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

March 13, 2012

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.

1   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

---

[*]A shorter version of this work has been published in [17].

paedophilia and induces a trivialisation of such content. Much work is devoted to these psychological and societal issues, see [21, 32].

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 [14]. This makes fighting these exchanges a key issue for law enforcement [31, 33]. 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 [6, 18].

For these reasons, knowledge of paedophile activity in P2P systems is a critical resource for law enforcement, child protection and policy making. See [21, 31, 32, 33] for surveys on these issues. However, current knowledge on this activity and its extent remains very limited and is subject to controversy [12, 16, 26, 29, 31, 32].

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 [22].

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 and publicly provide 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 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.

Figure 1 summarises our global methodology. We describe in Section 2 our dataset collection and anonymisation. Section 3 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 4 and 5. We discuss related work in Section 6.

Finally, we obtain, for the first time at this level of accuracy and reliability, estimates of the two most crucial statistics on paedophile activity in a large P2P system. We establish reference end-to-end methodology for tackling such issues rigorously. We moreover provide publicly our full datasets, our paedophile query detection tool, and the sets of detected paedophile queries [28]. We therefore open the way to more research on paedophile activity and other activities in P2P systems, which we discuss in Section 7.

2   Data

Although many extensions exist [30], the eDonkey system basically relies on a set of 100 to 300 servers indexing available files and providers for these files. Clients
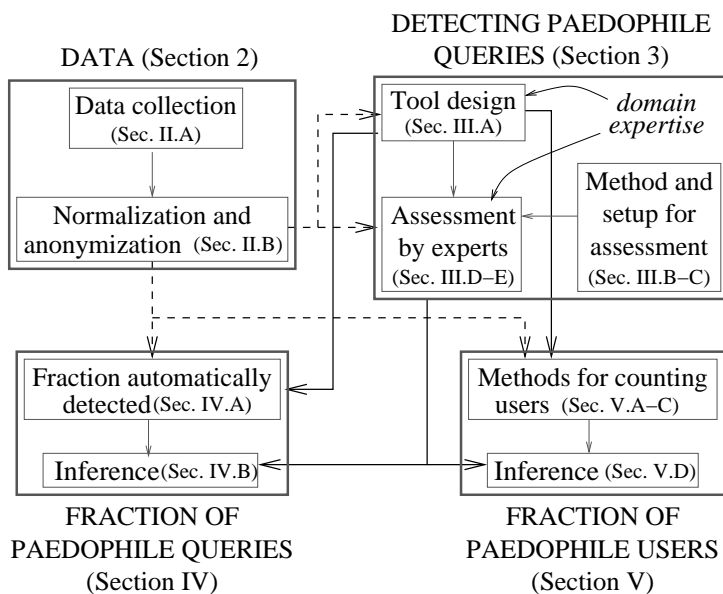
Figure 1: Structure of the paper and main contributions.

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 [22].

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 number used for sending each query. 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.

Key features of both datasets are summarised in Table 1. We detail the data collection, normalisation and anonymisation procedures below.

4

	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	*n/a*

Table 1: Main features of our two datasets after normalisation, anonymisation, and removal of empty queries.

2.1 Data collection

We collected the 2007 dataset on one of the main servers at that time, during a 10-week continuous measurement. It consisted in doing an IP-level capture of the UDP traffic on the server. We therefore had to decode this traffic into application-level messages, and then to select keyword-based queries. One advantage of this dataset is that we were able to record both IP addresses and connection ports of clients, which plays a key role in user quantification (Section 5). The procedure for this measurement is fully detailed in [2].

We collected the 2009 dataset by a 28-week continuous measurement on a medium-sized server at that time. We activated the log capability embedded in the standard server software, which records directly (in a human-readable textual format) timestamped keyword-based queries and the IP address they were received from. We observe this way both UDP and TCP traffic, but have no access to information on connection ports.

Figure 2 presents the roles of the client and server during an eDonkey session. Due to the measurement techniques used on the different servers, some information may or may not be available in our collected data.
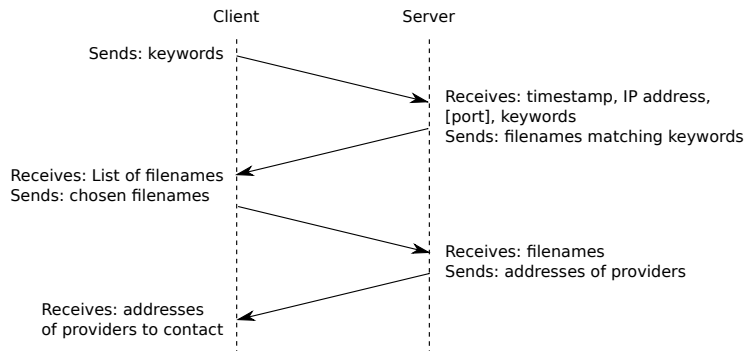

Figure 2: Client-server exchanges during an eDonkey session.

