

# Detection and analysis of paedophile activity in P2P networks

Raphaël Fournier-S'niehotta



CIRCAMP

February, 26th 2013

# Context

## Antipaedo project

<http://antipaedo.lip6.fr>

### Funding

- European commission MAPAP project
- French Research agency (ANR)
- Action Innocence Monaco

### Partners

- |                                       |         |
|---------------------------------------|---------|
| • University of Ljubjana Slovenia     | • CNRS  |
| • Nobody's Children Foundation Poland | • LORIA |
| • University College Cork             | • LEA   |

# P2P systems and paedophile activity



- tens of millions of users worldwide

# P2P systems and paedophile activity



- tens of millions of users worldwide

## Rationale

- Children victimization
- Danger for innocent users
- Policy making issues
- Societal problem

Very little is known

# Goals

## Help fight against P2P paedophile exchanges

- increase knowledge
- provide tools

### Detection

- Create an automatic tagging tool

### Analysis

- Rigorous quantification of queries
- Paedophile users
- Temporal evolution
- Comparison between different P2P systems

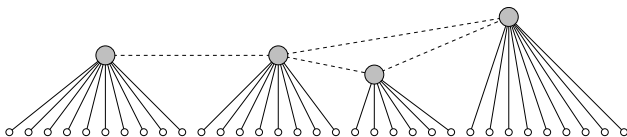
# Challenges

- Appropriate data collection  
size, dynamics, poorly documented protocols
- Automatic detection tool  
hidden activity, several languages
- Rigorous statistical inference  
low amount of paedophile queries
- User identification  
partial information, unreliable

# Outline

- 1 Data collection
- 2 Paedophile queries
- 3 Paedophile users
- 4 Temporal dynamics
- 5 Comparaison KAD eDonkey
- 6 Conclusion

# eDonkey system overview



peer		server
keywords	→	
	←	filelist
file	→	
	←	providers

## Measurements

- **server**
- client, queries
- client, honeypot



# Collected data

- Queries submitted to eDonkey search engine

- Five large datasets collected:

**2007** 10 weeks, 100 millions queries, 24 million IP addresses

**2009-2012** 147 weeks, 1,3 billion queries, 82 million IP addresses

**KAD** 10 days, 250 000 queries

**ed2k-FR** 60 days, 241 152 queries

**ed2k-UA** 60 days, 166 154 queries

Duly anonymised

# Examples of 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

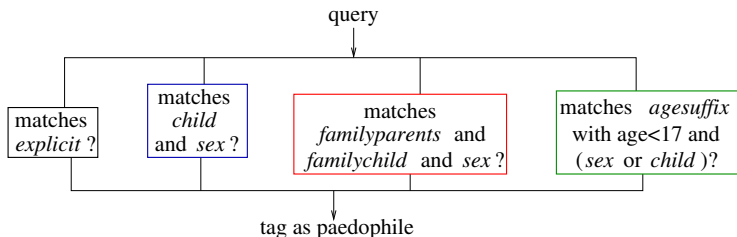
...

# Outline

- 1 Data collection
- 2 Paedophile queries
  - Tool design
  - Tool assessment
  - Fraction of paedophile queries
- 3 Paedophile users
- 4 Temporal dynamics
- 5 Comparaison KAD eDonkey
- 6 Conclusion

# Tool design

- Set of rules based on law-enforcement knowledge
- Manual inspection of our datasets
- Improve until negligible changes
- 4 categories of paedophile queries



raygold 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 12yo”*  
contains paedophile keywords that we don't search for

How to estimate false positive and false negative rates?

# Tool assessment – Survey

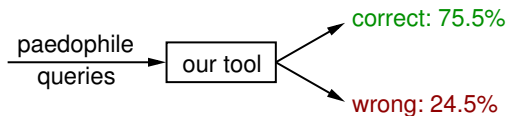
- Set of 21 volunteering experts from Europol, national authorities, NGOs
- Set of 3,000 **randomly selected** queries:
  - Paedophile
  - Not paedophile
  - *Neighbours* (submitted within the 2 previous or next hours of a paedophile query by the same user)
- Tag queries as *paedophile*, *probably paedophile*, *probably not paedophile*, *not paedophile* or *I don't know*

# Tool assessment – Survey results

<i>paedo</i>	<i>prob. paedo</i>	<i>don't know</i>	<i>prob. not</i>	<i>not paedo</i>	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
<b>1174</b>	<b>111</b>	<b>20</b>	<b>64</b>	<b>789</b>	<b>2158</b>	<b>99.1</b>
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

