Scientific openness with sensitive data

De l'ouverture des données scientifiques sensibles

Raphaël Fournier-S'niehotta

Journée SoData!, IGN

14 mars 2013

Contexte

- équipe ComplexNetworks : grands graphes de terrain et réseaux sociaux, mesure de l'Internet (carte)

  compétences en collecte de données

- projets MAPE (ANR) et MAPAP (CE) : Measurements and Analysis of P2P Activity against Paedophile content

  Antipaedo http://antipaedo.lip6.fr

  sujet sensible

Contexte (suite)

L'activité pédophile dans le pair-à-pair (P2P)

- Victimes directes
- Danger pour les utilisateurs non pédophiles
- Impact sur la régulation de l'Internet

Très peu de connaissances

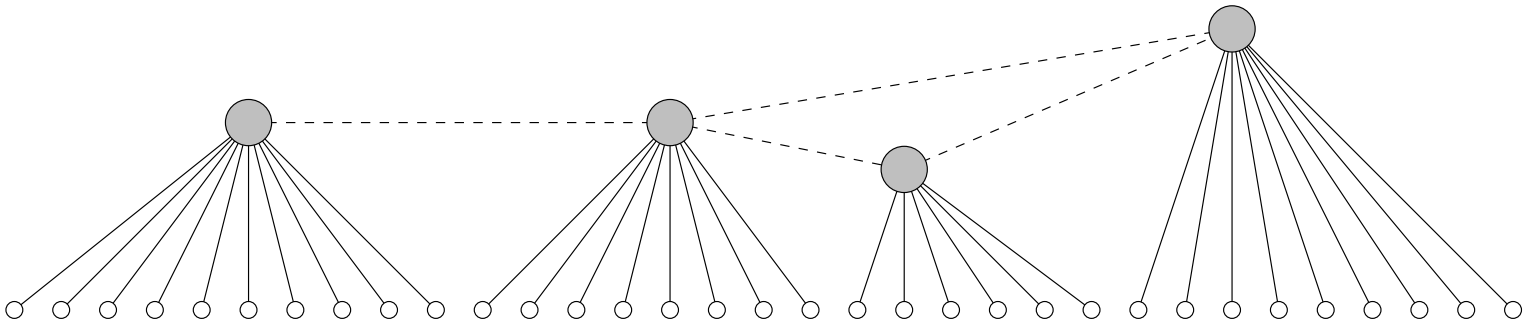
Objectifs

- Quantifier les requêtes et les utilisateurs
- Étudier l'évolution de l'activité
- Comparer différents réseaux

Améliorer significativement les connaissances sur l'activité pédophile dans le P2P

Réseaux P2P

- des millions d'utilisateurs



utilisateur		**serveur**
mots-clefs	$\longrightarrow$	
	$\longleftarrow$	liste de fichiers
fichier(s)	$\longrightarrow$	
	$\longleftarrow$	fournisseur(s)

Données

2 collectes en continu sur eDonkey :

2007 10 semaines, 100 millions de requêtes, 24 millions d'IP

2009 147 semaines, 1,3 milliard de requêtes, 82 millions d'IP (géolocalisées)

1 collecte pour étudier KAD [1] :

KAD 10 jours, $\sim$ 250 000 requêtes

ed2kFR 60 jours, $\sim$ 240 000 requêtes

ed2kUA 60 jours, $\sim$ 170 000 requêtes

Contient des informations sensibles

T. Cholez, I. Chrisment, and O. Festor. Monitoring and Controlling Content Access in KAD. *ICC 2010.*

Des données sensibles

12/03-02:48:08	once upon a time s02e16	 <Rio De Janeiro>
12/03-02:48:09	devenir male dominant	
12/03-02:48:09	la historia sin fin	<AR><Buenos Aires>
12/03-02:48:09	mario party 9	<DE><Enger>
12/03-02:48:09	gangster squad	<PT><Barcelos>
12/03-02:48:09	naruto	
12/03-02:48:09	desaparecidos fiesta loca	
12/03-02:48:09	secret life american vostfr	
12/03-02:48:10	pthc 12yo	
12/03-02:48:10	the mentalist s01e19	
12/03-02:48:10	ich mich nach deiner liebe soundtrack	
12/03-02:48:10	michael jackson bad man in the mirror	
28/02-01:25:02	pierre durand cancer	<FR><Talence>
28/02-01:25:14	college emile fournier de badonvillier	<FR><Talence>
3/03-18:50:29	julie fournier	<IT><Rovigo>
7/04-13:22:49	lilian moreno 06 17 79 18 35	<FR><Toulouse>

Ouverture

Préoccupation dès le début du projet

Motivation

- Fournir les données à la communauté scientifique
- Reproductibilité des résultats

Problèmes

- Satisfaire les exigences légales
- Ne pas divulguer d'informations personnelles

Trouver un compromis
entre richesse des données et anonymat

Anonymisation : procédure

Temps

- valeur relative plutôt qu'absolue

Adresses IP

- « seulement » 2^{32} possibilités
- fonction de hachage connue insuffisante
- anonymisation à la volée par des entiers
  - lenteur et demande en calculs
  - haut niveau d'anonymisation
  - usage ultérieur du jeu de données immédiat

Anonymisation : procédure (suite)

Requêtes

- distinguer le général du particulier (sensible)
  - peu de requêtes
  - ou beaucoup de requêtes du même utilisateur
  - seuil de 50 IP distinctes
- nombres : téléphone et cartes de crédit, mais aussi âges

Disponibilité

Complètement accessibles

- méta données
- échantillons
- outil (algorithme de détection)
- format standard, fichier texte formaté

Sur demande

- totalité des jeux de données

Résultats

Mise au point d'un outil de détection de requêtes pédophiles

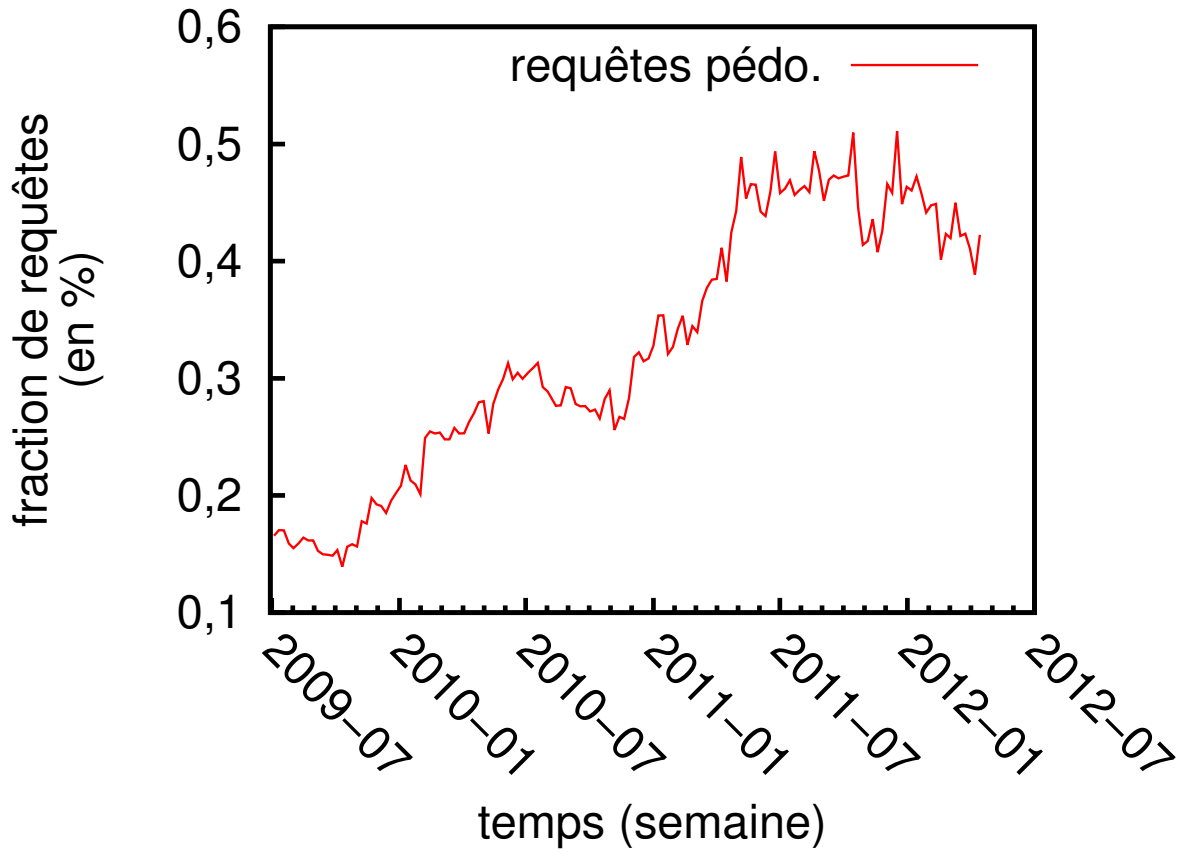
- conçu en collaboration avec forces de l'ordre
- validation
- connaissance des taux d'erreurs (FP/FN)
  - précision 98,6 %
  - rappel 76%