These datasets come from different servers, with different importance in the network (one of the main servers for 2007 and a medium one for 2009), and the measurement procedures were very different. In addition, three years elapsed between the collection of these two datasets, with many evolutions: users are not the same, the eDonkey software and protocols evolved significantly, many servers were stopped due to actions of copyright holders [6], other P2P systems increased

in popularity, etc. We therefore consider these two datasets as complementary, which gives some insight on the robustness of our observations to measurement conditions.

2.2   Normalisation and anonymisation

The traffic we observe contains much personal information, which we must remove to comply with ethical and legal restrictions. The main such information is the IP address from which each query was made, but text queries also contain personal information: some individuals enter their name or their friends' names, phone numbers, or even credit card numbers as search-strings [1, 2, 3, 19]. Finally, users interested in rare contents may enter very specific keywords, usable for identifying them.

Anonymisation of internet traces is a subtle issue in itself [1, 3, 19]. The challenge consists in obtaining rich data while fully preserving user privacy. Relying on state-of-the-art knowledge, we set up an appropriate anonymisation procedure which we describe below. We applied it independently to both datasets (as it had to be performed at measurement time).

First notice that anonymising IP addresses with a hash code is not satisfactory: one may decode addresses by applying the function to the 2^{32} possible addresses. Likewise, encryption relying on a secret key is unsure on such data. We therefore chose to encode addresses according to their order of appearance in the captured data: we replace the first observed IP address by 0, the second by 1 and so on. We proceed similarly for connection port numbers. This anonymisation is consistent: we always replace a same IP address or port by the same integer. Although computationally expensive, it has three key advantages: it ensures a very strong anonymisation level, it is feasible in real-time during the measurement (which is mandatory), and it makes further use of the dataset much easier.

We apply the following scheme to keyword-based queries. First we replace all accented characters by the corresponding unaccented letter. Then we convert all characters to lower-case, replace all non-alphanumeric characters by a space, and remove successive spaces. We call the obtained queries *normalised queries*. They contain only series of alphanumeric characters separated by spaces, which we call *words*.

In order to anonymise this normalised data, we have to distinguish between personal (sensitive) and general (non-sensitive) information. We assume that a same word entered by many users (and thus in many queries) is not sensitive [1, 3]. For instance, a sensitive name or phone number would appear only in a few queries, or in many queries but entered by a same user. We finally remove all words appearing in normalised queries from less than 50 distinct IP addresses, which ensures a very high level of anonymisation.

Numbers, in particular short numbers, tend to appear in many contexts. They therefore appear in clear after the procedure above, which raises anonymisation concerns. In particular, phone numbers often appear as space-separated series of two or three digits, which appear in queries from more than 50 IP

addresses and so are not removed. A solution would be to remove all numbers, but keeping such information is crucial here as they often indicate ages in paedophile queries, see Section 3.1. We therefore decided to remove numbers of more than two digits, and numbers with value greater than 16. This hides most numbers while keeping the age information we need.

Finally, short words raise similar concern. In particular, charset encoding problems sometimes lead to strings where a blank space is inserted between any two consecutive characters. Single characters are then considered as words, and they will most probably appear frequently. In the procedure above, this means that the corresponding queries will appear in clear. To avoid this, we remove all words composed of only one letter. Removing words with only two or three letters would also make sense, but this would discard important information, see Section 3.1. Inspection of our data showed that keeping them raises no significant anonymisation concern.

Figure 3 presents two examples of queries, and their counterparts after the normalisation and anonymisation procedures.

Photo Scarlett O'Hara été "Champs-Élysées"
↪ photo scarlett o hara ete champs elysees
↪ photo scarlett hara ete champs elysees

Credit card 1234 4567 7654 4321 johndoe7643@something.com
↪ credit card 1234 4567 7654 4321 johndoe7643 something com
↪ credit card something com

Figure 3: Two queries presented in their raw, normalised and anonymised versions, from top to bottom. Raw queries are italicised.
In the first query, accented letters are first replaced by their non-accented counterparts. All letters are transformed into lower-case, the apostrophe and quotation marks are replaced by spaces. We obtain a *normalised query*. The subsequent anonymisation procedure only removes the remaining *o*, now a single-letter word.
In the second query, the normalisation consists of putting the text in lower-case and replacing the arobace and dot by spaces. The anonymisation then removes the 4-digit groups, and *johndoe7643*, a word not sufficiently common to appear in queries from more than 50 distinct IP addresses. Consecutive spaces are eventually trimmed.

