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

Raphaël Fournier-S'Niehotta

UPMC
SORBONNE UNIVERSITÉS

THÈSE

dirigée par Matthieu Latapy

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

spécialité Informatique

Détection et analyse d'une thématique rare dans de grands ensembles de requêtes : l'activité pédophile dans le p2p

---

Raphaël Fournier-S'niehotta

soutenue publiquement le 21 décembre 2012 devant le jury composé de

Rapporteurs :	Emmanuel Viennet	Professeur, Université Paris-XIII
	Anne-Marie Kermarrec	Directrice de Recherche, INRIA
	Walid Dabbous	Directeur de Recherche, INRIA
Examinateurs :	Serge Abiteboul	Directeur de Recherche, INRIA
	Olivier Festor	Directeur de Recherche, INRIA
	Alix Munier Kordon	Professeur, UPMC
	Gilles Robine	Officier de Police, Ministère de l'Intérieur
Directeur :	Matthieu Latapy	Directeur de Recherche, CNRS

Remerciements

Cette thèse est le fruit d'un peu plus de trois années de travail et son aboutissement doit beaucoup à l'aide et au soutien de nombreuses personnes.

Je tiens tout d'abord à remercier chaleureusement Matthieu Latapy, qui m'a encadré en stage et a joué un rôle déterminant dans l'orientation de ma carrière vers la recherche. Son expérience, son soutien, ses remarques et encouragements ont été très précieux, depuis le début. Je souhaite aussi souligner sa gentillesse, sa disponibilité et sa force de travail. Travailler à ses côtés m'a enrichi professionnellement, scientifiquent et humainement.

J'ai aussi beaucoup appris aux côtés de Clémence Magnien, avec qui j'ai notamment collaboré pour tous mes articles. Ses idées et son travail ont grandement compté pour moi, de même que les discussions que nous avons eu lors des pauses et des repas. Je remercie aussi l'ensemble des membres de l'équipe ComplexNetworks, dans laquelle il a toujours été très agréable de travailler. J'ai en particulier eu le plaisir de faire des enseignements avec Jean-Loup Guillaume, je le remercie pour ses conseils et encouragements. Je remercie aussi Lionel Tabourier et Daniel Bernardes, avec qui nous avons eu de longues et stimulantes discussions. J'ai aussi partagé des repas, des pauses et des idées avec Alice Albano, Maximilien Danisch, Sébastien Heymann, Serguey Kirgizov, Élie Rotenberg et Fabien Tarissan. J'ai avancé durant ma thèse en observant ceux qui me précédaient et je tiens donc à remercier aussi Thibault Cholez, Thomas Aynaud, Lamia Benamara, Massoud Seifi, Oussama Allali et Abdelhamid Salah Brahim, mais aussi Christophe Crespelle et Guillaume Valadon.

J'ai eu la chance durant la rédaction de ce manuscrit de pouvoir compter sur plusieurs relecteurs dont les remarques m'ont permis d'améliorer grandement la qualité de mon manuscrit : un grand merci à Romain Campigotto, Jean-Loup Guillaume, Bénédicte Le Grand et Clémence Magnien.

J'ai encadré le stage de fin d'études de Master de Laure Fouard, ce qui a été une expérience formatrice. J'ai une pensée pour Élena Collado, Lucas Di Cioccio et Adrien Friggeri qui ont commencé et terminé leurs thèses à peu près en même temps que moi. Je remercie aussi vivement Véronique Varenne pour son efficacité sans faille et sa gentillesse.

Je remercie enfin ma famille et mes amis au sens large. Je partage ma vie avec Marie et je la remercie pour tout ce qu'elle a fait pour moi : son soutien et son amour m'ont accompagné du début à la fin de cette thèse.

Je tiens enfin à exprimer toute ma gratitude à Walid Dabbous, Anne-Marie Kermarrec et Emmanuel Viennet qui ont accepté d'être les rapporteurs de mon travail, ainsi qu'à Serge Abiteboul, Olivier Festor, Alix Munier Kordon et Gilles Robine qui ont bien voulu faire partie de mon jury.

Toutes mes excuses à ceux que j'ai pu oublier. S'ils m'ont lu jusque-là, ils méritent mes remerciements.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Pour de nombreux systèmes informatiques, il est essentiel de comprendre de quelle manière les utilisateurs s'en servent. Ceci permet en effet de proposer un meilleur service, d'augmenter les revenus générés ou encore de surveiller le fonctionnement du système. Par exemple, les responsables d'un moteur de recherche peuvent examiner quelles sont les requêtes entrées par les utilisateurs afin de mettre en place un système de suggestions. Le propriétaire d'un site de commerce en ligne peut quant à lui vérifier que ses produits sont convenablement présentés mais aussi que la navigation et le paiement sont accessibles au plus grand nombre. Enfin, un responsable informatique d'entreprise peut vouloir s'assurer que les employés ont un usage approprié des ressources et que les données de l'entreprise sont préservées des attaques.

Suivant les buts recherchés, de nombreuses méthodes peuvent être employées pour observer la manière dont les utilisateurs se servent d'un système : sondages, échantillonnages, questionnaires, etc. Une méthode de plus en plus répandue consiste à enregistrer automatiquement les interactions entre les utilisateurs et le système. Dans le cas de moteurs de recherche, on obtient ainsi des séries de requêtes envoyées au système qui contiennent une information extrêmement riche.

Au-delà des défis qu'ils soulèvent pour leur stockage, leur partage et leur visualisation, ces ensembles ouvrent des perspectives d'analyse nouvelles et prometteuses. Certaines sont très originales : par exemple, de tels ensembles de requêtes ont été utilisés récemment pour suivre la propagation des foyers d'une épidémie de grippe et même pour prédire avec fiabilité quelles personnes étaient susceptibles de tomber malades [28, 65].

Cette thèse se positionne dans ce contexte, l'objectif central étant d'utiliser de grands ensembles de requêtes pour étudier un sujet crucial pour notre société : l'activité pédophile dans les réseaux pair-à-pair (P2P).

La pédophilie désigne en général l'attirance sexuelle d'un adulte envers les enfants prépubères ou en début de puberté, un pédophile étant une personne qui éprouve ce type d'attirance. Le terme pédophilie peut aussi désigner par abus de langage la pédopornographie, c'est-à-dire la pornographie mettant en scène des enfants. Nous parlerons dans cette thèse d'*activité pédophile* pour désigner les

activités d'utilisateurs souhaitant obtenir des contenus à caractère pédopornographiques. Par extension, une *requête pédophile* sera donc une suite de mots-clefs envoyée à un moteur de recherche dans ce but. Il est clair que certaines de ces requêtes *ne proviennent pas* de pédophiles au sens médical ou judiciaire du terme. Nous supposons toutefois que c'est le cas de la grande majorité d'entre elles et désignerons donc dans la suite *toutes* les requêtes liées à la pédopornographie par l'expression *requêtes pédophiles* et les utilisateurs soumettant de telles requêtes seront appelés *utilisateurs pédophiles*.

Une meilleure connaissance de l'activité pédophile en général est essentielle d'abord pour la protection des enfants qui en sont victimes, l'arrestation des criminels concernés, mais aussi car de nombreux autres domaines sont impactés par ces activités. Notamment, il est souvent affirmé que l'Internet facilite les échanges de contenus pédopornographiques [16, 42, 62]. En particulier, les liens entre le P2P et l'activité pédophile sont régulièrement avancés afin de légiférer sur le filtrage des réseaux ou comme chef d'accusation à l'encontre des personnes responsables de la création de systèmes P2P [13, 62], ce qui a un effet sensible sur la popularité et l'existence même de ces réseaux [18].

Il existe toutefois extrêmement peu de connaissances fiables en la matière et l'objectif de cette thèse est de faire progresser significativement cette situation. Nous adopterons une approche originale reposant sur l'analyse de grands ensembles de requêtes. Celle-ci rompt avec les études qualitatives traditionnelles procédant par des entretiens avec des pédophiles arrêtés [26] ou par de la surveillance individualisée par les forces de l'ordre. Ces méthodes ont notamment pour défaut de se concentrer sur un nombre d'individus réduit, des durées limitées et/ou une zone géographique peu étendue.

Dans cette thèse, nous ferons face à un certain nombre de problématiques délicates, certaines générales et d'autres plus spécifiques, que nous allons développer maintenant. Nous utiliserons des données et feront appel à l'état de l'art que nous présenterons ensuite. Nous terminerons ce chapitre introductif par une présentation de l'organisation de cette thèse.

1.1 Problématiques

Nous considérons un *ensemble de n requêtes* de la forme $D = \{q_i, i = 1, \ldots, n\}$. Chaque requête q_i est un message soumis à un moteur de recherche, dans notre contexte un système indexant les fichiers dans un système P2P (voir section suivante). Une information temporelle $t(q_i)$ est associée à chaque requête q_i, correspondant à l'instant où celle-ci a été reçue par le moteur de recherche. On a : $i \leq j \Rightarrow t(q_i) \leq t(q_j)$ pour tout i et j, et nous utiliserons parfois cette relation d'ordre dans nos ensembles de requêtes. La fonction u associe à q_i l'utilisateur $u(q_i)$ qui en est à l'origine. On considérera en première approximation qu'un utilisateur

est un individu, mais il peut également s'agir d'un programme informatique (robot), d'un groupe d'individus, etc.

Faire progresser significativement les connaissances sur l'activité pédophile dans le P2P à partir de tels ensembles de requêtes, l'objectif de cette thèse, soulève plusieurs problématiques difficiles.

Collecte de grands ensembles de requêtes

Afin de pouvoir effectuer notre analyse, nous devons disposer de données adéquates, c'est-à-dire de grands ensembles de requêtes soumises au moteur de recherche d'un système P2P. Pour les obtenir, il faut procéder à une opération de collecte, qui peut s'avérer très complexe. Plusieurs facteurs contribuent à rendre cette tâche délicate, comme les contraintes légales, les protocoles des systèmes, (distribués, peu documentés, etc.) ou leur taille (plusieurs dizaines de millions d'utilisateurs dans le monde). Ces données sont précieuses pour la communauté scientifique et il faut donc également envisager de les rendre disponibles, ce qui introduit des contraintes supplémentaires.

Cette opération de *mesure* des réseaux P2P est un champ de recherche en soi qui, bien que très lié à notre travail, n'est cependant pas au cœur de cette thèse : nous avons pu accéder à des jeux de données collectés au préalable et de diverses manières (voir section 1.2). Nous avons cependant procédé à la normalisation des requêtes brutes et à l'anonymisation de celles-ci (voir annexe A).

Détection des requêtes pédophiles

La pédopornographie est une activité criminelle activement poursuivie par diverses organisations qui ont pour but d'anéantir les réseaux de production et de diffusion de tels contenus. Les individus qui participent à de tels échanges souhaitent donc généralement se cacher et ont donc recours à des pseudonymes et à des mots-clefs spécifiques, cachés, qu'un non-initié ne comprend pas ou ne considère pas comme faisant référence à de la pédopornographie [27]. Les réseaux P2P étant accessibles depuis la plupart des pays, les requêtes peuvent en outre être exprimées dans de nombreuses langues différentes. Un premier défi dans notre travail consiste donc à savoir détecter une *requête pédophile*, afin de pouvoir la discerner des autres et de pouvoir en estimer le nombre. Compte tenu de la taille des ensembles considérés, la classification des requêtes ne peut se faire de façon complètement manuelle et cette détection se traduit en pratique par l'élaboration d'un algorithme de classification des requêtes en deux classes, pédophiles et non pédophiles.

Une manière classique d'aborder ce type de problème de classification consiste à utiliser des techniques d'apprentissage automatique. Cependant, l'emploi de ces

méthodes nécessite de disposer au préalable de données étiquetées pour élaborer les modèles et entraîner les algorithmes à distinguer les requêtes appartenant à telle ou telle classe. Or, dans le contexte de l'activité pédophile dans le P2P, ce type de données n'existait pas au début de cette thèse. Au contraire, les travaux que nous avons effectués ayant permis d'obtenir de grands ensembles de requêtes pédophiles, de telles approches seront possibles dans le futur.

Une autre problématique importante de notre travail est l'évaluation de la qualité de la détection réalisée, c'est-à-dire des performances de notre outil de classification de requêtes pédophiles. En effet, malgré l'importance du sujet, peu de personnes disposent de l'expertise requise pour participer à la validation d'un tel outil. En outre, la faible présence relative de ces requêtes nécessite d'adopter des méthodes statistiques adaptées.

Classification des utilisateurs

Pour la société, disposer de statistiques fiables sur les utilisateurs qui participent aux échanges de contenus pédopornographiques est une problématique fondamentale, qu'ils en soient les fournisseurs ou les demandeurs. Dans notre contexte, étudier les utilisateurs soulève deux problématiques de recherche principales.

Dans un premier temps, il est fondamental de pouvoir *identifier* un utilisateur, c'est-à-dire de distinguer quelles requêtes proviennent de la même personne et quelle est cette personne. Sur l'Internet, les ordinateurs interagissent en étant identifiés par une adresse IP et un port de communication (qui ne sera souvent pas disponible dans nos données). Cependant, plusieurs utilisateurs peuvent se servir de la même machine (et donc de la même adresse) et une personne peut avoir à sa disposition plusieurs machines différentes (par exemple, à son domicile, sur son lieu de travail, des terminaux publics ou mobiles, etc.). En outre, pour diverses raisons techniques, les adresses IP peuvent être attribuées à des machines différentes au cours du temps, ce qui complique le problème de l'étude sur une durée relativement longue. L'identification des utilisateurs est donc en pratique quasiment impossible dans notre contexte. Dans la plupart des cas, nous pourrons cependant nous contenter de savoir seulement *distinguer* (c'est-à-dire ne pas mélanger) les requêtes de plusieurs utilisateurs, Cette problématique est loin d'être triviale, pour les mêmes raisons que précédemment.

L'autre difficulté essentielle à laquelle nous faisons face, même en sachant isoler les requêtes soumises par un utilisateur donné, est de décider si cet utilisateur est pédophile ou non. Nous ferons en première approximation l'hypothèse qu'un utilisateur ayant soumis au moins une requête pédophile est pédophile, mais dans ce cas les erreurs de classification des requêtes induisent des erreurs de classifications des utilisateurs. Bien entendu, des raffinements de cette hypothèse

sont envisageables, comme par exemple fixer un seuil différent pour être considéré comme pédophile, mais cela sort du cadre de notre thèse.

Notre étude doit faire face à plusieurs problématiques :
- l'activité pédophile est cachée et s'effectue dans plusieurs langues ;
- la grande taille des ensembles de requêtes et la faible proportion relative de requêtes pédophiles imposent des traitements adaptés ;
- il n'existe pas d'ensemble de référence et peu d'experts du sujet pouvant classer des requêtes comme pédophiles ou non pédophiles ;
- les informations disponibles ne permettent pas d'identifier ou de distinguer directement les utilisateurs.

1.2 Données

Les ensembles de requêtes que nous allons étudier dans ce manuscrit ont été collectés en observant à différents moments l'activité de deux des plus grands systèmes d'échanges de fichiers actuellement déployés, *eDonkey* [20, 46] et *KAD* [12] [1]. Sur ces deux réseaux, un utilisateur souhaitant télécharger un contenu auprès des autres utilisateurs peut se servir d'un moteur de recherche intégré. Il saisit alors sa requête dans un champ de recherche sous la forme d'une succession de mots-clefs. Le moteur de recherche propose en retour une liste de fichiers correspondant à ces mots-clefs, que l'utilisateur peut décider de télécharger auprès des pairs. De telles requêtes de recherche, que nous appelons plus simplement requêtes, sont intéressantes pour l'analyse car elles capturent l'intention de l'utilisateur à un moment donné.

Les deux systèmes fonctionnent de façons bien différentes du point de vue des recherches : alors qu'*eDonkey* est constitué d'un ensemble de serveurs, chacun indexant les fichiers dont il a connaissance et fournissant son propre moteur de recherche (sur ces fichiers), *KAD* a en revanche un moteur de recherche distribué, qui indexe tous les fichiers du réseau.

Nous présentons ci-dessous les différents jeux de données que nous utilisons par la suite ; leurs caractéristiques sont résumées dans le tableau 1.1.

1. Nous détaillerons dans le chapitre 5 les différences de fonctionnement entre ces deux réseaux P2P.

Pendant une durée de dix semaines en continu, en 2007, des membres de l'équipe *Complex Networks* du LIP6 ont observé le trafic destiné à un des plus importants serveurs *eDonkey* du moment [2]. En particulier, les requêtes auprès du moteur de recherche du serveur ont pu être collectées. Chaque requête comporte un horodatage, l'adresse IP et le port de communication de l'utilisateur, ainsi que les mots-clefs de la recherche. Nous appelons ce premier jeu de données `data-ed2k2007`.

En 2009, nous avons obtenu du responsable de deux serveurs *eDonkey* qu'il active la capacité d'enregistrement des requêtes soumises au moteur de recherche sur chacun de ses serveurs. Ceux-ci sont situés dans des pays différents (l'un en France, l'autre en Ukraine) et ont des politiques d'indexation de contenus [2] différentes. Le port de communication de l'utilisateur n'est cette fois pas disponible. En revanche, l'adresse IP est géolocalisée avant d'être anonymisée. La collecte sur le serveur en Ukraine a été arrêtée après quelques mois, celle sur le serveur en France est toujours opérationnelle en octobre 2012. Nous allons utiliser à plusieurs reprises dans ce manuscrit les données obtenues sur ces serveurs, en extrayant différents jeux de données selon les besoins.

Le jeu de données principal créé avec l'enregistrement du serveur français s'étend actuellement sur une période de 147 semaines (près de trois ans) et contient plus d'un milliard de requêtes. Nous le désignons sous le nom `data-ed2k0912`. Nous l'utiliserons notamment dans le chapitre 4 pour étudier l'évolution à long terme de l'activité pédophile. Auparavant, nous avons travaillé sur un jeu de données intermédiaire de 28 semaines, que nous appelons `data-ed2k2009` [48,49].

Nous avons également obtenu des données sur le réseau *KAD*, grâce à une collaboration avec des membres de l'équipe MADYNES du LORIA. Dans le cadre de la thèse de Thibault Cholez, les membres de l'équipe ont développé un système de supervision de *KAD*, appelé HAMACK [14, 15]. Celui-ci est capable de collecter dans *KAD* toutes les requêtes concernant un mot-clef donné. En revanche, HAMACK ne peut superviser l'intégralité des mots-clefs. Nous avons donc élaboré avec l'équipe MADYNES une liste de mots-clefs correspondant à nos besoins et nous avons ciblé la mesure sur ceux-ci (les détails sont présentés dans le chapitre 5). L'enregistrement des requêtes a duré dix jours en continu, en novembre 2010. Nous désignons ce jeu de données par l'appellation `data-KAD`.

Enfin, nous avons utilisé les mesures des serveurs *eDonkey* français et ukrainien, mais avons restreint les requêtes observées afin que les jeux de données soient comparables à *KAD* (voir chapitre 5). Nous appelons les ensembles ainsi créés `data-ed2k-FR` et `data-ed2k-UA` respectivement.

2. Ce qui signifie que certains fichiers partagés par les utilisateurs sont filtrés par les serveurs, qui ne les indexent alors pas.

	Durée	Nb. requêtes	Nb. IP	Spécificités
data-ed2k2007	10 semaines	107 226 021	23 892 531	port de communication disponible
data-ed2k0912	147 semaines	1 290 377 956	82 264 897	géolocalisation des IP
data-ed2k2009	28 semaines	205 228 820	24 413 195	géolocalisation des IP
data-KAD	10 jours	250 000	n/a	requêtes de longueur 1 uniquement
data-ed2k-FR	60 jours	241 152	n/a	requêtes de longueur 1 uniquement
data-ed2k-UA	60 jours	166 154	n/a	requêtes de longueur 1 uniquement

Tableau 1.1 – *Caractéristiques principales de nos jeux de données.*

Afin de pouvoir étudier ces ensembles de requêtes, nous avons dû les normaliser et adopter des formats similaires pour chacun. Cette procédure de standardisation comporte notamment l'anonymisation des informations que recelaient les requêtes, qui est un problème crucial et délicat. La figure 1.1 présente deux exemples de requêtes, dans leurs versions brute, normalisée et anonymisée. Les détails des procédures sont présentés en annexe A. Une partie des données a ensuite été mise à disposition de la communauté scientifique et publiée [80].

```
24/04-03:31:48.812965 ac95ebac883dd688e91d0d62f4af331d
'Photo Scarlett O'Hara été « Champs-Élysées »'

↪ 57635154 ac95ebac883dd688e91d0d62f4af331d photo scarlett o hara ete
champs elysees

↪ 57635154 425926 photo scarlett hara ete champs elysees


14/05-02:02:37.375921 63f063400f20715230f8ae4b5156588f
'Credit 1234 4567 7654 4321 johndoe7643@something.com'

↪ 381600 47337 credit card 1234 4567 7654 4321 johndoe7643 something
com

↪ 381600 47337 credit card something com
```

Figure 1.1 – *Illustration de notre procédure de normalisation et anonymisation sur deux exemples de requêtes. Chacune est présentée, de haut en bas, en version brute puis normalisée et anonymisée.*

Après ces étapes de normalisation et d'anonymisation, nous obtenons des ensembles de requêtes, de la forme :

$$q_i = (t, u, k_1, k_2, \ldots, k_n)$$

avec :

- t un horodatage, exprimé en secondes depuis le début de la mesure
- u une adresse IP, éventuellement complétée par un port de communication (jeu de données `data-ed2k2007`) ou une information de géolocalisation (jeux de données `data-ed2k2009` et `data-ed2k0912`)
- $(k_1, k_2, \cdots, k_n)$ une suite de mots-clefs.

Nous ferons usage des différentes composantes des requêtes selon les besoins de nos études. L'information sur l'utilisateur sera explorée spécifiquement dans le chapitre 3 et l'information temporelle sera fondamentale pour les travaux présentés dans le chapitre 4.

> Nous utilisons des ensembles de requêtes de recherches de contenus effectuées sur les systèmes P2P*eDonkey* et *KAD* ; chaque ensemble contient plusieurs millions et jusqu'à un milliard de requêtes sur presque 3 ans, dûment normalisées et anonymisées.

1.3 État de l'art

L'étude que nous avons menée ici constitue un cas pratique de détection et de caractérisation d'une thématique rare dans de grands ensembles de requêtes. C'est un problème ancré dans le réel, à l'intersection de plusieurs sujets : l'étude des ensembles de requêtes, la classification et la pédopornographie. Nous allons donc présenter un rapide état de l'art de ces différents domaines, en ciblant dans chaque cas les contributions les plus pertinentes pour notre travail.

1.3.1 Analyse d'ensembles de requêtes

Notre étude de la pédopornographie dans le P2P repose sur l'analyse de grands ensembles de requêtes effectuées pour la recherche de contenus. Ce type d'analyse se retrouve dans de nombreux contextes, et a fait à ce titre l'objet de nombreuses recherches depuis de nombreuses années.

Tout d'abord, ces ensembles de requêtes ont été présentés comme les « fichiers résultant de l'enregistrement des interactions entre des systèmes de Recherche d'Information (RI) et les personnes qui s'en servent » [56]. On appelle cette discipline *Transaction-Log Analysis*. Elle se limite initialement à l'évaluation des systèmes de RI en bibliothèque. Ce domaine a été profondément transformé par l'apparition de l'Internet, l'augmentation du nombre et de la puissance des machines interconnectées et l'explosion des usages. Ces facteurs ont en effet fait apparaître de

nouveaux types d'ensembles de requêtes et ont augmenté drastiquement la taille des ensembles considérés [3]. Ils ont aussi multiplié les domaines d'application.

Plusieurs auteurs ont alors utilisé les ensembles de requêtes pour analyser différents systèmes provenant du Web [17, 41]. En particulier, les entreprises possédant des moteurs de recherche ont observé les tendances et les effets produits par les améliorations apportées à leurs systèmes [30, 79]. Leurs objectifs sont d'améliorer la compréhension des interactions entre les contenus, la personne qui effectue la recherche et le moteur de recherche (ou les interactions entre deux de ces entités). Cela permet ensuite d'envisager l'amélioration du système de recherche, en proposant par exemple une l'aide à la recherche ou en modifiant l'ergonomie générale du système.

Mais l'étude de vastes ensembles de requêtes ne se limite pas à l'amélioration des systèmes chargés de les traiter. Elle ouvre également la porte à une bien meilleure connaissance des activités humaines et ouvre des perspectives très originales, inconcevables il y a encore peu de temps. Récemment, des chercheurs de Google ont montré par exemple qu'en analysant les requêtes relatives aux symptômes de la grippe effectuées sur leur moteur de recherche, ils pouvaient suivre la diffusion des foyers de l'épidémie plus rapidement qu'avec les méthodes classiques reposant sur les données cliniques [28]. D'autres ont utilisé les messages publiés par les membres du réseau Twitter et les liens entre ces membres pour calculer la probabilité qu'un individu donné tombe malade [65]. Ces travaux ouvrent la voie à des études complètement nouvelles (et très prometteuses), à la frontière de plusieurs domaines de recherche actuels.

La collecte et l'analyse de grands ensembles de requêtes provenant du pair-à-pair est aussi un champ de recherche très actif. Les travaux existants étudient principalement les propriétés des utilisateurs qui peuvent être utiles pour la conception et l'amélioration des protocoles, comme les durées de connexions, les comportements de partage ou les similarités d'activité de recherche de fichier ou de localisation géographique [31, 33, 54, 66].

D'autres travaux analysent les requêtes saisies par les utilisateurs [29, 44] mais s'en tiennent à des statistiques limitées (telles que la longueur des requêtes, leur nombre, leur redondance et les intervalles entre leurs arrivées). Peu de travaux étudient l'intérêt des utilisateurs [36, 67, 68].

Remarquons cependant que depuis les prémices de l'utilisation de l'analyse des requêtes comme méthode d'étude, des limites importantes ont été identifiées [11, 32, 57]. Elles résultent surtout du fait que les enregistrements sont effectués sans contact avec l'utilisateur. En particulier, il n'est pas possible pour le chercheur de connaître précisément le vécu de l'utilisateur au moment de la recherche ni le

3. L'expression anglophone *big data* (« données volumineuses ») est parfois employée, mais son acception est très large.

véritable besoin d'information sous-tendant sa recherche. En outre, on ne peut que difficilement mesurer par cette méthode la satisfaction de l'utilisateur vis-à-vis des résultats que lui a proposé le système de recherche. Ces critiques sont bien entendu valables dans notre contexte d'étude : nous ne pouvons par exemple pas distinguer une requête effectuée par les forces de police souhaitant surveiller les fichiers pédophiles d'une requête provenant d'un individu souhaitant participer à des échanges de tels fichiers.

La discipline s'est progressivement structurée et les informations contenues dans les ensembles de requêtes se sont standardisées [40]. Concernant les ensembles de requêtes provenant de moteurs de recherche, il est devenu assez classique qu'une requête comporte une information temporelle (en général, il s'agit de la date et de l'heure de l'interaction entre l'utilisateur et le système de recherche), une identification de l'utilisateur (une adresse IP, éventuellement complétée par un port de communication ou un identifiant attribué par le système) et les mots-clefs de la requête.

L'étude de ces éléments est maintenant devenue assez classique [39]. Une requête peut être analysée dans sa globalité ou plus finement au niveau de chacun des termes qui la composent. L'analyse des sessions de recherche, c'est-à-dire plusieurs requêtes d'une même personne pendant une courte durée, est également intéressante, en particulier pour comprendre le comportement des utilisateurs face à un système de recherche. Nous ferons usage de ces différentes notions dans notre travail.

1.3.2 Techniques de classification

Notre étude repose sur la classification de requêtes en deux catégories principales : pédophiles et non pédophiles. Celle-ci peut être effectuée soit manuellement, en se basant sur les connaissances d'experts humains, soit à l'aide d'un système automatique. Dans notre contexte, l'approche manuelle s'avère infaisable dans la pratique, compte tenu de la quantité phénoménale d'informations à traiter. Il est donc indispensable de recourir à un système automatisé.

Pour cela, des méthodes d'apprentissage supervisé pourraient être employées. Elles reposeraient sur l'entraînement d'un modèle capable de discriminer une requête *pédophile* d'une requête *non pédophile*. Un tel système attribue à chaque nouvelle requête une classe (pédophile ou non pédophile) tout en maximisant la performance de classement. On obtient un score pour chaque requête et c'est l'analyse de ces scores qui permet de distinguer les requêtes pédophiles des non pédophiles. Ces méthodes requièrent cependant de disposer d'un ensemble de requêtes connues comme étant pédophiles, pour la phase d'apprentissage. De tels ensembles n'étant pas disponibles quand nous avons commencé notre travail, nous avons donc écarté cette approche.

En outre, les modèles d'apprentissage souffrent dans notre contexte du problème dit « des classes déséquilibrées [35]. Typiquement, dans un contexte de classification binaire, on parle de déséquilibre lorsque le nombre d'éléments de la classe majoritaire présente un rapport de $100 : 1$ voire $1000 : 1$ avec celui de la classe minoritaire. Dans notre cas, on s'attend à ce que les requêtes non pédophiles soient largement plus nombreuses que les requêtes pédophiles. Cette situation de déséquilibre engendre des erreurs importantes pour le modèle, surtout lorsque l'on se trouve dans une situation où les tailles d'échantillons sont réduites, ce qui est notre cas.

L'évaluation de la performance d'un outil de classification dans un contexte à deux classes comme le nôtre se fait généralement à l'aide du taux d'erreur. En utilisant la convention classique où la classe en minorité est dite « positive » et la classe en majorité est « négative », une représentation de la performance du classificateur est donnée par une matrice de confusion comme celle du tableau 1.2.

	Positif	Négatif
Détecté positif	Vrai positif (TP)	Faux positif (FP)
Détecté négatif	Faux négatif (FN)	Vrai négatif (TN)

Tableau 1.2 – *Matrice de confusion pour un problème à deux classes.*

On estime la valeur de quatre indicateurs :
- les vrais positifs (TP, pour *true positive*), qui sont les éléments de la classe minoritaire détectés comme tels ;
- les vrais négatifs (TN, pour *true negative*), qui sont les éléments de la classe majoritaire détectés comme tels ;
- les faux négatifs, (FN, pour *false negative*), qui n'ont pas été détectés comme positifs alors qu'ils auraient dû l'être ;
- les faux positifs, (FP, pour *false positive*), détectés comme positifs mais qui n'auraient pas dû l'être.

Le taux d'erreur est alors défini par

$$TauxErreur = 1 - \frac{TP + TN}{TP + FP + TN + FN} \quad .$$

Cependant, dans le cas de classes déséquilibrées, l'utilisation de cet indicateur n'est pas approprié. Dans le cas de données où il y aurait 5% d'éléments de la classe minoritaire et 95% pour la classe majoritaire, une approche naïve qui consisterait à classer tous les éléments comme appartenant à la classe majoritaire n'aurait un taux d'erreur que de 1%. Ceci pourrait sembler très bon mais masquerait le fait qu'aucun élément de la classe minoritaire n'a été convenablement détecté. Cette métrique n'est donc pas adaptée à ce type de problème.

D'autres métriques peuvent être introduites pour fournir une meilleure évaluation des performances : la précision et le rappel. Avec les notations présentées

ci-dessus, la précision est définie comme :

$$Precision = \frac{TP}{TP + FP}$$

et le rappel comme :

$$Rappel = \frac{TP}{TP + FN} \quad .$$

Intuitivement, la précision est une mesure de l'exactitude : elle indique parmi les éléments qui ont été détectés comme positifs, la proportion d'entre eux qui le sont réellement. Le rappel est une mesure de complétude représentant la fraction des éléments de la classe positive qui ont été étiquetés correctement. La précision et le rappel nous permettent d'évaluer notre outil et de pouvoir l'utiliser pour quantifier la fraction de requêtes pédophiles dans différents jeux de données.

