One Year of Peer to Peer

Ron McLeod, BCSc, MCSc.
Director - Corporate Development Telecom Applications Research Alliance
Doctoral Student, Faculty of Computer Science, Dalhousie University

Presentation Summary

This presentation will profile the result of the growth in peer-to-peer applications on a sample network and describe the resultant massive increase in the diversity of traffic. This diversity impacts the ability to profile baseline normative behaviour using Blind Flow Analysis.

I will also briefly discuss the application of SiLKtools, Neural Networks and Bioinformatic strategies to Blind Flow Analysis of real world security problems and how that analysis is affected by the growth in recreational/user driven applications.

What began as a basic design principal of end-to-end management with popular applications in recreational computing is quickly becoming a dominant evolutionary force in network traffic patterns.

Traffic patterns are becoming emergent properties influenced by the voluntary adoption of new systems by individuals without any collective intent.

The network is evolving at the edges.

“Peer-to-Peer is the basic design of the Internet” – Christian Huitema

Sample Network Description

- A Multi-tenant Commercial Network consisting of:
  - ~ 40 user assigned hosts, actual number subject to minor fluctuations over time.
  - ~40 special hosts not assigned to individual users. These hosts form parts of various temporary development and experimental environments.
  - Users were apprised that Network flow data was now being captured for experimental and management reasons.
  - Payload data was neither collected nor examined.
  - Analysts did not have access to the content of specific hosts for further investigation.
  - For confidentiality reasons the identity of the Network is not specified in this Presentation.

A Review of Blind Flow Analysis

The Need for Classification Based on Minimal Information (the extreme case in the world of tomorrow)

- Capturing and examining payload contents is widely viewed as a potential violation of privacy and placed in a category similar to listening in on a telephone call.
- Even attempts to use information derived from the payload (such as ngrams) do little to alleviate the fundamental concern of the user surrounding access to the payload.
- In multi-tenant commercial environments this user concern may be based in protection of commercial confidentiality.
- There is less (although not zero) concern among the user community with regard to the capture and investigation of packet header data (some concern for Source and Destination IP's and MAC's).
- Therefore, the network analyst may be limited to examining a severely reduced subset of the packet header information in an attempt to determine if the system under their management (or monitoring) is operating properly or experiencing anomalous behavior.
- The loss of access to the originating address information means that the analyst no longer has access to a unique field in the data that identifies the individual hosts in the traffic (i.e. they cannot tell one computer from another by looking at the remaining flow record traffic alone).
- In such an environment, what is required is a method of classification that relies on minimal information and the development of traffic flow behaviour models that use only this information.

One Strategy for Comparing A Suspicious Host to a Standard Workstation Using Blind Flow Analysis

Local Baseline Workstation Behaviour (BWB)	**Suspicious Host**
Bytes Transferred in one month < 20 million per month	45 billion per month
Internal DIPs < 10 per month	3 per month
External DIPs < 20 per month	1.74 million per month
Protocols: 1 < 2 %	1 1 %
6 > 70 %	6 9 %
17 < 30 %	17 90 %
Number of Protocols < 5	3

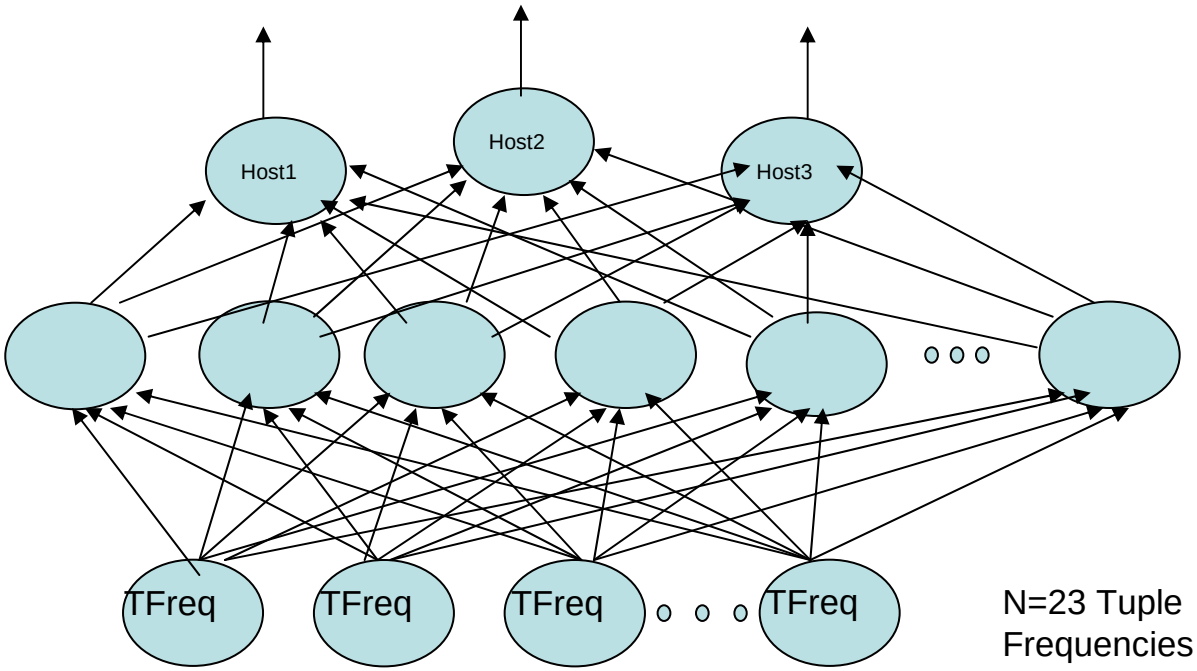
Port Number Range	# of Ports Accessed	%of Ports Accessed	%of Total Bytes Traffic	# of Ports Accessed	%of Ports Accessed	%of Total Bytes Traffic
<1024	< 7	20-50%	<1%	45	0.07%	
1024-5000	< 10	>30%	>90%	3,976	6%	1%
>5000	< 5	<20%	<9%	60,059	93%	99%

Impact of Peering Traffic on Blind Flow Analysis and the Uniqueness of Minimal Information

- In early 2006 Neural Network was used to classify workstation traffic based on a localized “Workstation Genome”.
- It was found workstation behaviour could be fully described by a set of 23 unique 3-tuples formed by the combination of Protocol, Destination Port, and Byte Range ID – Where Byte Range ID was one of five levels given by:

Bytes	Range
0 – 100	1
100 – 999	2
1000 – 9,999	3
10,000 – 49,999	4
50,000 +	5

Impact of Peering Traffic on Blind Flow Analysis and the Uniqueness of Minimal Information



Each input frequency vector contains an observed frequency for each 3-tuple for a 24 hour period.
Each 3-tuple is defined as Protocol, Destination Port, Byte Range.
All observed Workstations could be described by a 23 element Vector.

