Technical report on

Quantification of Paedophile Activity in a Large P2P System

Measurement and Analysis of P2P Activity Against Paedophile Content project
`http://antipaedo.lip6.fr`

Matthieu Latapy [1], Clémence Magnien, and Raphaël Fournier

Abstract

In this work, we explore two basic but crucial statistics: the fraction of paedophile queries entered by users of a large P2P file exchange system, and the fraction of involved users. In order to do so, we carefully inspect two huge datasets of more than one hundred million queries recorded in two very different contexts. We then use a state-of-the-art tool for automatic detection of paedophile queries. Using the known rate of false positives and negatives of this tool, we obtain an estimate of the fraction of paedophile queries. Turning this into statistics on users is challenging, as one has no clear way to identify a user in such measurements. We explore two natural approaches, obtaining clear indication of the fact that one is misleading and the other is relevant. We finally obtain approximately 2 per thousand for both the fraction of paedophile queries and users. Although the reliability of these values may be improved further, they constitute a significant progress over the current situation, providing quantitative information on the amount of paedophile activity in a P2P system for the first time at this level of precision and reliability.

1 Introduction

It is widely acknowledged that peer-to-peer (P2P) file exchange systems host large amounts of paedophile activity (users providing and/or looking for files with paedopornographic content) [6]. However, obtaining precise information on this activity, even very basic one such as the number of involved users, is extremely challenging. Indeed, P2P systems are distributed by nature, and so lack a central authority with a view of user activity. Moreover, their sheer size makes the study of P2P exchanges as a whole almost impossible: at least dozens of millions of users exchange at least millions of files on a daily basis, which accounts for the use of most internet capabilities (bandwidth). In addition, these systems are very dynamic: users and files arrive in the system and leave it very rapidly [7]. Last but not least, users are identified by their IP address and port in the best case, which leads to much ambiguity: several users may use the same computer (at home or in public access points in particular); several computers may use the same address (behind a firewall or

[1] Contact author: `Matthieu.Latapy@lip6.fr`

because of dynamic allocation of addresses); and one user may use several computers and thus several addresses (home and office, for instance) [1].

As a consequence, our current knowledge of P2P activity remains very limited, and the situation regarding *paedophile* activity in such systems is even more alarming: monitoring paedophile activity in particular is even more challenging as it means that one has to inspect more precisely the nature of exchanges. This means that more detailed data needs to be collected, which raises serious privacy concerns, and that ambiguity of terms, multilingual environments, fake files (files whose content differs significantly from their name), and many other difficulties have to be handled.

Finally, although previous studies succeeded in giving rough estimates of P2P traffic and user activities [3], almost nothing precise and rigorous is currently known regarding paedophile activity in P2P systems. Even simple quantities like the number of involved users, the amount of files with paedophile content, the existence and proportion of different kinds of such users and files, and more subtle ones, remain out of reach.

However, gaining knowledge of paedophile activity in P2P systems is extremely important. It is indeed a crucial resource for policy making [9], affectation of law enforcement personnel and resources, as well as P2P and internet regulation. Moreover, the wide availability of paedophile content provided by P2P systems and the fact that these contents may be easily accessed (by children in particular) is an important societal concern. It may have a strong impact on the public acceptance of paedophilia, and on real-world behaviours [2, 4].

We present the first large-scale study which succeeds in providing precise and rigorous quantification of paedophile activity in a large P2P system. In order to do so, we carefully examine two sets of queries entered by users of one of the largest P2P systems currently in use, which may be considered as representative (Section 2). We then use a tool for automatic identification of paedophile queries entered by users [2], which is the current state-of-the-art on this topic. This tool has a known rate of false positives (queries which it identifies as paedophile but are not) and a tight lower bound for its rate of false negatives (paedophile queries which it identifies as non-paedophile) [5]. Combining this information with appropriate statistical inference methods, we observe the fraction of paedophile queries in our dataset (Section 3). Going further, we explore our ability to obtain an estimate of the fraction of users entering paedophile queries. This is much more difficult as we have only IP address information in the datasets, as well as connection port for one of the datasets. We however show that, unlike IP address only, this pair of information seems to be reasonably efficient in characterising a user in our context (Section 4).