- Relevance rate: adequate knowledge of specific context

# Assessment results



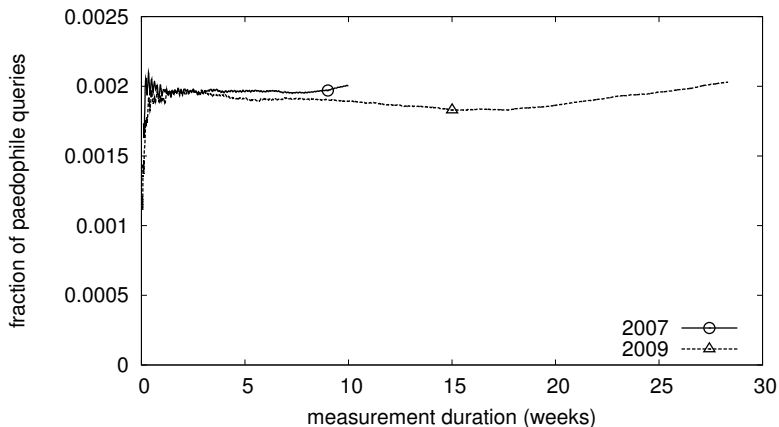


# Assessment results

$$\frac{|P^+|}{|D|} = \frac{(1 - f'^+)}{1 - f^-} \cdot \frac{|T^+|}{|D|}$$

- $P^+$  : paedophile queries
- $T^+$  : tagged paedophile queries
- $f'^+$  : false positive rate
- $f^-$  : false negative rate

# Fraction of detected paedophile queries



# Fraction of paedophile queries

## Validation result

- Detected queries: slightly above 0.19% for both datasets (2010)
- After correction: 2,5 queries out of 1,000 are paedophiles
- 1 paedophile query every 33 seconds



MATTHIEU LATAPY, CLÉMENCE MAGNIEN, AND RAPHAËL FOURNIER. Quantifying paedophile queries in a large P2P system. In *IEEE International Conference on Computer Communications (INFOCOM) Mini-Conference*, 2011.



MATTHIEU LATAPY, CLÉMENCE MAGNIEN, AND RAPHAËL FOURNIER. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, In press, 2012.

# Outline

- 1 Data collection
- 2 Paedophile queries
- 3 Paedophile users**
  - Distinguishing different users
  - Fraction of paedophile users
- 4 Temporal dynamics
- 5 Comparaison KAD eDonkey
- 6 Conclusion

# Distinguishing users

Possible approximation:  
user  $\sim$  IP address

## Problems

- Gateway/firewall (NAT) IP addresses
- Dynamic addresses allocation
- Several users per computer
- Several computers per user

# Distinguishing users

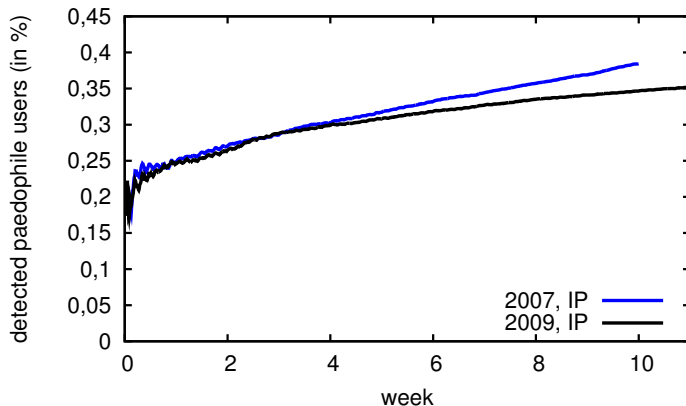
## Paedophile user

- User paedophile after one paedophile query
- All dynamic/public IP addresses may be considered as paedophile *after some time*

3 methods:

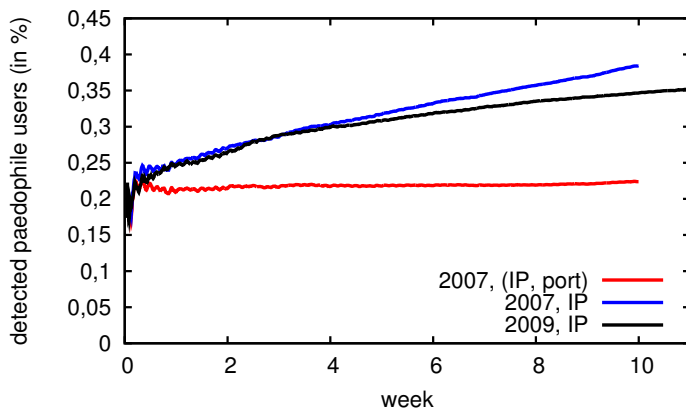
- User  $\sim$  IP adress + connection port
- Measurement duration
- Sessions

