Technical report on

Behaviours of Users Entering Paedophile Queries

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

Jean-Loup Guillaume[1], Matthieu Latapy, Bénédicte Le Grand and Clémence Magnien

Abstract

We use here three large-scale datasets about keyword-based queries and filenames in the *eDonkey* system to gain significant insight on behaviours of paedophile users. We identify paedophile queries and filenames using an automatic tool and then study the main statistics describing related users. We study in particular the proportion of users interested in paedophile content, as well as the specificity of their interest, i.e. whether they are interested only in paedophile content or in other kinds of content too. We then study age indications present in paedophile filenames and queries, which allows a finer analysis of user interests. We also observe in more details a few specific users, which makes it possible to gain more insight on their behaviours. We conclude that very different behaviours co-exist, with very different numbers of paedophile queries and focus on different kinds of paedophile content.

1 Preliminaries

We use here three datasets in order to study paedophile activity in the *eDonkey* network, in particular user behaviours. These datasets correspond to captures of traffic at two *eDonkey* servers during several weeks in 2007 and in 2009. We summarise their main features below. See the technical reports on data collection and quantification of paedophile activity [1, 2] for more details.

Our first measurement, which we call *qu_2007*, was performed in 2007. It consists in a 10 weeks (70 days) capture of the keyword-based queries received by a large *eDonkey* server. Each such query is described by a timestamp (in seconds), the anonymised IP address of the client who made the query, its port number and finally the list of keywords used to formulate the query. In the following, two distinct datasets will be derived from the initial *qu_2007* dataset: *qu_2007*-IP, where only the IP address is used to identify users, and *qu_2007*-IP+PORT, where both IP address and connection port are used to identify users. Indeed, the problem of identifying users in the internet over a long period of time is a challenge in itself, see the technical report on quantification of paedophile activity [2]. To avoid this issue, we will consider here both versions of the dataset.

During the same measurement, we also captured filenames associated to files available in the system. Each file is identified by a unique identifier and a list of filenames, each being composed of a sequence of keywords. In this report, we are only interested in filenames; we will therefore ignore the fact that a given file may have different names. The corresponding dataset is called *fid_2007*.

[1]Contact author: Jean-Loup.Guillaume@lip6.fr

Our second measurement, called *qu_2009*, was performed in 2009. It consists in a 102 days (14 weeks and a half) capture of keyword-based queries received by an *eDonkey* server. The format of each query is similar to the one in the *qu_2007* sample, without port number. This means that this dataset is similar to *qu_2007*-IP.

In addition to these datasets, we use a filter able to detect strings, i.e. queries or filenames, which are likely to correspond to paedophile content, see [3]. In the following, a *paedophile query* (resp. *paedophile filename*) is simply a query (resp. filename) which matches this filter. Similarly, and although it is certainly more subtle, we call any user who entered at least one paedophile query a *paedophile user*. These approximations will prove to be relevant in the following.

Finally, key features of the considered datasets are given in Table 1.

	qu_2009	*qu_2007*	*fid_2007*
number of items	106 344 062	127 316 861	24 666 569
number of paedophile items	172 524	204 746	18 357
fraction of paedophile items	0.16% = 1/616	0.16% = 1/622	0.07% = 1/1344

Table 1: Number and proportion of paedophile queries or filenames in each dataset.

In the following, we first focus in Section 2 on queries made by users so as to understand which proportion of them are of paedophile nature. In Section 3 we analyse age indications which often appear in queries (mainly paedophile ones). Finally, in Section 4 we study specific users to show the different kinds of behaviours which can be observed.

2 Queries by users

In this section we study users to understand how many make paedophile queries and, among them, what proportion of their queries are paedophile. This information is of prime importance to determine if users entering paedophile queries are mainly interested in such content, or mix such queries with other interests.

	qu_2009	*qu_2007*-IP	*qu_2007*-IP+PORT
users	15 277 005	28 395 243	61 683 559
users with paedophile queries	58 455	90 405	111 086
prop. of paedophile users	0.38% = 1/261	0.32% = 1/314	0.18% = 1/555

Table 2: Number and proportion of users performing queries/paedophile queries.

Table 2 presents the number of users in each dataset, depending on whether they make paedophile queries or not. Distinguishing users based on IP addresses yields fewer users than using IP+port, as expected, which is why there are much more users in the *qu_2007*-IP+PORT dataset than in the *qu_2007*-IP one. However, there are proportionally less users performing paedophile queries when the connection port is used, a strong indication of the fact that user identification is a key issue in our context, see also [2]. This will be detailed below.

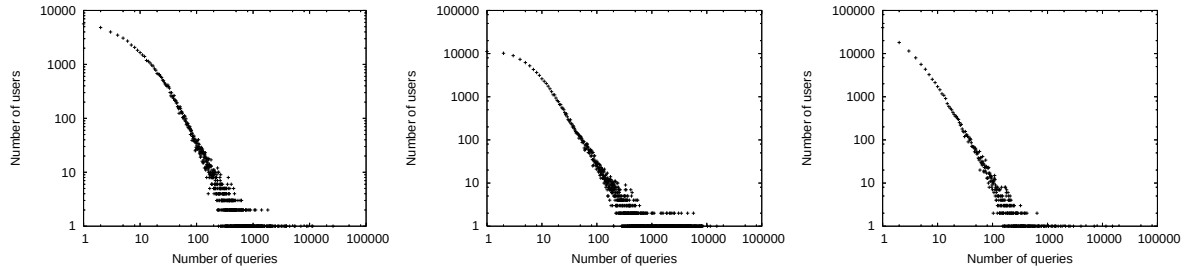

Figure 1: Distribution of the number of queries by user. From left to right: *qu_2009*, *qu_2007*-IP and *qu_2007*-IP+PORT datasets.

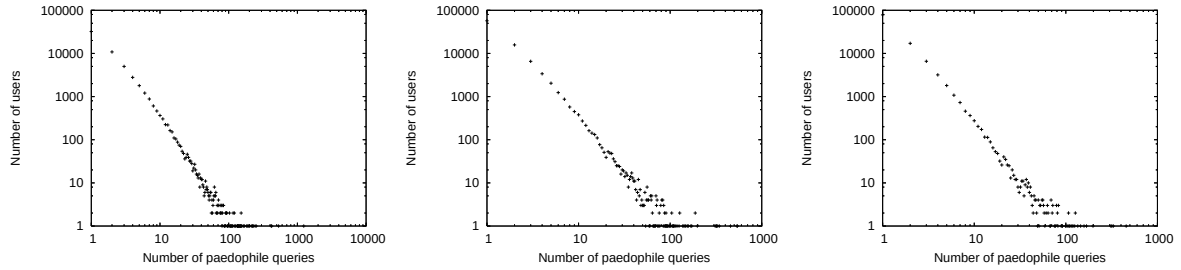

Figure 2: Distribution of the number of paedophile queries for each user. From left to right: *qu_2009*, *qu_2007*-IP and *qu_2007*-IP+PORT datasets.

