

# Quantifying Paedophile Activity in a Large P2P System

Matthieu Latapy, Clémence Magnien, and Raphaël Fournier  
CNRS and UPMC – 4 place Jussieu, 75005 Paris, France  
{*firstname.lastname*}@lip6.fr

**Abstract**—Increasing knowledge of paedophile activity in P2P systems is a crucial societal concern, with important consequences on child protection, policy making, and internet regulation. Because of a lack of traces of P2P exchanges and rigorous analysis methodology, however, current knowledge of this activity remains very limited. We consider here a widely used P2P system, eDonkey, and focus on two key statistics: the fraction of paedophile queries entered in the system and the fraction of users who entered such queries. We collect hundreds of millions of keyword-based queries; we design a paedophile query detection tool for which we establish false positive and false negative rates using assessment by experts; with this tool and these rates, we then estimate the fraction of paedophile queries in our data; finally, we design and apply methods for quantifying users who entered such queries. We conclude that approximately 0.25 % of queries are paedophile, and that more than 0.2 % of users enter such queries. These statistics are by far the most precise and reliable ever obtained in this domain.

## I. INTRODUCTION

It is widely acknowledged that peer-to-peer (P2P) file exchange systems host large amounts of paedophile content (mainly movies and pictures), which is a crucial societal concern. In addition to children victimisation, the wide availability of paedophile material is a great danger for regular users (including children and teenagers), who may be exposed unintentionally to extremely harmful content. In particular, this may lead initially innocent users to develop an interest in child pornography. It also has a strong impact on the public acceptance of paedophilia and induces a trivialisation of such content. Much work is devoted to these psychological and societal issues, see [1], [2].

Downloading and/or providing paedophile content is a legal offence in many countries, and there is a correlation between downloading paedophile content and having actual sexual intercourse with children. This makes fighting these exchanges a key issue for law enforcement [3], [4]. This also has much impact on P2P and internet regulation, and is used as a key allegation against people providing P2P facilities. For instance, people providing indexes of files available in P2P systems (including a small fraction of files with paedophile content) are often accused of helping and promoting paedophile exchanges, with strong penal threats [5], [6].

For these reasons, knowledge of paedophile activity in P2P systems is a critical resource for law enforcement, child protection and policy making. See [1], [3], [2], [4] for surveys on these issues. However, current knowledge on this activity and

its extent remains very limited and is subject to controversy [7], [8], [9], [10], [3], [2].

In this paper, we provide ground truth on paedophile activity in a large P2P system, at an unprecedented level of accuracy and reliability. We focus on two basic yet crucial statistics: the fraction of paedophile queries entered in the system and the fraction of users entering such queries. We establish reference methodology and tools for obtaining these values, and provide them in the case of the *eDonkey* system, which is one of the largest P2P systems currently used [11].

Obtaining precise such information on paedophile activity in P2P systems raises several challenges:

- *Appropriate data collection.* Obtaining large-scale data of activity in P2P systems is a difficult task in itself. The main reasons are the lack of central authority, the size of these systems and their high dynamics, the poor structure of the traffic, and limited user identification.
- *Paedophile activity identification.* As the relative amount of paedophile activity in P2P systems is very low, quantifying it by manually inspecting a random sample of the data is not feasible: this sample would have to be very large in order to contain a significant amount of paedophile activity. Moreover, this activity is often hidden (paedophiles use very specific keywords), and recognising it requires a deep expertise of the domain. Finally, machine learning approaches, though appealing, cannot be applied in this context because of the lack of prior knowledge of representative paedophile data.
- *Rigorous inference of statistics.* In a context where detection of paedophile activity as well as user identification are prone to errors, inferring reliable statistics is difficult. In addition, these statistics may fluctuate greatly with time, which makes their relevance unsure. Direct computations are not satisfactory to this regard, and the statistics must be carefully examined before concluding.

To address these challenges, we make the following contributions:

- *Datasets.* We collect and publicly provide two sets of keyword-based queries entered by *eDonkey* users, on two different servers in 2007 and 2009. Each spans several weeks of activity (10 and 28, respectively) and contains hundreds of millions keyword-based queries, involving millions of users. Using two datasets collected

on different servers and at different dates increases the generality of our results significantly.

- *Detection tool.* Using domain knowledge of paedophile keywords, we design a tool for automatic detection of paedophile queries. We evaluate its success rate by a rigorous assessment involving 21 experts having a deep knowledge of online paedophile activity. These experts work in various national and international law-enforcement agencies and well-established NGOs, including *Europol* and the *National Center for Missing & Exploited Children*.
- *Quantification.* Our tool detects hundreds of thousands of paedophile queries in our datasets. Using the error rates of the tool, we derive a reliable estimate of the actual fraction of paedophile queries they contain, which is approximately 0.25%. We then design several complementary approaches to estimate the fraction of observed users who enter paedophile queries and check both their statistical significance and their consistence. We finally establish a lower bound of 0.2% for users who enter paedophile queries in the 2007 dataset. Analysis of the 2009 dataset indicates that the 0.2% bound is also valid in this case.

We describe in Section II our datasets. Section III presents our tool for automatic detection of paedophile queries, our assessment methodology, and the estimates of its error rates by experts of the field. We finally establish the fractions of paedophile queries and users who entered them in Sections IV and V. We discuss related work in Section VI. We present our conclusions and the perspectives of our work in Section VII.

## II. DATA

Although many extensions exist [12], the *eDonkey* system basically relies on a set of 100 to 300 servers indexing available files and providers for these files. Clients send to these servers keyword-based queries (which may also contain meta-data such as a type of file) describing the content they search for. Servers answer with lists of files matching these keywords (typically, their filenames contain these keywords). Clients may then ask the server for providers of selected files. Once they have obtained this information, they may contact providers directly to obtain the files. Servers only play the role of directories; they do not store any exchanged file, and exchanges take place between clients, from peer to peer. *eDonkey* is currently one of the largest P2P systems used, and this has been true for several years [11].

