Anomaly Detection Through Blind Flow Analysis Inside a Local Network

Ron McLeod, BCSc, MCSc.
Director - Corporate Development Telecom Applications Research Alliance
Doctoral Student, Faculty of Computer Science, Dalhousie University
Vagishwari Nagaonkar, BCSc
Senior Systems Engineer, Wipro Technologies
Graduate Student, Faculty of Computer Science, Dalhousie University

Abstract

In the August of 2006, 4 months of Netflow records that were collected inside a small private network were subjected to a Blind Flow Analysis. Such an analysis is characterized by having access to the flow records from inside the network but no access to the payload data and no physical access to the hosts generating the traffic. Experiments were conducted to discover if useful behavioural clusters could be constructed with such minimal access and whether individual classes of hosts could be clustered into standard ranges including clusters indicative of compromised hosts. Early results are promising in that hosts may be clustered into User Workstations, Servers, Printers and hosts Compromised by Worms.

Overview

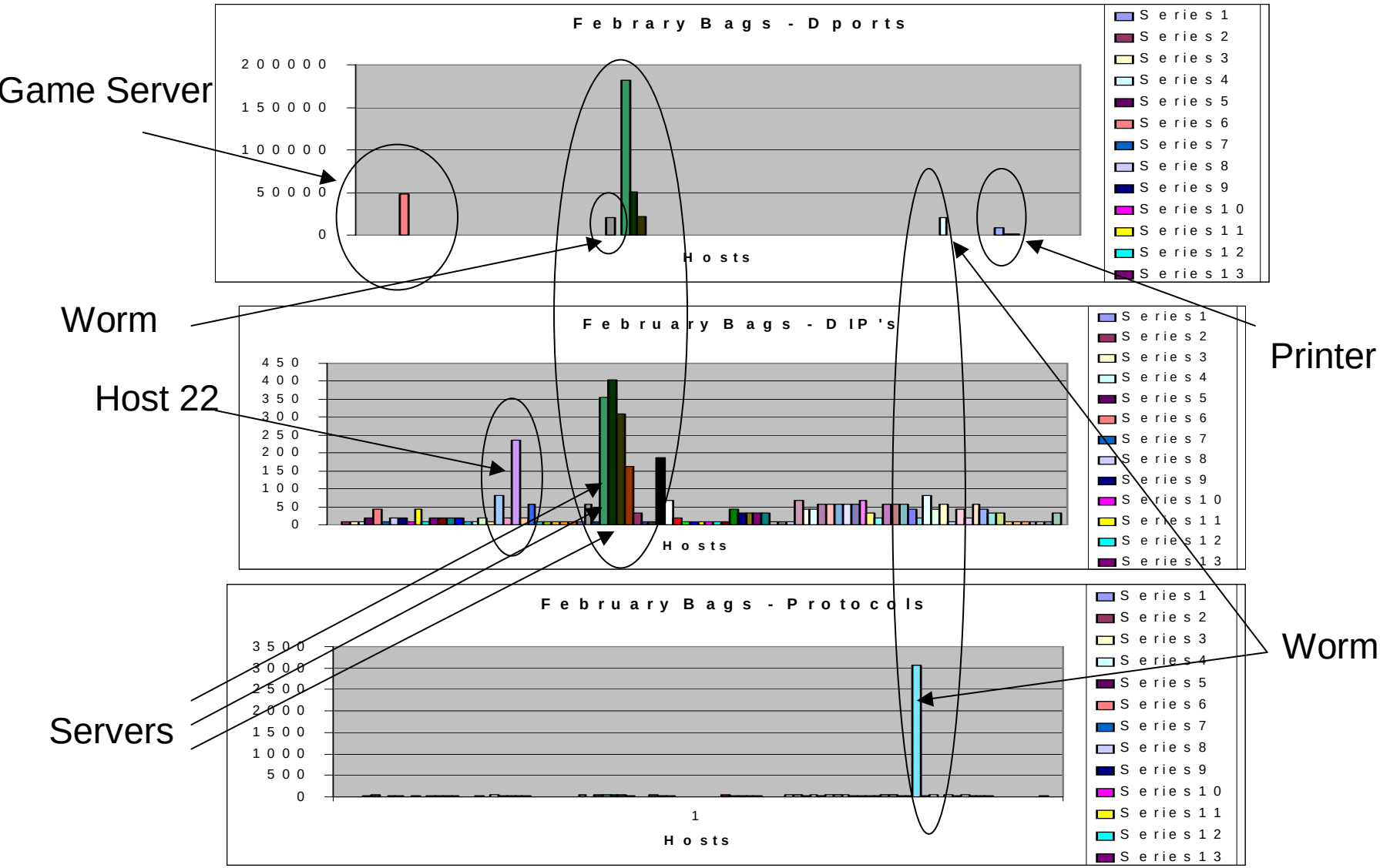
- Network Monitoring for Security In a Multi-tenant Environment network environment is problematic.
- Tenants (Including individuals and corporate entities) have specific concerns with respect to privacy or corporate confidentiality.
- A network analyst may be specifically forbidden from capturing the payload data. The analyst may not be granted access to specific hosts and may not even be able to receive information as to the type and nature of the host in question (i.e. is this a server, a workstation or a printer).
- In this environment the analyst may be restricted to analyzing only packet header data or flow records.
- The authors decided to test the ability of the analyst to form useful characterizations in such a restricted environment.

Clusters by File Size

- First Characterization was by Bag File Size

Feb Bag Data

File Size in Bytes



Procedure

- As each anomaly in the data was observed, A hypothesis was developed and the IP number of the host in question along with the hypothesis was sent to the network owner.
- For the purpose of testing experimental results, the network owners were asked to confirm (on a voluntary basis) the hypothesis.

Two Anomalous Cases Only Briefly Addressed

- The Anomaly labeled Game Server represented a compromised host on the network that was being used to support worldwide on-line gaming.
- The anomaly labeled Host 22 was believed to be a VLAN gateway, but this hypothesis has yet to be confirmed.

