Measurement and Analysis of P2P Exchanges Against Paedophile Activity

http://antipaedo.lip6.fr

Raphaël Fournier, Matthieu Latapy and Clémence Magnien

LIP6 (Computer Science laboratory)

CNRS and Université Pierre et Marie Curie (UPMC), France

INTERPOL, Lyon, September 28th, 2010

Outline

Outline

Rationale

Much paedophile activity in P2P systems

- Many offenders / interested users
- Danger for innocent users
- Correlation between viewing and acting out
- Policy making issues

Very little is known

Approach

Research $\longleftrightarrow$ **Applications**
scientific problems   relevance for enforcement issues

Balance between research and applications

interaction with authorities
problem identification
result assessment

Partners & Support

Partners:

CNRS - UPMC, France	Computer science
INRIA (Lorraine), France	Computer science
University College Cork, Ireland	Applied psychology
University of Ljubljana, Slovenia	Social sciences
Nobody's Children Fundation, Poland	NGO

Support:

- European Commission (Safer Internet Programme)
- French National Agency for Research (ANR)
- Action Innocence Monaco (NGO)

Activities

Improve knowledge
on paedophile activity

Measure   Collect appropriate data
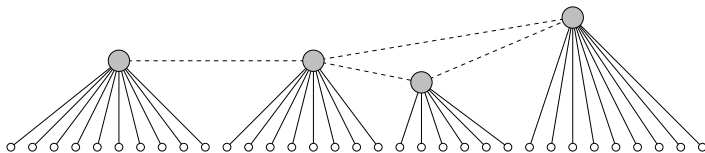
Analyse   Obtain reliable information from data

Outline

2. Measurement of P2P systems
   - Server measurements
   - Other measurements

eDonkey system – Basic overview

Popular clients: *eMule*, *MLDonkey*, *Shareaza*

- Server : file indexing and list of providers

- File transfers between peers only

eDonkey system – Server measurement principle

peer0	$\longrightarrow$	Keyword-based query *beatles yellow submarine*	$\longrightarrow$	**server**
peer0	$\longleftarrow$	List of files matching the keywords *beatles.yellow.submarine.192k.mp3* *beatles-yellow-submarine.wma* *Revolver.Yellow-Submarine.128vbr.mp3*	$\longleftarrow$	**server**
peer0	$\longrightarrow$	Desired file F *beatles.yellow.submarine.192k.mp3*	$\longrightarrow$	**server**
peer0	$\longleftarrow$	Providers of F *peer5678*	$\longleftarrow$	**server**
peer0	$\longrightarrow$	initiate file download	$\longrightarrow$	**peer5678**
peer0	$\longleftarrow$	File F	$\longleftarrow$	**peer5678**

First server measurement – 2007

- 10 weeks, in 2007
- Server importance: high

Collected data

- 89 million peers (IP addresses)
- 275 million files (hash code)
- 24 million distinct filenames
- 104 million distinct keyword queries

Second server measurement – 2009

- 28 weeks, in 2009 (still running)
- Server importance: average
- Keyword-based queries only
- Geolocation of IP addresses

Collected data

- 24 million peers (IP addresses)
- 228 million distinct queries

Data anonymisation

- User privacy and legal concern

- Strong anonymisation of IP addresses and connection ports

Other measurements

- Honeypot

- Client sending queries

- Kad (another protocol, less centralized)

Outline

Keyword-based queries

. . .
pagine
dvdrip xxx
carte europe pour pc pocket medion
10yo boy hard sex
a long dimanche the passion
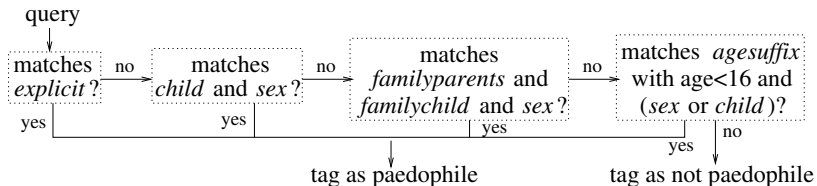der wald ist nicht genug
black affaire
raygold
dans la lune

. . .

Tool design

1. set of rules based on domain knowledge
2. manual inspection of our datasets
3. improve until negligible changes
4. 4 categories of paedophile queries

Tool design

qqaazz little girl

porno infantil

incest mom son video

12yo fuck video

Quality

False positive

"sexy daddy destinys child"
contains "sexy", "daddy" and "child"
but most likely a music-related query

False negative

"pjk kdv 12yo"
contains paedophile keywords that we don't search for

How to estimate false positive and false negative rates ?