Figure 1 (resp. Figure 2) displays the number of queries (resp. paedophile queries) sent by each user. As many human-driven behaviours, these distributions are very heterogeneous: while most users only perform few queries (paedophile or not), some make a large number of queries, and all intermediary behaviours are observed. This is an unsurprising, but important fact: there is no notion of *typical* user in such situations.

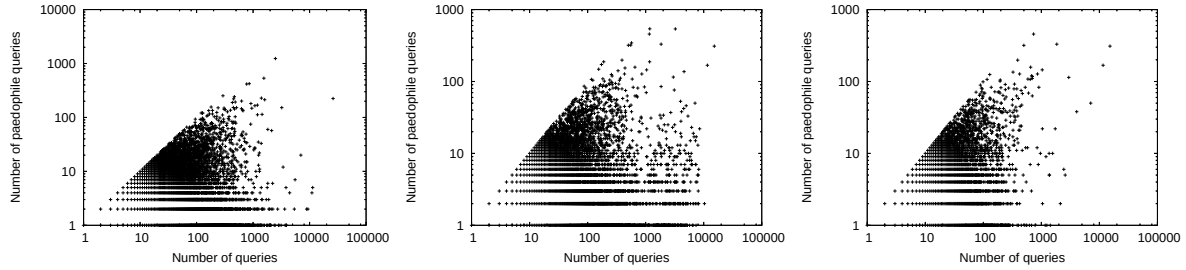

Figure 3: Scatter plot of the number of queries by user versus number of paedophile queries, for all users who made at least one paedophile query. From left to right: *qu_2009*, *qu_2007*-IP and *qu_2007*-IP+PORT datasets.

However, this gives no indication on which users make more or less paedophile queries relative to the total number of queries they perform. Figure 3 is a scatter plot of Figures 1 and Figure 2: each dot (x, y) represents one user, having performed x queries, among which y are paedophile. Figure 4 displays the average value of y, for all x.

We observe from these plots that different behaviours are possible: some users make many paedophile queries and some only a few, and there is no strong dependency on the total number of queries they make. However, on each dataset there are dots near the

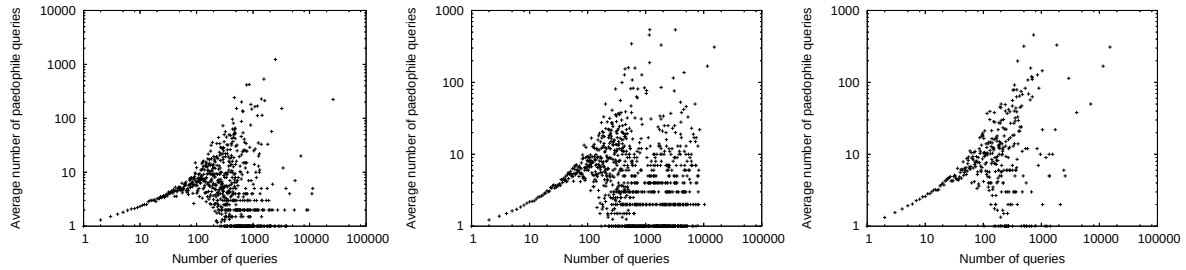

Figure 4: Average number of paedophile queries (vertical axis) of users having sent a given number of queries (horizontal axis), for all users who made at least one paedophile query. From left to right: *qu_2009*, *qu_2007*-IP and *qu_2007*-IP+PORT datasets.

diagonal $y = x$, which means that most queries of the corresponding users are paedophile. Conversely, dots close to the x axis correspond to user making very few paedophile queries in proportion to their total number of queries. Note also that most users have not made any paedophile query and therefore do not appear on these plots.

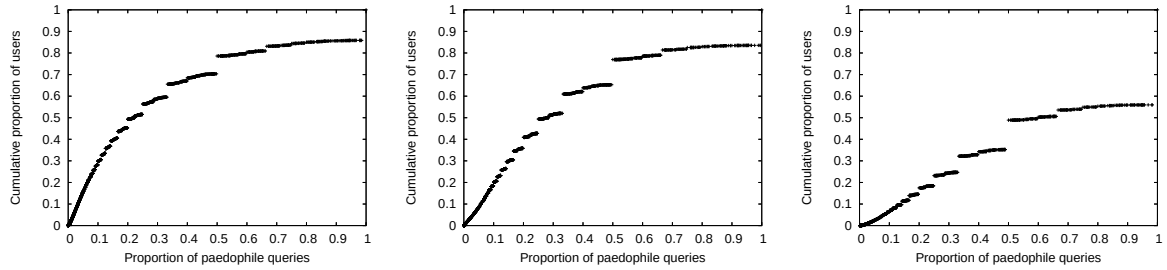

Figure 5: Cumulative distribution of the number of paedophile queries per user, for all paedophile users. From left to right: *qu_2009*, *qu_2007*-IP and *qu_2007*-IP+PORT datasets.

To detail this, Figure 5 displays the cumulative distribution of the proportion of paedophile queries per user for all users who have made paedophile queries. While *qu_2009* and *qu_2007*-IP are very similar, we can observe that the *qu_2007*-IP+PORT plot increases much more slowly. This is due to the fact that, in this last dataset, we distinguish between different users with the same IP address, and so the obtained users are more focused. Indeed, if two users have the same IP address and one of them is making paedophile queries, these users will be seen as one single user performing half as much paedophile queries in *qu_2009* and *qu_2007*-IP. In *qu_2007*-IP+PORT, on the contrary, both users will be distinguished by their connection port.

Finally, on *qu_2007*-IP+PORT, more than 40% of users who do paedophile queries never search for other contents of the network. Using IP information only therefore leads to both overestimating the proportion of paedophile users [2] and underestimating their interest in paedophile content.

3 Ages in filenames and queries

In this section we inspect age indications contained in queries and filenames, which provide strong indications on the type of content users are looking for or are sharing.

We mainly present co-occurrences of age indications with specific paedophile keywords (from the *explicit* keyword list, see [3]). A number is considered as an age indication if it is followed by "yo" or "yr" (with or without space between the age and "yo"/"yr"). For instance, "12yo" and "5 yr" are age indications.

Table 3 provides the number of queries containing age indications, as well as the corresponding average age, within all queries and within paedophile queries (lines 1 and 2 respectively). The other lines are similar but restricted to the co-occurrence of age indications with a specific paedophile keyword.

We first observe that age indications are much more frequent in paedophile queries or filenames than in general queries. Indeed, 1 query over 616 for *qu_2009* (and 1 query over 622 for *qu_2007*) are labelled as paedophile, while approximately one fifth to one third of queries or filenames containing an age indication are paedophile. In paedophile queries, it is not likely to find age indication co-occurring with explicit paedophile keywords. For instance, for the *qu_2009* dataset, only around 12% of paedophile queries containing an age indication also contain a specific paedophile word from the above list.