Impact of Peering Traffic on Blind Flow Analysis and the Uniqueness of Minimal Information

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

100% Success rate on uniquely classifying a small sample of the population

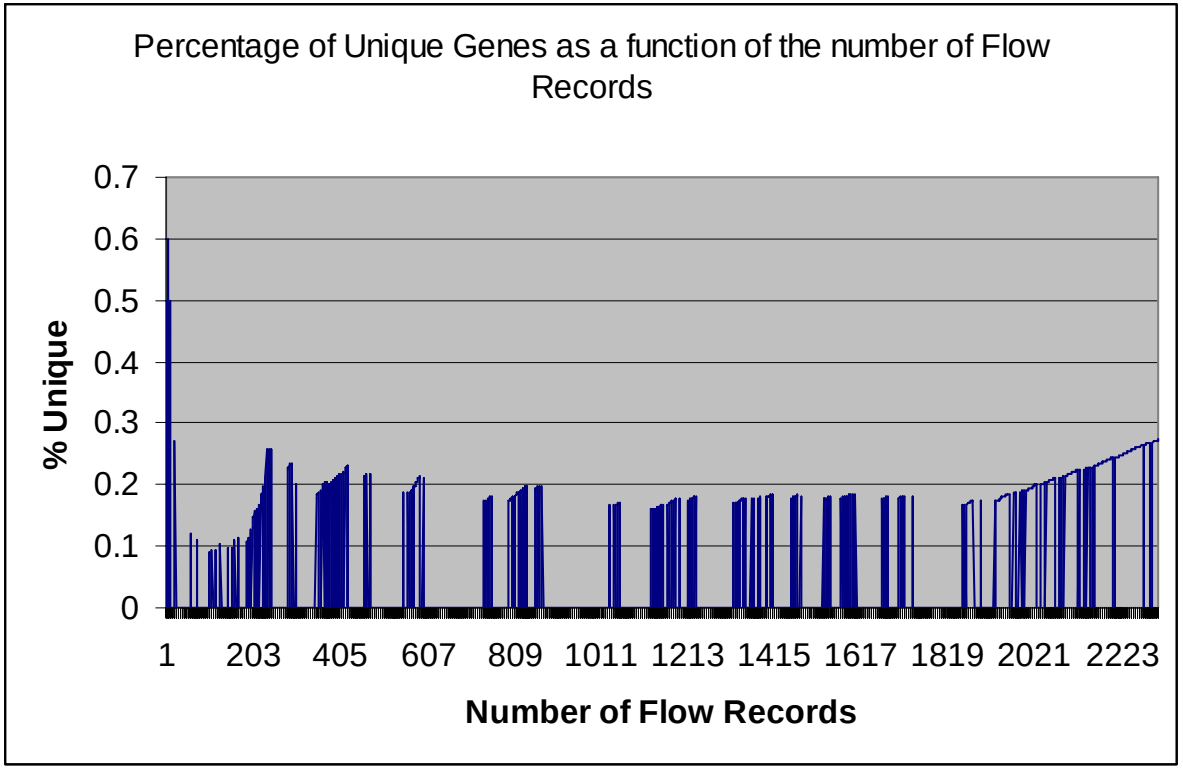
Impact of Peering Traffic on Blind Flow Analysis and the Uniqueness of Minimal Information

- In early 2007 a similar population of workstations was chosen with the goal of testing a Support Vector Machine approach to classification.

- *To the great surprise of the author, the number of unique 3-tuples required to uniquely describe the Workstation Genome had risen from 23 to over 600 in 16 months.*

- Subsequent investigation showed that the diversity of the observed behaviour increased as a function of both population size as well as the length of the sampling period.

Impact of Peering Traffic on Blind Flow Analysis and the Uniqueness of Minimal Information



By limiting the traffic to ICMP and TCP flow records, the number of unique tuples required to adequately describe the population reached a steady state of approximately 18% of the total number of all expressed tuples.

When UDP traffic was introduced into the sample, the percentage of unique tuples in the population did *not* reach a steady state in proportionality but rather the number of the unique tuples increased in linear proportion to the number of total tuples observed.

Impact of Peering Traffic on Blind Flow Analysis and the Uniqueness of Minimal Information

- What happened to the network traffic to create such diversity in such a short period of time?

- Expected monthly unique destination IPs =1200 (40 hosts * 30 external and internal DIP contacts).

  Actual values:

  Average monthly destination IPs = 140,000
  Average monthly number of flows = 2.8 million
  Average monthly byte volume of approximately 31 billion

- In addition to unusual volumes, two fundamental behaviours changed.
  - Protocol Ratio
    - From TCP 70% UDP 30%
    - To TCP 50% UDP 50%
  - Use of Unique Destination Ports by Workstations now parallels Server behaviour.

One Year of Peer-to-Peer

Much has been written lately of the growth and deployment of Peer-to-Peer Protocols

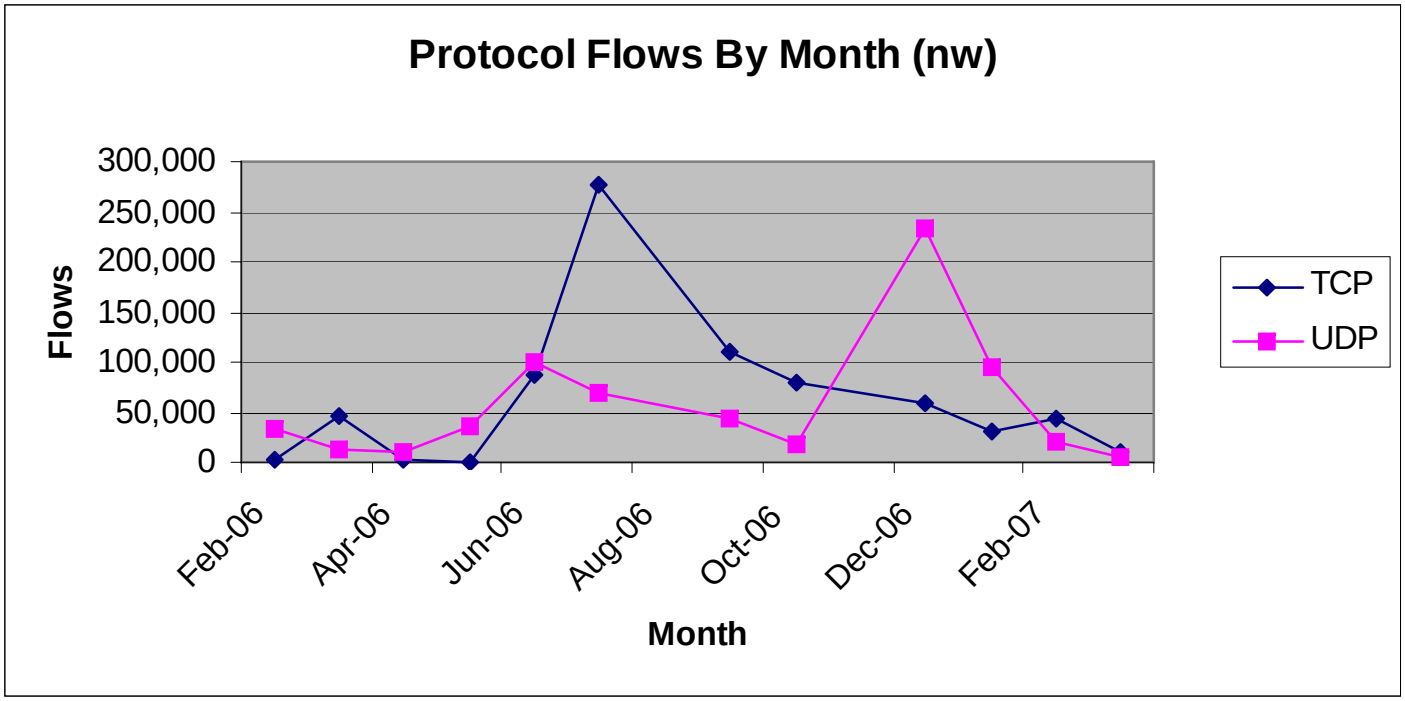
Recommended reading *“Transport Layer Identification of P2P Traffic”, Thomas Karagiannis, et al, IMC’ 04, 2004, Taorimina, Italy.*

Perhaps Peer-to-Peer is the culprit.