Tool assessment – Survey

- set of 21 volunteering experts (Europol, national authorities, NGOs)

- sets of randomly selected queries:
  - paedophile
  - not paedophile
  - *neighbours* (queries submitted by the same paedophile user within the 2 previous or next hours)

- tag queries as *paedophile*, *probably paedophile*, *probably not paedophile*, *not paedophile* or *I don't know*

Tool assessment – Survey results

paedo	prob. paedo	don't know	prob. not	not paedo	total	relevance
1530	149	25	66	1230	3000	99.5
1381	247	125	580	667	3000	98.5
1679	89	2	113	1117	3000	99.1
1603	201	99	174	923	3000	99.0
1598	5	15	1	1381	3000	98.8
128	81	1	26	124	360	100.0
216	154	0	142	132	644	98.4
1624	126	16	165	581	2512	99.8
351	16	2	16	27	412	100.0
647	119	71	40	439	1316	98.4
1174	111	20	64	789	2158	99.1
335	17	1	70	166	589	97.5
641	383	4	112	753	1893	97.8
1071	546	2	453	928	3000	88.4
1554	197	28	327	894	3000	97.6
1506	120	6	25	393	2050	98.3
305	270	24	89	181	869	99.0
371	1017	496	570	546	3000	95.7
976	936	405	594	89	3000	96.6
344	12	10	70	156	592	98.3
845	139	323	175	182	1664	97.9

Assessment results

Possible to estimate reliably:

$$f'^{+} = \frac{|F^{+} \cap P^{-}|}{|F^{+}|} \sim 1.39\%$$

$$f^{-} = \frac{|F^{-} \cap P^{+}|}{|P^{+}|} \sim 24.5\%$$

Expression:

$$|P^{+}| = \frac{|F^{+}|(1 - f'^{+})}{1 - f^{-}}$$

P^{+} is the number of paedophile queries in our set

F^{+} the number of queries tagged as paedophile by your tool.

Assessment results

Possible to estimate reliably:

$$f'^+ = \frac{|F^+ \cap P^-|}{|F^+|} \sim 1.39\%$$

$$f^- = \frac{|F^- \cap P^+|}{|P^+|} \sim 24.5\%$$

Expression:

$$|P^+| = \frac{|F^+|(1 - f'^+)}{1 - f^-}$$

P^+ is the number of paedophile queries in our set

F^+ the number of queries tagged as paedophile by our tool.

Fraction of paedophile queries

2.5 queries out of 1,000 are paedophile in our datasets

Fraction of paedophile users 1/2

Classical hypothesis:
user $\sim$ IP address
132.227.84.51
Dataset: 2007 and 2009

Problems: gateway/firewall (NAT) IP addresses, dynamic addresses
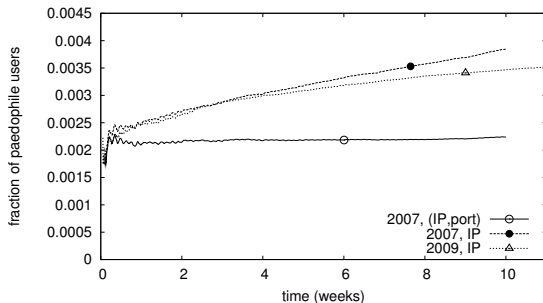allocation, several users per computer, several computers per user

Better:
user $\sim$ IP address + connection port
132.227.84.51:4964
Dataset: 2007 only

Fraction of paedophile users 2/2

- pollution: all dynamic/public IP addresses may be considered as paedophile *after some time*
- convergence when considering IP+ port

Approx. 2.2 users out of 1,000 are paedophile in our datasets

Deeper analysis

- Age-related queries

- Maps of paedophile users using IP geolocation

- Content rating system

- Temporal evolution of the use of paedophile keywords

- User behaviors

Outline

4 **Conclusion**

Conclusion (1/2)

Main outcomes

- Complementary measurement methods
  - huge datasets, long durations, available for research purposes
- Automatic detection tool
  - Rigourous quantification
    2.5 queries out of 1,000 are paedophile
    2.2 users out of 1,000 are paedophile
- More : ages, geolocation-based maps, content rating system, user behaviours, temporal evolution of keywords.

Conclusion (2/2)

Going further

- Other analyses of our datasets
  (particularly the set of paedophile queries)

- Implement monitoring using our tool

- Study other P2P systems

Resources

Contact address: Raphael.Fournier@lip6.fr
Project leader: Matthieu.Latapy@lip6.fr

```
http://antipaedo.lip6.fr
```

Thank you for your attention.