Détection et analyse d'une thématique rare dans de grands ensembles de requêtes : l'activité pédophile dans le P2P

Raphaël Fournier-S'niehotta

Soutenance de thèse
Sous la direction de Matthieu Latapy

21 décembre 2012

Contexte

Étude des ensembles de requêtes

- Des applications :
  - classiques (amélioration de systèmes)
  - moins classiques (suivi de la grippe)

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 de la thèse

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

- Quantifier les requêtes
- Quantifier les utilisateurs
- Étudier l'évolution de l'activité

Méthodologie générale
Thématique rare

Problématiques

- Collecte de données

  $\rightarrow$ taille, dynamicité, protocoles peu documentés

- Outil de détection automatique

  $\rightarrow$ peu d'experts, pas d'ensemble de référence

- Inférence statistique rigoureuse

  $\rightarrow$ faible quantité de requêtes pédophiles

- Identification des utilisateurs

  $\rightarrow$ information partielle, peu fiable

Données

- Requêtes envoyées au moteur de recherche d'eDonkey
- 2 collectes en continu :

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

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

- Normalisation et anonymisation des données brutes

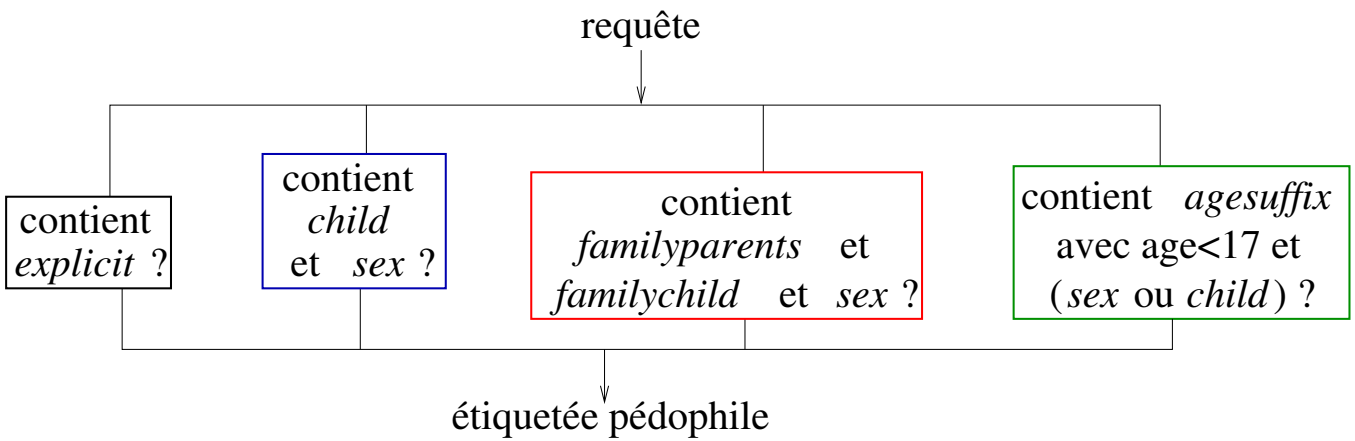
Séquence de requêtes : $q_i = (t, u, k_1, k_2, \ldots, k_n)$

- t horodatage
- u information sur l'utilisateur (adresse IP, port, géolocalisation)
- $(k_1, k_2, \ldots, k_n)$ suite de mots-clefs

Plan

Conception de l'outil

- Listes de mots-clefs élaborées avec les forces de l'ordre
- Inspection des ensembles de requêtes
- Amélioration des tests
- 4 types de requêtes pédophiles



raygold little girl porno infantil incest mom son video 12yo fuck video

Évaluation de la qualité

Faux positifs

“sexy daddy destinys child”
contient “sexy”, “daddy” et “child” $\longrightarrow$ étiquetée pédophile
probablement une recherche liée à la musique

Faux négatifs

“pjk 12yo” $\longrightarrow$ étiquetée non pédophile
contient un marqueur pédophile non reconnu

Comment estimer ces taux de faux positifs et faux négatifs ?

Validation – Sondage

- 21 experts volontaires (Europol, forces de l'ordre, ONG)
- 3 000 requêtes choisies aléatoirement parmi :
  - celles étiquetées pédophiles (1 000)
  - celles étiquetées non pédophiles (1 000)
  - les *voisines* (1 000)
    - soumises dans les 2h avant ou après une requête étiquetée pédophile, par la même adresse IP
- Réponses possibles : *pédophile*, *probablement pédophile*, *probablement non pédophile*, *non pédophile* ou *je ne sais pas*

Validation – Résultats du sondage

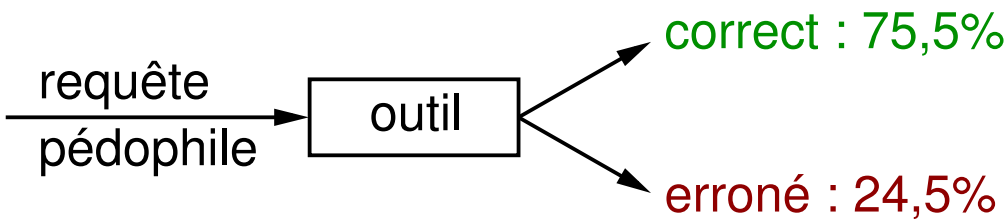
paedo	*prob. paedo*	*don't know*	*prob. not*	*not paedo*	total	pertinence
1530	149	25	66	1230	3000	99.5
1381	247	125	580	667	3000	98.5
1679	89	2	113	1117	3000	99.1
1603	201	99	174	923	3000	99.0
1598	5	15	1	1381	3000	98.8
128	81	1	26	124	360	100.0
216	154	0	142	132	644	98.4
1624	126	16	165	581	2512	99.8
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

