P2P Filesharing Population Tracking Based on Network Flow Data

Arno Wagner

wagner@tik.ee.ethz.ch

Communication Systems Laboratory

Swiss Federal Institute of Technology Zurich (ETH Zurich)

Outline

1. Motivation, Setting
2. The PeerTracker
3. Validation by Polling
4. Comments on Legal Aspects
5. Conclusion

Contributors

- Philipp Jardas, ”P2P Filesharing Systems: Real World NetFlow Traffic Characterization”, Barchelor Thesis, ETH Zürich, 2004
- Lukas Hämmerle, ”P2P Population Tracking and Traffic Characterization of Current P2P file-sharing Systems”, Master Thesis, ETH Zürich, 2004
- Roger Kaspar, ”P2P File-sharing Traffic Identification Method Validation and Verification”, Semester Thesis, ETH Zürich, 2005

PDFs available from
`http://www.tik.ee.ethz.ch/~ddosvax/sada/`

Motivation

- P2P traffic forms a large and dynamic part of the overall network traffic
- Identification allows blocking/shaping
- Identification allows P2P anomaly detection
- Identification allows better analysis of other traffic

The DDoSVax Project

`http://www.tik.ee.ethz.ch/~ddosvax/`

- Collaboration between SWITCH (www.switch.ch, AS559) and ETH Zurich (www.ethz.ch)
- Aim (long-term): Near real-time analysis and countermeasures for DDoS-Attacks and Internet Worms
- Start: Begin of 2003
- Funded by SWITCH and the Swiss National Science Foundation

