



Spam and Traffic Profiling

Institute Eurécom, France

Oliver HOHLFELD
Vincenzo SANTORO
Andre SCHROEDER
Paul WILCZYNSKI

December 2006



Spam and Traffic Profiling

- The Issue
- Spam Profiling
 1. Mail life cycle & spamming methods
 2. Existing mitigation techniques & drawbacks
 3. Characteristics of spam & spammers behaviour
- Traffic Profiling
 1. Existing approaches
 2. BLINC
 3. P2P



Spam and Traffic Profiling

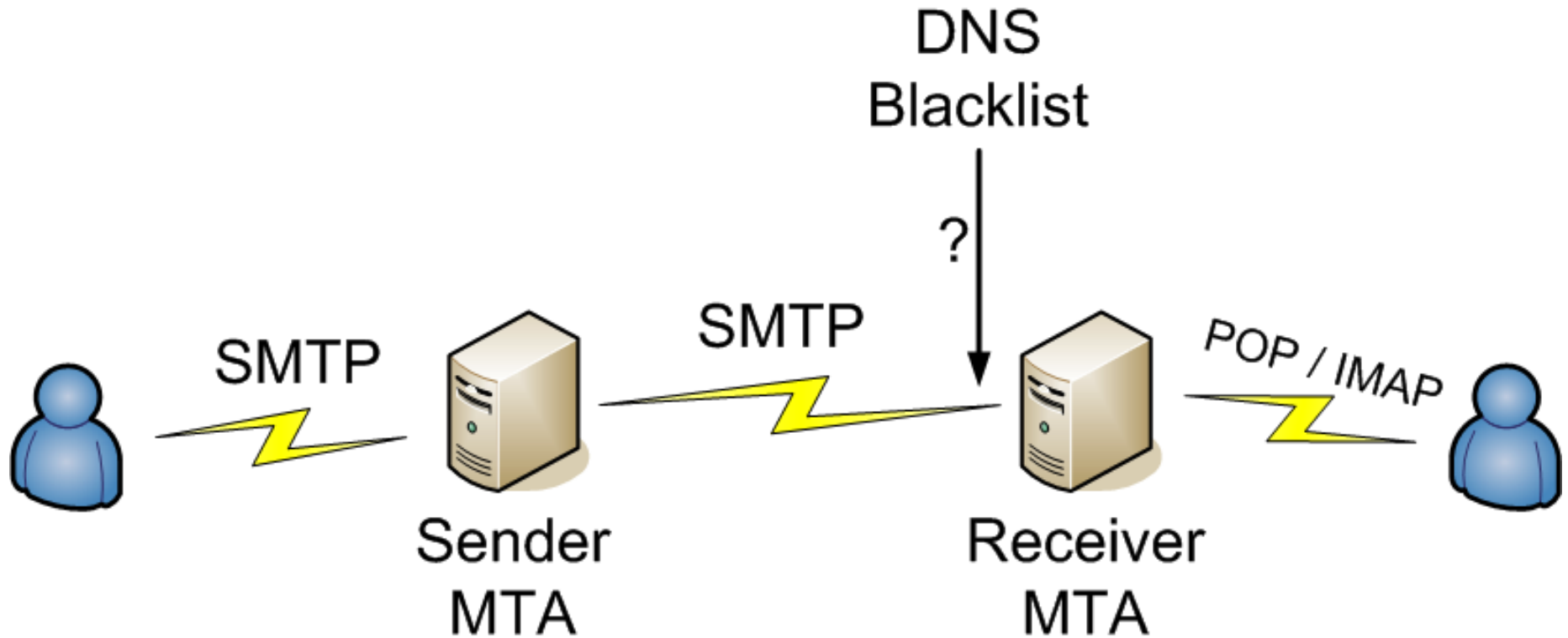
- ISPs & network administrators monitor & regulate traffic for:
 - Network design
 - Performance
 - Legal & policy issues
- Spam and uncontrolled P2P are responsible for:
 - Wasted bandwidth and congestion
 - Productivity loss
 - Legal issues
- Problems:
 - How to characterize the various flows going through a network?
 - How to identify P2P and spam traffic, and filter it ?



