Quantifying Paedophile Queries in a Large P2P System

Matthieu Latapy, Clémence Magnien and Raphaël Fournier

LIP6

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

IEEE INFOCOM Mini-Symposium, Shanghai

AGENCE NATIONALE DE LA RECHERCHE
ANR

What is the real extent of
paedophile activity
in peer-to-peer systems?

Rationale

- Children victimization
- Danger for innocent users
- Societal problem

Very little is known

Goal

- Detect and quantify paedophile queries

Challenges

- Appropriate data collection
  size, dynamicity
- Automatic detection tool
  hidden activity, several languages
- Rigorous statistical inference
  low amount of paedophile queries

Datasets

- eDonkey
- semi-centralized

	duration	queries
2007	10 weeks	107,226,021
2009	28 weeks	205,228,820

Main features of our two datasets

Duly anonymised

F. AIDOUNI, M. LATAPY, AND C.MAGNIEN. Ten weeks in the life of an edonkey server. *Proceedings of HotP2P'09*, 2009.

O. ALLALI, M. LATAPY, AND C. MAGNIEN. Measurements of *eDonkey* activity with distributed honeypots. *Proceedings of HotP2P'09*, 2009.

Collected 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
. . .

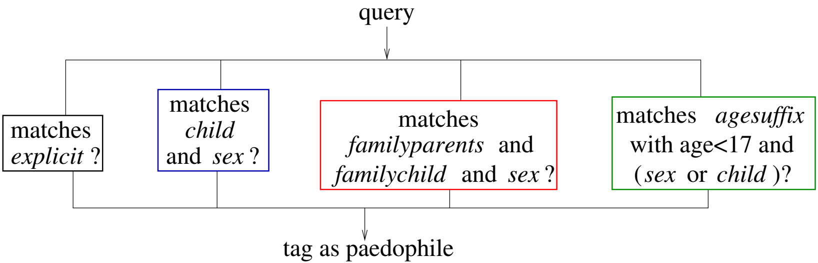
Outline

1. Detection of paedophile queries
   - Tool design
   - Tool assessment

Tool design

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

Tool design: detection steps


query
matches *explicit* ?
matches *child* and *sex* ?
matches *familyparents* and *familychild* and *sex* ?
matches *agesuffix* with age<17 and (*sex* or *child*)?
tag as paedophile

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 (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

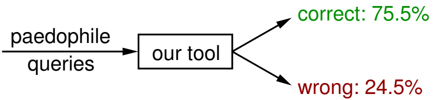
	prob.	*don’t*	*prob.*	*not*		
paedo	*paedo*	*know*	*not*	*paedo*	total	relevance
1530	149	25	66	1230	3000	99.5
1381	247	125	580	667	3000	98.5
1679	89	2	113	1117	3000	99.1
1603	201	99	174	923	3000	99.0
1598	5	15	1	1381	3000	98.8
128	81	1	26	124	360	100.0
216	154	0	142	132	644	98.4
1624	126	16	165	581	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

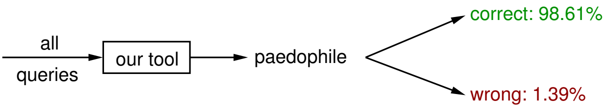
- relevance rate: adequate knowledge of specific context

Assessment results

Limited filter precision

- False negatives
- False positives



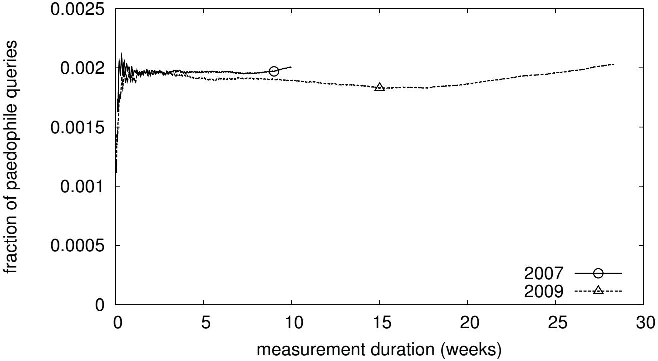


Outline

2 Quantification of paedophile queries

Fraction of paedophile queries

- slighlty above 0.19% for both datasets



Fraction of queries detected as paedophile since the beginning

Inference

Expression:

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

Inference

Expression:

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

- $\sim$ 2.5 queries out of 1,000 are paedophile in our datasets

Outline

3 Conclusion

Conclusion

Paper contributions

- General approach for detecting rare contents
- Automatic detection tool
- Set of paedophile queries
- Rigorous quantification
  2.5 queries out of 1,000 are paedophile

Deeper analysis

- User quantification (based on IP identification)
- Maps of paedophile users using IP geolocation
- Temporal evolution of the use of paedophile keywords
- Age-related queries

Resources

Thank you for your attention.