We collected for this study two independent datasets, in 2007 and 2009. Both consist of a recording of hundreds of millions keyword-based queries received by an *eDonkey* server during a period of time of several weeks. To each query is associated a timestamp and the IP address from which it was received. The 2007 dataset contains in addition the connection port used for sending each query. We performed the 2007 measurement on one of the main servers running at that time by capturing and decoding IP-level traffic [13]; we performed the 2009 measurement on a medium-sized server by activating

its log capabilities. Notice that we do not observe exchanges actually occurring between users, and have no access to file content. This is not obtainable in practice at a large scale and is not mandatory for our purpose as we focus on what users seek. Finally, both datasets have been carefully anonymised at collection time, in conformance with legal and ethical constraints.

Key features of our datasets are summarised in Table I. We provide them publicly at [14] together with more details on collection, anonymisation, and normalisation procedures.

	duration	queries	IP addresses	(IP,port)
2007	10 weeks	107,226,021	23,892,531	50,341,797
2009	28 weeks	205,228,820	24,413,195	<i>n/a</i>

TABLE I  
MAIN FEATURES OF OUR TWO DATASETS AFTER NORMALISATION,  
ANONYMISATION, AND REMOVAL OF EMPTY QUERIES.

## III. DETECTING PAEDOPHILE QUERIES

In this section we design a tool for automatically identifying paedophile queries in large sets of queries, most of which are not paedophile. As machine learning approaches require prior knowledge of a representative set of paedophile queries, they cannot be applied here. We therefore rely on domain knowledge of paedophile keywords and ad-hoc observations to manually design our tool (Section III-A). Such a tool is necessarily prone to errors: some paedophile queries may not be tagged as such, and some non-paedophile queries may be tagged as paedophile. It is therefore crucial to obtain precise estimates of our error rates in order to make quantification of paedophile activity possible. This raises specific challenges in our context, which we address in Section III-B. We then set up an assessment framework which we submit to several independent and highly qualified experts (Section III-C). Using the results of this assessment (Section III-D), we finally obtain reliable estimates of our tool's error rates (Section III-E).

### A. Tool design

Our tool for detecting paedophile queries consists in performing a series of simple lexical tests (matchings of keywords in queries), each aimed at detecting paedophile queries of a specific form. We built a first set of rules based on our expertise in the paedophile context acquired for several years of work on the topic with law-enforcement personnel. We then manually inspected the results, identified some errors, and corrected them by adding minor variants to these general rules. We iterated this until obtained improvements became negligible. We describe our final rules below, and outline the detection steps in Figure 1.

According to experts of paedophile activity, some keywords point out exclusively such activity in P2P systems, *i.e.* they have no other meaning and are dedicated to the search of paedophile content. Typical examples include *qqaazz*, *r@ygold*, or *hussyfan*. We therefore built a list of specific keywords, called *explicit*, and we tag any query containing at least one word from this list as paedophile.

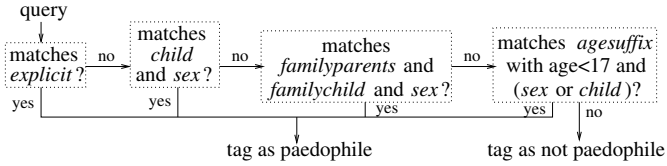


Fig. 1. Sequence of tests performed by our tool. Each matching consists in detecting if the query contains words from specific sets (named *explicit*, *child*, *sex*, *familyparents*, *familychild*, and *agesuffix*). See [14] for these sets of keywords and the tool source code.

Many paedophile queries contain words related to children or childhood and words related to sexuality, such as *child* and *sex*. We therefore constructed a list of keywords related to childhood, called *child*, and a list of keywords related to sexuality, called *sex*. We tag any query containing a keyword in both lists as paedophile. Notice that this may be misleading in some cases, for instance for queries like *destinys child sexy daddy* (a song descriptor).

A variant of this rule, which we added to the two previous ones, consists in tagging as paedophile the queries containing words related to family, denoting parents *and* children (stored in two lists called *familyparents* and *familychild*), and a word from the *sex* list.

Finally, many queries contain age indications under the form *n yo*, generally meaning that the user is seeking content involving *n years old* children. Other suffixes also appear in place of *yo*: *yr*, *years old*, etc. We identified such suffixes and built a list named *agesuffix*. Age indications are strong indicators of paedophile queries, but they are not sufficient in themselves: they also occur in many non-paedophile queries (e.g. when the user seeks a computer game for children). We decided to tag a query as paedophile if it contains age indication lower than 17 (greater ages appear in many non-paedophile queries) *and* a word in the *sex* or *child* lists.

In all situations above, although most keywords are in English, local language variations occur, in particular French, German, Spanish, and Italian versions. A few queries in rarer languages, such as Russian and Chinese, also occur. We included the most frequent translations in our sets of keywords.

We provide the exact rules implemented in our tool (including the sets of keywords we use) and the tool itself at [14].

## B. Method for tool assessment

Let us consider a set  $Q$  of queries, and let us denote by  $P^+$  (resp.  $P^-$ ) the set of paedophile (resp. non-paedophile) queries in  $Q$ . Let us denote by  $T^+$  (resp.  $T^-$ ) the subset of  $Q$  tagged as paedophile (resp. non-paedophile) by our tool.

Ideally, we would have  $T^+ = P^+$ , which would mean that our tool makes no mistake. In practice, though, there are in general paedophile queries which our tool mis-identifies, *i.e.* queries in  $T^- \cap P^+$ . Such queries are called *false negatives* (the tool produces an erroneous negative answer for them). *False positives*, *i.e.* queries in  $T^+ \cap P^-$ , are defined dually.