Pour améliorer un outil de classification donné, c'est-à-dire trouver un compromis acceptable entre précision et rappel, on utilise souvent des courbes ROC (*Receiver Operating Curve*) [23]. Celles-ci permettent d'observer les différentes versions de l'outil de classement sur une courbe et de sélectionner les plus performants. Le nombre de faux positifs est représenté en abscisse et le nombre de vrais positifs en ordonnée. La performance globale de l'outil de classification peut s'évaluer à l'aide de l'aire sous la courbe (*Area Under Curve*, ou AUC). Nous ne ferons cependant pas usage de cette méthode qui, dans notre contexte, aurait été très coûteuse en expertise humaine (voir chapitre 2).

1.3.3 Pédopornographie

Comme nous l'avons rappelé en préambule de cette thèse, la pédophilie désigne l'attirance sexuelle d'un adulte envers les enfants prépubères ou en début de puberté. Un pédophile est une personne qui éprouve ce type d'attirance. La pédophilie est classée comme un trouble de la préférence sexuelle (trouble mental) par la classification internationale des maladies (CIM [55]) et comme paraphilie par le manuel diagnostic et statistique des troubles mentaux (DSM-IV [6]). Le terme « pédophilie » peut aussi désigner la pédopornographie, c'est-à-dire la pornographie mettant en scène des enfants [73].

La plupart des législations dans le monde condamnent aujourd'hui fortement la pédophilie, qui est reconnue comme un grave délit ou un crime. L'adulte impliqué dans ces relations est considéré comme seul responsable et coupable. Contrairement à des rapports entre adultes, l'absence de consentement de l'enfant n'est pas requise pour que l'infraction soit constituée : la relation sexuelle est en elle-même interdite. Les documents pédopornographiques (photographies, films, etc.) sont également interdits, éventuellement par des lois spécifiques, comme en France (article 227-23 du Code Pénal). Les œuvres imaginaires sont le plus souvent condamnées, même si cela varie selon les pays : les textes pédopornographiques ne

sont pas interdits en France, les dessins (appelés *Lolicon* [72]) sont autorisés au Japon tant que de vrais modèles n'ont pas été impliqués dans le processus créatif.

Une diffusion à grande échelle de contenus pédopornographiques sur l'Internet pose des problèmes importants pour la société, le premier étant bien entendu que des adultes abusent sexuellement d'enfants pour produire ces contenus. Il y a aussi le fait que des usagers innocents peuvent se retrouver confrontés involontairement à des images d'une extrême violence. Cela peut aussi amener des individus à développer un intérêt pour le sujet et jouer un rôle important dans l'acceptation générale de la pédophilie [60, 78].

Des réseaux organisés de production et de distribution de contenus pédophiles ont été identifiés et analysés dès 1984 [47], bien avant que l'informatique ne devienne grand public. En revanche, dans la seconde moitié des années 1990, différents travaux ont commencé à laisser entendre que la généralisation de l'accès à l'Internet avait permis à ces réseaux d'accroître leur dimension et avait ainsi facilité l'accès à la pédopornographie [34, 45].

Récemment, la pédopornographie sur l'Internet a été étudiée sous les aspects économique et sociologique. En 2010, une étude [21] a détaillé les aspects économiques de la pédopornographie sur l'Internet, d'après le témoignage anonyme d'un informaticien ayant travaillé dans ce milieu. En 2010 également, le criminologue Patrice Corriveau a publié un travail approfondi sur les groupes de nouvelles [4] à caractère pédopornographique [16], qui permettent aux pédophiles de communiquer et de s'organiser pour échanger ensuite des contenus pédopornographiques. Enfin, Frank *et al.* [25] ont étudié différents sites Web sur le sujet et ont utilisé des méthodes d'analyse de réseaux sociaux pour déterminer ceux sur lesquels les investigations policières devraient porter en priorité.

D'autres auteurs se sont intéressés aux caractéristiques des échanges pédopornographiques dans les réseaux P2P [22, 71]. Ces premières études quantitatives reposent néanmoins sur des jeux de données de taille réduite et collectés à la main (typiquement en saisissant quelques requêtes et en examinant les résultats obtenus). Leur objectif est avant tout d'établir la présence d'activité pédophile dans le P2P, pas de la quantifier, ni de la caractériser.

Deux articles quantifient l'activité pédophile dans un réseau P2P d'une façon similaire à celle que nous développons dans cette thèse [36, 68].

Dans [36], les auteurs considèrent des ensembles de 10 000 requêtes collectées durant trois dimanches de 2005. Deux évaluateurs ont manuellement classifié les requêtes comme appartenant ou pas à la catégorie *pornographie illégale* (cette étude ne se limite donc pas à la pédopornographie, mais prend aussi en compte d'autres formes de violence). Ils concluent que $1,6\%$ des requêtes observées appartiennent à cette catégorie. Leur jeu de données est de taille réduite, avec donc un faible

4. L'appellation « groupes de nouvelles », ou *newsgroups*, désigne le système d'échange de messages Usenet [76].

nombre de requêtes pédophiles, ce qui le rend statistiquement peu significatif. De plus, leur méthodologie pour la classification est complètement manuelle (et ne passe donc pas à l'échelle) et repose sur seulement deux évaluateurs, dont l'expertise peut être questionnée.

Ce même groupe d'auteurs a proposé des techniques de découverte automatique de mot-clefs pédophiles [37]. Ces méthodes requièrent cependant un grand ensemble de requêtes connues comme étant pédophiles. Utiliser de telles méthodes constitue donc une perspective de notre travail, puisque nous avons depuis réuni des requêtes pouvant servir de base d'apprentissage.

Dans [68], l'auteur utilise un ensemble de $235\,513$ requêtes (soit environ $10\,000$ fois moins que nos données). Compte tenu de la fréquence des requêtes de cette nature, la taille de cet échantillon reste faible. Les requêtes sont classées comme pédophiles ou non selon qu'elles contiennent un mot clef provenant d'une liste particulière, similaire à notre liste *explicit* (voir section 2.1). Cette approche n'est cependant pas complètement satisfaisante, puisque de nombreuses requêtes ne contiennent pas de tels mots-clefs mais des combinaisons de mots qui, séparément, ne suffisent pas pour qu'une requête soit « pédophile », comme nous le verrons dans le chapitre 2. L'auteur conclut que près d'$1\,\%$ des requêtes examinées sont pédophiles, sans fournir la liste des mot-clefs utilisés ni ses données, ce qui rend la reproduction de ses expérimentations impossible.

Ces contributions doivent être vues comme pionnières pour la quantification et la caractérisation de l'activité pédophile dans le P2P, mais elles restent d'envergure limitée et ne s'intéressent notamment pas à la notion d'utilisateur, pourtant cruciale dans ce domaine.

Notre travail se trouve à l'intersection de trois domaines : l'analyse de grands ensembles de requêtes, la classification et, bien entendu, la pédopornographie. Nous avons présenté succintement les notions et résultats de ces domaines.

1.4 Organisation de la thèse

Dans cette thèse, nous avons pour objectif de faire progresser les connaissances sur l'activité pédophile dans les réseaux P2P, par une analyse de grands ensembles de requêtes.

La première étape importante consiste donc à être capable de détecter les requêtes pédophiles. Compte tenu de la taille des jeux de données considérés et de la relative rareté des requêtes pédophiles, il nous faut utiliser un filtre automatisé.

Nous détaillons la conception d'un tel outil de détection de requêtes pédophiles dans le chapitre 2. Nous introduisons une méthode d'évaluation de celui-ci, ce qui nous permet de connaître les erreurs de classification qu'il commet. Nous pouvons ensuite proposer une première estimation fiable de la fraction de requêtes pédophiles dans le réseau *eDonkey*.

Après avoir quantifié le nombre de requêtes pédophiles, il est naturel de s'intéresser aux utilisateurs qui en sont à l'origine. Nous montrons dans le chapitre 3 que les informations dont nous disposons permettent d'étudier plusieurs modélisations des utilisateurs. Nous les comparons et fournissons une estimation de la fraction d'utilisateurs pédophiles. Nous proposons également une estimation des performances de notre outil de détection lorsque l'on passe de la granularité des requêtes à celle des utilisateurs.

Cette notion d'utilisateur est également au cœur du chapitre 4 dans lequel nous prolongeons notre étude par l'examen de la dynamique temporelle de leurs comportements par rapport à la thématique pédophile. Dans un premier temps nous observons l'évolution du trafic pédophile sur un serveur *eDonkey* pendant près de trois ans. Dans un second temps, nous étudions l'intégration sociale des utilisateurs pédophiles en considérant la dynamique de leurs requêtes dans la journée, ce qui est une approche originale et prometteuse sur le sujet.

Ces résultats sont obtenus avec des données provenant d'*eDonkey*. Pour en accroître la généralité, il est important de les confronter à ceux d'autres réseaux. Nous présentons dans le chapitre 5 la méthodologie que nous avons élaborée et utilisée pour comparer les réseaux *KAD* et *eDonkey*. Nous terminons ce chapitre en inférant la fraction de requêtes pédophiles présentes dans *KAD* à partir de celle d'*eDonkey*.

Ce manuscrit s'achève par le chapitre 6 dans lequel nous présentons nos conclusions et les perspectives de recherche ouvertes par les travaux de cette thèse.

Soulignons que nos travaux sur l'activité pédophile dans le P2P ont été réalisés avec un souci de réutilisation dans d'autres contextes. Nous tentons autant que possible de passer du cas particulier de notre étude à des méthodes générales, appropriées à l'étude de classes rares de requêtes dans d'autres grands ensembles de requêtes. À titre d'exemple, remarquons que la recherche de transactions bancaires frauduleuses dans des grands ensembles de transactions est un sujet de recherche en soi, aux problématiques similaires : la notion d'utilisateur n'est pas évidente (les utilisateurs peuvent se servir d'une ou plusieurs cartes et plusieurs personnes peuvent payer avec la même carte), la rareté des fraudes rend nécessaire des traitements statistiques similaires, etc.

Chapitre 2

Requêtes pédophiles

Ce chapitre est consacré à la quantification de l'activité pédophile dans un réseau P2P en terme de requêtes.

Nous mettons d'abord au point un outil permettant de détecter les requêtes ciblant des contenus pédopornographiques (section 2.1). Ne disposant pas d'un ensemble de référence, c'est-à-dire des requêtes connues comme étant pédophiles, nous ne pouvons pas utiliser de techniques d'apprentissage, faute de pouvoir entraîner un modèle à classer les requêtes. En revanche, nous mettons à profit les connaissances d'experts du domaine et nos propres observations pour concevoir un outil de détection automatique. Celui-ci résulte d'une étude préliminaire qui nous a permis de distinguer quatre catégories de requêtes pédophiles.

Cet outil commet bien sûr des erreurs de classification, des requêtes non pédophiles pouvant être étiquetées comme pédophiles et des requêtes pédophiles n'étant pas détectées comme telles. En vue d'aboutir à une estimation précise de la quantité de requêtes à caractère pédophile, il est fondamental de connaître avec précision ces taux d'erreur. Nous verrons que se posent alors des questions délicates liées à la rareté de la thématique, ce que nous détaillons dans la section 2.2.

Nous présentons la mise en place d'un protocole de validation de notre outil par des experts qualifiés et indépendants dans la section 2.3 puis les résultats de cette validation en section 2.4. À l'aide des taux d'erreur de notre outil présentés dans la section 2.5, nous obtenons finalement une estimation fiable de la fraction de requêtes pédophiles présentes dans nos ensembles, que nous donnons en section 2.6.

2.1 Détection

Nous avons mis au point un outil de détection de requêtes pédophiles qui effectue une séquence de tests sur la suite (k_i) des mots-clefs d'une requête anonymisée q_i. Chaque étape vise à repérer la présence d'un ou plusieurs mots-clefs qui permettront d'établir si la requête appartient à au moins une des catégories de requêtes pédophiles que nous avons identifiées (détaillées ci-dessous).

Pour mettre en place notre outil, nous avons travaillé à partir des données `data-ed2k2007`, les seules disponibles à ce moment-là et suffisantes pour cette partie de l'étude.

La première étape de la conception de l'outil a reposé sur l'expérience acquise par les membres de l'équipe *Complex Networks* qui ont collaboré durant plusieurs années avec les forces de l'ordre travaillant sur le sujet de la pédopornographie (projets MAPAP et ANR MAPE [52]). Ces experts de la pédopornographie en ligne nous ont fourni une liste préliminaire de mots-clefs spécifiques à cette thématique. Ceux-ci sont destinés exclusivement à la recherche de contenus pédopornographiques dans les systèmes P2P ; ils n'ont pas d'autre usage que ce contexte spécifique. Citons par exemple *qqaazz*, *hussyfan* ou *r@ygold*. Le dernier de ces mots-clefs fait par exemple référence à Richard Goldberg, pédophile condamné [74], qui utilisait ce pseudonyme pour diffuser de tels contenus. Ce mot-clef est devenu ensuite un moyen d'indiquer dans le nom d'un fichier sa nature pédophile. Remarquons que certains de ces mots-clefs sont des abréviations ou des séquences de lettres qui n'ont pas nécessairement de sens dans une quelconque langue.

Cependant, cette liste préliminaire ne suffit pas car certaines requêtes pédophiles ne contiennent aucun de ces mots-clefs spécifiques. Nous avons alors exploré les co-occurrences de mots-clefs. Les noms de fichiers sur les réseaux P2P contiennent souvent des mots-clefs proches thématiquement. Un fichier musical d'une chanson de Madonna pourra ainsi contenir les mots-clefs « music » et « pop » (la nature générale du contenu et le genre de musique), « mp3 » (le format de fichier), ainsi que le nom de la chanson et éventuellement celui de l'album. En analysant les termes qui apparaissent fréquemment dans les requêtes contenant les mots-clefs spécifiques à la pédophilie, nous avons pu ajouter de nouvelles règles de détection à notre outil. Nous avons créé des listes de mots-clefs sémantiquement proches (par exemple liés à l'acte sexuel ou à l'enfance) et observé si la présence combinée de mots-clefs appartenant à diverses listes permettait de détecter davantage de requêtes pédophiles. En procédant à plusieurs itérations, jusqu'à ce que les améliorations introduites dans la nouvelle mouture de l'outil ne donnent pas de résultats sensiblement différents de la précédente version, nous avons identifié différentes catégories de requêtes à caractère pédophile et nous avons conçu un filtre de détection adapté.

Nous adoptons finalement quatre règles de détection, qui font appel à des mots-clefs répartis en six listes, que nous présentons en détail ci-dessous. La figure 2.1 illustre le fonctionnement de notre outil.

Nous appelons la liste construite avec les mots-clefs des experts *explicit*. Chaque requête qui contient au moins l'un de ces mots-clefs est étiquetée comme pédophile. Cela constitue notre **première catégorie** de requêtes pédophiles.

De nombreuses requêtes pédophiles contiennent des mots-clefs liés aux enfants (ou à l'enfance en général) et des mots-clefs liés à l'acte sexuel : pour définir la **deuxième catégorie de requêtes pédophiles**, nous avons élaboré deux listes relatives à ces thématiques, appelées respectivement *child* et *sex*. Nous étiquetons comme pédophile une requête qui contient au moins un mot-clef de chacune de

ces catégories. Cela peut introduire cette fois des classification trompeuses. En effet, une requête telle que « destiny's child sexy daddy » décrit vraisemblablement la chanson « Sexy Daddy » du groupe Destiny's Child. Cependant, comme elle contient les mots-clefs « Sexy » et « Child », l'outil la classe comme pédophile.

La **troisième règle de classement** est une variante de la précédente. Nous établissons deux nouvelles catégories de mots-clefs, relatives à la position des adultes et des enfants au sein de la famille, appelées respectivement *familyparents* et *familychild*. La première contient par exemple « *father* », la seconde contient « fille » ou « *daughter* ». L'outil étiquette comme pédophile une requête qui contient un mot de chacune de ces catégories ainsi qu'un mot-clef de la catégorie *sex*.

Ensuite, de nombreuses requêtes contiennent des indications d'âge, le plus souvent sous la forme anglophone *n yo*[1], traduisant la volonté de l'utilisateur de trouver des contenus mettant en scène des enfants âgés de *n* années. D'autres suffixes se trouvent parfois à la place de *yo*, comme *yr* ou *years old*. La liste qui contient les différentes variantes de ces suffixes s'appelle *agesuffix*. De telles indications d'âge sont des marqueurs fréquents qui indiquent que la requête est pédophile. Cependant, elles ne suffisent cependant pas, puisqu'elles sont aussi présentes dans des contextes très différents : par exemple, elles sont très utilisées par des personnes cherchant des jeux vidéos destinés à des enfants d'un certain âge. L'outil ne classe donc une requête comme pédophile que si elle contient une indication d'âge inférieur ou égal à 16 ans, ainsi qu'un mot-clef appartenant soit à la catégorie *sex* soit à la catégorie *child*.

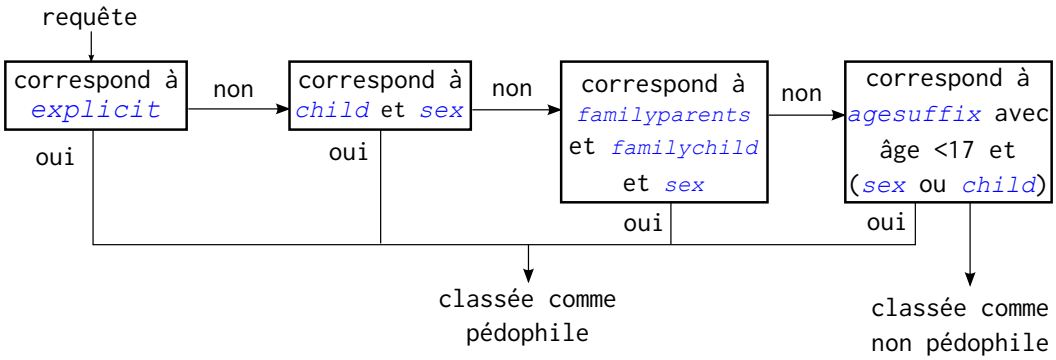


Figure 2.1 – *Séquence des tests réalisés par notre outil de détection. Chaque boîte évalue si la requête contient des mots-clefs appartenant à six listes de mots particulières, appelées :* explicit, child, sex, familyparents, familychild, *et* agesuffix*). Ces listes sont données en annexe D.*

Enfin, les requêtes sur les réseaux P2P sont fréquemment effectuées en anglais, y compris par des utilisateurs dont ce n'est pas la langue principale. Un certain nombre de requêtes sont tout de même effectuées dans d'autres langues. La plupart

1. La langue anglaise utilise fréquemment l'abréviation *yo* pour *years old*.

des mots-clefs de nos listes sont donc en anglais, mais nous avons inclus dans la mesure du possible leurs traductions les plus fréquentes dans différentes langues telles que le français, l'allemand, l'espagnol et l'italien.

Remarquons qu'une requête pédophile peut appartenir à plusieurs catégories. Par exemple, une requête qui contient un mot-clef de la liste *explicit* ainsi qu'un mot-clef de la catégorie *sex* et un de la catégorie *child* appartient à la première et à la deuxième catégories. L'implémentation de notre outil réalise les tests de chaque requête de façon séquentielle (voir figure 2.1). Nous étudions dans l'annexe C les effets de ce fonctionnement sur le nombre de requêtes détectées par catégorie.

La figure 2.2 présente un ensemble de requêtes provenant de nos jeux de données et illustre la détection des requêtes pédophiles par notre outil.

sirenia at sixes and sevens	capricorne one
dj coupe decale	dream dance vol
hannah montana clear	the mentalist s01e19
mino reitano discografia	ich mich nach deiner liebe soundtrack
qqaazz little girl	fine ukraine
el gallo sube	dash berlin man on the run
motherfucker of the year	*porno infantil*
h2o the last prime minister	kyle 4x07
devenir male dominant	billie jean body
kid rock fuck that	taviani
saghe mentali	una voce nella notte ost
gram parsons wild horses	paolo conte schiava del live
sheherazade korsakov	schiavi padroni
naruto	*12yo fuck video*
incest mom son video	amedeo minghi alla fine
desaparecidos fiesta loca	michael jackson bad man in the mirror
secret life american vostfr	

Figure 2.2 – *Exemples de requêtes de nos jeux de données. Les requêtes en italique souligné sont détectées comme pédophiles, les autres comme non pédophiles.*

En partant de mots-clefs fournis par des experts de la pédopornographie en ligne, nous avons réalisé une exploration préliminaire et distingué quatre catégories de requêtes pédophiles. Notre outil de détection repose sur l'identification de ces catégories par des mots-clefs ou des combinaisons de mots-clefs.

2.2 Méthode de validation

Soit D un ensemble de requêtes. Notons P^+ (respectivement P^-) l'ensemble des requêtes pédophiles (respectivement non pédophiles) de D et T^+ (respectivement T^-) le sous-ensemble de D de requêtes étiquetées par notre outil comme pédophiles (respectivement non pédophiles). La figure 2.3 illustre ces notations.

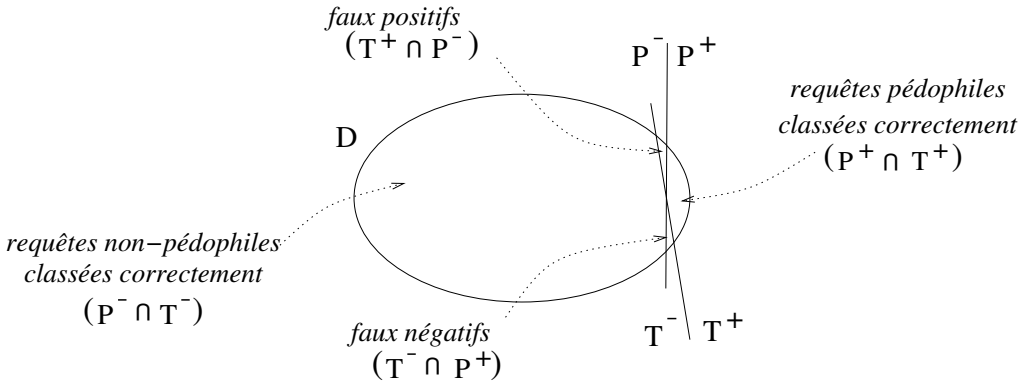


Figure 2.3 – *Illustration de nos notations. L'ellipse représente l'ensemble de toutes les requêtes D. La ligne verticale P^-/P^+ divise D entre les ensembles de requêtes non pédophiles P^- (à gauche) et pédophiles P^+ (à droite). De manière similaire, la ligne T^-/T^+ divise D entre les requêtes étiquetées comme non pédophiles et pédophiles.*

Idéalement, $T^+ = P^+$, dans le cas où toutes les requêtes sont classées correctement. En pratique, certaines requêtes pédophiles ne sont pas classées comme telles par notre outil, c'est-à-dire que $T^- \cap P^+$ est non vide. Ces requêtes sont des faux négatifs, l'outil fournissant de façon erronée une réponse négative pour ces requêtes. Nous définissons réciproquement les faux positifs.

Les nombres de faux positifs et de faux négatifs permettent de décrire la performance de notre outil sur l'ensemble D. Remarquons cependant que leurs valeurs dépendent fortement des tailles de P^+ et P^-. Dans notre situation, nous nous attendons à ce que P^+ soit très nettement plus petit que P^-, c'est-à-dire que la plupart des requêtes de D ne sont pas pédophiles. Dans le cas extrême où l'outil ne donnerait que des réponses négatives, ses réponses ne seraient erronées que sur un faible nombre de requêtes, car il y aurait peu de faux négatifs. Néanmoins, cela occulterait le fait qu'aucune requête pédophile n'ait été correctement détectée.

Pour évaluer la performance de notre outil dans ce contexte, nous faisons donc appel à deux notions de taux de faux positifs et de faux négatifs.

Tout d'abord, nous considérons f^- comme étant le taux de faux négatifs (respectivement positifs pour f^+) quand toutes les requêtes évaluées sont pédophiles (respectivement non pédophiles).

$$f^- = \frac{|T^- \cap P^+|}{|P^+|} \quad \text{et} \quad f^+ = \frac{|T^+ \cap P^-|}{|P^-|} \ .$$

Nous pouvons obtenir une estimation de f^+ en prenant un échantillon aléatoire X de requêtes de P^-, c'est-à-dire des requêtes non pédophiles aléatoires, puis examiner les résultats de l'étiquetage proposé par l'outil. Construire X est relativement simple : comme la plupart des requêtes sont non pédophiles, nous pouvons tirer aléatoirement des requêtes et éliminer manuellement celles qui sont pédophiles. Tant que X reste petit, le coût reste raisonnable. Néanmoins, la fraction de requêtes étiquetées comme pédophiles par l'outil dans X sera très faible et l'estimation de f^+ ainsi obtenue sera de faible qualité.

Réciproquement, l'estimation de f^- peut être effectuée avec un échantillon aléatoire Y de requêtes de P^+ et en analysant les réponses de l'outil sur Y. Cependant, P^+ étant de taille très réduite, la construction de Y n'est pas aisée.

Afin d'obtenir un échantillon de requêtes pédophiles aléatoires, nous faisons l'hypothèse que les requêtes consécutives des utilisateurs sont thématiquement proches et, en particulier, que les requêtes émises peu de temps avant ou après une requête pédophile ont une probabilité élevée d'être pédophiles (très nettement supérieure à la probabilité d'une requête tirée aléatoirement). Nous introduisons la notion de requêtes $voisines$, précédente et suivante, pour une requête donnée.

Considérons une requête q_i et la requête q_j telle que $j < i, u(q_j) = u(q_i)$ et j est le plus grand entier vérifiant ces conditions. Si q_j précède q_i de moins de deux heures, nous l'appelons *requête précédente* de q_i et la notons $pred(q_i)$. De façon similaire, nous définissons la *requête suivante* comme étant la première requête émise par le même utilisateur que q_i dans un délai de deux heures après q_i. Nous notons cette dernière $succ(q_i)$.

Nous avons choisi le seuil de deux heures en examinant les distributions de temps inter-requête, selon la méthodologie présentée dans [9]. La valeur doit être suffisamment élevée pour que des requêtes voisines existent dans de nombreux cas, mais assez petite pour que les requêtes *précédente* et *suivante* concernent des contenus similaires à ceux de q_i. Deux heures constitue un bon compromis entre ces deux exigences. Remarquons que, comme nous le verrons dans le Chapitre 3, les adresses IP ne suffisent pas à distinguer les utilisateurs. Cependant, ce qui importe vraiment dans cette étape c'est d'obtenir un nombre important de requêtes voisines qui soient pédophiles, ce qui est bien le cas.

Pour une requête donnée q_i, nous notons $N(q_i)$ l'ensemble des requêtes *précédente* et *suivante*. $N(q_i)$, qui peut être vide et contient au plus deux éléments. Nous introduisons aussi l'ensemble $N(S) = \cup_{q_i \in S} N(q_i)$ qui contient les requêtes voisines de toutes les requêtes d'un ensemble S.

Nous espérons que les requêtes de $N(P^+)$, et surtout les requêtes de $N(T^+)$, ont une probabilité très élevée d'être pédophiles. Nous le vérifions plus loin,

voir le tableau 2.3, page 39. Nous pouvons donc considérer $N(T^+) \cap P^+$ comme un ensemble de requêtes pédophiles aléatoires, et l'utiliser pour construire un ensemble X adapté à l'estimation de la valeur de f^-. Comme X ne contient que des requêtes pédophiles, la valeur de f^- est ainsi le nombre de requêtes non détectées comme pédophiles dans X, divisé par la taille de X ($|\frac{N(T^+)\cap P^+\cap T^-|}{|N(T^+)\cap P^+|}$).

Le fait que les requêtes de X soient tirées dans T^+ peut biaiser nos résultats. En effet, la probabilité qu'un utilisateur saisisse une requête que l'outil détecte comme pédophile est plus élevée si l'utilisateur a déjà saisi une requête similaire (contenant les mêmes mots-clefs). La valeur de f^- que nous calculons sous-estime donc la vraie valeur.

Pour évaluer les performances de notre filtre, nous ne pouvons pas obtenir une bonne estimation de f^+, alors que nous pouvons calculer f^-. Nous avons besoin cependant d'un taux de faux positifs et de faux négatifs. Pour dépasser cet obstacle, nous introduisons une seconde notion de taux de faux positifs et négatifs (notés respectivement f'^+ et f'^-), qui évalue la probabilité que l'outil donne une réponse erronée quand il fournit une réponse positive (respectivement négative) :

$$f'^+ = \frac{|T^+ \cap P^-|}{|T^+|} \quad \text{et} \quad f'^- = \frac{|T^- \cap P^+|}{|T^-|} \ .$$

Une estimation de f'^+ s'obtient en tirant aléatoirement un échantillon X de requêtes de T^+ et en évaluant le nombre de faux positifs contenus dans X. Nous pouvons nous attendre à ce que tous les ensembles requis pour ce calcul soient de taille raisonnable (et c'est confirmé par l'expérience : voir section 2.4). Une estimation de f'^+ peut donc facilement être calculée.

Réciproquement, nous pourrions évaluer f'^-, à l'aide d'un ensemble X de requêtes tirées aléatoirement dans T^-, en y recherchant les faux négatifs. Cependant, comme nous nous attendons à ce que les requêtes pédophiles soient très rares, le nombre de faux négatifs risque d'être très faible, ce qui rendrait l'évaluation difficile.

Pour évaluer la qualité de notre outil, nous allons donc utiliser de manière conjointe f'^+ (le taux d'erreurs de l'outil lorsque la réponse donnée est positive) et f^- (le taux de requêtes vraiment pédophiles non classées comme telles par l'outil).

Remarquons que nous pouvons utiliser les métriques classiques de précision et de rappel (voir section 1.3.2). Le calcul de f^- permet d'obtenir le rappel de notre outil et celui de f'^+ permet d'évaluer sa précision :

$$Precision = 1 - f^- \quad \text{et} \quad Rappel = 1 - f'^+ \ .$$

La section suivante fournit les détails pratiques de la construction des différents ensembles nécessaires ainsi que les outils nécessaires à l'évaluation de la nature

des requêtes qu'ils contiennent. Les résultats de cette procédure sont présentés dans la section 2.5.

Afin d'évaluer les performances de notre outil de détection de requêtes pédophiles, nous utilisons des notions de taux de faux positifs et de faux négatifs, ainsi qu'une méthode pour les estimer de manière fiable dans notre contexte.

2.3 Protocole de validation

Afin de pouvoir estimer la qualité de notre outil de détection de requêtes pédophiles selon la procédure décrite dans la section précédente, il faut en pratique identifier des requêtes pédophiles dans des ensembles spécifiques de requêtes. Nous avons fait appel à des experts indépendants qui ont examiné et classé de telles requêtes. En pratique, ce sera notre définition de requêtes pédophile : nous considérons qu'une requête est pédophile si les experts la considèrent comme telle. Nous présentons ici la construction de ces ensembles, la sélection des experts qui nous ont assisté, ainsi que l'interface que nous leur avons fournie.

2.3.1 Construction des échantillons

Pour effectuer la sélection des requêtes à valider, nous nous sommes servis de l'ensemble `data-ed2k2007`. Notons D l'ensemble de toutes les requêtes de cet ensemble et divisons D en trois ensembles (non disjoints) :

- T^- contient les requêtes étiquetées comme non pédophiles par notre outil ;
- T^+ contient les requêtes étiquetées comme pédophiles par notre outil ;
- $N(T^+)$ contient les requêtes voisines des requêtes étiquetées comme pédophiles.