Matthieu.Latapy@lip6.fr
Clemence.Magnien@lip6.fr
Raphael.Fournier@lip6.fr

`http://antipaedo.lip6.fr`

Client measurement (1/3) – principle

***fake* peer** ⟶ Keyword-based query ⟶ **server**

pthc

***fake* peer** ⟵ List of some files with the keywords ⟵ **server**

pthc-12yo-1.jpg
pthc-11yr-1.jpg
pthc-11yr-2.jpg

***fake* peer** ⟶ List of all these files ⟶ **server**

pthc-12yo-1.jpg
pthc-11yr-1.jpg
pthc-11yr-2.jpg

***fake* peer** ⟵ List of some peers providing the files ⟵ **server**

peer123,peer234,
peer345,. . . ,peer456

Client measurement (2/3) – Results

Measurement setup

- periodically sending of keyword queries
- for each discovered file, query server for providers
- geo-location of peers

Client measurement (2/3) – Results

Measurement setup

- periodically sending of keyword queries
- for each discovered file, query server for providers
- geo-location of peers

Obtained data with focus on paedophile activity

- 1 client, approx. 100 servers
- queries for paedophile and non paedophile keywords
- 7 months
- 3 million files (800 000 paedophile)
- 3.5 million peers (1.3 million providers of paedophile)

Client measurement (3/3)

Advantages

- No server authorization required
- Several servers studied simultaneously

Drawbacks

- Focus on some keywords only

eDonkey system – Honeypot measurements

honeypot	→	List of shared files	→	**server**
peer	→	Keyword-based query	→	**server**
peer	←	List of files matching the keywords	←	**server**
peer	→	Some of our files	→	**server**
peer	→	file download request	→	**honeypot**

Honeypot client

- Specific paedophile keywords or files
- Measurement length
- Distributed on several computers

First measurement

- one month
- distributed on 40 machines
- 24 649 peers

Honeypot client

- Specific paedophile keywords or files
- Measurement length
- Distributed on several computers

Second measurement

- one month
- 1 client, providing all known files
- 870 000 peers
- 275 000 files

Honeypot client

- Specific paedophile keywords or files
- Measurement length
- Distributed on several computers

Second measurement

- one month
- 1 client, providing all known files
- 870 000 peers
- 275 000 files

Effective but interfers with law-enforcement

Honeypot

Advantages

- No server authorization required
- Several servers studied simultaneously

Honeypot

Advantages

- No server authorization required
- Several servers studied simultaneously

Drawbacks

- Focus on some keywords only
- Interfers with law-enforcement monitoring

KAD network

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

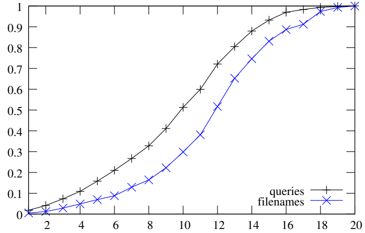
Principle:

- Precise and targeted injection of peers into the network to control files or keywords
- Peers catch queries and control replies

Applications:

- Which files are published for a given keyword? Which peers share them ?
- Eclipse : prevent peers from accessing content

Ages



x : ages *xyo*

y : fraction of occurrences with age $\leq$ *x*

$\leq$ 10 years old : 50% (queries) et 30% (files)

$\leq$ 5 years old : 15% (queries) et 7% (files)

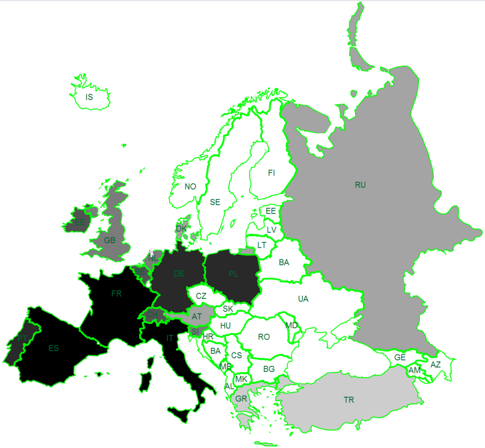
Geo-location: statistics

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:

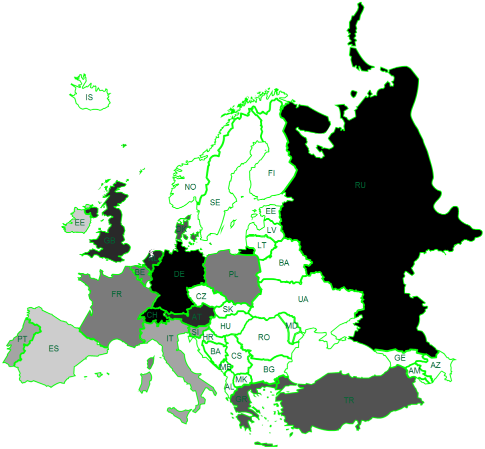
- language knowledge
- decoding problems

