This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2011

# Quantifying Paedophile Queries 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. We conclude that approximately 0.25 % of queries are paedophile. Our statistics[1] 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, which is a crucial societal concern. In addition to children victimisation, the wide availability of paedophile material is a great danger for regular users, who may be exposed unintentionally to extremely harmful content. It also has a strong impact on the public acceptance of paedophilia and induces a trivialisation of such content. 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 [2]. 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.

For these reasons, knowledge of paedophile activity in P2P systems is a critical resource for law enforcement, child protection and policy making. See [2], [3], [4] for surveys on these issues. However, current knowledge on this activity remains limited and subject to controversy [4], [5], [6], [7].

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 a basic yet crucial figure: the fraction of paedophile queries[2] entered in the system. We establish reference methodology and tools for obtaining this value, and provide it in the case of the *eDonkey* system.

Obtaining precise such information on paedophile activity in P2P systems raises several challenges such as an appropriate data collection, a careful paedophile activity identification and a rigorous inference of statistics. 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. Using both of them 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.
- *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 $0.25\,\%$.

## II. Data

Although many extensions exist [8], 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 whose names contain these keywords. Clients may then ask the server for providers of selected files and contact them directly to obtain the files. Servers do not store any file; exchanges only take place between clients, from peer to peer. *eDonkey* is currently one of the largest P2P systems in use worldwide, and this has been true for several years [9].

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

[1]This work is a short version of [1], which also carefully addresses paedophile *users* quantification.

[2]In this paper, we consider a query as *paedophile* if entering it in the system leads mostly to paedophile content (child abuse images and videos mainly).

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 [10]; we performed the 2009 measurement on a medium-sized server by activating its log capabilities. 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 [11] 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 | *n/a* |

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

## III. DETECTING PAEDOPHILE QUERIES

Our tool aims at automatically identifying paedophile queries in large sets of queries. It is crucial to obtain precise estimates of its error rates to make rigorous quantification of paedophile activity possible.

### A. Tool design

Our tool performs a series of simple lexical tests (matchings of keywords in queries), each aiming 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 out final rules below, and outline the detection steps in Figure 1.

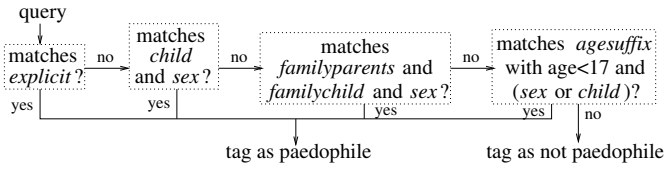


Fig. 1. Sequence of tests performed by our tool.

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.

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 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; few queries are in Russian or Chinese. 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 [11].

### 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.

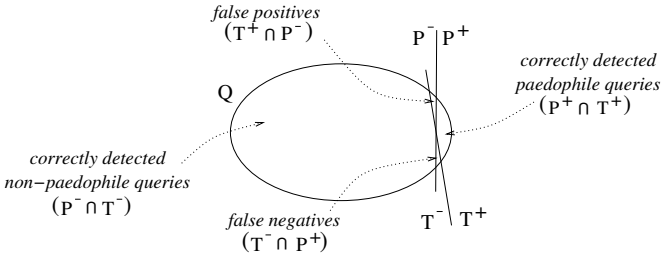


Fig. 2. Illustration of our notations. The ellipse represents the set of all queries, $Q$. The line labelled $P^-/P^+$ divides $Q$ into non-paedophile queries $P^-$ (left) and paedophile queries $P^+$ (right). Likewise, the line labelled $T^-/T^+$ divides $Q$ into the set of queries tagged as non-paedophile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

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 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 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 $X$. However, since the fraction of queries in $X$ which will be tagged as paedophile by our tool will be extremely small, 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* $q$. We therefore expect it was entered by the same user as $q$, seeking similar content. 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 a much higher probability 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 II.

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.* 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.

### C. Assessment setup

We resort to independent experts of paedophile activity who manually inspect and tag queries to identify actual paedophile queries in some specific sets.

*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.* which are tagged as paedophile by the tool, are composed of only one word. Then, this word is necessarily in the *explicit* paedophile keywords list described in Section III-A. If such a keyword appears in a query, then it surely is a paedophile one. We therefore increase the efficiency of our assessment by not submitting these one-keyword queries to experts. We denote by $T_1^+$ this set of queries, 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 we 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.

*Interface.* We set up a web interface 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. There were 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.

*Experts.* 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. We contacted many specialists and obtained a set of 21 volunteers for participating to our assessment task. They 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. Their approach of paedophile activity is different and, as such, complementary.