Decided to check for the presence of known P2P in the traffic

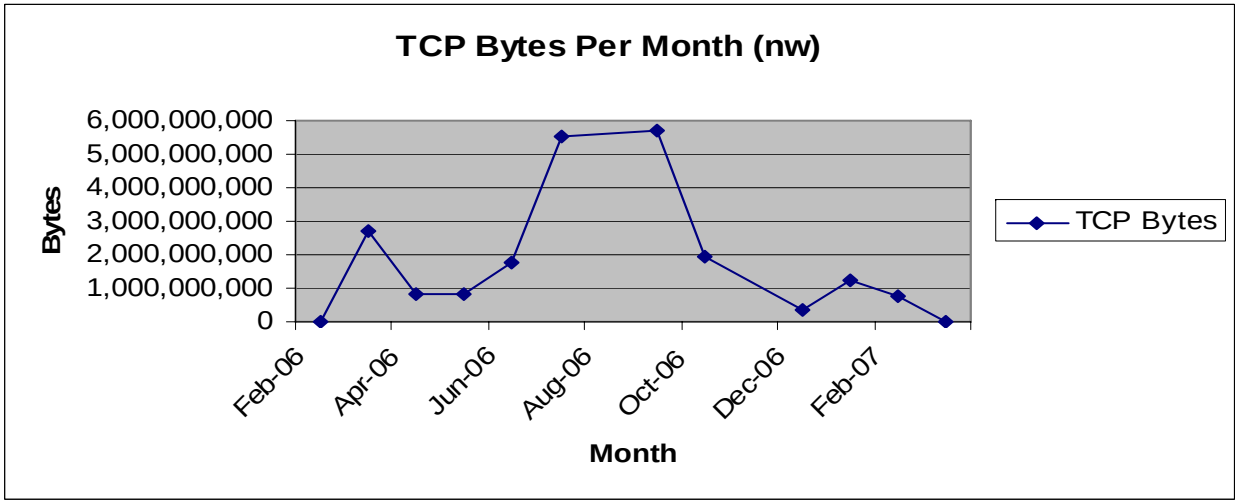
eDonkey2000
Fasttrack
Bittorent
Gnutella
MP2P

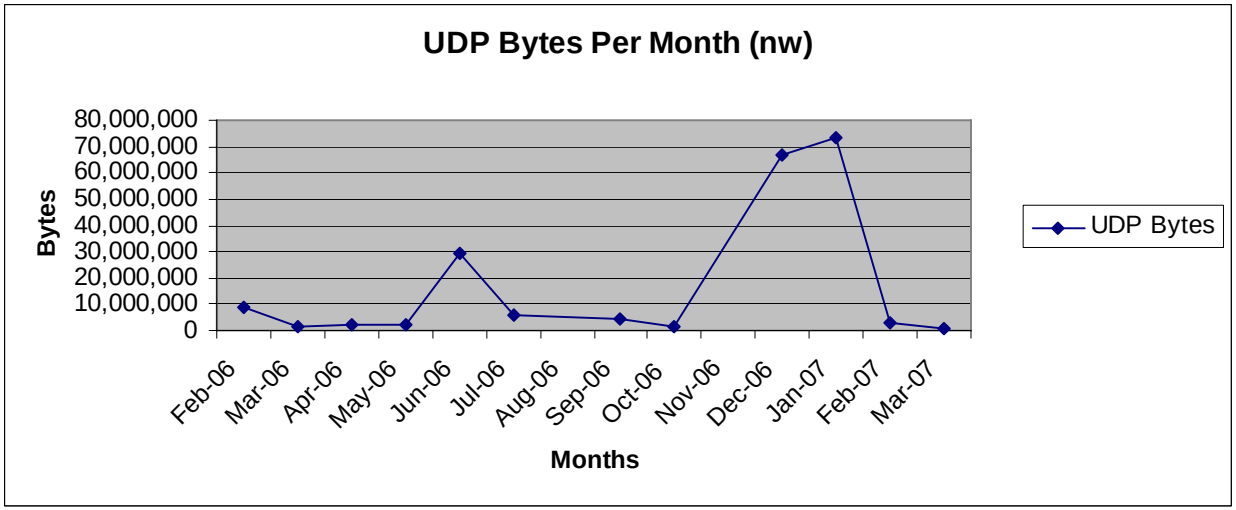
One Year of Peer-to-Peer



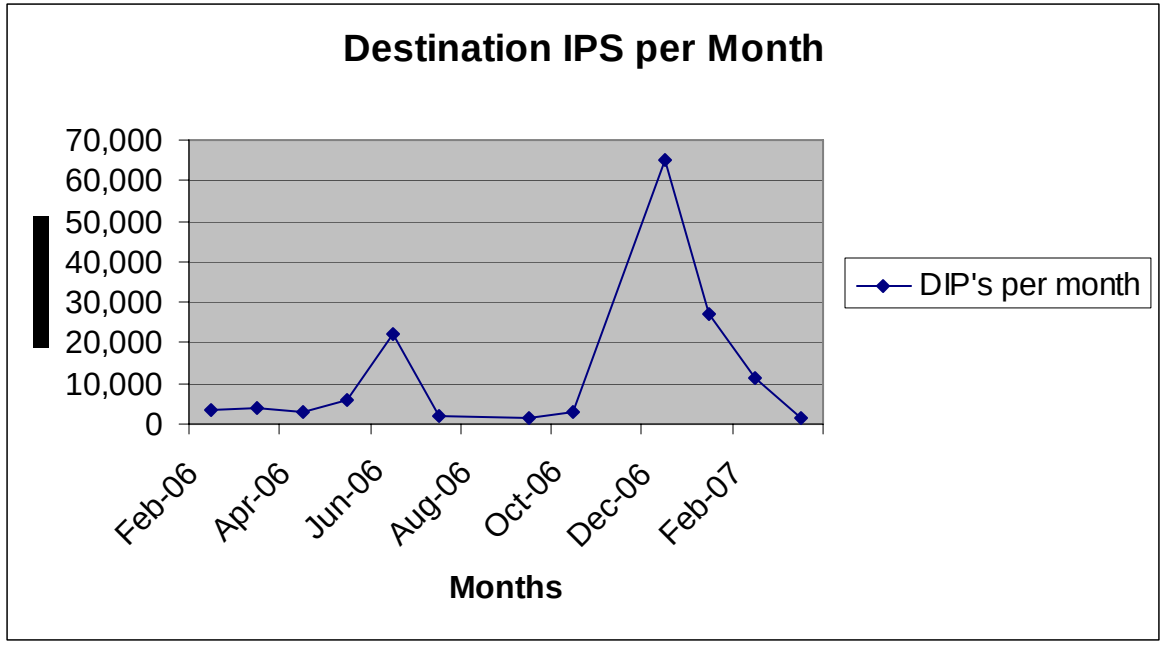
The graph above shows the pattern of flows by protocol for one year for the Target network.

One Year of Peer-to-Peer



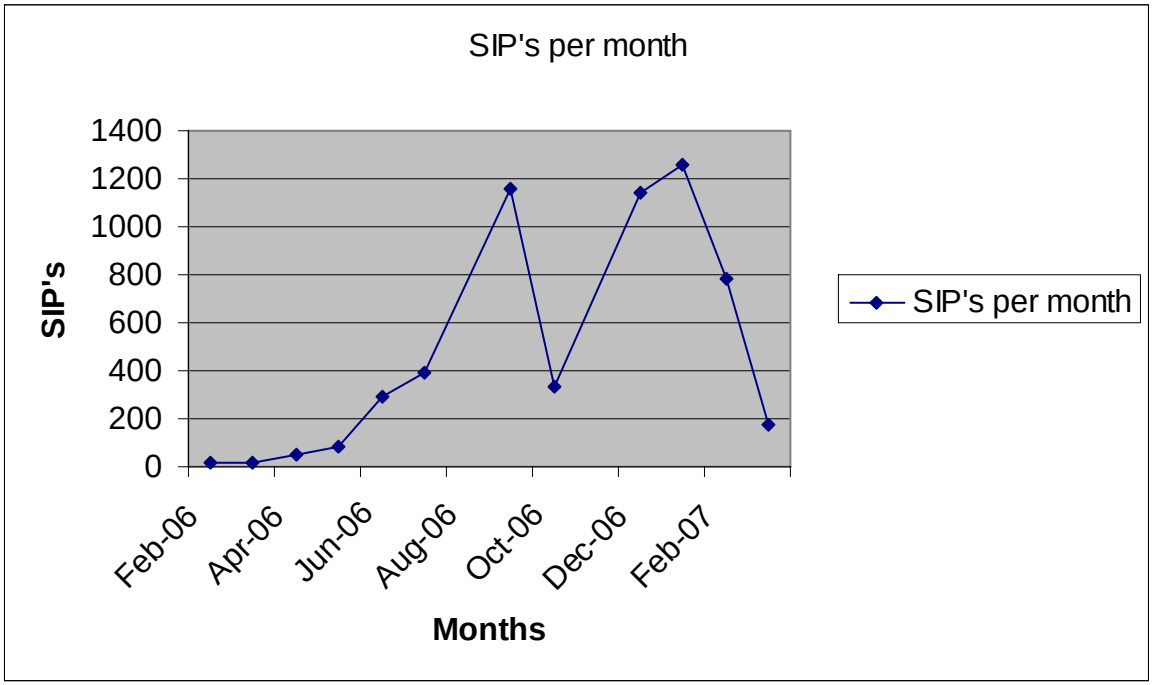


One Year of Peer-to-Peer



For a small network they talked to quite a few friends.

