Measurement of paedophile activity in eDonkey

Guillaume Valadon - http://valadon.complexnetworks.fr

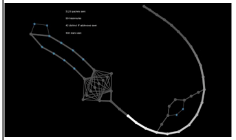
http://antipaedo.lip6.fr

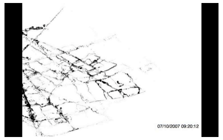
LIP6 (CNRS - UPMC)

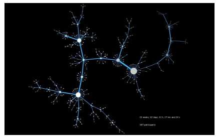
Complex Networks team
http://complexnetworks.fr

The team

- http://complexnetworks.fr : plots & videos
  - 4 permanent members : Jean-Loup Guillaume, Matthieu Latapy, Bénédicte Le Grand, Clémence Magnien
  - 2 postdocs, 9 Ph.D. students







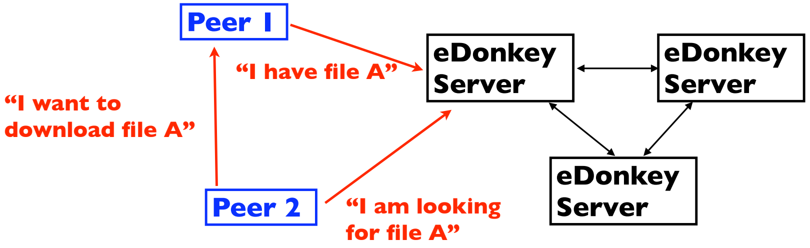
- Focus & interests:
  - Internet topology, P2P networks, social networks
  - measurements
  - analysis

What is peer-to-peer ?

- exchanges do not rely on a server
  - content is inside peers or created by peers
- peers are equal
  - clients and servers at the same time
- peer removal not problematic

- Usages:
  - file sharing: eDonkey, bittorrent
  - telephony: skype
  - video streaming: joost, PPlive

What is eDonkey ?



- servers are catalogs of files
- peers:
  - inform the servers
  - search files on servers
  - download files from peers

Context

- study *exchanges* in eDonkey
  - files diffusion
  - communities of interests
  - popularity

- some motivations
  - understand users' behaviour
  - blind content detection
  - detect paedophile activities

Outline

1. eDonkey measurements
   1. server side
   2. honeypot
   3. client side

2. results & statistics

3. community detection

eDonkey exchanges: what can be observed ?

1. inter-server communications

   statistical data about servers usages & peers

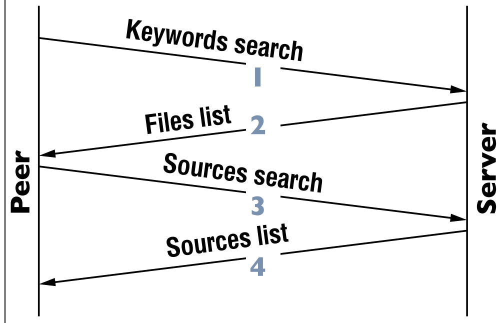
2. peer-server communications

   index of files, file search, source search

3. inter-peer communications

   file downloads, retrieve lists of files

Looking for a file



1. “music mp3”

2. name:ACDC.mp3

   ID:101050

   name:rock.mp3

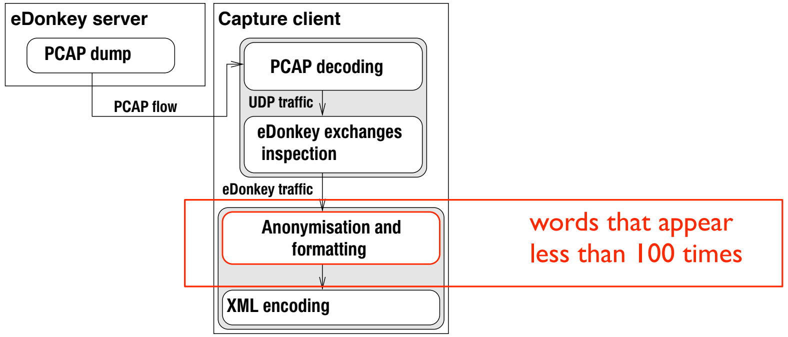
   ID:B16V100

3. ID:B16V100

4. 192.168.0.2

   1.2.3.4

Capturing traffic on a real server



words that appear less than 100 times

```
<opcode dir="received" TS="2786402.373146" IP="0045125351"
type="high" port="02029"><OP_GLOBSEARCHREQ>
<tags count="1"><anon-string>3108886</anon-string></tags>
</OP_GLOBSEARCHREQ></opcode>
```