Statistiques globales

- eDonkey [2]
  - environ 2,5 requêtes pédophiles pour 1 000 (en 2009)
  - 1 requête pédophile toutes les 33 secondes environ
  - environ 2,2 utilisateurs pour 1 000 sont pédophiles
- environ 2 fois moins sur KAD

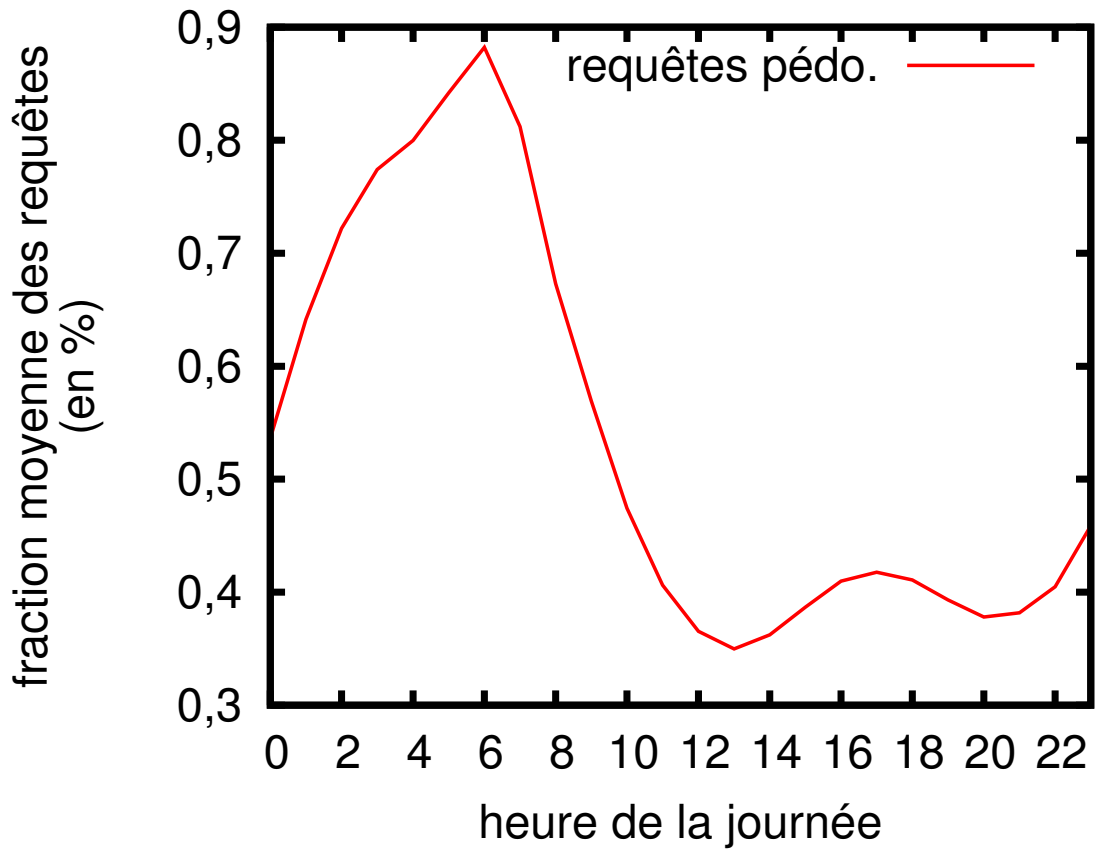
M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, 2012.

Évolution temporelle



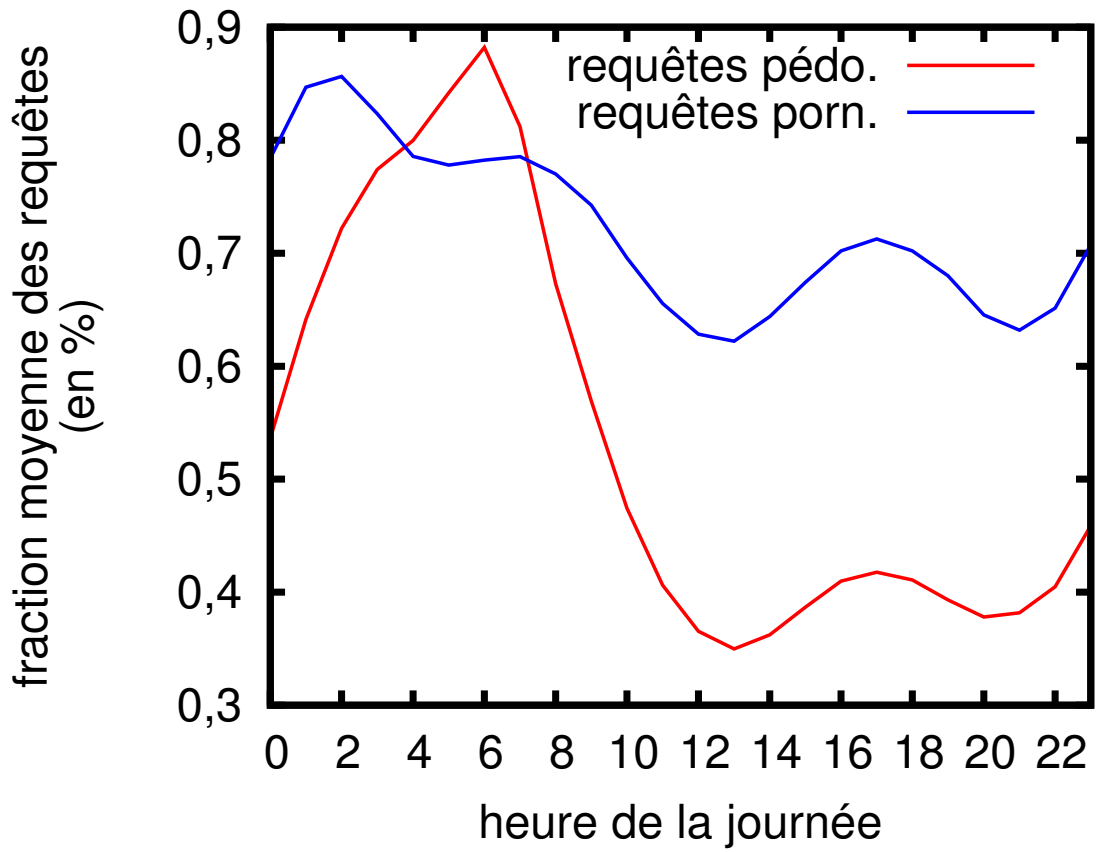
- Trafic global stable sur 3 ans
- Trafic pédophile en forte croissance

Intégration sociale



- Effet jour/nuit
- Pic de fraction de requêtes pédophiles vers 6 heures
- Différent pour les requêtes pornographiques

Intégration sociale



- Effet jour/nuit
- Pic de fraction de requêtes pédophiles vers 6 heures
- Différent pour les requêtes pornographiques

Ouverture des données

Puissance de ce type d'analyse (*Transaction-Log Analysis*)

- avancées significatives
- contribution qualitative avec une approche quantitative
- utile dans de nombreux autres contextes

Mais :

- responsabilité / éthique
  - Google Flu ([1])
  - Étude sur Twitter ([2])
- erreurs
  - AOL
  - notre outil


J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.