One Year of Peer-to-Peer

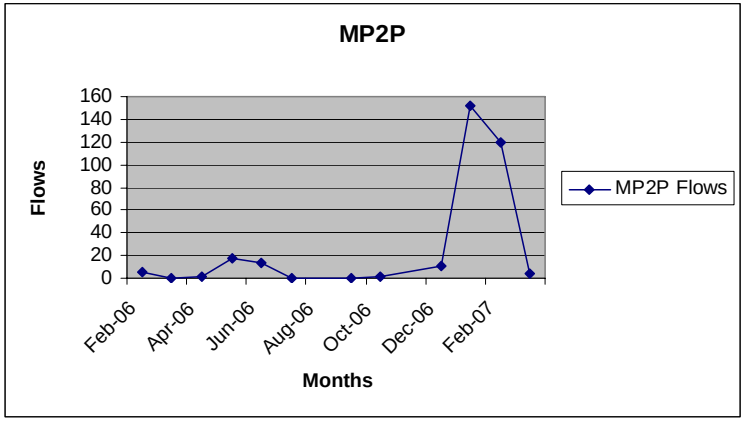


The feeling was mutual.

One Year of Peer-to-Peer

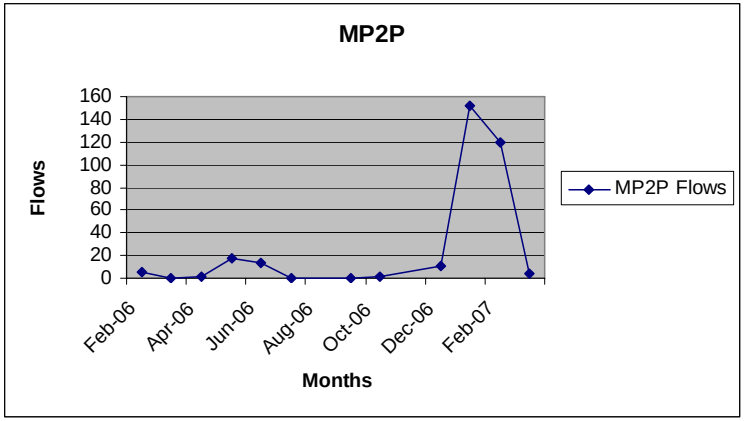
Let’s consider the traffic contribution for each P2P Application in the table.

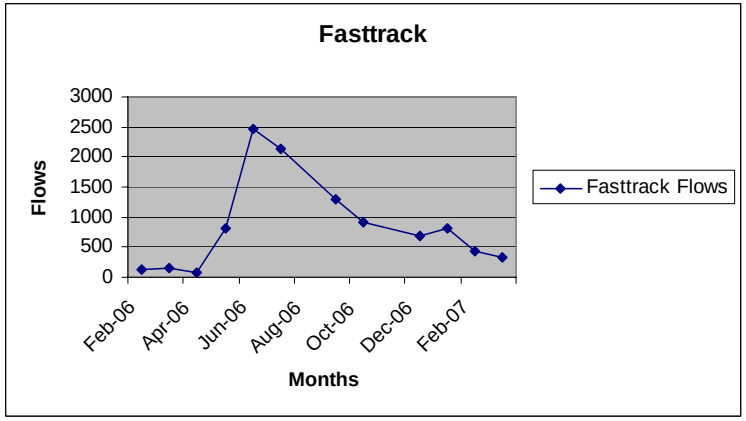
One Year of Peer-to-Peer



MP2P, or Manolito, is a P2P system primarily used to share music files. MP2P traffic was the least contributor to the overall network traffic among the observed systems. This traffic reached a peak flow count of just under 160 in January 2007.

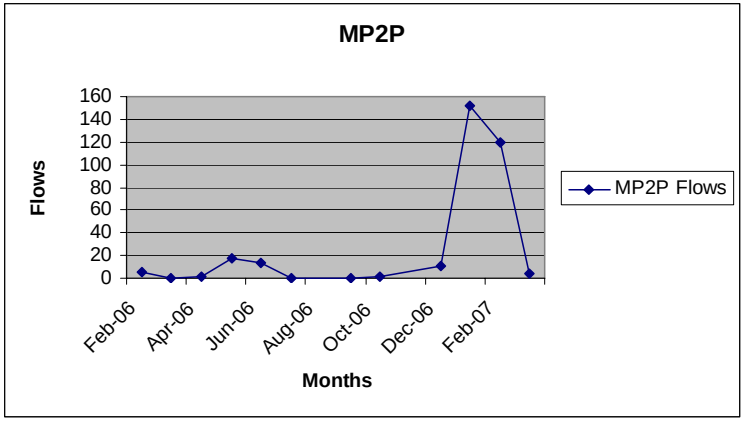
One Year of Peer-to-Peer

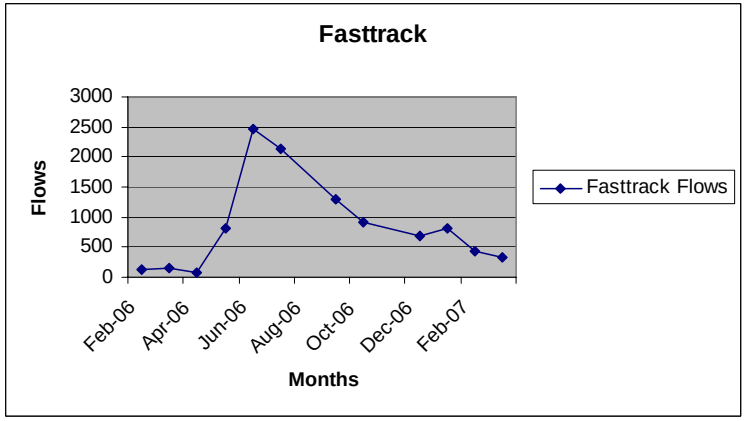


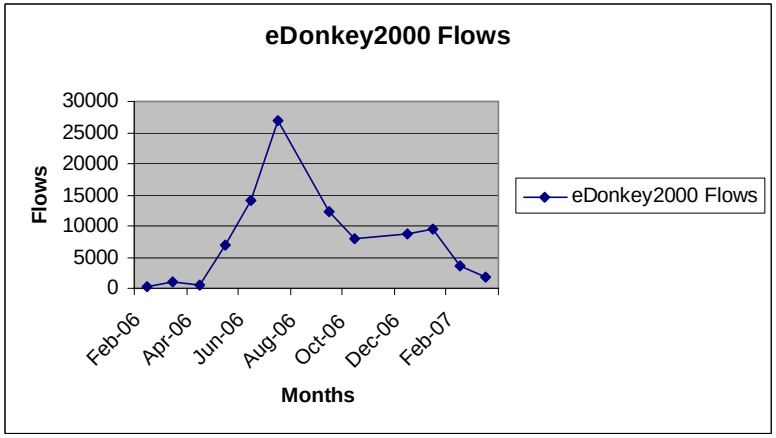


The Fasttrack P2P system is primarily used by Kazaa and its variants to exchange mp3 music files. Fasttrack traffic reached a peak flow count of 2,500 in July 2006.