Resulting data set in numbers

[HotP2P'09]

- 10 weeks measurements
- ~500 GB of compressed XML
- ~ 10 billions messages
- ~ 90 millions peers
- ~ 280 millions of distinct files

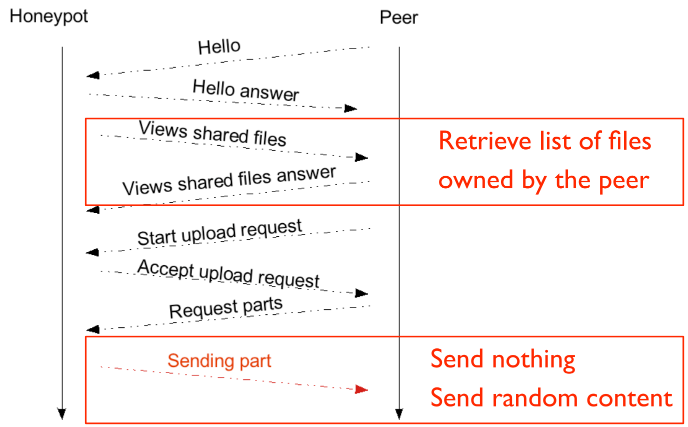
➡ anonymized data available online at http://antipaedo.lip6.fr

Honeypot based measurements

[HotP2P'09]

- eDonkey honeypot:
  - customized eDonkey client
  - announce files to a server (filename, ID, size)
  - log queries made by regular peers
- Manager:
  - control distributed honeypots
  - send commands to honeypots: server to connect to, files to exchange, ...

Downloading a file



Retrieve list of files owned by the peer

Send nothing

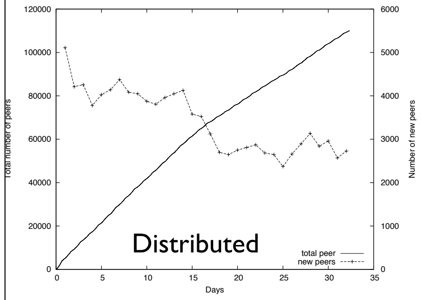
Send random content

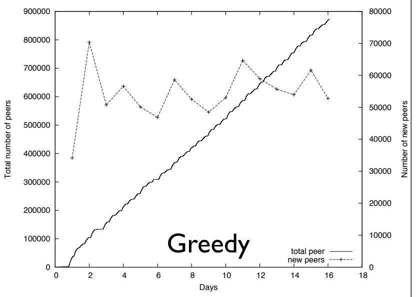
Methodology

- 24 PlanetLab nodes, running distributed honeypots:
  - 12 sending *no content*
  - 12 sending *random content*
- 1 greedy honeypot:
  - learn files during the first day
  - afterwards, announce these files

	distributed	greedy
Honeypots	24	1
Duration in days	32	15
Shared files	4	3 175
Distinct peers	110 049	871 445
Distinct files	28 007	267 047

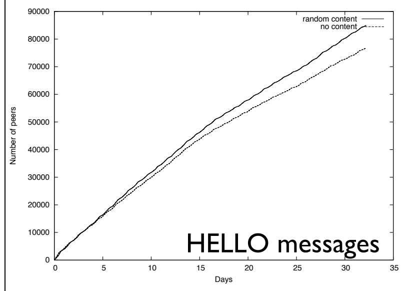
Parameters : distributed or greedy

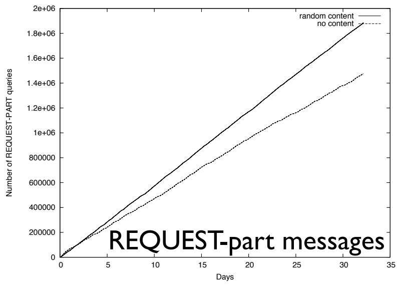




- long measurements are relevant
- effects of blacklisting and file popularity

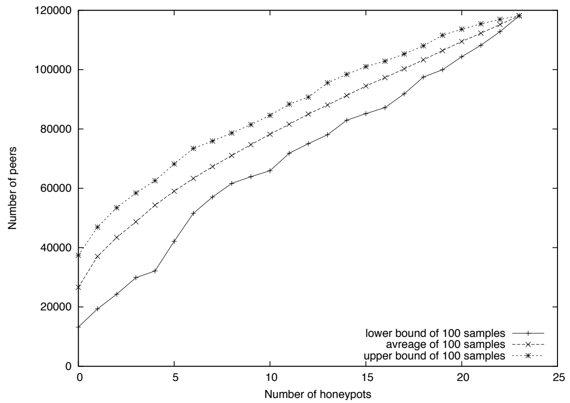
Parameters : no-content & random-content