- Pertinence : connaissance appropriée du contexte

Résultats de la validation

Précision

Rappel

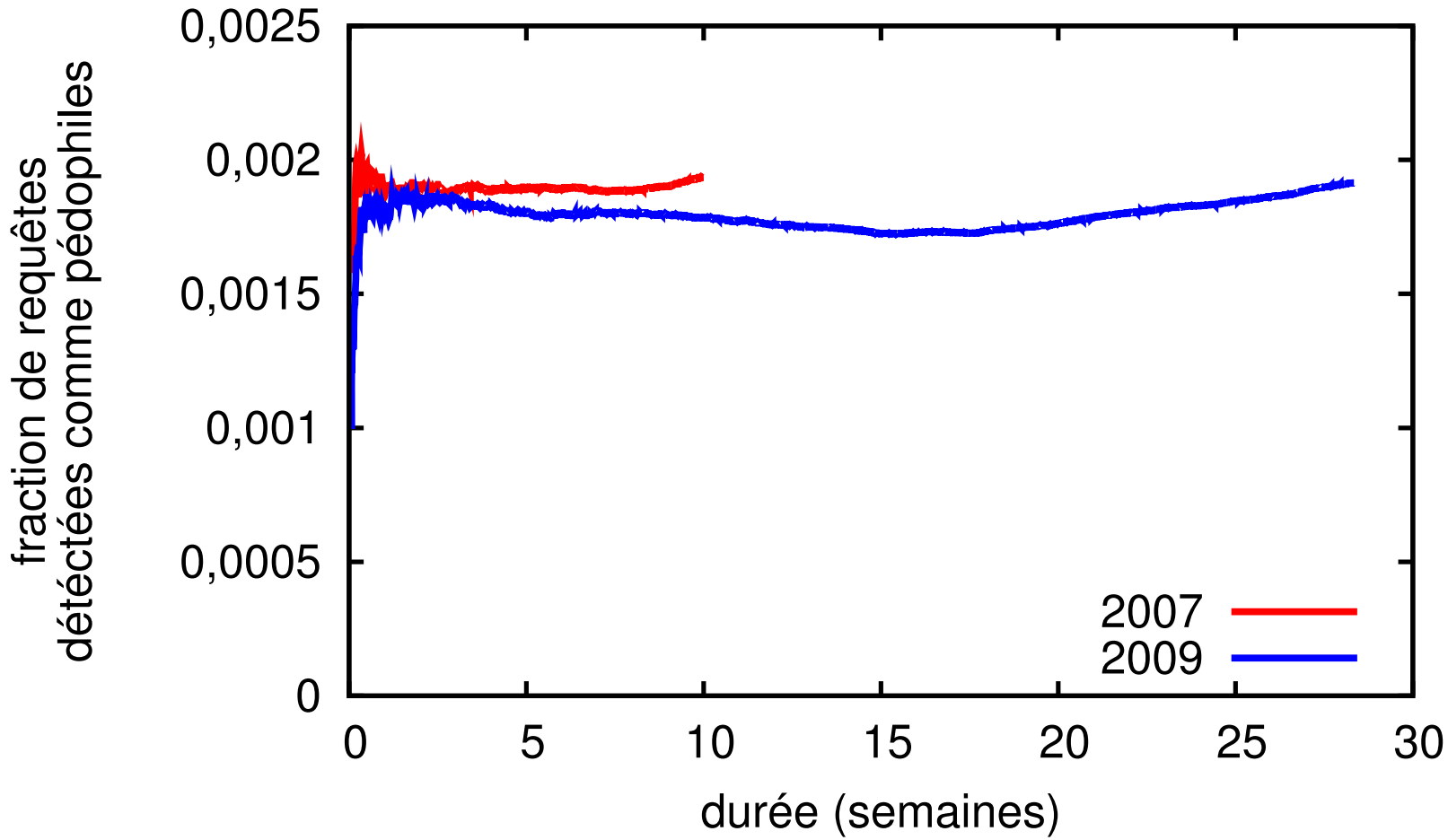


Résultats de la validation

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

- P^+ : nombre de requêtes pédophiles
- T^+ : nombre de requêtes étiquetées pédophiles
- f'^+ : taux de faux positifs
- f^- : taux de faux négatifs

Fraction de requêtes détectées comme pédophiles



Fraction de requêtes détectées comme pédophiles

Résultat

- Détection : légèrement au-dessus de 1,9 pour 1 000
- Après correction : 2,5 requêtes pour 1 000 sont pédophiles
  - 1 requête pédophile toutes les 33 secondes 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.

Plan

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

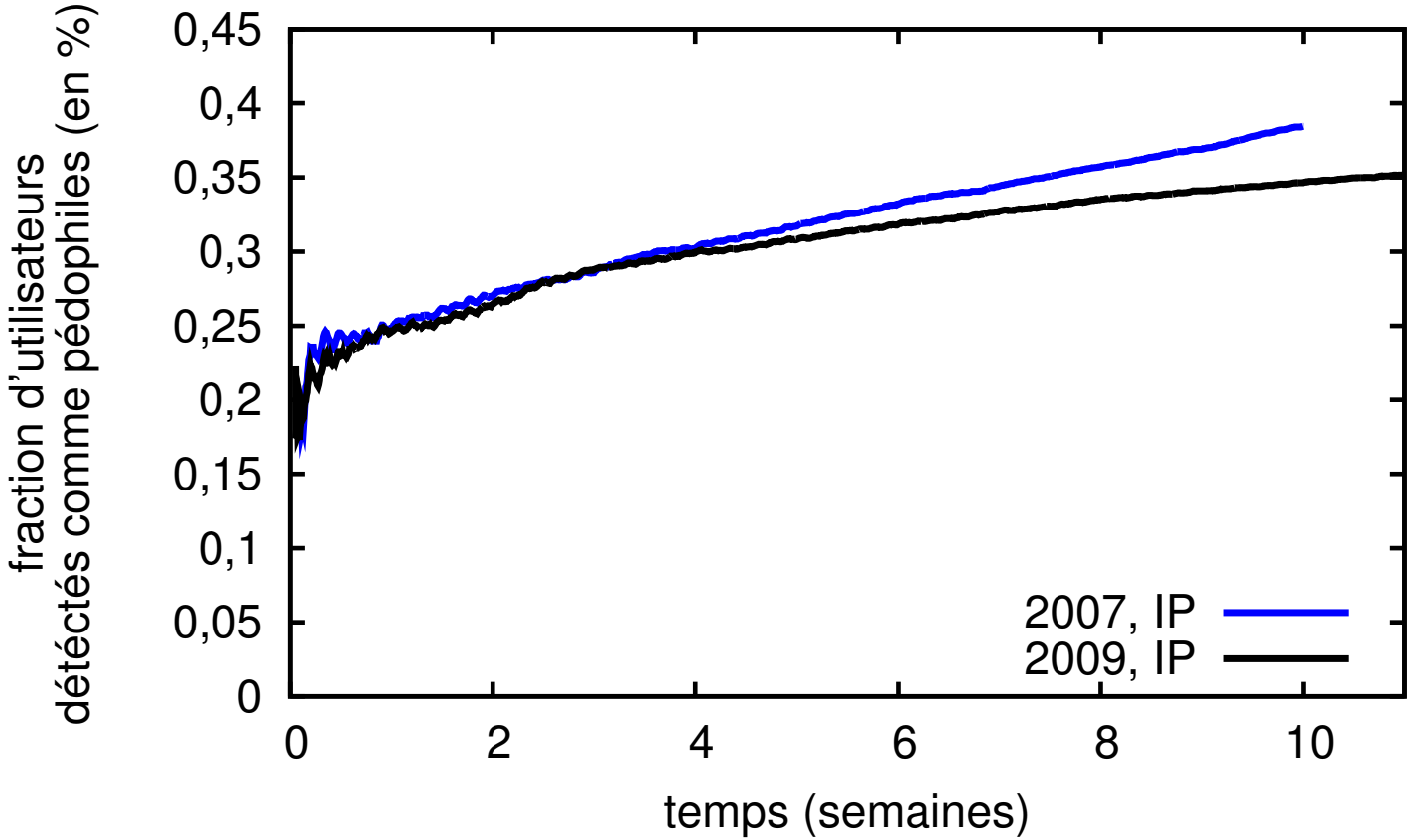
Utilisateur pédophile

- 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 :