A. Sadilek, H. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. *AAAI Conference on Artificial Intelligence*, 2012.

Merci.

raphael.fournier@lip6.fr

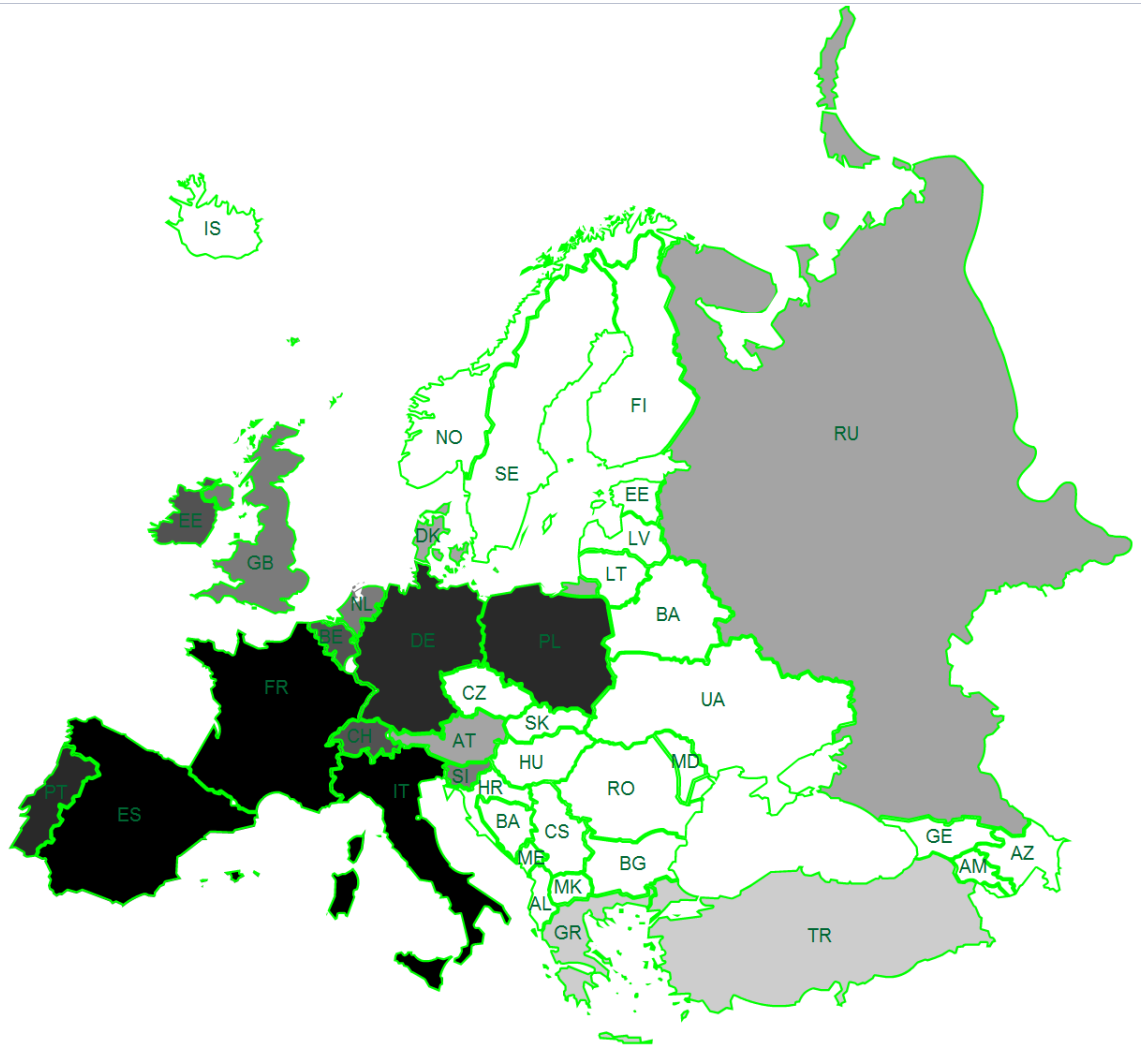
Géolocalisation

pays	# requêtes	# pédo.	fraction
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 %

Problèmes :

