Paedophile activity in large P2P systems

R. Fournier, T. Cholez, C. Magnien, I. Chrisment, M. Latapy and O. Festor
LIP6 (CNRS/UPMC) and LORIA, France

Measurements

- *eDonkey* and *KAD* P2P network, two of the largest in use
- recordings of queries submitted to search engines

examples →

  - ▷ first measurement: collection of all keyword-based queries on two *eDonkey* servers, at different times

	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

  - ▷ second measurement: collection of 1-keyword-long queries in the *KAD* network and on two *eDonkey* servers (in 2010)

	KAD	$eDonkey_1$	$eDonkey_2$
queries	250,000	241,152	166,154

    - passive monitoring of *KAD* with the HAMACK[**?**] architecture: distributed honeypots exploit the DHT mechanisms to collect keyword-indexed requests.
    - supervision of 72 keywords, in 3 categories: paedophile, mixed (mostly ages) and not paedophile.

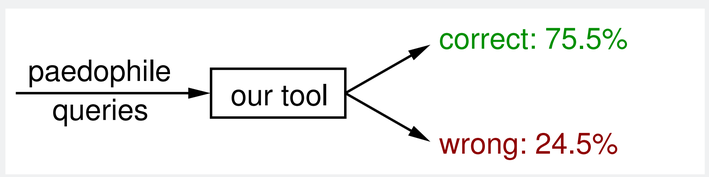
Detection tool

- tags any query as paedophile or not paedophile
- 4 categories of paedophile queries:
  - ▷ with "explicit" keywords (*qqaazz little girl*)
  - ▷ with "child" and "sex" related keywords (*porno infantil*)
  - ▷ with "parents", "child" and "sex" keywords (*incest mom son video*)
  - ▷ with an age below 16 and a "child" or "sex" keyword (*12yo fuck video*)

examples →

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:



↪ few errors, rigorous quantification

Quantification results

- 0.25% queries are paedophile in our *eDonkey* datasets
- 0.10% queries are paedophile in our *KAD* datasets
  ↪ 2.5 times less paedophile queries in *KAD* than in *eDonkey*
- 0.22% users are paedophile in our *eDonkey* datasets

References

[1] T. Cholez, I. Chrisment, and O. Festor. Monitoring and Controlling Content Access in KAD. In *IEEE International Conference on Communications ICC*, May 2010.

[2] R. Fournier, T. Cholez, M. Latapy, I. Chrisment, I. Daniloff, and O. Festor. Comparing paedophile activtity in different P2P systems. *submitted*, 2011.

[3] M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile queries in a large P2P system. In *IEEE INFOCOM Mini-Conference*, 2011.

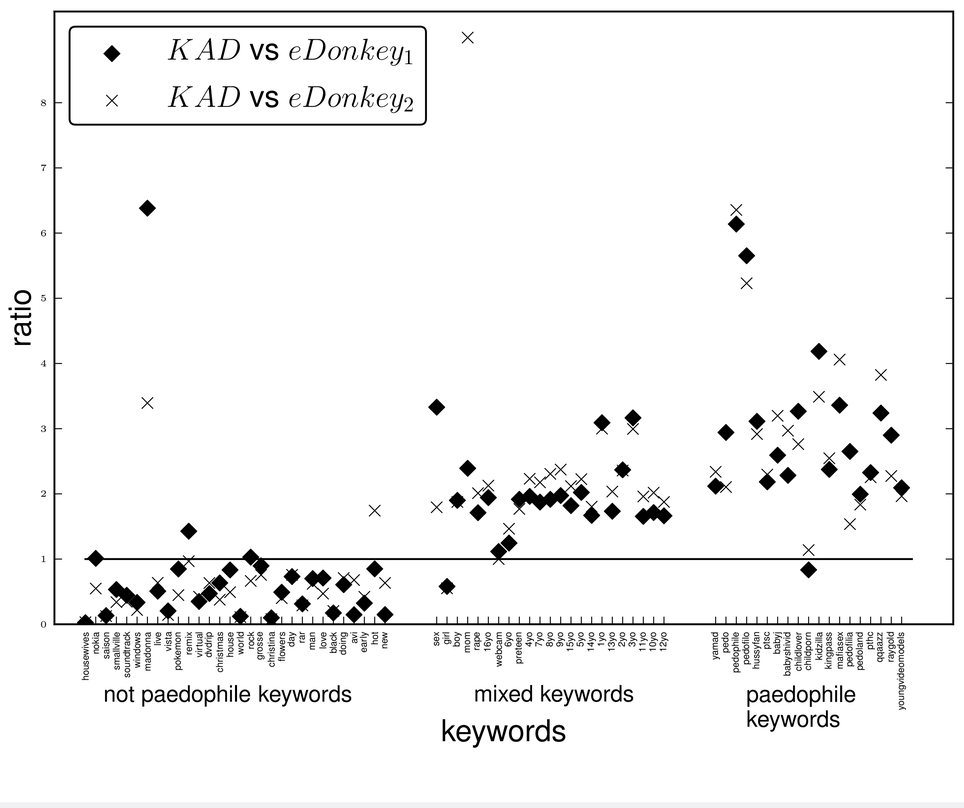
[4] M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, To appear.