Furthermore, the average age in these queries is lower in paedophile queries than in general queries. This comes from the fact that some non paedophile queries (but generally pornographic ones) contain such indications, for instance: `cochonne 41yr`, `60yo`; or the following filenames: `18yo rape and crying with feet bound bd sm bdsm torture slave bondage wmv`, `granny -142 60yo mature oma s und opa s im sex rausch 1 mpg`, `janet jackson 20yo 02 so excited ft khia mp3`, are not paedophile but contain age indications.

Finally, the average age in queries is lower by more than one year than the average age in filenames. Therefore, there is a demand for paedophile content involving younger children than what is actually offered.

Concerning specific keywords, one can see that their use is very variable. Some keywords are very often used in conjunction with age indications, while some are not. Furthermore, the average age recorded in conjunction with these keywords can vary a lot. For instance, the average age for "babyj" is lower than 6, while it is around 9 for more classical paedophile keywords such as "pthc" or "ptsc". Although direct average values are not conclusive, this confirms that different paedophile keywords are used depending on the context.

3.1 Ages with specific paedophile keywords

To be more precise than simply representing average ages in queries or filenames, Figures 6 and 7 display age distributions in the datasets.

Furthermore, even in paedophile queries, older ages can be found, for instance `hussyfan pthc new old man 58yr fuck a young girl 12yr privat 2009 wmv` or `pthc family dad 43yo mom 38yo girl 13yo boy 9yo girl`.

Some age indications can also be found in non paedophile queries. For instance, a query such as: `english kids educ pc game caillou s birthday party win98 2`

	qu_2009		*qu_2007*		*fid_2007*	
	nb	average	nb	average	nb	average
all	81 668	10.87	74 470	10.33	13 662	11.61
pedo	18 945	9.82	24 048	10.19	5 837	11.54
babyj	30	5.96	82	5.29	70	5.75
babyshivid	15	3.33	93	6.09	66	3.72
childlover	52	11.51	287	9.81		
childporn	5	4.6	18	9.55		
childsex	20	11.5				
childfugga	4	12	47	10.08		
ddoggprn	3	11.33	69	10.59	12	13.75
hussyfan	93	9.33	560	9.71	150	10.89
kdquality	10	9.9	204	9.22	12	11.33
kidzilla	50	9.14	113	9.07	21	9.52
kingpass	32	10.53	107	9.05	70	10.22
mafiasex	7	9	56	8.76		
pedo	663	9.20	2218	9.53		
pedofilia	28	10.39	201	11.14		
pedofilo						
pedoland	7	10	2	12		
pedophile	3	7.66				
pedophilia	1	13	3	5		
pedophilie						
pthc	1124	9.53	3868	9.84	21	12
ptsc	141	8.78	517	10.21	629	10.51
qqaazz	18	3.66	92	3.34	4	2.75
raygold	10	8.2	66	11.42	107	11.48
reelkiddymov	2	12	76	11.65		
yamad	2	6	52	10	27	6.85
youngvideomodels	8	10.87	46	9.54	104	11.05

Table 3: Co-occurrence of age indications with explicit paedophile keywords. For each word and each dataset, the corresponding two cells give the number of queries containing this word together with an age indication, and the average age indication. Empty cells correspond to keywords which never co-occur with age indications, even if their occur frequently in the dataset.

`6 yrs rar` is clearly not paedophile, and a query such as: `15yo model` may not be considered as paedophile without any other keyword.

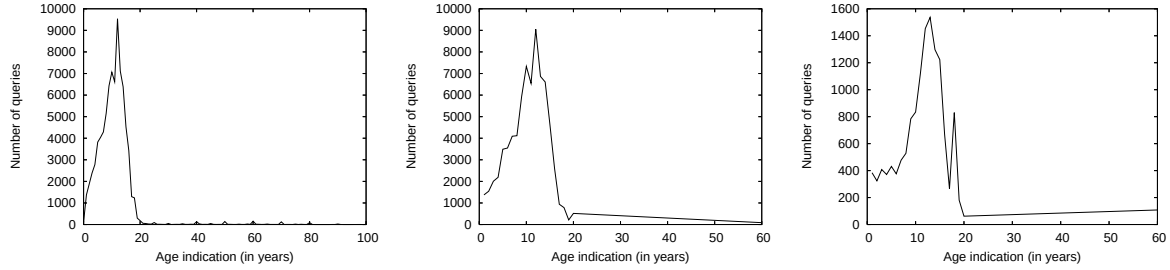

Figure 6: Age distribution in all queries and filenames. From left to right: *qu_2009*, *qu_2007* and *fid_2007* datasets.

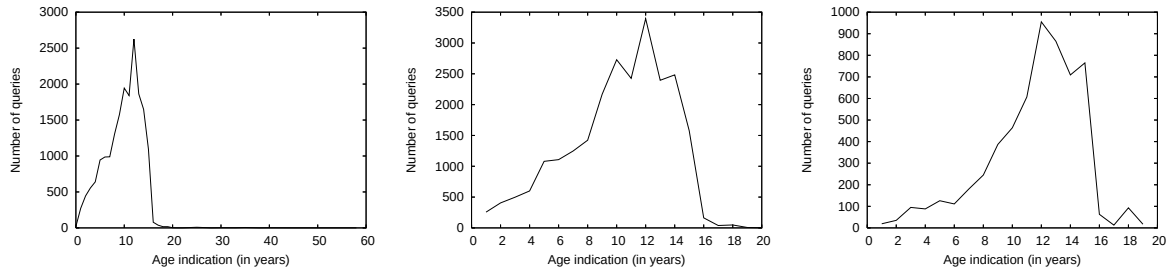

Figure 7: Age distribution in all paedophile queries and filenames. From left to right: *qu_2009*, *qu_2007* and *fid_2007* datasets.

Age distributions of queries and filenames which contain specific paedophile keywords are also interesting. We display in Figure 8 and 9 these distributions for the most interesting words.

For "babyj" (line 1 of Figure 8), while queries are centred on very young ages (around 5), filenames are either for very young ages (4 to 5 years) of for older ones (around 9). Such a difference can also be observed for "babyshivid" (line 2 of Figure 8) on *qu_2007* dataset.

Some keywords are centred around two different ages in both queries and filenames. This is for instance the case of "kingpass" (line 4 of Figure 8) which is often associated with 9 and 12 years in all datasets.

Finally, other keywords are mainly centred around 9 years ("hussyfan", "pthc", "ptsc", "youngvideomodels") with a large variation between 1 and 20 years.