3   Detecting paedophile queries

In this section we design a tool for automatic identification of paedophile queries in large sets of queries, most of which are not paedophile. The standard way for doing so is to use machine learning approaches, which rely on the prior

knowledge of a representative set of paedophile queries. No such dataset is currently available, though, and constructing it without an automatic tool would require expert inspection of huge sets of queries, which is by far too time-consuming and costly.

We therefore rely on domain knowledge of paedophile keywords and ad hoc observations to manually design our tool (Section 3.1). 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 3.2. We then set up an assessment framework which we submit to several independent and highly qualified experts (Section 3.3). Using the results of this assessment (Section 3.4), we finally obtain reliable estimates of our tool's error rates (Section 3.5).

3.1   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 [28]. 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 out final rules below, and outline the detection steps in Figure 4.
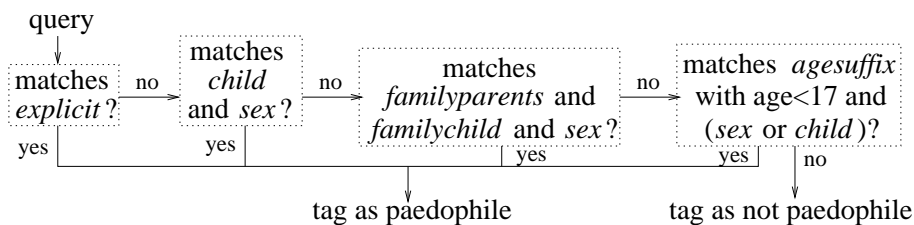

Figure 4: Sequence of tests performed by our tool. Each matching consists in detecting if the query contains words from specific sets (named *explicit*, *child*, *sex*, *familyparents*, *familychild*, and *agesuffix*). See [28] for these sets of keywords and the tool source code.

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.

8

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 (for example 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 [28].

3.2   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 which is tagged as paedophile (resp. non-paedophile) by our tool. Figure 5 provides an illustration of our notations.

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. These notions are classical in data mining, see for instance [8].

The numbers of false positives and false negatives describe the performance of our tool on Q. Notice however that they 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
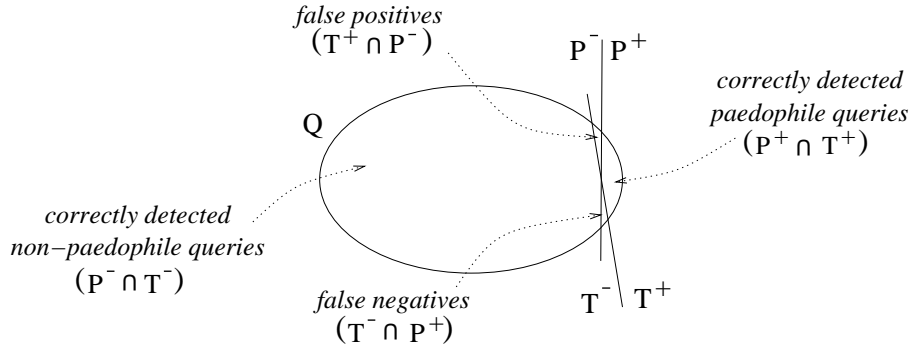
Figure 5: Illustration of our notations. The ellipse represents the set of all queries, Q. The vertical line labelled P^-/P^+ divides Q into the set of non-paedophile queries P^- (left) and the set of paedophile queries P^+ (right). Likewise, the vertical line labelled T^-/T^+ divides Q into the set of queries tagged as non-paedophile by the tool, T^- (left), and the set of queries it tags as paedophile, T^+ (right).

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, if it exists [1]. We therefore expect it was entered by the same user as q, seeking the

---

[1]We chose this threshold after examining the distributions of query interarrival 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

same kind of content [2]. 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 3.4, Table 4.

Obviously, $N(T^+) \cap P^+ \subseteq P^+$, but $N(T^+) \cap P^+ \nsubseteq T^+$ 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 3.4), so there is no obstacle in computing a reasonable estimate for f'^+.

---

regard, two hours is a good compromise (which we confirm in Section 3.4, Table 4), but a wide range of values around this specific value lead to similar observations. The value we chose for the threshold is consistent with the conclusions of [4], a study dedicated to inter-query times in eDonkey.