[2] Manually inspecting a random set of queries would also lead to an estimate of the fraction of paedophile queries but, because such queries are relatively rare compared to others, this would need manual inspection of a huge sample, which is not feasible in practice.

2 Data

The data used for this work consists in recordings of keyword-based queries received by two *eDonkey* servers during two different periods of time of several weeks each [8]. Each query is associated to a timestamp and the IP address from which it was received. One dataset contains in addition the connection port used, but the other does not provide this information. Their key features are summarised in Table 1.

	date	duration	nb queries	nb IP	nb IP+port
First measurement	2007	10 weeks	127 316 861	28 395 244	61 683 017
Second measurement	2009	15 weeks	106 344 062	16 020 976	*unknown*

Table 1: Main features of the two datasets we use here.

Almost three years elapsed between the collection of these two datasets, and P2P protocols and uses evolved much during this period (users are not the same, *eDonkey* evolved significantly, other protocols became much more used, etc). Using two datasets with so many differences is important as this will make it possible to confront our estimations on both datasets and thus to evaluate their robustness to changes in input data.

In both cases, the data are carefully anonymised: IP addresses are replaced by integers which reflect their order of appearance (the 1-st address observed is replaced with 0, the 2-nd with 1, the 3-rd with 2, and so on). Likewise, the text queries are normalised (all non alphanumeric characters are replaced by spaces, and the queries are splitted into words according to spaces), and words which appear less than 100 times are anonymised. Notice however that this anonymisation is coherent: a same IP address or word will always be replaced by the same integer, thus making it easy to recognise that a query was received from the same address (which will be crucial in Section 4).

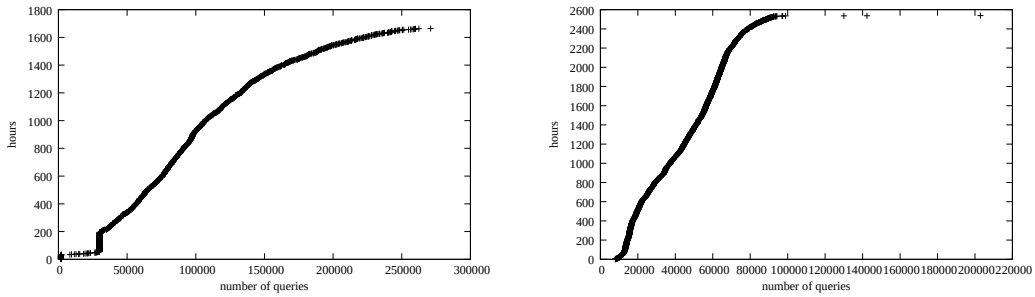

Figure 1: Cumulative distribution of the number of queries observed in each one-hour time slice: a point at coordinates (x, y) indicates that we saw y one-hour time-slices with less than x queries. We observe that some hours led to the observation of only very few queries, thus leading to non significant statistics. We will therefore remove them from our computations. Left: first dataset; right: second one.

Finally, notice that such long-term measurements are subject to interruptions due to server or network failures, upgrades, and other technical reasons. As a consequence, some

time slices in the measurements may not be significant, and we must discard them when we compute statistics. For instance, we display in Figure 1 the distribution of the number of queries received in each time-slice of one hour. It shows that some one-hour slices contain abnormally low numbers of queries, and so we will discard these slices when we will compute statistics on one-hour slices. This technique ensures that our statistics will not be biased by such abnormal events occurring to the server.

3 Fraction of paedophile queries

Any set Q of queries may be divided into two disjoint sets: the set P^+ of paedophile queries and the set $P^- = Q \setminus P^+$ of non-paedophile queries.

Estimating $\frac{|P^+|}{|Q|}$, *i.e.* the fraction of paedophile queries in Q, 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.

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

3.1 Fraction of automatically detected queries

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

This may be trivially obtained by computing the set $\overline{P}$ of queries tagged as paedophile by the tool, and then divide it by the total number of queries. We obtain this way ratios slightly lower than 0.15 % for both datasets. In order to ensure the relevance of this estimation, though, we will enter in more details in the results.

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 2. It clearly shows that this fraction converges rapidly to a reasonably steady value, slightly lower than 0.15 %; changing this value significantly would need a drastic change in the data.