Ces trois ensembles sont simples à construire en utilisant notre outil de détection.

Remarquons que, dans l'ensemble T^+, quelques requêtes ne sont composées que d'un seul mot. Il s'agit alors nécessairement d'un mot qui appartient à la liste *explicit* de mots-clefs qui sont utilisés uniquement dans le contexte pédopornographique (voir section 2.1). Nous savons donc avec certitude que la requête est bien de nature pédophile. Nous pouvons donc accroître l'efficacité de notre procédure de validation en ne soumettant aucune de ces requêtes de longueur 1 à nos experts. Nous notons ainsi T_1^+ l'ensemble de ces requêtes et $T_{>1}^+ = T^+ \backslash T_1^+$ l'ensemble des requêtes de T^+ qui contiennent plus d'un mot.

Nous construisons finalement les ensembles de requêtes nécessaires pour le calcul de f^+ et f^-, en tirant aléatoirement $1\ 000$ requêtes dans chacun des ensembles suivants : T^-, $T^+_{>1}$ et $N(T^+)$. Cela représente donc $3\ 000$ requêtes au total, réparties dans 3 ensembles que nous notons par $\overline{T^-}$, $\overline{T^+_{>1}}$, et $\overline{N(T^+)}$.

Comme l'ensemble $N(T^+)$ recouvre partiellement les autres ensembles, il aurait pu arriver que nos échantillons se recouvrent aussi et nous aurions obtenu moins de $3\ 000$ requêtes au total. Les tailles des échantillons étant cependant petites devant les tailles des ensembles à partir desquels ils sont construits, cela n'est pas arrivé.

2.3.2 Experts

Nous devons maintenant classer les requêtes contenues dans les trois ensembles comme pédophiles ou non pédophiles. Nous avons pour cela fait appel à des experts qui ont effectué une classification manuelle. Le choix des experts eux-mêmes est une étape importante. Bien entendu, une connaissance approfondie de la pédopornographie en ligne est indispensable, avec si possible une spécialisation dans les réseaux P2P. Ce profil est extrêmement rare, non seulement en France mais aussi dans le monde. De telles personnes travaillent auprès des forces de l'ordre et des ministères de différents pays, où il existe des départements dédiés à la lutte contre la pédopornographie (éventuellement en ligne). Elles peuvent aussi être employées par des organisations non-gouvernementales qui œuvrent à cette cause, avec une approche différente. Des consultants en sécurité informatique peuvent également avoir de telles compétences.

Nous avons contacté la principale liste de diffusion mondiale de personnes travaillant sur la cybercriminalité et nous avons trouvé 21 personnes ayant la capacité d'être experts pour notre étude. Ces participants, bénévoles, sont employés par Europol, les ministères français et danois de l'Intérieur, des ONG reconnues (telles que *National Center for Missing & Exploited Children*, *Nobody's Children Foundation*, *Action Innocence Monaco* et le *International Association of Internet Hotlines*) ; d'autres encore sont consultants en sécurité. Nous présentons plus loin une évaluation de nos experts, pour nous assurer de leurs compétences (voir section 2.4.1).

2.3.3 Interface

Afin de faciliter la participation des experts à l'étiquetage, nous avons mis en place un site Web dont une capture d'écran est présentée figure 2.4. Les $3\ 000$ requêtes étaient présentées dans un ordre aléatoire, différent pour chaque participant, afin de réduire un éventuel biais lié à l'ordre des requêtes (qui aurait pu exister si nous avions par exemple présenté toutes les requêtes étiquetées comme

pédophiles par l'outil, puis toutes celles étiquetées comme non pédophiles). Il était possible pour les experts de ne pas examiner toutes les requêtes, de façon à ce que ceux qui disposaient de peu de temps puissent tout de même contribuer.

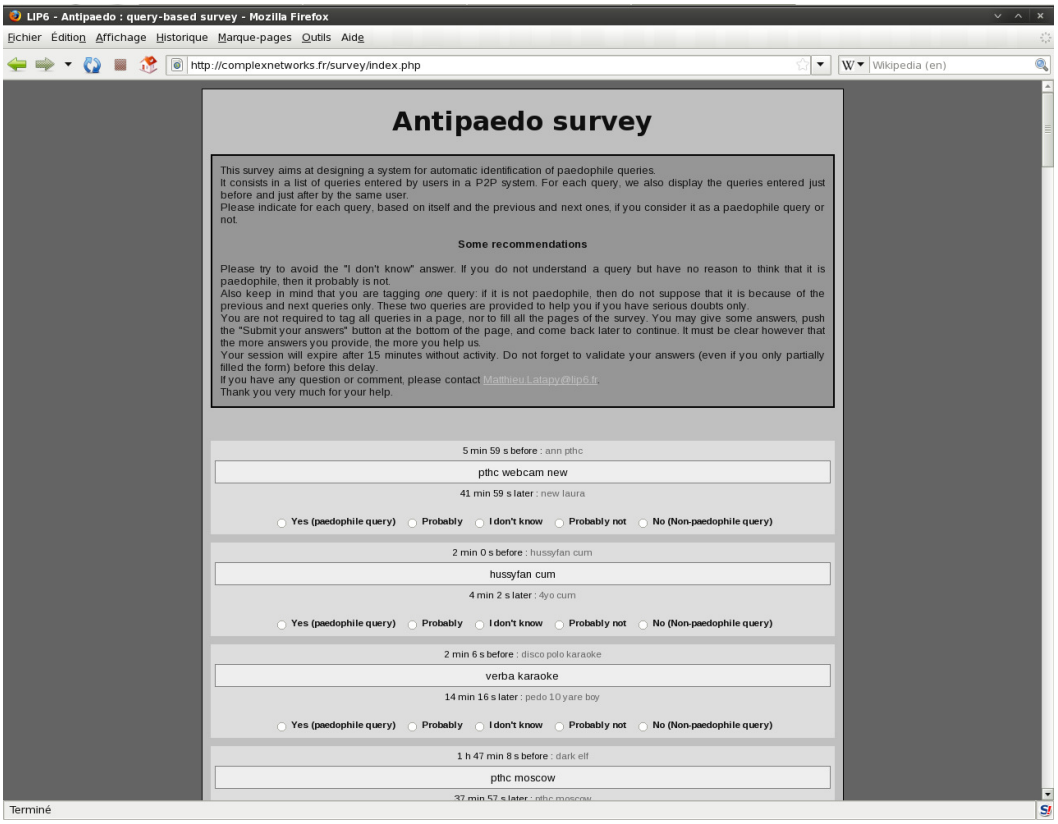


Figure 2.4 – *Capture d'écran du site Web mis en place pour la classification des requêtes par les experts.*

Le site, en anglais, proposait 5 réponses possibles quant à la nature de chaque requête :

- *paedophile* (pédophile),
- *probably paedophile* (probablement pédophile),
- *probably not paedophile* (probablement non pédophile),
- *not paedophile* (non pédophile),
- *I dont know* (je ne sais pas).

Le site affichait également, si elles existaient dans les données, les requêtes précédente et suivante (dont nous avons introduit les définitions dans la section 2.2), afin d'aider l'expert à décider de la nature de la requête.

Nous avons construit des sous-ensembles de requêtes particuliers pour valider notre outil. Nous avons sélectionné 21 experts indépendants qui ont effectué une classification via un site Web que nous avons mis à leur disposition.

2.4 Résultats fournis par les experts

Les réponses que nos experts ont fournies sont présentées dans le tableau 2.1. Chacun d'entre eux a étiqueté plus de 300 requêtes (soit 10% du total) et 12 en ont évalué plus de 2 000.

paedo.	*prob. paedo.*	*don't know*	*prob. not*	*not paedo.*	total	pertinence (%)
1530	149	25	66	1230	3000	99,6
1381	247	125	580	667	3000	98,5
1679	89	2	113	1117	3000	99,2
1603	201	99	174	923	3000	99,3
1598	5	15	1	1381	3000	98,9
128	81	1	26	124	360	96,3
216	154	0	142	132	644	98,6
1624	126	16	165	581	2512	99,7
351	16	2	16	27	412	99,6
647	119	71	40	439	1316	98,9
1174	111	20	64	789	2158	99,3
335	17	1	70	166	589	97,1
641	383	4	112	753	1893	96,6
1071	546	2	453	928	3000	87,3
1554	197	28	327	894	3000	98,2
305	270	24	89	181	869	98,3
371	1017	496	570	546	3000	95,7
976	936	405	594	89	3000	95,5
344	12	10	70	156	592	99,0
845	139	323	175	182	1664	98,1
1506	120	6	25	393	2050	98,3

Tableau 2.1 – *Résultats de la validation par chacun de nos experts. Chaque ligne correspond à un participant et donne le détail des différentes réponses fournies, le nombre total de réponses et la pertinence de l'expert.*

2.4.1 Sélection des experts

En dépit de nos efforts pour sélectionner des experts *a priori* compétents pour notre validation, certains auraient pu ne pas être suffisamment familiers avec le contexte très spécifique de notre étude (le P2P) et qu'ils auraient risqué de compromettre le processus de validation avec des réponses erronées. Pour repérer ces cas éventuels, nous avons examiné les réponses de chaque participant pour les requêtes qui contenaient au moins un mot-clef de la liste *explicit*.

Parmi les 3 000 requêtes de l'échantillon, 1 003 étaient des requêtes de ce type. Nous appelons « pertinence » le pourcentage de ces requêtes pour lesquelles l'expert a fourni la réponse « pédophile » ou « probablement pédophile ». Pour tous les experts sauf un, ce pourcentage est supérieur à 95%, ce qui est satisfaisant et montre que l'expert est familier avec ces mots-clefs. Le dernier expert a une pertinence de 87,3% et nous l'avons pros en compte pour nos évaluations. Cette pertinence figure dans la colonne la plus à droite du tableau 2.1.

Si un participant avait étiqueté comme « pédophile » toutes les requêtes qui lui étaient présentées, il aurait eu une pertinence de 100%. La lecture détaillée du tableau 2.1 nous permet de constater que seuls 3 experts ont étiqueté une majorité de requêtes comme pédophiles, la plupart des autres ayant fourni des réponses équilibrées entre les différentes possibilités. L'examen des réponses des 3 experts en question montre qu'ils se sont en fait concentrés sur les requêtes pédophiles (voir par exemple la dernière ligne du tableau), délaissant les autres, ce qui ne fausse donc pas les résultats. Nous avons donc conservé ces trois experts.

Finalement, nous avons recueilli 42 059 réponses provenant de 21 experts, avec plus de 300 réponses pour chacun, ce qui donne en moyenne un peu plus de 14 avis d'experts par requête.

	échantillon		
	$\overline{T^-}$	$\overline{T^+_{>1}}$	$\overline{N(T^+)}$
paedophile	63	11 530	8 286
probably paedophile	237	2 303	2 395
I don't know	1 009	208	458
probably not paedophile	2 294	336	1 242
not paedophile	9 537	241	1 920
Total	13 140	14 618	14 301

Tableau 2.2 – *Nombre de votes dans chaque catégorie.*

La distribution des réponses pour chaque catégorie est présentée dans le tableau 2.2. Cela correspond à ce que nous pouvons attendre si l'outil étiquette correctement la plupart des requêtes et si notre hypothèse que $\overline{N(T^+)}$ contient de nombreuses requêtes pédophiles est valide. Nous examinons cela plus en détail dans ce qui va suivre.

2.4.2 Classification des requêtes

Pour chaque requête q_i proposée à l'examen de nos experts, nous notons q_i^{++} la fraction d'experts qui l'ont étiquetée comme « pédophile », parmi ceux qui ont fourni une réponse pour q_i, et q_i^+ la fraction d'experts qui l'ont étiquetée comme « pédophile » ou « probablement pédophile ». Nous définissons q_i^{--} et q_i^- de manière similaire.

Remarquons que, comme il y a parfois des réponses « je ne sais pas », on a $q_i^+ + q_i^- < 1$ en général pour une requête donnée. Notons aussi que pour chaque requête, nous avons $q_i^+ \geq q_i^{++}$ et $q_i^- \geq q_i^{--}$.

Afin de classer les requêtes selon les réponses des experts, il est souhaitable que chaque requête ait soit une valeur élevée de q_i^+ (respectivement q_i^{++}) soit une valeur élevée de q_i^- (respectivement q_i^{--}), mais pas simultanément deux valeurs

élevées ou deux valeurs faibles (ce qui signifierait qu'il n'y a pas de consensus entre les experts sur la nature de q_i). La figure 2.5 présente la distribution de la différence entre q_i^+ et q_i^- et entre q_i^{++} et q_i^{--} pour toutes les requêtes. Ces courbes indiquent une croissance lente pour les valeurs faibles de l'axe horizontal, ce qui signifie que très peu de requêtes présentent une différence faible. En revanche, pour beaucoup de requêtes, la différence est importante : pour 1 305 requêtes, la différence est supérieure à 0,8 dans le cas de q_i^{++} et q_i^{--}. Elle est supérieure à 0,8 pour 2 308 requêtes dans le cas de q_i^+ et q_i^-.

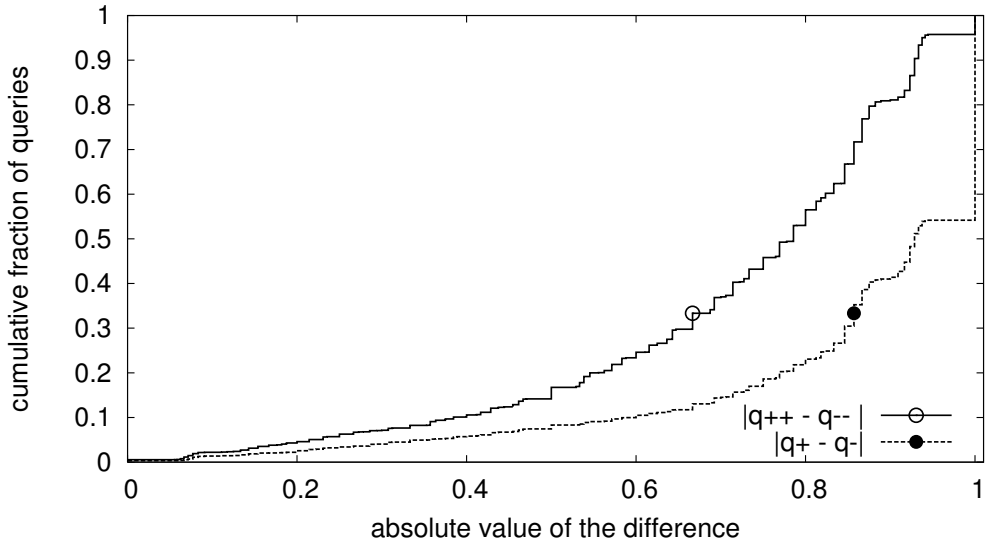


Figure 2.5 – *Distribution cumulative (CDF) de la valeur absolue de la différence entre q_i^{++} et q_i^{--} et entre q_i^+ et q_i^- pour chaque requête : un point de coordonnées (x, y) indique qu'une fraction y de requêtes ont une différence inférieure à x.*

Seules 31 requêtes présentent une différence $|q_i^+ - q_i^-|$ inférieure à $0,1$, ce qui est déjà significatif[2]. Nous choisissons finalement de considérer une requête comme pédophile si $q_i^+ - q_i^- > 0,1$. Le classement final des requêtes figure dans le tableau 2.3.

	échantillon		
	$\overline{T^-}$	$\overline{T^+_{>1}}$	$\overline{N(T^+)}$
requêtes pédophiles	1	985	754
requêtes non pédophiles	999	15	246

Tableau 2.3 – *Nombre de requêtes étiquetées comme pédophiles ou non pédophiles par les experts, pour chaque échantillon.*

2. Ces requêtes pour lesquelles les avis des experts sont divergents sont principalement liées à de la pornographie classique, éventuellement violente

Nous avons vérifié la qualité des réponses fournies par les experts. Nous avons finalement obtenu une caractérisation de chacune des requêtes présentes dans nos échantillons, indiquant leur nature pédophile (ou non), en agrégeant les réponses des experts.

2.5 Résultats de la validation

Grâce aux résultats de la validation et aux expressions présentées dans la section 2.2, nous pouvons maintenant calculer les taux de faux positifs et de faux négatifs de notre outil de détection et ainsi en évaluer la qualité.

Remarquons tout d'abord que, comme attendu, le nombre de requêtes pédophiles dans l'ensemble des requêtes étiquetées comme non pédophiles est très bas : $|\overline{T^-} \cap P^+| = 1$. Il en résulte qu'estimer

$$f'^- = \frac{|T^- \cap P^+|}{|T^-|}$$

avec

$$\frac{|\overline{T^-} \cap P^+|}{|\overline{T^-}|} = \frac{1}{1\,000}$$

donnerait un résultat peu significatif.

L'estimation de f'^+ est de nettement meilleure qualité. Elle repose sur l'expression suivante

$$\begin{aligned} f'^+ &= \frac{|T^+ \cap P^-|}{|T^+|} \\ &= \frac{|T_1^+ \cap P^-| + |T_{>1}^+ \cap P^-|}{|T^+|} \\ &= \frac{|T_{>1}^+ \cap P^-|}{|T^+|} \end{aligned}$$

(puisque $T_1^+ \cap P^- = \emptyset$, toutes les requêtes de T_1^+ étant pédophiles, voir la section 2.3).

Une estimation de $|T_{>1}^+ \cap P^-|$ est fournie par $|\overline{T_{>1}^+} \cap P^-| \cdot \frac{|T_{>1}^+|}{|\overline{T_{>1}^+}|}$. Il y a 207 340 requêtes étiquetées comme pédophiles dans `data-ed2k2007`, dont 14 795 de longueur 1, ce qui donne

$$\begin{aligned} f'^+ &\sim \frac{|\overline{T_{>1}^+} \cap P^-|}{|T^+|} \cdot \frac{|T_{>1}^+|}{|\overline{T_{>1}^+}|} \\ &= \frac{15}{207\,340} \cdot \frac{192\,545}{1\,000} \\ &\sim 1,39\,\% \ . \end{aligned}$$

Cette estimation est de bonne qualité non seulement car $|\overline{T^+_{>1}} \cap P^-| = 15$ est significatif, mais aussi parce que nous l'avons calculée à l'aide d'un échantillon de requêtes de $T^+_{>1}$ qui est nettement plus petit (500 fois) que T^-, qui intervient dans le calcul de f'^-.

Réciproquement, les résultats de la procédure de validation confirment qu'estimer $f^+ = \frac{|T^+ \cap P^-|}{|P^-|}$ avec nos données serait de mauvaise qualité, puisque $|T^+ \cap P^-|$ est petit (il y a peu de requêtes pédophiles), tout comme l'échantillon devant la taille de P^-.

Il est possible d'évaluer plus précisément f^-

$$\begin{aligned} f^- &= \frac{|T^- \cap P^+|}{|P^+|} \\ &\gtrsim \frac{|T^- \cap (\overline{N(T^+)} \cap P^+)|}{|\overline{N(T^+)} \cap P^+|} \\ &= \frac{185}{754} \\ &\sim 24,5\,\% \ . \end{aligned}$$

Cette valeur sous-estime cependant légèrement la vraie valeur, puisque nous avons utilisé des requêtes « proches » lors de la validation, en lieu et place de requêtes pédophiles aléatoires.

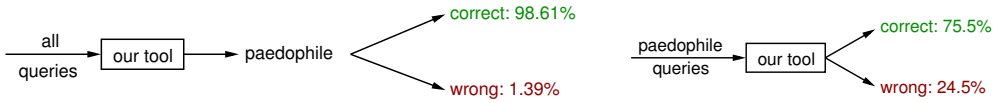


Figure 2.6 – *Taux d'erreurs de notre outil. Dans le premier cas, il s'agit du taux d'erreur quand l'outil déclare une requête comme pédophile. Dans le second cas, c'est la fraction de requêtes pédophiles que notre outil omet de détecter.*

À l'aide de la classification réalisée par les experts, nous obtenons les taux de faux positifs et faux négatifs de notre outil, qui sont nécessaires à l'estimation de la fraction de requêtes pédophiles présentes dans nos jeux de données (voir figure 2.6).

2.6 Fraction de requêtes pédophiles

Dans cette section, nous estimons la fraction de requêtes pédophiles $\frac{|P^+|}{|D|}$ dans deux jeux de données D, `data-ed2k2007` et `data-ed2k2009`.

Comme nous l'avons dit précédemment, compte tenu de la faible taille de l'ensemble P^+ des requêtes pédophiles, nous ne pouvons procéder en tirant aléatoirement un échantillon de requêtes de D et en les soumettant à des experts pour décider si elles sont ou non pédophiles : l'obtention d'un nombre significatif de requêtes pédophiles nécessiterait des ensembles de taille beaucoup trop grande pour qu'une inspection manuelle soit possible.

Nous utilisons donc l'outil de détection automatique élaboré dans la section 2.1 et pour lequel nous connaissons désormais les taux d'erreur. Nous estimons d'abord de manière rigoureuse la fraction de requêtes étiquetées comme pédophiles par notre outil, avant d'inférer la valeur de $\frac{|P^+|}{|D|}$.

2.6.1 Fraction de requêtes étiquetées comme pédophiles

L'outil de détection automatique divise D en deux ensembles disjoints, T^+, les requêtes étiquetées pédophiles, et T^-, les requêtes étiquetées non pédophiles. La valeur de $\frac{|T^+|}{|D|}$ pour nos deux jeux de données `data-ed2k2007` et `data-ed2k2009` est proche de $0,19\,\%$.

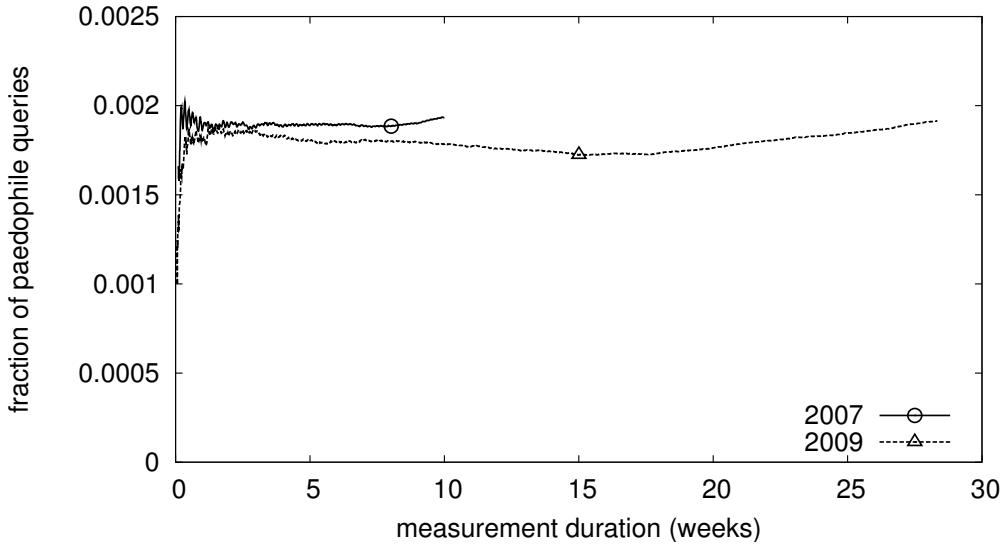


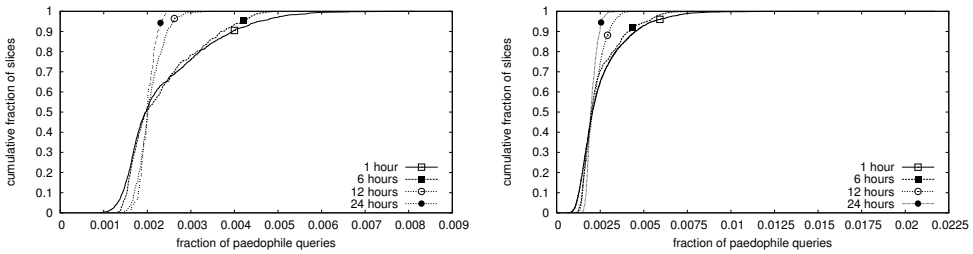
Figure 2.7 – *Fraction de requêtes détectées comme pédophiles dans les jeux de données `data-ed2k2007` et `data-ed2k2009`, en fonction de la durée de la mesure.*

Nous vérifions la robustesse de ces calculs en regardant tout d'abord si la durée de mesure est suffisante, en étudiant la courbe de la fraction de requêtes étiquetées

comme pédophiles en fonction de la durée de mesure (voir la figure 2.7). Celle-ci montre clairement que la fraction de requêtes pédophiles converge rapidement vers une valeur raisonnablement stable, légèrement inférieure à 0,2%.

Pour aller un peu plus loin, nous traçons sur la figure 2.8 la distribution cumulative des requêtes étiquetées comme pédophiles dans des fenêtres de différentes durées : 1 heure, 6 heures, 12 heures et 24 heures. Nous constatons qu'il existe une notion de normalité pour chaque taille de fenêtre de temps, et que la valeur médiane est relativement indépendante de la taille, proche de 0,2%, en accord avec ce qui a été calculé précédemment.

Les serveurs étant sujets à diverses interruptions de service au cours de mesures de longue durée, remarquons qu'il arrive que certaines fenêtres de temps ne contiennent pas ou peu de requêtes : nous ne les avons pas prises en compte pour ces calculs.



(a) `data-ed2k2007` (b) `data-ed2k2009`

FIGURE 2.8 – *Distribution cumulative des fraction de requêtes pédophiles observées dans des fenêtres de temps de 1, 6, 12 et 24 heures (chaque courbe correspond à une taille de fenêtre), pour les jeux de données `data-ed2k2007` (a) et `data-ed2k2009` (b). Un point de coordonnées (x, y) indique que $y\%$ de fenêtres contiennent moins de $x\%$ de requêtes pédophiles. Une augmentation verticale rapide autour de x indique donc que de nombreuses pfenêtres de temps contiennent autour de $x\%$ de requêtes pédophiles.*

Nous concluons finalement que la fraction de requêtes étiquetées comme pédophiles par notre outil est de $\frac{|T^+|}{|D|} \sim 0{,}2\%$ dans chaque jeu de données.

2.6.2 Fraction de requêtes pédophiles

Nous avons obtenu dans la section 2.5 des estimations fiables pour f^- et f'^+. Nous pouvons alors inférer la taille de P^+ à partir de ces taux. Nous utilisons l'expression suivante

$$\begin{aligned} |P^+| &= |P^+ \cap T^+| + |P^+ \cap T^-| \\ &= |T^+|(1 - f'^+) + |P^+|f^- \end{aligned}$$

et nous obtenons

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

En utilisant $f^- \gtrsim 24,5\,\%$ et $f'^+ \sim 1,39\,\%$ (section 2.5), nous obtenons

$$\frac{|P^+|}{|D|} \gtrsim 0,25\,\%$$

pour les deux jeux de données.

Nous utilisons le signe $\gtrsim$ car nous avons probablement légèrement sous-estimé la valeur de f^-. Remarquons que si nous avions pris $f^- \sim 50\,\%$ (ce qui serait une sur-estimation probablement très importante de la valeur réelle de f^-), nous obtiendrions une valeur de $\frac{|P^+|}{|D|}$ de $0,38\,\%$. Il est donc raisonnable de penser que la valeur réelle est proche de la valeur calculée.

En d'autres termes, au moins une requête sur 400 est de nature pédophile dans les ensembles considérés, ce qui correspond à environ une requête toutes les 33 secondes.

> Nous obtenons une estimation fiable de la fraction de requêtes pédophiles dans un grand jeu de données provenant du P2P. La valeur est de l'ordre de 0,25%.

2.7 Conclusion

Dans ce chapitre, nous avons présenté les fondations de notre étude quantitative de l'activité pédophile dans le P2P, avec le développement d'un outil de détection de requêtes pédophiles.

Le caractère spécifique et caché de l'activité pédophile était une difficulté majeure, que nous avons surmontée en adoptant tout d'abord une approche manuelle, reposant sur des connaissances acquises sur le terrain par des spécialistes du domaine. Nous avons ensuite évalué les performances de cet outil, afin de nous assurer de sa qualité et de fiabiliser les résultats ultérieurs. Nous avons pour cela fait appel à des experts indépendants qui ont classé des requêtes sur une interface Web spécialement conçue à cet effet. Nous avons pu obtenir les taux de faux positifs et de faux négatifs, indispensables pour obtenir l'estimation de la quantité de requêtes pédophiles dans les jeux de données dont nous disposons.

Ce résultat est important car c'est la première fois que l'activité pédophile est mesurée avec une telle précision et sur un système d'une telle échelle. Le chiffre

de 3 requêtes pour 1 000 a des conséquences importantes sur la protection de l'enfance, mais aussi sur la législation relative à l'Internet et sur la connaissance de la pédophilie en général.

Il est cependant indispensable d'aller plus loin. Notre étude s'en tient pour le moment au réseau *eDonkey*, sur deux périodes précises et relativement courtes. Si les similarités entre les résultats sur les deux jeux de données donnent une certaine fiabilité à notre estimation, une étude sur une durée plus longue montrera certainement des phénomènes intéressants. Nous examinons cela dans le chapitre 4. De même, de nombreux facteurs influent certainement sur l'intensité de l'activité pédophile et notamment le type de réseau P2P. C'est pourquoi nous comparerons *eDonkey* et *KAD* dans le chapitre 5.

Le travail présenté offre des perspectives de recherche intéressantes. Le fonctionnement de l'outil pourrait être amélioré. Par exemple, de la même façon que nous présentions les requêtes *précédente* et *suivante* aux experts pour les aider à trancher la nature d'une requête, l'outil pourrait utiliser ces requêtes *précédente* et *suivante* pour effectuer sa classification. Il serait également intéressant d'accroître le nombre de langues que l'outil est capable de traiter (il est pour l'instant limité aux langues les plus parlées en Europe occidentale). La constitution d'ensembles de requêtes pédophiles permet aussi d'envisager d'utiliser des techniques d'apprentissage pour affiner la classification des requêtes. Ces ensembles ouvrent aussi la voie à des recherches pour mieux comprendre la pédophilie, en examinant par exemple en détail quels sont les autres sujets auxquels s'intéressent les pédophiles ou comment évolue leur intérêt pour les contenus pédopornographiques au fil du temps.

Chapitre 3

Utilisateurs pédophiles

La fraction de *requêtes pédophiles* que nous avons obtenue au chapitre 2 est un indicateur important pour connaître l'ampleur de l'activité pédophile dans les réseaux P2P. Quantifier les *utilisateurs* qui participent à ces échanges est naturellement un indicateur encore plus important mais plus difficile à obtenir. Nous étudions cette problématique dans ce chapitre. Nous définissons un *utilisateur pédophile* comme un utilisateur ayant effectué *au moins une* requête pédophile.

Quantifier les utilisateurs pédophiles passe *a priori* par deux grandes étapes. Il s'agit d'abord de regrouper les requêtes émanant d'un même utilisateur, c'est-à-dire de reconnaître un utilisateur. Nous étudions cette question en section 3.1. Il faut ensuite décider pour chaque utilisateur s'il doit être considéré comme pédophile ou non. Nous considérons un utilisateur comme pédophile dès lors qu'il a soumis au moins une requête pédophile et utiliserons l'outil et la méthodologie du chapitre 2 pour détecter de telles requêtes. Ceci nous permettra d'obtenir une quantification rigoureuse des utilisateurs pédophiles en section 3.2.