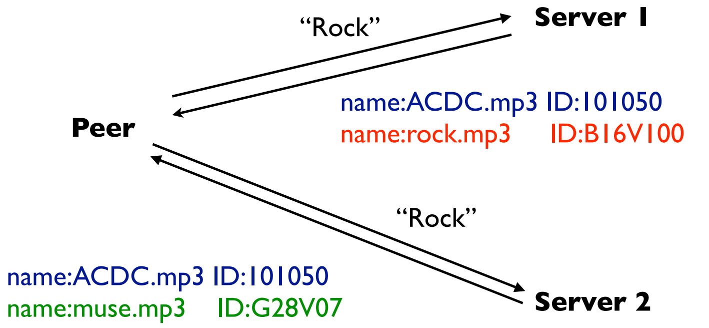
- advantage of sending random content
- global and local blacklisting

Parameters: number of honeypots



- important benefit in using several honeypots

Peer based measurements



"Rock"

Server 1

Peer

name:ACDC.mp3 ID:101050
name:rock.mp3 ID:B16V100

"Rock"

name:ACDC.mp3 ID:101050
name:muse.mp3 ID:G28V07

Server 2

Querying multiple servers helps to discover more files
More IP addresses can be seen in the same way

Methodology & data

- a modified client connects to multiple servers
  - queries servers with 15 keywords (8 are paedophile)
  - retrieves all filenames and IP addresses
  - restarts every 12 hours

- Resulting data set
  - 140 days measurements (October 2008 to February 2009)
  - ~ 3 millions peers
  - ~ 3 millions of distinct files
  - ~1.5 millions different filenames

Outline

1. eDonkey measurements
   1. server side
   2. honeypot
   3. client side

2. results & statistics

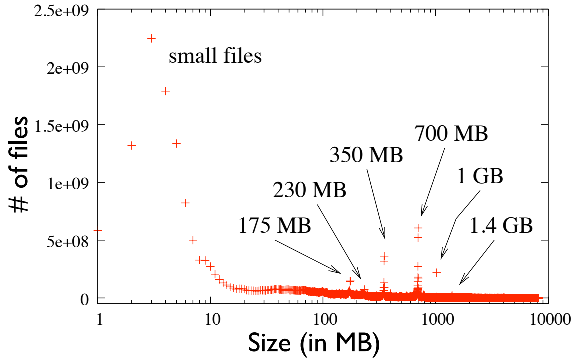
3. community detection

Goal & limitations

- rigorous evaluation of several elements
  – peers, queries, filenames, ...

- difficulties
  – IP equals user ?
  – one file, several names
  – paedophile query ? paedophile user/IP ?
  – no access to files’ content

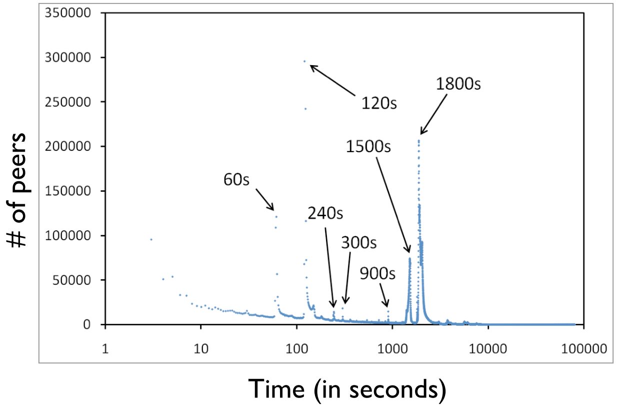
Basic analysis: file sizes



- obtained from the server answers
- CD-ROM size and fractions (1/2, 1/3, and 1/4)