Going further, we plot in Figure 3 the distribution of the fraction of queries tagged as paedophile in all relevant one-hour, 3-hour, 12-hour, 24-hour and 36-hour slices of the measurements. 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 (in particular for the second

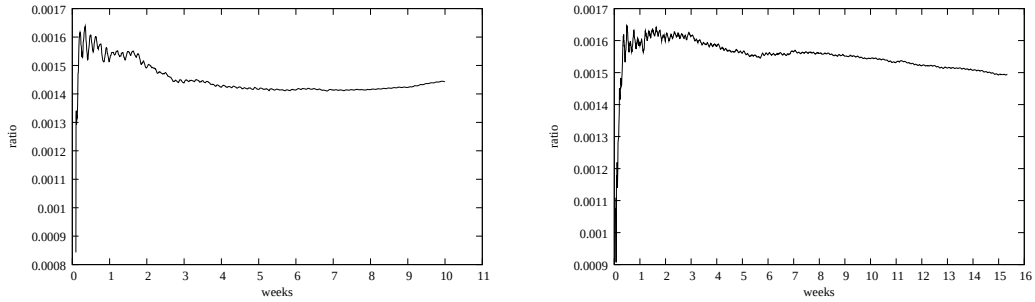

Figure 2: Fraction of paedophile queries (vertical axis) observed in the datasets we consider, as a function of the measurement duration (horizontal axis, in weeks). Left: first dataset; right: second one.

measurement). Again, this ratio is close to 0.15 %, in accordance with the computations above, but slightly above this value.

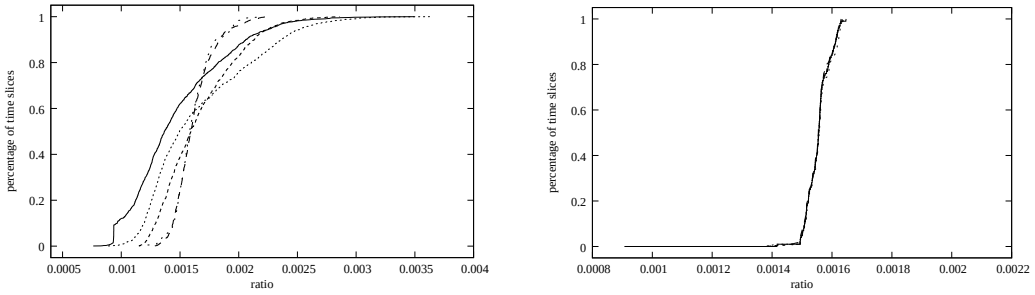

Figure 3: Cumulative distribution of the fraction of paedophile queries observed in time-slices of 1, 3, 12, 24 and 36 hours (each plot corresponds to a size of time slice). A point at coordinates (x, y) means that we observed y slices with less than a fraction x of paedophile queries. A sharp vertical increase around x therefore indicates that many slices were observed with a fraction of paedophile queries close to x. Left: first dataset; right: second one.

Finally, we conclude that the fraction of queries tagged as paedophile may be approximated by $\frac{|F^+|}{|Q|} \sim 0.15\,\%$.

3.2 Inference

Let us consider a set Q of queries, and let us denote by P^+ (resp. P^-) the set of paedophiles (resp. non-paedophile) queries in Q. Let us denote by F^+ (resp. F^-) the subset of Q which our filter tags as paedophile (resp. non-paedophile).

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

In such situations, two natural definitions of false positive and negative rates coexist. Both will prove to be useful here.

First, one may consider the rate of false positives (resp. negatives) when all inspected queries are non-paedophile (resp. paedophile). If we run the filter on Q, this leads to:

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

If f^+ and f^- are known, as well as the size of F^+, one may derive from this an estimate of the size of P^+ as follows:

$$\begin{aligned} |F^+| &= |F^+ \cap P^+| + |F^+ \cap P^-| \\ &= f^+|P^-| + |P^+|(1-f^-) \\ &= f^+(|Q| - |P^+|) + |P^+|(1-f^-) \\ &= f^+|Q| + |P^+|(1-f^- - f^+) \end{aligned}$$

and so

$$|P^+| \sim \frac{|F^+| - f^+|Q|}{1-f^- - f^+}$$

The other natural approach consists in considering the probability that the filter is wrong when it gives a positive (resp. negative) answer:

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

If f'^+ and f'^- are known, as well as the size of F^+ and F^-, one may derive from this an estimate of the size of P^+ as follows:

$$\begin{aligned} |P^+| &= |P^+ \cap F^+| + |P^+ \cap F^-| \\ &= |F^+|(1-f'^+) + |F^-|f'^- \end{aligned}$$

In our situation, though, a reliable estimate is available only for f^- and f'^+. Indeed, the fact that P^+ is very small compared to Q makes it prohibitive to try to estimate f^+ and f'^-. See [5].

As a consequence, we have to infer the size of P^+ from f^- and f'^+. This may be done as follows:

$$\begin{aligned} |P^+| &= |P^+ \cap F^+| + |P^+ \cap F^-| \\ &= |F^+|(1-f'^+) + |P^+|f^- \end{aligned}$$

and so

$$|P^+| = \frac{|F^+|(1-f'^+)}{1-f^-}$$

3.3 Result

As $f^- \sim 24.9\%$ and $f'^+ \sim 1.13\%$ are given [5], and as we have $\frac{|P^+|}{|Q|} \sim 0.15\,\%$ for both datasets from previous section, we obtain:

$$\frac{|P^+|}{|Q|} \sim 0.2\,\%$$

for both datasets.

In other words, approximately 2 queries over 1000 are paedophile in our two datasets.

4 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 [1, 7]. Computers are 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 and/or that two computers are using 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;
- at home or at offices, as well as in various places where public internet access is provided (internet coffees, parks, libraries, etc), various 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 will lead to interpret the obtained 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 underlying users entered paedophile queries, then the whole series will be considered as coming from a paedophile user. Note that since the overall fraction of users entering paedophile queries is very small, it happens very infrequently 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 will call this phenomenon *pollution*, and we will observe this in practice below.

In the data we consider, only two pieces of recorded information may lead to distinguish between users: the IP address from which they sent the queries, and the connection port used by the application. This last information is important: it makes it possible to distinguish between several users in a same local network with a firewall. However, connection port information is available only in the first dataset we consider, therefore we will focus on it.

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

We display in Figure 4 the fraction of paedophile users observed during the measurements, under each hypothesis.

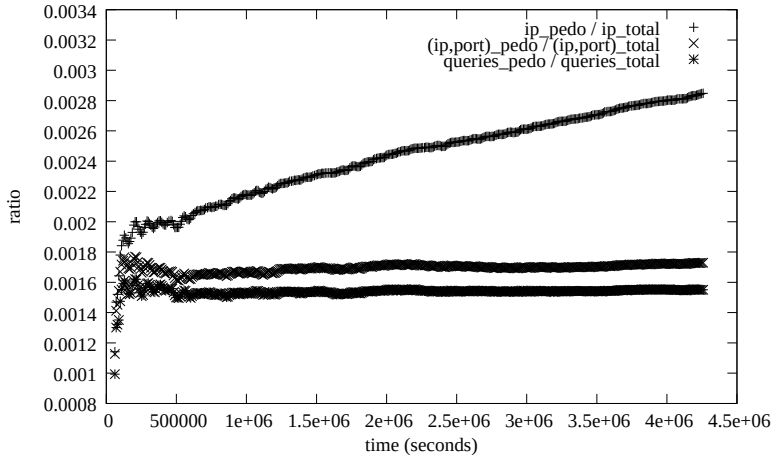

Figure 4: Fraction of paedophile users (vertical axis) observed in the datasets we consider, as a function of the measurement duration (horizontal axis). The two topmost plots correspond to two assumptions: IP addresses are sufficient to distinguish between users (upper plot); and IP addresses and ports are sufficient (middle plot). We also display for comparison the fraction of detected paedophile queries (lower plot), already plotted in Figure 2.