[5] G. Montassier, T. Cholez, G. Doyen, R. Khatoun, I. Chrisment, and O. Festor. Content Pollution Quantification in Large P2P networks : a Measurement Study on KAD. In *IEEE International Conference on Peer-to-Peer Computing*, 2011.

Typical data

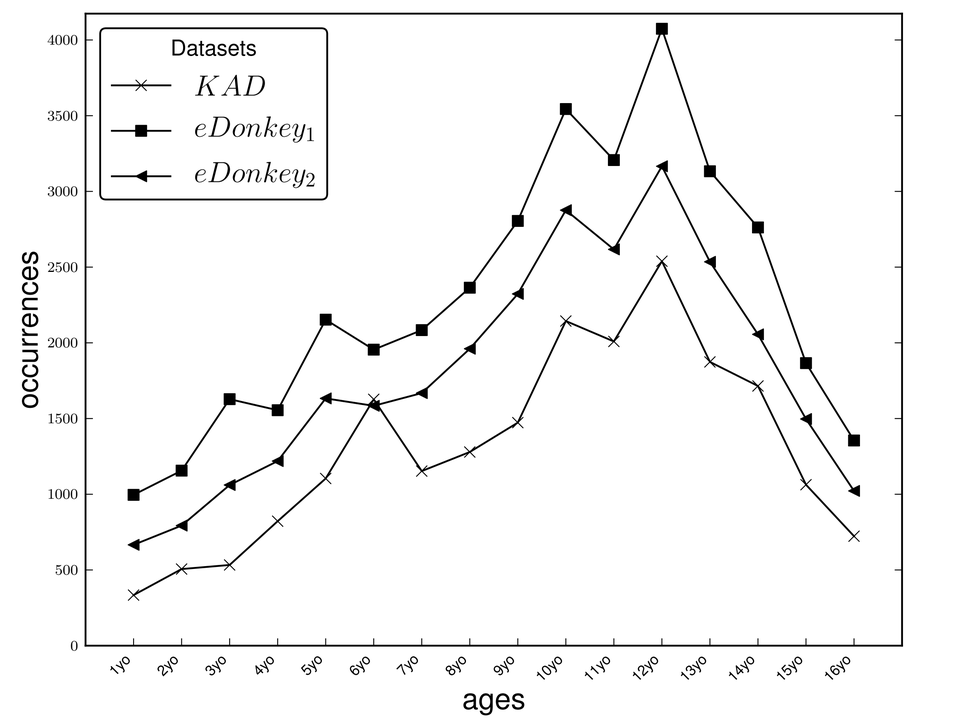
sirenia at sixes and sevens
dj coupe decale
hannah montana clear
mino reitano discografia
qqaazz little girl
el gallo sube
motherfucker of the year
h2o the last prime minister
devenir male dominant
saghe mentali
gram parsons wild horses
sheherazade korsakov
naruto
capricorne one
dream dance vol
the mentalist 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
incest mom son video
desaparecidos fiesta loca
secret life american vostfr
12yo fuck video
amedeo minghi alla fine
michael jackson bad man in the mirror

Comparison KAD vs eDonkey



Keywords below the horizontal $\mathbf{y = 1}$ line are more present in *KAD* than in an *eDonkey* dataset. Paedophile keywords are significantly more present in eDonkey.

Ages in queries



Distributions of ages in all three datasets are very similar: although the amount of paedophile activity varies between systems, its nature is similar, at least regarding ages.

Conclusions and perspectives

- Large-scale datasets on P2P activity
- Largest sets of paedophile queries
- Reference methodology
- Automatic detection tool
- Rigourous quantification
- Temporal evolution
- Behaviours of paedophile users
- Content rating system