[2] IP addresses are not enough to distinguish between users (see Section 5) but *many* neighbours of paedophile queries are themselves paedophile (see Section 3.4, Table 4), which is what we need.

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.

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 3.5.

3.3 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. To do so, 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 3.2. 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). These three sets are easy to compute from Q using our tool.

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 3.1. 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 [3]).

---

[3] Since there is an overlap between $N(T^+)$ and the other sets, we could have sampled some queries more than once, leading to less than $3,000$ queries in total. Since $1,000$ is small

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 for which we need expert classification, 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. It is present mainly in law-enforcement institutions, where special units are devoted to fighting (online) paedophile activity, and in NGOs dedicated to similar tasks (but with a different approach, in general). Some security consultants also have this kind of knowledge.

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 a large number of specialists who may play the role of experts in our study. We were for instance able to send a call for experts on the main international mailing-list of law-enforcement personnel working on cybercrime.

We finally 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 Hotlines*). A few security consultants also contributed.

We later conducted an assessment of participants themselves to ensure that we use only answers from relevant experts, see Section 3.4.

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 for 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 3.2), when they existed. This was of great help in tagging ambiguous queries.

---

compared to the total sizes of the three sets, this did not happen here.

3.4 Expert results

The answers collected from our 21 participants are summarised in Table 2. Each of them tagged more than 300 queries (*i.e.* 10 % of the whole), and 12 tagged more than 2,000.

paedo	prob. paedo	don't know	prob. not	not paedo	total	relevance
1,530	149	25	66	1,230	3,000	99.6
1,381	247	125	580	667	3,000	98.5
1,679	89	2	113	1,117	3,000	99.2
1,603	201	99	174	923	3,000	99.3
1,598	5	15	1	1,381	3,000	98.9
128	81	1	26	124	360	96.3
216	154	0	142	132	644	98.6
1,624	126	16	165	581	2,512	99.7
351	16	2	16	27	412	99.6
647	119	71	40	439	1,316	98.9
1,174	111	20	64	789	2,158	99.3
335	17	1	70	166	589	97.1
641	383	4	112	753	1,893	96.6
1,071	546	2	453	928	3,000	87.3
1,554	197	28	327	894	3,000	98.2
305	270	24	89	181	869	98.3
371	1,017	496	570	546	3,000	95.7
976	936	405	594	89	3,000	95.5
344	12	10	70	156	592	99.0
845	139	323	175	182	1,664	98.1
1,506	120	6	25	393	2,050	98.3

Table 2: Assessement results for each participant. Each line corresponds to a participant, and gives the number of answers of each kind he/she provided, his/her total number of answer, and his/her relevance.

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 3.1). 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 which contain at least one explicit paedophile keyword. We provide in Table 2 (right-

most column) 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. Table 2 gives precise insight on this. The answers of most contributors are well balanced between all possible answers, except for three contributors (see for instance the last line of Table 2). 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		
	$\overline{T^-}$	$\overline{T^+_{>1}}$	$\overline{N(T^+)}$
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

Table 3: 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 3. 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. Figure 6 displays the difference between q^+ and q^- and between q^{++} and q^{--} for all queries. These plots grow very slowly for small values on the horizontal axis,

15

showing that only very few queries have a small difference. On the contrary, 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^{-}.
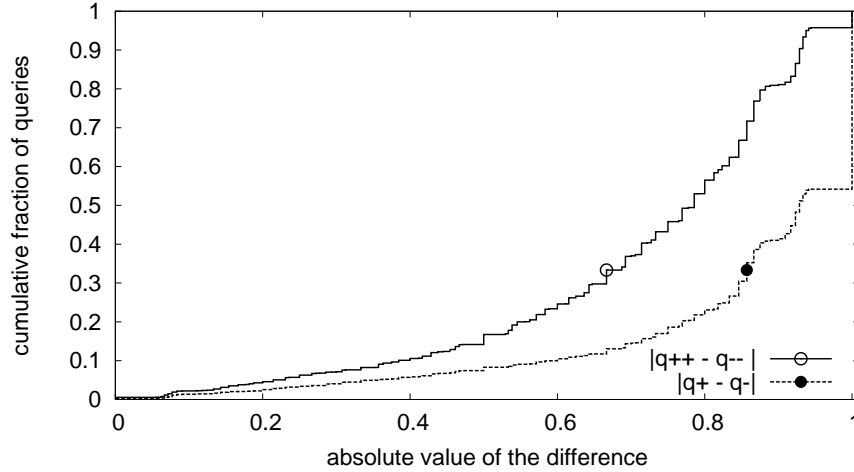

Figure 6: The cumulative distribution (CDF) of the absolute value of differences between q^{++} and q^{--} and between q^{+} and q^{-} for each query: a point at coordinate (x, y) means that a fraction y of queries have a difference lower than x.