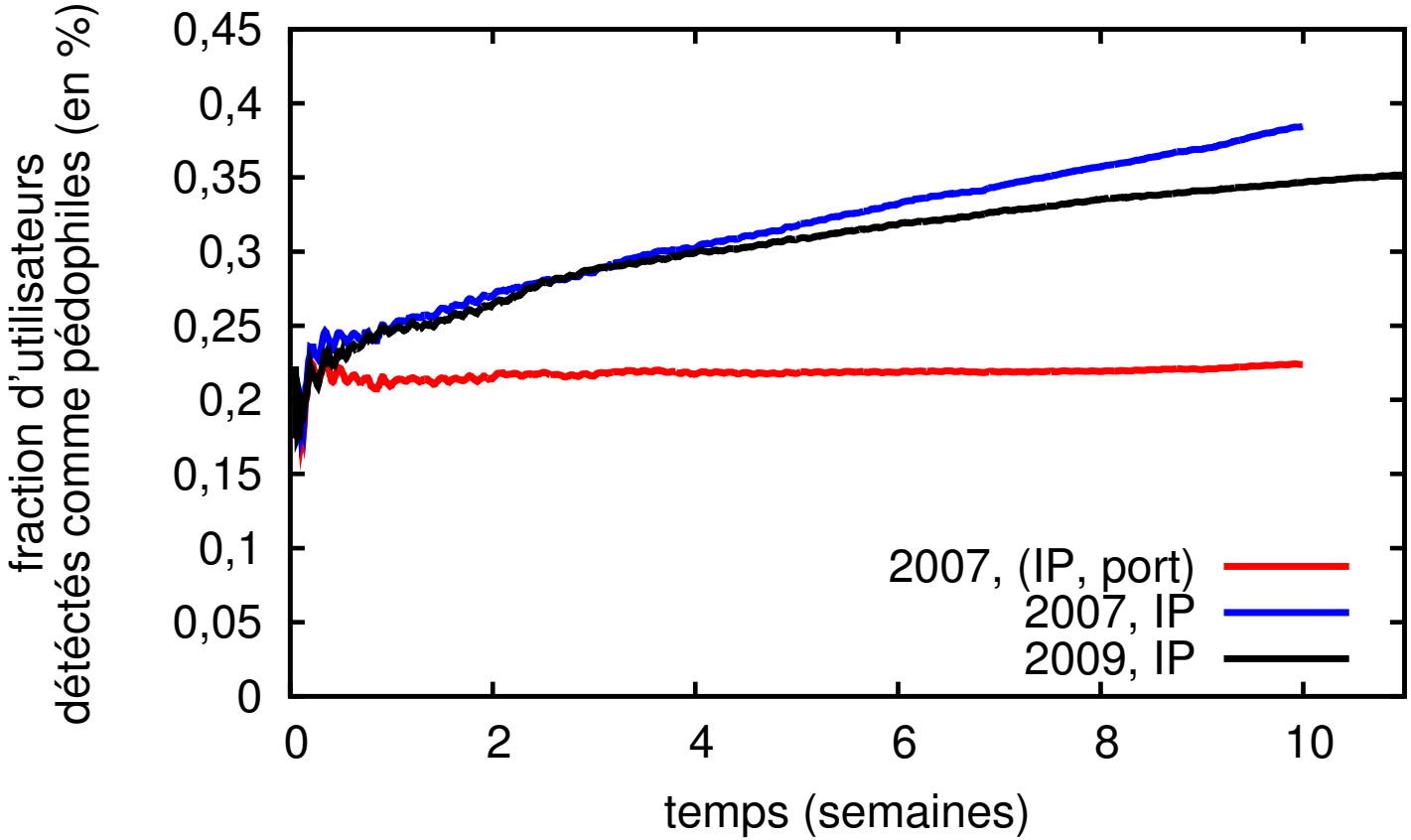
- utilisateur $\sim$ adresse IP + port de connexion
- sessions temporelles
- durée de la mesure

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



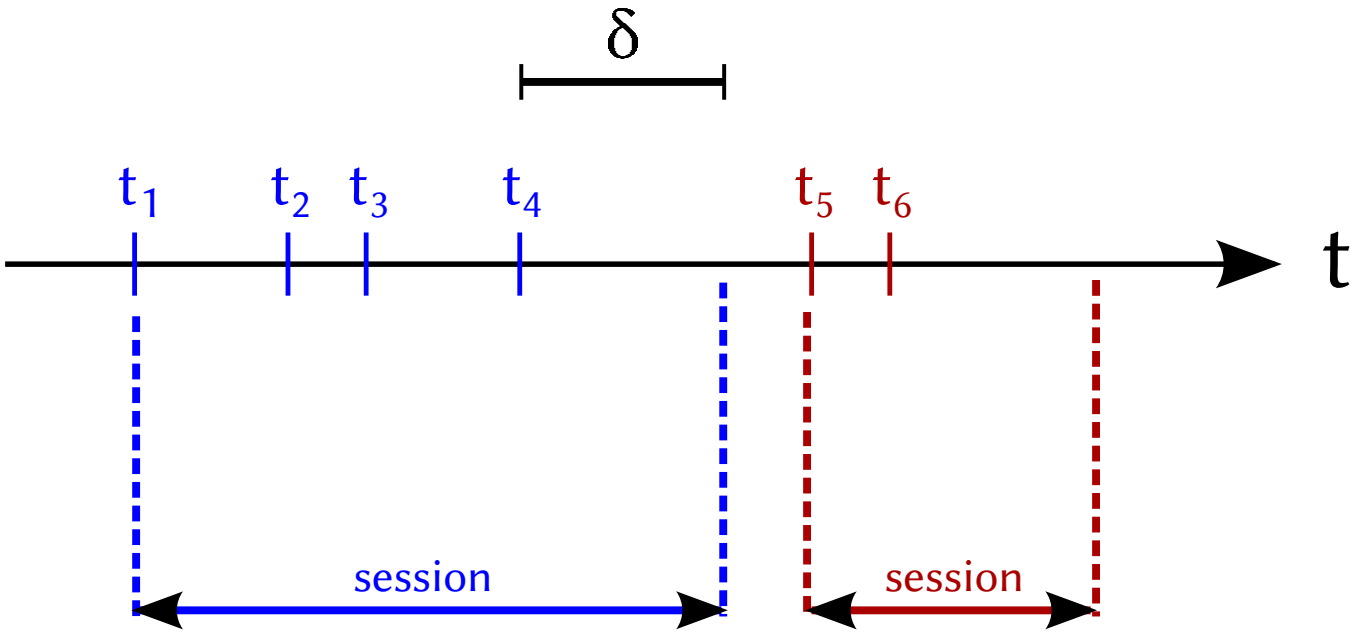
- (IP, port) permet d'éviter la pollution