3.1 Différentes notions d'utilisateurs

Identifier un utilisateur, c'est être capable de repérer un individu et de le reconnaître sans ambiguïté. Dans notre contexte, c'est un défi en soi [10, 69]. À un instant donné, une machine donnée connectée à l'Internet est identifiée par une adresse IP. Il est cependant impossible, en général et à large échelle, de détecter plus tard que cette machine utilise une adresse IP différente ou qu'une autre machine utilise maintenant cette adresse IP. Un individu peut par ailleurs disposer de plusieurs machines, et une même machine peut être utilisée par plusieurs personnes, ce qui rend l'identification d'autant plus délicate. Plus précisément, il arrive souvent que :

- plusieurs ordinateurs d'un même réseau local soient connectés à l'Internet derrière un routeur qui effectue une « traduction d'adresses réseau » (ou *Network Address Translation* ou « NAT »), un mécanisme qui fait correspondre à plusieurs adresses IP privées non routables (et non uniques sur l'Internet) une adresse publique routable. Ainsi, tous les ordinateurs connectés depuis ce sous-réseau au serveur *eDonkey* sont vus avec la même adresse IP ; la machine proposant le NAT se charge de distribuer le trafic arrivant de l'extérieur à chacun, à l'aide de ports de communication ;

- les fournisseurs d'accès à Internet (FAI) peuvent attribuer dynamiquement les adresses IP à leurs clients : une machine se connectant à des moments différents pourra avoir une adresse IP différente, et différentes machines seront susceptibles d'avoir la même IP (à des moments distincts) ;
- dans des lieux publics disposant d'un accès Internet, ou à leur domicile pour les différents membres d'un foyer, plusieurs utilisateurs peuvent se succéder sur un poste et partager la même adresse ;
- réciproquement, une personne peut utiliser plusieurs machines (à la maison, au travail, etc.).

L'*identification* des utilisateurs à grande échelle est donc très délicate, voire impossible en pratique. Cependant, dans le contexte de notre étude, nous pouvons nous contenter d'un niveau de précision plus faible : nous cherchons à *dénombrer* des utilisateurs et nous devons donc avant tout nous assurer que nous ne mélangeons pas les requêtes de plusieurs d'entre eux, ou que les erreurs sont négligeables.

En effet, si nous interprétons une série de requêtes provenant de plusieurs utilisateurs comme si elle venait en fait d'un seul et, puisque nous considérons un utilisateur comme pédophile dès lors qu'il a effectué une requête de ce type, il suffit qu'un des utilisateurs « mélangés » ait effectué une requête pédophile pour que toutes les requêtes soient considérées comme provenant d'un *utilisateur pédophile*. Puisque la fraction d'utilisateurs pédophiles est très faible, cependant, il arrive très rarement que deux utilisateurs pédophiles soient mélangés. En considérant que les requêtes de plusieurs utilisateurs appartiennent à un seul, on a donc tendance à sous-estimer le nombre total d'utilisateurs. Comme le nombre d'utilisateurs pédophiles reste sensiblement le même, cela conduit à surestimer la fraction d'utilisateurs pédophiles dans les jeux de données. Nous appelons ce phénomène *pollution* et nous en détaillons les effets dans la suite.

Dans cette section, nous explorons plusieurs approches pour définir et compter les utilisateurs. Nous examinons la définition qui consiste à identifier un utilisateur à son adresse IP seule et constatons que ceci est insuffisant pour distinguer des utilisateurs. Nous montrons ensuite que la prise en compte du port de communication en plus de l'adresse IP constitue une meilleure définition, adaptée à notre contexte. Nous considérons aussi la notion de *session*, définie comme une succession de requêtes provenant de la même adresse IP dans un intervalle de temps réduit.

3.1.1 Adresse IP et port de communication

Nous disposons dans l'ensemble de requêtes `data-ed2k2007` de deux informations relatives aux utilisateurs : l'adresse IP et le port de communication, lequel permet de distinguer des utilisateurs d'un même réseau local avec NAT. Dans le jeu de données `data-ed2k2009`, les ports de communication ne sont pas disponibles.

Dans un premier temps, nous supposons que l'adresse IP seule permet de distinguer des utilisateurs, puis que c'est le couple (IP, port) qui permet de le faire. Nous savons *a priori* que la seconde définition sera plus précise que la première, alors la comparaison des deux peut invalider la première.

La figure 3.1 présente la fraction d'adresses IP et de couples (IP, port) dont la séquence de requête contient au moins une requête pédophile.

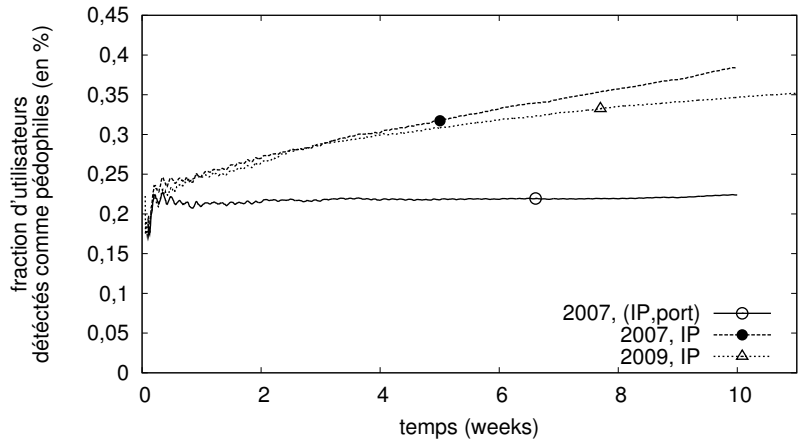


Figure 3.1 – *Fraction d'utilisateurs détectés comme pédophiles en fonction de la durée de la mesure.*

Pour les deux ensembles de requêtes, la fraction d'adresses IP détectées comme pédophiles augmente sensiblement avec la durée de mesure. Cela traduit un phénomène de *pollution* : puisque une adresse donnée peut correspondre à différents utilisateurs et puisqu'une seule requête suffit pour être considéré comme pédophile, la probabilité qu'une adresse soit déclarée comme pédophile augmente avec la durée de mesure (si la rotation des IP était importante, nous pourrions, après un certain temps, nous trouver dans le cas limite où toutes les adresses pourraient être déclarées comme pédophiles). Cela démontre que l'adresse IP seule ne suffit pas.

En revanche, la fraction des couples (adresse IP, port) converge rapidement, de façon très similaire à la fraction des requêtes (voir figure 2.7, page 42). Le phénomène de pollution est cette fois négligeable.

3.1.2 Effet de la durée de la mesure

La figure 3.1 montre qu'augmenter la durée de la mesure conduit à une pollution des adresses IP par les utilisateurs détectés comme pédophiles. Il est alors naturel de penser qu'en réduisant la mesure à des fenêtres de temps plus petites, nous pouvons maîtriser cet effet. Évidemment, en faisant cela, nous travaillons sur

des sous-ensembles de requêtes d'autant plus petits, ce qui fournit des résultats moins fiables. Il y a donc un compromis à trouver.

Nous divisons nos jeux de données en périodes de temps plus courtes et calculons pour chacune la fraction de d'adresses IP et de couples (IP, port) détectés comme pédophiles. La distribution de ces fractions (non présentée ici) est homogène et la moyenne est donc significative. La figure 3.2 présente cette moyenne en fonction de la taille de la fenêtre de temps.

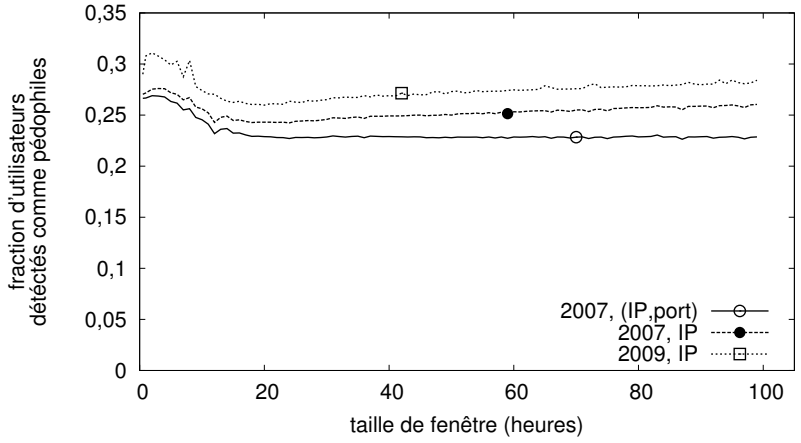


Figure 3.2 – *Fraction d'utilisateurs détectés comme pédophiles en fonction de la taille de la fenêtre temporelle considérée.*

La fraction de couples (IP, port) détectés comme pédophiles pour le jeu de données `data-ed2k2007` fluctue tout d'abord pour les fenêtres de petite taille avant de converger rapidement vers un régime stationnaire, très proche de la fraction totale de couples (IP, port) dans `data-ed2k2007`.

Il est toujours possible qu'un couple (IP, port) soit en fait utilisé par plusieurs personnes, par exemple les membres d'une même famille. Néanmoins, la probabilité que cela arrive est d'autant plus faible que la période de temps considérée est courte. Comme la fraction de couples (IP, port) dans les petites fenêtres de temps est très proche de la fraction globale calculée ci-dessus, nous pouvons donc penser que cette dernière constitue une estimation raisonnable de la fraction d'utilisateurs pédophiles.

Cela est confirmé par la fraction d'adresses IP détectées comme pédophiles pour le même jeu de données en fonction de la durée de la mesure. Après quelques fluctuations initiales, la valeur tombe à un peu moins de $0,25\,\%$, avant de croître linéairement avec la durée de la mesure.

Considérer des fenêtres de temps de taille plus petite semble donc réduire la pollution temporelle. À un instant donné, il y a cependant des utilisateurs qui partagent la même adresse IP (derrière un NAT par exemple). Ils utilisent

cependant des ports de communications différents, ce qui explique que la fraction de couples (IP, ports) soit nécessairement inférieure à la fraction d'adresses IP pédophiles.

La courbe pour la fraction d'IP détectées comme pédophiles dans `data-ed2k2009` a la même allure que celle pour `data-ed2k2007`, mais présente des valeurs supérieures. Cela pourrait s'expliquer par le fait que la fraction d'utilisateurs pédophiles a augmenté entre 2007 et 2009 sur *eDonkey*, mais cela peut aussi provenir de simples différences entre serveurs [1].

3.1.3 Sessions temporelles

Nous définissons une session comme l'ensemble maximal de requêtes provenant de la même adresse IP tel que deux requêtes ne sont pas séparées de plus de δ secondes.

La figure 3.3 illustre notre définition de session. Nous considérons six requêtes, pour une adresse IP donnée. Les quatre premières requêtes q_1, q_2, q_3 et q_4 sont telles que : $|t_2 - t_1| \leq \delta$, $|t_3 - t_2| \leq \delta$, et $|t_4 - t_3| \leq \delta$. En revanche, pour q_4 et q_5, nous avons $|t_5 - t_4| > \delta$: la requête q_5 ne fait pas partie de la première session mais en amorce une nouvelle. Comme $|t_6 - t_5| \leq \delta$, les requêtes q_5 et q_6 sont dans la même session.

Pour un δ donné, l'étude des sessions est un moyen de réduire la pollution temporelle, puisqu'il est probable qu'il y ait un délai entre les requêtes d'utilisateurs faisant successivement usage de la même adresse IP. Il n'y a pas de raison *a priori* pour que les utilisateurs pédophiles fassent davantage de sessions que les autres utilisateurs et nous pouvons donc penser que la fraction d'utilisateurs pédophiles est proche de la fraction de sessions détectées comme pédophiles.

La figure 3.4 présente la fraction de sessions contenant au moins une requête détectée comme pédophile en fonction de δ. Les fractions pour les faibles valeurs de δ sont peu significatives, puisqu'il s'agit des cas où les requêtes d'une même personne sont considérées comme faisant partie de sessions différentes. Pour de plus grandes valeurs de δ, la fraction de sessions détectées comme pédophiles converge très rapidement vers la valeur de la fraction de couples (IP, port) pédophiles détectés dans le jeu de données. Cela confirme que la notion de couple (IP, port) est suffisamment fiable pour quantifier les utilisateurs pédophiles du jeu de données `data-ed2k2007`.

La fraction de sessions détectées comme pédophiles correspondant à des adresses IP seulement est plus élevée, pour la même raison que précédemment : plusieurs personnes peuvent utiliser la même adresse IP, mais pas le même port.

1. Rappelons que les données de `data-ed2k2007` et de `data-ed2k2009` n'ont pas été collectées sur le même serveur.

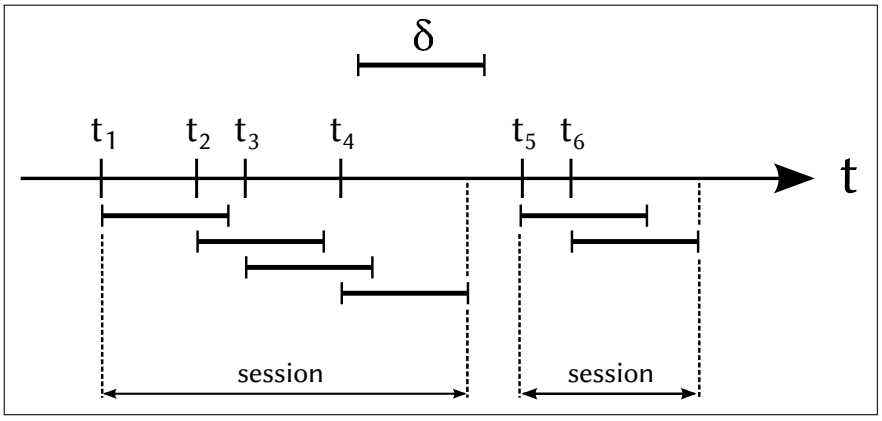


Figure 3.3 – *Illustration de la notion de session que nous utilisons. Les requêtes provenant d'une adresse IP sont disposées sur une ligne de temps. Avec le délai δ considéré, il y a ici deux sessions ; la première comporte quatre requêtes et la seconde session en a deux.*

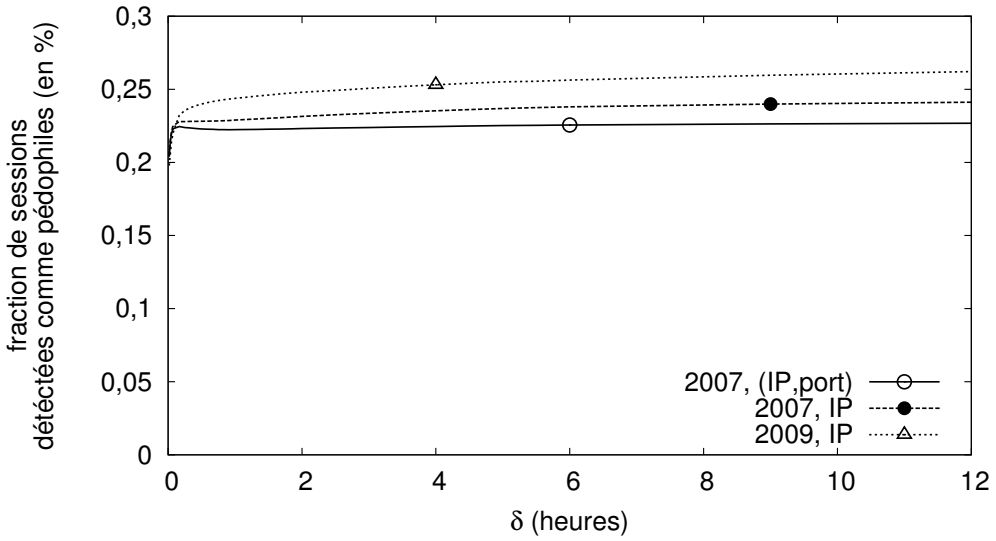


Figure 3.4 – *Fraction de sessions détectées comme pédophiles en fonction de δ, le délai maximal entre deux requêtes consécutives de la même session.*

Nous conjecturons également que la différence entre les deux jeux de données provient d'une augmentation de l'activité pédophile entre 2007 et 2009 (mais il peut s'agir seulement de différences entre les serveurs sur lesquels ont été collectées les requêtes).

Nous proposons plusieurs approches pour compter les utilisateurs à l'aide des informations présentes dans nos ensembles de requêtes. L'adresse IP seule n'est pas suffisante, alors que le couple (adresse IP, port de communication) semble fournir une approximation satisfaisante. Celle-ci peut également être obtenue en faisant varier la durée d'observation ou les sessions temporelles.

3.2 Quantifier les utilisateurs pédophiles

Après avoir vu comment regrouper les requêtes appartenant à un utilisateur donné, nous cherchons dans cette section à estimer la fraction d'utilisateurs pédophiles, dans le jeu de données `data-ed2k2007`. Nous utilisons pour cela l'outil de classement de requêtes que nous avons conçu et validé au chapitre précédent et étudions comment prendre en compte les erreurs de classement afin d'aboutir à une classification rigoureuse des utilisateurs. Rappelons que nous considérons en première approximation qu'un utilisateur est pédophile dès lors qu'il a entré (au moins) une requête pédophile. Utiliser d'autres hypothèses que celle-ci fait partie de nos principales perspectives, voir la section 3.3.

Nous présentons dans un premier temps un calcul simple pour disposer d'une borne inférieure à cette valeur, avant d'estimer les taux de faux positifs et de faux négatifs de notre outil sur les utilisateurs et de pouvoir ainsi obtenir une estimation rigoureuse de la fraction d'utilisateurs pédophiles.

3.2.1 Une borne inférieure

À l'aide des taux de faux positifs et de faux négatifs de notre outil pour les requêtes, qui indiquent le nombre de requêtes que l'outil a mal classées, nous pouvons proposer une borne inférieure de la fraction d'utilisateurs pédophiles.

Comme nous ne connaissons pas précisément quelles requêtes ont été mal classées, nous ne savons pas à quelle fraction d'utilisateurs cela correspond. En supposant qu'un utilisateur donné emploie des termes de recherche similaires dans différentes requêtes, si l'outil se trompe pour l'une des requêtes d'un utilisateur, il a même probablement un risque plus important de se tromper sur d'autres requêtes du même utilisateur.

En utilisant le taux de faux positifs, nous établissons une borne inférieure pour la fraction d'utilisateurs pédophiles. Une fraction f'^{+} des requêtes détectées comme pédophiles par notre outil sont en fait non pédophiles, ce qui représente un nombre

n de requêtes. Il est clair que le nombre d'utilisateurs auxquels correspondent ces requêtes est d'au plus n (il est égal à n si chacune des requêtes mal classées est saisie par un utilisateur différent). Réciproquement, l'outil a manqué un certain nombre de requêtes pédophiles. Si toutes ces requêtes provenaient d'utilisateurs déjà détectés comme pédophiles, alors aucun utilisateur pédophile n'est mal classé par l'outil. Ce dernier détecte $|T^+| = 207\ 340$ requêtes pédophiles dans `data-ed2k2007`, correspondant à 112 712 utilisateurs. Le nombre de requêtes classées par erreur comme pédophiles est $|T^+| \cdot f'^+ = 2\ 882$. Ainsi, le nombre d'utilisateurs pédophiles est d'au moins $112\ 712 - 2\ 882 = 109\ 830$, ce qui conduit à une fraction d'utilisateurs légèrement inférieure à $0,22\%$.

Puisque nous avons précédemment observé que les valeurs des fractions d'adresses IP pédophiles étaient similaires entre les jeux de données `data-ed2k2009` et `data-ed2k2007`, mais plus élevées pour `data-ed2k2009`, nous pouvons penser que cette borne inférieure établie pour `data-ed2k2007` peut s'appliquer pour `data-ed2k2009`.

3.2.2 Taux d'erreur sur les utilisateurs

Nous allons maintenant tenter d'évaluer précisément la fraction d'utilisateurs pédophiles en étudiant les taux de faux positifs et de faux négatifs au niveau des utilisateurs.

Pour un utilisateur qui effectue seulement une requête, les taux d'erreurs de notre outil sont connus : ce sont les mêmes que ceux calculés dans le chapitre 2. Notre outil commet une erreur dans 1,39% des cas quand il classe de tels utilisateurs comme pédophiles et il ne détecte pas 24,5% des utilisateurs pédophiles qui lui sont proposés.

Le cas d'un utilisateur qui effectue plus d'une requête est plus complexe car les erreurs de notre outil peuvent se compenser ou s'additionner. Par exemple, pour un utilisateur qui fait deux requêtes dont une seulement est pédophile, si notre outil se trompe sur chacune d'elles, l'utilisateur est correctement détecté comme pédophile, malgré deux erreurs. Il est donc important d'examiner en détail ce qui se passe concernant les taux d'erreurs de notre outil à la granularité des utilisateurs.

Notons U l'ensemble des utilisateurs. On peut le diviser en U^+ et U^-, qui sont les ensembles des utilisateurs pédophiles et non pédophiles respectivement ou en V^+ et V^- qui sont les ensembles d'utilisateurs *détectés* comme pédophiles et comme non pédophiles respectivement. Ces ensembles sont définis par analogie avec les ensembles P^+, P^- et T^+, T^-, que nous utilisons pour les requêtes (voir la figure 2.3, page 31). Notons également $V(n, k)$ l'ensemble des utilisateurs qui ont

effectués n requêtes dont k sont détectées comme pédophiles et $U(n)$ l'ensemble des utilisateurs ayant fait n requêtes. Rappelons qu'un utilisateur est considéré comme pédophile s'il a effectué au moins une requête pédophile.

À l'aide de notre outil, il est possible d'obtenir facilement $\frac{|V^+|}{|U|}$, la fraction d'utilisateurs détectés comme pédophiles. Nous cherchons ici à connaître la valeur de $\frac{|U^+|}{|U|}$, la vraie fraction d'utilisateurs pédophiles. Nous devons pour cela prendre en compte les erreurs que commet l'outil lorsqu'il considère un utilisateur comme pédophile, ou non pédophile. Dans le chapitre précédent, nous avons introduit les définitions suivantes : une fraction f'^+ des requêtes détectées comme pédophiles n'auraient pas dû l'être et une fraction f'^- des requêtes détectées comme non pédophiles sont en fait pédophiles.

Nous cherchons à repérer les cas où l'outil identifie un utilisateur comme pédophile alors qu'il n'aurait pas dû, et inversement. Nous détaillons ci-dessous ce qu'il advient pour $n = 1$ et $n = 2$, avant d'en venir au cas général.

Pour $n = 1$, quand l'utilisateur a effectué une requête, on peut être dans les cas suivants :

- $k = 0$: la requête (et donc l'utilisateur) n'est pas détecté comme pédophile mais l'est en réalité avec une probabilité f'^-. La probabilité qu'il soit correctement détecté comme non pédophile est de $1 - f'^-$.
- $k = 1$: de manière similaire, l'utilisateur est non pédophile avec une probabilité f'^+ et est correctement détecté comme pédophile avec une probabilité $1 - f'^+$.

Pour $n = 2$, nous pouvons avoir :

- $k = 0$: les deux requêtes étiquetées comme non pédophiles l'ont été avec succès avec une probabilité $(1 - f'^-)^2$. Une erreur a été commise sur les deux à la fois avec une probabilité $(f'^-)^2$. Et une erreur a été commise sur la première (mais pas la deuxième), ou sur la deuxième mais pas la première, avec une probabilité $2f'^-(1 - f'^-)$.
  Un tel utilisateur, étiqueté comme non pédophile par l'outil, l'est en fait avec une probabilité

$$p(u \in U^- \mid u \in V(2,0)) = (1 - f'^-)^2 \ .$$

  La probabilité qu'il soit en fait pédophile est de

$$p(u \in U^+ \mid u \in V(2,0)) = 1 - (1 - f'^-)^2 = 2f'^- - (f'^-)^2 \ .$$

- $k = 1$: une requête est étiquetée comme pédophile, l'autre comme non pédophile. L'outil peut n'avoir commis aucune erreur, avec une probabilité $(1-f'^+)\cdot(1-f'^-)$. Sa réponse est erronée sur la requête pédophile et correcte sur la non pédophile avec une probabilité $f'^+ \cdot (1 - f'^-)$. Inversement, sa réponse est correcte sur la requête pédophile et erronée sur la non pédophile

avec une probabilité $(1 - f'^{+}) \cdot f'^{-}$. Enfin, l'outil a commis deux erreurs avec une probabilité $f'^{+} \cdot f'^{-}$.
L'utilisateur est donc en réalité non pédophile avec une probabilité

$$p(u \in U^- \mid u \in V(2,1)) = f'^{+} \cdot (1 - f'^{-}) \ .$$

– $k = 2$: les deux requêtes de l'utilisateur sont considérées comme pédophiles. Ce classement est correct avec une probabilité $(1 - f'^{+})^2$. Une seule requête est pédophile s'il y a une erreur sur la première ou sur la deuxième requête, avec une probabilité $(1 - f'^{+}) \cdot f'^{+} + f'^{+} \cdot (1 - f'^{+})$. Enfin, l'utilisateur n'avait en fait effectué aucune requête non pédophile avec une probabilité $f'^{+} \cdot f'^{+}$.
Nous avons alors :

$$p(u \in U^- \mid u \in V(2,2)) = 2f'^{+} \ .$$

Plus généralement, pour les utilisateurs vus comme non pédophiles, k a pour valeur 0, quel que soit n : l'outil n'a détecté aucune requête comme pédophile. Le classement est correct si l'utilisateur n'a effectivement soumis aucune requête pédophile, ce qui veut dire que toutes les requêtes ont été correctement classées. Ceci survient avec une probabilité $p(u \in U^- \mid u \in V(n,0)) = (1 - f'^{-})^n$. L'outil se trompe donc sur un utilisateur déclaré non pédophile avec une probabilité de

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

Dans le cas d'un utilisateur détecté comme pédophile, nous avons $k > 0$. L'outil commet une erreur si l'utilisateur est en fait non pédophile, c'est-à-dire qu'il n'a fait aucune requête pédophile. Il y a donc une erreur sur les k requêtes détectées comme pédophiles et une réponse correcte sur les $n - k$ requêtes non pédophiles. Cela arrive avec une probabilité de

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

Nous avons donc obtenu les taux d'erreurs sur les utilisateurs, en fonction du nombre n de requêtes qu'ils ont effectuées et du nombre k de requêtes détectées comme pédophiles par notre outil.

3.2.3 Fraction d'utilisateurs pédophiles

Nous allons alors pouvoir estimer la fraction d'utilisateurs pédophiles présents dans le jeu de données `data-ed2k2007`, c'est-à-dire $\frac{|U^+|}{|U|}$. Nous savons que $U^+ = \bigcup_n U(n)$ et que $|U(n)| = \sum_{k=0}^{n} |V(n,k)|$. Notons N le nombre maximal de requêtes d'un utilisateur dans notre jeu de données. Dans le cas de `data-ed2k2007`,

N vaut 17 756. Remarquons cependant, comme le montre la figure 3.5, que le nombre de requêtes par utilisateur est en général assez faible : parmi les 50 341 798 utilisateurs de l'ensemble, plus de 63% d'entre eux n'effectuent qu'une seule requête en 10 semaines, et plus de 98,5% en font moins de 10.

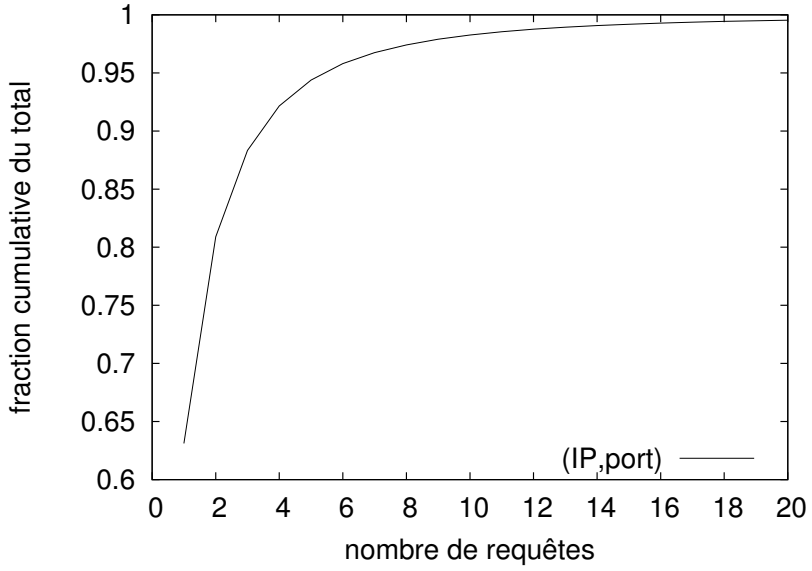


FIGURE 3.5 – *Distribution cumulative du nombre de requêtes par couple (IP,port) dans l'ensemble* `data-ed2k2007`*, restreinte aux nombres de requêtes entre 1 et 20.*

Nous allons calculer les nombres d'utilisateurs pédophiles pour chaque ensemble $U(n)$. Nous savons que $|U^+| = |U^+ \cap V^+| + |U^+ \cap V^-|$, ce qui nous permet d'avoir :

$$|U^+ \cap V^+| \;=\; \sum_{n=1}^{N}\sum_{k=1}^{n}(1-(f'^+)^k(1-f'^-)^{n-k})\cdot|V(n,k)| \;, \qquad (3.1)$$

et

$$|U^+ \cap V^-| = \sum_{n=1}^{N}(1-(1-f'^-)^n)\cdot|V(n,0)| \;.$$

Les expressions ci-dessus dépendent de la valeur de f'^-, que nous n'avons pas calculée précédemment. Néanmoins, nous pouvons en proposer une approximation. Avec les notations du chapitre 2, nous avons en effet :

$$\begin{aligned} f'^- &= \frac{|T^- \cap P^+|}{|T^-|} \\ &= f^- \cdot \frac{|P^+|}{|T^-|} \;. \end{aligned}$$

Avec les valeurs numériques précédemment calculées, nous obtenons la valeur de f'^{-} :

$$f'^{-} = 0,245 \cdot \frac{0,0025 \cdot 107\,226\,021}{107\,226\,021 - 207\,340} = 0,06\% \ .$$

Rappelons que la valeur de f'^{+} était de :

$$f'^{+} = 1,39\% \ .$$

En effectuant la somme pour tous les ensembles $U(n)$ (voir la formule (3.1)), nous obtenons finalement la valeur de $\frac{|U^+|}{|U|}$ recherchée :

$$\frac{|U^+|}{|U|} = 0,2217\% \ .$$

Celle-ci est légèrement au-dessus de la borne inférieure que nous avons obtenue précédemment (voir section 3.2.1).

Après en avoir établi une borne inférieure, nous déterminons la fraction d'utilisateurs pédophiles dans nos données en tenant compte des erreurs de classification des requêtes. Nous estimons que la fraction d'utilisateurs pédophiles dans `data-ed2k2007` est de 0,22%, soit environ un sur 450.

3.3 Conclusion

Dans ce chapitre, nous avons quantifié les utilisateurs qui participent à l'activité pédophile dans un grand système P2P. Nous avons d'abord considéré plusieurs manières de distinguer les requêtes de différents utilisateurs. L'adresse IP complétée par le port de communication, de même que la variation de la durée de mesure et les sessions temporelles permettent d'obtenir des résultats satisfaisants. Nous avons ensuite pris en compte les erreurs de détection de notre outil et estimé la proportion d'utilisateurs pédophiles dans `data-ed2k2007`, proche de 0,22%. C'est la première fois qu'une étude de ce type est menée et elle permet de disposer d'un chiffre-clef pour la connaissance de l'activité pédophile dans le P2P.