Only 41 queries have a difference $|q^{+} - q^{-}|$ smaller than or equal to 0.1, which already is significant. Moreover, this number increases very slowly when the difference grows. 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 4.

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

Table 4: Number of queries classified as paedophile or not by experts for each considered set.

3.5 Tool assessment results

Thanks to the assessment results in Table 4 and the expressions given in Section 3.2, 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: $|\overline{T^-} \cap P^+| = 1$. As a consequence, approximating $f'^- = \frac{|T^- \cap P^+|}{|T^-|}$ by $\frac{|\overline{T^-} \cap P^+|}{|\overline{T^-}|} = \frac{1}{1,000}$ would yield very poor quality result.

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

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

(since $T_1^+ \cap P^- = \emptyset$, because all queries in T_1^+ are paedophile, see Section 3.3).

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

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

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:

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

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.

4   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 3.2, the dataset under concern being given by the context). This may be done by sampling a random subset of Q and then submit the queries it contains to experts able to decide whether they are paedophile or not. As we expect that P^+ is very small compared to Q (the fraction of paedophile queries is low), though, this is not feasible in practice: the size of a random set large enough to contain a representative number of paedophile queries is prohibitive for manual inspection.

We therefore use here the automatic paedophile query detection tool designed in Section 3, for which precise information on its error rates is available. We first estimate the fraction of queries in Q tagged as paedophile by the tool, and then infer from it an estimate of the fraction $\frac{|P^+|}{|Q|}$.

4.1   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 by the tool; 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 ratios slightly above $0.19\,\%$ for both datasets. In order to ensure the relevance of this estimate, though, we go into details below.
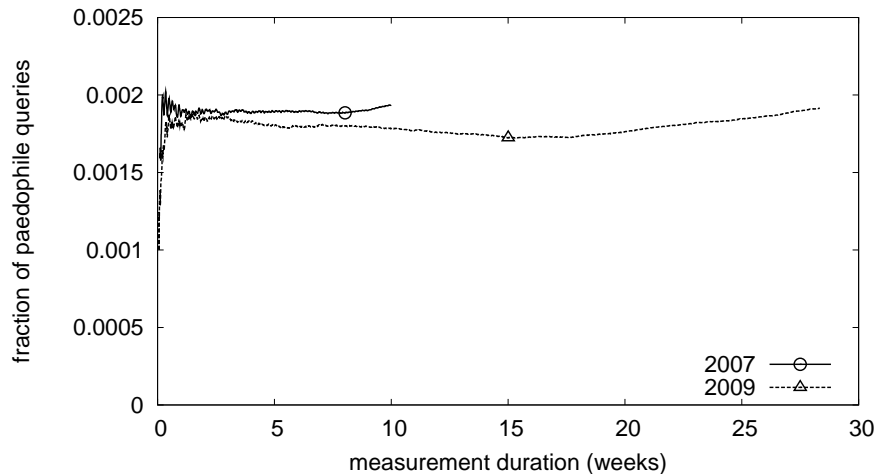

Figure 7: Fraction of paedophile queries detected in our datasets as a function of the measurement duration.

18

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 7. 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 plot in Figure 8 the cumulative distribution of the fraction of queries tagged as paedophile in all relevant one-hour, 6-hour, 12-hour and 24-hour slices of the measurements [4]. This clearly shows 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.
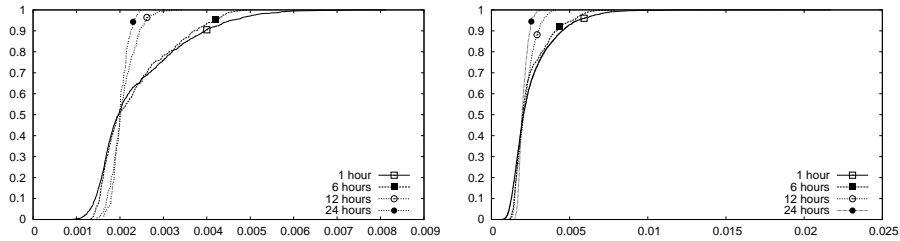

Figure 8: Cumulative distribution of the fraction of paedophile queries observed in time-slices of 1, 6, 12 and 24 hours (each plot corresponds to a size of time slice). A point at coordinates (x, y) means that we observed a fraction y of slices with less than a fraction x of paedophile queries. A sharp vertical increase around x therefore indicates that many slices were observed with a fraction of paedophile queries close to x.