One Year of Peer-to-Peer

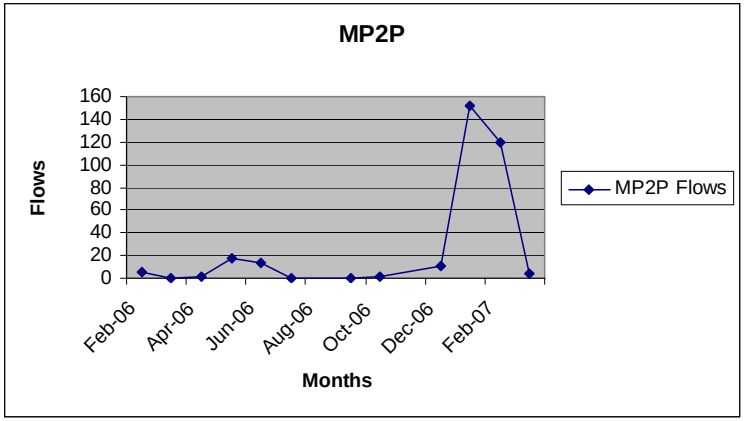


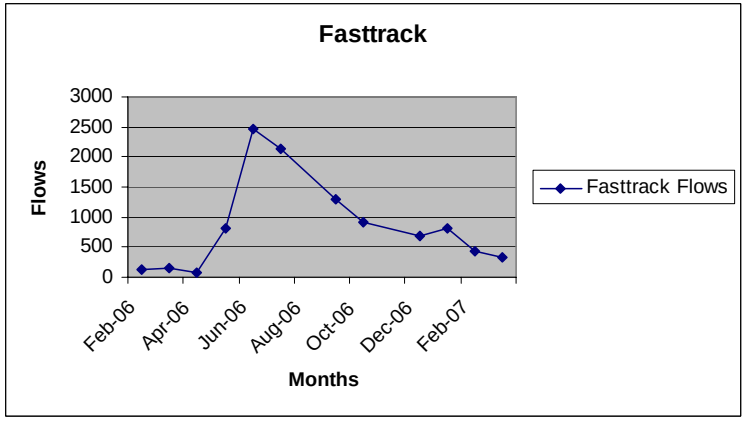




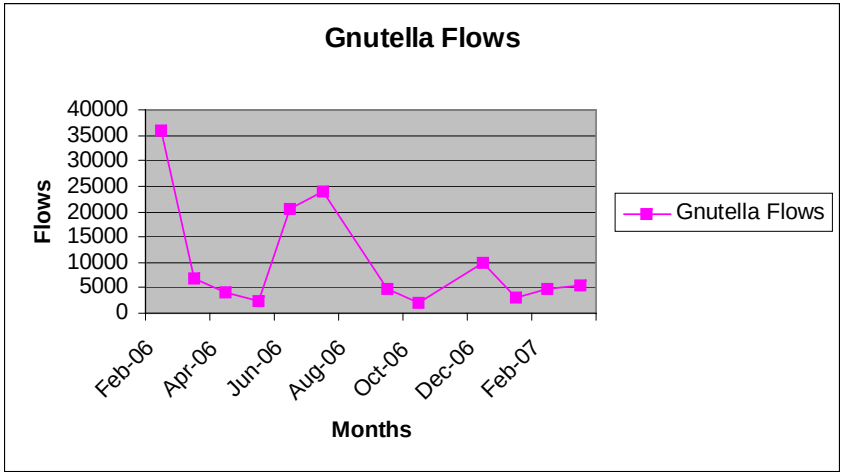
EDonkey2000 was a peer-to-peer system primarily used to distribute large images, video games and software. Although officially discontinued in September 2005 due to legal action brought by the Recording Industry Association of America (RIAA), we speculate, based on our profiling, that we observed eDonkey2000 communication during 2006. EDonkey traffic passed 25,000 flows in July 2006.

One Year of Peer-to-Peer

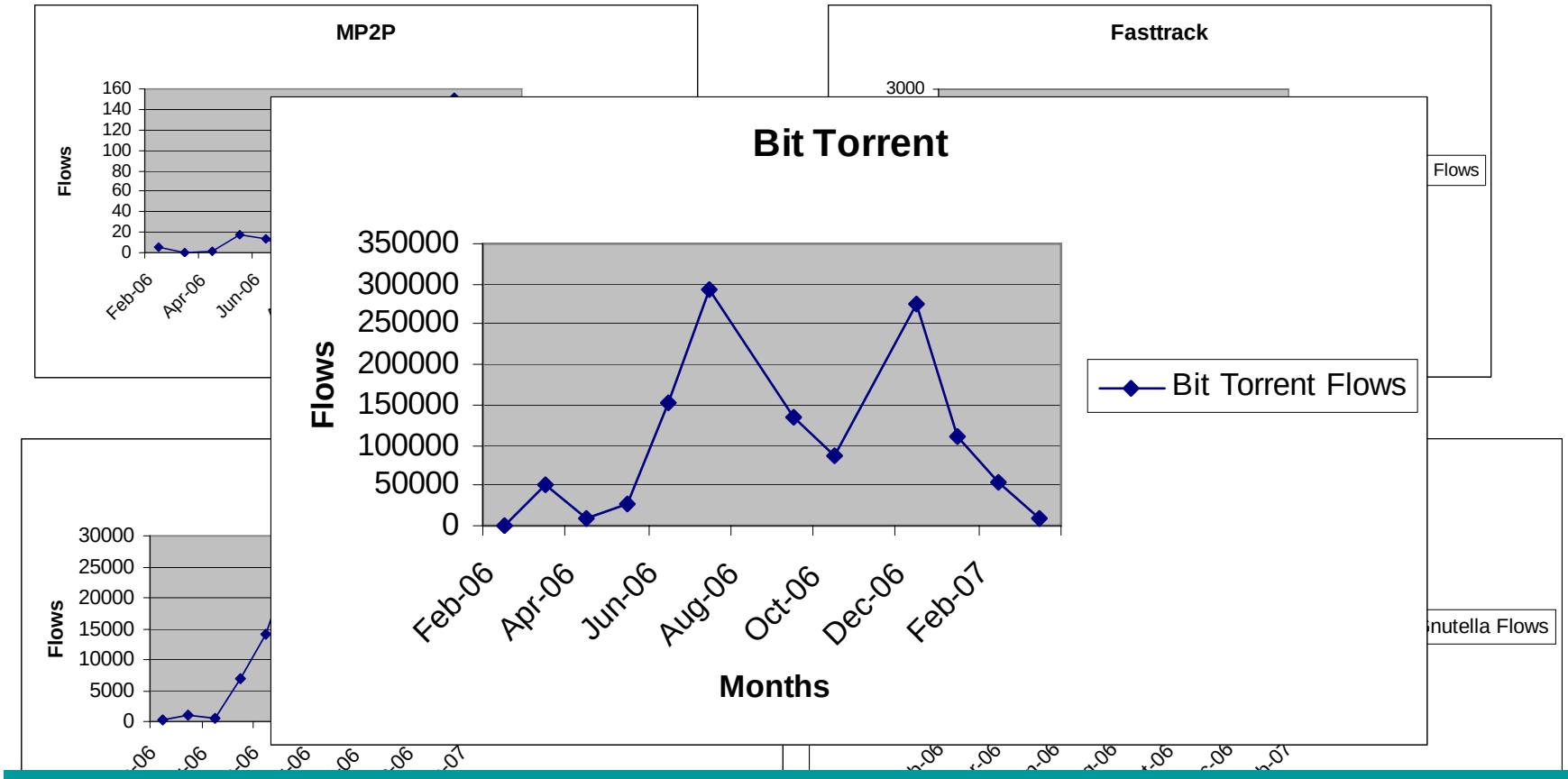




Gnutella is a multi-tier Peer based file exchange system. Traffic from Gnutella ranged from 5,000 to 35,000 flows per month.