SPAM: Unsolicited Bulk Email

- The volume of unsolicited e-mails containing commercial content is increasingly at a very fast rate:
 - July 2002 35% of all e-mail in US were spam
 - 2003: 10 spams / day (business user)
2008 expected to be 40 / day
 - AOL and MSN block 2.4 billion spam mails / day (80% of mails)
- Impact on productivity
 - US\$50 to US\$1400 costs per year and worker
 - Total annual cost associated with spam to American business in the range of US\$ 10 billion to US\$ 87 billion.

Mail Life-Cycle





Spamming Methods

- Direct spamming
 - Spammers purchase upstream connectivity from “spam-friendly ISPs”
- Open relays and proxies
 - Mail servers that allow unauthenticated Internet hosts to connect and relay email through them
- Botnets (e.g. Bobax)
 - Collections of machines acting under one centralized controller
- BGP spectrum agility
 - Spammers announce IP address space from which they send spam and withdraw the routes for that space once the spam has been sent.



Mitigation Techniques

- Pre-acceptance method
 - Blacklist of known spammers, open relays and open proxies. Lookups to determine whether the sending IP address is in a “blacklist”
- Post-acceptance method
 - Content-based filtering uses features of the contents of an e-mail’s headers or body to determine whether it is likely to be spam.



Mitigation Techniques ISSUES

- Existing spam detection and filtering techniques have very high success rates: up to 97% of spams are detected
- But they suffer from two limitations:
 - Rate of false positives can be as high as 15%
 - Lifetime of existing techniques is compromised by spammers frequently changing their mode of operation. Constant upgrades and new developments are necessary.
- A quantitative analysis of the determinant characteristic of SPAM is still in demand...



Characteristics of Spam

- Goal of study
 - Develop an understanding of the fundamental characteristics of spam traffics and spammer's behavior.
 - Identify quantitative and qualitative differences between spam and non spam.



Methods used

- Collect a large amount of email headers, filtered into spam & non spam categories.
- Perform statistical analysis on different header data (time, senders, receivers..)



Main findings

- Spam traffic, unlike non-spam, is mostly stable during the whole day, all days.
- No clear weekly & daily patterns.
- Non-spam traffic is concentrated during business hours



Main findings

- The average size of non-spam is 6 to 8 times larger than the average size of spam.
- Non-spam emails have a 3 times higher coefficient of variability of their size.



Main findings

- 15% of spam have more than 1 recipient.
- Only 5% of non-spam do.



