Paedophile activity in a large P2P system

Matthieu Latapy, Clémence Magnien and Raphaël Fournier
LIP6, CNRS and Université Pierre et Marie Curie (UPMC), Paris, France

1 paedophile query every 33 seconds
0.22 % users enter such queries

Measurements

► eDonkey P2P network, one of the largest in use
► Recording of keyword queries submitted to the search engine

Two measurements

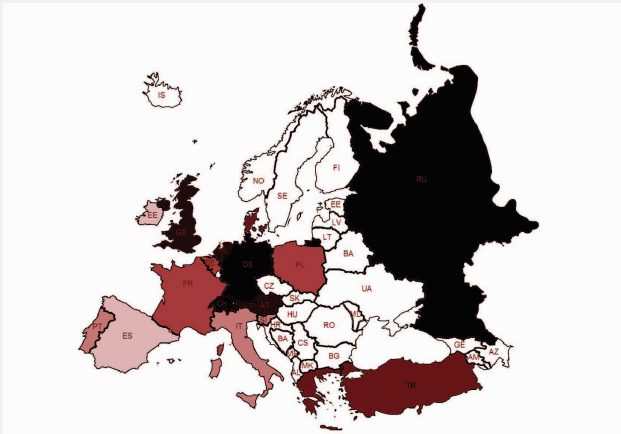|	2007	2009
duration	10 weeks	28 weeks
observed queries	107 226 021	205 228 820
IP addresses involved	23 892 531	24 413 195

+ strong anonymisation

examples

Detection tool

Tags any query as paedophile or not paedophile

4 categories of paedophile queries:
► with "explicit" keywords (ex.: qqaazz little girl)
► with "child" and "sex" related keywords (ex.: porno infantil)
► with "parents", "child" and "sex" keywords (ex.: incest mom son video)
► with an age below 16 and a "child" or "sex" keyword (ex.: 12yo fuck video)

examples

Map of paedophile activity

Paedophile activity in Europe. The darker a country is coloured, the higher the fraction of paedophile queries received from this country.

WARNING: language and encoding issues, erroneous geolocation, etc.

Contributions

► Large-scale datasets on P2P activity
► Largest sets of paedophile queries
► Reference methodology
► Automatic detection tool
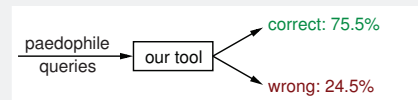► Rigourous quantification

References

[1] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier.
Quantifying paedophile activity in a large P2P system.
IEEE Infocom Mini-Conference, 2011.
To appear. Preprint available at http://antipaedo.lip6.fr.

Typical data

sirenia at sixes and sevens
dj coupe decale
hannah montana clear
mino reitano discografia
qqaazz little girl
el gallo sube
motherfucker of the year
h2o the last prime minister
devenir male dominant
saghe mentali
gram parsons wild horses
sheherazade korsakov
naruto
incest mom son video
desaparecidos fiesta loca
secret life american vostfr
hysteric ego want love
cuatro vidas
komedie pl
wwe the bash
blue water high ita
cheryl ladd
der komtur
le carte leggendarie gba
pittbul
er medici in prima linea cover
beethoven cello sonatas rostropovich richter
cafe del mar

capricorne one
dream dance vol
vedrai vedrai
andy moor faces
the mentalist s01e19
ich mich nach deiner liebe soundtrack
fine ukraine
dash berlin man on the run
porno infantil
kyle 4x07
billie jean body
taviani
una voce nella notte ost
paolo conte schiava del live
schiavi padroni
12yo fuck video
amedeo minghi alla fine
michael jackson bad man in the mirror
film alexandre le grand theme
shema plantation interieur doc
dragon ball ita video
warriors of terra
cd sertao
computer idea ita
boxing fights video
mirko crocop the story video
la maledizione del forziere fantasma
neil hannon

Tool assessment

21 independent experts (Europol, national authorities, NGOs)

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

Error rates:

all queries → our tool → paedophile → correct: 98.61% / wrong: 1.39%

paedophile queries → our tool → correct: 75.5% / wrong: 24.5%

↪ few errors, rigorous quantification

Quantification results

► 0.25% queries are paedophile in our datasets

► 0.22% users are paedophile in our datasets

Perspectives

► Study of other networks (KAD, Gnutella, etc.)
► Behaviours of paedophile users
► Temporal evolution
► Age-related queries
► Content rating system