The numbers of false positives and false negatives describe the performance of our tool on  $Q$ . Notice however that they

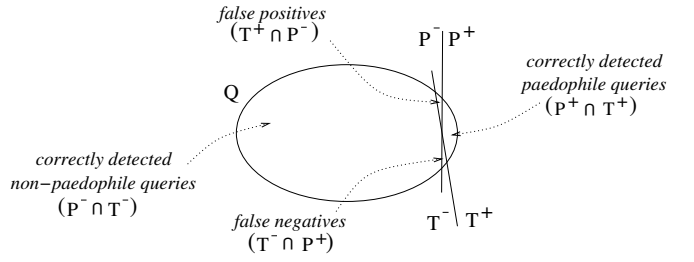


Fig. 2. Illustration of our notations. The ellipse represents the set of all queries,  $Q$ . The vertical line labelled  $P^-/P^+$  divides  $Q$  into the set of non-paedophile queries  $P^-$  (left) and the set of paedophile queries  $P^+$  (right). Likewise, the vertical line labelled  $T^-/T^+$  divides  $Q$  into the set of queries tagged as non-paedophile by the tool,  $T^-$  (left), and the set of queries it tags as paedophile,  $T^+$  (right).

strongly depend on the size of  $P^+$  and  $P^-$ . In our situation, we expect  $P^+$  to be much smaller than  $P^-$  (most queries are not paedophile), which automatically leads to small numbers of false negatives, even in the extreme (and useless) case where the tool would give only negative answers.

To evaluate the performance of a tool in such situations, two natural notions of false positive and false negative rates coexist. Both will prove to be useful here.

First, one may consider the false negative (resp. positive) rate when all inspected queries are paedophile (resp. non-paedophile):

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

An estimate of  $f^+$  may then be obtained by sampling a random subset  $X$  of  $P^-$  (*i.e.* random non-paedophile queries) and manually inspecting the results of the tool on  $X$ . Constructing  $X$  is easy: as most queries are non-paedophile, one may sample random queries and then manually discard the ones which are paedophile. As long as  $X$  is small, this has a reasonable cost. However, the fraction of queries in  $X$  which will be tagged as paedophile by our tool will be extremely small. As a consequence, an estimate of  $f^+$  obtained this way would be of poor quality.

Conversely, an estimate of  $f^-$  may be obtained by sampling a random subset  $X$  of  $P^+$  (*i.e.* random paedophile queries) and manually inspecting the results of the tool on  $X$ . As  $P^+$  is very small and unknown, sampling  $X$  is a difficult task. We may however approximate it using the notion of *neighbour* queries as follows.

Given a query  $q$  in  $Q$ , its *backward neighbour* is the last query in  $Q$  which was received from the same IP address as  $q$  less than two hours *before*  $q$ <sup>1</sup>. We therefore expect it

<sup>1</sup>We chose this threshold after examining the distributions of query inter-arrival times; it must be large enough to lead to many cases where neighbour queries exist, while being small enough to make it probable that neighbours of a query are related to this query. To this regard, two hours is a good compromise (see Section III-D, Table III), but a wide range of values around this specific value lead to similar observations.

was entered by the same user as  $q$ , seeking similar content<sup>2</sup>. Likewise, we define the *forward neighbour* of  $q$  as the first query in  $Q$  which was received from the same IP address as  $q$  within two hours *after*  $q$ .

We denote by  $N(q)$  the set containing the backward and forward neighbours of a query  $q$ . This set may be empty, and contains at most two elements. We denote by  $N(S) = \cup_{q \in S} N(q)$  the set of neighbour queries of all queries in set  $S$ , for any  $S$ . We guess that queries in  $N(P^+)$ , *i.e.* the neighbours of paedophile queries, are also paedophile with high probability (much higher than random queries in  $Q$ ). We expect this to be also true for queries in  $N(T^+)$ , which is confirmed in Section III-D, Table III.

Obviously,  $N(T^+) \cap P^+ \subseteq P^+$ , but  $N(T^+) \cap P^+ \not\subseteq T^+$  in general. In other words,  $N(T^+)$  probably contains queries in  $P^+$  (*i.e.* paedophile queries) which are *not* detected by our tool. If we consider the queries in  $N(T^+) \cap P^+$  as random paedophile queries, then they may be sampled to construct a set  $X$  of random paedophile queries suitable for estimating  $f^-$ . As  $X$  contains only paedophile queries, this estimate is equal to the number of queries in  $X$  not detected as paedophile by our tool divided by the size of  $X$ .

Notice that the queries in  $X$  may actually be biased by the fact that they are derived from  $T^+$ : the probability that a user enters a paedophile query which the tool is able to detect is higher if this user already entered one such query (he/she may enter in both cases keywords detected by our tool). As a consequence, our estimate of  $f^-$  may be an under-estimate.

Finally, one cannot, in our context, evaluate  $f^+$  properly; on the contrary, we are able to give a reasonable (under-)estimate for  $f^-$ . But both  $f^+$  and  $f^-$  are needed to evaluate the performance of our tool.

In order to bypass this issue, we consider the following variants of false negative and false positive rates, which capture the probability that the tool gives an erroneous answer when it gives a positive (resp. negative) one:

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

An estimate of  $f'^+$  may be obtained by sampling a random subset  $X$  of  $T^+$  (*i.e.* a random set of queries for which our tool gives a positive answer) and by manually inspecting this subset in order to obtain the number of false positives. We expect all sets involved in these computations to be of significant size (which is confirmed in Section III-D), so there is no obstacle in computing a reasonable estimate for  $f'^+$ .

Conversely, an estimate of  $f'^-$  may be obtained by sampling a random subset  $X$  of  $T^-$  and inspect it to determine the number of false negatives, *i.e.* the number of queries in  $X$  which actually are paedophile. However, as paedophile queries are expected to be very rare, the number of observed false negatives will be extremely small as long as  $X$  is of reasonable size.

<sup>2</sup>IP addresses are not enough to distinguish between users (see Section V) but *many* neighbours of paedophile queries are themselves paedophile (see Section III-D, Table III), which is what we need.

Therefore, one may easily obtain a significant estimate of  $f'^+$ , but computing a reasonable estimate for  $f'^-$  is not tractable in our case.

Finally, the quantities we will use for evaluating the quality of our tool are  $f'^+$  (the rate of errors when our tool decides that a query is paedophile) and  $f^-$  (the rate of paedophile queries that our tool mis-classifies as non-paedophile), which we are able to properly estimate. We describe our practical procedure for computing these estimates in the following sections and provide the obtained estimates in Section III-E.