➡ related to classical sizes of storage support

Basic analysis: time between queries



- regularities of queries

Basic analysis: top 10 words

Rank	Keyword	# occurrences
1	mp3	12 121 052
2	avi	2 860 225
3	the	2 657 349
4	rar	1 610 669
5	de	1 607 634
6	jpg	1 296 610
7	la	1 236 001
8	of	1 082 521
9	a	1 039 469
10	mpg	993 077

filenames

Rank	Keyword	# occurrences
1	the	4 147 197
2	de	3 382 473
3	la	2 337 404
4	a	1 761 179
5	of	1 751 848
6	2	1 398 154
7	i	1 153 601
8	ita	1 101 964
9	2006	1 075 982
10	el	1 025 315

queries

Basic analysis: names per file

- different filenames but same content

  - translation, commas/spaces, fake files, ...

- up to 82 different filenames for one file
- ~ 16 millions files with 1 filename
- ~ 3 millions files with 2 or more filenames

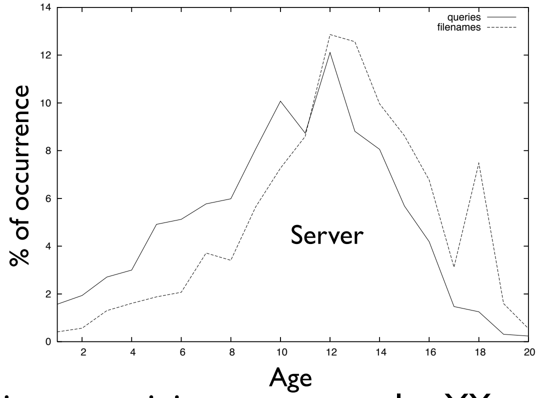
➡ problematic for content detection and fake detection systems

name:kung-fu-panda.avi ID:1234

name:kungfupanda-FR.avi ID:1234

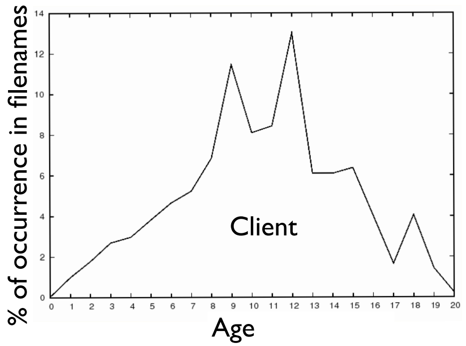
name:paedophile-keyword.jpg ID:1234

Paedophile content: age detection



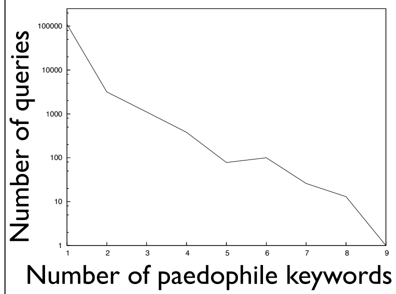
- strings containing ages; example: XYyo
- queries targets *younger* ages than filenames
- more demand than supply

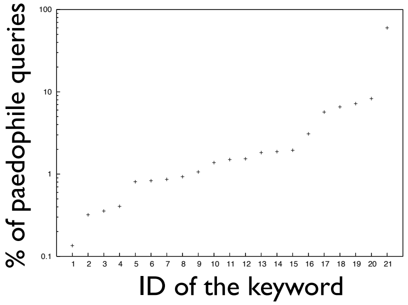
Paedophile content: age detection



- similarity in server and client measures
- spikes at 9,12 and 18 years old

Paedophile keywords in queries

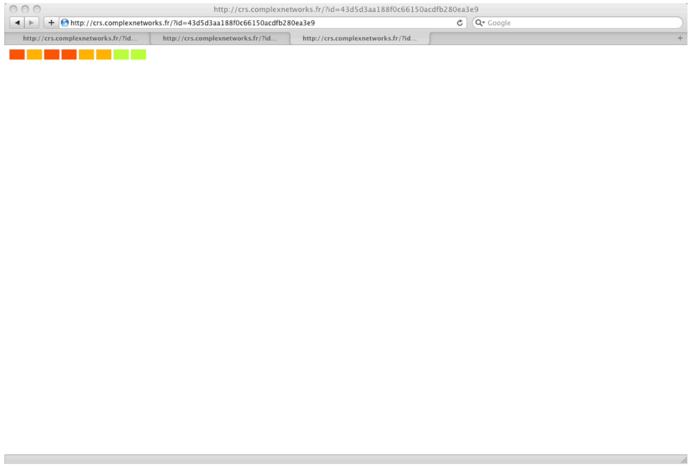




- 94% of queries contain only one keyword
- keyword 21 is used in 60% of the queries
  - is it still a paedophile keyword ?

