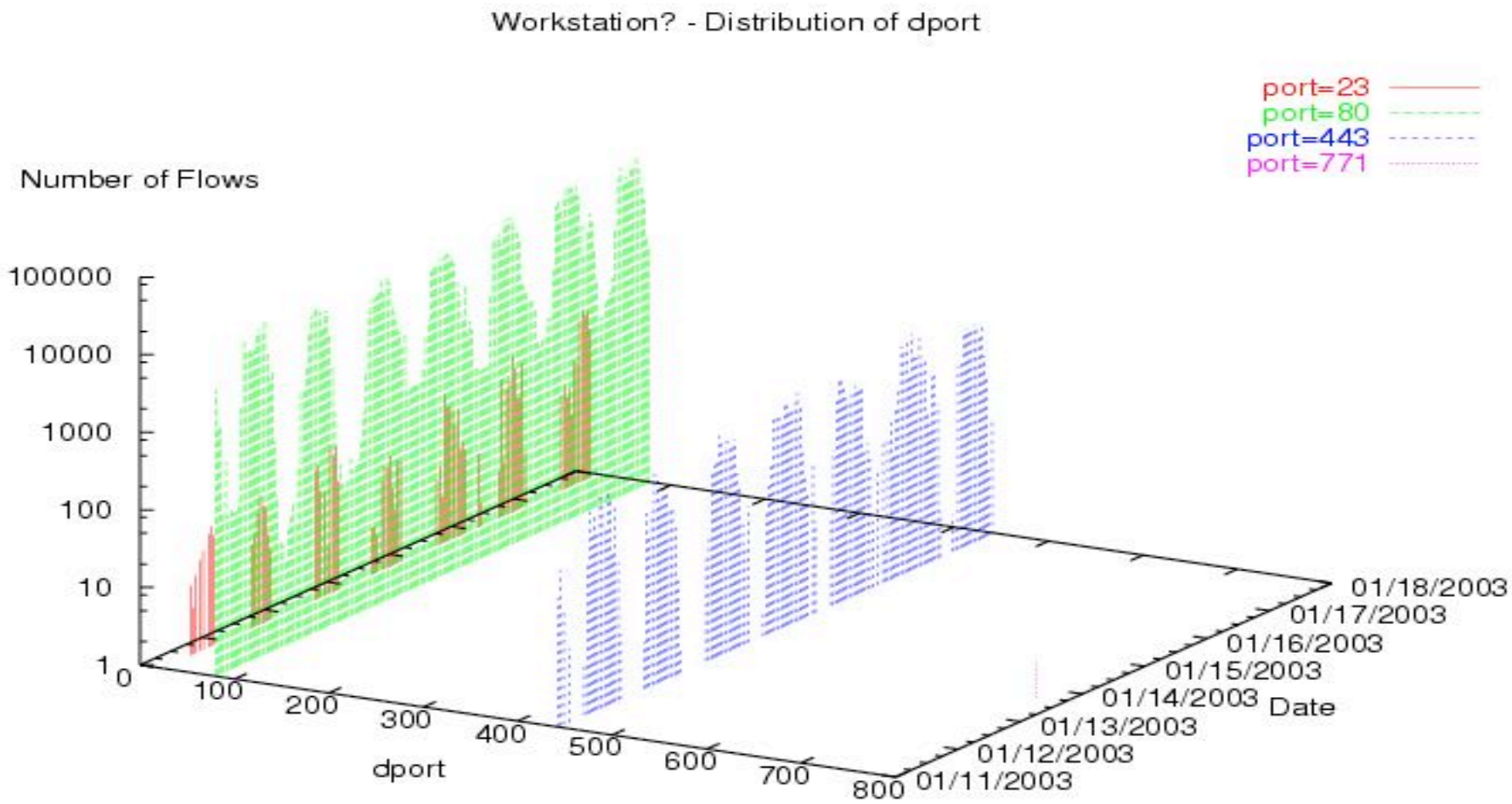
One Day of Inside to Outside