Explanation

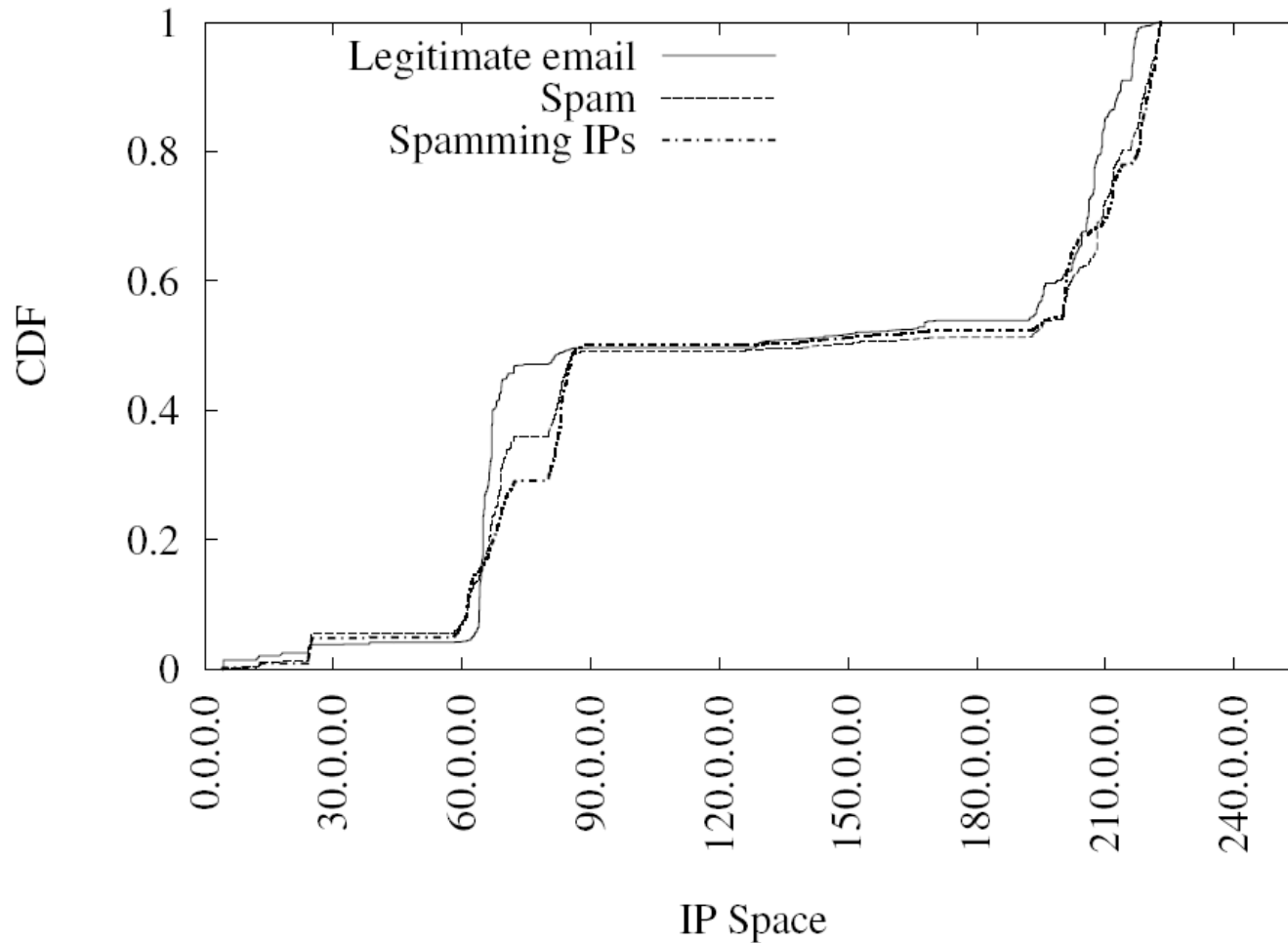
- Non-spam email transmission is the result of a social interaction between 2 humans.
 - Special Case: Newsletters etc.
- Whereas spam is a unilateral action typically performed by automatic tools.