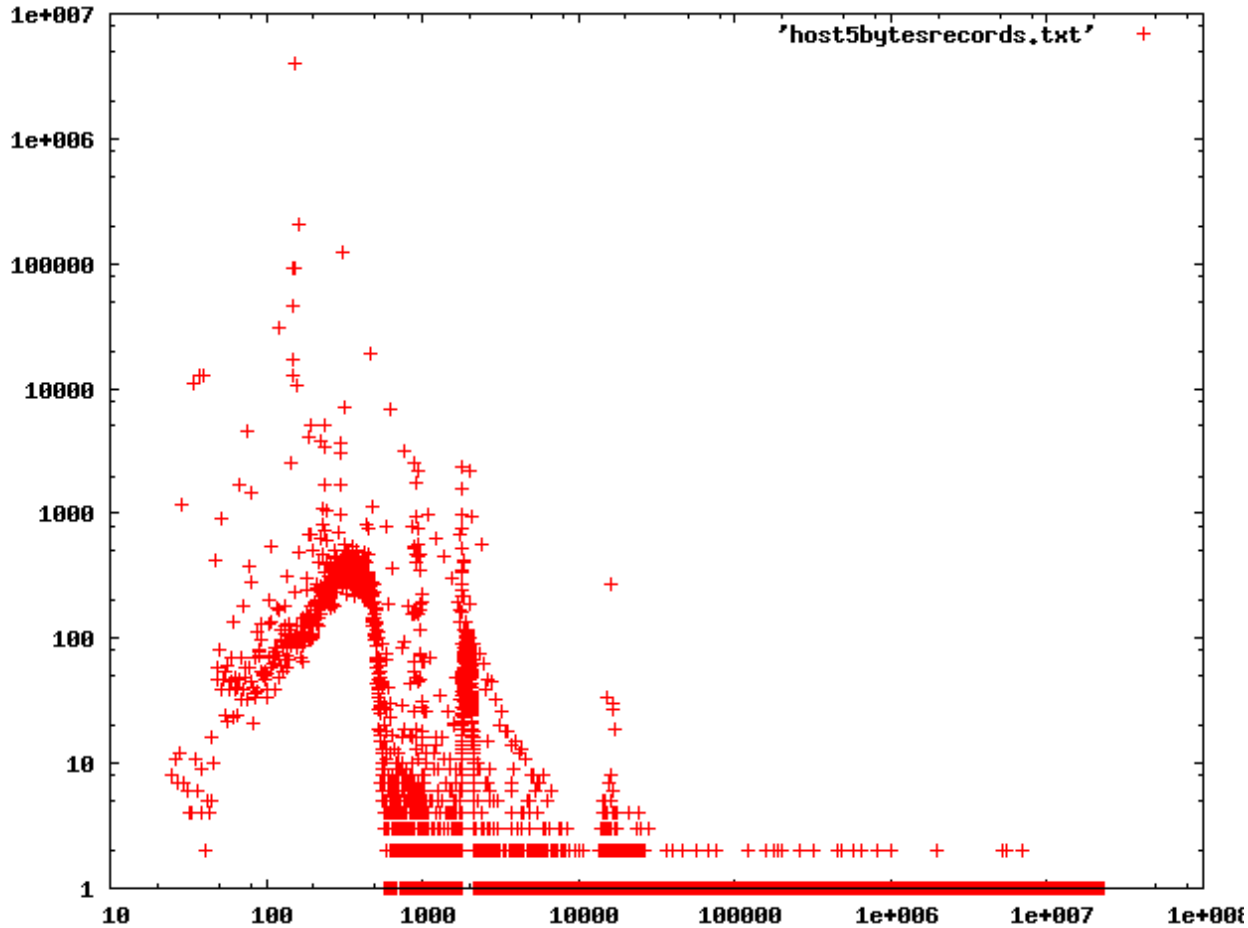
Game Server Behaviour

Game Server Profile from Bags for Bytes, Destination IP's and Destination Ports

- Outbound Byte Transfers per month: 45 Billion Bytes.
- Destination IP's per month: 2 million external hosts
- Large numbers of flow records of small byte size coupled with less number of records with very large Byte size. Accessed to virtually every destination Port