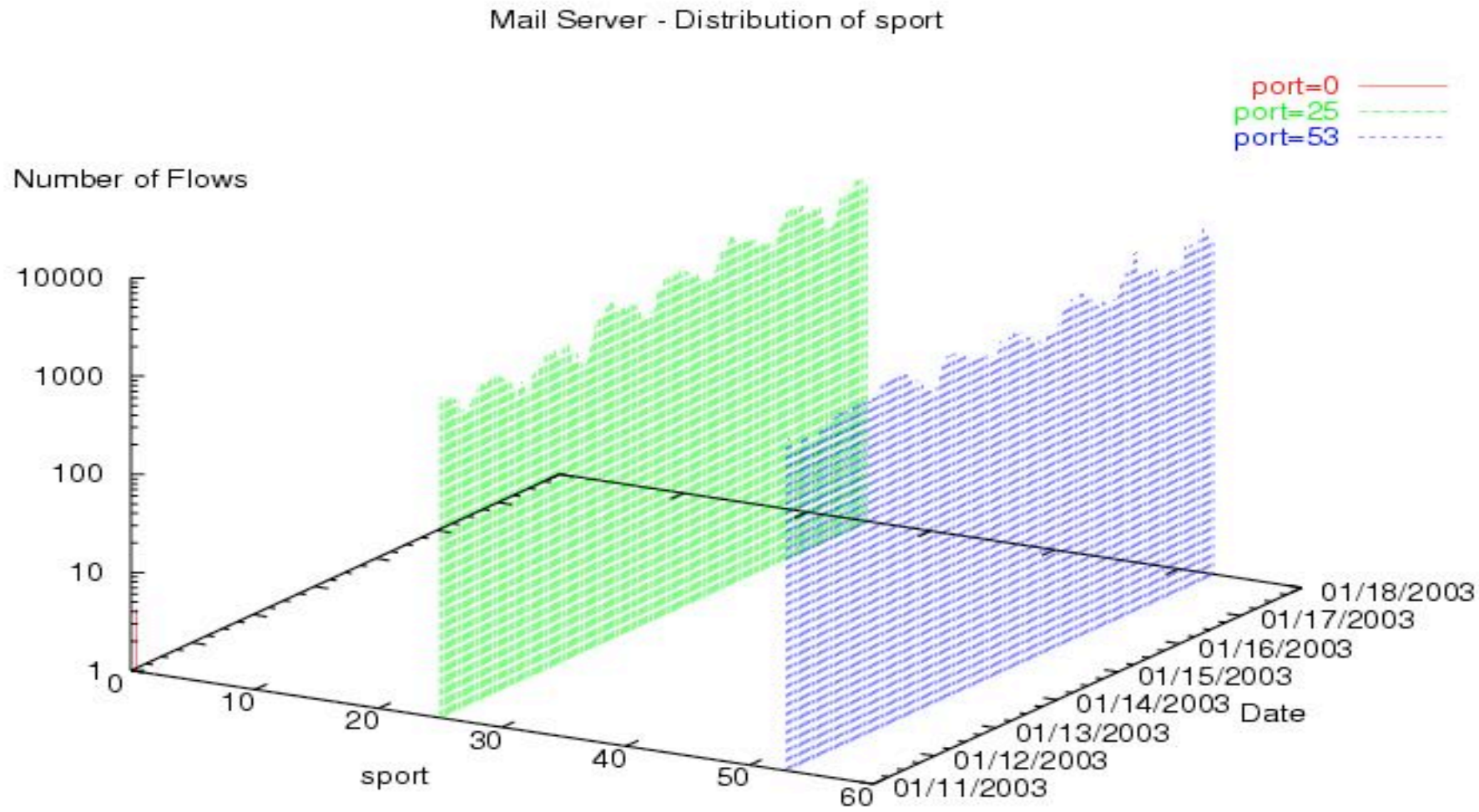
The Week's Out/In Surface