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

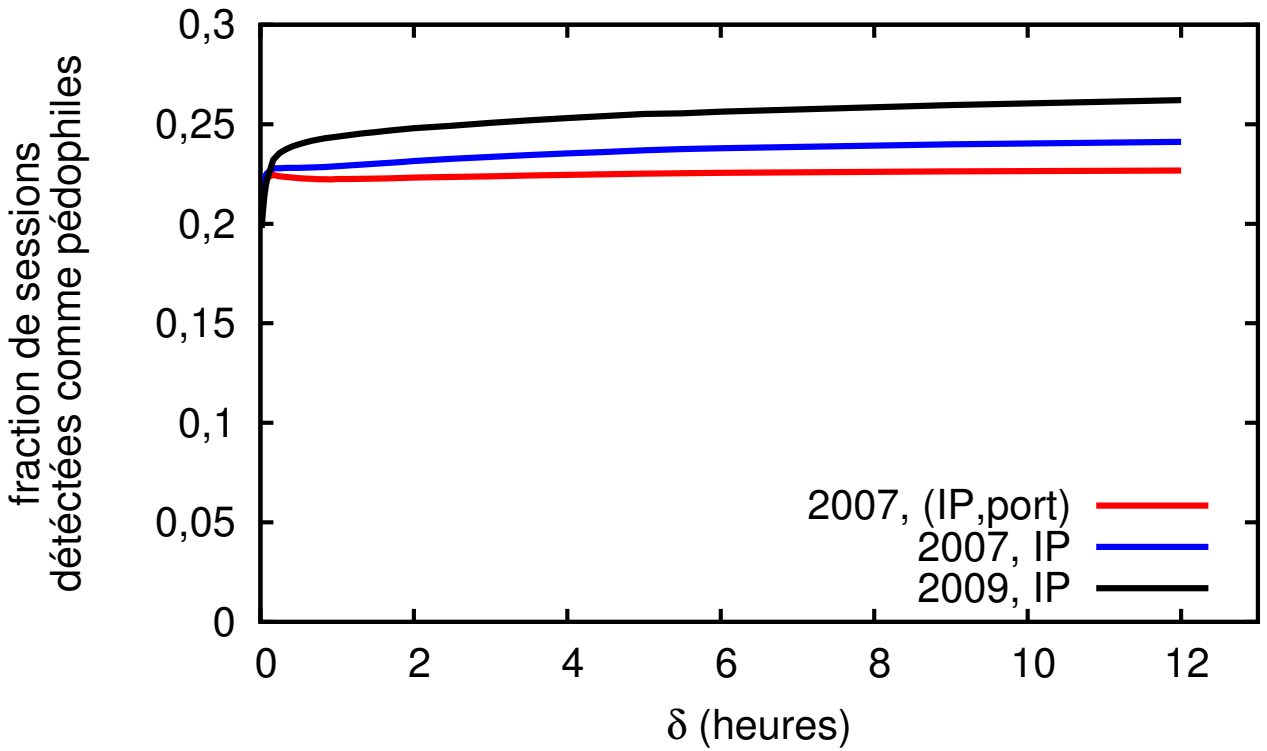


- (IP, port) permet d'éviter la pollution

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

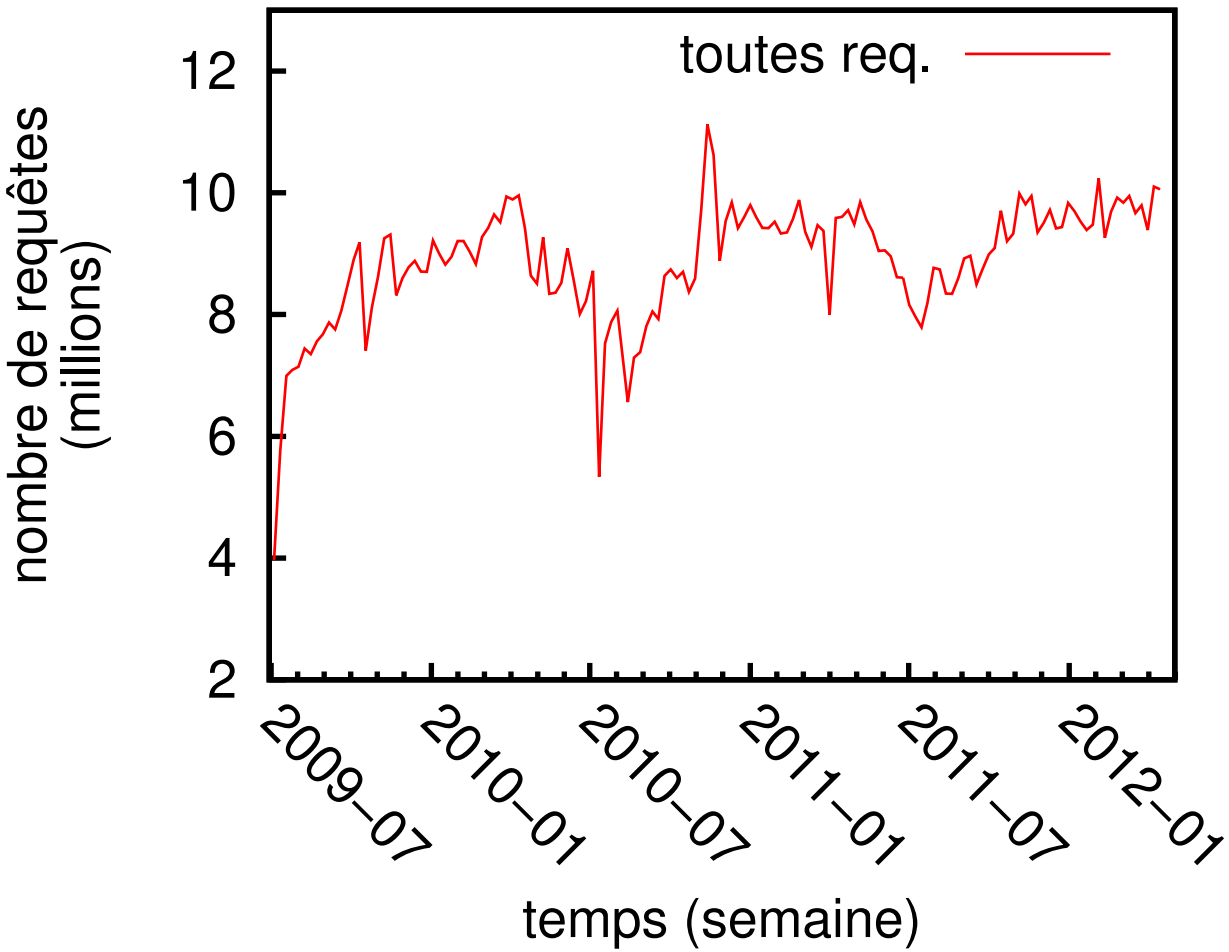
- Fraction d'utilisateurs pédophiles proche de 0,22% [2007]
- 1 utilisateur pédophile sur 450 environ

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.