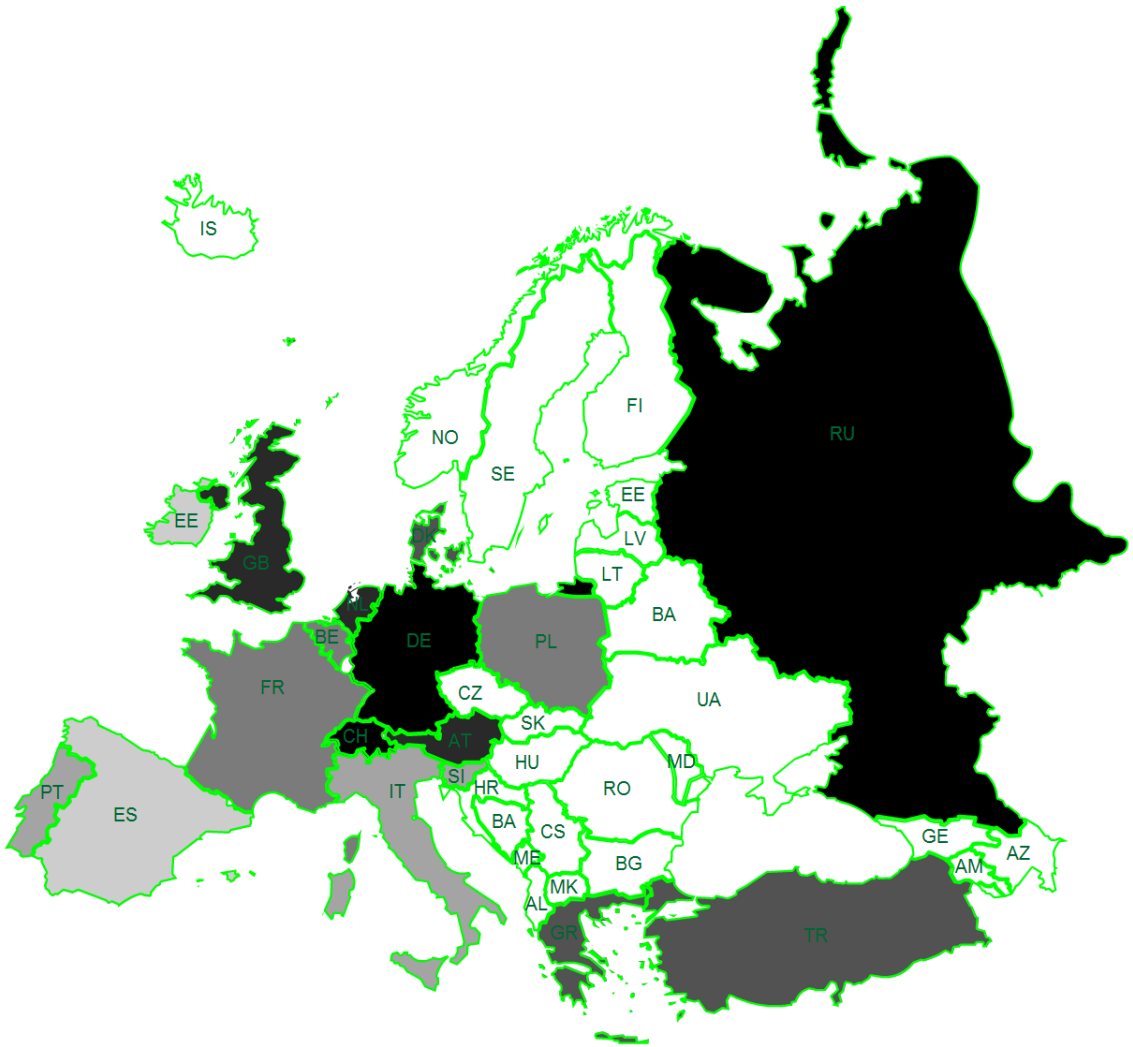
- langues
- encodage
- VPN

Géolocalisation



total des requêtes

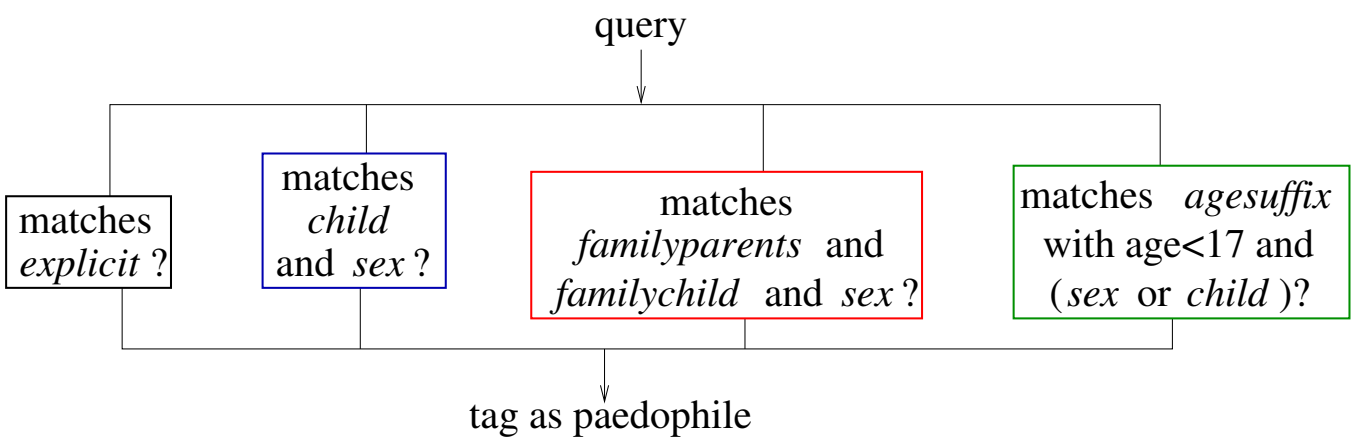
Géolocalisation



fraction de requêtes pédophiles

Tool design

- 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

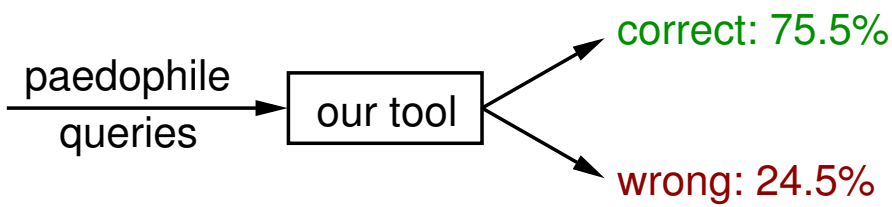
- set of 21 volunteering experts (Europol, national authorities, NGOs)
- set of 3,000 randomly selected queries:
  - paedophile
  - not paedophile
  - *neighbours* (submitted within the 2 previous or next hours of a paedophile query by the same user)
- tag queries as *paedophile*, *probably paedophile*,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

pédo	*prob. pédo*	*je ne sais pas*	*prob. pas*	*pas pédo*	total	pertinence
...	...	...	...	...	...	...
1174	111	20	64	789	2158	99.1
...	...	...	...	...	...	...

Assessment results

Limited filter precision

- False negatives
- False positives



Notion d'utilisateur

Approximation possible :
utilisateur $\sim$ adresse IP

Problèmes

- Traduction d'adresse (NAT)
- Renouvellement d'adresses
- Plusieurs utilisateurs par ordinateur
- Plusieurs ordinateurs par utilisateur

Notion d'utilisateur

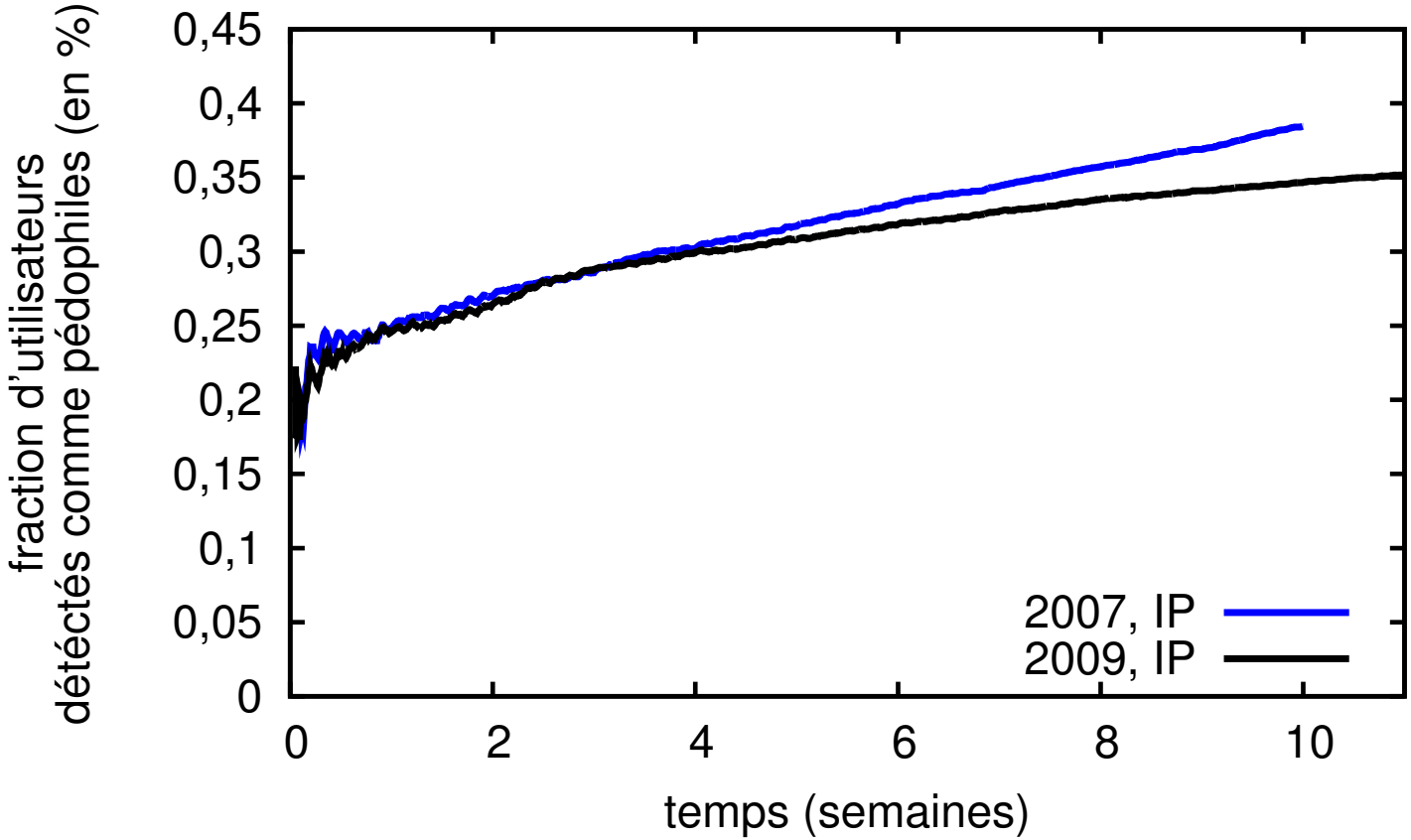
Utilisateur pédophile

- Un utilisateur est pédophile s'il a fait une requête pédophile
- Pollution : toutes les adresses IP vues comme pédophiles, après *un certain temps*

3 approches :