Blind content detection

http://crs.complexnetworks.fr/?id=HASH


http://crs.complexnetworks.fr/?id=43d5d3aa188f0c66150acdfb280ea3e9
http://crs.complexnetworks.fr/?id=43d5d3aa188f0c66150acdfb280ea3e9
Google
http://crs.complexnetworks.fr/?id...
http://crs.complexnetworks.fr/?id...
http://crs.complexnetworks.fr/?id...

porn1; porn2; pedo1; pedo2; fake1; fake2; fake3; fake4

Perspectives and ongoing work

- definition of paedophile IPs and queries
  - one keyword ? several ?
  - once count as paedophile always paedophile ? (NAT)

- what is the number of file and paedophile users ?
  - “9000 paedophiles on the Internet” http://tinyurl.com/c78vlu

    “including 1000 in germany”

Outline

1. eDonkey measurements
   1. server side
   2. honeypot
   3. client side

2. results & statistics

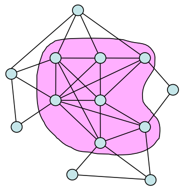
3. community detection

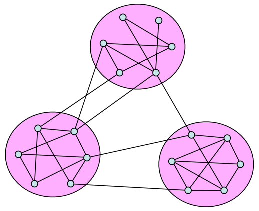
Goals

- analysis based on the structure
- represent data as a graph (nodes & edges)
  - words in queries ? filenames ?
  - file-ID ? client-ID ?
- Motivations
  - understand the structure
  - detect communities of interest
  - graph visualization
  - improve our knowledge of paedophile keywords

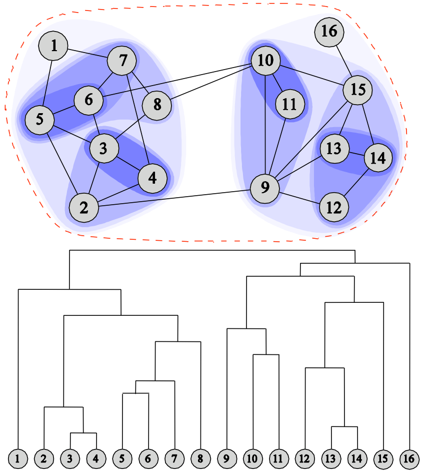
What is a community?

- A community is a set of nodes
  - nodes share something,
  - high level of connection,
  - more links inside than outside.