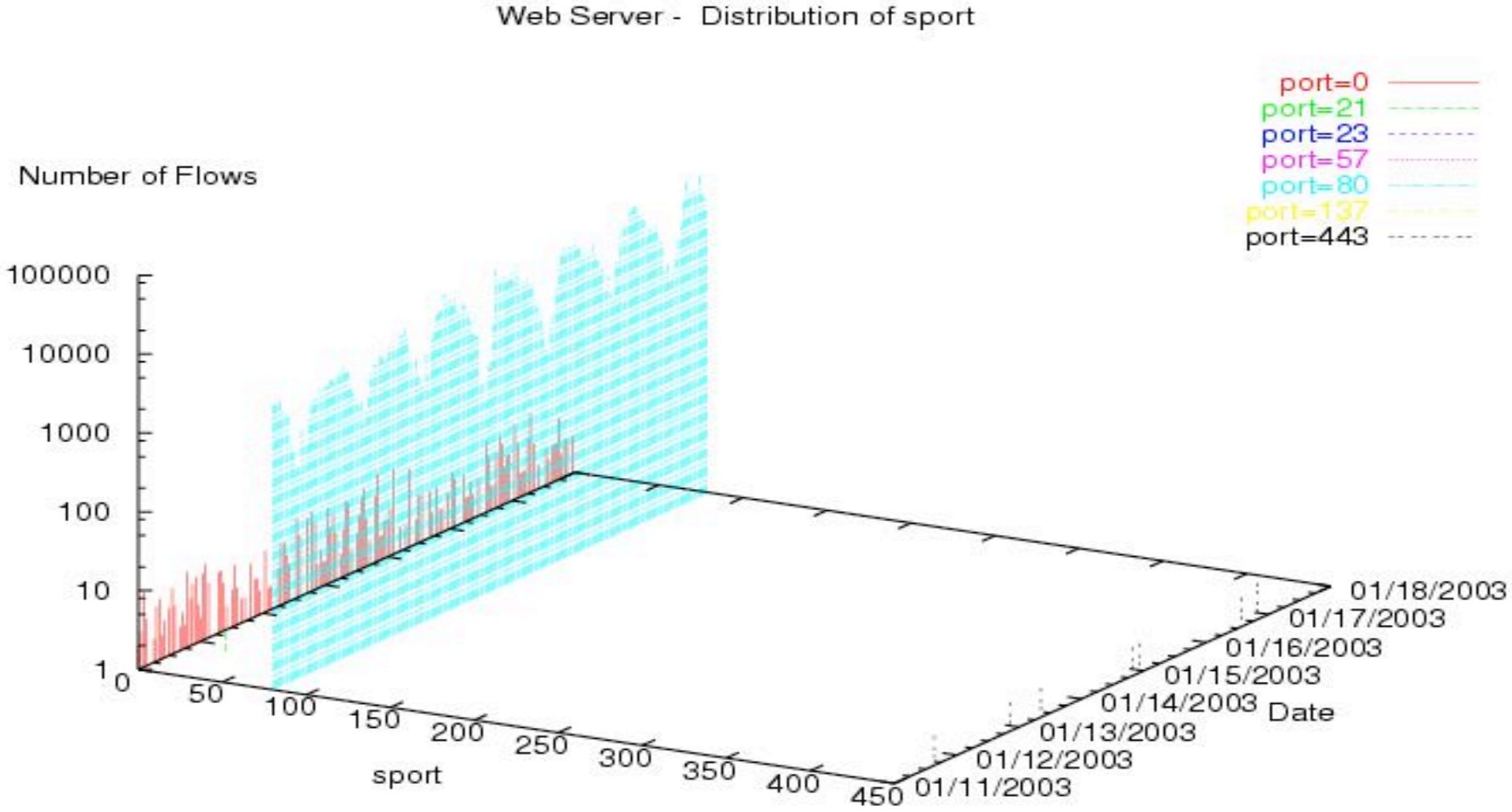
Workstation?