The plot obtained when we suppose that IP addresses are sufficient to distinguish between users clearly grows with the measurement duration. This reveals the *pollution* phenomena sketched above: as IP addresses may host different users during time, and as one paedophile user is sufficient to make us consider the corresponding address as paedophile, then the probability that any given address will 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 plot obtained when we distinguish between users with both their IP address and port has a very different behaviour: it rapidly reaches a steady regime, very similar to the fraction of paedophile queries studied in Section 3. This shows that pollution due to dynamic allocation of addresses and ports, and to change of users for a same computer, is not significant in this case: although it may have some impact, it is negligible in a measurement of the scale and duration of the one we observe here.

Still, a given user may use several IP addresses and/or ports; then, either he/she sends similar queries from all his/her addresses, and then this does not impact our estimates; or he/she only uses some addresses for paedophile queries, and then we underestimate the fraction of paedophile users. The fact that the filter for automatic detection of paedophile queries has a very small false positive rate and a much larger false negative rate also biases the results in this direction. We therefore conclude that the fraction of users which we observe sending paedophile queries in our dataset is an underestimate of the true value.

We finally conclude that distinguishing between users using IP address and port seems sufficient in our context. The observed fraction of paedophile users (paedophile pairs of IP address and port) is above 0.17 %, indicating that of the order of 2 users over 1000 enter paedophile queries in our observations.

5 Conclusion and Perspectives

Relying on two large-scale measurements of keyword-based queries submitted to the *eDonkey* P2P system, and using an automatic paedophile query tool for which false positive and false negative rates are known, we evaluated two quantities of prime importance: the fraction of paedophile queries entered in the system; and the fraction of users sending such queries. Both are close to a rate of 2 per thousand.

It is the first time that quantitative information on paedophile activity in P2P systems is obtained at this level of precision, reliability, and at such a scale. This information may help in policy making, and significantly improves awareness on what actually occurs regarding paedophile activity in P2P systems.

It must be clear, though, that our work may be improved in several ways.

First, other datasets should be considered, in particular datasets from other P2P systems. Technical features of such systems may indeed have an influence on their use by paedophiles, and thus the amount of paedophile activity may vary between systems.

Another possible improvement deals with the notion of user in such a system. We have shown here that IP addresses and ports seem sufficient, but it would be interesting to

deepen this. In particular, one may observe the fraction of users who send several, or many paedophile queries; one may investigate further the influence of measurement duration on our observation; and one may explore the impact of language and local encodings (using geolocation information, for instance).

Another direction of interest is the study of *sessions*, *i.e.* sets of queries from the same IP address (and port) such that two consecutive queries are not separated by more than a given delay (to determine). Determining the fraction of paedophile sessions (and studying them) would be easier than the fraction of paedophile users, although more difficult to interpret.

Acknowledgements. We warmly thank the administrator of the `peerates.net` *eDonkey* server for his help in collecting data. This work is supported in part by the MAPAP SIP-2006-PP-221003 and ANR MAPE projects.

References

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

[2] P. Greenfield. Inadvertent exposure to pornography on the internet: Implications of peer-to-peer file-sharing networks for child development and families. *Journal of Applied Development Psychology*, 2004.

[3] D. Hughes, S. Gibson, J. Walkerdine, and G. Coulson. Is deviant behavior the norm on p2p file sharing networks? *IEEE Distributed Systems Online 7(2)*, 2006.

[4] C. Kim. From fantasy to reality: The link between viewing child pornography and molesting children. *Prosecutor 39(2): 17-18,20,47*, 2005.

[5] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Technical report on the *Automatic Detection of Paedophile Queries*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[6] United State General Accounting Office. File sharing programs: Child pornography is readily accessible over peer-to-peer networks. 2003.

[7] Daniel Stutzbach and Reza Rejaie. Understanding churn in peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM on Internet measurement - IMC '06*, page 189. ACM Press, 2006.

[8] F. Aidouni, M. Latapy, and C. Magnien. Ten weeks in the life of an edonkey server. *Proceedings of HotP2P'09*, 2009.

[9] J. Wolak, K. J. Mitchell, and D. Finkelhor. Internet sex crimes against minors: The response of law enforcement. *National Center for Missing and Exploited Children*, 2003.

Project MAPAP SIP-2006-PP-221003.

http://antipaedo.lip6.fr

Supported in part by the European Union
through the *Safer Internet plus Programme*.

http://ec.europa.eu/saferinternet