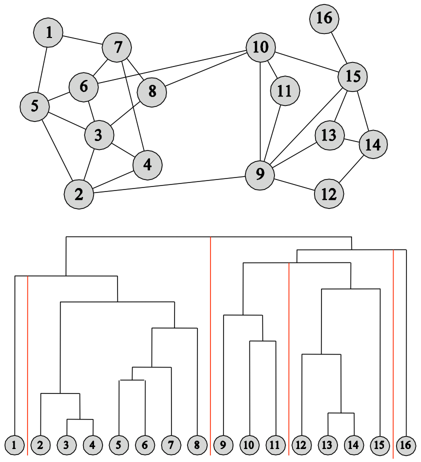


Community detection: hierarchical clustering



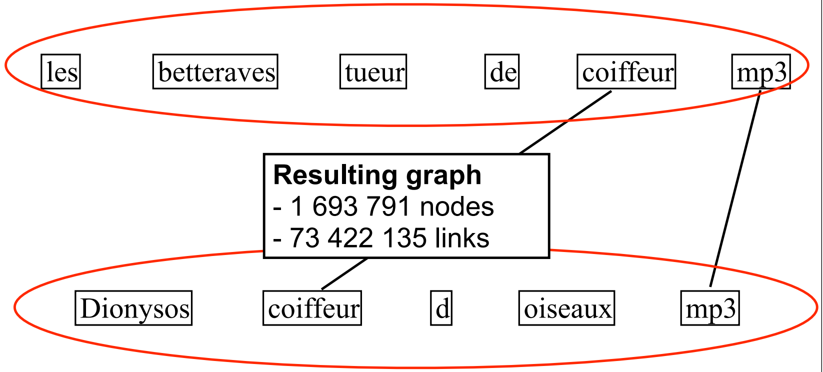
Slide courtesy of Pascal Pons

Community detection: divisive approach

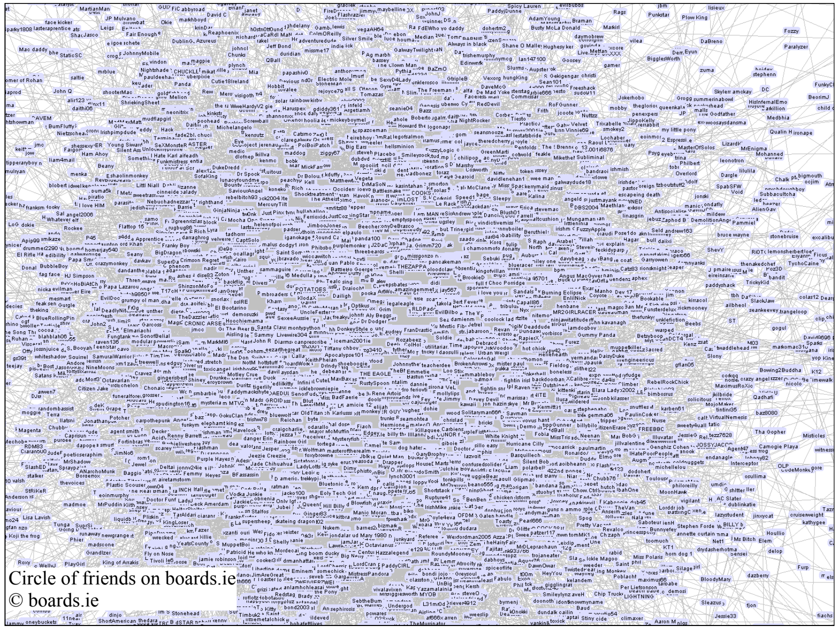


Adapted from slide courtesy of Pascal Pons

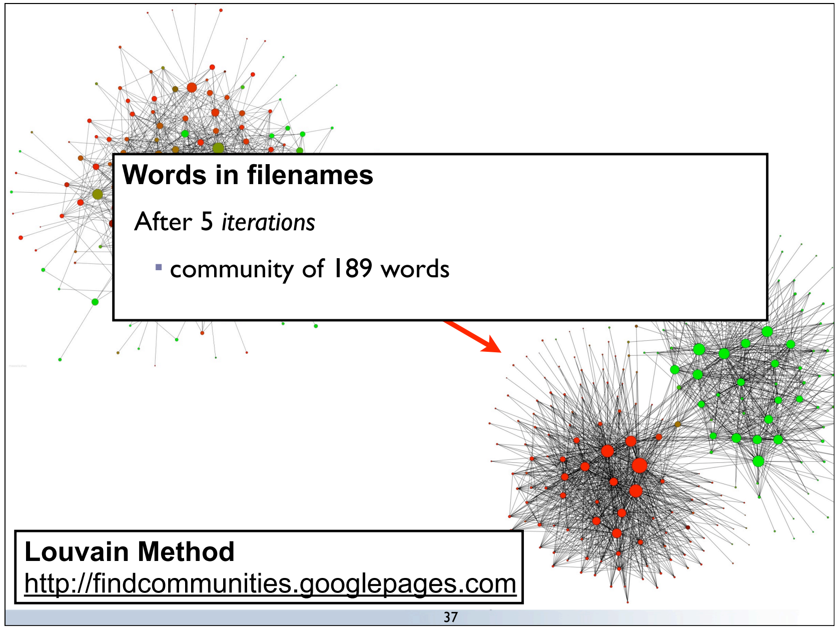
Words in filenames:
how to build a graph ?