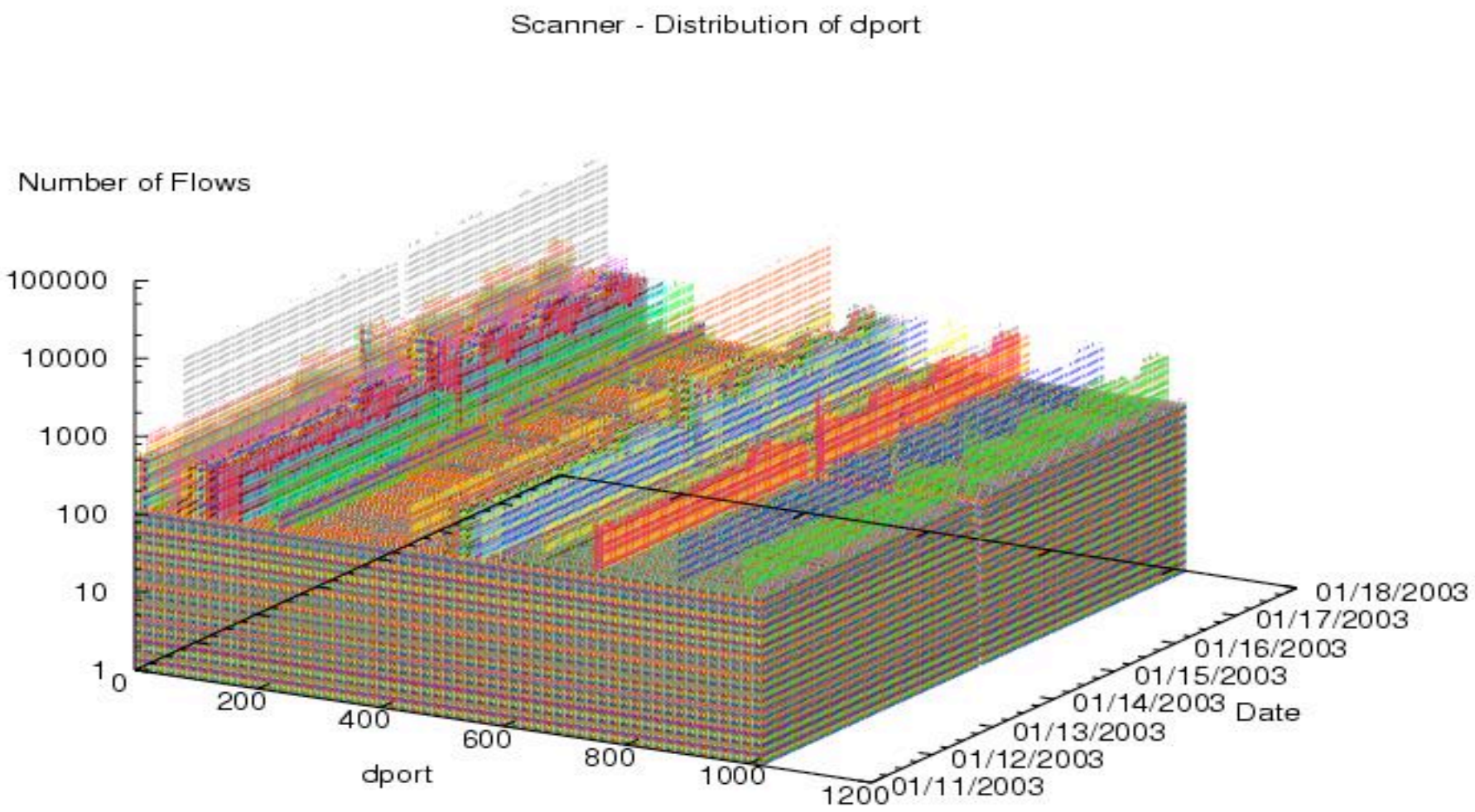
Mail Server?


Mail Server - Distribution of sport
port=0
port=25
port=53
Number of Flows
10000
1000
100
10
1
0
10
20
30
40
50
60
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

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

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

What does it mean

- There is a forensic component to anomaly detection when it is used for detecting intrusions.
- In evaluating an anomaly detector used this way, it is incumbent on the evaluator to show that the features on which the detection is based are not serendipitous.
  - In addition to the above, I have reviewed numerous papers based on irrelevant detections.

Consequences

- Artificial data such as the Lincoln data is useful, but:
  - Don't take the results too seriously
  - Good results on artificial data should
    - Encourage you to see if the results hold up in the wild (If you can't do this was the idea really good?)
    - Discourage you from seeking to publish, at least in the intrusion detection literature

Conclusions

- Anomalies are not necessarily benign or malicious, they are just different or rare
- Just because an intrusion is associated with an anomalous feature does not mean that it has to be that way.
- If you don't understand the forensics of intrusions, you are likely to be part of the problem, not part of the solution.

Pyrite or Gold?