Les travaux que nous avons présentés dans ce chapitre ouvrent plusieurs perspectives de travail. Tout d'abord, nous nous sommes servis des couples (IP, port) pour estimer la proportion d'utilisateurs pédophiles, sur `data-ed2k2007`. Nos

résultats doivent être corroborés par une étude avec les autres méthodes pour distinguer des utilisateurs, sur le jeu de données `data-ed2k2007`, mais aussi sur les autres ensembles.

Nous n'avons pas exploité les catégories de requêtes pédophiles définies dans le chapitre précédent. Pourtant, il est possible que certains utilisateurs effectuent des requêtes dans seulement l'une ou l'autre de ces catégories. Une telle étude permettrait de comprendre plus finement les intentions de recherche des utilisateurs pédophiles. De même, examiner la dynamique des requêtes à l'aide de l'information temporelle des requêtes apporte un éclairage supplémentaire sur l'activité pédophile dans le P2P, comme nous le verrons dans le chapitre suivant.

Une autre perspective intéressante consiste à utiliser des hypothèses différentes pour décider qu'un utilisateur est pédophile. Nous avons considéré qu'une seule requête de cette nature suffisait. Il serait certainement pertinent d'utiliser des seuils différents. Ceux-ci pourraient être absolus (par exemple : « au moins deux requêtes pédophiles »), ou relatifs au nombre de requêtes soumises par l'utilisateur (par exemple : « au moins $\frac{n}{2}$ requêtes pédophiles »). La fraction d'utilisateurs pédophiles en fonction de tels seuils serait intéressante à observer.

D'autre part, étudier la distribution des nombres de requêtes par utilisateur pourrait aider à approfondir notre étude des utilisateurs. Certains d'entre eux effectuent un très grand nombre de requêtes, et il peut s'agir de robots, que nous ne devrions peut-être pas prendre en compte dans nos études. Inversement, certains utilisateurs n'effectuent qu'une seule requête sur des périodes de temps très longues telles que celles que nous étudions, ce qui est surprenant. Examiner leurs requêtes permettrait certainement de mieux comprendre leur usage des réseaux P2P.

La problématique plus générale de la quantification des utilisateurs dans un contexte comme le nôtre reste à explorer plus en détail. Nos travaux peuvent être prolongés en adoptant d'autres approches ou en utilisant des données différentes. Il serait très pertinent de confronter les méthodes que nous avons présentées (couple (IP, port), variation de la durée de mesure, sessions temporelles) sur des sujets où on dispose de manières fiables d'identifier des utilisateurs. On peut penser par exemple à des données provenant du Web, où les utilisateurs disposent souvent de *cookies* ou d'un identifiant et d'un mot de passe (complété éventuellement par un *user-agent*[2]).

2. Un *user agent* est transmis par la machine de l'utilisateur au serveur et donne des informations sur l'environnement logiciel de l'utilisateur : nom et version des applications, système d'exploitation, langue, etc.

Chapitre 4

Évolution temporelle

Dans les chapitres précédents, nous avons étudié la détection de requêtes pédophiles, ainsi que les utilisateurs qui les soumettent, et nous les avons quantifiés. Nous avons jusqu'ici peu exploité l'information temporelle contenue dans les requêtes, à l'exception notable de la section 3.1.3 où nous prenons en compte celle-ci pour distinguer des utilisateurs à l'aide de *sessions*. Nous allons présenter dans ce chapitre plusieurs utilisations possibles de cette information pour améliorer la connaissance et la compréhension de l'activité pédophile dans le P2P. En effet, compte tenu de la taille des données dont nous disposons, nous pouvons observer sur une période de temps étendue le comportement de très nombreux pédophiles provenant de différents pays. L'échelle de temps, de plusieurs années, est sans commune mesure avec celle des études antérieures. En outre, les données collectées reflètent directement les recherches de contenus pédophiles, et non le ressenti des pédophiles *a posteriori*, comme dans de nombreuses études qualitatives [16, 77]. Nous espérons donc proposer un éclairage nouveau et pertinent sur le comportement de ces personnes.

Nous utilisons ici la temporalité à deux échelles très différentes.

Dans un premier temps, nous étudions la question naturelle de l'évolution à long terme de l'activité pédophile. Nous nous intéressons donc aux changements d'usage sur le long terme des utilisateurs pédophiles : la quantité d'activité pédophile, et sa proportion, augmente-t-elle, diminue-t-elle ou reste-t-elle stable sur le long terme ? Quels autres changements significatifs observe-t-on d'année en année ?

La seconde problématique vise à utiliser nos approches quantitatives pour essayer de répondre à une question de nature qualitative. Étant donnée la nature criminelle de cette activité, il est important d'étudier l'intégration sociale des utilisateurs qui s'y livrent : ont-ils des horaires indiquant une activité professionnelle, une vie de famille, etc. ? Pour apporter un éclairage sur ces questions, nous examinons les heures auxquelles les requêtes pédophiles sont effectuées afin de savoir si elles le sont à des horaires spécifiques de la journée, ou si elles sont, au contraire, mêlées au reste des requêtes.

Grâce aux jeux de données dont nous disposons, nous pouvons proposer la première étude de cette envergure sur le sujet. Pour tous les travaux du chapitre, nous utilisons en effet l'ensemble `data-ed2k0912`, dont la collecte s'est étendue

sans interruption de juin 2009 à avril 2012 sur un serveur *eDonkey* (voir section 1.2). Il contient plus d'un milliard de requêtes, avec pour chacune un horodatage, une adresse IP anonymisée et géolocalisée au préalable, ainsi que les mots-clefs recherchés.

4.1 Évolution à long terme

Dans cette section, nous étudions l'évolution de l'activité pédophile sur un serveur *eDonkey* français pendant plusieurs années. Nous nous intéressons d'abord au nombre total de requêtes (pas seulement les requêtes pédophiles), avant d'examiner la fraction de requêtes et d'adresses IP pédophiles.

4.1.1 Évolution globale

Pour commencer, nous analysons l'évolution hebdomadaire du nombre total de requêtes sur le serveur de juin 2009 à avril 2012, présentée sur la figure 4.1. Il y a en moyenne 8,7 millions de requêtes par semaine (soit un peu plus de 14 par seconde), avec un maximum de 11 008 609 lors de la 71^{e} semaine et un minimum de 5 332 788 lors de la 54^{e} semaine [1]. Il y a entre 7 389 984 et 9 968 603 requêtes pour 90% des semaines.

Nous observons sur la figure 4.1 que le nombre de requêtes est relativement stable mais pas constant. Au-delà des petites fluctuations d'une semaine à l'autre, un motif semble se répéter d'année en année : le nombre de requêtes est stable entre mi-octobre et mi-mai, décroissant de mi-mai à fin juillet puis augmente d'août à octobre. Ces similarités s'expliquent peut-être par des changements de rythmes réguliers des utilisateurs, par exemple liés aux vacances scolaires.

Il est remarquable que les valeurs atteintes pendant ces périodes de stabilité soient très proches pour les trois années considérées. En effet, différents événements auraient pu avoir une influence notable sur le nombre de requêtes hebdomadaire. En particulier, la loi HADOPI [51] [2] a été mise en place à la fin de l'année 2009 en France, avec pour objectif de mettre un terme aux échanges de fichiers protégés par le droit d'auteur sur les réseaux P2P. Les réseaux P2P, et en particulier *eDonkey*, ont commencé à être surveillés, ce qui a amené les utilisateurs à se tourner vers des solutions alternatives, comme des réseaux P2P plus difficiles à surveiller qu'*eDonkey*, ou des plateformes de téléchargement direct, comme

1. Le serveur a subi une interruption de service en juillet 2010 pendant une vingtaine d'heures, correspondant à la chute brutale observée. De même, le début de la mesure correspond au démarrage du serveur qui n'avait donc pas encore un trafic établi, ce qui explique la faiblesse des valeurs relevées pour les deux premières semaines.

2. Rappelons que les requêtes étudiées ici ont été collectées sur un serveur français.

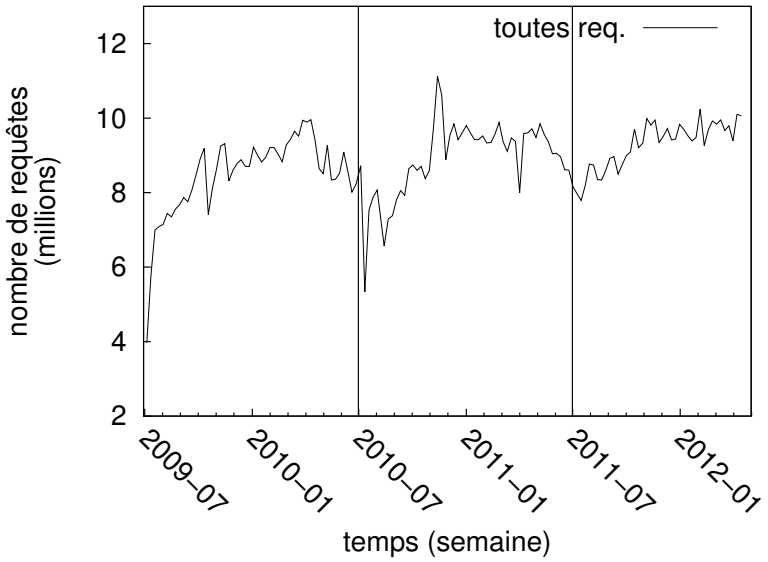


Figure 4.1 – *Nombre total de requêtes par semaine, entre juillet 2009 et avril 2012.*

MegaUpload. La fermeture de MegaUpload [59] début 2012 aurait aussi pu inciter de nombreux utilisateurs de cette plateforme à utiliser *eDonkey*. Mais la courbe de la figure 4.2 ne montre pas de variation importante du volume général lors de ces deux évènements. On peut imaginer que les variations de trafic dans un pays ont été compensées par celles d'autres pays, que les nouveaux utilisateurs du P2P se tournent vers d'autres systèmes, ou encore qu'*eDonkey* a atteint un niveau de saturation de son public.

4.1.2 Activité pédophile

La figure 4.2 présente la fraction de requêtes détectées comme pédophiles chaque semaine. Nous constatons que celle-ci évolue de façon importante au cours de la mesure. D'abord en deçà de 2 pour mille, la fraction augmente ensuite significativement pendant les six premiers mois de 2010. Elle se stabilise alors, et décroît même légèrement, avant une nouvelle augmentation sensible pour les six premiers mois de 2011. Enfin, la fraction de requêtes pédophiles semble se stabiliser entre 4,5 et 5 pour mille.

Sachant que le nombre total de requêtes est stable, cette étude montre que la fraction de requêtes pédophiles observées a sensiblement augmenté (presque triplé) entre 2009 et 2012. Cette augmentation peut être due à une augmentation du nombre de personnes liées à cette activité ou à une augmentation du nombre de requêtes de ces personnes.

Pour différencier ces deux situations, nous examinons ce qu'il en est pour les utilisateurs. Compte tenu des informations disponibles dans les données

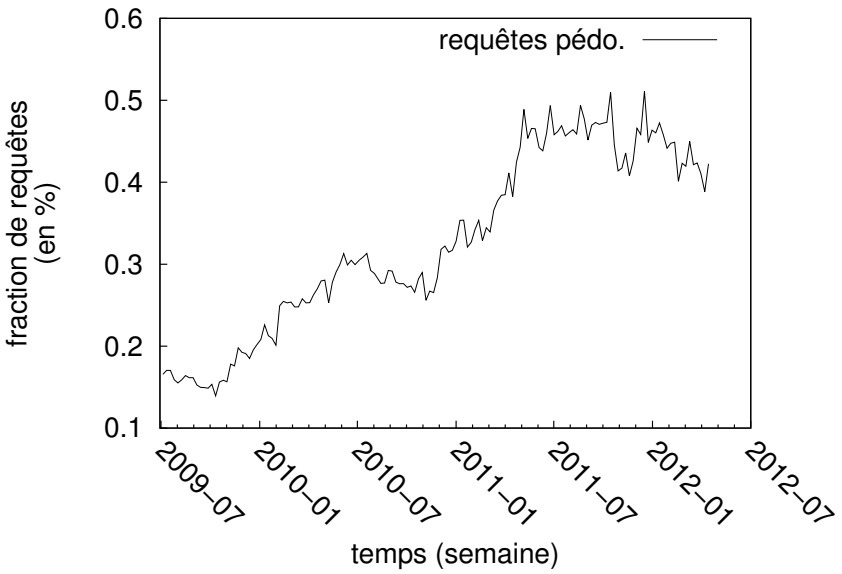


Figure 4.2 – *Fraction de requêtes détectées comme pédophiles par semaine, entre 2009 et 2012.*

`data-ed2k0912`, nous traçons l'évolution de la fraction d'adresses IP pédophiles détectées chaque semaine, c'est-à-dire d'adresses IP ayant effectué au moins une requête pédophile durant la semaine [3]. Le résultat est présenté en figure 4.3. La valeur est proche de 0,25% jusqu'à fin 2009, puis augmente progressivement, jusqu'à atteindre 0,8% mi-2011. Elle diminue ensuite légèrement et se rapproche de 0,7% à la fin de la mesure. Au-delà des petites variations d'une semaine à l'autre, la tendance générale est celle d'une croissance importante entre 2009 et 2012. C'est donc un début d'explication de l'augmentation de la fraction de requêtes pédophiles observée : il y a davantage de requêtes pédophiles car davantage d'adresses IP soumettent ce type de requêtes. Bien que le nombre total d'utilisateurs soit stable, donc, de plus en plus d'entre eux entrent des requêtes pédophiles. Ceci peut-être dû à l'arrivée de nouveaux utilisateurs pédophiles ou à des utilisateurs initialement non pédophiles se découvrant ensuite cette tendance. Ceci peut indiquer une dangereuse banalisation de ce type de contenu. Il est aussi envisageable que certains mots-clefs utilisés par notre outil de détection (notamment ceux de la liste *explicit*) perdent, en partie ou totalement, leur caractère spécifique à ce contexte et soient alors employés par davantage d'utilisateurs. Remarquons néanmoins que des utilisateurs qui commenceraient à employer fortuitement de tels mots-clefs dans un contexte non pédophile décideraient sans doute rapidement d'en changer, afin d'éviter de s'exposer aux contenus pédophiles. Enfin, des phénomènes d'offre

3. L'adresse IP permet d'obtenir une approximation légèrement surestimée de la fraction d'utilisateurs. Considérer des semaines permet de réduire légèrement la pollution due à la réallocation des adresses IP entre utilisateurs (voir chapitre 3).

et de demande influent certainement sur ces chiffres : les pairs désireux d'obtenir des contenus pédophiles vont privilégier les réseaux P2P sur lesquels ils vont trouver plus facilement ce qu'ils cherchent (et en délaisseront d'autres). De même, les fournisseurs sont susceptibles de diffuser leurs contenus préférentiellement sur certains réseaux. Nous explorons en partie ces questions dans le chapitre suivant, en comparant deux systèmes P2P.

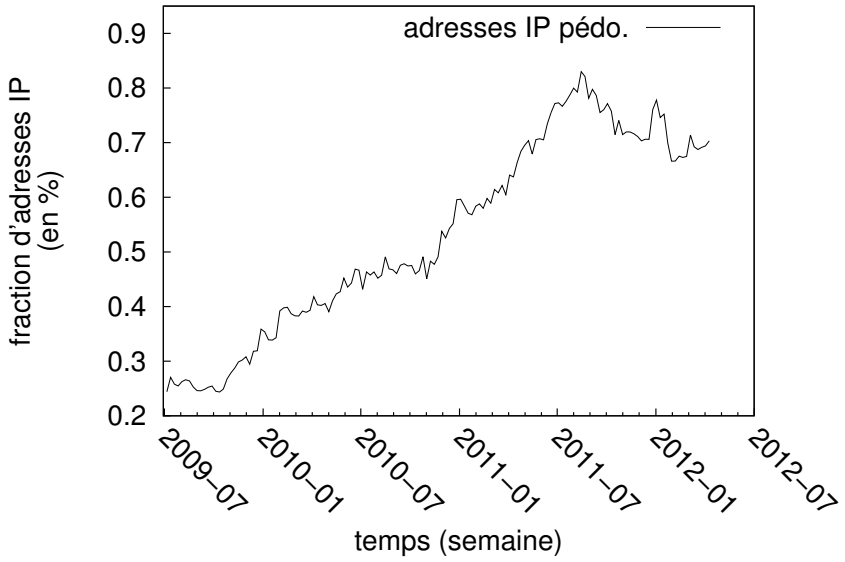


Figure 4.3 – *Fraction d'adresses IP détectées comme pédophiles par semaine, entre 2009 et 2012.*

Nous avons étudié l'évolution de la fraction de requêtes pédophiles de 2009 à 2012. Elle a significativement augmenté, passant de moins de 0,2% à 0,5%, alors que le nombre total de requêtes est resté stable. La fraction d'adresses IP impliquées dans ces requêtes a aussi quasiment triplé sur cette période, indiquant que de plus en plus d'utilisateurs (nouveaux ou anciens) sont intéressés par des contenus à caractère pédo-pornographique.

4.2 Dynamique journalière

Dans cette section, nous utilisons l'information temporelle pour analyser la dynamique des utilisateurs pédophiles au cours de la journée.

Notre objectif est de déterminer si les utilisateurs pédophiles effectuent généralement ce type de requêtes en même temps que les autres, ou au contraire à des moments distincts. Ceci permet d'avoir un éclairage original sur l'intégration sociale des utilisateurs pédophiles : s'ils souhaitent se cacher, on peut supposer qu'ils soumettent ces requêtes lorsque les membres de leur entourage a le moins de chances de les découvrir, par exemple parce que ces derniers sont couchés. C'est un aspect de l'activité pédophile qui est pour l'instant très peu étudié et essentiellement avec des approches qualitatives [26]. Remarquons que les auteurs de [64] affirment avoir effectué brièvement une telle étude (sur des requêtes du réseau BitTorrent), mais ils ne détaillent ni leurs méthodes, ni leurs conclusions. Dans [8], les auteurs présentent une étude de requêtes Web, dans laquelle ils analysent notamment la fréquence de certaines catégories de requêtes.

Nous effectuons d'abord une étude générale de la dynamique journalière, en examinant les volumes et fractions de requêtes au cours de la journée. Nous procédons ensuite à une vérification de ces résultats en restreignant géographiquement notre analyse. Nous terminons cette section en comparant la dynamique de l'activité pédophile à celle de l'activité pornographique.

4.2.1 Étude générale

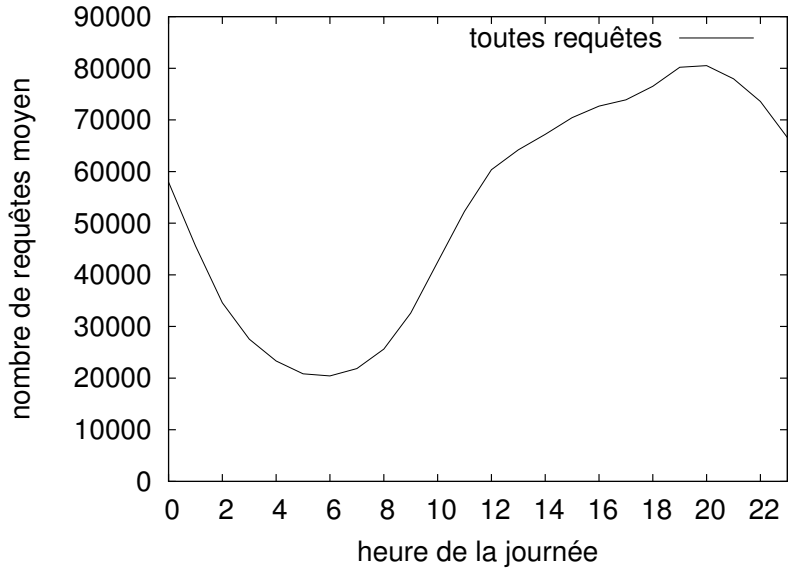


Figure 4.4 – *Évolution du nombre total de requêtes en fonction de l'heure de la journée.*

La figure 4.4 présente la variation du nombre de requêtes en fonction de l'heure de la journée. Cette courbe est réalisée en prenant en compte l'ensemble des requêtes de `data-ed2k0912` et en calculant, pour chaque heure de la journée,

le nombre moyen de requêtes reçues par le serveur. C'est l'heure de ce dernier qui sert de référence et il est réglé à l'heure française [4].

Nous constatons que le nombre moyen de requêtes est très variable en fonction de l'heure de la journée. Le minimum (20 423,6) est atteint à 6 heures du matin et il croît ensuite jusqu'à 20 heures où il atteint son maximum (80 518,8). La croissance est forte jusqu'à 12 heures, puis s'infléchit. Le nombre de requêtes décroît ensuite régulièrement entre 20 heures et 6 heures du matin.

La chute du nombre de requêtes que nous observons entre 20 heures et 6 heures du matin correspond à la période de la nuit en Europe de l'Ouest. Le serveur connaît son pic d'activité entre 17 et 22 heures, donc en soirée. Le nombre de requêtes sur le serveur suit donc le rythme journalier des pays de l'Europe de l'Ouest. Ceci est confirmé par l'étude de l'information de géolocalisation des adresses IP (voir annexe B), qui montre que les requêtes sur ce serveur proviennent majoritairement d'Italie, de France et d'Espagne.

Connaissant la dynamique des requêtes sur le serveur, nous étudions ensuite l'évolution des requêtes pédophiles pour chaque heure de la journée.

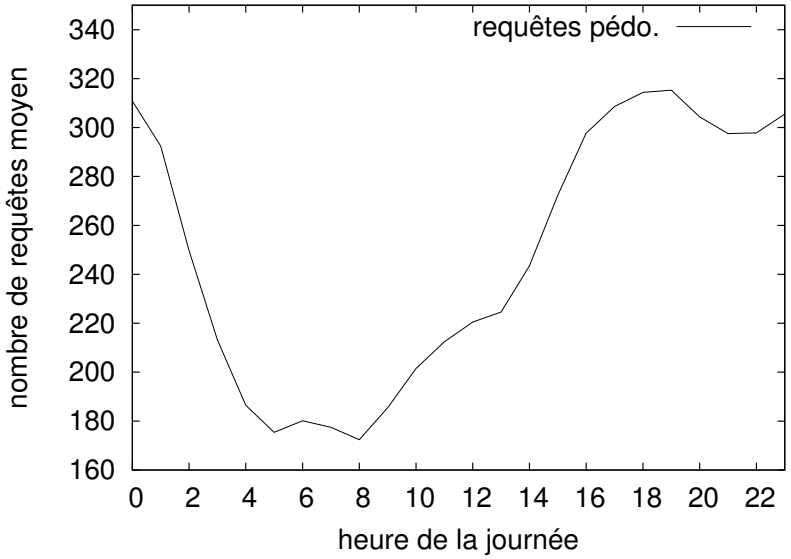


Figure 4.5 – *Nombre de requêtes détectées comme pédophiles en fonction de l'heure de la journée.*

La figure 4.5 présente le nombre moyen de requêtes pédophiles observées pour chaque heure de la journée [5]. L'allure générale de cette courbe est très

4. Pour éviter de biaiser les mesures en prenant en compte les quelques heures où le serveur a partiellement cessé de collecter des requêtes, les 2% d'heures avec le moins de requêtes ont été enlevées.

5. Compte tenu de l'augmentation importante de la fraction de requêtes pédophiles au cours

similaire à celle du nombre total de requêtes. Il y a un effet jour-nuit prononcé : la valeur atteint un minimum vers 6 heures du matin et un maximum à 19 heures. Néanmoins, nous constatons que les amplitudes sont sensiblement différentes : pour le volume total, le rapport entre le nombre de requêtes à 19 heures et à 6 heures est proche de 4, alors que pour les requêtes pédophiles, ce rapport est proche de 2.

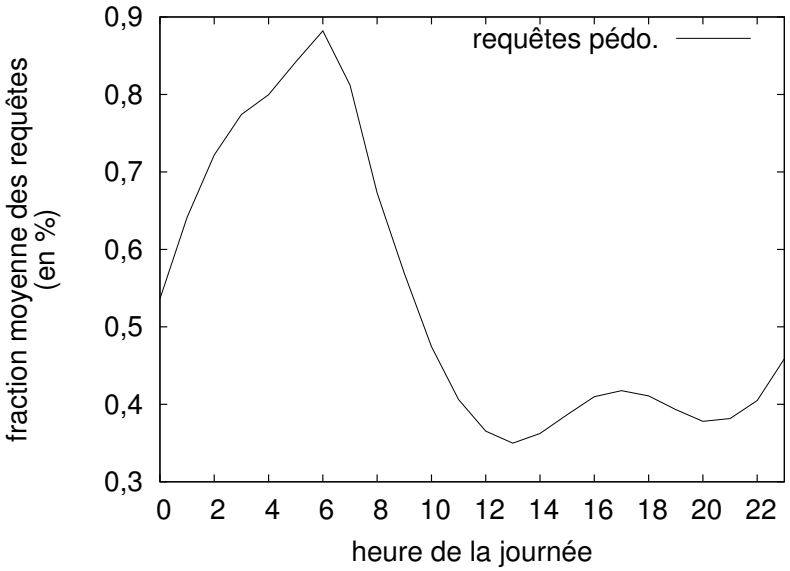


Figure 4.6 – *Fraction de requêtes détectées comme pédophiles en fonction de l'heure de la journée.*

Ceci est mis en évidence sur la figure 4.6 qui présente la fraction moyenne de requêtes considérées comme pédophiles pour chaque heure. La valeur varie sensiblement au cours de la journée. Elle atteint un maximum de près de 0,9% vers 6 heures du matin et un minimum de 0,35% à 13 heures. Elle croît régulièrement de 22 heures à 6 heures du matin avant de décroître rapidement dans la matinée, de 6 heures à midi. Il y a ensuite une deuxième période de croissance entre 13 heures et 17 heures, où la fraction atteint un maximum local de 0,41%.

Si les requêtes pédophiles étaient réparties uniformément dans la journée, la fraction devrait être constante. Le pic important que présente la fraction de requêtes pédophiles vers 6 heures du matin confirme nos observations ci-dessus : si les utilisateurs pédophiles semblent avoir des rythmes d'activité liés au cycle jour-nuit en général, les requêtes pédophiles sont sensiblement sur-représentées vers la fin de la nuit. Nous remarquons aussi la présence d'une légère augmentation

de la mesure `data-ed2k0912`, les moyennes calculées sur l'ensemble de la mesure pourraient être biaisées. Cependant, une étude (non présentée dans ce manuscrit) réalisée sur les périodes de février à novembre 2010 et de mai à décembre 2011, où le trafic est relativement constant, donne des résultats similaires.

autour de 17 heures, qui correspond au moment où il y a un pic du nombre de requêtes pédophile et une croissance légèrement ralentie du nombre total de requêtes.

Nous constatons sur la figure 4.6 que la fraction d'adresses IP pédophiles en fonction de l'heure de la journée évolue de la même façon que celle des requêtes. Il ne semble donc pas y avoir de phénomène sous-jacent tel que la présence vers 6 heures du matin d'utilisateurs effectuant significativement plus de requêtes pédophiles que les autres, qui pourrait expliquer le pic de requêtes autour de 6 heures du matin.

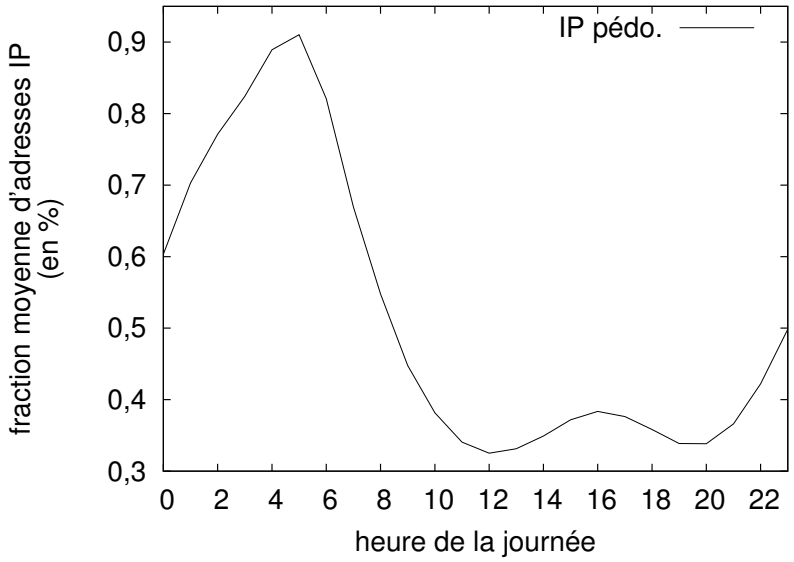


Figure 4.7 – *Évolution de la fraction moyenne d'adresses IP détectées comme pédophiles, en fonction de l'heure (du serveur).*

4.2.2 Restriction géographique

Les résultats précédents reposent sur l'horodatage de la requête par l'outil de collecte. Pour en accroître la fiabilité, nous pouvons mettre à profit la géolocalisation des adresses IP des utilisateurs et regarder si nous observons la même allure de courbe pour des pays ayant des fuseaux horaires différents.

Il nous faut pour cela choisir des pays pour lesquels le nombre de requêtes est suffisamment élevé pour avoir des valeurs significatives et dont les habitants parlent des langues pour lesquelles la fiabilité de notre outil est bonne (typiquement l'anglais, le français et les autres langues européennes). Il faut aussi que les pays présentent un décalage horaire significatif, typiquement de plusieurs heures. Après examen des données, nous choisissons d'une part la France, dont le fuseau horaire

est UTC+1 [6] et dont les habitants parlent majoritairement le français, et d'autre part, le groupe de pays Argentine et Brésil, qui utilisent tous les deux le fuseau horaire UTC-3, et dont les habitants parlent des langues latines. Il y a donc, selon les périodes de l'année, entre 4 et 5 heures de décalage horaire entre les pays choisis.

La figure 4.8 montre la fraction moyenne de requêtes pédophiles pour la France d'une part et le groupe Argentine-Brésil d'autre part. Les allures des courbes sont exactement les mêmes mais avec un décalage de quelques heures. Ceci nous permet de confirmer nos observations préalables : les requêtes pédophiles représentent une proportion nettement plus importante du trafic à un moment précis de la journée, en début de matinée.

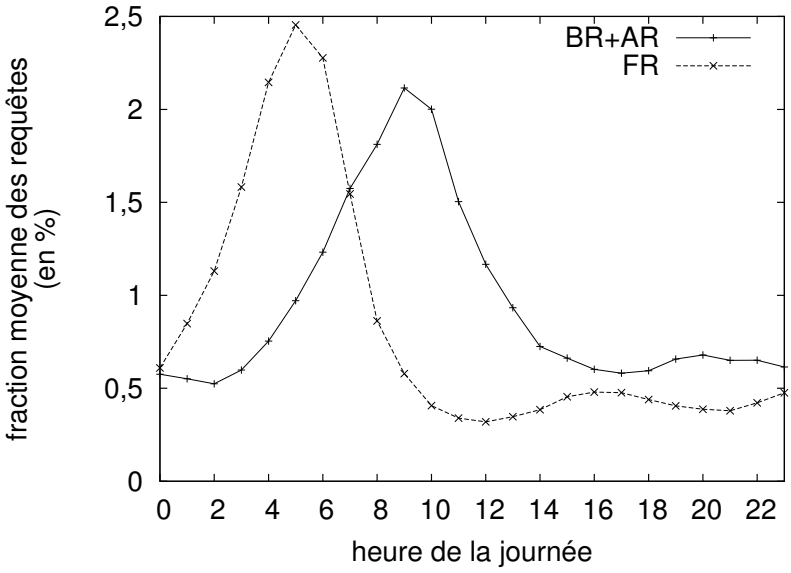


Figure 4.8 – *Évolution des fractions moyennes de requêtes détectées comme pédophiles en France et dans le groupe de pays Argentine-Brésil, en fonction de l'heure (du serveur).*

4.2.3 Comparaison thématique

La pédopornographie est souvent considérée comme une sous-partie de la pornographie classique [64], qui représente une proportion très importante du trafic (12% des sites Web y sont consacrés, selon [61]). Il est donc intéressant de comparer l'activité des utilisateurs pédophiles et celle des utilisateurs cherchant de la pornographie. Nous pouvons pour cela comparer l'évolution des demandes visant ces deux types de contenus.