# User: IP vs (IP,port)



- (IP, port) reduces pollution (bias)

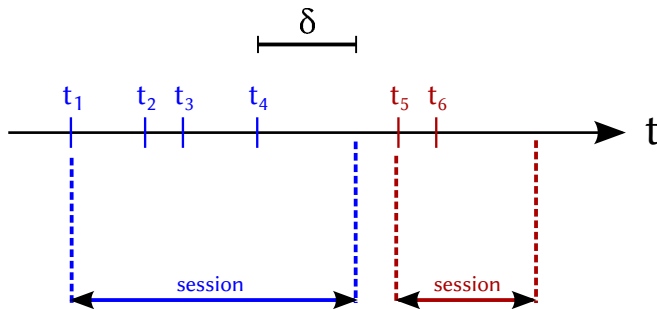
# User: IP vs (IP,port)



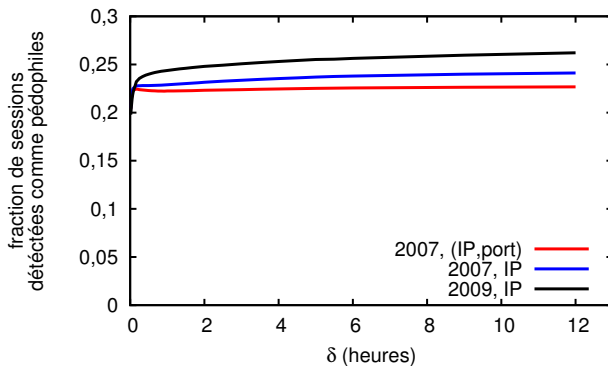
- (IP, port) reduces pollution (bias)



# User: sessions



# User: sessions



# Fraction of paedophile users

## Result

- Close to 0,22% for both datasets (2010)
- 1 paedophile user out of 450



MATTHIEU LATAPY, CLÉMENCE MAGNIEN, AND RAPHAËL FOURNIER. Quantifying paedophile queries in a large P2P system. In *IEEE International Conference on Computer Communications (INFOCOM) Mini-Conference*, 2011.

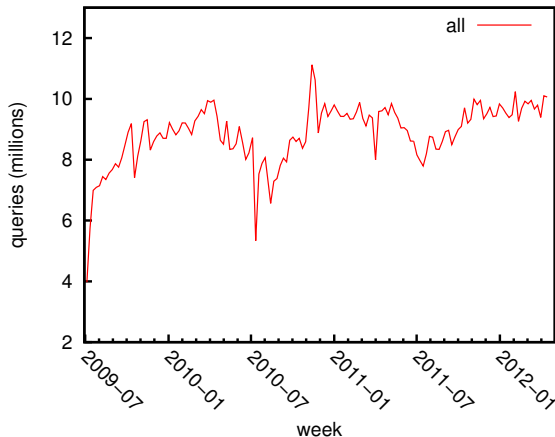


MATTHIEU LATAPY, CLÉMENCE MAGNIEN, AND RAPHAËL FOURNIER. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, In press, 2012.

# Outline

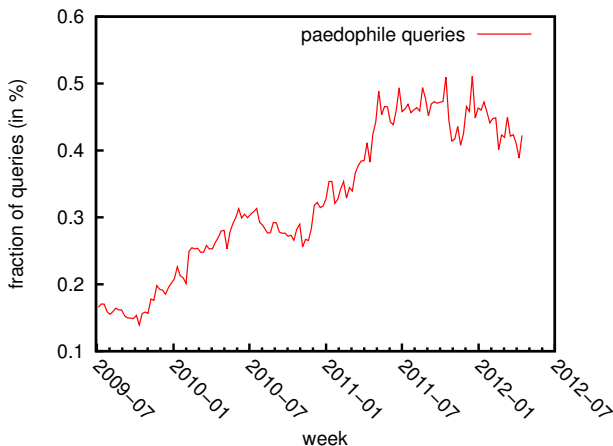
- 1 Data collection
- 2 Paedophile queries
- 3 Paedophile users
- 4 Temporal dynamics**
  - Long-term evolution of paedophile activity
  - Daily evolution of paedophile activity
- 5 Comparaison KAD eDonkey
- 6 Conclusion