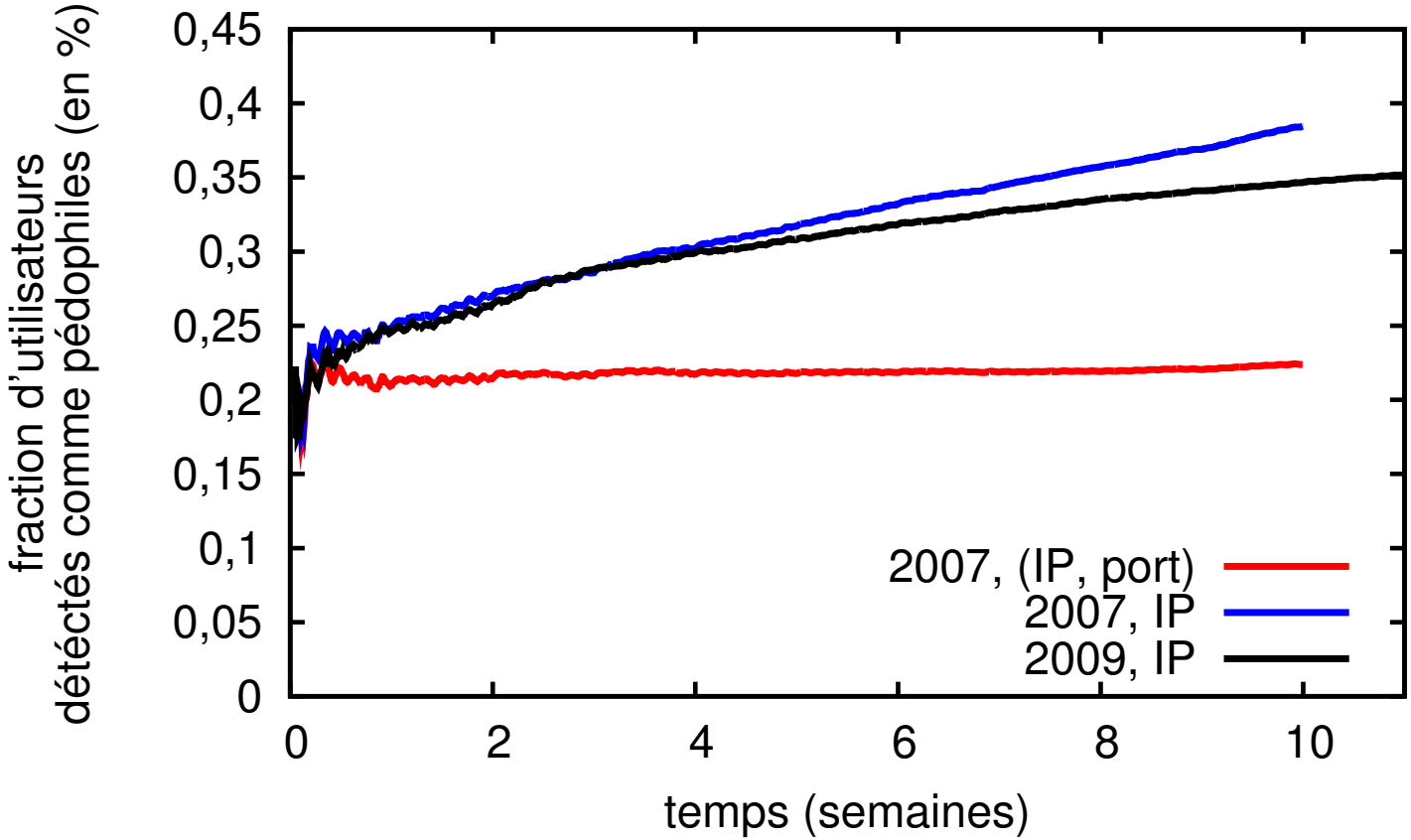
- utilisateur $\sim$ adresse IP + port de connexion
- sessions temporelles
- durée de la mesure

Notion d'utilisateur : IP *vs* (IP,port)



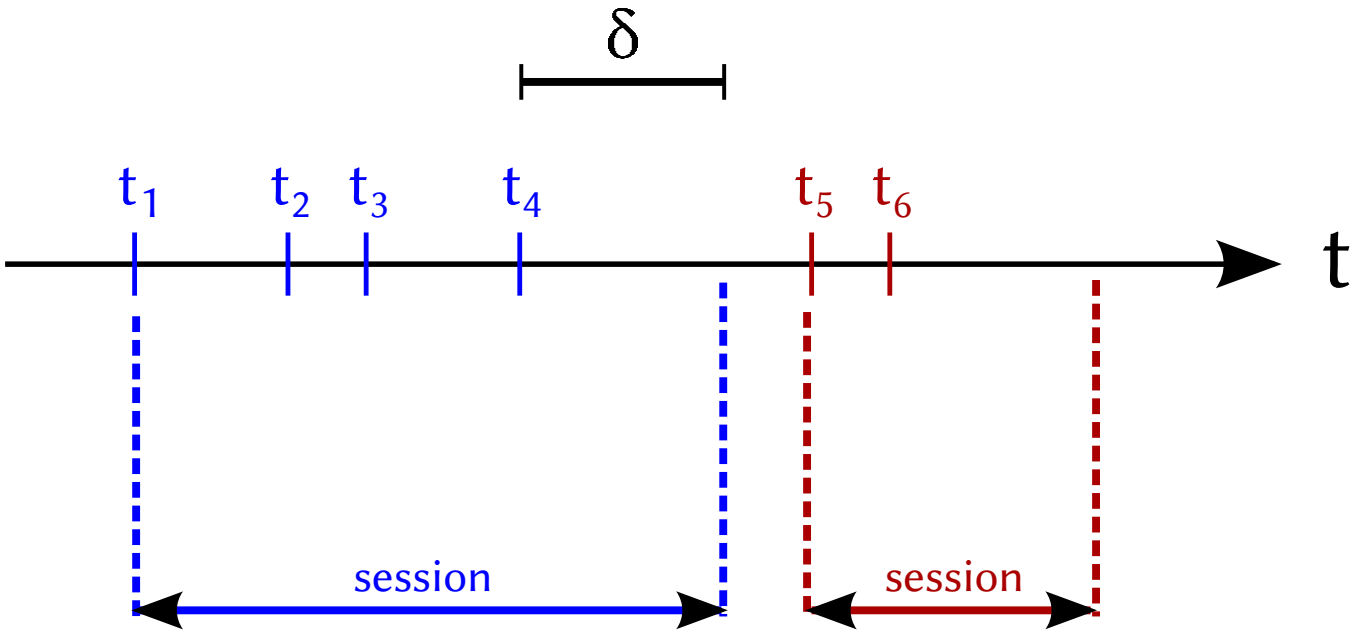
- (IP, port) permet d'éviter la pollution

Notion d'utilisateur : IP *vs* (IP,port)