6. C'est-à-dire décalé d'une heure par rapport au Temps universel coordonné [75]. En raison d'un changement d'heure, ce décalage est porté à deux heures en été.

Cela nécessite d'avoir un outil pour détecter les requêtes ciblant la pornographie. Notre objectif est avant tout de savoir, en première approximation, s'il existe des moments privilégiés pour effectuer des requêtes visant des contenus pornographiques. Nous restreignons notre étude aux 6 mots-clefs les plus fréquents de la liste *sex* définie dans le chapitre 2 [7] et nous considérons une requête comme pornographique si elle contient l'un de ces six mots-clefs. Nous procédons ensuite à l'analyse de l'ensemble `data-ed2k0912` avec ce filtre.

Nous traçons sur la figure 4.9 la fraction de requêtes pédophiles (déjà présentée sur la figure 4.6) et la fraction de requêtes pornographiques au cours de la journée.

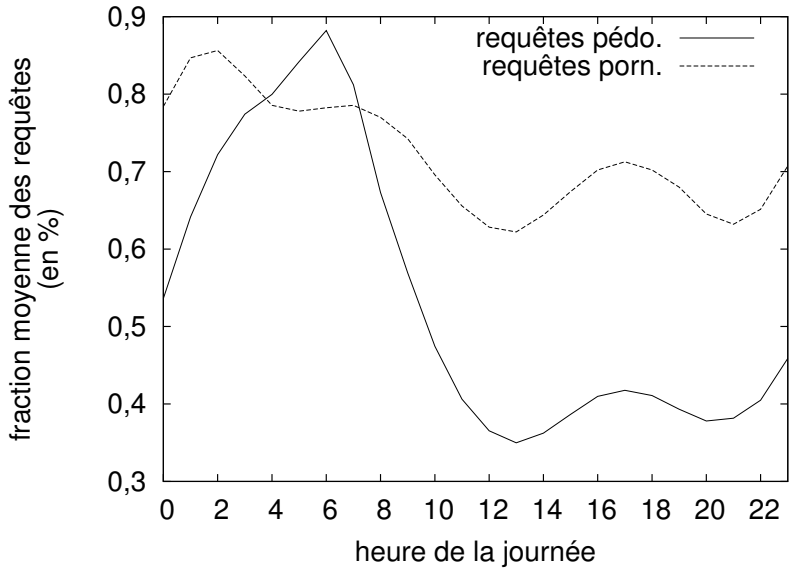


Figure 4.9 – *Fraction moyenne de requêtes détectées comme pédophiles et pornographiques en fonction de l'heure de la journée.*

Les ordres de grandeur des deux fractions sont similaires, mais cela est peu significatif, puisque nous avons volontairement limité le nombre de mots-clefs pornographiques pris en compte [8].

Les bornes entre lesquelles évoluent les fractions de requêtes pédophiles et pornographiques calculées sont très intéressantes. Pour la pédophilie, la valeur est, nous l'avons dit, comprise entre 0,35% et 0,89%, alors que pour la pornographie la valeur se situe entre 0,62% et 0,85%. Cette faible amplitude montre que la pornographie est presque uniformément répartie dans les requêtes de la journée, ce qui n'est pas du tout le cas pour la pédopornographie qui, comme nous l'avons

7. Cette liste contient de nombreux termes non spécifiques à la pornographie et qui, employés seuls, ne suffiraient pas à caractériser une requête pornographique.

8. Remarquons cependant que l'allure de notre courbe pour les requêtes pornographiques est très similaire à celle présentée dans [8], ce qui laisse penser que les mots-clefs choisis sont pertinents pour notre étude.

vu, présente un pic important vers 6 heures du matin. Il y a donc bien une différence significative entre les deux thématiques, indiquant que les utilisateurs concernés par l'une et l'autre ont des comportements sensiblement différents. Bien que ceci dépasse le cadre de cette thèse, ce type d'observation pourrait fournir des arguments importants aux chercheurs souhaitant évaluer dans quelle mesure la pédopornographie est une sous-thématique de la pornographie ou caractériser le comportements des utilisateurs pédophiles. Les horaires de plus forte présence des utilisateurs pédophiles indiquent par exemple une intégration sociale assez forte, puisqu'ils sont compatibles avec des horaires de travail et d'activités familiales typiques (ce qui n'aurait pas été le cas si nous avions observé un pic autour de 10, 14 ou 18 heures par exemple).

L'évolution de la fraction de requêtes pédophiles au cours de la journée montre que 6 heures du matin est un moment d'activité pédophile accru. Cela caractérise l'activité pédophile et la distingue de la pornographie qui est répartie plus uniformément.

4.3 Conclusion

Dans ce chapitre, nous avons étendu notre étude de l'activité pédophile dans les réseaux P2P en analysant son évolution temporelle. Nous avons dans un premier temps examiné les changements au long terme, constatant que les chiffres de 2009, obtenus aux chapitres 2 et 3, ont significativement évolué dans les années suivantes. Nous avons ensuite présenté une étude des requêtes pédophiles en fonction de l'heure de la journée et avons constaté que la fin de la nuit est privilégiée pour effectuer des recherches pédophiles. Ce n'est pas le cas pour des requêtes liées à la pornographie classique, les deux activités étant donc sensiblement distinctes.

Ces travaux introduisent de nouvelles approches d'analyse et de nouvelles connaissances sur l'activité pédophile dans les réseaux P2P. Les statistiques obtenues sont très importantes pour alimenter en informations précises et rigoureuses les débats sur les législations et réfléchir aux moyens à mettre en œuvre concernant l'Internet et de la protection des enfants.

L'évolution des requêtes pédophiles sur une longue durée pourrait être poussée plus loin, notamment en utilisant la géolocalisation de manière plus précise (voir annexe B). Il serait par exemple intéressant d'étudier si des événements relatifs à la pédopornographie et localisés dans un pays (par exemple, une opération de police ou la diffusion d'un documentaire sur le sujet) se traduisent par des augmentations

ou des chutes de trafic. L'étude détaillée de l'évolution des requêtes en fonction des catégories définies au chapitre 2 pourrait aussi éclairer la compréhension du phénomène (voir annexe C). Enfin, l'analyse de la dynamique journalière pourrait être prolongée, par exemple en observant la dynamique hebdomadaire ou à d'autres granularités, pour examiner notamment les différences éventuelles entre les jours ouvrables et le week-end.

Chapitre 5

Comparaison de *KAD* et *eDonkey*

Jusqu'à présent, nous avons effectué notre travail d'analyse sur des ensembles de requêtes provenant uniquement du réseau *eDonkey*. On peut se demander si ces résultats sont généralisables à d'autres systèmes et, notamment, si l'activité pédophile diffère entre les réseaux P2P. Comparer des systèmes très différents s'avère toutefois difficile : il est nécessaire de disposer de données adaptées, afin que les différences observées traduisent bien des usages différents. Nous avons pu obtenir cela pour *eDonkey* et pour *KAD* et nous proposons donc dans ce chapitre une comparaison de l'activité pédophile entre ces deux réseaux.

De nombreuses raisons peuvent guider un utilisateur pour choisir tel ou tel système P2P : la législation de son pays, l'intensité de la répression envers les responsables des réseaux, l'offre de contenus dans telle ou telle langue, etc. [63] Ces facteurs influencent donc aussi les échanges pédophiles. Le niveau d'anonymat offert peut notamment être un élément important pris en compte par les utilisateurs pour choisir tel ou tel réseau, compte tenu de l'illégalité de cette activité.

KAD et *eDonkey* sont deux réseaux P2P très répandus qui ont toutefois une différence architecturale majeure : *eDonkey* repose sur des serveurs alors que *KAD* est complètement distribué et ne repose que sur les pairs qui s'y connectent. Cette absence de centralisation laisse penser qu'un utilisateur de *KAD* est plus anonyme qu'un utilisateur d'*eDonkey*, système dans lequel les administrateurs du serveur peuvent très facilement observer les utilisateurs.

La comparaison des deux systèmes que nous présentons ici permet donc d'évaluer l'influence de l'architecture du réseau P2P sur l'activité pédophile et d'accroître les connaissances sur l'activité pédophile dans le P2P en général.

La section 5.1 décrit les spécificités des données que nous utilisons. La section 5.2 présente une comparaison de la *quantité* de requêtes pédophiles sur *KAD* et *eDonkey*. Nous étudions en détail dans la section 5.3 les indications d'âges, régulièrement présentes dans les requêtes pédophiles, afin d'explorer de possibles différences dans la *nature* de l'activité pédophile dans les deux systèmes. Enfin, nous estimons la fraction de requêtes pédophiles dans *KAD* à partir de celle d'*eDonkey* en section 5.4.

5.1 Données

Nous présentons dans cette section les données que nous utilisons pour réaliser notre comparaison de *KAD* et d'*eDonkey*. Nous entrons davantage dans les détails que dans la section 1.2, dans laquelle nous avions brièvement introduit les ensembles étudiés.

Dans *eDonkey*, comme nous l'avons vu, les pairs se connectent à un serveur, chargé d'indexer les contenus partagés par chacun et ainsi capable de répondre aux demandes des utilisateurs qui souhaitent obtenir des fichiers. Dans *KAD*, en revanche, l'indexation des contenus est répartie entre les différents pairs du réseau, sans autorité centrale. Quand un utilisateur effectue une recherche, le système contacte les pairs de proche en proche pour déterminer s'ils possèdent des fichiers correspondant à la requête, à l'aide de l'empreinte [1] de la requête [12].

Pour collecter les données du réseau *KAD*, nous avons collaboré avec l'équipe MADYNES du LORIA qui a développé le système HAMACK au cours de la thèse de Thibault Cholez [14, 15]. Celui-ci introduit des sondes dans la table de hachage distribuée du réseau, de façon à observer tout le trafic relatif à un mot-clef donné. Nous avons supervisé 72 mots-clefs, que nous avons choisis pour représenter la variété des requêtes soumises au système et satisfaire les besoins de notre expérience. Nous avons tenu compte des capacités de HAMACK et des fréquences d'apparition des mots-clefs dans les requêtes de différentes catégories de `data-ed2k2009`.

Les mots-clefs retenus sont classés en trois groupes. Le premier groupe contient 19 mots-clefs provenant de la liste *explicit* définie dans le chapitre 2 : *babyj, babyshivid, childlover, childporn, hussyfan, kidzilla, kingpass, mafiasex, pedo, pedofilia, pedofilo, pedoland, pedophile, pthc, ptsc, qqaazz, raygold, yamad, youngvideomodels*. Le deuxième groupe contient 23 mots-clefs qui apparaissent parfois dans les requêtes pédophiles de `data-ed2k2009`, mais aussi dans les requêtes non pédophiles : *1yo, 2yo, 3yo, 4yo, 5yo, 6yo, 7yo, 8yo, 9yo, 10yo, 11yo, 12yo, 13yo, 14yo, 15yo, 16yo, boy, girl, mom, preteen, rape, sex, webcam*. Les âges sont fréquemment utilisés dans les requêtes pédophiles (voir les chiffres présentés dans l'annexe C) mais le sont aussi dans d'autres contextes, comme par exemple des parents cherchant des jeux pour enfants adaptés à un certain âge. Le dernier groupe est considéré comme groupe-témoin et contient 30 mots-clefs non spécifiques et très rarement présents dans les requêtes pédophiles de `data-ed2k2009` : *avi, black, christina, christmas, day, doing, dvdrip, early, flowers, grosse, hot, house, housewives, live, love, madonna, man, new, nokia, pokemon, rar, remix, rock, saison, smallville, soundtrack, virtual, vista, windows, world*. Nous notons respectivement ces trois groupes de mots-clefs

1. Une empreinte, ou *hash*, est un identifiant de la requête qui permet de la stocker dans la table de hachage de *KAD*. Dans ce réseau, les pairs et les contenus ont des identifiants, les pairs étant responsables des contenus qui ont les identifiants les plus proches (au sens de la distance XOR). Pour obtenir un fichier, il faut contacter le pair qui en est responsable.

`paedophile`, `mixed` et `not paedophile`. Remarquons que la plupart des mots-clefs sont anglais (*love, early, flowers*), mais certains proviennent d'autres langues (*saison, pedofilia*), et certains sont même des marques commerciales (*pokemon, nokia*).

Sur le réseau *eDonkey*, une mesure permet d'obtenir toute l'activité d'un serveur, c'est-à-dire d'une partie du réseau. Afin d'accroître la portée de nos résultats, nous avons cependant utilisé ici les requêtes de deux serveurs, grâce à la collaboration entamée en 2009 avec le responsable de deux serveurs. Le premier, français, filtre les contenus et n'indexe qu'une partie d'entre eux. Le second, situé en Ukraine, indexe tout ce que les utilisateurs proposent, sans filtrage.

Pour *KAD*, HAMACK ne permet de superviser qu'un ensemble restreint de mots-clefs. En contrepartie, il enregistre la totalité des requêtes du réseau concernant ces mots-clefs. Cependant, différents mécanismes de recherche ont été implémentés dans les applications clientes. Dans certaines, lorsqu'un utilisateur saisit une requête, l'application divise celle-ci en plusieurs mots-clefs et calcule l'empreinte en se fondant sur le premier d'entre eux. Pour les autres, le mot-clef retenu n'est plus le premier mais le plus long de la requête. Par exemple, « *the matrix revolution* » est indexé selon les clients sur « the » ou sur « revolution ». Or, HAMACK collecte les requêtes en fonction du mot-clef sur lequel elles sont indexées. Cela a donc une conséquence directe sur l'enregistrement effectué : pour un mot-clef très court, comme « avi » par exemple, HAMACK ne collecte quasiment que des requêtes où celui-ci est le seul mot-clef de la requête. En effet, dans les autres requêtes, ce mot-clef n'est pas le plus long et n'est pas fréquemment placé en début de requête (« avi » est une extension du nom du fichier et apparaît donc généralement à la fin). Afin de rendre les jeux de données comparables, nous avons donc restreint tous les jeux de données initiaux aux requêtes composées d'un seul mot, celui-ci appartenant aux 72 que nous supervisons sur *KAD*.

Nous obtenons finalement trois jeux de données, appelés `data-KAD`, `data-ed2k-FR` et `data-ed2k-UA`. Leurs caractéristiques sont rappelées dans le tableau 1.1. La construction de ces ensembles assure qu'ils sont comparables : dans chaque cas, ils contiennent toutes les requêtes d'une forme donnée parvenues dans chacun des systèmes. Leur grande taille permet de rendre pertinentes les études que nous présentons dans la suite.

Nous décrivons l'obtention d'ensembles de requêtes comparables entre *KAD* et deux serveurs *eDonkey*. Celles-ci sont composées d'un seul mot-clef parmi les 72 étudiés.

5.2 Quantité de requêtes pédophiles dans *eDonkey* et dans *KAD*

La manière la plus immédiate de comparer l'activité pédophile entre différents systèmes est certainement de comparer la fraction de requêtes pédophiles dans chacun d'entre eux. Nous allons étudier les fréquences des mots-clefs des différentes catégories.

La figure 5.1 présente la fraction de requêtes dans les trois catégories de mots-clefs que nous avons définies. Près de 70% des requêtes de `data-KAD` proviennent de la catégorie `not paedophile`, un peu moins de 20% sont des requêtes contenant des mots de `mixed` et enfin un peu plus de 10% sont des mots-clefs appartenant à `paedophile`.

Nous observons nettement que les comportements de recherche sont distincts entre les deux réseaux : les valeurs obtenues pour les catégories `paedophile` et `not paedophile` diffèrent sensiblement entre `data-KAD` et les deux ensembles de requêtes d'*eDonkey*. Nous observons que la fraction de requêtes pédophiles est significativement plus faible dans `data-KAD` que dans `data-ed2k-FR` et `data-ed2k-UA`, ce qui est contraire à notre intuition, puisque *KAD* est supposé garantir un anonymat supérieur. Cet histogramme montre aussi qu'il existe des valeurs similaires pour les deux ensembles `data-ed2k-FR` et `data-ed2k-UA`. Les différences d'indexation entre les serveurs semblent donc avoir peu d'influence sur la fraction de requêtes visant des contenus pédopornographiques.

En vue de mieux comprendre les phénomènes sous-tendant ces résultats, nous étudions séparément les fréquences de chaque mot-clef dans chacun des jeux de données.

Nous voulons vérifier que la nature pédophile d'un mot-clef influe sur sa fréquence d'apparition dans un système donné. Nous devons donc d'abord évaluer dans quelle mesure un mot-clef est pédophile. Nous utilisons pour cela le jeu de données `data-ed2k2009` aisni que notre outil de détection de requêtes.

Soit D l'ensemble de toutes les requêtes et $D(k)$ l'ensembles des requêtes contenant un mot-clef k. Pour chaque mot-clef, nous divisons $D(k)$ en deux sous-ensembles $P(k)$ et $N(k)$, composés respectivement des requêtes étiquetées comme pédophiles et non pédophiles. Ensuite, nous définissons le *coefficient pédophile* $\pi(k)$ de chaque mot-clef k comme suit :

$$\pi(k) = \frac{|P(k)|}{|D(k)|}.$$

Si toutes les requêtes contenant k sont pédophiles, $\pi(k) = 1$, si aucune ne l'est, $\pi(k) = 0$. Tous les mots-clefs de la catégorie `not paedophile` ont un *coefficient*

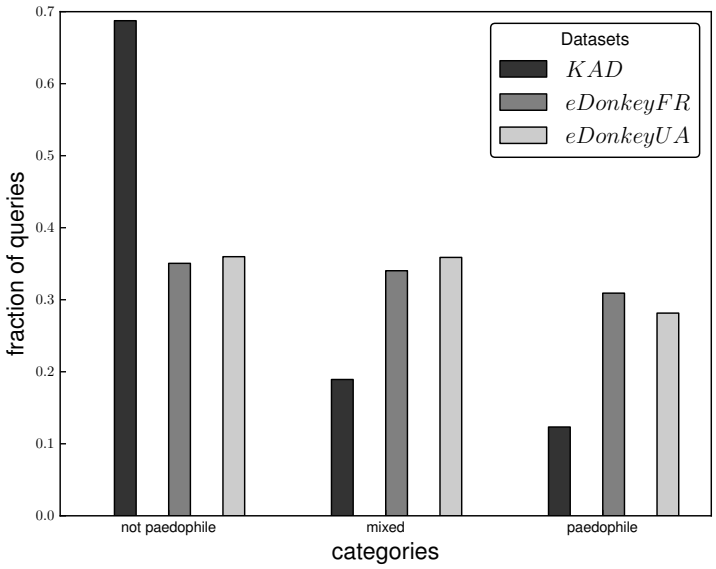


Figure 5.1 – *Fraction de requêtes dans chaque catégorie, pour les trois jeux de données étudiés.*

pédophile inférieur à $0,6\%$. Ceux de la catégorie `mixed` ont un *coefficient pédophile* compris entre 1 et 40%. Tous les mots-clefs de la catégorie `paedophile` ont un *coefficient pédophile* supérieur à $88,5\%$.

Nous traçons sur la figure 5.2 les fractions $\frac{f_{edonkey1}(k)}{f_{kad}(k)}$ et $\frac{f_{edonkey2}(k)}{f_{kad}(k)}$, où $f_s(k)$ est le nombre de requêtes contenant un mot-clef k dans un ensemble s donné, pour chacun des 72 mots-clefs. Ceux-ci sont ordonnés en abscisse par *coefficient pédophile* croissant. La ligne horizontale représente *y = 1*, ce qui permet de comparer visuellement les valeurs entre les ensembles : lorsque le point est situé sous la ligne, le mot-clef apparaît plus fréquemment dans `data-KAD`, sinon dans `data-ed2k-FR` ou `data-ed2k-UA` selon le cas.

Cette courbe montre une corrélation claire entre la nature pédophile d'un mot-clef et une présence plus importante dans les requêtes d'*eDonkey* que de *KAD*. En outre, les fréquences sur chacun des ensembles `data-ed2k-FR` et `data-ed2k-UA` sont similaires, ce qui traduit que le phénomène évoqué plus haut existe également au sein des catégories.

Ces résultats pourraient toutefois être biaisés par le fait que la validation a eu lieu sur *eDonkey* et l'on pourrait penser que les requêtes sur *KAD* sont simplement de nature différente. Remarquons cependant que les mots-clefs pédophiles les plus rares auraient alors des valeurs différentes des plus évidents (comme *childporn*, *pedo*). Cependant, les valeurs sont homogènes pour toute la catégorie `paedophile`. La nature des requêtes pédophiles entre les deux réseaux P2P semble donc similaire.

Un argument supplémentaire en faveur de cette hypothèse est présenté dans la section 5.3.

Nous pouvons conclure que le niveau d'anonymat fourni par défaut par un réseau P2P n'est pas le facteur déterminant pour choisir un système ou l'autre, puisque ni l'architecture décentralisée de *KAD*, ni les politiques d'indexation de contenus ne font augmenter les fractions de requêtes pédophiles. Au contraire, celles-ci sont plus fréquentes sur *eDonkey* que sur *KAD*. L'explication de ce phénomène reste ouverte. L'utilisation de *KAD* requiert *a priori* des compétences techniques supérieures que celles nécessaires à *eDonkey*, et cela peut être un facteur. Les pairs peuvent aussi effectuer leurs recherches sur *eDonkey* en protégeant leur vie privée avec d'autres outils, tels que des réseaux privés virtuels (VPN) ou TOR [70]. Les requêtes sur *KAD* sont envoyées en UDP et ne peuvent bénéficier de l'anonymisation de TOR, ce qui peut aussi expliquer les différences d'usage observées. Selon les pays, la connaissance de l'existence de l'un ou l'autre des réseaux peuvent être différentes. *KAD* étant apparu plus récemment qu'*eDonkey*, il peut aussi être moins connu.

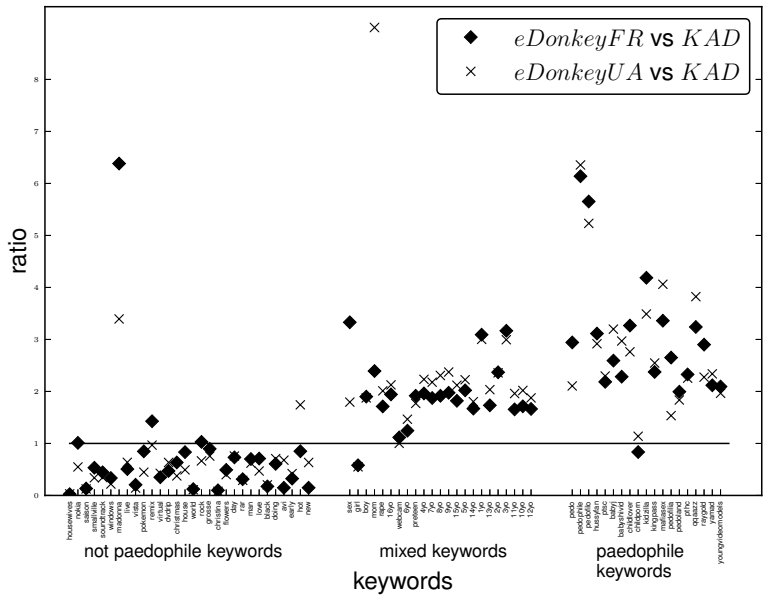


Figure 5.2 – *Rapport des fréquences d'apparition des mots-clefs dans* KAD *et dans* eDonkey*, ordonnés par* coefficient pédophile *croissant. Les points au-dessus de la ligne horizontale* $y = 1$ *indiquent des mots-clefs davantage présents dans* eDonkey *; en-dessous de la ligne se trouvent les mots-clefs plus présents dans* KAD*.*

Nous avons comparé le nombre de requêtes des différentes catégories obtenues pour *KAD* et pour les deux serveurs *eDonkey*. Contrairement à notre supposition initiale, il y a davantage de requêtes pédophiles sur *eDonkey*.

5.3 Indications d'âge

Nous pouvons comprendre plus en détail les échanges pédophiles observés en étudiant la distribution des indications d'âge présentes dans les requêtes [68]. Pour chaque entier n inférieur à 17, nous traçons sur la figure 5.3 le nombre de requêtes de la forme « *n*yo ».

Les trois courbes présentent une allure similaire, avec des valeurs qui augmentent de 1 à 10, une légère inflexion pour 11, un pic à 12 et une chute entre 13 et 16. Les valeurs pour `data-KAD` sont inférieures à celles des ensembles `data-ed2k-FR` et `data-ed2k-UA`, ce qui peut s'expliquer par le fait que l'ensemble est légèrement plus petit et qu'il contient de façon générale moins de requêtes pédophiles. Les requêtes avec des indications d'âge représentent respectivement 8,4%, 15,2% et 17,3% des ensembles `data-KAD`, `data-ed2k-FR` et `data-ed2k-UA` . Remarquons que, si ces âges sont utilisés dans des contextes différents de la pédopornographie, la figure 5.2 montre tout de même que les *coefficients pédophiles* des mots-clefs de la forme *n*yo sont similaires à ceux du groupe `paedophile`, ce qui traduit une apparition fréquente dans les requêtes pédophiles.

Les indications d'âge sont souvent présentes dans les requêtes pédophiles. Les distributions similaires des fréquences dans `data-KAD`, `data-ed2k-FR` et `data-ed2k-UA` indiquent que l'activité pédophile des deux systèmes est de même nature.

5.4 Quantifier l'activité pédophile dans KAD

Nous avons vu que dans `data-KAD` nous n'avons accès qu'à une fraction réduite et biaisée de l'ensemble des requêtes émises dans le système durant la collecte. Ceci ne permet pas d'estimer directement, comme nous l'avons fait au chapitre 2 pour *eDonkey*, la fraction de requêtes pédophiles. Néanmoins, nous

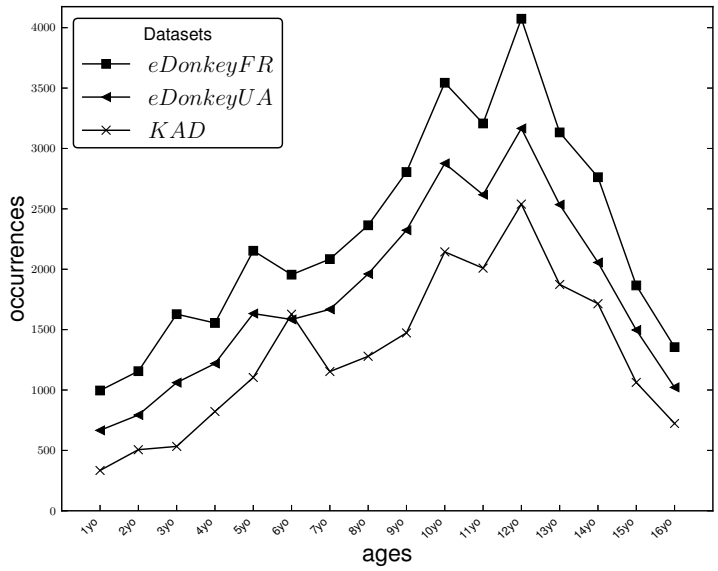


FIGURE 5.3 – *Distribution des indications d'âge pour les trois ensembles de requêtes.*

montrons dans cette section qu'il est possible d'estimer la fraction de requêtes pédophiles dans *KAD* à partir de nos données et du travail effectué pour *eDonkey*.

Notons D l'ensemble de toutes les requêtes d'un jeu de données considéré, P le sous-ensemble de celles-ci qui sont pédophiles, $\overline{Q}$ l'ensemble des requêtes composées d'un seul mot-clef parmi les 72 que nous surveillons dans *KAD* et $\overline{P}$ le sous-ensemble des requêtes composées d'un mot-clef de la liste `paedophile`. On a en particulier $\overline{P} = \overline{Q} \cap P$.

Définissons également $\alpha = \frac{|\overline{Q}|}{|Q|}$ et $\beta = \frac{|\overline{P}|}{|P|}$. Le coefficient α estime la fraction de requêtes d'un mot-clef parmi ceux étudiés dans l'ensemble total des requêtes. Le coefficient β représente la fraction de requêtes d'un mot-clef de la liste `paedophile` parmi toutes les requêtes pédophiles.

Avec ces notations, la fraction que nous cherchons à obtenir est $\frac{|P|}{|Q|}$ et elle peut être obtenue à l'aide de l'expression suivante :

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

Pour les ensembles `data-ed2k-FR` et `data-ed2k-UA` (provenant d'*eDonkey*) nous pouvons disposer directement des valeurs de α, β et de la fraction de requêtes $\frac{|P|}{|Q|}$, comme on l'a vu dans le chapitre 2. Les résultats sur ces ensembles sont présentés dans le tableau 5.1.

En revanche, pour `data-KAD`, nous ne pouvons calculer directement que $\frac{|\overline{P}|}{|\overline{Q}|}$. Cependant, compte tenu de ce qu'estiment α et β, il n'y a pas de raison *a priori*

de supposer que les valeurs de ces coefficients diffèrent significativement entre *eDonkey* et *KAD*, et nous utilisons donc dans la suite pour `data-KAD` les valeurs de $\frac{\alpha}{\beta}$ des ensembles `data-ed2k-FR` et `data-ed2k-UA` (voir tableau 5.1) et obtenons pour *KAD* :

Ensemble	$\frac{\lvert P\rvert}{\lvert Q\rvert}$	$\lvert\overline{P}\rvert$	$\lvert\overline{Q}\rvert$	α	β	$\frac{\alpha}{\beta}$
`data-ed2k-FR`	$0,2554\%$	74 557	241 152	$2,067 \cdot 10^{-3}$	0,2502	$8,256 \cdot 10^{-3}$
`data-ed2k-UA`	$0,2668\%$	46 763	166 154	$2,134 \cdot 10^{-3}$	0,2251	$9,481 \cdot 10^{-3}$
`data-KAD`	n/a	30 821	250 000	n/a	n/a	n/a

Tableau 5.1 – *Caractéristiques de nos ensembles de requêtes.*

$$0.10\% \gtrsim \frac{|P|}{|Q|} \gtrsim 0.12\%.$$

Cette valeur est d'un ordre de grandeur similaire à celle que l'on avait obtenue pour le réseau *eDonkey*, mais deux fois inférieure.

Les explications de cette différence sont les mêmes que celles exposées à la fin de la section 5.2 : la disponibilité des contenus et la demande peuvent varier selon les langues et les périodes de l'année, en fonction des législations nationales, etc.

À l'aide des valeurs obtenues pour *eDonkey* dans le premier chapitre, nous avons estimé la fraction de requêtes pédophiles dans *KAD*. Nous obtenons une valeur proche de 0.11%, qui est donc sensiblement inférieure à celle d'*eDonkey*.

5.5 Conclusion

Nous avons présenté dans ce chapitre la première étude comparative de deux réseaux P2P de grande envergure concernant les échanges pédophiles. Nous avons pour cela préparé des ensembles de données aux caractéristiques identiques, afin de pouvoir comparer *KAD* et *eDonkey* de façon fiable. Nous avons obtenu le résultat contre-intuitif qu'il y a significativement plus de requêtes pédophiles dans *eDonkey* que dans *KAD*, pourtant plus anonyme *a priori*. En revanche, une étude des âges dans les requêtes tend à montrer que l'activité dans les deux réseaux est de nature similaire. Enfin, nous avons estimé la fraction de requêtes pédophiles dans *KAD* à environ 0.11%, ce qui est d'ordre de similaire à celle pour *eDonkey*, mais tout de même près de deux fois moins.