### C. Assessment setup

In order to apply the method for quantifying our tool quality described above, we need to identify actual paedophile queries in some specific sets. We resort to independent experts of paedophile activity who manually inspect and tag these queries. We describe here the construction of these sets, the experts who helped us, and the interface we provided to them.

*Query selection.* Because the 2009 dataset was not yet available when we designed our tool and assessed it, we used the 2007 dataset for sampling queries to assess. We denote by  $Q$  the whole set of queries, and use the formalism of Section III-B. We divide  $Q$  into three sets (with overlap):  $T^-$  (queries tagged as not paedophile by our tool),  $T^+$  (queries it tagged as paedophile), and  $N(T^+)$  (neighbours of queries it tagged as paedophile).

Notice that some queries in  $T^+$ , *i.e.* some queries which are tagged as paedophile by the tool, are composed of only one word. Then, this word is necessarily a word in the *explicit* paedophile keywords list described in Section III-A. These keywords are known to have a very strong paedophile nature. Therefore, if such a keyword appears alone in a query, then this query surely is paedophile. We therefore increase the efficiency of our assessment by not submitting these one-keyword queries to experts. We denote by  $T_1^+$  the set of queries in this set, and by  $T_{>1}^+$  the queries in  $T^+$  composed of more than one word. Our optimisation consists in using the fact that  $T_1^+ \subseteq P^+$ , and so use only  $T_{>1}^+$  for assessment.

We finally construct the sets of queries to assess by selecting 1,000 random queries in each of the sets  $T^-$ ,  $T_{>1}^+$  and  $N(T^+)$  (thus 3,000 queries in total). This leads to three subsets which we denote by  $\overline{T}^-$ ,  $\overline{T}_{>1}^+$ , and  $\overline{N(T^+)}$  respectively. Notice that carefully tagging 3,000 queries already is a heavy task for experts. For this reason, we did not reproduce the assessment on the 2009 dataset and simply checked manually that its outcome would be very similar.

*Experts.* Once we selected sets of queries, the choice of experts is a crucial step. Indeed, deep knowledge of online paedophile activity is needed, if possible with a focus on P2P activity and/or query analysis. Such expertise is extremely rare, even at the international level. Thanks to our involvement in international research projects on paedophile activity for several years, with partners in various law-enforcement agencies and NGOs in several countries, we were able to contact many specialists who may play the role of experts in our study.

We obtained a set of 21 volunteers for participating to our assessment task. These participants are personnel of various law-enforcement institutions (including Europol and the main French and Danish national agencies) and well-established NGOs (including the *National Center for Missing & Exploited Children*, *Nobody’s Children Foundation*, *Action Innocence Monaco* and the *International Association of Internet Hot-lines*). A few security consultants also contributed. Their approach of paedophile activity are different and, as such, complementary. However, to ensure that we use only answers from relevant experts, we later conducted an assessment of participants themselves, see Section III-D.

*Interface.* We set up a web interface to make it convenient for participants to tag queries. All 3,000 queries were presented in a different random order to each participant, thus avoiding possible bias due to a specific order. Moreover, it was possible for participants to tag only a part of the 3,000 proposed queries, thus allowing them to contribute even if they had limited time.

We proposed five possible answers for each query: *paedophile*, *probably paedophile*, *probably not paedophile*, *not paedophile*, and *I don’t know*. To help participant’s choice, we displayed each query with its backward and forward neighbours (defined in Section III-B), when they existed. This was of great help in tagging ambiguous queries.

#### D. Expert results

Each of our 21 participants tagged more than 300 queries (*i.e.* 10 % of the whole), and 12 tagged more than 2,000.

*Expert selection.* Despite our efforts to select appropriate contributors, some may have an inadequate knowledge of our particular context (paedophile queries in a P2P system), and lower the quality of our results by entering erroneous answers. In order to identify such cases, we examined the answers of each participant to the queries which contain an *explicit* paedophile keyword, *i.e.* a word in our *explicit* list (defined in Section III-A). As already said, these keywords are well acknowledged paedophile keywords, which all experts of the field consider as strong indicators of paedophile queries.

The set of all queries submitted to contributors contains 1,003 queries with at least one explicit paedophile keyword. We computed the percentage of these queries which the corresponding contributor tagged as *paedophile* or *probably paedophile*. For all contributors except one, this percentage is above 95 %, thus showing that these contributors recognise these keywords. The remaining contributor only slightly disagrees with a ratio of 87.3 %.

The ratios discussed above may be misleading if a contributor tags all or almost all queries as paedophile. Actually, the answers of most contributors are well balanced between all possible answers, except for three contributors. Manual inspection shows that these contributors focused preferentially on paedophile queries (they did not tag all queries), which does not invalidate their answers. We therefore keep them in our expert set.

Finally, we obtain 42,059 answers provided by 21 experts who contributed at least 300 answers each. This leads to an average of slightly more than 14 experts assessing each query, which is sufficient for our purpose.

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

TABLE II  
NUMBER OF VOTES OF EACH KIND FOR EACH CONSIDERED SET.

The distribution of these answers among the queries of each considered set is given in Table II. It is in accordance with what one would expect if our tool performs well, and if our assumption that  $\overline{N(T^+)}$  should contain many paedophile queries is verified. We analyse this in more details now.

*Classification of queries.* For each query  $q$  submitted to experts in our assessment procedure, we denote by  $q^{++}$  the fraction of experts (among the ones who provided an answer for  $q$ ) which tagged it as *paedophile* and by  $q^+$  the fraction of experts which tagged it as *paedophile* or *probably paedophile*. We define  $q^-$  and  $q^{--}$  dually. Notice that  $q^+ + q^- < 1$  in general, as some *I don’t know* answers were provided (the fraction of such answers is  $1 - q^+ - q^-$ ). Moreover,  $q^+ \geq q^{++}$  and  $q^- \geq q^{--}$  for all  $q$ .