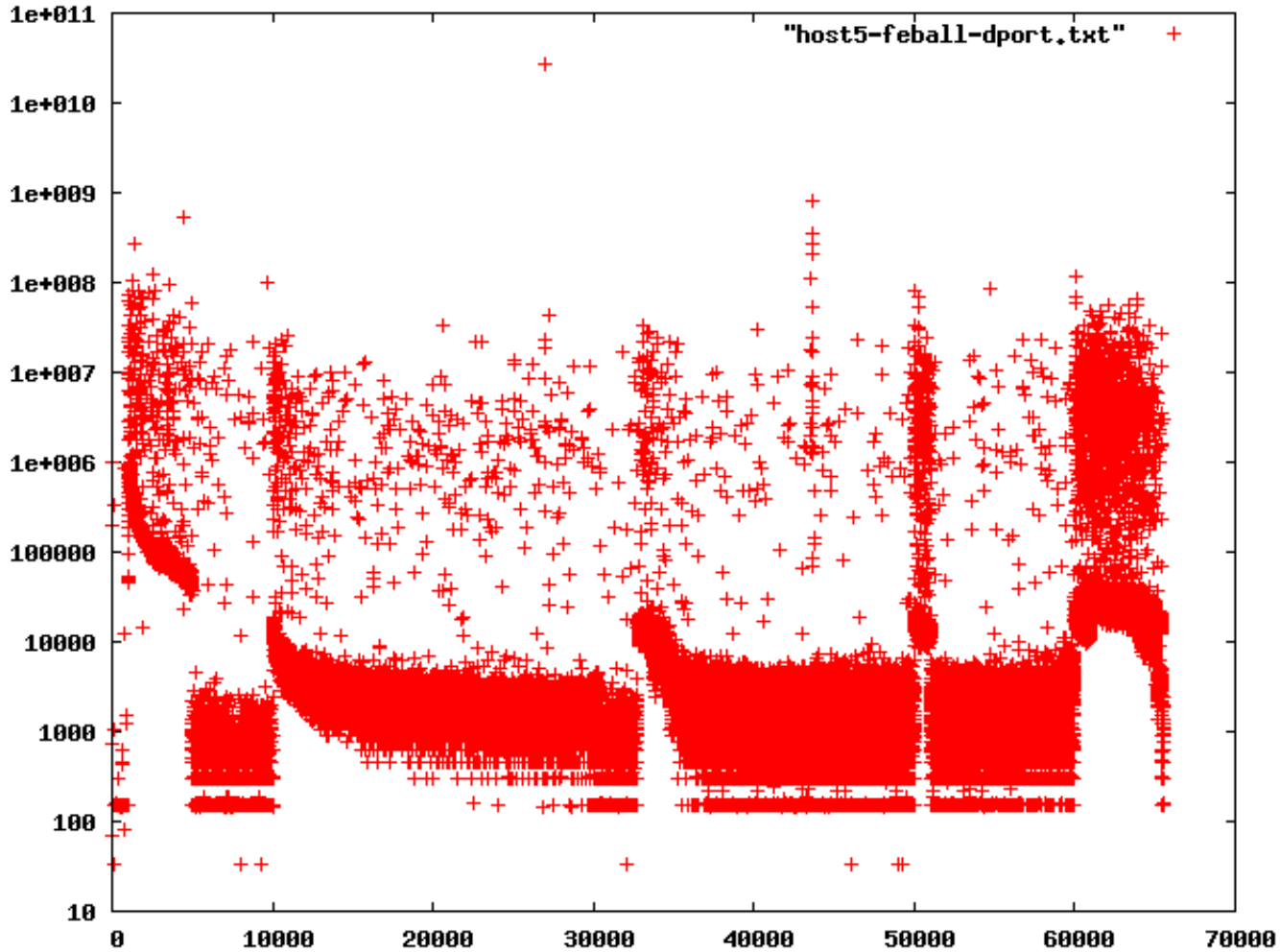
Games Server Record Size

Number Of Records

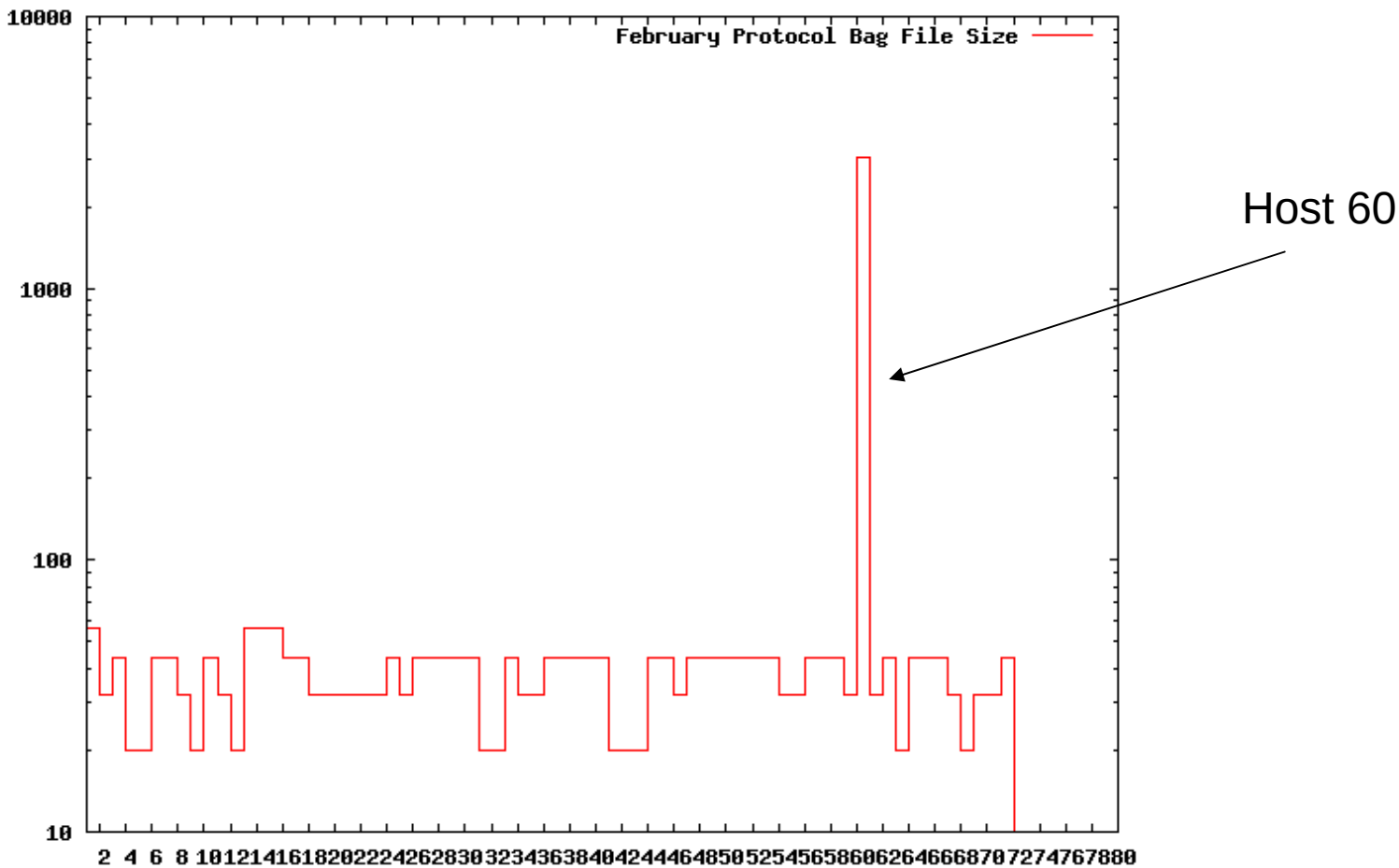


Byte Size of Record

Game Server Dport Distribution



Protocol Bag File Size



80 Hosts for one Month (Feb)

Protocol Bag File Size

Host 60 Behaviour

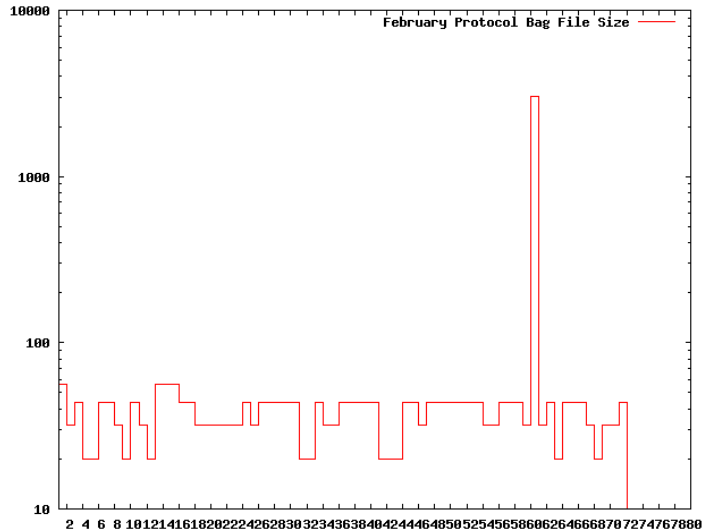
Sequential access to entire /24.

Sequential access to /24's within specific /16's (Microsoft).

Small uniform byte volumes sent to every port using every protocol to every machine.

Host was a user workstation and the user complained that their machine was slow and the cpu seems to be busy even when they are doing nothing.

Hard drive was restored from earlier backup. Performance improved.



Protocol Bag File Size

Host 60 Behaviour

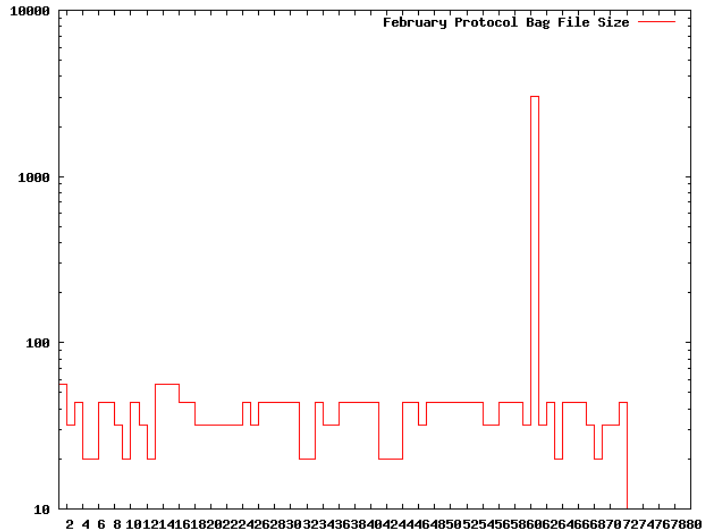
Sequential access to entire /24.