Resulting graph
- 1 693 791 nodes
- 73 422 135 links



Circle of friends on boards.ie
© boards.ie



Words in filenames

After 5 *iterations*

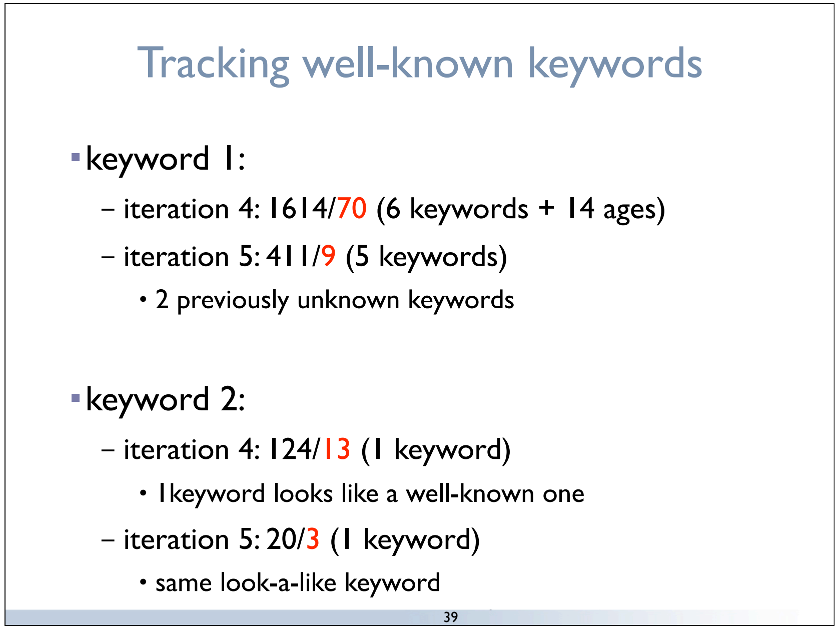
- community of 189 words

Louvain Method
http://findcommunities.googlepages.com

Community containing 12yo

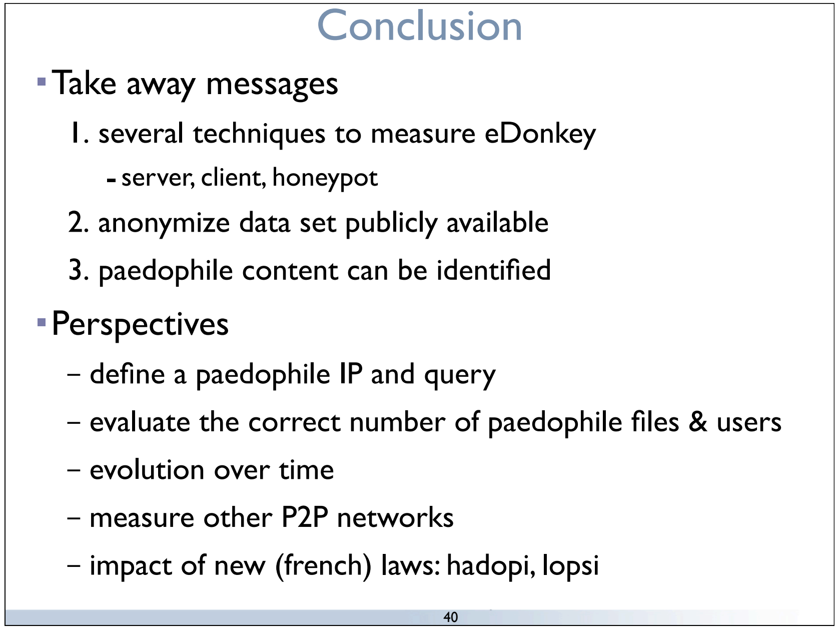
	Queries	Filenames
graph	3 380 213	1 693 791
iteration 1	2 469 134/107 688	828 913/97 111
iteration 2	785 606/23 680	314 665/30 367
iteration 3	232 620/3 954	13 540/1259
iteration 4	51 026/836	1310/94
iteration 5	1614/70	189/11

BLACK words/nodes in the community
RED well-known (> 100 times in the data set)



Tracking well-known keywords

- keyword 1:
  - iteration 4: 1614/70 (6 keywords + 14 ages)
  - iteration 5: 411/9 (5 keywords)
    - 2 previously unknown keywords
- keyword 2:
  - iteration 4: 124/13 (1 keyword)
    - 1keyword looks like a well-known one
  - iteration 5: 20/3 (1 keyword)
    - same look-a-like keyword



Conclusion

- Take away messages
  1. several techniques to measure eDonkey
     - server, client, honeypot
  2. anonymize data set publicly available
  3. paedophile content can be identified
- Perspectives
  - define a paedophile IP and query
  - evaluate the correct number of paedophile files & users
  - evolution over time
  - measure other P2P networks
  - impact of new (french) laws: hadopi, lopsi

Questions ?

Community detection: Louvain method