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.

4.2 Inference

We established in Section 3.5 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:

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

and so

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

_____

[4]Long-term measurements are subject to interruptions (server or network failures of upgrades, for instance). As a consequence, some time slices in the measurements are not significant (no or few queries are captured during these slices), and thus we discard them when we compute slice statistics.

Using $f^- \gtrsim 24.5\,\%$ and $f'^+ \sim 1.39\,\%$ (Section 3.5), 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 $0.25\,\%$.

5   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 [5, 27]. 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.

More precisely, the following situations occur:

- several computers in a local network are connected to the internet through a gateway or firewall which performs *network address translation* (NAT): they all appear to have the IP address of the gateway or firewall, which is responsible for redistributing the traffic coming from the internet (using ports);

- internet service providers (ISP) may allocate IP addresses dynamically, *i.e.* allocate different addresses to a same computer when it connects to the internet at different times, and also allocate the same address to different computers during time;

- in various places where public internet access is provided (internet cafes, parks, libraries, etc.) or at home, different users (temporarily) have the same address;

- and dually, a same user may use several computers (at home, at work, in public places, etc.).

This makes user identification at a large scale extremely challenging, and even impossible in practice. Notice however that, in our context, 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. Notice that 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*, and we observe this in practice below.

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: the bias due to dynamic addressing may be controlled with such an approach. 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.

5.1   IP addresses and connection port numbers

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 number they used. The latter is important: it 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 9 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.

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 make us 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.

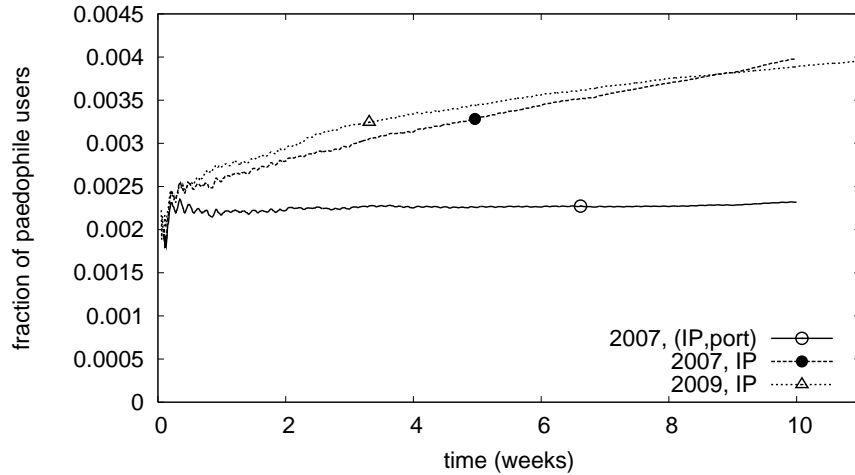On the other hand, the fraction of paedophile (IP, port) pairs in the 2007

Figure 9: Fraction of paedophile users detected in our datasets as a function of the measurement duration.

dataset has a very different behaviour: it rapidly reaches a steady regime, very similar to the fraction of paedophile queries studied in Section 4, Figure 7. 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, ports) pairs is meaningful, and that this fraction is slightly above 0.22 % here.

5.2   Varying measurement duration

Figure 9 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 phenomena. 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 10 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 (family computers, for instance). However, the probability that this happens *within a short time span* is greatly reduced. The fact that the fraction of paedophile (IP, port) pairs in windows of limited duration is very close to the overall fraction
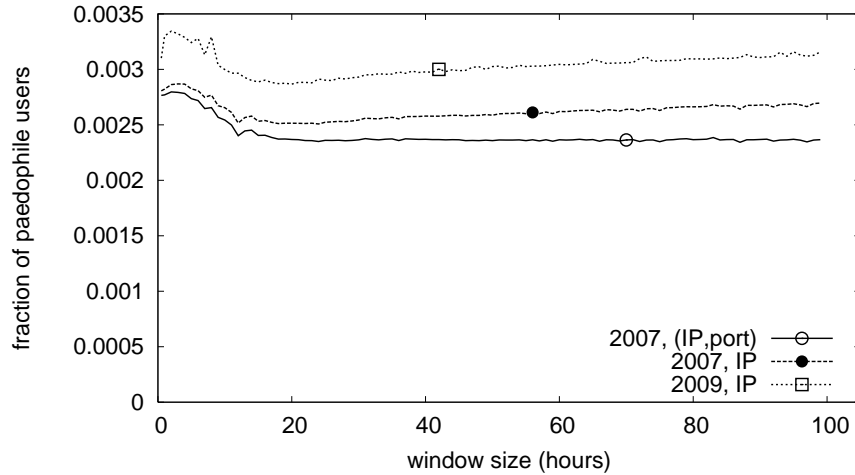
22

Figure 10: Fraction of paedophile users as a function of the measurement window size.