Sequential access to /24's within specific /16's (Microsoft).

Small uniform byte volumes sent to every port using every protocol to every machine.

Host was a user workstation and the user complained that their machine was slow and the cpu seems to be busy even when they are doing nothing.

Hard drive was restored from earlier backup. Performance improved.



Where did it come from?

Protocol Bag File Size

Host 60 Behaviour

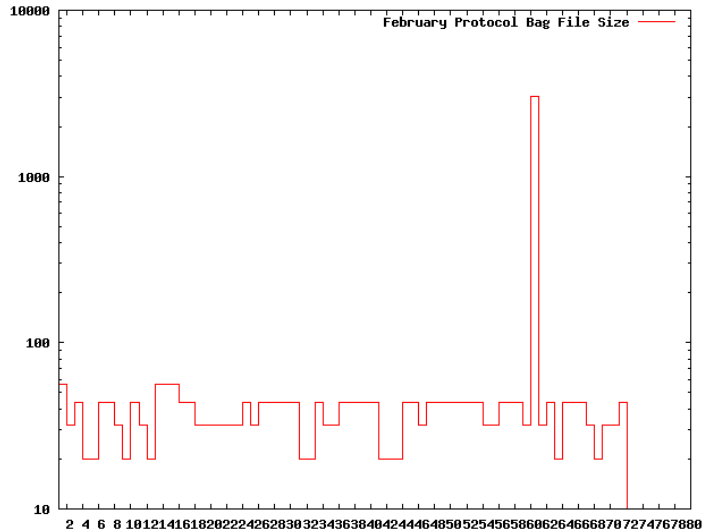
Sequential access to entire /24.

Sequential access to /24's within specific /16's (Microsoft).

Small uniform byte volumes sent to every port using every protocol to every machine.

Host was a user workstation and the user complained that their machine was slow and the cpu seems to be busy even when they are doing nothing.

Hard drive was restored from earlier backup. Performance improved.

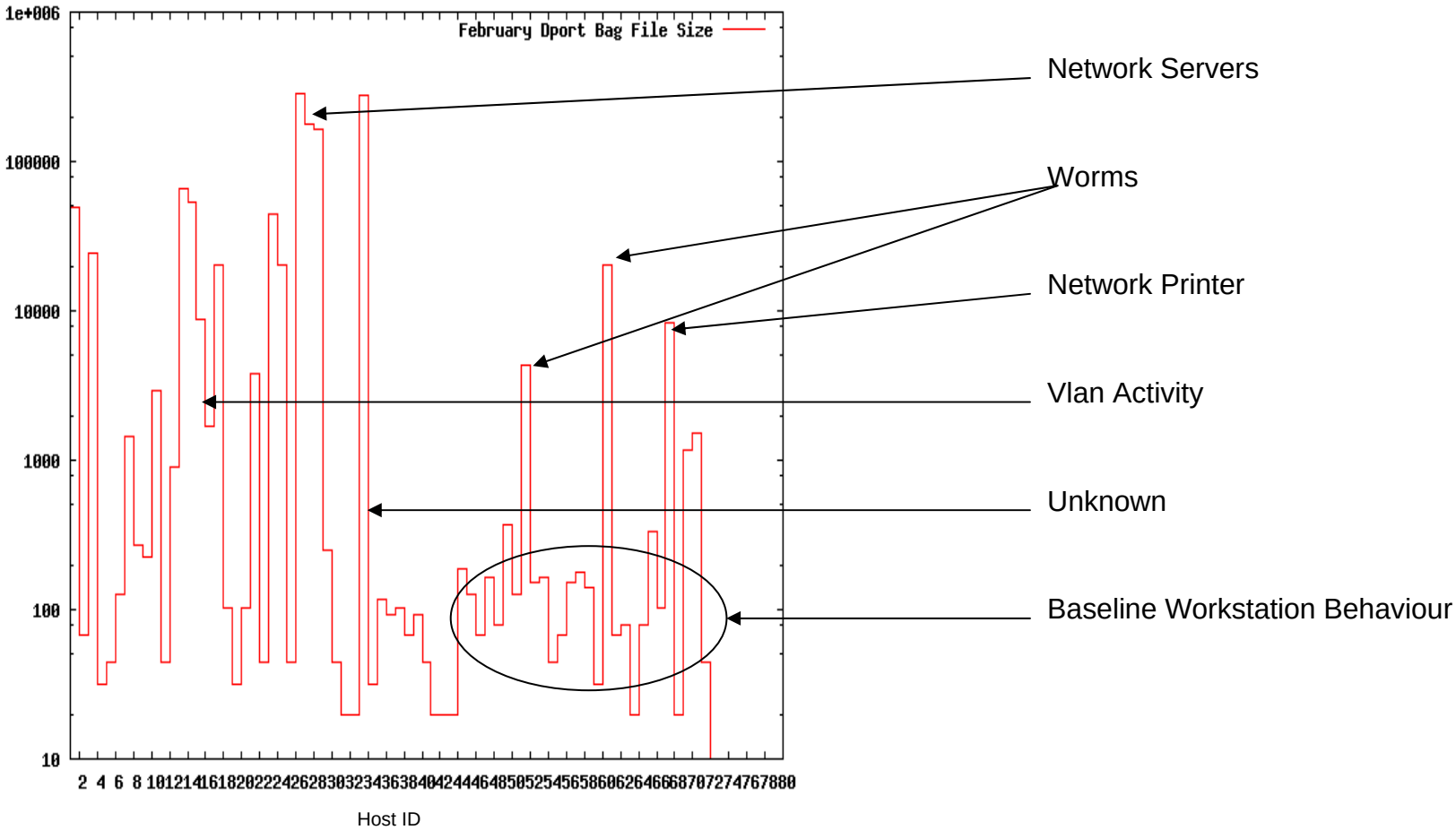


Where did it come from?