DDoSVax Data Source: SWITCH

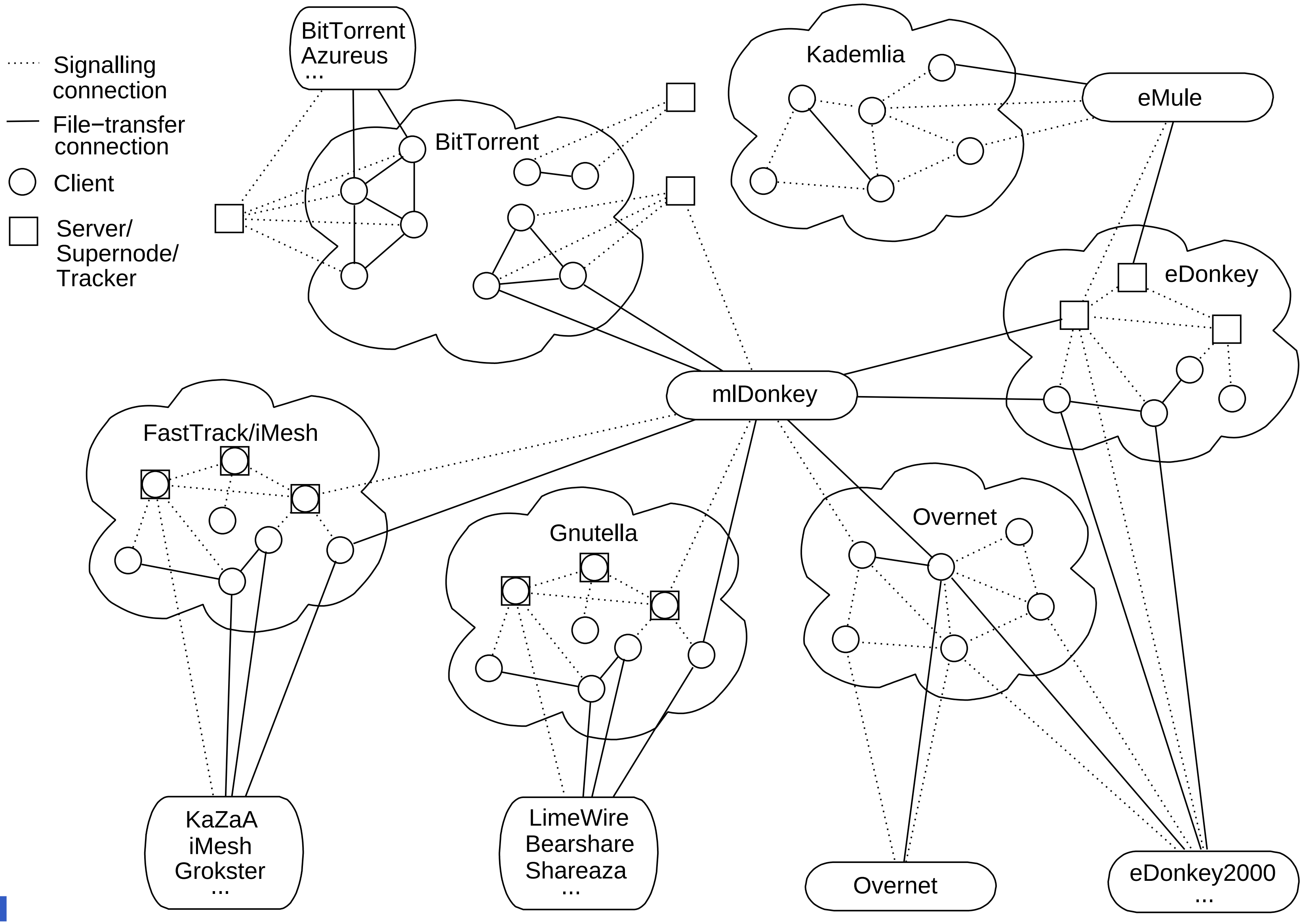
The Swiss Academic And Research Network

- .ch Registrar
- Links most Swiss Universities and CERN
- Carried around 5% of all Swiss Internet traffic in 2003
- Around 60.000.000 flows/hour
- Around 200GB...300GB traffic/hour
- Flow archive since May 2003
- Only few home users (`via-eth.ch`)

Network Flow Data

- Exported by routers or special sensors
- Aggregated source and destination IP and port, byte and packet count, start and end time
- No payload information
- Limited aggregation for UDP/ICMP and other non-TCP traffic

P2P Networks Considerd



PeerTracker Algorithm

Idea: ”Traverse” network from seeds

- Seeds are peers using default ports (TCP and UDP)
  ”Most used remote ports” better than ”local ports”
- Keep a pool of peers for each network
- Add hosts that communicate with the pool
- Remove hosts that are idle

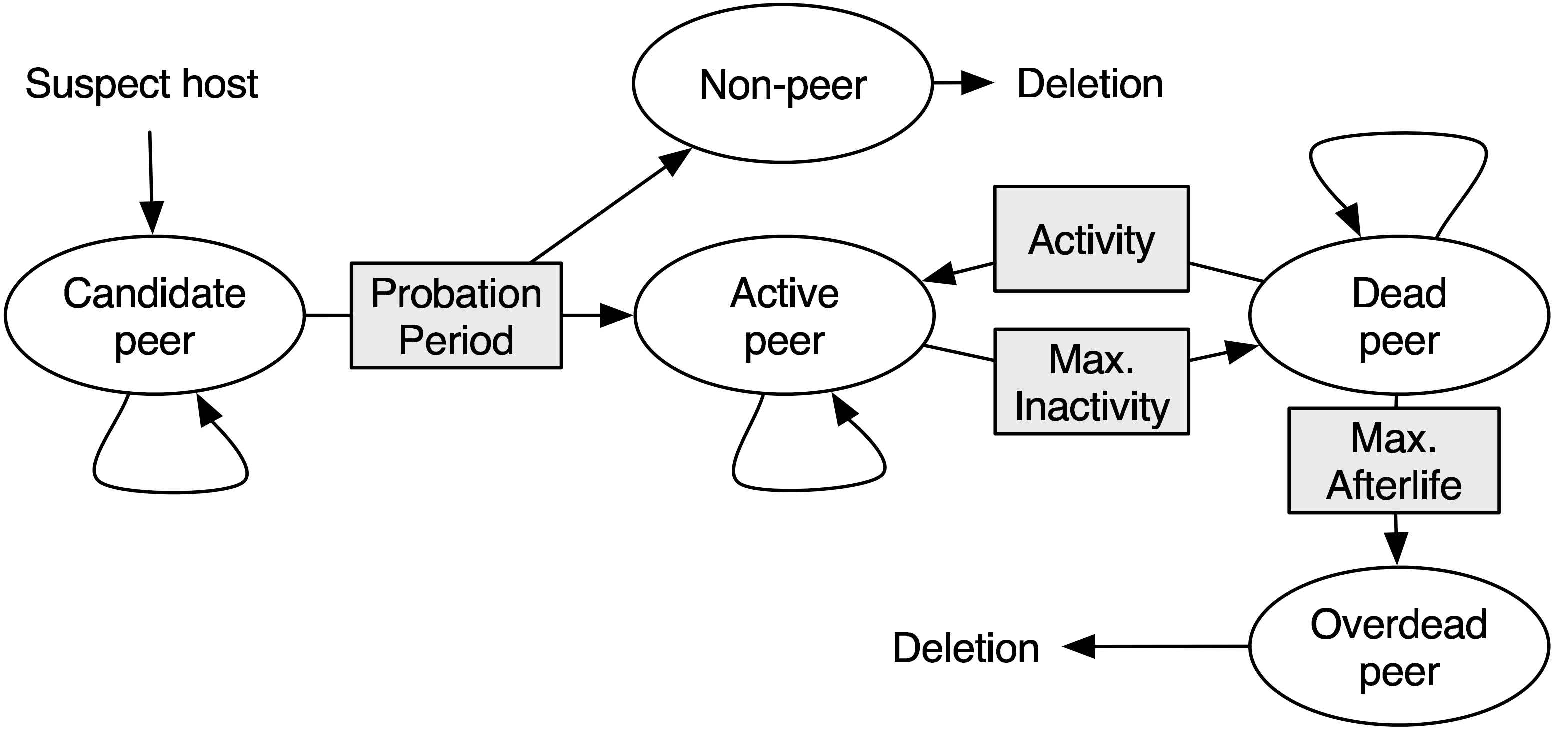
Notes:
Standard PC enough for SWITCH network
PeerTracker code is available under GPL