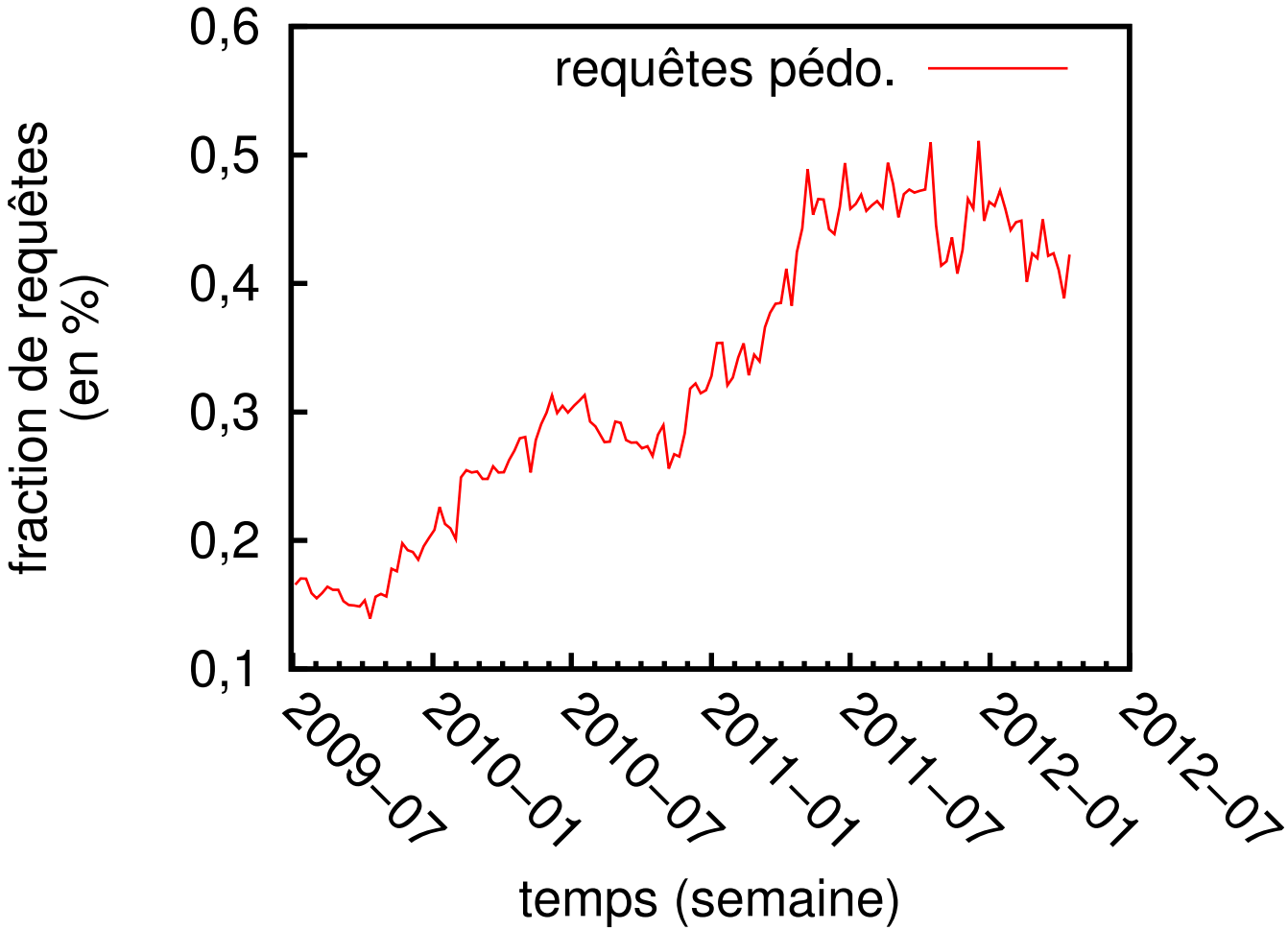
Plan

Évolution sur une longue période



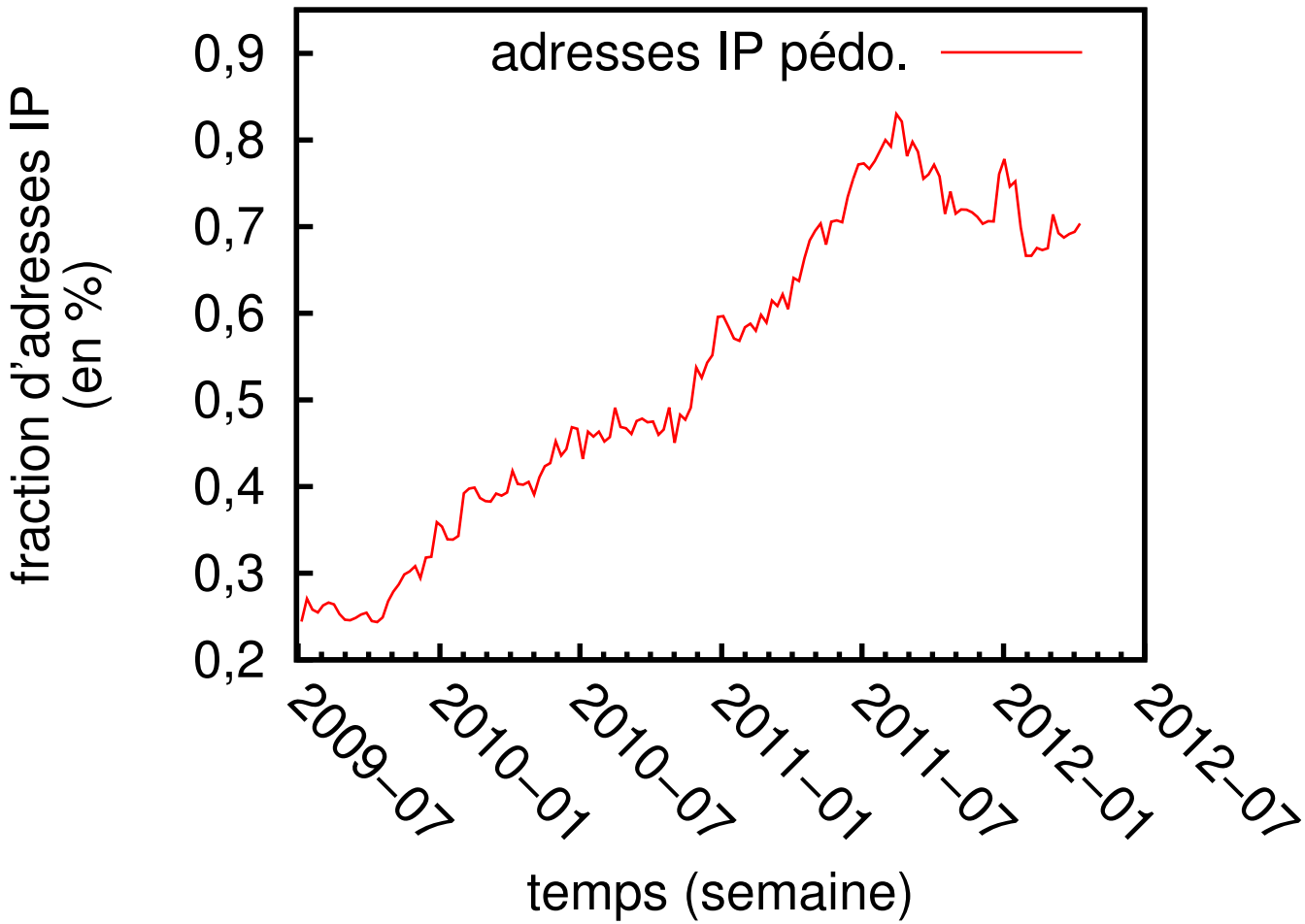
- Trafic global stable sur 3 ans

Évolution sur une longue période