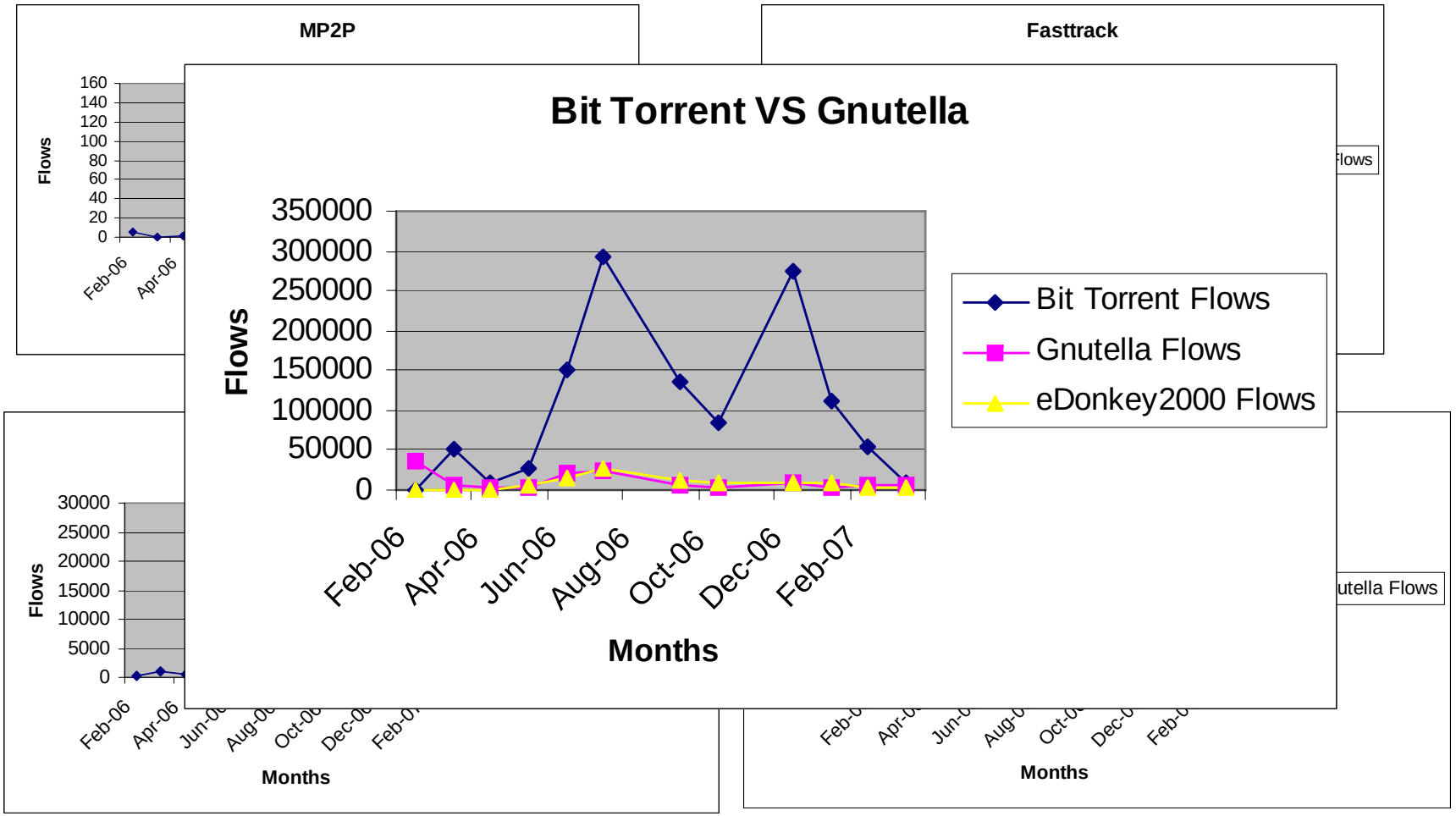


One Year of Peer-to-Peer



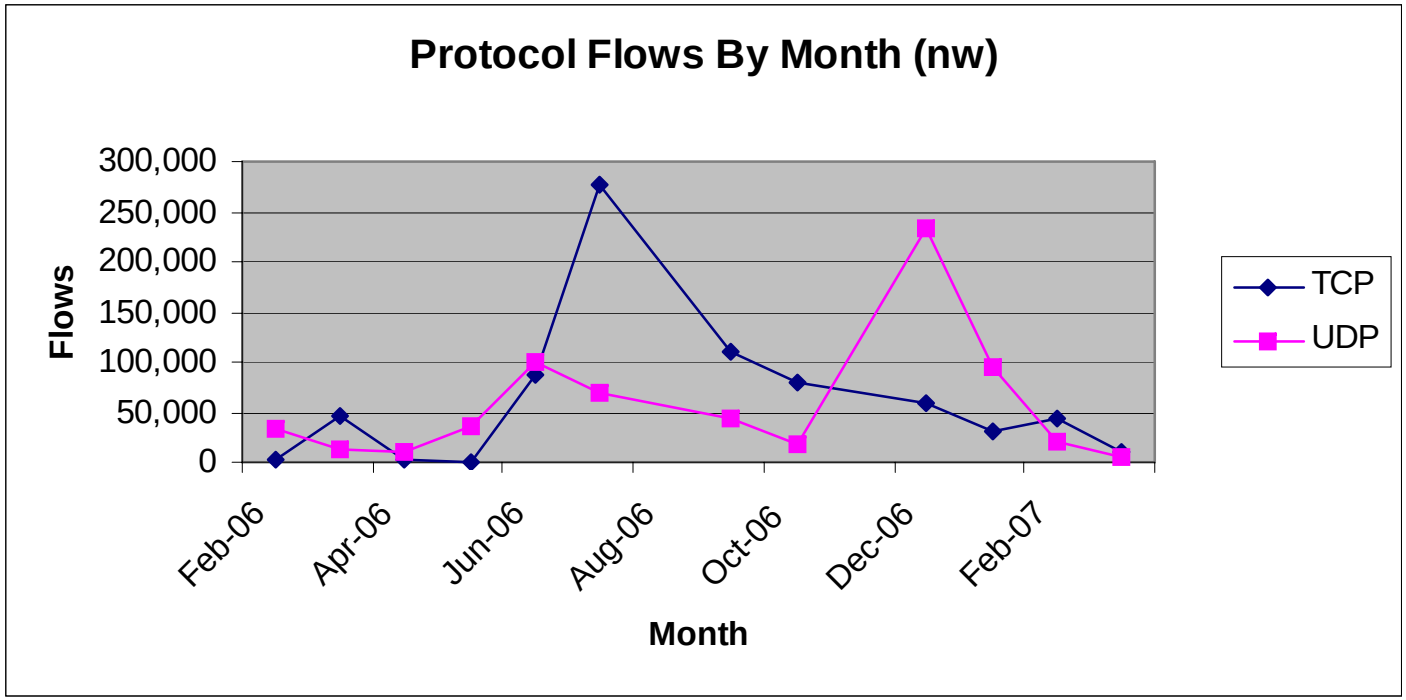
BitTorrent is an ever increasing popular P2P system used for exchanging large data files. Many open source software releases are distributed using BitTorrent. It is also used to distribute legal movie and music downloads. BitTorrent traffic eclipsed most P2P traffic at 300,000 flows.