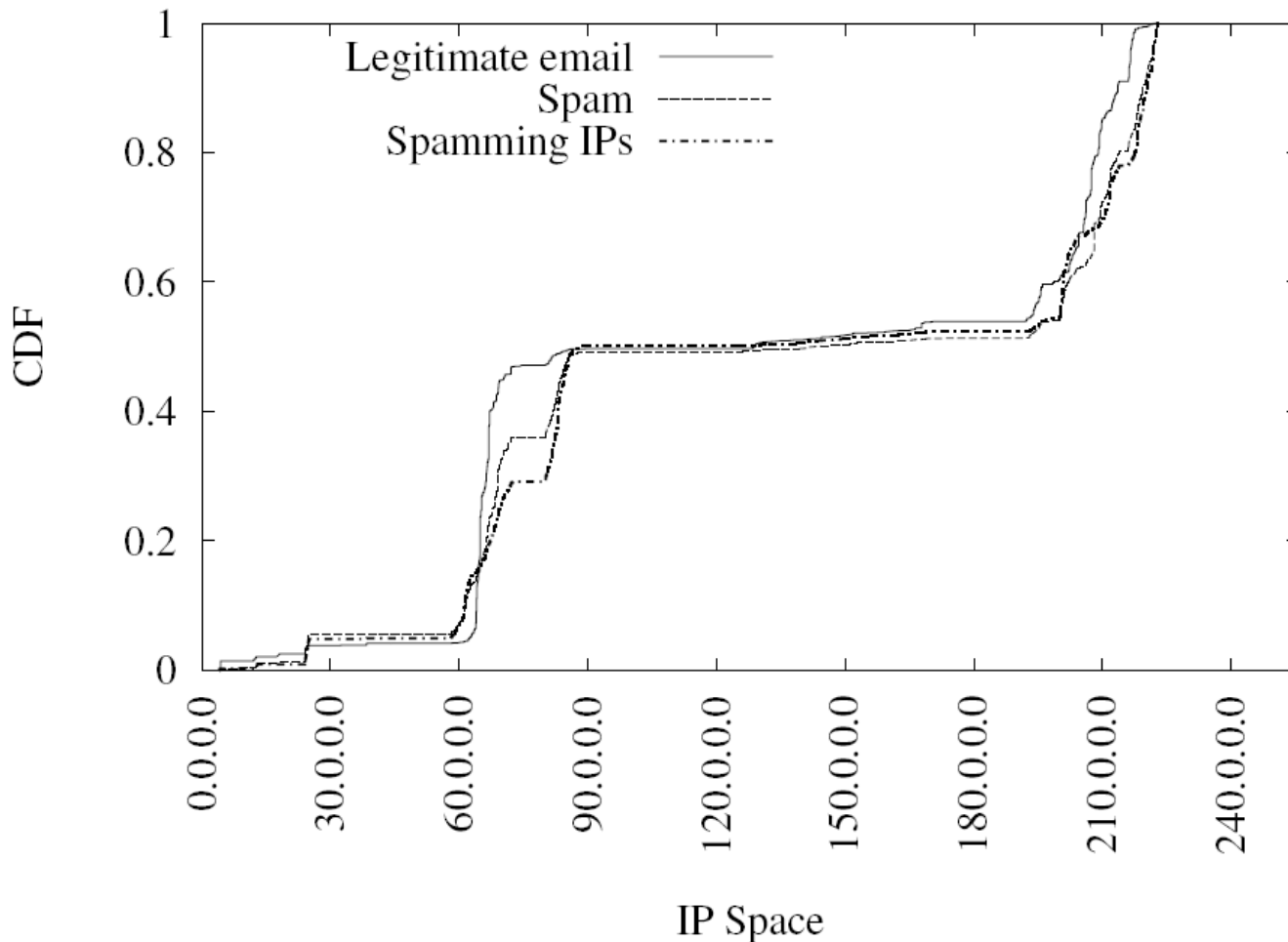
Network Level Behavior of Spammers

- Main Idea:
Information about the network level behavior of spam could be useful for designing spam filters based on spammers' network-level behaviour:
 - spammers have far less flexibility to alter the network-level properties of the spam they send
- Little is known...

Distribution across Networks 1/2



Distribution across Networks 1/2



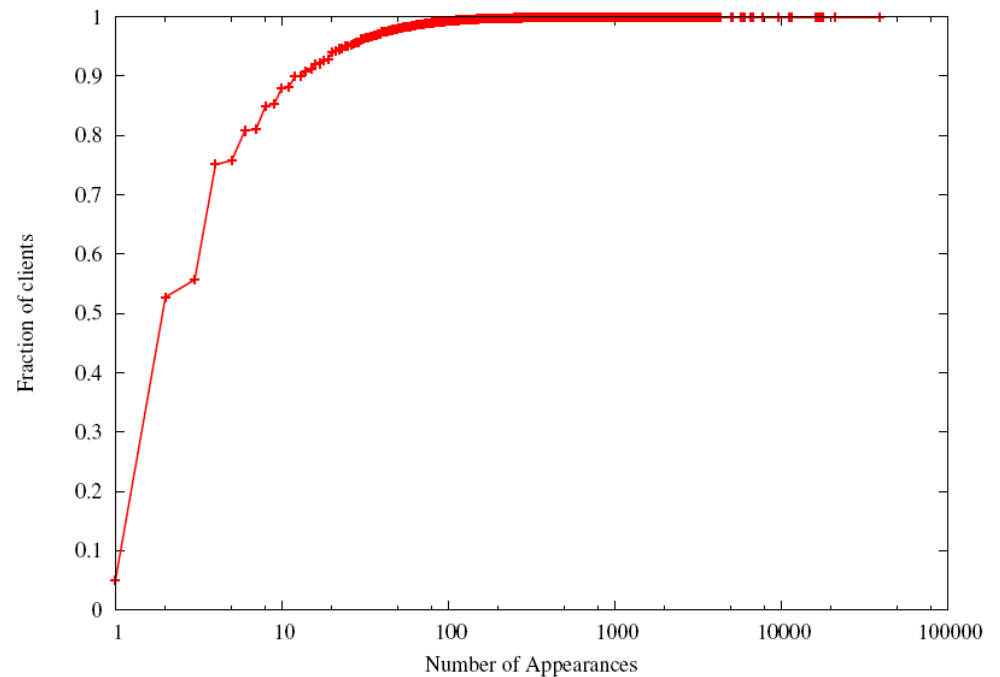
- Unused IP subnets instantly appear and disappear
- BGP announcements compromised!
- (BGP thought to be “safe” so far)



It's possible to use IP address ranges to distinguish spam from legitimate email.