# Global traffic on server



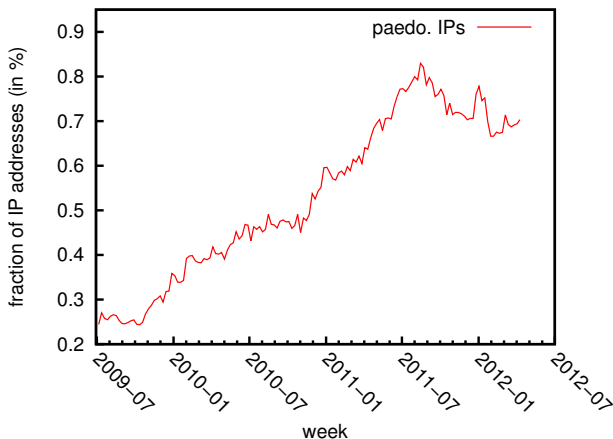
- Stability of global traffic over 3 years

# Fraction of paedophile queries



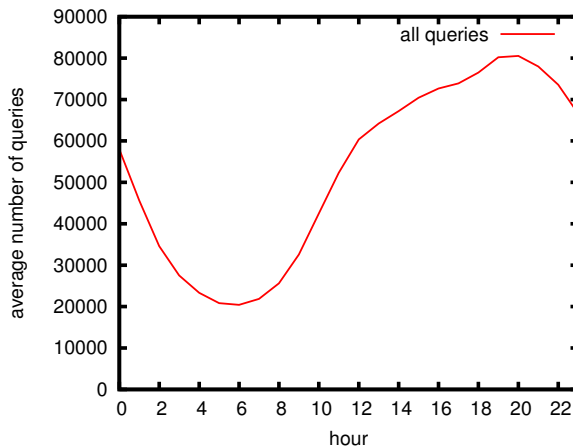
- Fraction of paedophile queries strongly increasing

# Fraction of paedophile users



- Fraction of paedophile users also increasing

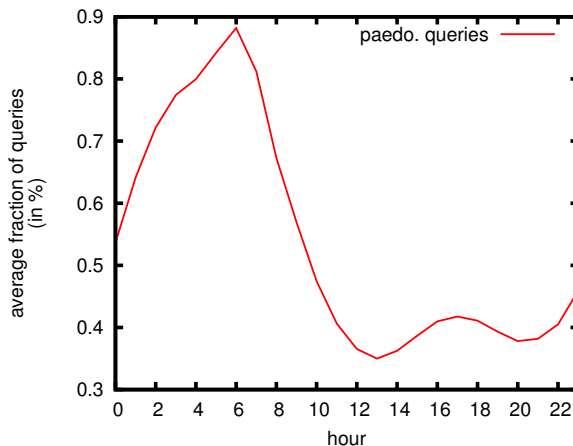
# Daily traffic



- Circadian cycle (day/night effect)

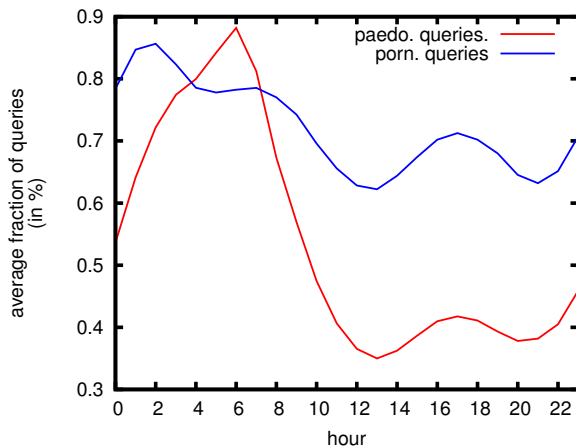


# Fraction of paedophile activity



- Fraction of paedophile queries peaks at 6 AM

# Pornography vs paedophile activity



- Paedopornography and traditional pornography differ

# Evolution of paedophile activity

## Results

- Important growth of paedophile activity between 2009 and 2012
- Fraction of paedophile queries peaks at 6 AM
- Qualitative contribution with quantitative approach

# Outline

- 1 Data collection
- 2 Paedophile queries
- 3 Paedophile users
- 4 Temporal dynamics
- 5 Comparaison KAD eDonkey**
- 6 Conclusion

# KAD network

## Overview

- Completely distributed protocol of clients
- No server for file indexing
- Some peers are in charge of some files and keywords

more anonymous and prone to illegal activity?