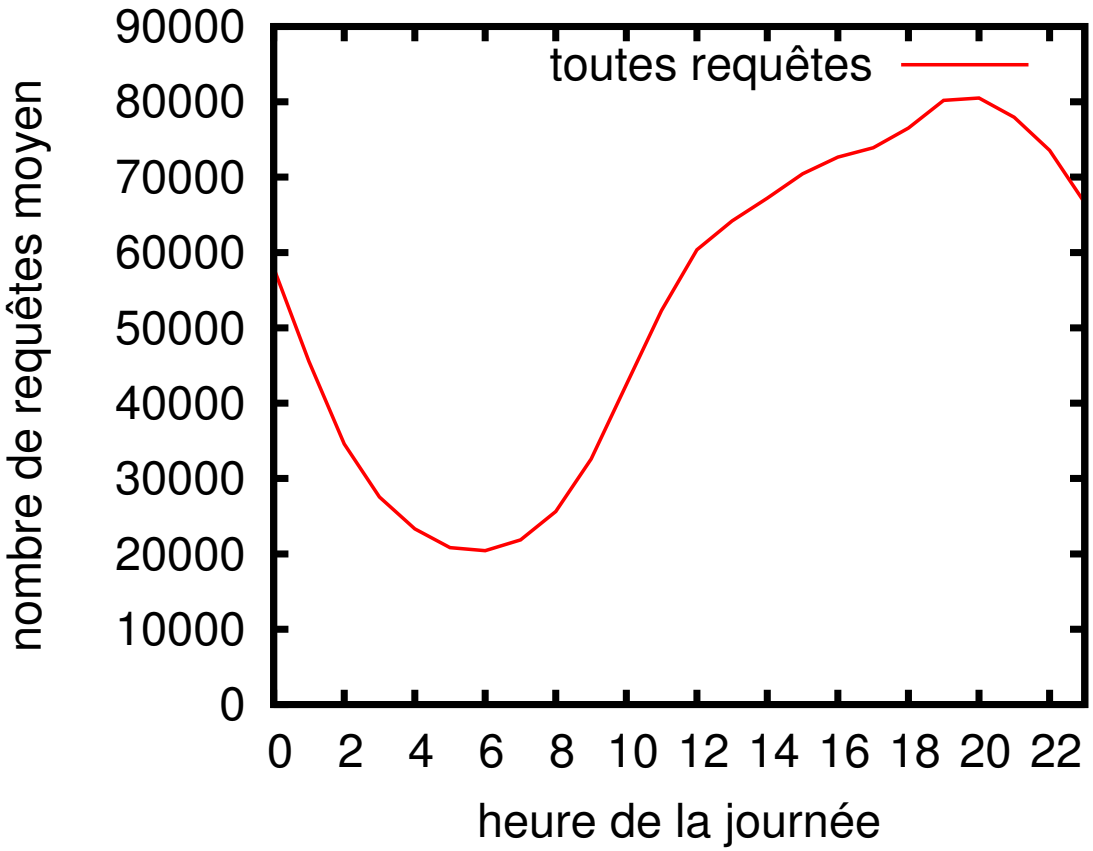
- Trafic pédophile en forte croissance

Évolution sur une longue période



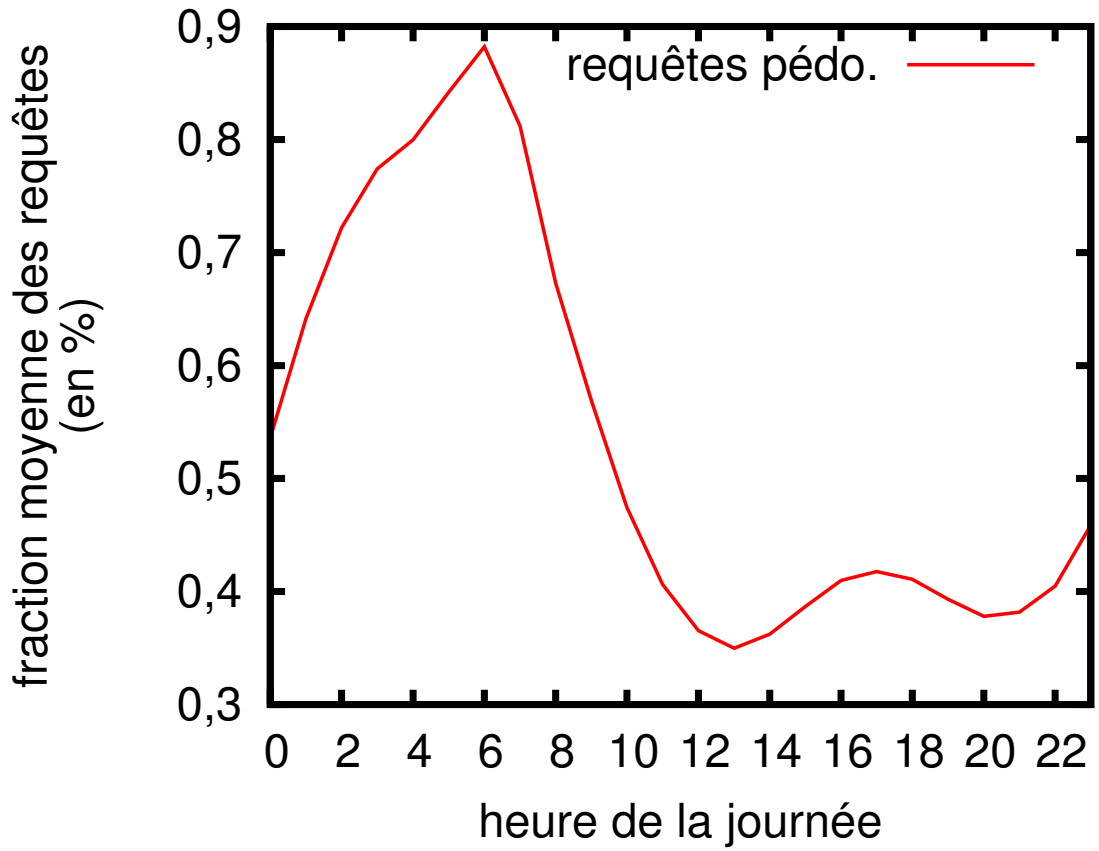
- Augmentation du nombre d'utilisateurs pédophiles