Geo-location: maps



# queries

Geo-location: maps



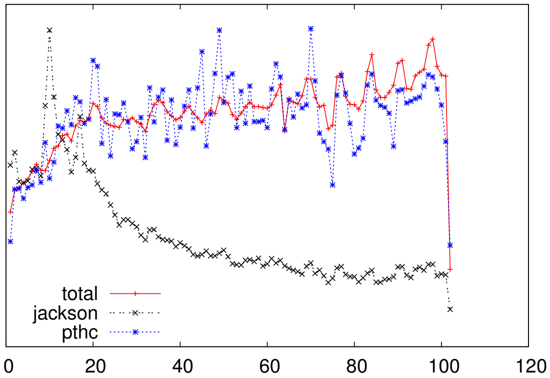
ratio # paedophile queries / # queries

Keyword dynamics

Does the rate of use of different keywords **evolve over time?**

- Used keyword detection methods at different times
- No significant change

Example



Comparative evolution of a general and a paedophile keyword

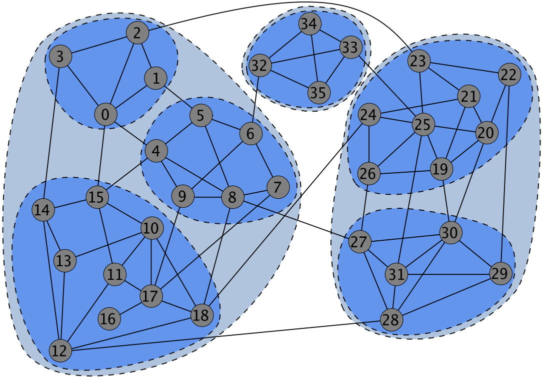
Content rating and *fake detection* systems

Automatic methods for deciding if

- a given file has pornographic/paedophile content
- the file's content is significatively different from its name

Goals: protect users
help for classification

Principle

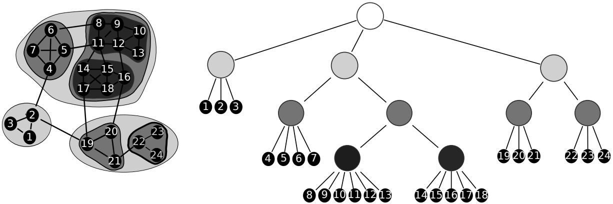


Two files are linked if many peers provide them both

Principle

File rating method

1. a graph capturing user interests
2. partition in *communities* (similar files)
3. a set of files known to be paedophile
4. for each community: % of files known to be paedophile
5. file rating: average of % for all its communities



Fake detection

Fake: file with name different from content

Method

- Relies on paedophile file name detection
- Paedophile name, low rating: fake
- Non paedophile name, high rating: fake

Same applies for pornographic content rating and fake detection

Validation

Two approaches:

- seek information about some files
  - files considered as paedophile, but low ratings?
  - unkown files with high ratings?
  - known files with high ratings?
- remove part of the initial information

positive results

Web interfaces – Demo

`http://antipaedo.complexnetworks.fr`

Paedophile keyword detection

Framework

- Comparison of 7 different methods
- initial knowledge: known paedophile keywords
    - 2 scenarios
- All rely on keywords co-occurring in filenames

Methods

- Span the currently existing techniques
- Three developed within the project (CNRS, UL)
- Involvement of linguists

Validation

Methodology

- 10 international experts (all partners)
- Rating for 189 words (given by at least one method)
- *Specific paedophile*, *paedophile*, *I don't know* or *general*

Results

- General agreement among experts
- All methods give promising results
- Two methods perform really well
- Manual inspection of results needed

Frequencies in obtained lists

frequencies: **lsm** (7), **ygold** (6), **qqaazz** (6), **ptsc** (6), **pedo** (6), **lsbar** (6), **ls** (6), **childlover** (6), **underage** (5), sandra (5), **pthc** (5), mylola (5), magazine (5), lsn (5), **kleuterkutje** (5), **kdquality** (5), jenny (5), **hussyfan** (5), daughter (5), **childfugga** (5), child (5), **babyj** (5), vicky (4), boy (4), vdbest (3), tori (3), rbv (3), **preteen** (3), novinhas (3), newer (3), **mafiasex** (3), little (3), **kingpass** (3), **kdv** (3)