3.2 Non-specific words appearing frequently with ages

In order to compare the distribution of ages co-occurring with paedophile keywords to the distributions of ages co-occurring with non-paedophile keywords, we performed the same experiments with the keywords which appear the most frequently with age indications. From the obtained list for each dataset, we removed all stop words (for instance "and"), all file extensions ("avi", "mpg", "jpg", etc.) and furthermore we removed explicit paedophile keywords of the above list (the only such word was "pthc"). We also

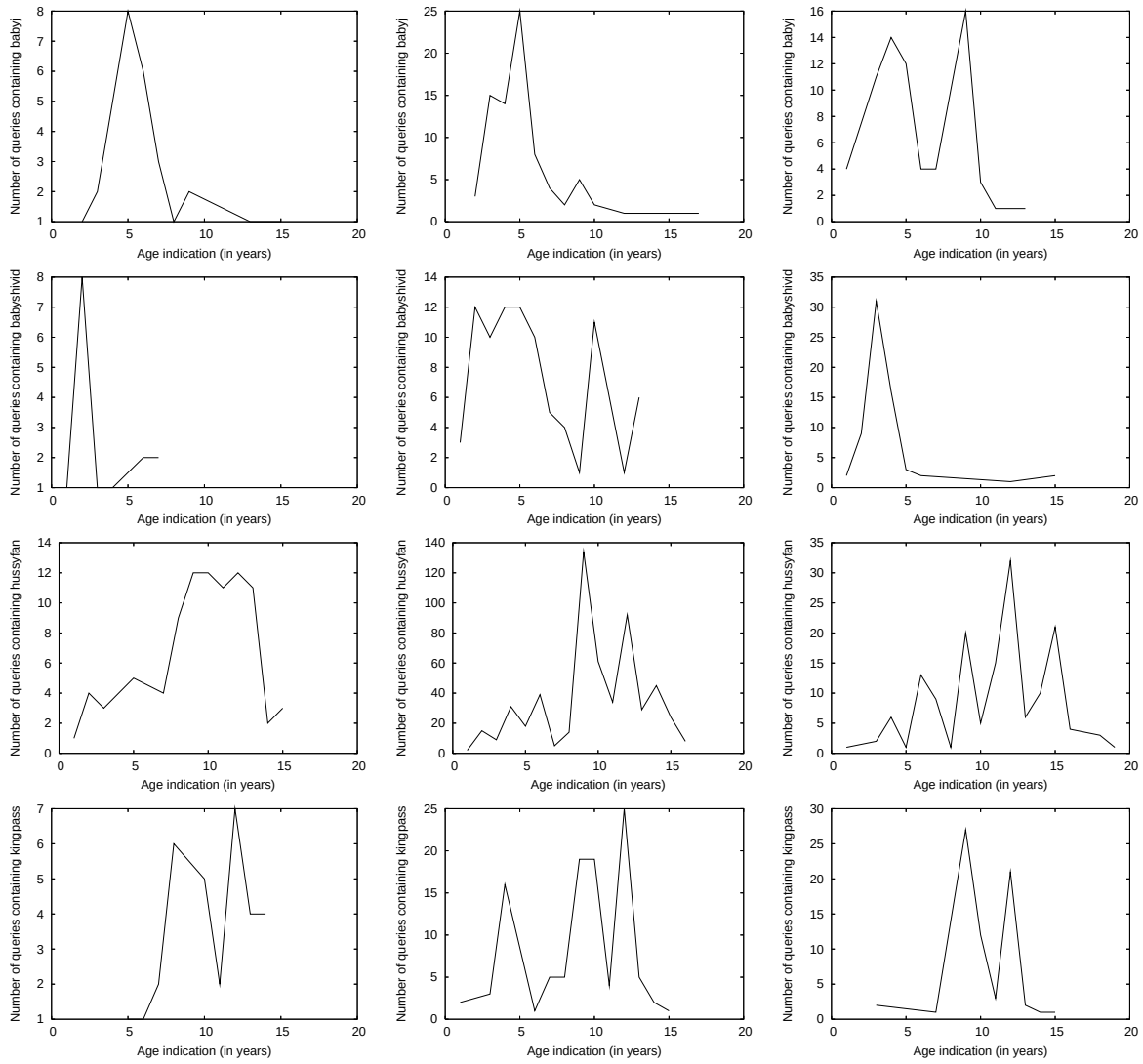

Figure 8: Distribution of ages co-occurring with babyj, babyshivid, hussyfan and kingpass. From left to right: *qu_2009*, *qu_2007* and *fid_2007* datasets.

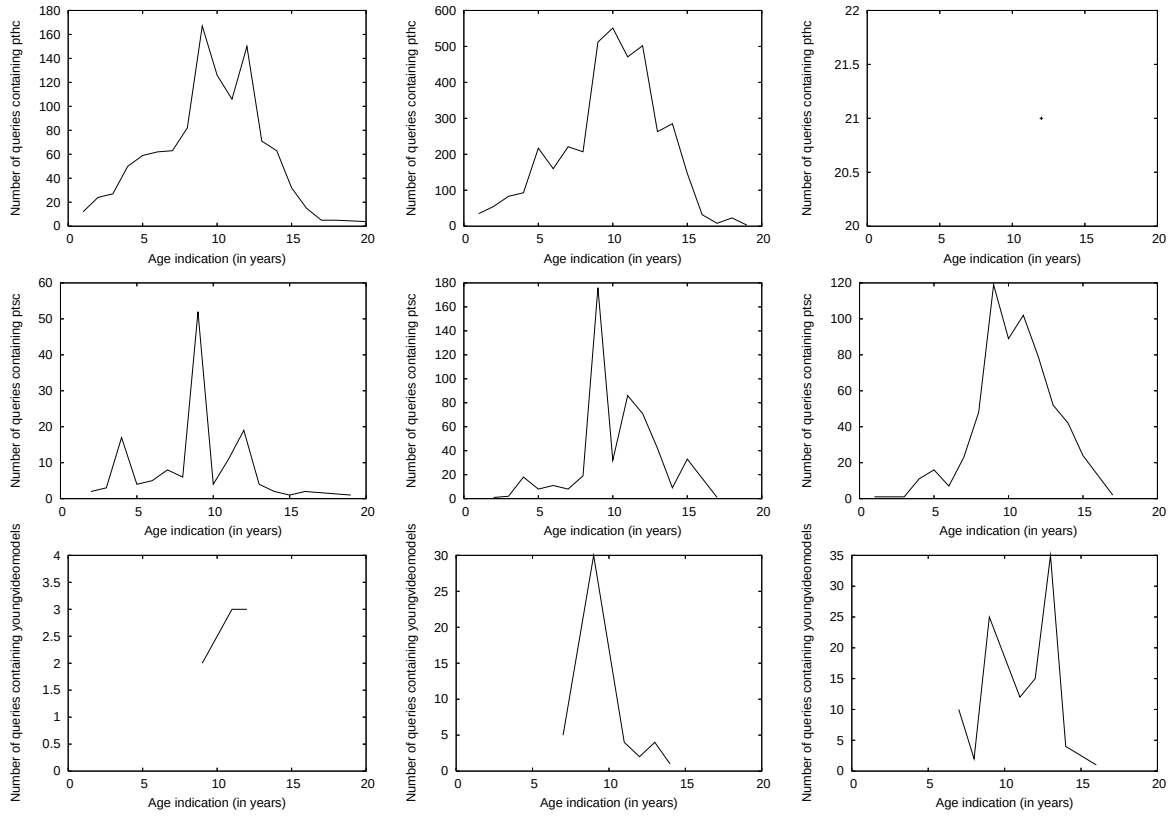

Figure 9: Distribution of ages co-occurring with pthc, ptsc and youngvideomodels. From left to right: *qu_2009*, *qu_2007* and *fid_2007* datasets.