However, 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 with erroneous answers. In order to identify such cases, we computed for each contributor the percentage of queries with at least one explicit paedophile keyword tagged as *paedophile* or *probably paedophile*. This percentage is above $95\,\%$ for all contributors but one ($87.3\,\%$), thus showing that they recognise these keywords.

Finally, we obtained 42,059 answers for the 3,000 queries with an average of slightly more than 14 experts assessing each query.Each of our 21 participants tagged more than 300 queries, and 12 tagged more than 2,000.

### D. Expert classification of queries

For each query $q$ submitted to experts, 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. In general, we have $q^{+}+q^{-}<1$ as some *I don't know* answers were provided. 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 $q$. We check that by computing $|q^{+}-q^{-}|$ and $|q^{++}-q^{--}|$ for each query. The difference $|q^{++}-q^{--}|$ is above $0.8$ for $1,305$ queries (over $3,000$) and $|q^{+}-q^{-}|$ is above $0.8$ for $2,308$ queries. Only $41$ queries have a difference $|q^{+}-q^{-}|$ smaller than or equal to $0.1$, a value which shows a significant agreement among experts. 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 II.

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

TABLE II
EXPERT CLASSIFICATION OF QUERIES FOR EACH CONSIDERED SET.

### E. Tool assessment results

We may now compute estimates of the false positive and false negative rates of our tool.

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:

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

(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 $|\overline{T^{+}_{>1}}\cap P^{-}|\cdot\frac{|T^{+}_{>1}|}{|\overline{T^{+}_{>1}}|}$ which leads to:

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

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, since $|T^{+}\cap P^{-}|$ 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:

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

However, as mentioned in Section III-B, this value is an under-estimate, because we assessed neighbours of detected paedophile queries instead of random paedophile queries. Though, we expect this bound to be reasonably tight and discuss this carefully in the following.

## IV. FRACTION OF PAEDOPHILE QUERIES

We estimate the fraction of paedophile queries in our two datasets by first computing the fraction of queries tagged as paedophile by our tool, and then infering from it an estimate of $\frac{|P^{+}|}{|Q|}$ (notations defined in Section III-B).

### A. Fraction of automatically detected queries

For both datasets, the fraction of queries tagged as paedophile, *i.e.* $\frac{|T^+|}{|Q|}$, may be trivially obtained by computing the set $T^+$ of queries tagged as paedophile by the tool, and then dividing it by the total number of queries. This rate is slightly above $0.19\,\%$ for both datasets. In order to ensure the relevance of this estimate, though, we go into details below.

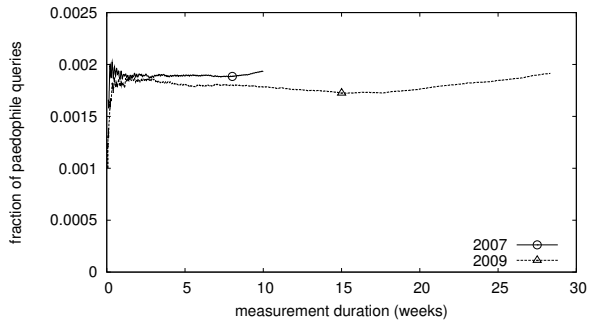


Fig. 3. Fraction of paedophile queries detected in our datasets as a function of the measurement duration.

We 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.

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. Related work

Up to our knowledge, only two papers deal with the quantification of paedophile queries in a P2P system in a similar way as the work presented here ([5], [6]). They both use datasets almost 1,000 times smaller than ours and do not describe precisely their methods. Therefore, they may be seen as pioneering but limited work on paedophile activity quantification when compared to our own work.

## VI. 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 in both of our datasets.

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 [5], [6] 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 (*e.g.*, methodology) are not specific to paedophile activity and/or P2P systems, and could be used to detect rare contents in other contexts.

## References


[1] M. Latapy, C. Magnien, and R. Fournier, "Quantifying paedophile activity in a large P2P system," *Submitted*.

[2] R. Wortley and S. Smallbone, "Child pornography on the Internet," 2006, report of the US Department of Justice.

[3] 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.

[4] 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).

[5] 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.

[6] C. M. Steel, "Child pornography in peer-to-peer networks," *Child Abuse & Neglect*, 2009.

[7] F. Waters, "Child sex crimes on the Internet," 2007, report of State of Wyoming Attorney General.

[8] Wikipedia, "EDonkey network," http://en.wikipedia.org/wiki/EDonkey_network.

[9] H. Schulze and K. Mochalski, "Ipoque Internet study," 2009, http://www.ipoque.com.

[10] 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.

[11] "Supplementary material," http://www-rp.lip6.fr/~latapy/antipaedo/.