In order to classify queries according to expert answers, we expect to observe that each query  $q$  has either a high  $q^+$  (*resp.*  $q^{++}$ ) or a high  $q^-$  (*resp.*  $q^{--}$ ), but not both or neither, meaning that experts agree on the nature of  $q$ .

For many queries, the difference is very large: above 0.8 for 1,305 queries (over 3,000) in the case of  $q^{++}$  and  $q^{--}$ , and for 2,308 queries in the case of  $q^+$  and  $q^-$ . Only 41 queries have a difference  $|q^+ - q^-|$  smaller than or equal to 0.1, which already is significant. We therefore classify a query as paedophile if  $q^+ - q^- > 0.1$  and as non-paedophile otherwise. We finally obtain the query classification by experts presented in Table III.

	random subset		
	$\overline{T^-}$	$\overline{T^+_{>1}}$	$\overline{N(T^+)}$
paedophile queries	1	985	754
non-paedophile queries	999	15	246

TABLE III  
NUMBER OF QUERIES CLASSIFIED AS PAEDOPHILE OR NOT BY EXPERTS FOR EACH CONSIDERED SET.

#### E. Tool assessment results

Thanks to the assessment results in Table III and the expressions given in Section III-B, we may now compute estimates of the false positive and false negative rates which describe the quality of our tool.

First notice that, as expected, the number of paedophile queries in the set of queries tagged as non-paedophile by the tool is very low:  $|T^- \cap P^+| = 1$ . As a consequence, approximating  $f'^- = \frac{|T^- \cap P^+|}{|T^-|}$  by  $\frac{|T^- \cap P^+|}{|T^-|} = \frac{1}{1,000}$  would yield very poor quality result.

The estimate obtained for  $f'^+$  is of much better quality. It relies on the following expression:

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

(since  $T_1^+ \cap P^- = \emptyset$ , because all queries in  $T_1^+$  are paedophile, see Section III-C).

An estimate of  $|T_{>1}^+ \cap P^-|$  is given by  $|\overline{T_{>1}^+} \cap P^-| \cdot \frac{|T_{>1}^+|}{|T_{>1}^+|}$  which leads to:

$$f'^+ \sim \frac{|\overline{T_{>1}^+} \cap P^-|}{|T^+|} \cdot \frac{|T_{>1}^+|}{|T_{>1}^+|} = \frac{15}{207,340} \cdot \frac{192,545}{1,000} \sim 1.39\%.$$

The quality of this estimate is good not only because  $|\overline{T_{>1}^+} \cap P^-| = 15$  is significant, but also because we evaluate it using a sample of queries in  $T_{>1}^+$ , which is much (more than 500 times) smaller than  $T^-$ , involved in the estimate of  $f'^-$ .

Conversely, the assessment results confirm that estimating  $f^+ = \frac{|T^+ \cap P^-|}{|P^-|}$  with our data would yield poor quality approximate, as  $|T^+ \cap P^-|$  is small (there are very few paedophile queries), as well as the sample size, compared to the size of  $P^-$ .

It is possible to estimate  $f^-$  much more accurately:

$$f^- = \frac{|T^- \cap P^+|}{|P^+|} \gtrsim \frac{|T^- \cap (\overline{N(T^+) \cap P^+})|}{|\overline{N(T^+) \cap P^+}|} = \frac{185}{754} \sim 24.5\%.$$

This value however is an under-estimate, because we assessed neighbours of detected paedophile queries instead of random paedophile queries. It is equal to the probability that our tool erroneously tags such a neighbour as non-paedophile. There is no *a priori* reason to suppose that this leads to huge differences, though, and we therefore expect this bound to be reasonably tight. We will handle this with care in the following.

#### IV. FRACTION OF PAEDOPHILE QUERIES

In this section, we estimate the fraction of paedophile queries in our two datasets, *i.e.*  $\frac{|P^+|}{|Q|}$  for each  $Q$  (we use the notations defined in Section III-B).

We use the automatic paedophile query detection tool designed in the previous section and its error rates. We first estimate the fraction of queries in  $Q$  tagged as paedophile by the tool, and then infer from it an estimate of  $\frac{|P^+|}{|Q|}$ .

##### A. Fraction of automatically detected queries

The automatic paedophile query detection tool divides  $Q$  into two disjoint subsets:  $T^+$ , the set of queries tagged as paedophile; and  $T^-$ , the set of queries tagged as non-paedophile. We estimate here the fraction of queries tagged as paedophile, *i.e.*  $\frac{|T^+|}{|Q|}$ , in both datasets.

This may be trivially obtained by computing the set  $T^+$  of queries tagged as paedophile by the tool, and then divide it by the total number of queries. We obtain this way rates slightly above 0.19% for both datasets. In order to ensure the relevance of this estimate, though, we go into details below.

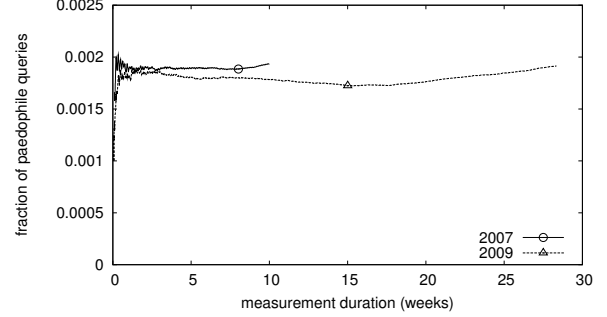


Fig. 3. Fraction of paedophile queries detected in our datasets as a function of the measurement duration.

We first check that the measurement duration is large enough by plotting the fraction of queries tagged as paedophile as a function of the measurement duration, see Figure 3. It clearly shows that this fraction converges rapidly to a reasonably steady value, slightly below 0.2%; changing this value significantly would need a drastic change in the data.

Going further, we studied the cumulative distribution of the fraction of queries tagged as paedophile in all 1-hour, 6-hour, 12-hour and 24-hour slices of the measurements (not presented here). It clearly appears that there is a notion of *normal*, or *median* behaviour for each slice size, and that it is quite independent of slice sizes. The averages of these distributions are all close to 0.2%, in accordance with our previous computations.