One Year of Peer-to-Peer

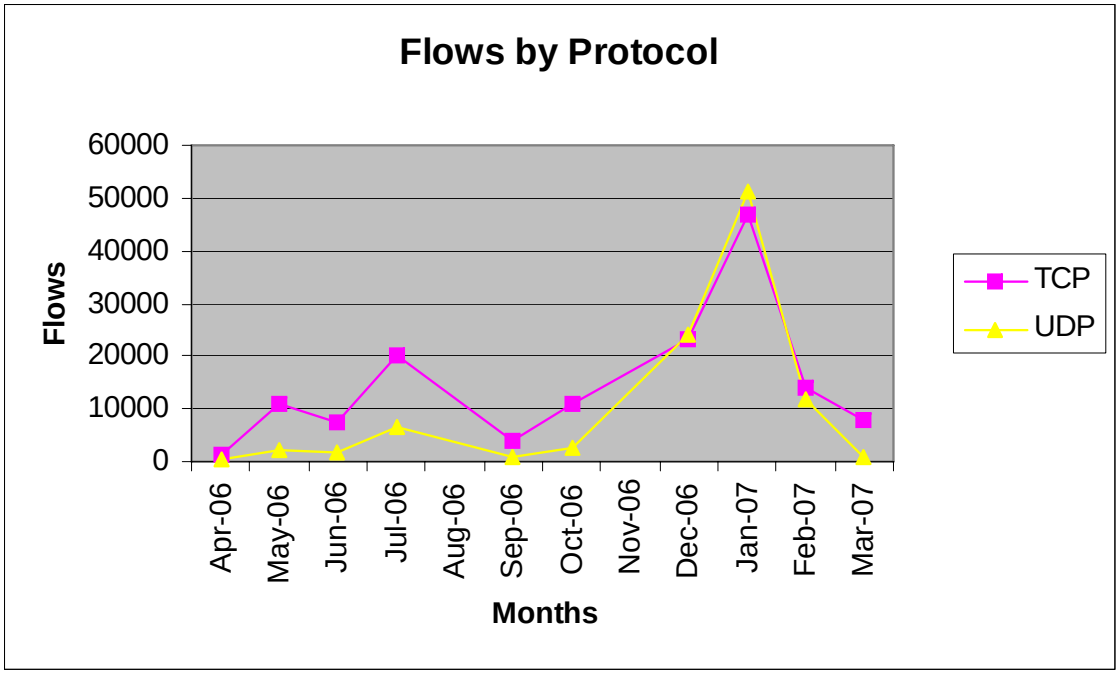


One Year of Peer-to-Peer

Unfortunately the overall Peer-to-Peer flow pattern did not match the pattern that we were seeking. That being a 50/50 ratio of TCP to UDP.