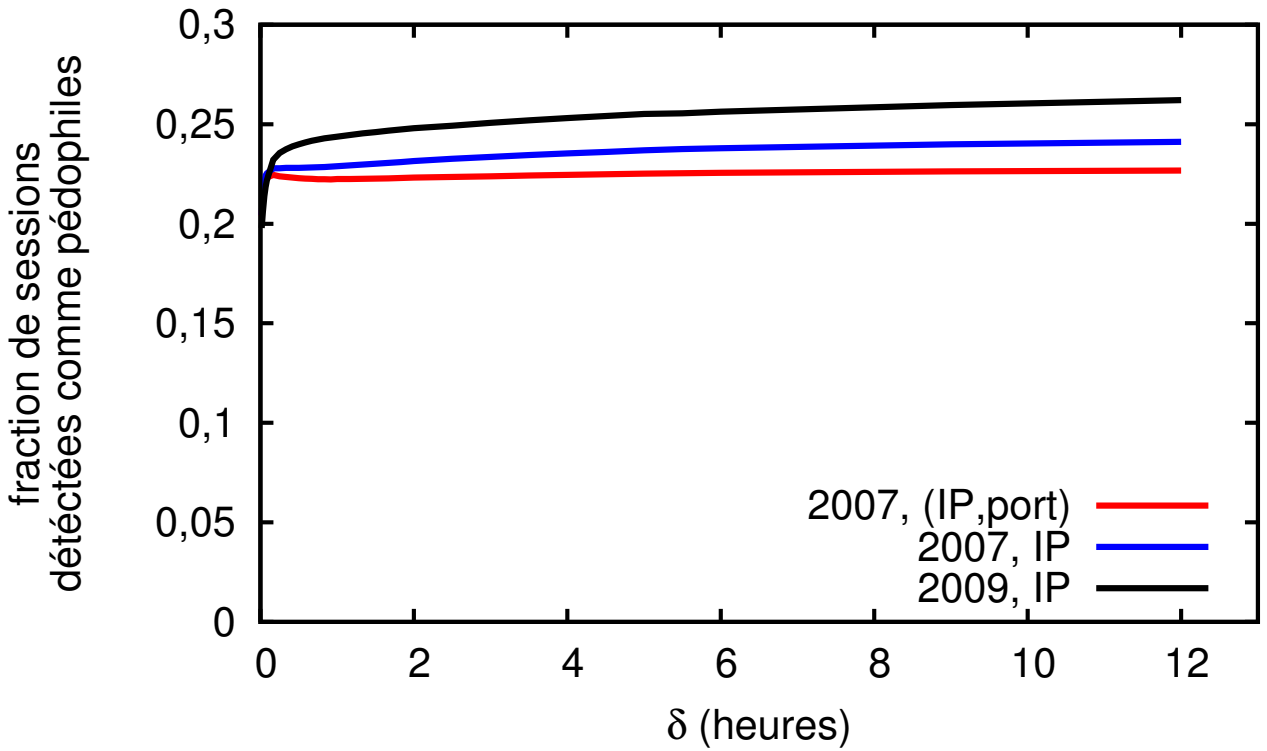


- (IP, port) permet d'éviter la pollution

Notion d'utilisateur : sessions temporelles



Notion d'utilisateur : sessions temporelles



Fraction d'utilisateurs pédophiles

- faux positifs et négatifs sur les utilisateurs

$$p(u \in U^+ \mid u \in V(n,0)) = 1 - (1 - f'^-)^n$$

$$p(u \in U^- \mid u \in V(n,k)) = (f'^+)^k (1 - f'^-)^{n-k}$$

  - U^+, U^- : ensemble des utilisateurs pédophiles/non pédophiles
  - V^+, V^- : ensemble des utilisateurs détectés comme pédophiles/non pédophiles
  - n : nombre de requêtes d'un utilisateur
  - k : nombre de requêtes détectées comme pédophiles

- $\frac{|U^+ \cap V^+|}{|D|} = \sum_{n=1}^{N} \sum_{k=1}^{n} (1 - (f'^+)^k (1 - f'^-)^{n-k}) \frac{|V(n,k)|}{|D|}$

Fraction d'utilisateurs pédophiles

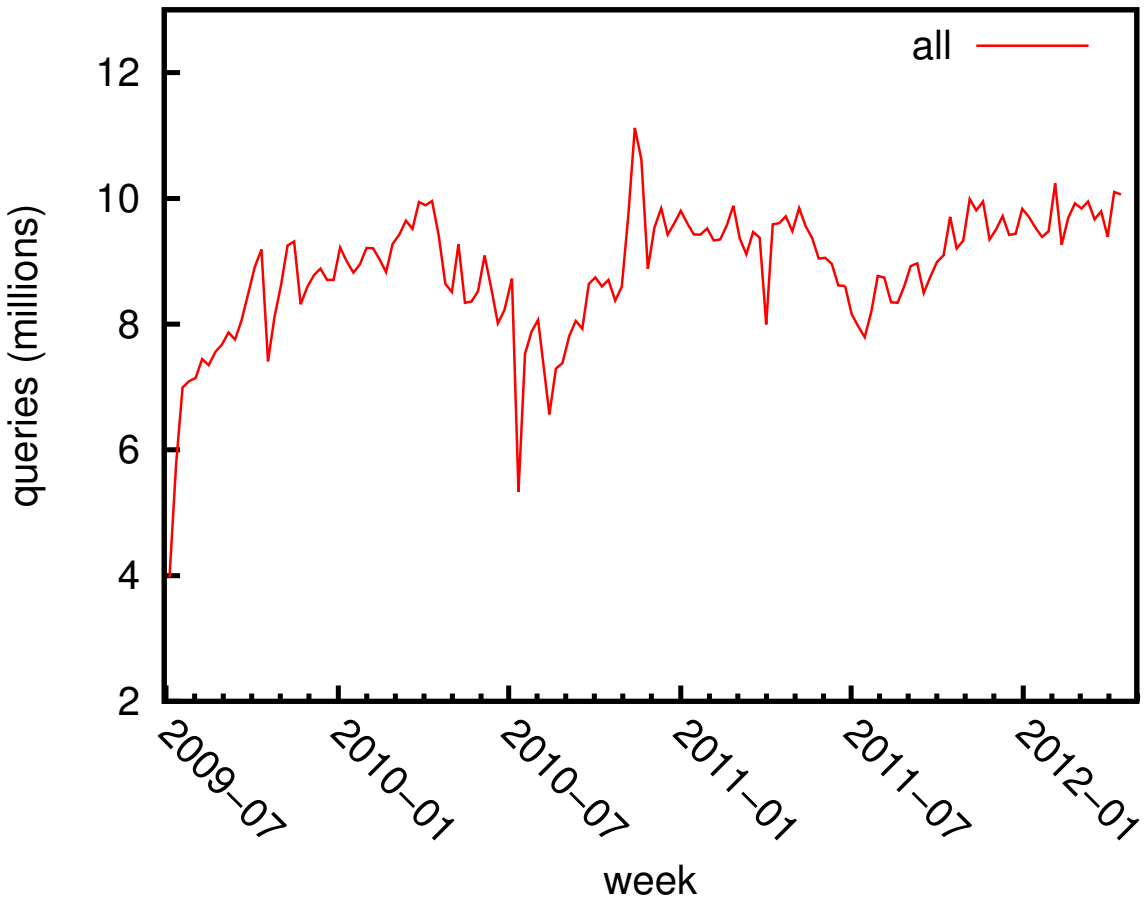
Résultat

- Fraction d'utilisateurs pédophiles proche de 0,22% [2007]
- 1 utilisateur pédophile sur 450 environ

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.

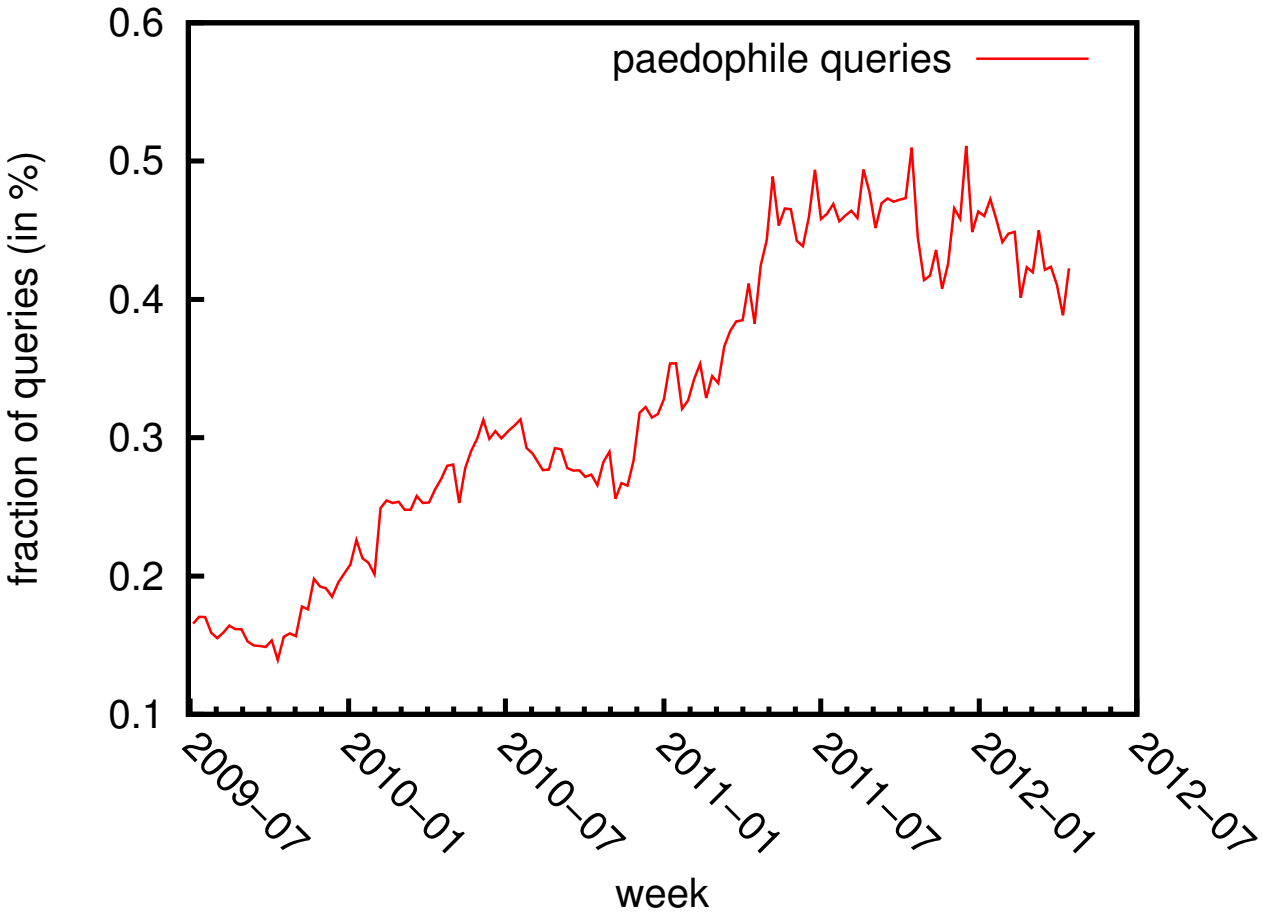
Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, In press, 2012.