Default Port Usage

P2P System	Default port usage (TCP)
BitTorrent	70.0 %
FastTrack	8.3 %
Gnutella	58.6 %
eDonkey	55.6 %
Overnet	83.9 %
Kademlia	66.6 %

From 8 day PeerTracker measurement end of 2004

PeerTracker: Internal Hosts



Some PeerTracker Details

- Peers in the SWITCH network form ”core”
- External hosts only identified if they contact core
- Different ageing for core and external hosts
- Dead core peers are still contacted from external peers (state ”dead”)

Some Measurements

P2P System	P2P default ports		PeerTracker method	
BitTorrent	55.4 Mbit/s	(12.2 %)	90.1 Mbit/s	(19.9 %)
FastTrack	1.8 Mbit/s	(0.4 %)	12.3 Mbit/s	(2.7 %)
Gnutella	5.1 Mbit/s	(1.1 %)	10.7 Mbit/s	(2.4 %)
eDonkey, Kademlia, Overnet	47.7 Mbit/s	(10.5 %)	82.1 Mbit/s	(18.1 %)
Total P2P	110.0 Mbit/s	(24.4 %)	195.2 Mbit/s	(43.1 %)

Measurements taken August 2004

Peers by Domain

ethz.ch	43%
via-eth.ch	26%
epfl.ch	10%
unil.ch	4%
zhwin.ch	4%

For more data see the referenced theses.

Validation

The PeerTracker does not look at Payloads
$\Rightarrow$ Large error possible

Validation Approach:

- Run PeerTracker
- Poll found peers immediately
- Compare polling and tracker results

Peer Polling Methods

P2P System	Polling method	
FastTrack	Request:	`GET /.files HTTP/1.0`
	Response:	`HTTP 1.0 403 Forbidden` <number 1> <number 2>
	or	`HTTP/1.0 404 Not Found/nX-Kazaa-`<username>
Gnutella	Request:	`GNUTELLA CONNECT/`<version>
	Response:	`Gnutella` <status>
eDonkey,	Request:	Binary: 0xE3 <length> 0x01 0x10 <MD4 hash> <ID> <port>
Kdemlia,	Response:	Binary: 0xE3 . . .
Overnet		
eMule	Same as eDonkey, but replace initial byte with 0xC5.	
BitTorrent	Unsolved. Seems to need knowledge of a shared file on the target peer.	

Polling Results

P2P System	TCP Connect	P2P-client found
eDonkey, Kademlia, Overnet	50%	41%
Gnutella	53%	30%
FastTrack	51%	41%
Total	51%	38%

Table 1: Positive polling answers

Polling Remarks

- Delay to polling 10 ... 15 Minutes
- About 50% unreachable via listening port $\Rightarrow$ NAT
- Other errors: Peer variation (esp. Gnutella), classification into wrong network, tracker error
- BitTorrent not really pollable

Legal Aspects

(Warning: I am no legal expert)

- Flow data likely not subject to privacy laws (unless attempts to identify people are made)
- PeerTracker does not identify content shared
  $\Rightarrow$ output unproblematic, no action needs to be taken
- Identification of heavy hitters unproblematic
- Polling unproblematic, since similar to running a peer (some users get nervous though...)

PeerTracker for Law Enforcement

Situation for CH!

- Massive private file-sharing is done
- Law enforcement is not really interested in copyright infringement
- Illegal contents (child pornography and the like) can done far better with modified P2P clients

⇒ Not really suitable

Conclusion

- Peer identification feasible with flow data
- Standard PC enough for fast links
- Nodes with little traffic problematic
- BitTorrent problematic
- P2P filesharing still evolves fast

Thank You!