Finally, we conclude that the fraction of queries tagged as paedophile by our tool may be approximated by  $\frac{|T^+|}{|Q|} \sim 0.2\%$  in both datasets.

##### B. Inference

We established in Section III-E reliable estimates for  $f^-$  and  $f'^+$ . As a consequence, we have to infer the size of  $P^+$  from these rates, which may be done as follows:

$|P^+| = |P^+ \cap T^+| + |P^+ \cap T^-| = |T^+|(1 - f'^+) + |P^+|f^-$  and so:

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

Using  $f^- \gtrsim 24.5\%$  and  $f'^+ \sim 1.39\%$ , we obtain:

$$\frac{|P^+|}{|Q|} \gtrsim 0.25\%$$

for both datasets. In other words, at least one query over 400 is paedophile in our two datasets.

Notice that taking  $f^- \sim 50\%$ , which most certainly is a huge over-estimate, leads to a ratio of approximately 0.38% paedophile queries. We therefore conclude that the true ratio is not much larger than 0.25%.

## V. FRACTION OF PAEDOPHILE USERS

Although the fraction of paedophile queries is of high interest in itself, the key question when quantifying paedophile activity actually is the fraction of paedophile *users*, which we define as users who entered at least one paedophile query.

However, identifying a user in an internet-like environment is a challenge in itself [15], [16]. Any computer is identified by an IP address at a given time, but even this may change and we are unable in general to detect that a same computer has two different addresses at different times and/or that two computers use the same address. In addition, a same user may use several computers, and several users may use the same computer, making identification of users even more challenging.

Notice however that what we need is slightly weaker: we need to make the difference between two users in our dataset in order to avoid mixing their queries. Indeed, mixing the queries of several users would lead to interpret the corresponding series of queries as a unique series, and thus a unique user. As we consider a user as paedophile as soon as he/she entered one paedophile query, if one of the corresponding users entered paedophile queries, then the whole series is considered as coming from a paedophile user. Since the overall fraction of users entering paedophile queries is very small, it is very unlikely that two paedophile users are mixed in this way. Therefore, mixing the queries of several users leads to a decrease of the total number of observed *users*, but in general the number of observed *paedophile users* stays the same. This leads to an over-estimate of the fraction of paedophile users: we call this phenomenon *pollution*.

We explore below different approaches to count users who sent paedophile queries in our datasets. First, we show that identifying users with their IP address only is not sufficient, but that considering the pair composed of their IP address and their connection port provides relevant information. We then study the influence of the measurement duration using sliding windows of different lengths. Finally, we consider series of queries received from a same IP address with small inter-query times, which we call *sessions*. Indeed, the fraction of sessions containing paedophile queries may be considered as an estimate of the fraction of users entering such queries.

### A. IP addresses and connection ports

Two pieces of information in our datasets may lead to distinguish between users: the IP address from which they sent queries, and the connection port they used. The latter makes it possible to distinguish between several users in a same local network with a NAT.

Therefore, we consider here two approximations of the notion of user: we first assume that the IP address is sufficient to distinguish between different users, and then that the pair (IP address, connection port) is sufficient. Notice that this last assumption is necessarily better than the previous one, but comparing the two is enlightening.

We display in Figure 4 the fraction of IP addresses and (IP, port) pairs from which at least one paedophile query (as detected by our tool) was entered. We call them paedophile IP

and paedophile (IP, port) pairs for simplification. Notice that only IP addresses are available in the 2009 dataset.

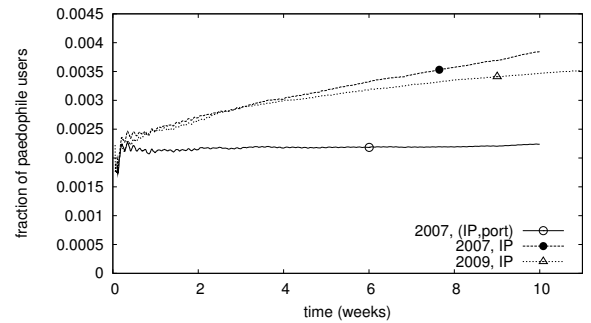


Fig. 4. Fraction of paedophile users detected in our datasets as a function of the measurement duration.

For both datasets, the fraction of paedophile IP addresses clearly grows with the measurement duration. This reveals the *pollution* phenomenon sketched above: as IP addresses may correspond to different users over time, and as one paedophile user is sufficient to consider the corresponding address as paedophile, then the probability for any given address to be considered as paedophile grows with measurement time (all IP addresses may eventually be considered as paedophile). This confirms that using IP address alone is misleading in this case.

On the other hand, the fraction of paedophile (IP, port) pairs in the 2007 dataset has a very different behaviour: it rapidly reaches a steady regime, very similar to the fraction of paedophile queries studied in Section IV, Figure 3. This shows that pollution due to dynamic allocation of addresses and ports is negligible in this case.

We finally conclude that the fraction of paedophile (IP, port) pairs is meaningful, and that this fraction is slightly above 0.22 % here.

### B. Varying measurement duration

Figure 4 shows that increasing the measurement duration leads to an increase of the pollution of IP addresses by paedophile users. Therefore, considering shorter measurement windows leads to a better handling of the pollution phenomenon. On the other hand, this leads to less observed data, and therefore to less reliable results.

To study this, we divide our datasets into small measurement windows, and compute the observed fraction of paedophile IP addresses or (IP, port) pairs for all windows. The distribution of these fractions for all windows (not presented here) are homogeneous, and therefore their mean is representative. We present in Figure 5 this mean as a function of the window size.