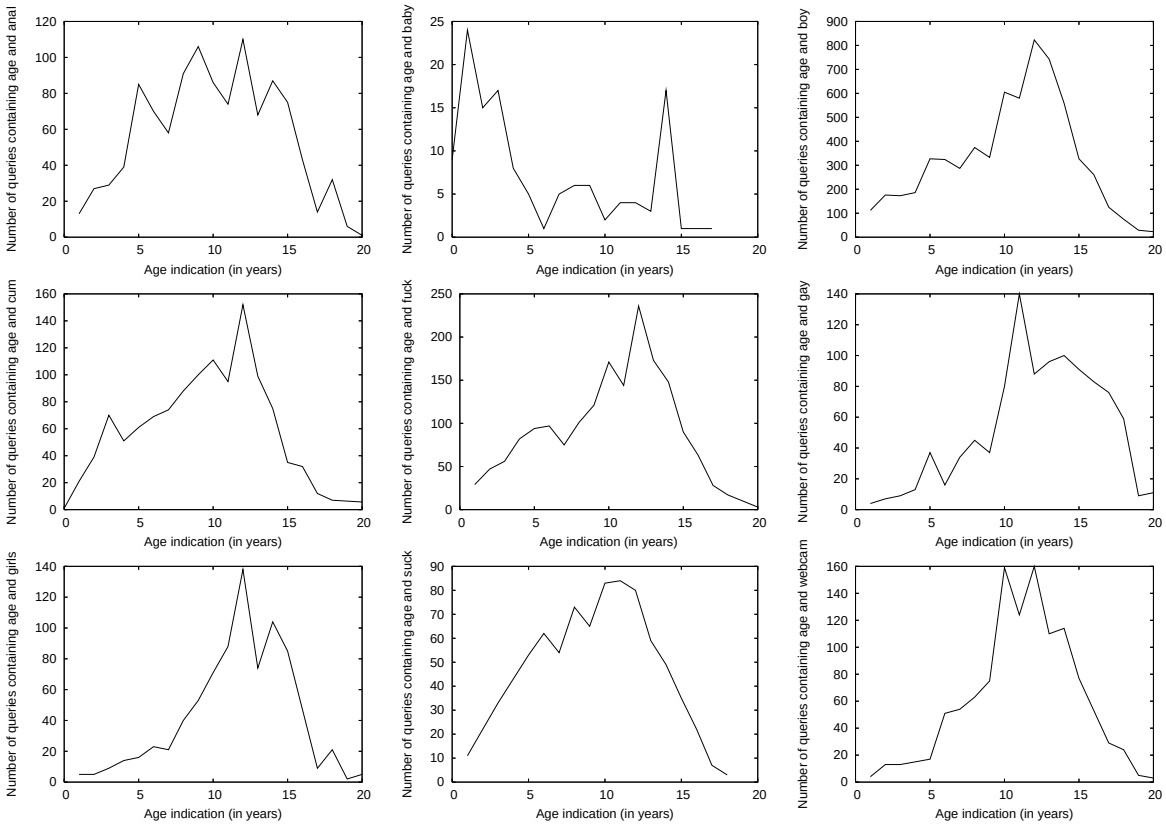

Figure 10: Distribution of ages (truncated at 20 years) which co-occur with anal, baby, boys, cum, fuck, gay, girls, suck, webcam in *qu_2009* dataset.

added the keyword "baby" in all lists. Figures 10, 11 and 12 present the lists of obtained keywords, together with the corresponding distributions, for the *qu_2009*, *qu_2007* and *fid_2007* datasets respectively.

The first thing to notice is that the obtained words clearly belong to the pornographic context, indicating that age is mostly used in pornographic queries or filenames (including paedophile ones). Then, we can observe that, since these words are not typical paedophile words, the average age associated with them is higher than the one observed for typical paedophile keywords. Note that, as expected, the "baby" keyword is generally related with younger ages than the other keywords, for all datasets.

Finally, notice that, though the obtained keywords are in general not specifically paedophile, the observed ages are clearly mostly below 18, indicating that the majority of queries or filenames containing an age are not only pornographic, but also are paedophile.

4 Specific users

Finally, we studied a few specific users who made a high number of paedophile queries. For these users, we extracted all queries containing age indications and we displayed the age-related queries using two different methods. Left plots in figures of this section always use non temporal information: we simply display the age for the 1-st query, then for the 2-nd one, and so on. Right plots on the contrary are time-related and indicate for each