Cette étude est la première du genre sur *KAD*. Elle permet de donner davantage de généralité aux résultats que nous avons obtenus pour le réseau *eDonkey* : l'ordre de grandeur de la fraction de requêtes reste valable entre les deux réseaux.

KAD est toutefois un réseau très proche d'*eDonkey*, puisqu'ils sont utilisés souvent par des logiciels similaires. Quelques aspects de la pédopornographie dans d'autres réseaux tels que *Gnutella* ont été étudiés [38, 50], mais avec des approches différentes. Très récemment, des auteurs ont présenté des analyses similaires aux nôtres pour le réseau *BitTorrent* [64] (concluant que 0,04% environ des requêtes étaient pédophiles sur ce réseau).

Il pourrait être intéressant de refaire une expérience de mesure à l'aide de HAMACK, avec les mêmes mots-clefs, pour comparer les résultats deux années plus tard sur le réseau *KAD*, comme nous l'avons fait sur *eDonkey* dans le chapitre 4. Modifier d'autres paramètres de la collecte, comme augmenter le nombre de mot-clefs étudiés ou la durée de la période d'observation, requiert de pouvoir mobiliser en continu de nombreuses machines réparties dans le monde, ce qui n'est pas évident.

Enfin, un intérêt du réseau *KAD*, qui n'a pas été exploité dans notre travail, est l'existence d'un identifiant pour les utilisateurs, distinct de l'adresse IP. Il est calculé lors de leur première connexion et il est responsable de la localisation dans le réseau. Il serait très intéressant d'obtenir et d'étudier cette information en vue de poursuivre l'exploration de la notion d'utilisateur, voir chapitre 3.

Chapitre 6

Conclusions et perspectives

Dans cette thèse, nous avons utilisé de grands ensembles de requêtes pour étudier l'activité pédophile dans les réseaux P2P.

6.1 Contributions

Nous avons abordé dans un premier temps la problématique de la détection des requêtes pédophiles, dans des ensembles de très grande taille qui en contiennent une proportion très réduite (chapitre 2). Nous avons fait appel aux compétences de spécialistes du domaine pour définir différentes catégories de requêtes pédophiles. Grâce à cette étude préliminaire, nous avons pu **concevoir un outil de détection de requêtes pédophiles**, qui analyse les termes saisis par l'utilisateur et classe chaque requête comme pédophile ou non. Pour évaluer les performances de notre outil, nous avons mis en place une procédure reposant sur la classification d'échantillons de requêtes par des experts indépendants. Notre outil atteint une précision très élevée (plus de 98%) et un bon rappel (proche de 75%). La connaissance des performances de l'outil nous permet d'**estimer la fraction de requêtes pédophiles** dans nos jeux de données, proche de 0,25%, ce qui correspond à une requête pédophile toutes les 33 secondes en moyenne.

La fraction d'utilisateurs entrant des requêtes pédophiles est un indicateur encore plus pertinent, mais l'identification des utilisateurs est en pratique impossible dans notre contexte : nous ne disposons que d'adresses IP, éventuellement accompagnées du port de communication. Nous avons néanmoins exploré différentes manières de **quantifier les utilisateurs soumettant des requêtes pédophiles**. En utilisant l'adresse IP et le port de communication notamment, nous disposons d'une manière satisfaisante de distinguer les requêtes de deux utilisateurs différents. Nous étudions alors les performances de notre outil lorsque l'on passe de la granularité des requêtes à celle des utilisateurs, afin d'obtenir une quantification rigoureuse (chapitre 3). Nous estimons que la fraction d'utilisateurs pédophiles est proche de 0,22%.

Nous avons poursuivi ces travaux de quantification de l'activité pédophile en étudiant son **évolution temporelle** (chapitre 4). Nous avons d'abord présenté l'évolution de la fraction de requêtes pédophiles, qui a notablement augmenté entre 2009 et 2012, alors que le trafic total est resté sensiblement stable, malgré les

changements de législation et l'apparition et la disparition d'alternatives au P2P. Cette augmentation de la fraction de requêtes pédophiles s'accompagne d'une augmentation de la fraction d'utilisateurs pédophiles. Dans un deuxième temps, nous avons étudié un aspect important de la pédopornographie : l'intégration sociale des utilisateurs. Nous avons montré que leurs requêtes suivent un rythme journuit classique mais qu'ils privilégient sensiblement les heures autour de 6 heures du matin pour effectuer des requêtes pédophiles. En cela, ils diffèrent des autres utilisateurs, y compris de ceux qui soumettent des requêtes pornographiques.

Nous avons enfin donné davantage de généralité aux résultats obtenus en proposant **une comparaison de l'activité pédophile entre les réseaux *eDonkey* et *KAD*** (chapitre 5). Nous avons défini une méthodologie pour obtenir des ensembles de requêtes comparables sur chacun des systèmes. Alors que l'on pourrait supposer que le niveau d'anonymat proposé par un système complètement décentralisé comme *KAD* inciterait les utilisateurs se livrant à des échanges illicites à privilégier *KAD*, nous avons obtenu le résultat contraire : les mots-clefs pédophiles étudiés sont significativement plus présents sur *eDonkey* que sur *KAD*. Les activités pédophiles sur les deux réseaux semblent cependant concerner des contenus similaires. Nous avons enfin estimé la fraction de requêtes pédophiles dans *KAD* à l'aide de celle d'*eDonkey* et conclu qu'elle était proche de $0,1\%$.

Les travaux que nous avons présentés dans cette thèse sont en rupture par rapport aux études antérieures de l'activité pédophile. Notre approche quantitative, reposant sur des ensembles de requêtes de très grande taille, ainsi que la rigueur des méthodologies que nous avons développées et mises en œuvre ont permis d'obtenir des statistiques plusieurs ordres de grandeur plus fiables que celles qui étaient disponibles auparavant. Même s'ils peuvent bien sûr être améliorés de nombreuses façons (voir section 6.2.1), nos résultats permettent de prendre la mesure de l'activité pédophile dans le P2P, avec des conséquences directes et indirectes sur les législations sur l'Internet ou la pédopornographie, et sur les actions à mener. L'étude de grands ensembles de requêtes s'avère ainsi très précieuse car elle permet, outre la seule analyse des systèmes dont ils proviennent, d'améliorer significativement les connaissances sur des activités humaines. Comme l'illustre notre analyse de l'évolution journalière de l'activité pédophile, il est même possible d'apporter des éclairages très originaux sur des questions clefs du domaine.

Au-delà des statistiques que nous avons obtenues, notre outil de détection de requêtes constitue en lui-même une contribution significative. Les listes de mots-clefs et la typologie des requêtes pédophiles sont une base de travail pour la mise au point d'outils applicatifs variés (dédiés au filtrage, à l'analyse d'ordinateurs saisis, etc.). De même, les ensembles de requêtes étiquetées comme pédophiles que nous avons obtenus sont aussi une contribution, puisqu'il s'agit de la première fois que des ensembles de cette taille sont disponibles pour l'étude. Pour ces raisons,

nous avons été conviés à présenter nos travaux au siège mondial d'INTERPOL [1] et avons collaboré étroitement avec les forces de l'ordre françaises.

L'étude des utilisateurs que nous avons réalisée apporte un éclairage original sur la façon de rechercher des contenus des pédophiles et leur intégration sociale. Les requêtes capturent en effet l'*intention* des individus, un élément crucial pour comprendre leurs actions et les motivations sous-jacentes. Mais nous avons aussi proposé un travail plus général sur l'identification des utilisateurs dans un contexte où l'information pour les distinguer est très limitée, une question qui se retrouve dans de nombreux contextes. Notre travail constitue un premier pas important dans cette direction et le pousser plus loin est une de nos perspectives principales (voir section 6.2.2).

Outre ces résultats sur l'activité pédophile dans le P2P, cette thèse a étudié un cas particulier d'une problématique beaucoup plus générale : la détection, la quantification et l'étude d'une thématique rare dans de grands ensembles de requêtes. La méthodologie que nous avons employée est spécifique pour certains points, mais elle peut être généralisée dans une large mesure et ainsi être appliquée à de nombreux autres contextes ; c'est une de nos perspectives principales, que nous détaillons ci-dessous (voir section 6.2.3).

6.2 Perspectives

Les résultats de cette thèse ouvrent de très nombreuses perspectives, aussi bien dans la continuité du travail déjà réalisé que dans des directions plus ambitieuses et à plus long terme. Nous présentons ci-dessous celles qui nous semblent les plus prometteuses, regroupées en trois catégories.

6.2.1 Améliorer la détection de l'activité pédophile

Dans un premier temps, il est possible de d'étendre nos travaux sur la détection de requêtes pédophiles de différentes façons.

L'approche de l'apprentissage automatique, que nous avons écartée au départ faute de disposer d'ensemble de requêtes connues comme pédophiles, peut maintenant être abordée. Les ensembles de requêtes que notre outil a détectées comme pédophiles dans nos grands ensembles de requêtes fournissent en effet une base sur laquelle peut s'appuyer une procédure d'apprentissage supervisé. L'ensemble des requêtes étiquetées comme pédophiles par les experts (de taille significativement plus petite toutefois) peut également être utilisé. Il serait ainsi possible d'améliorer la détection des requêtes pédophiles, notamment dans les

1. Conférence du « Groupe spécialisé d'INTERPOL sur la criminalité contre l'enfance », Lyon, 19-22 avril 2010.

cas où notre outil est mis en faute par des flexions linguistiques ou des fautes d'orthographe (si nos listes intègrent par exemple « childern » pour « children », elles ne sont évidemment pas exhaustives). D'autre part, les ensembles de requêtes étiquetées comme pédophiles que nous avons obtenus constituent un corpus qui pourra être analysé avec des méthodes de traitement automatique du langage, afin d'affiner par exemple les catégories de requêtes que nous avons définies (ou en trouver des nouvelles).

L'un des objectifs de cette thèse était de quantifier les requêtes pédophiles, ce qui supposait de connaître les performances de notre outil. Nous les avons estimées grâce à l'évaluation d'experts indépendants, ce qui nous a permis de disposer de résultats fiables. Nous avons ensuite cessé de modifier notre outil. Une perspective intéressante réside dans l'obtention de nouvelles versions de l'outil, lesquelles pourraient attribuer à chaque requête une probabilité d'être pédophile plutôt qu'une réponse binaire. Ceci pourrait être réalisé en pondérant par exemple l'importance de certains mots-clefs ou de certaines catégories, ou en tenant compte de la nature des requêtes précédente et suivante, comme nous l'avons proposé aux experts. En faisant évaluer les résultats ainsi obtenus par des experts, il serait ensuite possible de tester différents seuils de détection et ainsi de développer d'autres versions de notre outil, avec des performances diverses en précision et en rappel.

Enfin, l'évolution temporelle des mots-clefs pédophiles constitue également une problématique que nous n'avons pas traitée en détail mais qui peut jouer un rôle important. Les pédophiles utilisent en effet des mots-clefs codés, afin de dissimuler la nature de leurs échanges, et qui sont évidemment susceptibles de changer au cours du temps : des mots-clefs peuvent apparaître et d'autres disparaître.

6.2.2 Étude des utilisateurs

Approfondir l'étude des utilisateurs est certainement une des perspectives les plus intéressantes de notre travail. Nos travaux sont avant tout quantitatifs, mais il est possible d'aller plus loin.

La littérature sur la pédophilie parle souvent des « communautés » d'utilisateurs pédophiles, généralement sous l'aspect sociologique ou criminologique [16, 47]. À l'aide des données dont nous disposons, il est envisageable de réaliser cette caractérisation des utilisateurs pédophiles de plusieurs façons. Les plus proches de notre travail consistent sans doute à analyser les catégories de requêtes pédophiles que saisissent les utilisateurs ainsi que les moments de la journée auxquels ils procèdent à ces requêtes. Il existe certainement des profils très variés mais l'on peut probablement découvrir des similarités entre utilisateurs. On pourrait aussi modéliser les interactions des utilisateurs sous la forme de graphes (par exemple,

deux utilisateurs sont reliés s'ils emploient un même mot-clef) et analyser ceux-ci avec des méthodes de détection de communautés [24]. Cela pourrait apporter un éclairage très intéressant sur les utilisateurs pédophiles et leurs comportements de recherche : il existe peut-être des sous-ensembles d'utilisateurs qui ne se servent que de quelques mots-clefs particuliers, indiquant des sous-catégories de la pédophilie.

Pour accroître les connaissances sur les utilisateurs pédophiles, il est aussi possible d'analyser en détail les séquences de requêtes entrées par un même utilisateur (pédophile ou non). Nous avons utilisé cette idée de façon très limitée, en présentant aux experts les requêtes précédente et suivante d'une requête à classer (voir section 2.2). Nous avons également examiné plusieurs méthodes pour isoler ces sessions dans notre contexte (réduction de la fenêtre temporelle, adresse IP complétée ou non par le port de communication, voir chapitre 3), mais nous ne les avons pas exploitées en détail. Elles permettent pourtant de mieux comprendre la « stratégie de recherche d'information » [5] d'une personne, ce qui est éclairant pour mieux connaître le profil des utilisateurs pédophiles : effectuent-ils des sessions de recherche uniquement dédiées à cette activité ? Des utilisateurs initialement non pédophiles développent-ils un intérêt pour cette thématique au cours du temps ?

D'autres méthodes d'étude existent et pourraient permettre de compléter nos résultats sur ces utilisateurs. Des mesures actives, telles que la mise en place de « pots de miel » sont par exemple possibles [3]. Un client fictif déclare posséder des contenus pédophiles et enregistre les demandes qui lui sont faites. Ces mesures sont cependant difficiles à mettre en place à grande échelle et sont très encadrées légalement. La collecte et l'analyse des téléchargements effectués pourraient aussi s'avérer très utiles, permettant de savoir notamment dans quelle mesure les utilisateurs qui font des requêtes pédophiles accèdent véritablement aux contenus proposés.

Au-delà de l'étude des utilisateurs pédophiles, l'identification des utilisateurs dans un *contexte Internet* reste un défi à relever. Il serait très intéressant d'appliquer nos méthodes à des données dans lesquelles on sait distinguer les différents utilisateurs (grâce à un couple identifiant/mot de passe par exemple, un *cookie*, etc.). On pourrait alors confronter à la réalité les résultats fournis par nos méthodes.

6.2.3 Aller plus loin

Au-delà des extensions de nos résultats que nous avons présentées ci-dessus, nous pensons qu'une des perspectives majeures de ce travail est sa généralisation méthodologique. En effet, de très nombreux cas pratiques sont très proches du cas que nous avons étudié dans cette thèse. On pourrait par exemple étudier d'autres thématiques rares dans les mêmes ensembles de requêtes ou dans d'autres.

Même si la problématique de la détection ne se pose pas (parce qu'un mot suffit à caractériser les requêtes concernées, typiquement), les autres problématiques (identification des utilisateurs, évolution temporelle, etc.) se posent toujours. On peut penser par exemple à l'activité autour d'un artiste ou d'un courant musical.

Mais les possibilités de généralisation sont beaucoup plus vastes car des domaines très éloignés de l'étude des requêtes présentent aussi des problématiques très similaires. À titre d'exemple, nous pouvons citer la fraude bancaire en ligne [19, 43, 58]. Comme pour la pédopornographie, des réseaux de criminels internationaux sont fortement impliqués dans cette activité et la lutte contre ce phénomène est un enjeu important. Les taux de fraudes sont similaires aux taux de requêtes pédophiles (proches de 0,3%) et leur détection passe par une typologie des fraudes similaire à notre typologie des requêtes pédophiles. Améliorer la connaissance des utilisateurs « fraudeurs » constitue également une problématique centrale et délicate. Leur identification reste difficile : elle repose sur le numéro de carte bancaire, comme l'identification dans notre contexte repose sur l'adresse IP. Un même individu peut utiliser de nombreux numéros de cartes et un même numéro peut être utilisé par plusieurs personnes, comme pour les adresses IP. Dans le cadre de cette généralisation à d'autres problèmes, nous pouvons espérer que des avancées sur un sujet profite aux autres. Les limites et le potentiel de telles généralisations restent toutefois entièrement à explorer.

Annexes

Annexe A

Normalisation et anonymisation des données

Dans cette annexe, nous décrivons les procédures de normalisation et d'anonymisation des données que nous avons utilisées dans notre travail.

Pour *eDonkey*, la collecte a permis de capturer pour chaque requête un certain nombre d'informations : heure de réception, adresse IP de l'utilisateur, port de communication, séquences des mots-clefs. Les données brutes sont, selon les procédures d'enregistrement, éventuellement réparties en différents fichiers (par exemple, un fichier pour chaque heure ou jour de mesure). La syntaxe d'enregistrement diffère également entre les collectes : la mesure `data-ed2k2007` produit des fichiers au format XML, alors que les mesures collectées depuis 2009 donnent des fichiers en texte brut (avec un léger marquage pour distinguer les champs d'information).

Nous avons donc dû construire des ensembles de requêtes à partir de ces données, afin de pouvoir les étudier et les comparer facilement. S'il faut trouver un format de présentation des informations adapté, cela n'est toutefois pas suffisant. Les requêtes contiennent en effet de nombreuses informations personnelles, que nous devons enlever pour satisfaire aux exigences éthiques et légales de notre contexte de travail. Les utilisateurs saisissent dans leurs requêtes des informations sensibles, liées à leur vie privée, telles que des noms (les leurs ou ceux de leurs amis), des numéros de téléphone, voire des numéros de cartes de crédit [1, 4, 53]. L'adresse IP d'où provient une requête est aussi une information personnelle que nous ne pouvons pas conserver en clair. Enfin, des utilisateurs intéressés par des contenus très rares peuvent taper des mots-clefs très spécifiques qui permettent de les identifier [7].

L'anonymisation de requêtes produites par des systèmes informatiques est une problématique à part entière [1, 4, 53]. La difficulté réside dans l'obtention de données riches en informations tout en préservant la vie privée des utilisateurs. Dans notre cas, cela était d'autant plus important que nous souhaitions dès le début rendre disponibles nos jeux de données. Nous avons donc mis en place une procédure de normalisation et d'anonymisation adaptée.

En ce qui concerne les adresses IP, remarquons que l'encodage par une fonction de hashage n'est pas complètement satisfaisant puisqu'un attaquant pourrait

décoder les adresses IP en appliquant la fonction à l'ensemble des 2^{32} adresses possibles (en IPv4). Nous avons donc choisi d'encoder les adresses en fonction de l'ordre dans lequel elles apparaissent dans les données : le première est remplacée par 0, la seconde par 1, et ainsi de suite. Nous procédons de manière similaire pour les numéros de port de communication. Cette procédure d'anonymisation est cohérente, puisqu'on remplace toujours la même adresse ou le même port par le même entier. Quoique coûteuse en mémoire, cette opération permet une anonymisation forte des adresses et rend aisée la manipulation des données par la suite.

Nous normalisons les requêtes en remplaçant tout d'abord tous les caractères accentués par leurs équivalents dépourvus de diacritique. Nous passons ensuite toutes les lettres en bas-de-casse et nous remplaçons tous les caractères non-alphanumériques par des espaces. Les espaces consécutives sont supprimées (et remplacée par une seule). Nous appelons *requêtes normalisées* les requêtes ainsi obtenues. Elles contiennent seulement des séquences de caractères alphanumériques séparés par des espaces, que nous appelons *mots(-clefs)*.

Afin d'anonymiser ces requêtes normalisées, nous devons distinguer entre les informations personnelles (et sensibles) et celles qui ne le sont pas. Nous faisons l'hypothèse qu'un mot saisi par de nombreux utilisateurs différents (et donc dans de nombreuses requêtes) n'est pas sensible [1, 4]. Par exemple, un nom ou un numéro de téléphone n'apparaît que dans peu de requêtes, ou dans beaucoup de requêtes mais saisies par un seul utilisateur. Nous enlevons finalement tous les mots qui apparaissent dans les requêtes normalisées provenant de moins de 50 adresses IP différentes, ce qui assure un bon niveau d'anonymat.

Les nombres, en particulier les petits nombres, apparaissent dans des contextes très divers. En particulier, les numéros de téléphone peuvent apparaître comme des séries de deux ou trois entiers séparées par des espaces. Ces séquences se retrouvent dans les requêtes provenant de plus de 50 adresses IP et ne sont donc pas enlevées lors de l'opération décrite ci-dessus, ce qui pose des problèmes d'anonymat. Une solution consisterait à enlever tous les nombres, mais l'on perdrait alors beaucoup d'information. Les indications d'âge disparaîtraient notamment, alors qu'elles sont cruciales pour la détection de nombreuses requêtes pédophiles (voir chapitres 2 et 5). Nous enlevons donc seulement les nombres de plus de deux chiffres et les nombres supérieurs à 16, ce qui enlève la plupart des nombres mais préserve les âges dont nous avons besoin.

Enfin, les mots-clefs très courts posent des problèmes similaires. En particulier, des problèmes d'encodages induisent parfois que des espaces soient insérées entre deux caractères consécutifs. Des caractères sont alors considérés comme des mots et, puisqu'ils apparaissent souvent, sont préserver. Avec la procédure préalable, cela signifie que de telles requêtes apparaissent en clair. Afin d'éviter cela, nous enlevons tous les mots composés d'une seule lettre. Remarquons que l'on aurait pu enlever également les mots-clefs de deux ou trois lettres, mais cela aurait enlever

trop d'informations importantes de nos données. L'inspection de nos ensembles de requêtes a montré que les conserver ne posait pas de problème majeur concernant la vie privée.

Les caractéristiques finales des jeux de données obtenus sont présentées dans la section 1.2 de ce manuscrit.

Annexe B

Géolocalisation des utilisateurs

Notre étude de l'activité pédophile dans les systèmes P2P nous a conduit à examiner les requêtes d'utilisateurs provenant de différents endroits dans le monde. Nous avons procédé en agrégeant toutes les requêtes, à l'exception de la section 4.2.2 dans laquelle la séparation des requêtes des utilisateurs selon les pays nous a permis de confirmer un résultat sur la dynamique des comportements. Pourtant, l'information relative à la localisation des utilisateurs est précieuse, à plusieurs titres.

Tout d'abord, les législations sont souvent élaborées à l'échelle nationale (ou une échelle transnationale plus ou moins limitée, comme par exemple l'Union Européenne). Les forces de l'ordre opèrent aussi majoritairement une échelle nationale. Analyser la provenance géographique des requêtes permet de connaître l'ampleur du phénomène dans une région donnée et d'évaluer l'efficacité des éventuelles mesures mises en place pour lutter pour le contrer.

Une autre raison, d'ordre technique, motive l'étude de cet aspect géographique : notre capacité à détecter et quantifier l'activité pédophile est directement liée à notre capacité à interprêter convenablement les requêtes examinées. Les problèmes de langues ou d'encodage de caractères rendent notre étude de l'activité pédophile dépendante de la localisation. En se restreignant à un ensemble défini de pays, nous pouvons espérer rendre notre outil plus fiable.

L'objectif de cette étude est d'obtenir une meilleure connaissance de l'activité pédophile et de la répartition géographique des utilisateurs, en nous concentrant sur l'Europe.

Données

Les données que nous utilisons ici proviennent de l'ensemble `data-ed2k0912`, présenté dans le chapitre 1. Il s'agit des premières semaines contenant l'information de géolocalisation, du 29 août 2009 au 14 octobre 2009. Il y a 54 274 002 requêtes, provenant de 214 pays différents (voir la figure B.1 pour une distribution des requêtes par pays). La géolocalisation de l'adresse IP des utilisateurs, réalisée avant l'anonymisation, recourt à la base de données GeoIP de MaxMind [1]. Notre outil de détection classe 77 548 requêtes comme pédophiles dans cet ensemble.

Nous présentons dans un premier temps ...

1. `http://www.maxmind.com`

Statistiques

Remarquons tout d'abord que les utilisateurs des réseaux P2P ne sont pas uniformément répartis dans le monde. La population totale et la population de personnes disposant d'un accès à l'Internet varie grandement et il est notoire que les systèmes P2P les plus populaires ne sont pas les mêmes en Asie, en Europe ou en Amérique du Nord. Notre ensemble de requêtes contient donc vraisemblablement des nombres de requêtes très variables en fonction du pays.

La figure B.1 confirme cette hypothèse en présentant la distribution du nombre de requêtes par pays, qui suit une loi de puissance (qui indique une hétérogénéité importante).

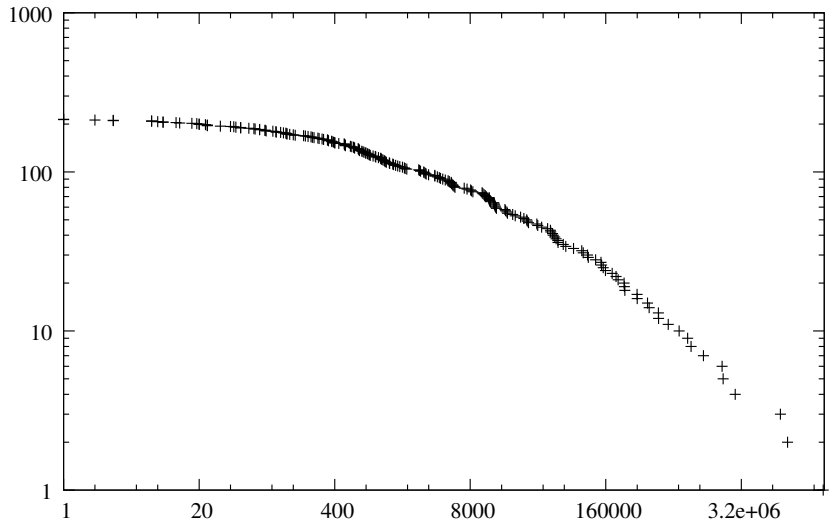


Figure B.1 – *Distribution du nombre de requêtes par pays. Pour chaque valeur x en abscisse, nous traçons le nombre y de pays pour lesquels nous disposons de x requêtes.*

Dans ce contexte, calculer des statistiques sur l'activité pédophile d'un pays n'a de sens que si nous disposons de suffisamment de requêtes provenant de ce pays. Par exemple, le Groenland présente un taux de requêtes pédophiles très élevé mais ce chiffre n'est pas significatif car le nombre de requêtes observé est très faible. Pour éviter ce type de biais, nous restreignons notre étude aux 30 pays qui contiennent plus de 100 000 requêtes, qui représentent 97,05% du total. Les statistiques sont présentées dans le tableau B.1 et dans le tableau B.2, respectivement ordonnés par ordre décroissant de requêtes reçues et de fraction de requêtes détectés comme pédophiles.

Remarquons que les encodages de caractères propres à certains pays (et les langues utilisées) nous amènent à avoir, pour certains pays, beaucoup de requêtes dont très peu sont détectées comme pédophiles.

pays	nb requêtes	nb 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 %
DK	159041	468	0.29 %
GR	150984	536	0.36 %
TR	145714	368	0.25 %
CL	143785	299	0.21 %
JP	127915	178	0.14 %
VE	108758	380	0.35 %
AU	106882	401	0.38 %

Tableau B.1 – *Pour chaque pays de notre jeu de données pour lequel nous possédons suffisamment d'information : nombre de requêtes reçues pour ce pays, nombre de requêtes pédophiles et fraction de requêtes pédophiles.*

Cartes

Afin de visualiser d'une manière plus intuitive et plus attrayante les résultats précédents, nous traçons des cartes qui reflètent ces statistiques. Les figures B.2, B.3 et B.4 présentent des cartes de l'Europe sur lesquelles les couleurs traduisent, pour chaque pays, le nombre de requêtes collectées, le nombre de requêtes pédophiles et la fraction de requêtes pédophiles. Plus la couleur tend vers le noir,

pays	nb requêtes	nb pédo.	fraction
RU	200958	2712	1.35 %
US	1053183	6184	0.59 %
DE	2093106	11238	0.54 %
CH	320054	1710	0.53 %
GB	319386	1698	0.53 %
AT	184248	977	0.53 %
MX	210504	1098	0.52 %
CA	241460	1233	0.51 %
NL	243646	1131	0.46 %
AU	106882	401	0.38 %
GR	150984	536	0.36 %
VE	108758	380	0.35 %
DK	159041	468	0.29 %
TR	145714	368	0.25 %
CL	143785	299	0.21 %
AR	810466	1465	0.18 %
BR	2795090	4849	0.17 %
BE	402565	646	0.16 %
JP	127915	178	0.14 %
IL	2139697	2618	0.12 %
PL	975170	1178	0.12 %
FR	7583815	8059	0.11 %
PT	513327	434	0.08 %
IT	19569361	15426	0.08 %
SI	239572	167	0.07 %
ES	8881405	5177	0.06 %
CN	635392	337	0.05 %
IE	511185	54	0.01 %

Tableau B.2 – *Pour chaque pays de notre jeu de données pour lequel nous possédons suffisamment d'information : nombre de requêtes reçues pour ce pays, nombre de requêtes pédophiles et fraction de requêtes pédophiles.*

plus la valeur est élevée. Les pays pour lequel nous manquons d'informations sont laissés en blanc.

Nous constatons que les pays d'Europe présentent des situations très diverses, avec des fractions de requêtes pédophiles très significativement plus élevées dans certains pays. Les pays d'Europe du Sud et de l'Ouest apparaissent comme ayant soumis le plus de requêtes au serveur. En revanche, les fractions de requêtes pédophiles observées sont plus importantes pour les pays d'Europe centrale et de l'Est (à l'exception de la Pologne) et le Royaume-Uni.