Distribution across Networks 2/2

Individual IP addresses from which spam is received is very transient



Targeting an individual IP address might not help mitigate spam without sharing information across domains.



Spam from Botnets

- Spamming hosts & Bobax drones
 - similar distributions across IP address space
 - much of the spam received may be due to botnets (e.g. Bobax)
 - > Evidence of correlation
- 65% of IP addresses of hosts infected with Bobax send spam only once
- 75% send spam for less than two minutes
 - But several emails !!!



Spam from Transient BGP Announcements

- Spectrum agility
 - Spammer can use a wide variety of IP addresses to send spam
- IPs of mail relays sending spam are widely distributed across the IP address space
- IPs appear only once
 - not reachable by traceroute
- Mail relay IPs located in unused IP address space



Better Spam Mitigation

- Network-level properties have two important properties that could potentially lead to more robust filtering:
 1. **Less malleable** than those based on an email's contents
 2. **Observable in the middle of the network** which may allow spam to be disposed before it ever reaches a destination mail server



Part II

Network Application
Identification &
Network Traffic Classification



Content

- Classification
 - Port Numbers
 - Payload Analysis
 - Statistical Analysis
- New Approaches
 - BLINC
 - PTP



Identification by Port Numbers

- Well known port numbers (transport layer)
 - 80 HTTP
 - 4661-4665 eDonkey2000
 - 1214 Fasttrack (KaZaA)
 -
- Problems
 - Inconsistent (e.g. port 80 used for p2p traffic)
 - Dynamically used = arbitrary ports used
 - To overcome firewalls
 - To hide because of legal aspects (P2P)
 - e.g. used by todays P2P applications



Payload Analysis

- Application layer analysis
 - Find typical patterns in packet payload, e.g.:
 - “Get /.hash” - Fasttrack
 - “GNUT”, “GIV” - Gnutella
- Problems
 - Privacy problems!
 - Does not work with encrypted payload
 - A-priori knowledge about concrete protocols required
 - Reverse engineering of some protocols



Statistical Analysis

- Classification as statistical problem, e.g.
 - Machine Learning
 - Statistical Clustering
- Problems
 - Difficult to validate effectiveness
 - Has not been done in large scale yet



New Approaches

- BLINC
 - General traffic classification
- PTP
 - Classification of P2P traffic



Constraints

- Not using payload
- Not using port numbers
- Not using external information
 - Provided by flow collectors



BLINC

- BLINC = BLINd Classification
 - blind: not looking at payload
- Multilevel Traffic Classification
 - Using several techniques for classification

No accurate method present so far



BLINC - Purpose

- Tool for network developers
 - Meaningful traffic classification per application
 - Insight into traffic behaviour
 - Detection of abnormalities
 - Malicious behaviour
 - Identification of novel applications
- effective network planning and design
- monitoring networks



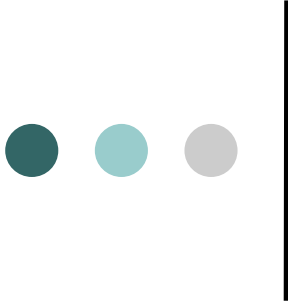
BLINC - Approach

- Not classifying individual flows
 - but associating host with application and then classify flow
- 3 Levels
 - Social
 - Functional
 - Application
- 6 Heuristics



BLINC – Host Characteristics

- Social (interaction with other hosts)
 - Popularity (# of communication partners)
 - Communities of hosts (collaborative applications)
- Functional
 - Is provider, consumer, or both
- Application
 - Transport layer interactions using 4-tuples:
 - (source ip, source port, dest ip, dest port)
 - Empirically derived library of “graphlets”



BLINC – Heuristics

- Transport Layer
 - TCP, UDP, both
- Cardinality of Sets
 - Destination Ports / IPs ratio
- Average package size
- ...