The fraction of paedophile (IP, port) pairs for the 2007 dataset first fluctuates for small window sizes, and quickly converges to a steady regime, very close to the overall fraction of paedophile (IP, port) pairs in the dataset. Notice that it is possible that a same (IP, port) pair corresponds to several users (*e.g.*: family computers). However, the probability that this happens *within a short time span* is greatly reduced. The fact

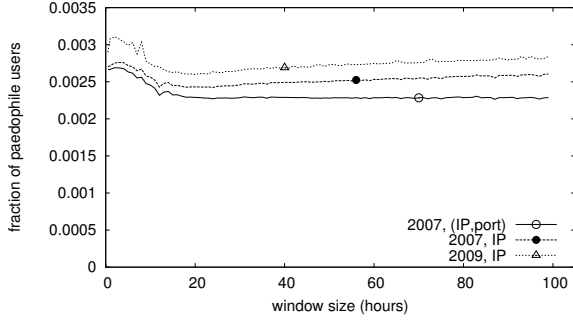


Fig. 5. Fraction of paedophile users as a function of the measurement window size.

that the fraction of paedophile (IP, port) pairs in windows of limited duration is very close to the overall fraction therefore shows that it is close to the fraction of actual detected users.

This is confirmed by the fraction of paedophile IP addresses as a function of the window size. After some initial fluctuations, this value drops to slightly less than 0.25%, then increases linearly with the window size<sup>3</sup>. Considering shorter windows therefore reduces temporal pollution. At any given time there are nonetheless several users *simultaneously* using the same IP address, because they are behind a NAT for instance. They will however use different ports, which is why the fraction of paedophile (IP, port) pairs is always lower than the fraction of paedophile IP addresses.

The plot for the fraction of paedophile IP addresses in the 2009 dataset has the same behaviour as for the 2007 dataset, but is larger than it. This could be because the fraction of paedophile users is larger than in the 2007 dataset. However, as the fraction of paedophile *queries* in both datasets are very similar, we suspect that this is because more users use the same IP address simultaneously in 2009 than in 2007.

### C. Sessions

A *session* is a maximal set of queries from the same IP address (or (IP, port) pair) such that two consecutive queries are not separated by more than a given delay  $\delta$ . Studying sessions reduces temporal pollution, as there will probably be a gap between the queries of two users who use the same IP address successively. Moreover, there is no *a priori* reason why paedophile users would make more sessions than other users, and so we approximate the fraction of paedophile users by the fraction of paedophile sessions.

We present in Figure 6 the fraction of paedophile sessions for different choices of  $\delta$ . This fraction for very small values of  $\delta$  is not relevant, because series of queries entered by a same user then belong to several sessions. For large values of  $\delta$ , this fraction becomes closer and closer to the overall fraction of paedophile users in the dataset<sup>4</sup>. This again confirms that

<sup>3</sup>This increase is not obvious on the figure because the slope is very small. If the  $x$  axis extended to the whole 10 weeks of measurement though, the plot would reach 0.38% which is the overall fraction of paedophile IP addresses in the dataset.

<sup>4</sup>If  $\delta$  is equal to the measurement duration, all queries entered from the same IP address or (IP, port) pair will belong to a single session.

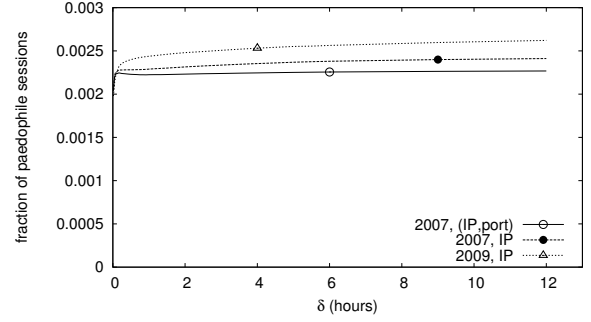


Fig. 6. Fraction of paedophile sessions as a function of  $\delta$ , the maximal delay between two consecutive queries in a same session.

considering IP addresses and connection ports seems to be enough to identify users in this dataset.

The fraction of paedophile sessions corresponding to IP addresses is higher. This again comes from the fact that several users are simultaneously connected from the same IP address, but do not use the same port.

The fraction of paedophile sessions is larger for the 2009 dataset than for the 2007 dataset. Again, we conjecture that this is because a higher number of users use simultaneously the same IP address.

### D. Inference

The fact that the three methods used above for user quantification are in accordance shows that considering (IP, port) pairs is relevant for identifying users in our context. The fraction of such users entering queries detected as paedophile by our tool is equal to 0.22% in the 2007 dataset. We now use the false positive and false negative rates of our tool to infer the actual fraction of paedophile users.

These rates give the number of queries that the tool misclassified. However, since we do not know which precise queries are mis-classified, we do not know what fractions of users they represent. If queries were mis-classified with uniform probability, they would correspond to a similar fraction of users. This is however probably not true, as a same user tends to enter similar queries. Therefore, if one of his/her queries is mis-classified, probably many others are.

We however establish, using the false positive rate, a lower bound for the fraction of paedophile users. A fraction  $f'^{+}$  of the queries detected as paedophile by the tool are in fact not paedophile, which represents a given number  $n$  of queries. Clearly, the corresponding number of users which the tool mis-identified as paedophile is at most  $n$  (it is equal to  $n$  if all mis-identified queries are entered by different users). Conversely, the tool failed to detect some paedophile queries. If all these queries were entered by users who were nonetheless detected as paedophile (because they entered other paedophile queries which were correctly identified), then no paedophile user is missed. The tool detected  $|T^+| = 207,340$  paedophile queries in the 2007 dataset, which correspond to 112,712 different users. The number of queries erroneously tagged as paedophile is  $|T^+| \cdot f'^{+} = 2,882$ . Finally, the number of paedophile users



is at least  $112,712 - 2,882 = 109,830$ , which leads to a fraction of paedophile users slightly lower than 0.22 %.

It is not possible to establish such a lower bound for the fraction of paedophile users in the 2009 dataset, because we do not have access to the connection ports of the users. However, we observe that when we reduce the pollution caused by users successively using the same IP address (by studying measurement windows and sessions, see Sections V-B and V-C), the obtained values are close for both datasets, but larger for the 2009 dataset. We therefore estimate that a lower bound of 0.2 % of paedophile users applies to both datasets.