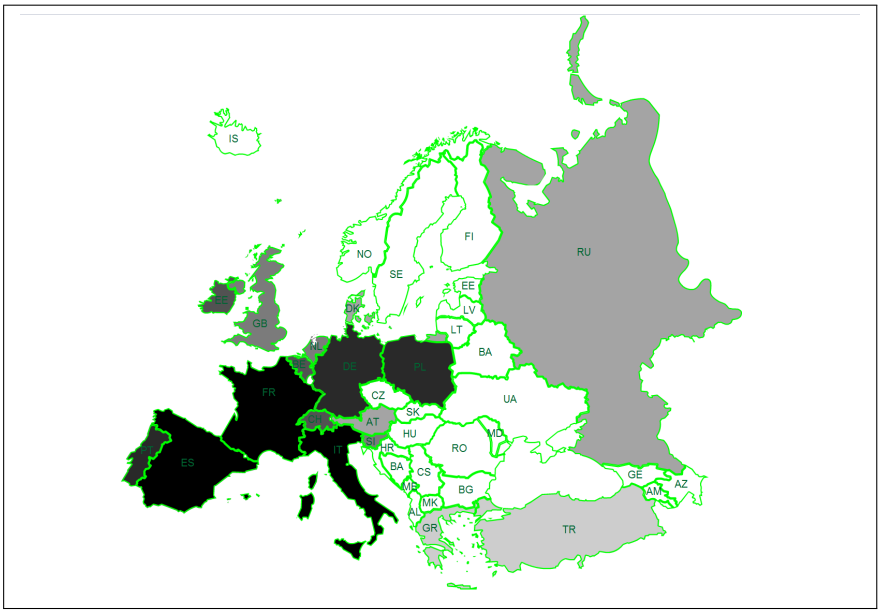


Figure B.2 – *Carte de l'Europe, sur laquelle chaque pays est coloré en fonction du nombre de requêtes reçues d'utilisateurs de ce pays.*

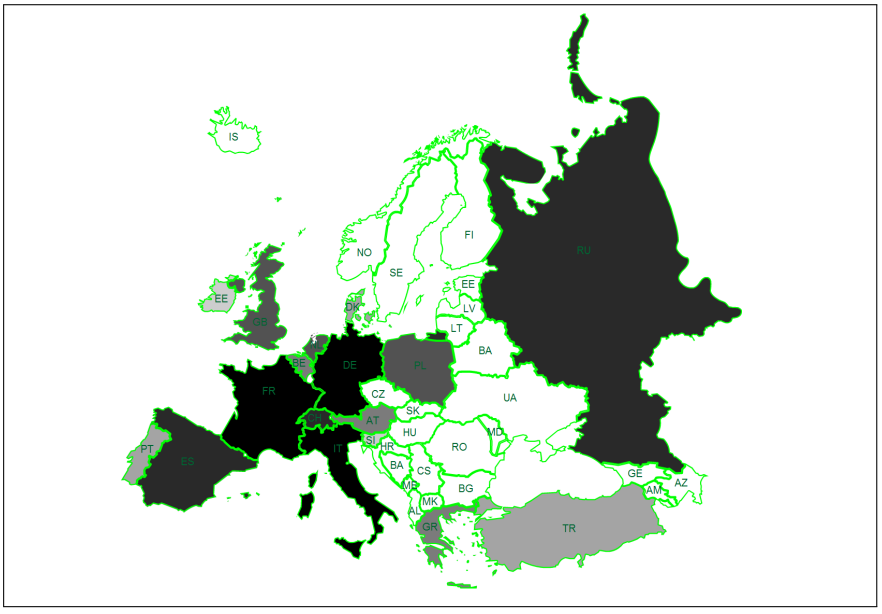


Figure B.3 – *Carte de l'Europe, sur laquelle chaque pays est coloré en fonction du nombre de requêtes pédophiles provenant des utilisateurs de ce pays.*

Bilan

Ces résultats reposent sur la géolocalisation de l'adresse IP, une information sujette à caution. En effet, il est assez facile techniquement d'obtenir une adresse

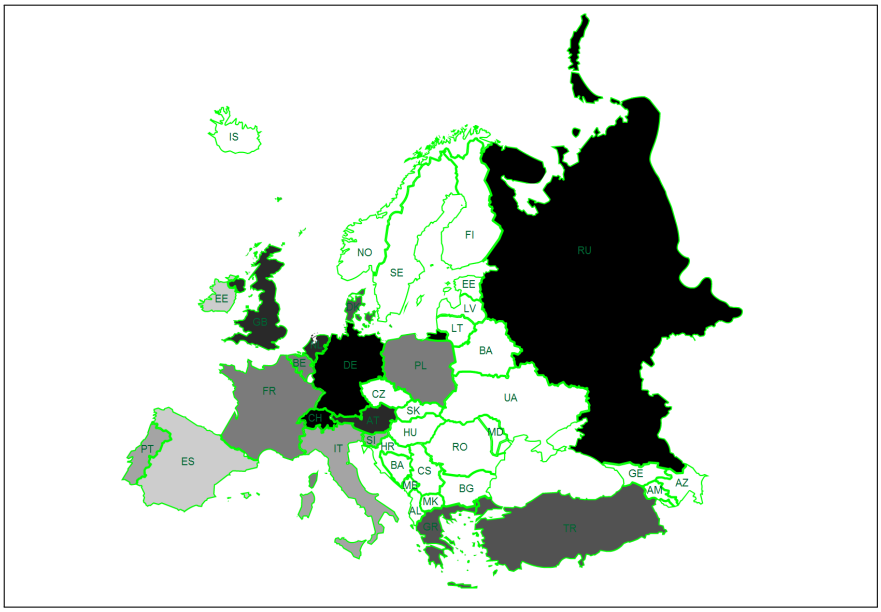


Figure B.4 – *Carte de l'Europe, sur laquelle chaque pays est coloré en fonction de la fraction de requêtes pédophiles provenant des utilisateurs de ce pays.*

IP dans un pays différent (nœud TOR, VPN), ce qui biaise les résultats. Si les utilisateurs pédophiles d'Europe de l'Ouest utilisaient massivement des VPN situés dans des pays d'Europe de l'Est, par exemple, cela pourrait expliquer (au moins en partie) ce que nous observons sur ces cartes. La géolocalisation souffre aussi des performances (notamment de la fréquence de mise à jour) de l'outil MaxMind.

Ces résultats constituent cependant une première étape pour mieux connaître quantitativement l'activité pédophile en Europe.

Annexe C

Catégories de requêtes pédophiles

Lors de l'étude préliminaire que nous avons réalisée pour concevoir notre outil de détection de requêtes pédophiles, nous avons identifié quatre catégories de requêtes pédophiles (voir chapitre 2). Nous avons ensuite utilisé ces différentes catégories pour réaliser des tests lexicaux sur les requêtes et détecter celles qui sont pédophiles.

Nous nous sommes cependant concentrés sur la nature de la requête, pédophile ou non pédophile. Notre outil effectue ainsi les tests de classification séquentiellement et affecte une seule catégorie à chaque requête. Une requête qui appartient aux catégories 1, 2 et 3 est alors classée seulement dans la catégorie 1 [1]. Nous avons modifié l'outil pour obtenir des statistiques sur la répartition des requêtes pédophiles en catégories, pour le jeu de données `data-ed2k0912`. Celle-ci est présentée dans le tableau C.1.

Ce tableau révèle de grandes disparités en général entre les différentes catégories : près de 77% des requêtes détectées comme pédophiles entrent dans la catégorie 1, 15% et 7% environ sont dans les catégories 3 et 2 respectivement. La catégorie 4 ne contient que 0.8% des requêtes. Et moins de 1.8% des requêtes appartiennent à plus d'une catégorie.

Nous observons ensuite l'évolution hebdomadaire des nombres de requêtes correspondant à chacune des catégories, présentée sur les figures C.1, C.2, C.3 et C.4. Les nombres de requêtes hebdomadaires sont très différents pour chacune des catégories : alors que jusqu'à 35 000 requêtes d'une semaine peuvent être dans la catégorie 1, seulement 500 au maximum peuvent être classées dans la catégorie 4. Les catégories 2 et 3 sont d'ordre de grandeur similaires, avec quelques milliers de requêtes par semaines dans chacune, la catégorie 3 ayant tout de même deux fois plus de requêtes environ que la catégorie 2.

Les quatre catégories de requêtes évoluent aussi différemment. Le nombre de requêtes dans la catégorie 4 a presque doublé entre mi-2009 et début 2012, mais présente en moyenne une stabilité pendant une grande partie de la mesure. Le profil d'évolution des catégories 1 et 2 sont similaires (figures C.1 et C.2), avec des

1. Nous identifions les catégories selon l'ordre de présentation de la figure 2.1. La catégorie 1 correspond ainsi aux requêtes contenant un mot-clef de la liste *explicit*, la catégorie 2 à celles qui contiennent un mot-clef de *child* et un de *sex*.

catégories	nb de requêtes	part des requêtes pédophiles
1	3 241 756	75,5334
3	645 404	15,038
2	294 684	6,86618
1,3	36 138	0,84202
4	33 919	0,790317
1,2	16 128	0,375785
2,3	8 577	0,199845
1,2,3	6 953	0,162006
3,4	3 697	0,0861406
1,4	1 694	0,0394704
2,4	1 447	0,0337153
1,3,4	789	0,0183838
1,2,4	250	0,00582503
2,3,4	235	0,00547553
1,2,3,4	149	0,00347172

Tableau C.1 – *Répartitions des requêtes dans les différentes catégories avec une version modifiée de l'algorithme autorisant la classification dans plusieurs catégories.*

pics d'activités aux mêmes moments et une multiplication par 3 des nombres de requêtes entre le début et la fin du jeu de données. Le nombre de requêtes dans la catégorie 3 augmente régulièrement mais ne présente pas les mêmes pics d'activité que les catégories 1 et 2 (figure C.3).

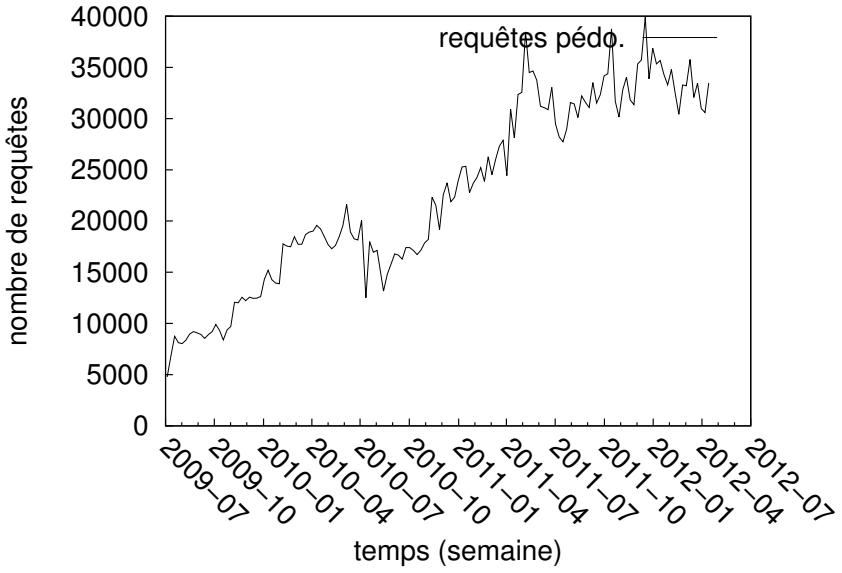


Figure C.1 – *Nombre de requêtes par semaine pour la catégorie 1*

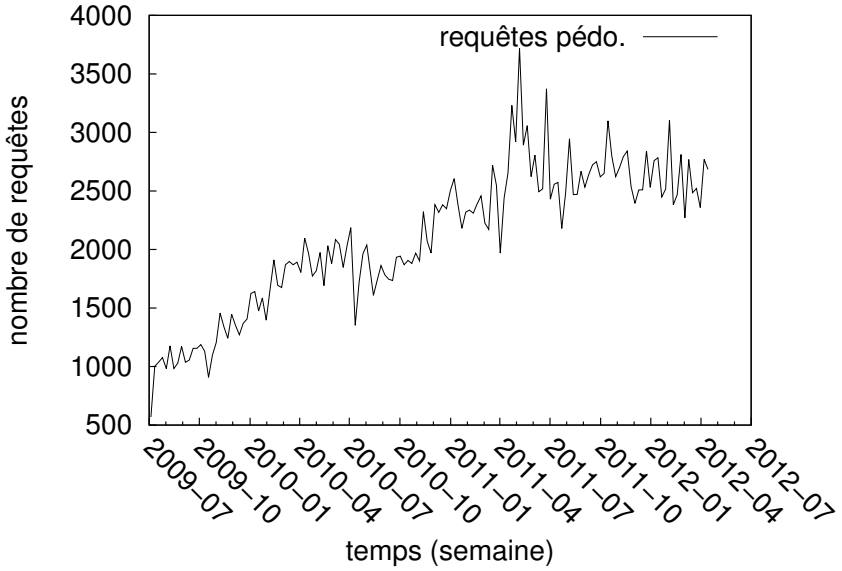


Figure C.2 – *Nombre de requêtes par semaine pour la catégorie 2*

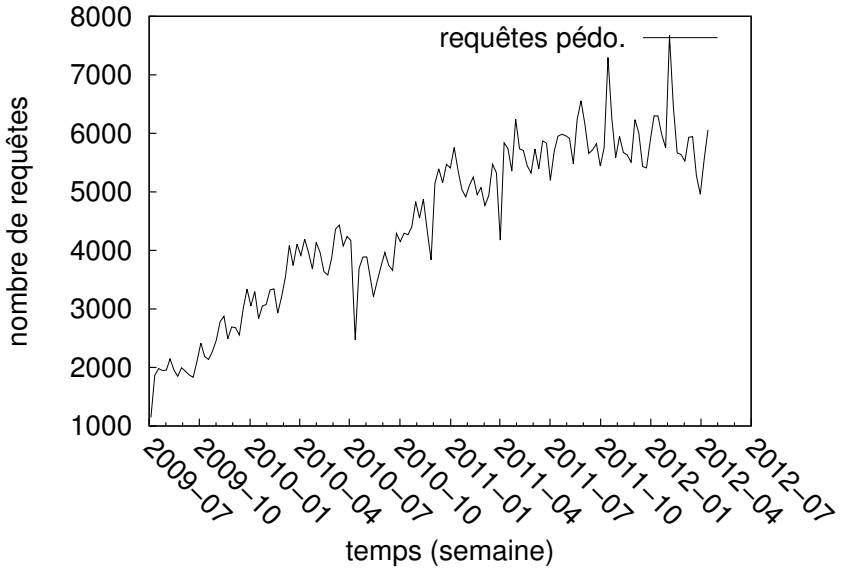


Figure C.3 – *Nombre de requêtes par semaine pour la catégorie 3*

Bilan

Comme nous l'avons dit en conclusion de ce manuscrit, les catégories de requêtes pédophiles sont constituent une contribution pour la détection et la quantification de l'activité pédophile dans le P2P. Il reste cependant de nombreuses pistes à explorer et cela pourrait notamment permettre d'améliorer les connaissances sur les utilisateurs pédophiles.

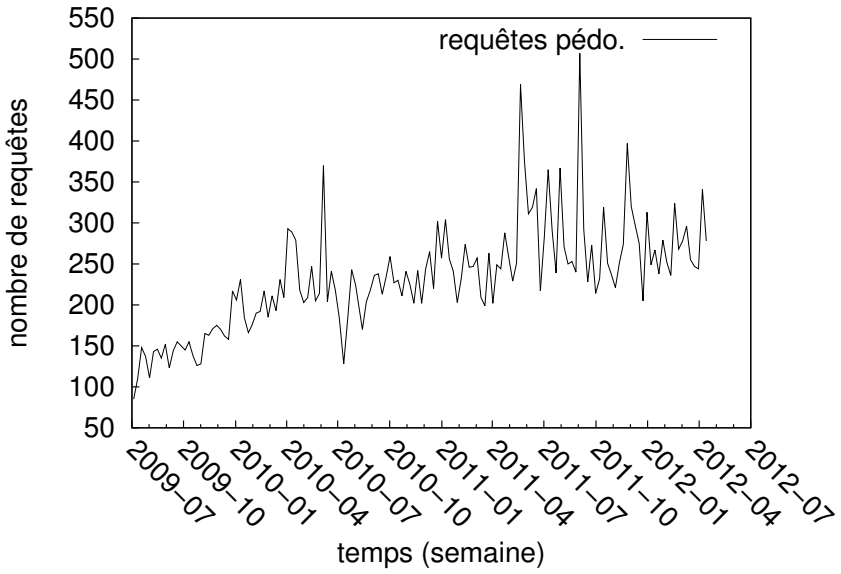


Figure C.4 – *Nombre de requêtes par semaine pour la catégorie 4*

Annexe D

Listes de mots-clefs utilisés par notre algorithme

---

explicit babyj, babyshivid, childlover, childporn, childsex, childfugga, ddoggprn, hussyfan, kdquality, kidzilla, kingpass, mafiasex, pedo, pedofilia, pedofilo, pedoland, pedophile, pedophilia, pedophilie, pthc, ptsc, qqaazz, raygold, ygold, reelkiddymov, yamad, youngvideomodels.

child adolescent, adolescente, child, childern, children, childrens, childs, enfant, fillette, gamine, infant, infantil, infantile, infantiles, kid, kiddy, kids, kinder, kindergarten, menor, menores, mineur, mineure, mineures, mineurs, nino, ninos, ninas, preteen, preteens, underage.

sex abuse, abused, abuso, anal, animalsex, ass, assfuck, asslick, avale, bath, bibcam, bitch, blowjob, cum, cumshot, defloration, dildo, dogsex, encule, enculer, eurosex, ficken, ficker, fickt, fuck, fucked, fucks, fucking, gay, groupsex, handjob, hard, hardcore, homemade, incest, inzest, kdv, lesbian, lickin, licking, loli, lolita, lolitaguy, lolitas, lolitasex, lover, masterbate, masterbates, masterbating, masterbation, masturb, masturbate, masturbates, masturbating, masturbation, masturbe, nackt, nackte, nackten, naked, naturist, nude, nudist, nudiste, orgasm, penetration, penis, porn, porno, prostitute, prostitue, pussy, rape, raped, salope, sado, sex, sexe, sexo, sexual, sexually, shower, sodom, sodomie, sodomise, sodomy, sodomized, spank, spanked, spanking, sperm, suce, suck, sucker, sucks, swallow, teensex, transexual, viol, viole, violee, webcam, whore, xxx, zoofilia, amamter, amateur, amateurs, amatoriale, amatrice, anale, anus, arsch, baise, bdsm, bondage, chatte, culo, cunt, ejac, ejaculation, erotic, exhib, facial, fellation, fetish, fick, fisting, gangbang, gode, hentai, hure, lesbienne, lingerie, naakt, orgy, partouze, piss, pornstar, putas, pute, scato, shemale, sm, soumise, sperma, sperme, suceuse, tournante, uro, vicieuse, voyeur.

familychild baby, bebe, boy, daughter, fille, girl, son, toddler.

familyparents dad, daddy, father, grandpa, grandma, mom, mommy, mum, mummy.

agesuffix yo, yr, an, ans, anni, anos, ano, jahr, jahre, jahres, old, yearold, old, yearsold.

Bibliographie

[1] Eytan Adar. User 4xxxxx9 : Anonymizing query logs. In *Query Logs Analysis Workshop, International Conference on World Wide Web*, 2007. (Cité en pages 93 et 94.)

[2] Frédéric Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an eDonkey server. *International Workshop on Hot Topics in P2P Systems*, 2009. (Cité en page 16.)

[3] O. Allali, M. Latapy, and C. Magnien. Measurements of *eDonkey* activity with distributed honeypots. *HotP2P'09*, 2009. (Cité en page 89.)

[4] Mark Allman and Vern Paxson. Issues and etiquette concerning use of shared measurement data. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2007. (Cité en pages 93 et 94.)

[5] H. Assadi and V. Beaudouin. Comment utilise-t-on les moteurs de recherche sur internet ? *Réseaux*, 20(116) :171–198, 2002. (Cité en page 89.)

[6] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision)*. American Psychiatric Association Publications, 4th edition, July 2000. (Cité en page 22.)

[7] Michael Barbaro and Tom Zeller, Jr. A face is exposed for AOL searcher no. 4417749. *New York Times*, August 2006. `https://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0`. (Cité en page 93.)

[8] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David A. Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized Web query log. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *SIGIR*, pages 321–328. ACM, 2004. (Cité en pages 66 et 71.)

[9] Lamia Benamara and Clémence Magnien. Estimating properties in dynamic systems : the case of churn in P2P networks. In *Second International Workshop on Network Science for Communication Networks (NetSciSom 2010), in conjuction with IEEE INFOCOM*, 2010. (Cité en page 32.)

[10] Ranjita Bhagwan, Stefan Savage, and Geoffrey M. Voelker. Understanding availability. In *International Workshop on Peer-To-Peer Systems (IPTPS)*, 2003. (Cité en page 47.)

[11] Deborah D. Blecic, Nirmala S. Bangalore, Josephine L. Dorsch, Cynthia L. Henderson, Melissa H. Koenig, and Ann C. Weller. Using transaction log analysis to improve OPAC retrieval results. *College and Research Libraries*, 59(1) :39–50, 1998. (Cité en page 19.)

[12] Rene Brunner. A performance evaluation of the Kad-protocol. Master's thesis, University of Mannheim and Institut Eurécom, Sophia-Antipolis, France, Nov. 2006. `http://www.eurecom.fr/~btroup/BThesis/MasterThesisBrunner.pdf`. (Cité en pages 15 et 76.)

[13] Guillaume Champeau. Hadopi 2 : Alliot-Marie met sur le même plan pédophilie et piratage, 2009. `http://www.numerama.com/magazine/13514-hadopi-2-alliot-marie-met-sur-le-meme-plan-pedophilie-et-piratage.html`. (Cité en page 12.)

[14] Thibault Cholez. *Supervision des réseaux pair à pair structurés appliquée à la sécurité des contenus*. ThÈse, Université Henri Poincaré – Nancy I, June 2011. (Cité en pages 16 et 76.)

[15] Thibault Cholez, Isabelle Chrisment, and Olivier Festor. Monitoring and Controlling Content Access in KAD. In *International Conference on Communications - ICC 2010*, Capetown, Afrique Du Sud, May 2010. IEEE. (Cité en pages 16 et 76.)

[16] Patrice Corriveau. Les groupes de nouvelles à caractère pédopornographique : une sous-culture de la déviance. *Déviance et Société*, 34(3) :381–400, 2010. (Cité en pages 12, 23, 61 et 88.)

[17] W. Bruce Croft, Robert Cook, and Dean Wilder. Providing government information on the Internet : Experiences with THOMAS. In *The Second International Conference on the Theory and Practice of Digital Libraries*, 1995. (Cité en page 19.)

[18] James Delahunty. eD2K razorback servers seized. *Afterdawn*, 2006. `http://afterdawn.com`. (Cité en page 12.)

[19] E-Fraud. `http://www.systematic-paris-region.org/fr/projets/e-fraud`, 2012. (Cité en page 90.)

[20] eMule-project. `http://www.emule-project.net/`. (Cité en page 15.)

[21] Fabrice Epelboin. *Le commerce de la pédopornographie sur Internet de 2000 à 2010*. Read Write Web, 2010. (Cité en page 23.)

[22] Paulo Fagundes. Fighting internet child pornography – the Brazilian experience. *The Police Chief Magazine*, LXXVI(9), 2009. (Cité en page 23.)

[23] Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8) :861–874, June 2006. (Cité en page 22.)

[24] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5) :75 – 174, 2010. (Cité en page 89.)

[25] Richard Frank, Bryce Westlake, and Martin Bouchard. The structure and content of online child exploitation networks. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ISI-KDD '10, pages 1–9, New York, NY, USA, 2010. ACM. (Cité en page 23.)

[26] A. Frei, N. Erenay, V. Dittmann, and M. Graf. Paedophilia on the Internet – a study of 33 convicted offenders in the Canton of Lucerne. *Swiss Medical Weekly*, 135 :488–494, 2005. (Cité en pages 12 et 66.)

[27] Diego Gambetta. *Codes of the Underworld : How Criminals Communicate*. Princeton University Press, 2011. (Cité en page 13.)

[28] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457 :1012–1014, 2009. (Cité en pages 11 et 19.)

[29] Adam Shaked Gish, Yuval Shavitt, and Tomer Tankel. Geographical statistics and characteristics of P2P query strings. In *International Workshop on Peer-to-Peer Systems (IPTPS)*, 2007. (Cité en page 19.)

[30] Google. Zeitgeist. `http://www.google.com/zeitgeist/`, 2012. (Cité en page 19.)

[31] Krishna P. Gummadi, Stefan Saroiu, and Steven D. Gribble. A measurement study of Napster and Gnutella as examples of peer-to-peer file sharing systems. *Computer Communication Review*, 32(1) :82, 2002. (Cité en page 19.)

[32] M. Hancock-Beaulieu, S. Robertson, C. Neilson, British Library. Research, and Development Dept. *Evaluation of Online Catalogues : An Assessment of Methods*. British Library Research Paper. British Library Research and Development Department, 1990. (Cité en page 19.)

[33] Sidath B. Handurukande, Anne-Marie Kermarrec, Fabrice Le Fessant, Laurent Massoulié, and Simon Patarin. Peer sharing behaviour in the eDonkey network, and implications for the design of server-less file sharing systems. In *EuroSys*, 2006. (Cité en page 19.)

[34] R. Karl Hanson and Heather Scott. Social networks of sexual offenders. *Psychology, Crime and Law*, 2(4) :249–258, 1996. (Cité en page 23.)

[35] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, 21(9) :1263–1284, 2009. (Cité en page 21.)

[36] Daniel Hughes, James Walkerdine, Geoff Coulson, and Stephen Gibson. Is deviant behavior the norm on P2P file-sharing networks ? *IEEE Distributed Systems Online*, 7(2), 2006. (Cité en pages 19 et 23.)

[37] Danny Hughes, Paul Rayson, James Walkerdine, Kevin Lee, Phil Greenwood, Awais Rashid, Corinne May-Chahal, and Margaret Brennan. Supporting law enforcement in digital communities through natural language analysis. In *International Workshop on Computational Forensics*, 2008. (Cité en page 24.)

[38] Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls, Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine, Marc Liberatore, and Brian

Lynn. Measurement and analysis of child pornography trafficking on Gnutella and eMule. Technical Report UM-CS-2012-016, University of Massachusetts Amherst, Department of Computer Science, May 2012. (Cité en page 84.)

[39] B. J. Jansen. Search log analysis : What is it ; what's been done ; how to do it. *Library and Information Science Research*, 28(3) :407–432, 2006. (Cité en page 20.)

[40] B. J. Jansen and U Pooch. Web user studies : A review and framework for future work. *Journal of the American Society for Information Science and Technology*, 52(3) :235 – 246, 2001. (Cité en page 20.)

[41] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs : A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2) :207–227, 2000. (Cité en page 19.)

[42] Kirsty Johnston. FBI says child porn found on Dotcom's Megaupload servers, 2012. `http://www.stuff.co.nz/world/americas/6995377/FBI-says-child-porn-found-on-Dotcoms-Megaupload-servers`. (Cité en page 12.)

[43] Piotr Juszczak, Niall M. Adams, David J. Hand, Christopher Whitrow, and David J. Weston. Off-the-peg and bespoke classifiers for fraud detection. *Comput. Stat. Data Anal.*, 52(9) :4521–4532, May 2008. (Cité en page 90.)

[44] Alexander Klemm, Christoph Lindemann, Mary K. Vernon, and Oliver P. Waldhorst. Characterizing the query behavior in peer-to-peer filesharing systems. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2004. (Cité en page 19.)

[45] Linda D. Koontz. File sharing programs – child pornography is readily accessible over peer-to-peer networks, 2003. Report of the US General Accounting Office. (Cité en page 23.)

[46] Yoram Kulbak and Danny Bickson. The eMule Protocol Specification. Technical report, The Hebrew University of Jerusalem, Jerusalem, School of Computer Science and Engineering, 2005. (Cité en page 15.)

[47] K.V. Lanning and A.W. Burgess. *Child pornography and sex rings*. Federal Bureau of Investigation, U.S. Dept. of Justice, 1984. (Cité en pages 23 et 88.)

[48] 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. (Cité en page 16.)

[49] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, In press, 2012. (Cité en page 16.)

[50] Marc Liberatore, Robert Erdely, Thomas Kerle, Brian Neil Levine, and Clay Shields. Forensic Investigation of Peer-to-Peer File Sharing Networks. In *Proc. DFRWS Annual Digital Forensics Research Conference*, August 2010. (Cité en page 84.)

[51] Loi HADOPI. *Légifrance*, 2012. (Cité en page 62.)

[52] Projets MAPE et MAPAP. `http://antipaedo.lip6.fr/`, 2012. (Cité en page 28.)

[53] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large datasets. In *IEEE Symposium on Security and Privacy*, 2008. (Cité en page 93.)

[54] Linh Thai Nguyen, Dongmei Jia, Wai Gen Yee, and Ophir Frieder. An analysis of peer-to-peer file-sharing system queries. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2007. (Cité en page 19.)

[55] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems*. World Health Organization, 2007. 10th Revision, `http://apps.who.int/classifications/apps/icd/icd10online/`. (Cité en page 22.)

[56] Thomas A. Peters. The history and development of transaction log analysis. *Library Hi Tech*, 11(2) :41 – 66, 1993. (Cité en page 18.)

[57] A. Phippen, L. Sheppard, and S. Furnell. A practical evaluation of Web analytics. *Internet Research : Electronic Networking Applications and Policy*, 14 :284–293, 2004. (Cité en page 19.)

[58] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of Data Mining-based Fraud Detection Research. *Artificial Intelligence Review*, 2005. (Cité en page 90.)

[59] Associated Press. Megaupload's Kim Schmitz arrested in Auckland, site shut down. *3 News*, 2012. [En ligne ; accès le 22 Juillet 2012]. (Cité en page 63.)

[60] Ethel Quayle, Lars Loof, and Tink Palmer. Child pornography and sexual exploitation of children online. In *World Congress III against Sexual Exploitation of Children and Adolescents*, 2008. (Cité en page 23.)

[61] Jerry Ropelato. Internet pornography statistics. *TopTenReviews*, 2007. (Cité en page 70.)

[62] Romain Roubaty. Rapport technique d'expertise de l'affaire Razorback, 2007. `http://www.numerama.com/media/pdf/Ratiatum-Razorback_expertise.pdf`. (Cité en page 12.)

[63] Yves Roumazeilles. P2P – lequel choisir ? `http://www.roumazeilles.net/news/nw/news0074.php`, 2008. (Cité en page 75.)

[64] Moshe Rutgaizer, Yuval Shavitt, Omer Vertman, and Noa Zilberman. Detecting pedophile activity in BitTorrent networks. In Nina Taft and Fabio Ricciato, editors, *PAM*, volume 7192 of *Lecture Notes in Computer Science*, pages 106–115. Springer, 2012. (Cité en pages 66, 70 et 84.)

[65] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. (Cité en pages 11 et 19.)

[66] Subhabrata Sen and Jia Wang. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Transactions on Networking*, 12(2) :219–232, 2004. (Cité en page 19.)

[67] Yuval Shavitt and Udi Weinsberg. Song clustering using peer-to-peer co-occurrences. In *IEEE International Symposium on Multimedia*, 2009. (Cité en page 19.)

[68] Chad M.S. Steel. Child pornography in peer-to-peer networks. *Child Abuse & Neglect*, 33(8) :560 – 568, 2009. (Cité en pages 19, 23, 24 et 81.)

[69] Daniel Stutzbach and Reza Rejaie. Understanding churn in peer-to-peer networks. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2006. (Cité en page 47.)

[70] The Tor Project, Inc. TOR project. `https://www.torproject.org/`. (Cité en page 80.)

[71] Flint Waters. Child sex crimes on the Internet, 2007. Report of State of Wyoming Attorney General. (Cité en page 23.)

[72] Lolicon — Wikipedia. `https://en.wikipedia.org/wiki/Lolicon`, 2012. (Cité en page 23.)

[73] Pédophilie — Wikipedia. `https://fr.wikipedia.org/wiki/Pédophilie`, 2012. (Cité en page 22.)

[74] Richard Goldberg — Wikipedia. `https://en.wikipedia.org/wiki/Raygold`, 2012. (Cité en page 28.)

[75] Temps universel coordonné — Wikipedia. `https://fr.wikipedia.org/wiki/Temps_universel_coordonné`, 2012. (Cité en page 70.)

[76] Usenet — Wikipedia. `https://fr.wikipedia.org/wiki/Groupe_de_nouvelles`, 2012. (Cité en page 23.)

[77] Janis Wolak, Kimberley Mitchell, and David Finkelhor. Child-pornography possessors arrested in internet-related crimes : Findings from the national juvenile online victimization study, 2005. Report of the National Center for Missing & Exploited Children (NCMEC). (Cité en page 61.)

[78] Janis Wolak, Kimberley Mitchell, and David Finkelhor. Online victimization of youth : five years later, 2006. Report of the National Center for Missing & Exploited Children (NCMEC). (Cité en page 23.)

[79] Yahoo ! Buzz log. 2012. (Cité en page 19.)

[80] Équipe ComplexNetworks. Antipaedo project website. `http://antipedo.lip6.fr`, 2012. (Cité en page 17.)