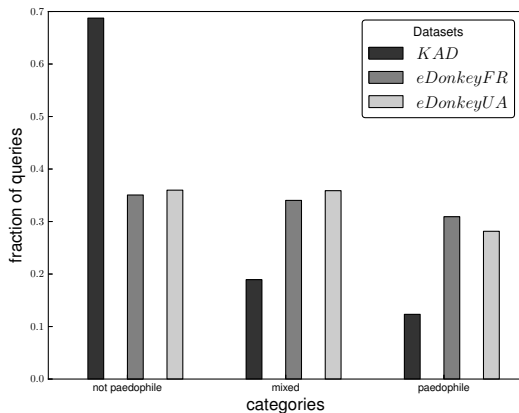
## Monitoring

- Targeted injection of fake peers to control files or keywords
- Peers catch queries and control replies
- Limited to keywords for comparison with eDonkey



R. FOURNIER, T. CHOLEZ, M. LATAPY, C. MAGNIEN, I. CHRISMENT, I. DANILOFF AND O. FESTOR.  
Comparing paedophile activity in different P2P systems. *Submitted*.

# KAD network



- overall: approx. **twice less** paedophile activity in KAD

# Outline

- 1 Data collection
- 2 Paedophile queries
- 3 Paedophile users
- 4 Temporal dynamics
- 5 Comparaison KAD eDonkey
- 6 Conclusion**

# Conclusion (1/2)

## 1 Paedophile queries

- automatic detection tool
- set of paedophile queries
- estimated fraction of paedophile queries

## 2 Paedophile users

- general study of user identification
- quantification of paedophile users



# Conclusion (2/2)

## ③ Temporal dynamics

three-year study  
user social integration

## ④ Comparing KAD and eDonkey

adequate methodology  
analysis with partial information

# Perspectives

## Tool improvement

- previous/next queries
- languages, word order, categories
- machine learning

## Analysis

- different thresholds for paedophile users
- community detection (users' social network)
- detailed study of sequences of queries
- file exchanges (supply)

# Final word

## Fruitful collaboration between LEA and academic labs

- keywords  $\leftrightarrow$  filters
- measure  $\leftrightarrow$  data/results
- field opened for other scientists

# Contact

Thank you for your attention.

[raphael.fournier@lip6.fr](mailto:raphael.fournier@lip6.fr)

<http://antipaedo.lip6.fr>





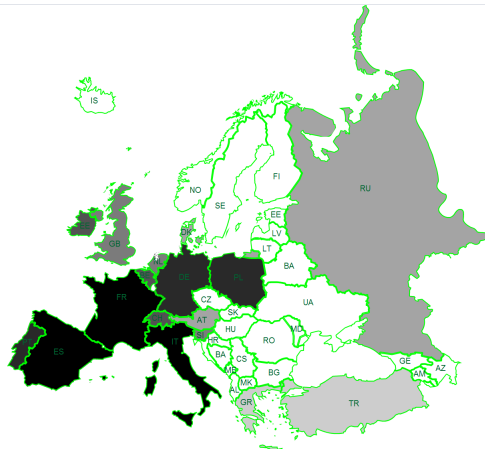
# Geo-location: statistics

country	# queries	# paedo	ratio
IT	19569361	15426	0.08 %
ES	8881405	5177	0.06 %
FR	7583815	8059	0.11 %
BR	2795090	4849	0.17 %
IL	2139697	2618	0.12 %
DE	2093106	11238	0.54 %
KR	1386799	336	0.02 %
US	1053183	6184	0.59 %
PL	975170	1178	0.12 %
AR	810466	1465	0.18 %
CN	635392	337	0.05 %
PT	513327	434	0.08 %
IE	511185	54	0.01 %
TW	417893	138	0.03 %
BE	402565	646	0.16 %
CH	320054	1710	0.53 %
GB	319386	1698	0.53 %
NL	243646	1131	0.46 %
CA	241460	1233	0.51 %
SI	239572	167	0.07 %
MX	210504	1098	0.52 %
RU	200958	2712	1.35 %
AT	184248	977	0.53 %

Biased by:

- language knowledge
- decoding problems

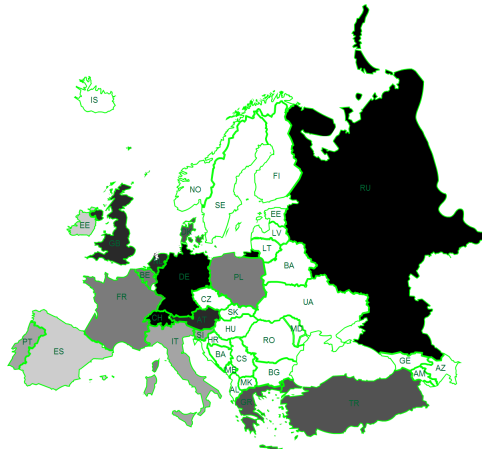
# Geo-location: maps



# queries



# Geo-location: maps



ratio # paedophile queries / # queries