## VI. RELATED WORK

Collection and analysis of large P2P traces is a very active field. Studies mainly focus on peer properties which are useful for protocol design, such as their connection time, sharing behaviour, or similarity regarding searched files and geographical location, see for instance [17], [18], [19], [20]. Some works also analyse queries entered by users [21], [22] but consider limited statistics (typically query length, number, interarrival time or redundancy). Only very few studies examine user interests in detail [7], [23], [9].

On the other hand, many papers discuss the amount and features of paedophile activity in P2P systems but they rely on very small datasets collected manually (typically by entering a few queries and examining obtained results), e.g. [24], [8], [10]. They aim at establishing the alarming presence of paedophile activity in P2P systems, not at quantifying it.

Up to our knowledge, only two papers deal with the quantification of paedophile activity in a P2P system in a similar way as the work presented here [7], [9]. Both use datasets almost 1 000 times smaller than ours, do not describe precisely their methods and do not address *user* quantification. Therefore, they may be seen as pioneering but limited work on paedophile activity quantification when compared to our own work.

## VII. CONCLUSION AND PERSPECTIVES

We addressed the problem of rigorously and precisely quantifying paedophile activity in a large P2P system. We first set up a methodology and designed a tool for automatic detection of paedophile queries. Thanks to independent highly-qualified experts of the field, we estimated its false positive and false negative rates. We collected two different datasets containing hundreds of millions keyword-based queries entered in the *eDonkey* system, and established that approximately 0.25 % of them are paedophile. We then designed several complementary methods for quantifying involved users; we established that at least 0.2 % of observed users sent paedophile queries in our 2007 dataset, similarly to our 2009 dataset.

It is the first time that quantitative information on paedophile activity in a large P2P system is obtained at this level of precision, reliability, and at such a scale. This significantly improves awareness on this topic, with important implications for child protection, policy making and internet regulation.

Moreover, our contributions open several promising directions. First, one may extend our results to other systems. One may for instance collect *Gnutella* queries like in [7], [9] and inspect them with our tool. We also open the way to studies and actions critical for understanding and fighting paedocriminality. Finally, many of our contributions are not specific to paedophile activity and/or P2P systems, and could be used in other contexts.

## REFERENCES

- [1] E. Quayle, L. Loof, and T. Palmer, "Child pornography and sexual exploitation of children online," in *World Congress III against Sexual Exploitation of Children and Adolescents*, 2008.
- [2] J. Wolak, K. Mitchell, and D. Finkelhor, "Online victimization of youth: five years later," 2006, report of the National Center for Missing & Exploited Children (NCMEC).
- [3] —, "Internet sex crimes against minors: the response of law enforcement," 2003, report of the National Center for Missing & Exploited Children (NCMEC).
- [4] R. Wortley and S. Smallbone, "Child pornography on the Internet," 2006, report of the US Department of Justice.
- [5] J. Delahunty, "eD2K razorback servers seized," *Afterdawn*, 2006, <http://afterdawn.com>.
- [6] S. Murray, "Peer-to-peer networks: file sharing software," 2005, bill number SB 96 of the California State Senate.
- [7] D. Hughes, J. Walkerdine, G. Coulson, and S. Gibson, "Is deviant behavior the norm on P2P file-sharing networks?" *IEEE Distributed Systems Online*, vol. 7, no. 2, 2006.
- [8] L. D. Koontz, "File sharing programs – child pornography is readily accessible over peer-to-peer networks," 2003, report of the US General Accounting Office.
- [9] C. M. Steel, "Child pornography in peer-to-peer networks," *Child Abuse & Neglect*, 2009.
- [10] F. Waters, "Child sex crimes on the Internet," 2007, report of State of Wyoming Attorney General.
- [11] H. Schulze and K. Mochalski, "Ipoque Internet study," 2009, <http://www.ipoque.com>.
- [12] Wikipedia, "EDonkey network," [http://en.wikipedia.org/wiki/EDonkey\\_network](http://en.wikipedia.org/wiki/EDonkey_network).
- [13] F. Aidouni, M. Latapy, and C. Magnien, "Ten weeks in the life of an edonkey server," *International Workshop on Hot Topics in P2P Systems*, 2009.
- [14] "Supplementary material," <http://www-rp.lip6.fr/~latapy/antipaedo/>.
- [15] R. Bhagwan, S. Savage, and G. M. Voelker, "Understanding availability," in *International Workshop on Peer-To-Peer Systems (IPTPS)*, 2003.
- [16] D. Stutzbach and R. Rejaie, "Understanding churn in peer-to-peer networks," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2006.
- [17] K. P. Gummadi, S. Saroiu, and S. D. Gribble, "A measurement study of Napster and Gnutella as examples of peer-to-peer file sharing systems," *Computer Communication Review*, vol. 32, no. 1, p. 82, 2002.
- [18] S. B. Handurukande, A.-M. Kermarrec, F. L. Fessant, L. Massoulié, and S. Patarin, "Peer sharing behaviour in the eDonkey network, and implications for the design of server-less file sharing systems," in *EuroSys*, 2006.
- [19] L. T. Nguyen, D. Jia, W. G. Yee, and O. Frieder, "An analysis of peer-to-peer file-sharing system queries," in *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2007.
- [20] S. Sen and J. Wang, "Analyzing peer-to-peer traffic across large networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 2, pp. 219–232, 2004.
- [21] A. S. Gish, Y. Shavitt, and T. Tankel, "Geographical statistics and characteristics of P2P query strings," in *International Workshop on Peer-to-Peer Systems (IPTPS)*, 2007.
- [22] A. Klemm, C. Lindemann, M. K. Vernon, and O. P. Waldhorst, "Characterizing the query behavior in peer-to-peer filesharing systems," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2004.
- [23] Y. Shavitt and U. Weinsberg, "Song clustering using peer-to-peer co-occurrences," in *IEEE International Symposium on Multimedia*, 2009.
- [24] P. Fagundes, "Fighting internet child pornography – the Brazilian experience," *The Police Chief Magazine*, vol. LXXVI, no. 9, 2009.