Dynamique journalière



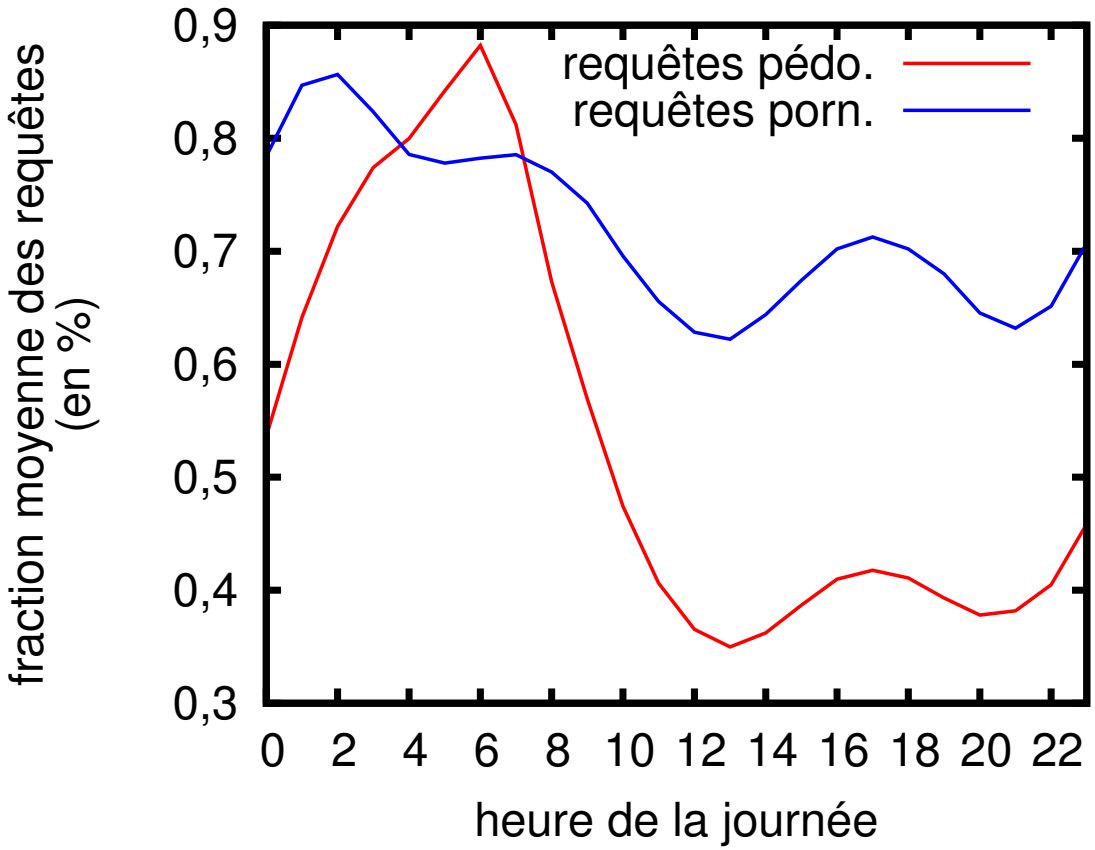
- Effet jour/nuit du trafic

Dynamique journalière



- Pic de fraction de requêtes pédophiles vers 6 heures

Dynamique journalière



- Différent pour les requêtes pornographiques

Évolution de l'activité

Résultat

- Augmentation importante entre 2009 et 2012
- Pic de requêtes autour de 6 heures du matin
- Contribution sur un aspect qualitatif avec une approche quantitative

Plan

Conclusion (1/2)

1. Requêtes pédophiles

   Outil de détection
   Grand ensemble étiqueté
   Estimation de la fraction de requêtes pédophiles

2. Utilisateurs pédophiles

   Étude de la notion d'utilisateur en général
   Estimation de la fraction d'utilisateurs pédophiles

Conclusion (2/2)

3. Dynamique temporelle

Évolution sur 3 ans
Intégration sociale des utilisateurs

4. Comparaison KAD eDonkey

Estimation à l'aide de données partielles

R. Fournier, T. Cholez, M. Latapy, C. Magnien, I. Chrisment, I. Daniloff and O. Festor.
Comparing paedophile activity in different P2P systems. Soumis.

Perspectives (1/2)

- Amélioration de l'outil de détection
  - requêtes précédente/suivante
  - langues, ordre des mots, catégories
  - apprentissage automatique
- Utilisateurs pédophiles
  - seuil différent pour considérer comme pédophile
  - détection de communautés
  - séquences de requêtes
  - échanges de fichiers

Perspectives (2/2)

- Méthodologie appliquée dans un autre contexte
  - d'autres thématiques
  - fraude bancaire

KAD

Comparer deux réseaux P2P différents

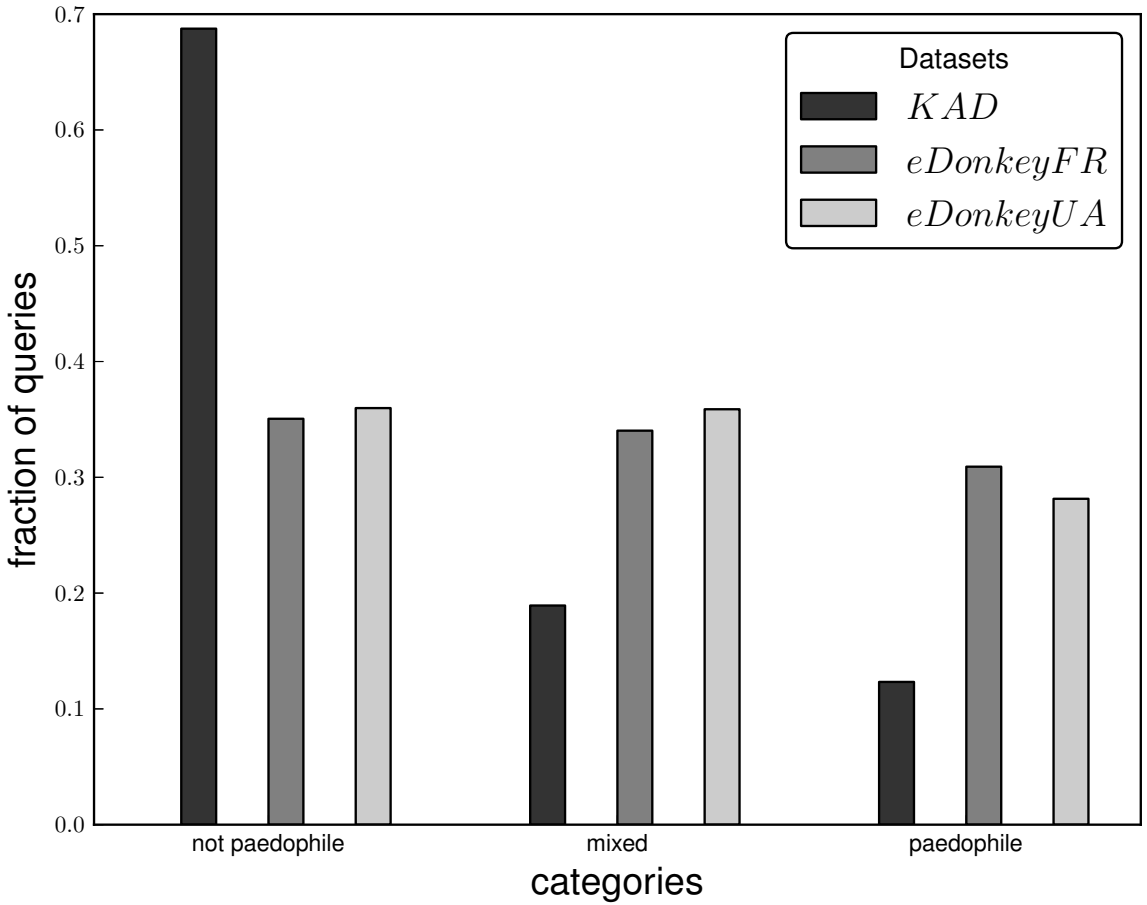
Description

- Réseau complètement distribué
- Les pairs sont responsables des mots-clefs et des fichiers

Mesure

- collaboration avec l'équipe MADYNES (LORIA)
- injection ciblée de pairs pour surveiller des mots-clefs
- 72 mots-clefs, répartis en 3 catégories
- 10 jours, 250 000 requêtes
- 2 serveurs eDonkey
- requêtes de longueur 1