Global traffic on server



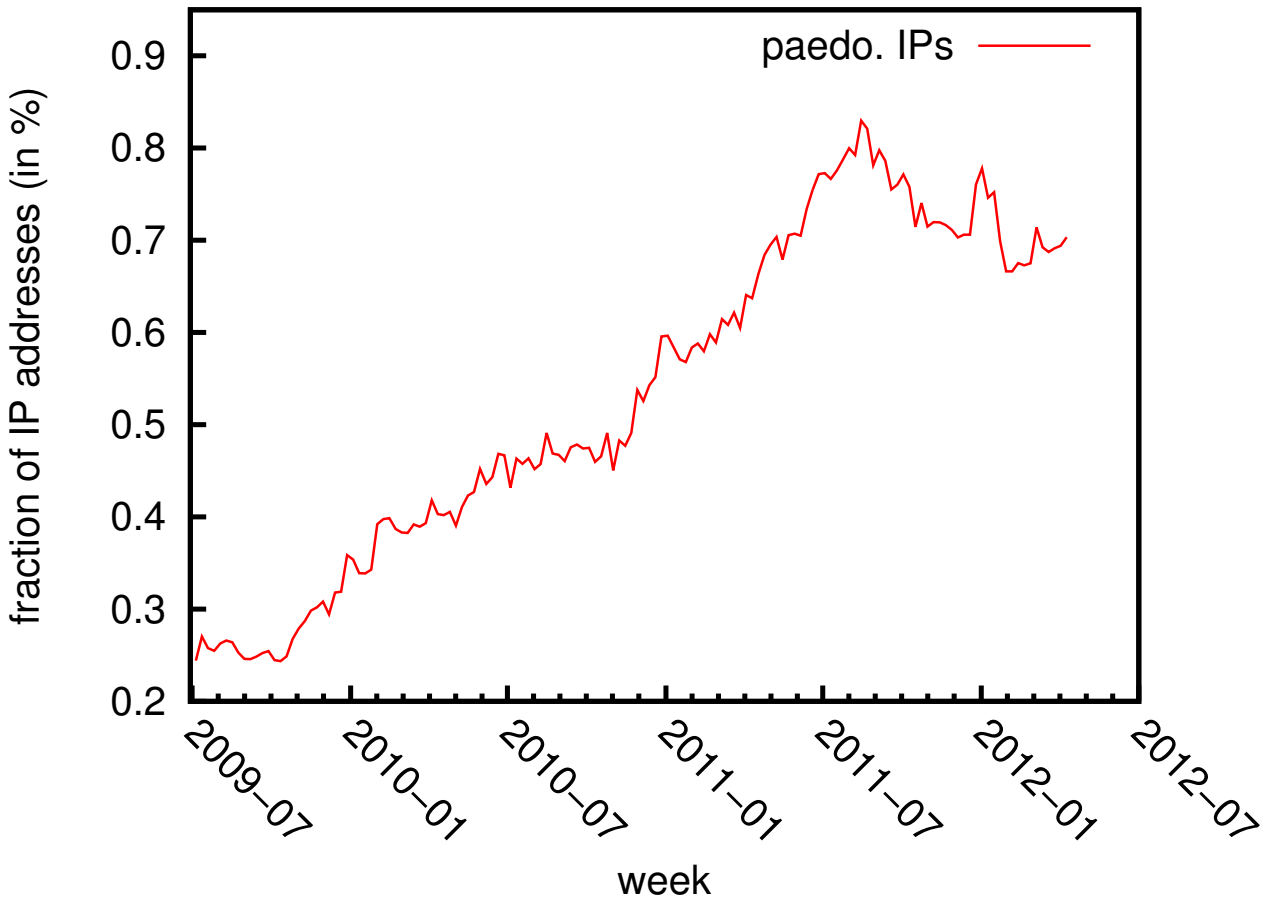
- Stability of global traffic over 3 years

Fraction of paedophile queries



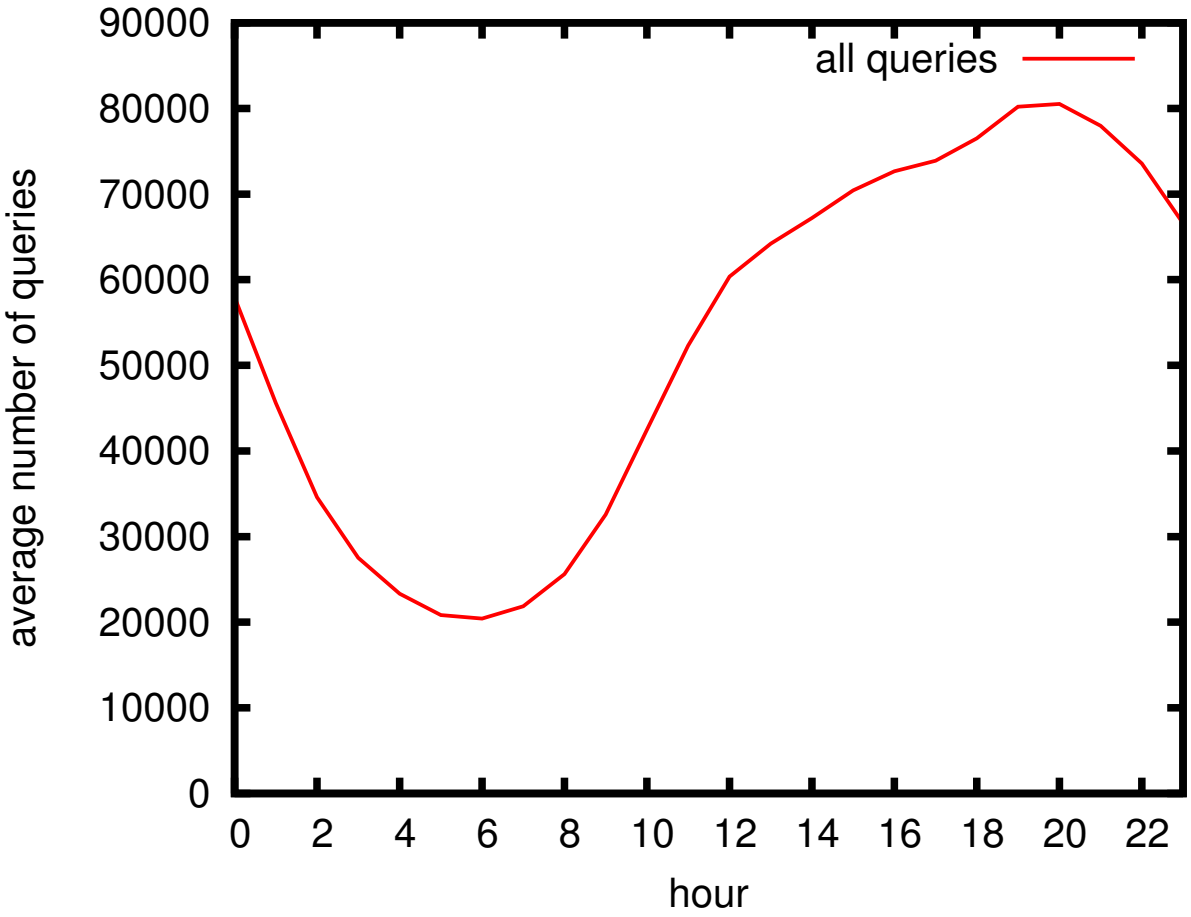
- Fraction of paedophile queries strongly increasing