BLINC - Results

- Traffic traces: Access links
- Completeness (amount of classified traffic)
 - 80% - 90%
- Accuracy
 - 95%
- New protocols were identified

Trade off between Completeness & Accuracy
can be fine-tuned

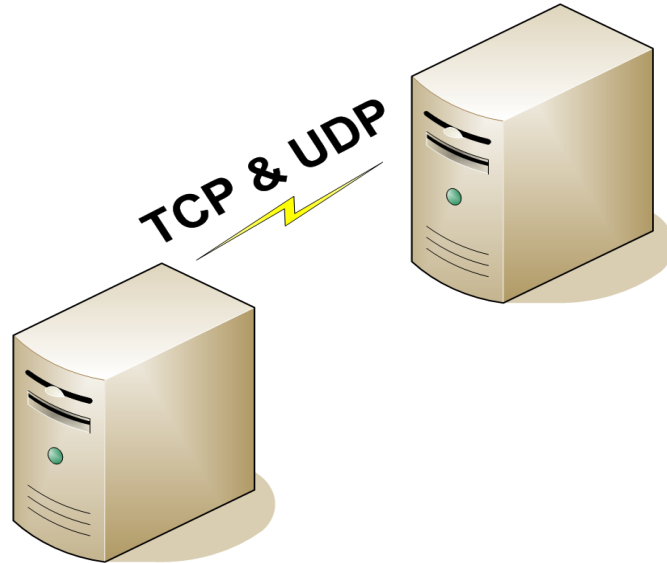
Classification reference: Payload based



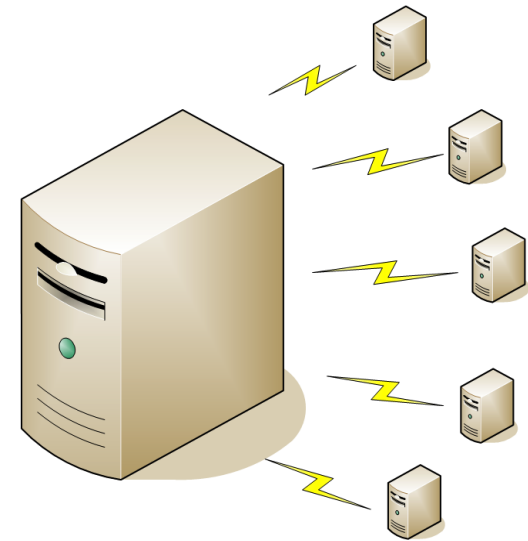
PTP: Transport Layer Identification of P2P Traffic

- Identify typical P2P network flow patterns
- Uses heuristics
 - No guarantees
 - False positives possible
- Independent of protocols
 - Can detect unknown protocols which follow the same pattern
 - No a-priori knowledge about concrete protocols required

PTP: Base Heuristics



Heuristic I



Heuristic II



PTP: Reduce False-Positives

- Additional heuristics to identify
 - Mail
 - DNS
 - Game and Malware
 - Portscans
- Drawback: too many heuristics
 - long computation time
 - new heuristic may not reduce false-positives



PTP: Results

- Measurement done on Backbone traces
- Compared to a payload analysis
 - 99% of P2P flows correctly identified
 - 95% of P2P Bytes correctly identified
- False-positives rate of 8 - 12 %
 - Decreases with time
 - More knowledge about network flows
- 3 unknown P2P protocols found



BLINC & PTP Issues

- Cannot identify specific application sub-types
 - P2P, but not eMula, Gnutella, KaZaA, ...
- Cannot work on encrypted transport layer headers



The End

- Questions?



References

- T. Karagiannis, A. Broido, M. Faloutsos, and K. claffy, “*Transport layer identification of p2p traffic*”. In IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, pages 121–134, New York, NY, USA, 2004. ACM Press.
- T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “*BLINC: Multilevel traffic classification in the dark*” (2005). ACM SIGCOMM. 35 (4), pp. 229-240.
- L. Gomes, C. Cazita, J. Almeida. “*Characterizing a Spam Traffic*”, In: Proc. of the IMC'04, Oct. 25-27, 2004, Taormina, Sicily, Italy. pp. 356-369
- J. Jung and E. Sit, “*An Empirical Study of Spam Traffic and the Use of DNS Black Lists*”. In Proc. of the Internet Measurement Conference, Taormina, Sicily, Italy, October 2004
- A. Ramachandran and N. Feamster, “*Understanding the network-level behavior of spammers*”. In Proc. ACM SIG-COMM, September 2006. Technical Report Georgia Tech TR GT-CSS-2006-001