KAD


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
fraction of queries
Datasets
KAD
$eDonkeyFR$
$eDonkeyUA$
not paedophile
mixed
paedophile
categories

KAD : estimations

- estimer la fraction de requêtes pédophiles dans KAD

  $\longrightarrow$ avec données partielles

- $\alpha = \frac{|Q'|}{|Q|}$
- $\beta = \frac{|\overline{P}|}{|P|}$
- $\frac{|P|}{|Q|} = \frac{|P'|/\beta}{|\overline{Q}|/\alpha} = \frac{\alpha}{\beta}\frac{|\overline{P}|}{|\overline{Q}|}$

- fraction de requêtes proche de 0,11%
- contraire à l'intuition initiale

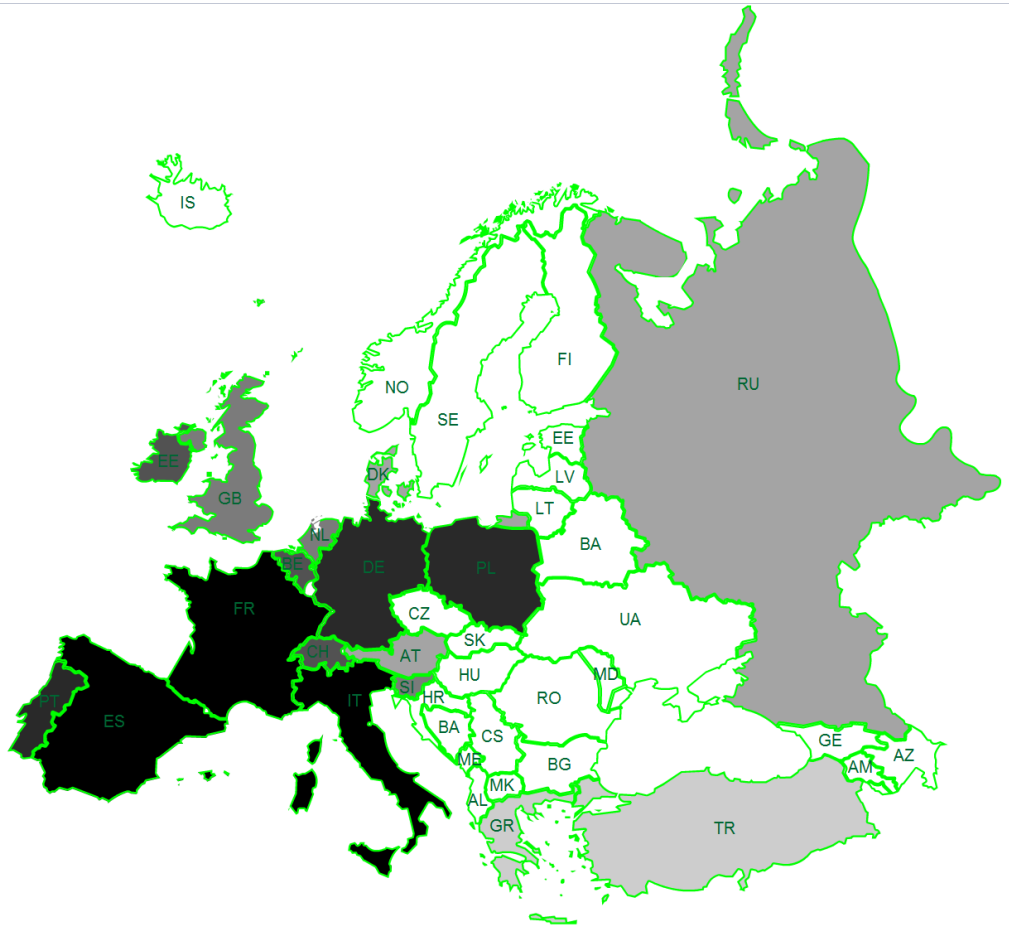
Geolocalisation

pays	# requêtes	# pedo.	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 :

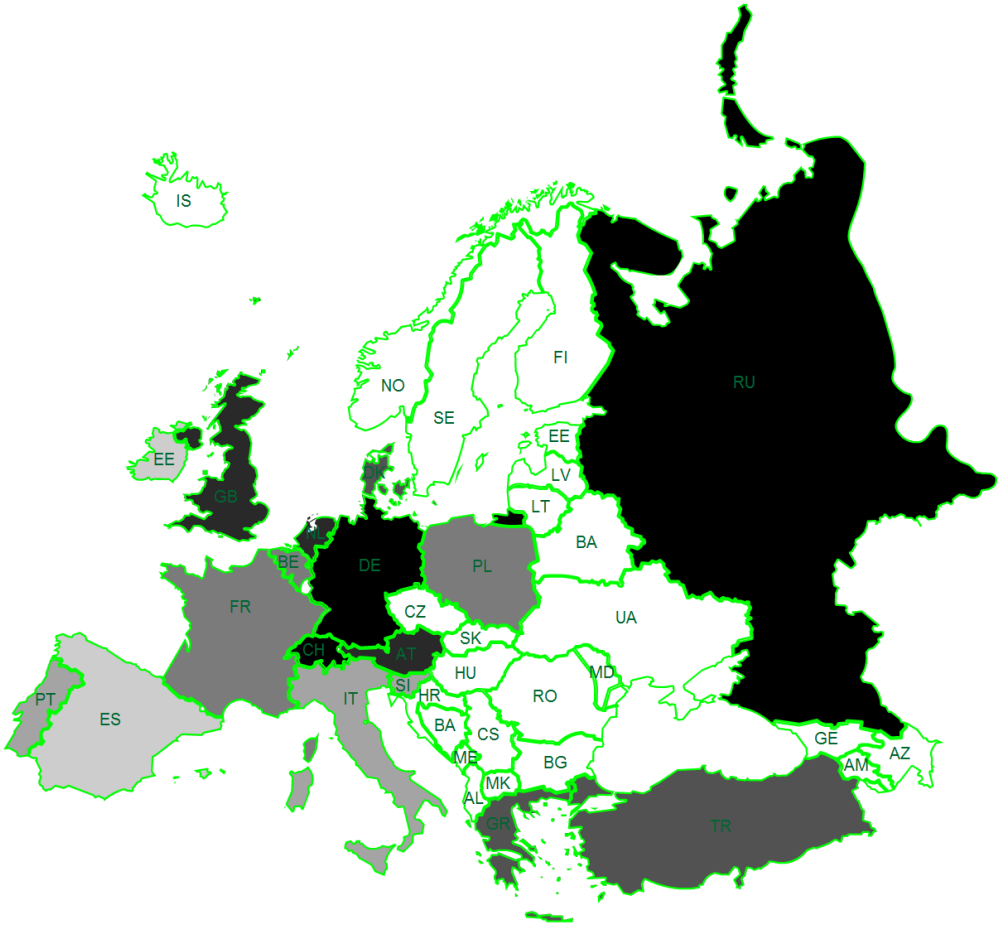
- langues
- encodage
- VPN

Geolocalisation



requêtes

Geolocalisation



fraction de requêtes pédophiles