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 [5]. 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.

5.3 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 δ. 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. On the other hand, there is no *a priori* reason why paedophile users would make

---

[5]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.

more sessions than other users, and so we consider the fraction of paedophile sessions as an approximation of the fraction of paedophile users.
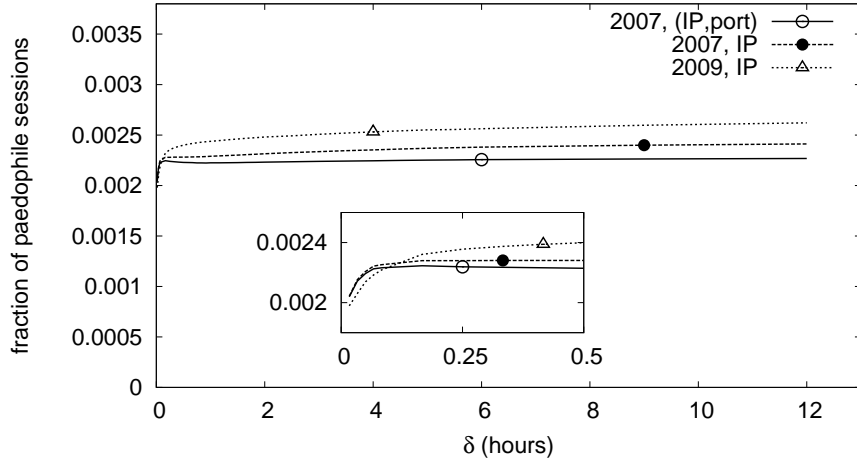

Figure 11: Fraction of paedophile sessions as a function of δ, the maximal delay between two consecutive queries in a same session.

We present in Figure 11 the fraction of paedophile sessions for different choices of δ. The fraction of paedophile sessions for very small values of δ is not relevant, because series of queries entered by a same user then belong to several sessions. For large values of δ, this fraction becomes closer and closer to the overall fraction of paedophile users in the dataset[6]. This again confirms that 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.

5.4 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.

---

[6]If δ is equal to the measurement duration, all queries entered from the same IP address or (IP, port) pair will belong to a single session.

These rates give the number of queries that the tool mis-classified. 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 Figures 10 and 11), 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.

6   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 [10, 11, 20, 23]. Some works also analyse queries entered by users [9, 15] but consider limited statistics (typically query length, number, interarrival time or redundancy). Only very few studies examine user interests in detail [12, 24, 26]. Besides, classification of search queries is an active field, with articles such as [25] in which authors aim at identifying the proportion of sexually-related web queries.

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.* [7, 16, 29]. They aim at establishing the alarming presence of paedophile activity in P2P systems, not at quantifying it, and as such cannot be compared to our work.

Up to our knowledge, only two papers deal with the quantification of pae-

dophile activity in a P2P system in a similar way as the work presented here [12, 26]. Both analyse *Gnutella* traces containing keyword-based queries.

In [12] the authors consider three sets of 10,000 queries captured during three consecutive sundays in 2005. Two reviewers (whose qualifications are unknown) manually classified these queries as related to either *illegal pornographic content* or not (they do not focus specifically on paedophile queries, but also include incest, rape and bestiality). They conclude that 1.6 % of the observed queries are related to illegal pornography. However their dataset is very small and therefore contains a very low number of paedophile queries, thus limiting the significance of these statistics. Moreover, their methodology for query classification is not automated nor fully specified, and it relies on two reviewers only.

Further work of this group on the same dataset proposes techniques for automatic discovery of paedophile keywords [13]. However, as already explained, applying such methods in our context requires a significant set of known paedophile queries, which was not available before our work. Using such methods to improve our results is one of our main perspectives.

In [26] the author considers a set of 235,513 queries, which is approximately 1,000 times smaller than our datasets (again, given the small rate of paedophile queries in such datasets, this leads to estimates with limited statistical significance). He classifies queries as paedophile or not based only on the fact that they contain a keyword in a specific list, similar to our *explicit* list introduced in Section 2. This is not sufficient, as many paedophile queries contain no such keyword, the paedophile nature arising from a combination of non-explicit keywords. Moreover, the author provides neither his dataset nor his keyword list, thus making it impossible to reproduce his results. He concludes that almost 1 % of queries in his dataset are related to child pornography.