DIP Bag for Game Server
1,434,898 External Hosts

Host 57	233,952
Host 60	**3,891,482**
Host 34	1,021,287

Some Identified Anomalies



Destination Port Bag File Sizes

Eight Workstation Hosts
Byte Bag Behaviour

Host ID	Outbound Bytes Recorded in February	
44	5,261,662	
47	10,521,361	
48	2,122,423	
50	2,935,836	
51	2,493,552,524	
52	8,251,245	
56	15,126,755	
60	7,869,147	

Byte Bag Characteristics

First Component of Local Workstation BWB Rule:

Byte Bag for Month will be Less than 20 million bytes

First Component of Local Workstation Worm Rule:

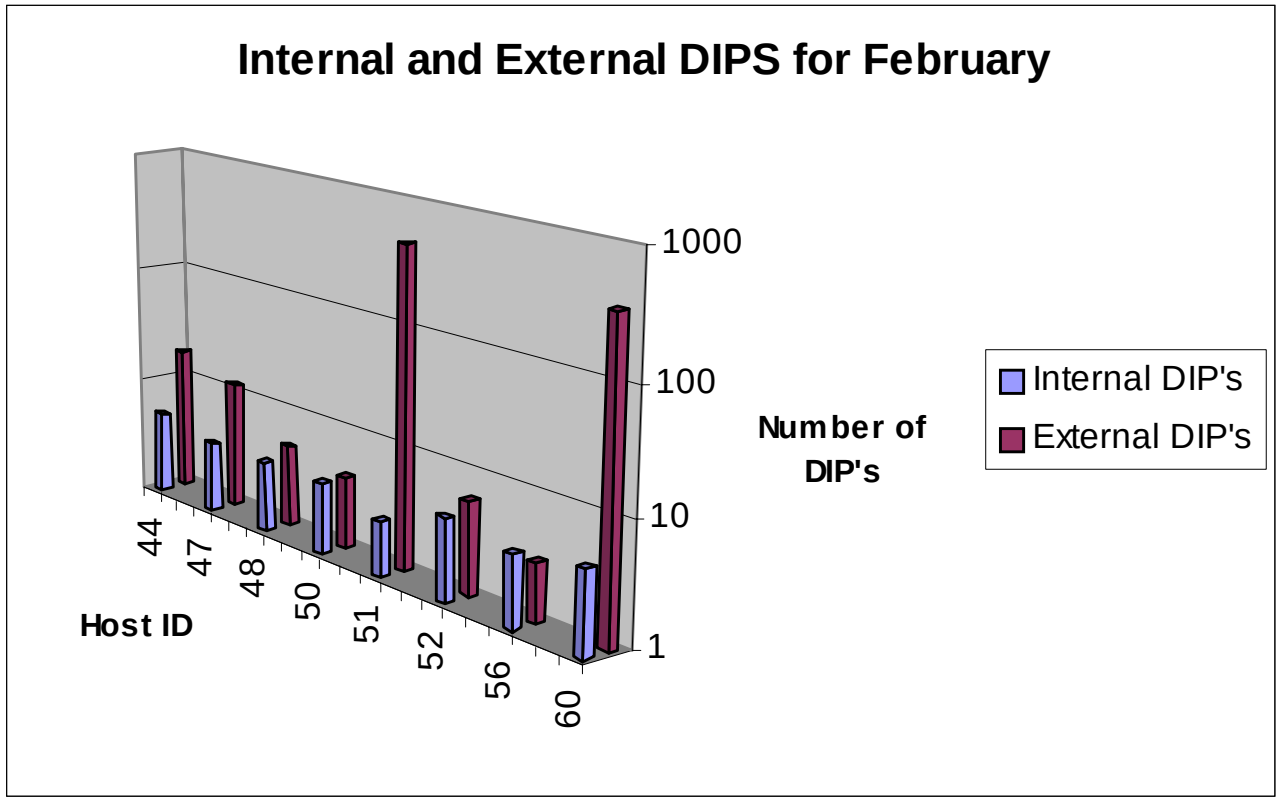
*Byte Bag for Month **may** be greater than 20 million bytes.*

Eight Workstation Hosts

Destination IP Bag Behaviour

Host ID	February Destination IP's		
	Internal	External	
44	5	17	
47	4	12	
48	4	5	
50	4	4	
51	3	467	
52	5	6	
56	4	3	
60	5	351	

Eight Workstation Hosts
Destination IP Bag Behaviour



Destination IP Bag Characteristics

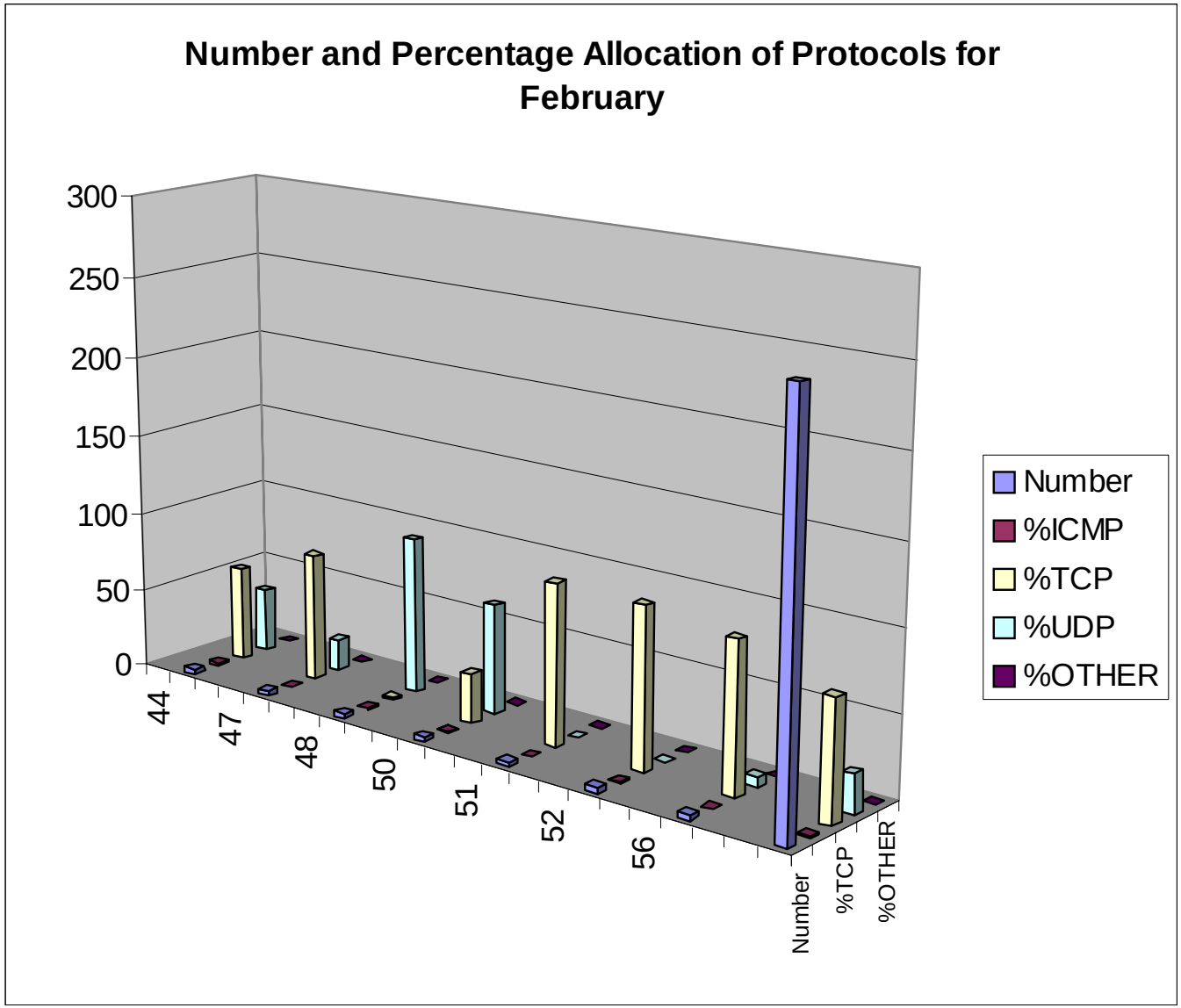
Second Component of Local Workstation BWB Rule:

Internal DIP's less than 10 per month and External DIP's less than 20 per month.