Fraction of paedophile users



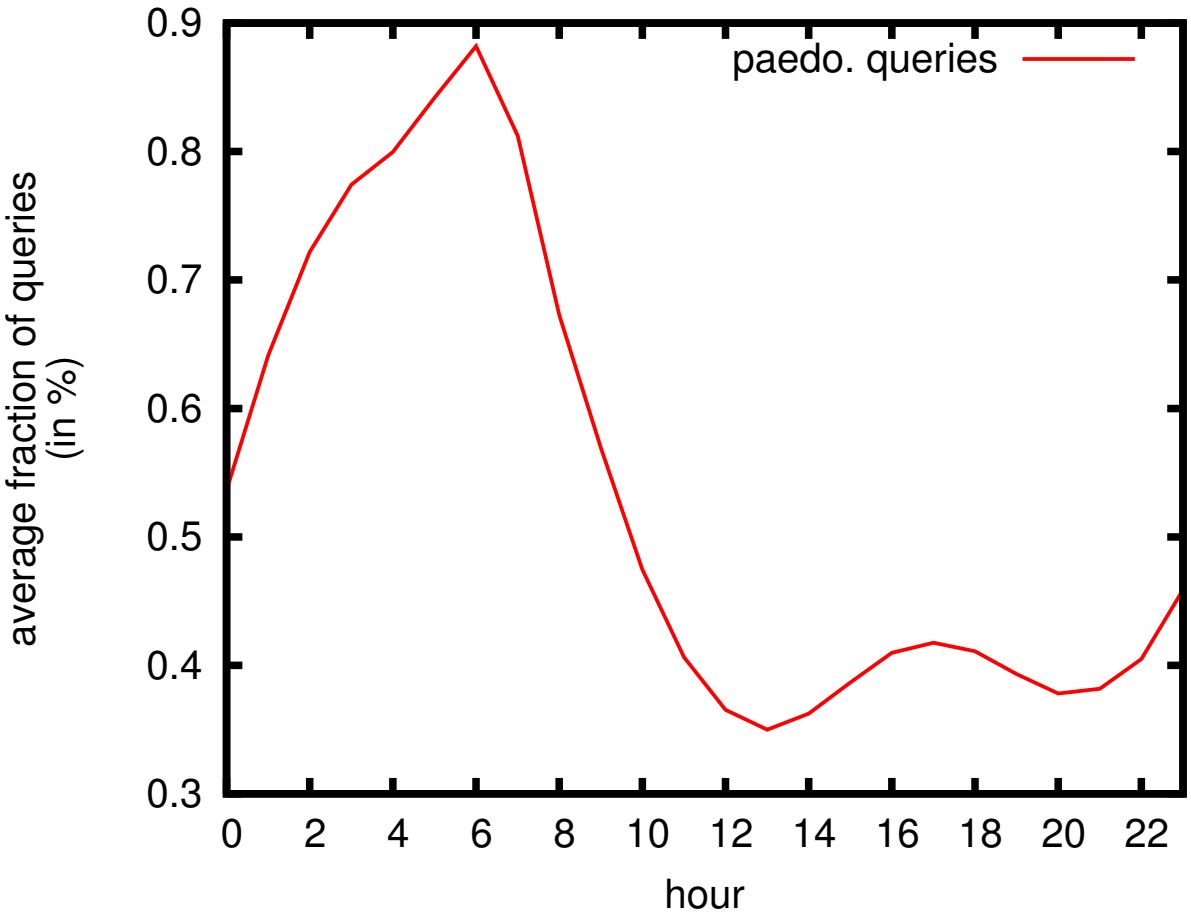
- Fraction of paedophile users also increasing

Daily traffic



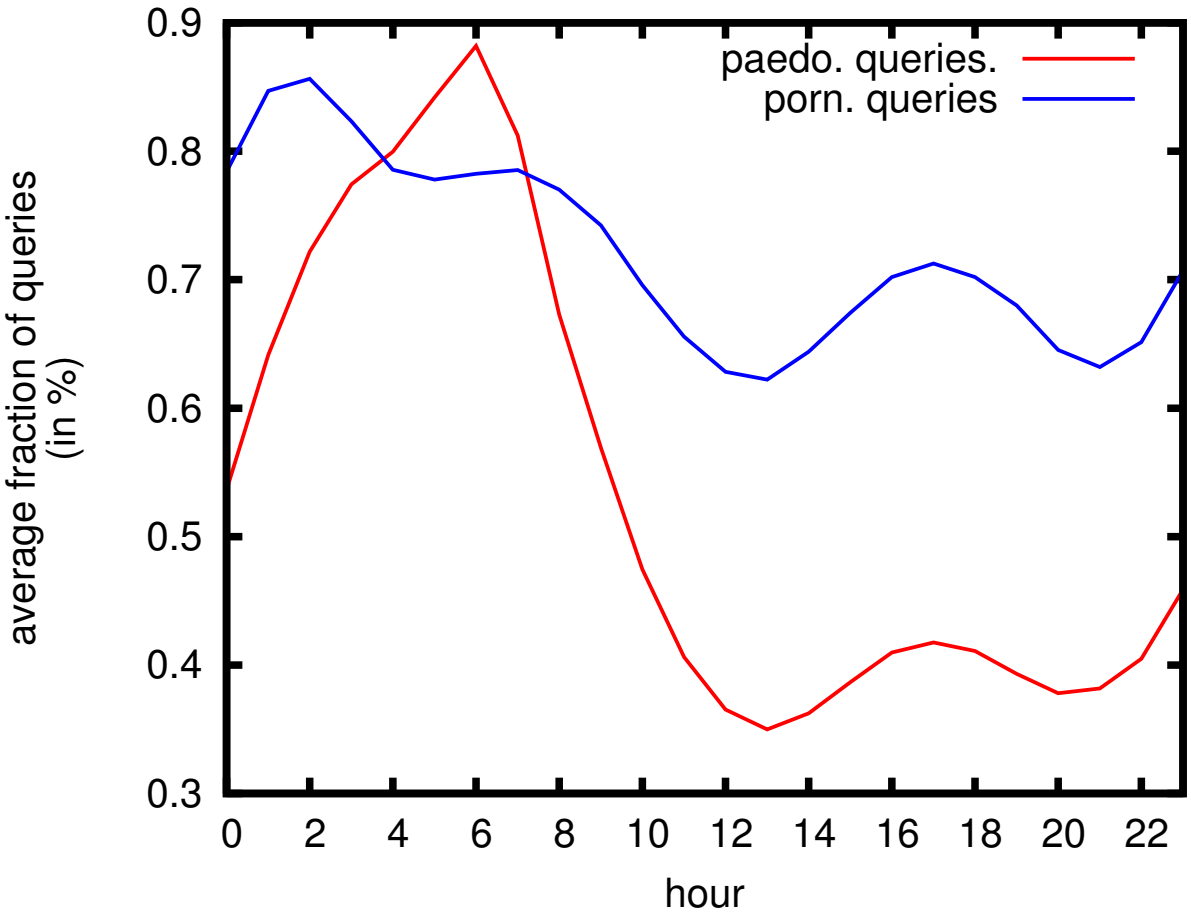
- Circadian cycle (day/night effect)

Fraction of paedophile activity



- Fraction of paedophile queries peaks at 6 AM

Pornography vs paedophile activity



- Paedopornagraphy 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