Finally, these contributions may be seen as pioneering but limited work on paedophile query quantification when compared to our own work. Moreover, although they discuss other interesting issues like age indications in queries and filenames, presence of sub-communities, and geographic location of users, none of them address the key question of paedophile *user* quantification (Section 5).

7 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 the involvement of 21 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 for future work:

- *Extending our results to other systems.* Although our observations are consistent for two datasets collected at different times and conditions, which gives them a high level of generality, it is important to obtain similar quantifications for other datasets and/or P2P systems. Indeed, different amounts of paedophile activity may occur in different systems, for various reasons (in particular, secrecy is easier in some systems than others). One may for instance collect *Gnutella* queries like in [12, 26] and inspect them with our tool.

- *Improving knowledge and fight against paedophile activity.* We open the way to many studies and actions critical for understanding and fighting paedocriminality. For instance, the low false positive rate of our paedophile query detection tool (which may be reduced further if more false negatives are allowed) makes it suitable for filtering at server level and more generally on any search engine. The large sets of paedophile queries we provide also open the way for the study of key questions regarding paedophile activity. One may analyse for instance age indications in these queries, other topics paedophile users are interested in, and how they start and develop their interest in paedophile content. A deeper investigation of the combinations of keywords used in paedophile queries may also significantly help to improve our paedophile query detection tool. All these issues are crucial for fighting paedocriminality and designing appropriate clinical responses.

- *Applying our contributions to other contexts.* Finally, many of our contributions are not specific to paedophile activity and/or P2P systems, and could be used for other purposes. First, our methodology is directly applicable to any specific interest present in P2P systems (and more generally search engine logs) which represents a small fraction of the overall activity. This includes most deviant behaviours, such as zoophilia, rape or incest. Analysing further our datasets with appropriate methodology may also shed light on activity regarding various topics in P2P systems, like software, movie, music and pornographic contents. Notice that these datasets also allow trace-based simulations at an unprecedented scale, with great potential impact for protocol design and testing.

Acknowledgments

References

[1] Eytan Adar. User 4xxxxx9: Anonymizing query logs. In *Query Logs Analysis Workshop, International Conference on World Wide Web*, 2007.

[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.

[3] Mark Allman and Vern Paxson. Issues and etiquette concerning use of shared measurement data. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2007.

[4] 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 (Netscicom 2010), in Conjuction with IEEE Infocom*, 2010.

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

[6] James Delahunty. eD2K razorback servers seized. *Afterdawn*, 2006. `http://afterdawn.com`.

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

[8] Andrew Foss and Osmar R. Zaïane. Estimating true and false positive rates in higher dimensional problems and its data mining applications. In *IEEE ICDM Workshops*, 2008.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Candice Kim. From fantasy to reality: The link between viewing child pornography and molesting children. *Prosecutor*, 39(2), 2005.

[15] 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.

[16] Linda D. Koontz. File sharing programs – child pornography is readily accessible over peer-to-peer networks, 2003. Report of the US General Accounting Office.

[17] 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.

[18] Senator Murray. Peer-to-peer networks: file sharing software, 2005. Bill number SB 96 of the California State Senate.

[19] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large datasets. In *IEEE Symposium on Security and Privacy*, 2008.

[20] 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.

[21] 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.

[22] Hendrik Schulze and Klaus Mochalski. Ipoque Internet study, 2009. `http://www.ipoque.com`.

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

[24] Yuval Shavitt and Udi Weinsberg. Song clustering using peer-to-peer co-occurrences. In *IEEE International Symposium on Multimedia*, 2009.

[25] A. Spink, A. Koricich, and B.J. Jansen. Sexual searching on web search engines. *CyberPsychology and Behavior*, 7(1):65–72, 2004.

[26] Chad M.S. Steel. Child pornography in peer-to-peer networks. *Child Abuse & Neglect*, 2009.

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

[28] Supplementary material, 2011. `http://www-rp.lip6.fr/~latapy/antipaedo/`.

[29] Flint Waters. Child sex crimes on the Internet, 2007. Report of State of Wyoming Attorney General.

[30] Wikipedia. EDonkey network. `http://en.wikipedia.org/wiki/EDonkey_network`.

[31] Janis Wolak, Kimberley Mitchell, and David Finkelhor. Internet sex crimes against minors: the response of law enforcement, 2003. Report of the National Center for Missing & Exploited Children (NCMEC).

[32] 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).

[33] Richard Wortley and Stephen Smallbone. Child pornography on the Internet, 2006. Report of the US Department of Justice.