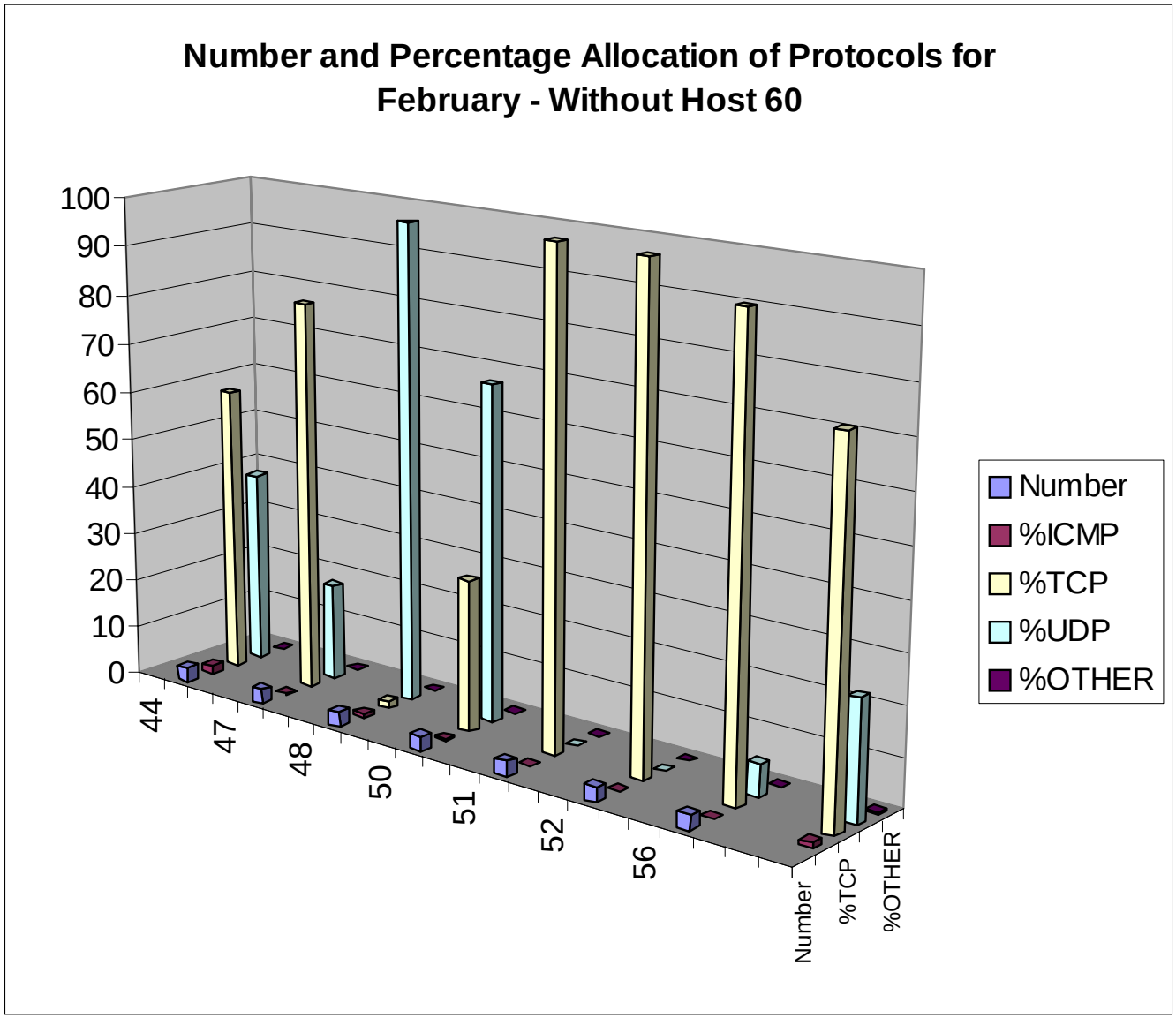
Second Component of Local Workstation Worm Rule:

External IP's contacted will greater than 20 per month.

Eight Workstation Hosts
Data Derived From Protocol Bag



Eight Workstation Hosts
Data Derived From Protocol Bag



Eight Workstation Hosts
Data Derived From Protocol Bag

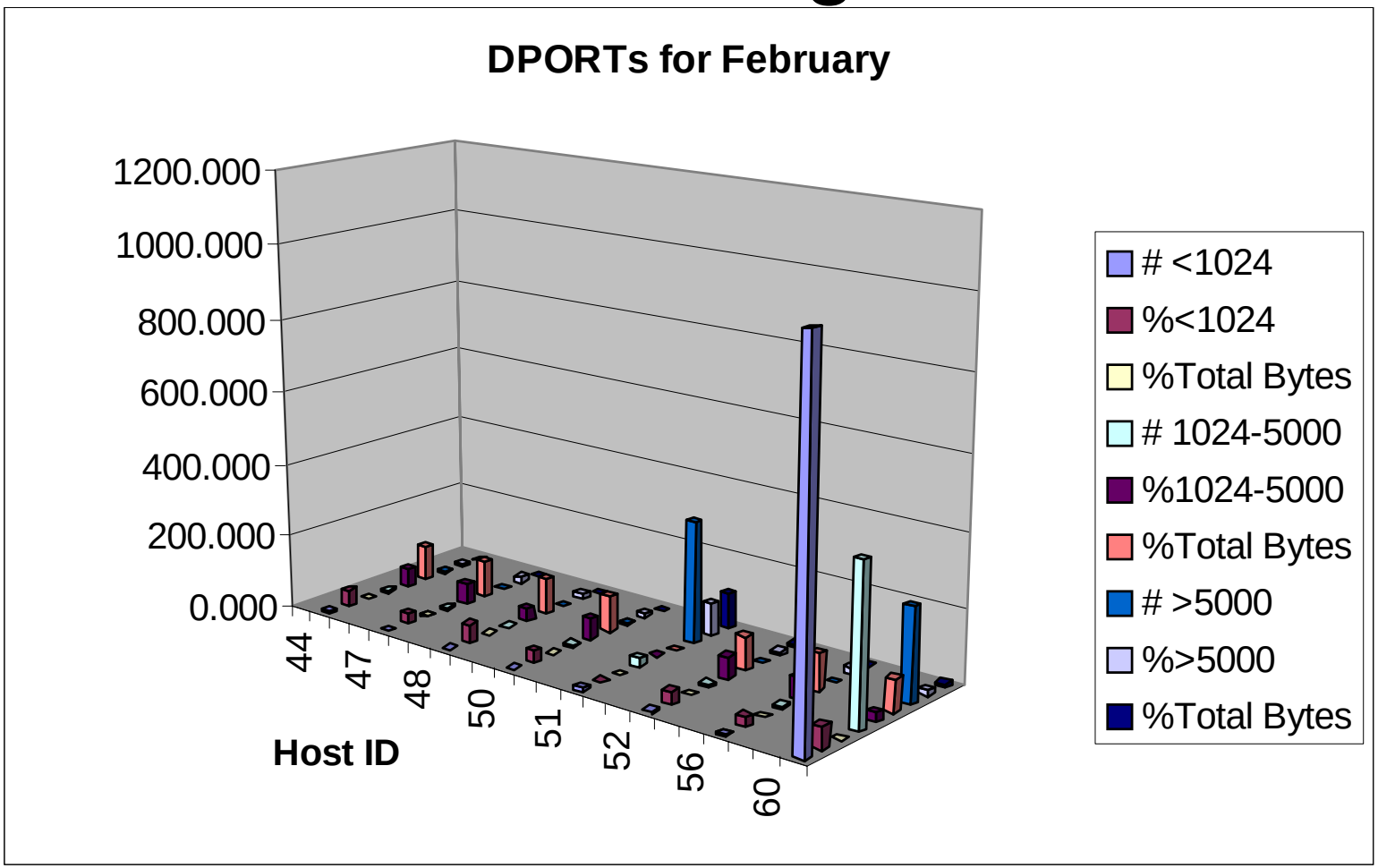
Relative Protocol Use
Third Component of Local Workstation BWB Rule:

Protocol Distribution will be as to TCP > 70%, UDP < 30% and ICMP <2% and Number of Protocols will be less than 5.

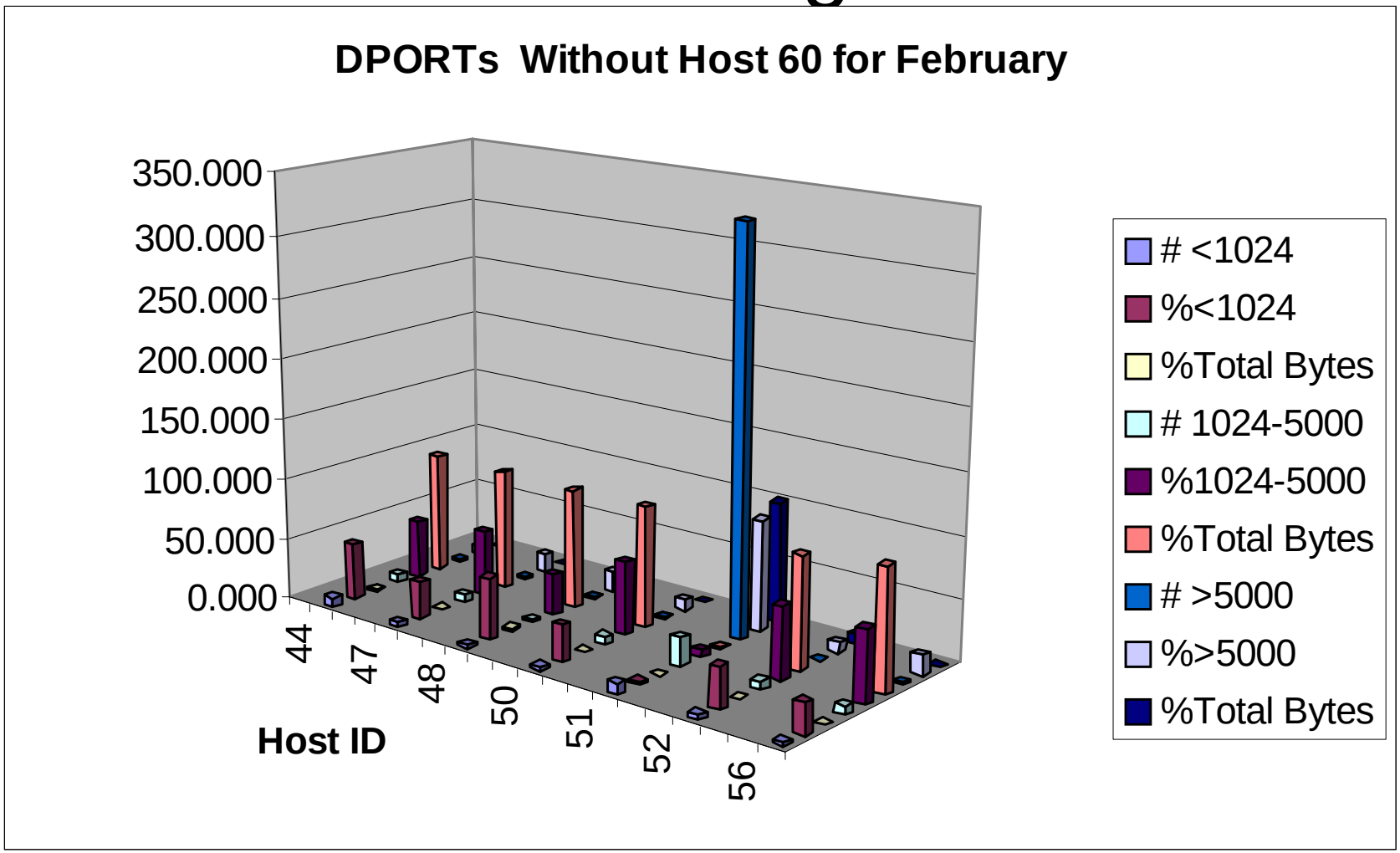
Third Component of Local Workstation Worm Rule:

Number of Protocols will be greater than 5.
(did not observer any greater than 4)

Eight Workstation Hosts
Data Derived From Destination Port Bag



Eight Workstation Hosts
Data Derived From Destination Port Bag



Eight Workstation Hosts
Data Derived From Destination Port Bag

Relative Destination Port Use
Fourth Component of Local Workstation BWB Rule:

NA -> Data Appeared to overlap so was deemed Not Applicable

Port # Range	# of Ports Accessed		%of Ports Access		%of Total Bytes	
	[BWB]	[WORM]	[BWB]	[WORM]	[BWB]	[WORM]
<1024	[< 7]	[> 7]	[20-50%]	[<20% || >50%]	NA	NA
1024-5000	[< 10]	[>10]	[>30%]	[<30%]	[>90%]	[<90%]
>5000	[< 5]	[> 5]	[<20%]	[>90%]	[<9%]	[>9%]

Local Baseline Workstation Behaviour (BWB)

