Technical report on

Dynamics of Paedophile Keywords in *eDonkey* Queries

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

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

Abstract

This technical report synthesizes the results of the analysis of paedophile keywords' dynamics in two sets of *eDonkey* queries, collected during several months in 2007 and 2009 respectively. The goal of this work is to study the evolution of paedophile keywords' frequency and popularity over several weeks (i.e. within a given dataset), as well as between the two different datasets. Moreover, specific periods are investigated in order to detect a potential periodicity over one week or over one day. Results show that the popularity of selected paedophile keywords is stable throughout both datasets where they can be found in 0,1% of the queries. No significant impact of the day of the week on the frequency of these paedophile keywords has been observed. Instead, the number of occurrences of paedophile keywords in *eDonkey* queries depends on the hour of the day. All these observations have been compared to those obtained with a selection of non paedophile keywords.

1. Context

1.1. Description of the datasets

Two datasets have been analyzed in order to study the dynamics of paedophile keywords in *eDonkey* queries. These datasets correspond to captures of queries received by *eDonkey* servers during several months in 2007 and in 2009 (denoted *2007_queries_capture* and *2009_queries_capture* respectively).

- *2007_queries_capture* consists in a 70 days' (10 weeks) capture. Each query received from the server is described by a timestamp (in seconds, starting at the beginning of the capture), the anonymized IP address of the client who made the query, its port number and finally the list of keywords used to formulate the query.
- *2009_queries_capture* consists in 102 days' capture (14 weeks and a half). The format of each query is similar to the one in the 2007 sample, without the port number. The format of the timestamp in this dataset is day/month-hour:min:sec. This dataset presents an advantage over the 2007 dataset, as days and months are explicitly indicated, which makes the analysis of specific days or weeks easier.

Presence or absence of port numbers had no impact for this study as only keywords and time stamps have been used. Indeed, the IP addresses and the port numbers have not been taken into account as

[1] Contact author: benedicte.le-grand@lip6.fr

the goal was to analyze the dynamics of paedophile keywords' occurrences in the queries, independently of the users who made these queries. Statistics about users (identified by the couple [IP address, port number]) can be found in [2].

Moreover, it should be noted that for privacy reasons, all IP addresses in the datasets have been anonymized, as well as keywords appearing less than 100 times (as unfrequent keywords may contain names, phone numbers and so on), as described in [1].

1.2. Analysis methodology

Both datasets have been analyzed in various ways in order to study the relative use of paedophile keywords in queries (with regard to common keywords). We also compared the frequencies of paedophile keywords and studied the evolution of their popularity over several weeks (i.e. within a capture) and over 3 years (i.e. between the 2 captures).

1.3. Selected 'paedophile' keywords

The list of 'paedophile' keywords used for this study has been consolidated throughout the project. The statistics and figures presented in this technical report are based upon the following list (composed of 26 keywords): "*babyj*", "*babyshivid*", "*childlover*", "*childporn*", "*childsex*", "*childfugga*", "*ddoggprn*", "*hussyfan*", "*kdquality*", "*kidzilla*", "*kingpass*", "*mafiasex*", "*pedo*", "*pedofilia*", "*pedofilo*", "*pedoland*", "*pedophile*", "*pedophilia*", "*pedophilie*", "*pthc*", "*ptsc*", "*qqaazz*", "*raygold*", "*reelkiddymov*", "*yamad*", and "*youngvideomodels*". These terms have been identified as the most typical of paedophile activity [3].

Our study of paedophile keywords' dynamics in *eDonkey* queries is organized as follows. In a first step, Section 2 discusses the frequency of paedophile keywords in the overall queries. Then the evolution of these keywords' popularity within a given dataset and between the 2 collected datasets is compared is Section 3. Finally a potential periodicity of paedophile keywords' occurrences over a week and over a day is investigated in Section 4.

2. Frequency of paedophile keywords in *eDonkey* queries

The total number of queries containing at least one paedophile keyword in the 2007 dataset is 141663, out of 127 316 861 queries, which represents 0.11% of the queries.

The total number of queries containing at least one paedophile keyword in the 2009 dataset is 117621, out of 106 344 062 queries, which also represents 0.11% of the queries.

The percentage of queries containing the selected paedophile keywords is therefore identical is both datasets. We conjecture that this percentage has probably remained stable from 2007 to 2009.

However, this value (0.11%) shows that the proportion of paedophile queries in the dataset (i.e. queries containing at least one of the selected paedophile keywords) is very small.

In order to illustrate the frequency of paedophile keywords in *eDonkey* queries, the number of occurrences of these keywords has been compared to the number of occurrence of selected non paedophile keywords (see Figure 1 for the 2007 dataset and Figure 2 for the 2009 dataset).

The chosen keywords belong to 3 categories:

- sex-oriented keywords (*sex*, *porn*),
- thematic keywords (*madonna*, *jackson*),
- keywords corresponding to popular file types extensions (*avi*, *divx*, *mp3*, *mpg*).