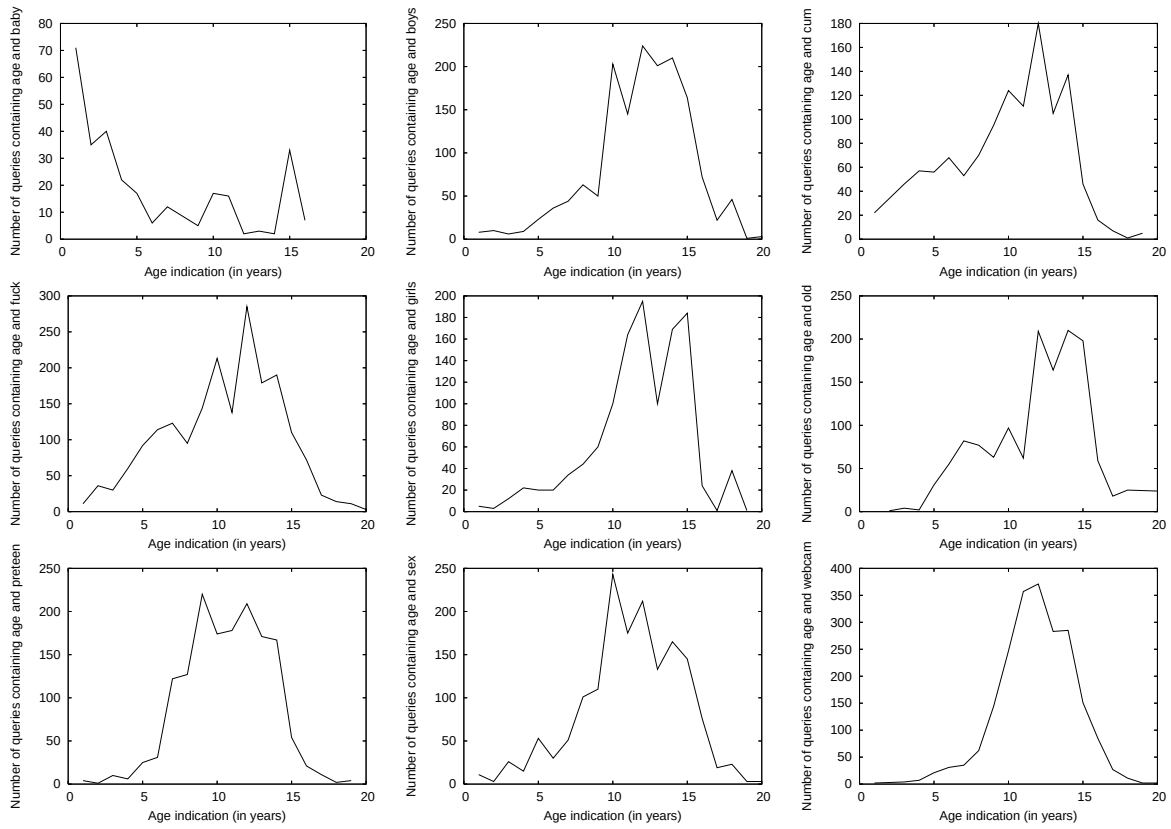

Figure 11: Distribution of ages (truncated at 20 years) which co-occur with baby, boys, cum, fuck, girls, old, preteen, sex, webcam in *qu_2007* dataset.

query the time at which it was received and the indicated age it contains.

The first user, *c72678062655c8a9d851701779aee85d* from the *qu_2009* dataset, had six sessions of age-related queries (on July 25, Aug. 8, Aug. 16, Sept. 26, Oct. 1, and Oct. 4), each lasting for at most 2 minutes. For instance, the queries performed on the 1-st of October are:

```
20:34:56 webcam 8yr
20:35:05 webcam 9yr
20:35:14 webcam 10yr
20:35:22 webcam 11yr
20:35:30 webcam 12yr
```

The session on the 16th of August is very similar, with a sequence of queries `xyo masturbate` for x from 6 to 13.

This is displayed in Figure 13. The left plot shows the sequence of ages that the user searched for. On the right plot, we can see vertically aligned dots, each vertical alignment corresponding to a short session.

This user seems very methodical, since during another session (not involving ages), he typed the query `pthc first-name`, for first-name in (sara, sarah, emma, alisa, alissa, aida, alica, alice, alison, amy, ann, ashlee, audrey, betty, bonny, bonnie, breeze, brenda, bristol, britney, britaney, brooke, carilyn, carla, carleigh, carly, catherine, carrie, charlee, chelsea, cherry, chrissy, chrissie, courtenay, courteney, courtney, daisy, daniella, danielle,

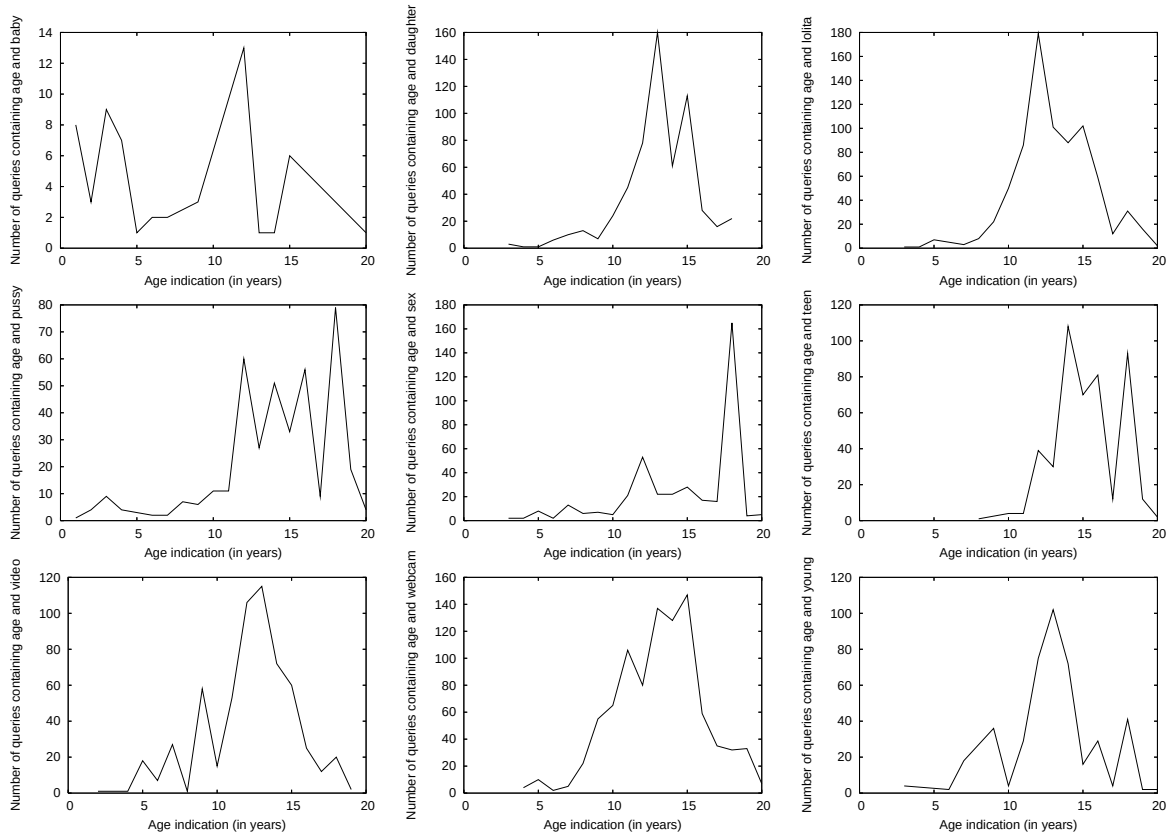

Figure 12: Distribution of ages (truncated at 20 years) which co-occur with baby, daughter, lolita, pussy, sex, teen, video, webcam, young in *fid_2007* dataset.

dawn, darla, edina, edith) in about 14 minutes.