Possible Workstation Rule for Classification

IF Bytes Transferred in one month < 20 million per month

AND Internal DIPs < 10 per month
AND External DIP's < 20 per month

AND Protocols: 1 < 2 %
6 > 70 %
17 < 30 %
Number of Protocols < 5

AND

Port Number Range	# of Ports Accessed	%of Ports Accessed	%of Total Bytes Traffic
<1024	< 7	20-50%	NA
1024-5000	< 10	>30%	>90%
>5000	< 5	<20%	<9%

THEN HOST is a user Workstation

Local Workstation Worm Behaviour

Possible Worm Rule for Classification

IF Bytes Transferred in one month > 20 million per month

AND Internal DIPs < 10 per month
AND External DIP's > 20 per month

AND Protocols: 1 < 2 %
6 > 70 %
17 < 30 %
Number of Protocols > 5

AND

Port Number Range	# of Ports Accessed	%of Ports Accessed	%of Total Bytes Traffic
<1024	> 7	<20 || >50%	NA
1024-5000	> 10	<30%	<90%
>5000	> 5	>90%	>9%

THEN HOST is a user Workstation Compromised by a Scanning Worm

Local MS Server Behaviour

Bytes per Month : > 93 million

DIP Bag:

Med-large Volume of external hosts
All internal Hosts on the same subnet as the Server host are contacted.
Uniform Medium-large level of byte volume to internal hosts.
Lesser byte volume to external hosts but still somewhat uniform.

Proto Bag

1	2%
6	65 %
17	33%

Dport Bag

Medium level of near uniform byte volume across every DPORT

Local Laser Printer Behaviour on MS Network

Bytes Per month: < 3 million

DIP Bag

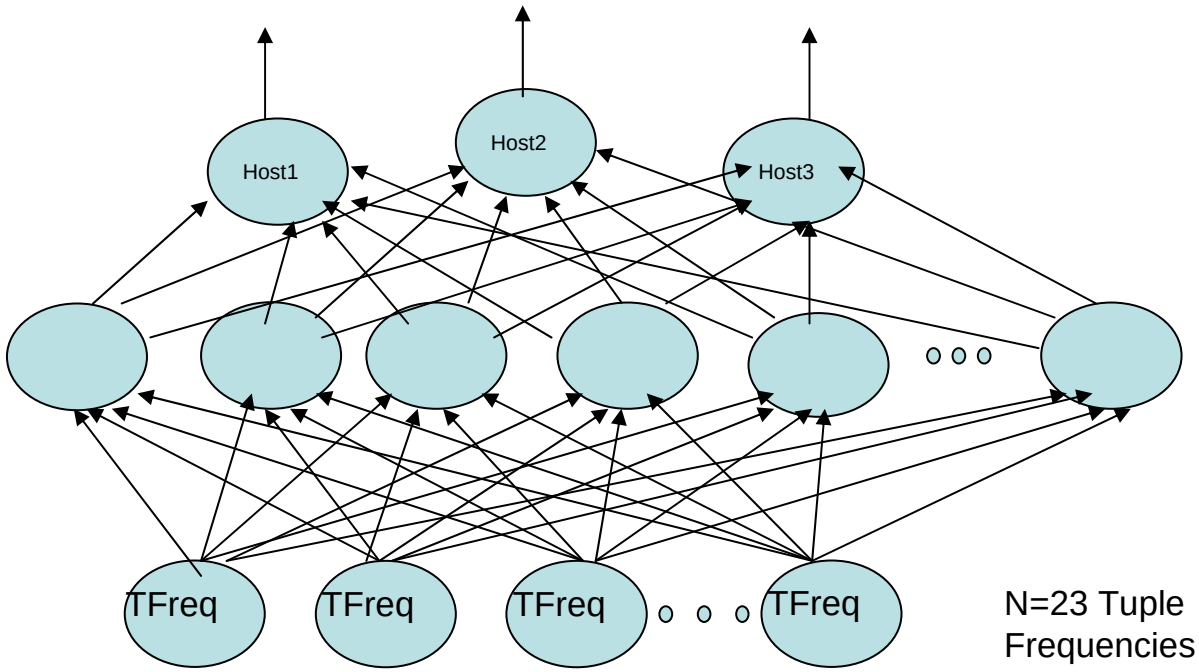
All Traffic sent to ~3 internal Hosts and ~5 external hosts
Internal traffic is medium-large and more-or-less uniform.
External traffic is small and somewhat uniform on targeted hosts.

Proto Bag	1	0.4%
	6	99.6%
	17	0%

Dport Bag

No regular Traffic below 1032 with the exception of some traffic at port 0.
All traffic is medium level bytes uniformly distributed across Ports 1032-58,361.
Ports are not sequential but appear to be clustered in specific ranges.

Uniqueness of Minimal Information A Neural Classification Approach



Each input frequency vector contains an observed frequency for each tuple for a 24 hour period.
Each tuple is defined as Protocol, Destination Port, Byte Range.
All observed Workstation (BWB) hosts could be described by a 23 element Vector.

Results of Neural Testing

Host ID	Day	Output Vector	Classification (Hit/Miss/Unknown
1 [0 1 0]	1	[0.04 0.86 0.08]	HIT
	2	[0.17 0.97 0.00]	HIT
	3	[0.10 0.91 0.02]	HIT
	4	[0.09 0.95 0.01]	HIT
2 [1 0 0]	1	[0.95 0.06 0.00]	HIT
	2	[0.96 0.04 0.00]	HIT
	3	[0.95 0.06 0.00]	HIT
	4	[0.95 0.07 0.00]	HIT
3 [0 0 1]	1	[0.00 0.09 0.92]	HIT
	2	[0.00 0.00 0.99]	HIT
	3	[0.00 0.12 0.92]	HIT
	4	[0.00 0.00 0.99]	HIT

Results of Neural Testing

Host ID	Day	Output Vector	Classification (Hit/Miss/Unknown)
1 [0 1 0]	5	[0.03 0.57 0.25]	HIT (barely)
2 [1 0 0]	5	[0.47 0.92 0.00]	MISS (anomaly detected)
3 [0 0 1]	5	[0.00 0.00 0.99]	HIT

Anomaly detected in the behaviour of Host 2 (misclassifying it as Host 1).
Network 47% confident that behaviour is coming from Host 2.
Network 92% confident that Host 2 is really Host 1
Network only 57% confident that Host 1 was in fact Host 1.
The anomaly in question was created when the owner (regular user) of Host 1 moved to work on Host 2 for part of the day. If these artifacts are more then just Coincidence (and I must strongly state that this is not proven)
then the network *may* have detected this movement and correlated the functioning of the workstation to the user;
Thereby discovering which user was working at Host 2.

Conclusion

The authors feel that they have satisfied the original fundamental research question described in the abstract, that being that useful clusters of network host behaviour can indeed be described by the use of flow record data alone, while recognizing the severely limited locality of the test data. Questions of scalability and portability of the results remain unanswered.