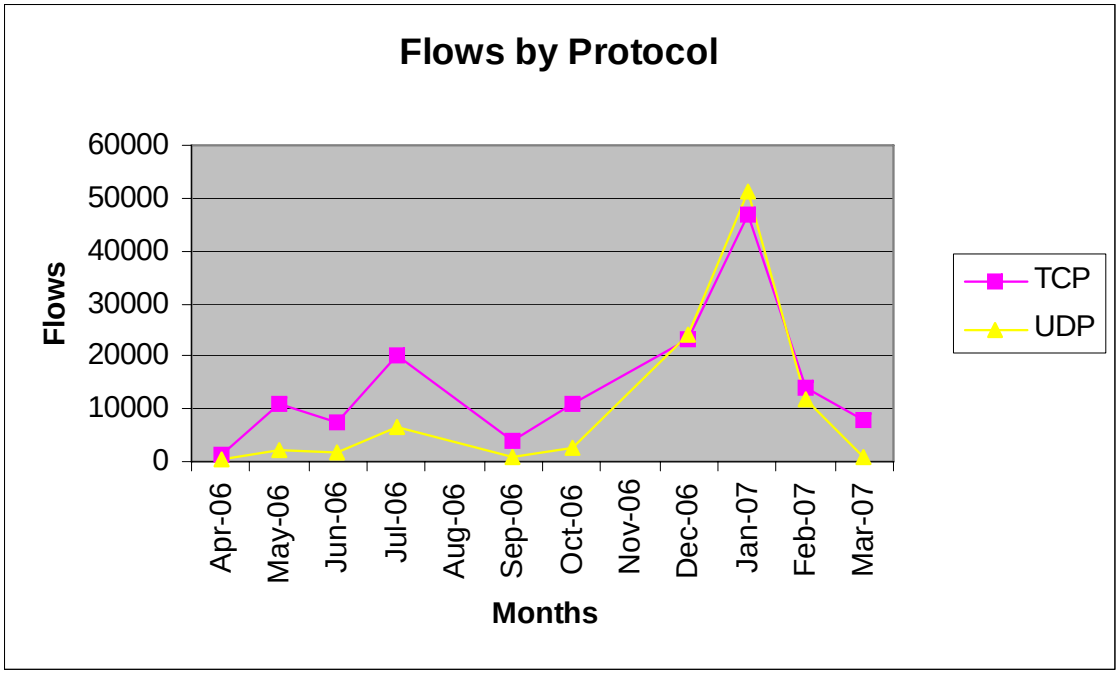


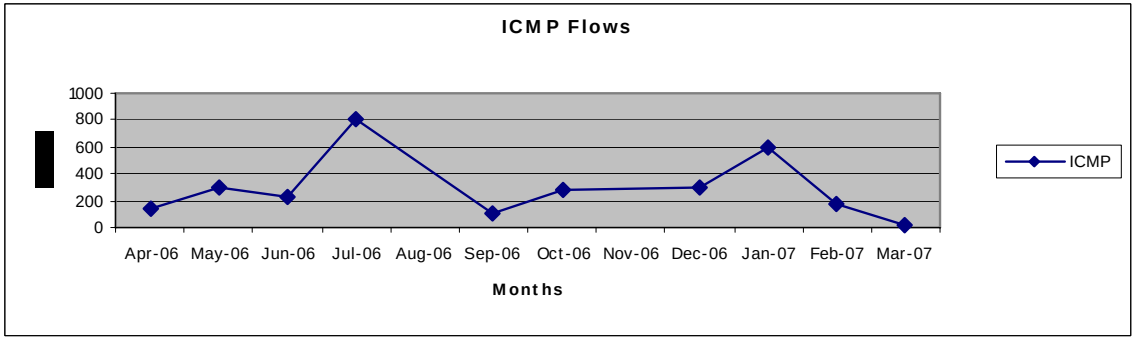
One Year of Peer-to-Peer



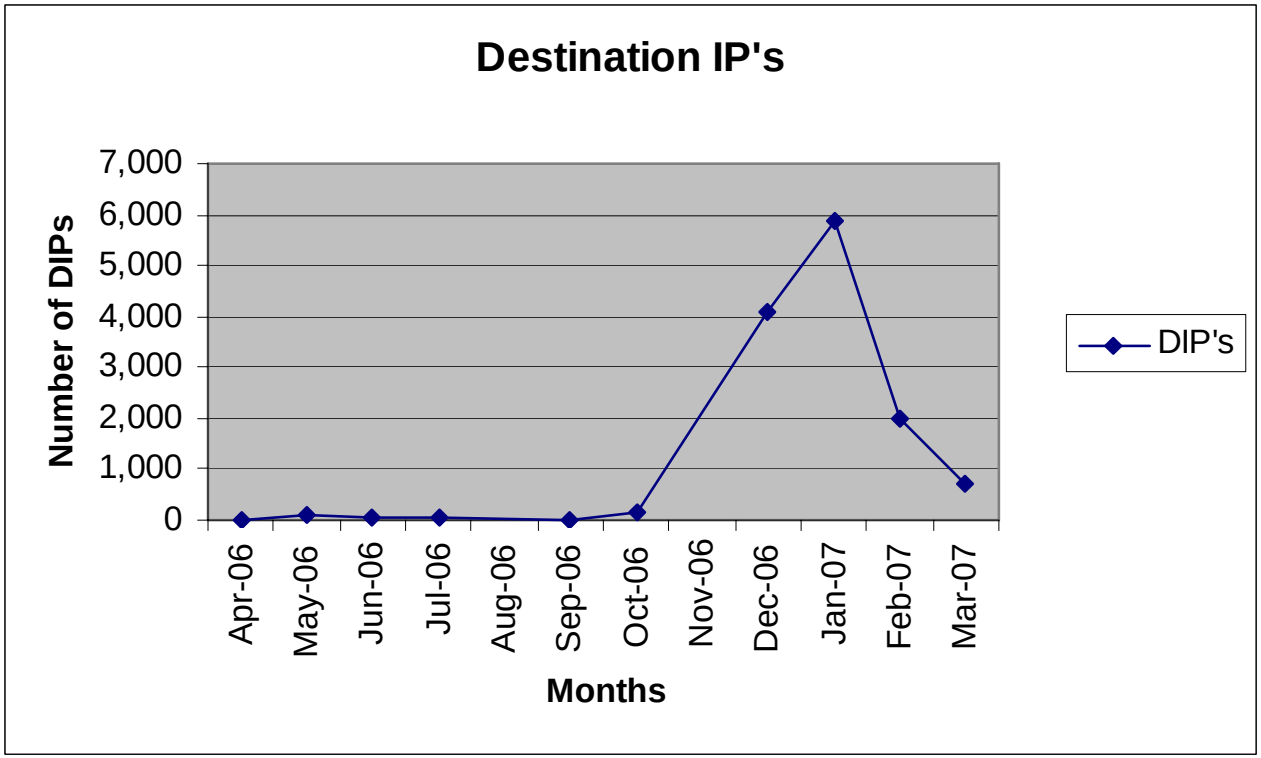
The graph above shows the pattern for which we were searching. This is the traffic from a single user workstation, with a peak flow count of 50,000 flows per month.

One Year of Peer-to-Peer



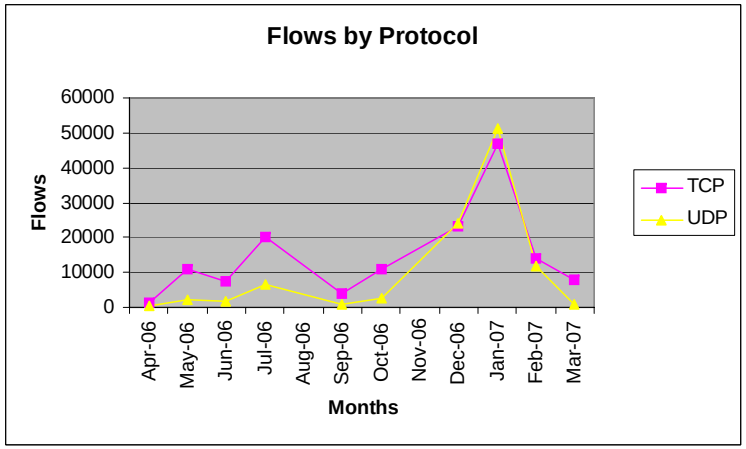


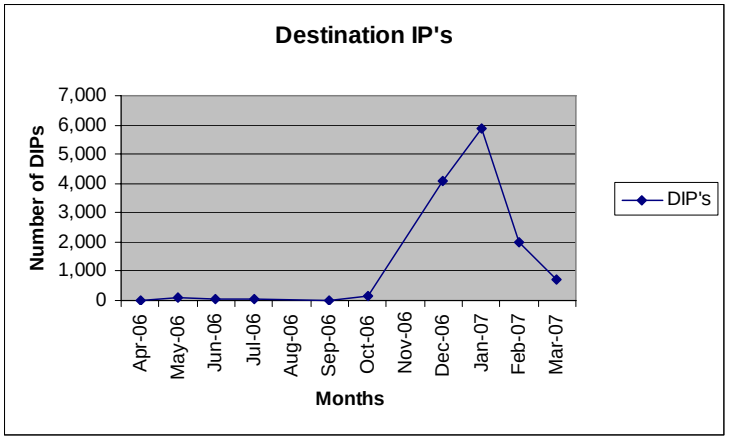
One Year of Peer-to-Peer



This workstation changed its behaviour in late fall 2006 from talking to less than 100 DIPs per month to 6,000 DIPs per month.

One Year of Peer-to-Peer





Who am I ?

One Year of Peer-to-Peer

SKYPE

This traffic pattern is driven by the adoption of Voip by a single user in the target network.

Disclaimer: It is important to point out that since the experimenter had no access to the actual machine or payload data this conclusion is simply conjecture based on known user Behaviour within the target network. (Skype is a wonderful App)

Observations on Traffic for Clients and Peers

- Consumes considerable Resources.
- Represents an Application Level WAN Network for Communication.
- Provides a channel to hide Malicious Activity.

"*McAfee suggested hackers were likely to create malicious software to target instant messaging services, Voice over Internet Protocol (VoIP) telephony services and online gaming sites.*" **Hackers will target social networking sites: security firms - Thursday, November29, 2007, CBC News http://www.cbc.ca**

Evidence that all is not as it Appears

- One day in February a conversation took place between a user host on the Network and a host compromised by an on-line game server.
- Two hours later the user host was attempting to contact a few friends….

Sequentially….

Destination IP	sPort	dPort	Proto	bytes
XXX.XXX.026.000	0	2048	1	56
XXX.XXX.026.000	0	2048	1	168
XXX.XXX.026.001	0	2048	1	56
XXX.XXX.026.001	0	2048	1	168
XXX.XXX.026.002	0	2048	1	56
XXX.XXX.026.002	0	2048	1	168
XXX.XXX.026.003	0	2048	1	56
XXX.XXX.026.003	0	2048	1	168
XXX.XXX.026.004	0	2048	1	56
XXX.XXX.026.004	0	2048	1	168
XXX.XXX.026.005	0	2048	1	56
XXX.XXX.026.005	0	2048	1	168
XXX.XXX.026.006	0	2048	1	56
XXX.XXX.026.006	0	2048	1	168
XXX.XXX.026.007	0	2048	1	56
XXX.XXX.026.007	0	2048	1	168
XXX.XXX.026.008	0	2048	1	56
XXX.XXX.026.008	0	2048	1	168
XXX.XXX.026.009	0	2048	1	56
XXX.XXX.026.009	0	2048	1	168
XXX.XXX.026.010	0	2048	1	56
XXX.XXX.026.010	0	2048	1	168
XXX.XXX.026.011	0	2048	1	56
XXX.XXX.026.011	0	2048	1	168
XXX.XXX.026.012	0	2048	1	56
XXX.XXX.026.012	0	2048	1	168
XXX.XXX.026.013	0	2048	1	56
XXX.XXX.026.013	0	2048	1	168
XXX.XXX.026.014	0	2048	1	56
XXX.XXX.026.014	0	2048	1	168
XXX.XXX.026.015	0	2048	1	56
XXX.XXX.026.015	0	2048	1	168
XXX.XXX.026.016	0	2048	1	56

We Need to Re-Consider our Willingness to be a Peer

- Users willingly download and install client/peer/server software.
- They even participate in strategies to avoid barriers and impediments (like Nat'ing).
- There is an implied trust that the communication is exclusively what it claims to be.
- "When they thought they were playing at war craft, they were actually playing at war craft."

Concluding Notes

- The network is evolving at the edges
- This means that network architectures, management and provisioning strategies are now more responsive then ever.
- Global communication resources are primarily influenced by the uncoordinated activities of individuals.
- Traffic patterns are emergent properties without intent.

Future Work

- Study the growth in diversity of patterns in traffic.
- Study the form and distribution of applications and participants.
- Track Unidentified Anomalies.

- February 2008, TARA will announce the InTARA project

Intelligent Network Traffic Analyzers for Reconstructive and Real Time Analysis

- InTARA will be a multi-million dollar, multi-year project to develop intelligent traffic analysis capabilities for the good guys.
- We are seeking global collaborative research and commercialization partners. Early stage interest from Australia, India, Switzerland, Canada.