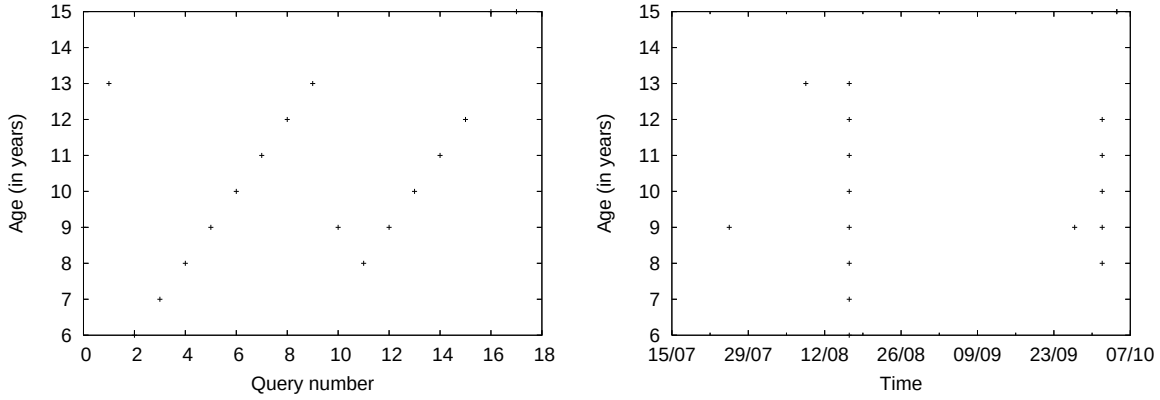

Figure 13: Age related queries of user c72678062655c8a9d851701779aee85d.

The second user we detail (see Figure 14) has a similar behaviour, showing sequences of queries with variation of ages and different sessions. However, this user starts by entering queries with ages above 10 during a few days and then starts a decreasing age sequence (from 9 to 2). Afterwards the user enters a few queries above 10 again and finally sticks to queries below 10. For this user also the sessions are very short, only a few minutes in general with gaps which can last for weeks between two sessions.

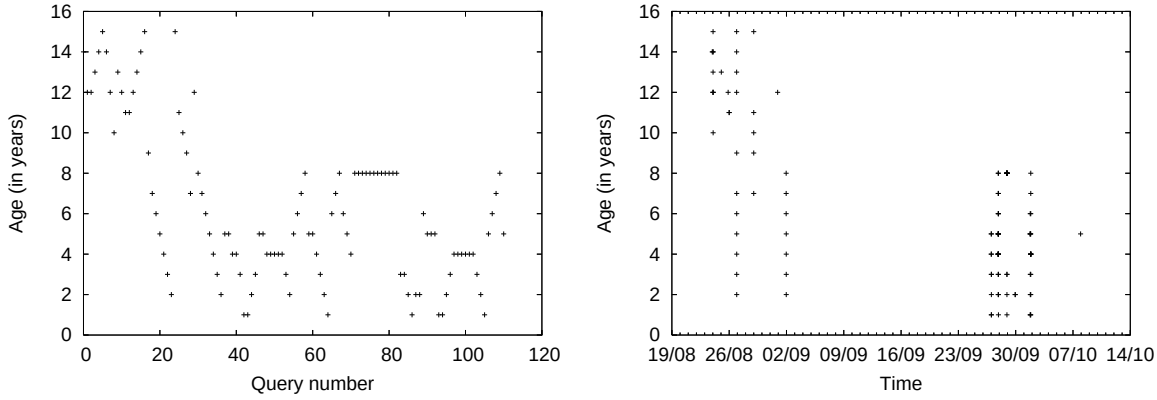

Figure 14: Age related queries of user 9db123d5fd2a8f0abd51582ceca776c8.

The third user we observe (see Figure 15) is more focused on specific topics and makes very frequent queries: half of the days of measurement.

More than one fourth of the queries of this user contain the word "daughter", half of which contain also the word "dad". In particular, this user has entered the query `incest dad wanks his 15yo daughter` 12 times from the 5th of August to the 22nd of September.

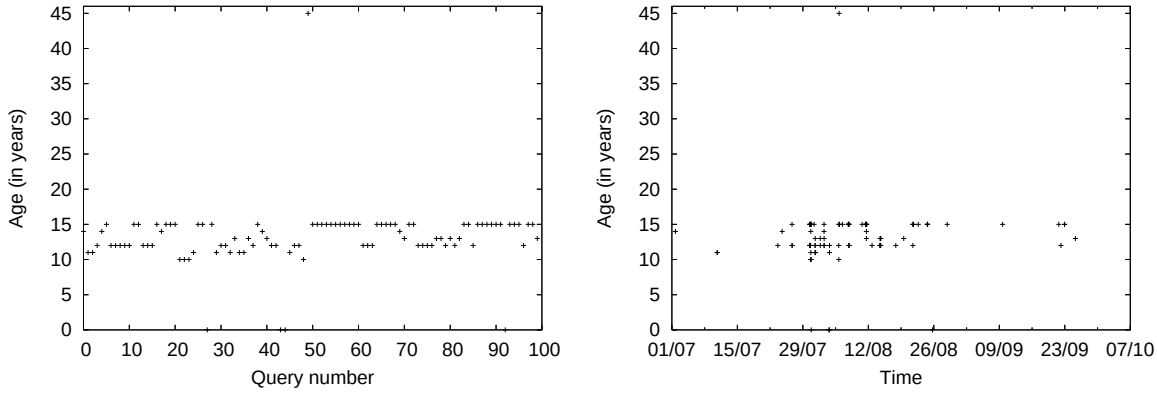

Figure 15: Age related queries of user 5d712719f1deea24f4c027c9ca73633a.

The fourth user we detail (see Figure 16) has a very specific behaviour since he only types 3 distinct queries: `13yo boy`, `14yo boy`, and `15yo boy`, respectively 32, 65 and 29 times.

Some users also search for explicit filenames lists that they probably found on specialised web sites or obtained in chatrooms. For instance, user in Figure 17 has typed in less than 7 minutes the following queries:

- `hussyfan pthc tante und -85371 inzest 1 mpg`
- `pthc composite 01 father and his 10yo twins -80703 mpg`
- `childlover pthc pedofilia part4 1 mpg`
- `pthc young 15yo bitch webcam strip fuck avi`
- `12y webcam new 10yo girl masturbation with sound pthc mpg`
- `video -433515 mylola pthc r ygold lsm -916 05 avi`

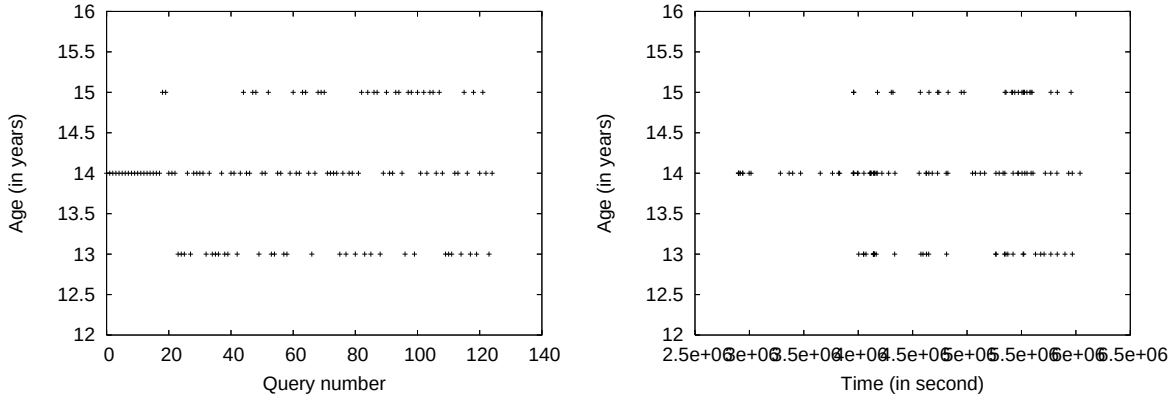

Figure 16: Age related queries of user 619795 6346.