For each initial dataset, the subset of queries containing at least one of the following keywords has therefore been extracted: "*babyj*", "*babyshivid*", "*childlover*", "*childporn*", "*childsex*", "*childfugga*", "*ddoggprn*", "*hussyfan*", "*kdquality*", "*kidzilla*", "*kingpass*", "*mafiasex*", "*pedo*", "*pedofilia*", "*pedofilo*", "*pedoland*", "*pedophile*", "*pedophilia*", "*pedophilie*", "*pthc*", "*ptsc*", "*qqaazz*", "*raygold*", "*reelkiddymov*", "*yamad*", "*youngvideomodels*", "*sex*", "*porn*", "*madonna*", "*jackson*", "*avi*", "*divx*", "*mp3*", "*mpg*".

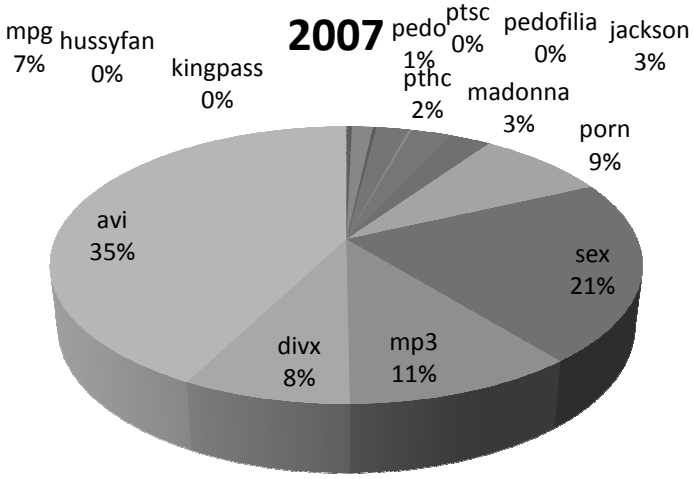

Figure 1. Frequency of paedophile keywords with regard to other types of keywords in the 2007 dataset

Figures 1 and 2 show that in both extracted subsets, the *relative* frequency of paedophile keywords (i.e. with regard to the number of queries of the subsets) is very low compared to most of these keywords (even *pthc*, which is the most frequent paedophile keyword as explained below, represents only 2% in both datasets).

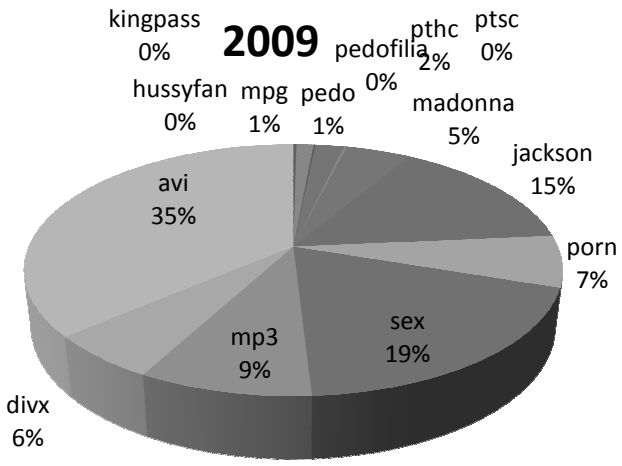

Figure 2. Frequency of paedophile keywords with regard to other types of keywords in the 2009 dataset

The ranking of paedophile keywords in *eDonkey* queries is provided in Table 1 (based on the ratio of the number of occurrences of each paedophile keyword divided by the total number of occurrences of paedophile keywords). The order of keywords in the table is based on 2009's ranking (3rd column of the table).

	2007	2009
pthc	**42,2%**	**48,5%**
pedo	**27,4%**	**28,6%**
ptsc	**4,0%**	**4,0%**
hussyfan	*5,1%*	*3,6%*
pedofilia	*4,6%*	*3,0%*
kingpass	*2,2%*	*1,4%*
babyshivid	1,3%	1,2%
raygold	1,3%	1,1%
mafiasex	1,6%	0,9%
yamad	1,1%	0,9%
babyj	1,3%	0,9%
childporn	0,5%	0,9%
childlover	1,4%	0,8%
youngvideomodels	1,6%	0,7%
qqaazz	1,0%	0,6%
pedoland	0,5%	0,6%
pedofilo	0,3%	0,6%
kidzilla	0,7%	0,4%
pedophile	0,3%	0,3%
pedophilia	0,3%	0,2%
pedophilie	0,0%	0,1%
childfugga	0,3%	0,1%
kdquality	0,6%	0,1%
childsex	0,0%	0,1%
reelkiddymov	0,3%	0,1%
ddoggprn	0,3%	0,1%

Table 1. Ranking of paedophile keywords in *eDonkey* queries

We observe 3 groups of paedophile keywords:

- *pthc* and *pedo*, which represent more than 25% of the occurrences of paedophile keywords (in both datasets),
- *ptsc*, *hussyfan* and *pedofilia* which have frequencies comprised between 3 and 5.1% (in both datasets),
- remaining keywords, with frequencies lower than 2% (in both datasets).