In addition, 94% of his queries contain either "mpg", "jpg" or "avi".

As it can be seen in the figure, this user still has very short sessions.

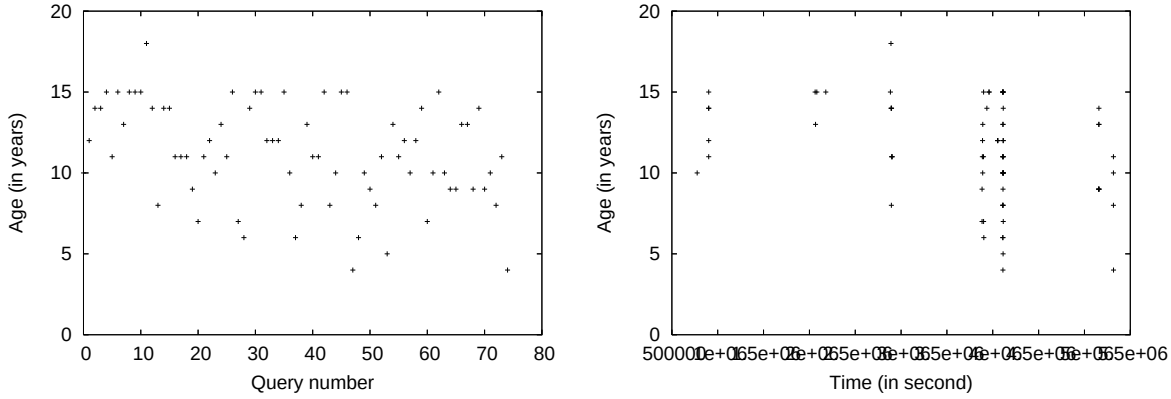

Figure 17: Age related queries of user 2898707 6346.

Finally, the last user we discuss (see Figure 18), who makes a lot of age-related queries, simply repeats over and over the same query `pthc new linda rides on dick 9yo fucks like pro`, 137 times, and 86 more times for the subquery `pthc new linda rides on dick`. This user is probably looking for a very specific content.

5 Conclusions and perspectives

The main conclusions of this study are threefold.

First, we found out that there is a strong heterogeneity between users. This is true regarding both the number of paedophile queries they enter and the proportion this represents among the total number of queries they enter. However, we found out that a large proportion of users interested in paedophile content have a strong focus on this type of content: on the *qu_2007*-IP+PORT dataset, for instance, more then 40% of paedophile users entered *only* paedophile queries.

Furthermore, users not only have different interest, but they also have different behaviours as it has been showed in Section 4. Some users mainly focus on specific ages,

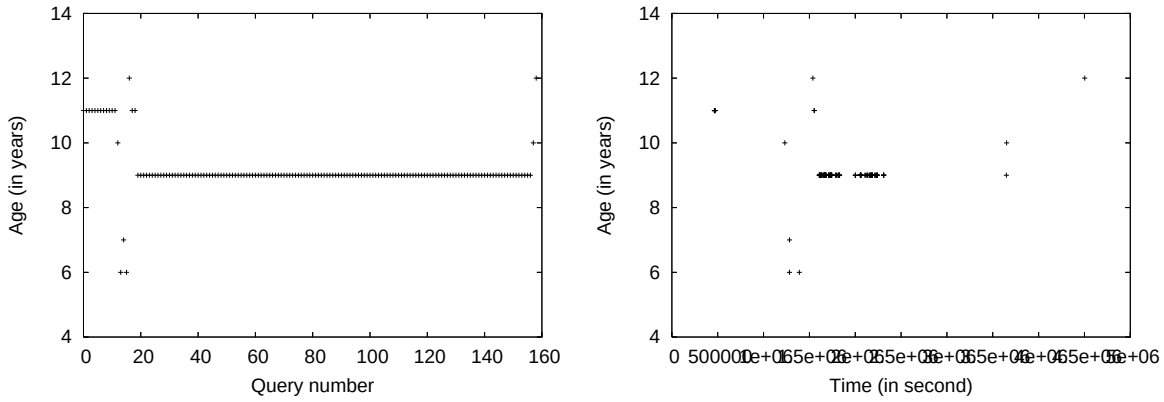

Figure 18: Age related queries of user 72846 6346.

while others are looking for much larger ranges, going from 1 to 16 years old. We also noted that some users are doing very short sessions of queries, lasting no more than a few minutes, and that others can spend more time searching. Finally, we also detected some users looking for very specific contents, repeatedly entering specific keywords or filenames.

Ages in queries and filenames also give deep insight on the behaviour of users. It must be first noted that queries generally indicate younger ages than filenames. This means that the demand is for younger ages than the offer, with a difference of about 1 year. We have also showed that depending on the type of queries or filenames, the encountered ages can vary a lot. Some keywords are generally associated with very young ages (for instance "baby" or "babyshivid"), while others are mostly related to ages between 9 and 12.

One key interest of this work also is that it shows that inspecting in depth series of queries entered by users provide much insight, and that the available data is relevant for doing so (as well as automatic paedophile query detection tools). It is therefore a first step towards a significant improvement of our knowledge of paedophile user behaviours and, as such, it opens many interesting perspectives.

One of the main perspectives relies in the inspection of other topics of interest of paedophile users, and how their interest in paedophile content evolves with time. Indeed, a key question is whether this interest evolves towards younger ages, and if some topics tend to conduct people to develop an interest in paedophile content. Inspecting long series of queries entered by paedophile users may help in answering these questions, in an unprecedented way.

Another direction is the search for user profiles, with a focus on paedophile users. Indeed, there are probably several kinds of behaviours, which may constitute distinct classes with their own interests. We made a first step in this direction by examining a few specific users, thus proving the relevance of this question. However, much remains to be done.

Finally, let us notice that we focused here on keyword-based queries, which seemed the most direct way to gain insight on user behaviours. However, depending on the data, we also have information on which users provide which files, which ones introduce new (paedophile) files, which files users choose to download, etc. Including these different

aspects in more detailed studies is very promising.

Acknowledgements. This work is supported in part by the MAPAP SIP-2006-PP-221003 and ANR MAPE projects.

References

[1] Frédéric Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an edonkey server. In *Proceedings of HotP2P'09*, 2009.

[2] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Technical report on *Quantification of Paedophile Activity in a Large P2P system*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[3] 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.

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