Figure 3 shows that keywords *pthc* and *pedo* are proportionally more used (with regard to the other paedophile keywords) in 2009 than in 2007.

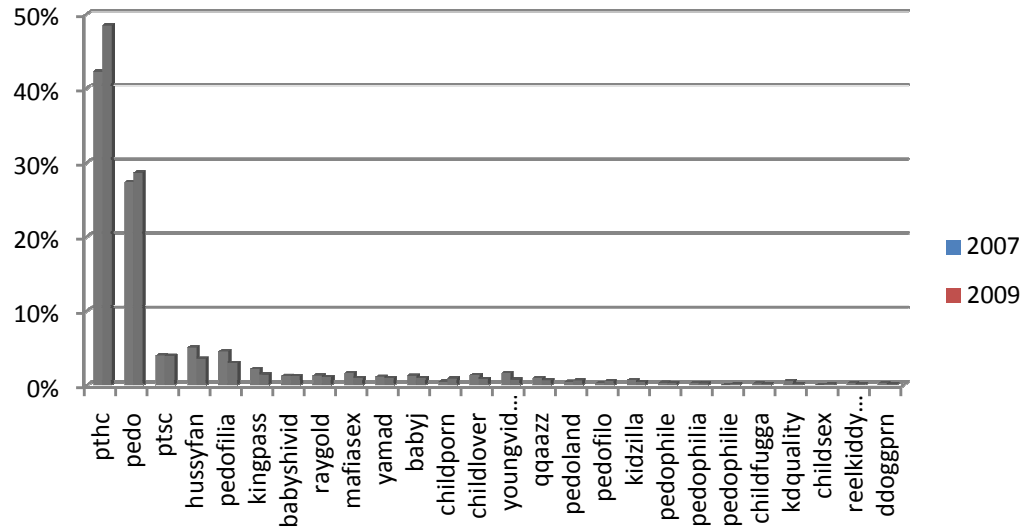

Figure 3. Ranking of paedophile keywords in 2007 and 2009 queries

Conversely, the relative proportion of most other paedophile keywords has decreased from 2007 to 2009 (see Figure 4).

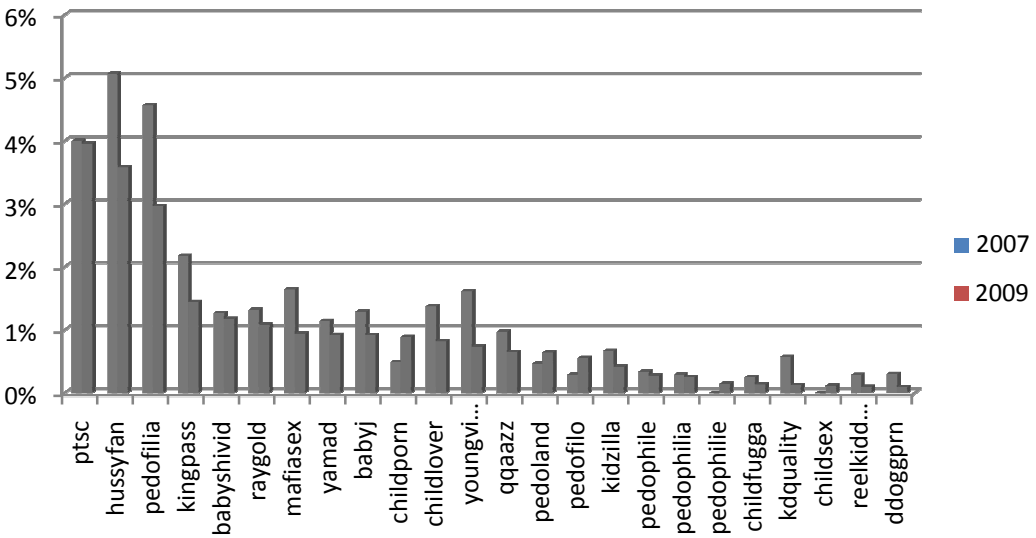

Figure 4. Ranking of less frequent paedophile keywords in 2007 and 2009's queries

3. Dynamics of paedophile keywords' popularity (over the whole duration of the captures)

This section first studies the evolution of paedophile keywords' popularity in *eDonkey* queries (i.e. their number of occurrences) over the whole duration of 2007 and 2009 datasets (in Section 3.1). These results are then compared in Section 3.2 with those obtained for non paedophile keywords.

3.1. Evolution of paedophile keywords' popularity during the whole captures

Figure 5 represents the evolution of paedophile keywords' occurrences in *eDonkey* queries over the 70 days of the 2007 dataset capture. This figure may therefore be interpreted as the evolution of the "popularity" of these keywords over several weeks. Similarly, Figure 6 represents the evolution of paedophile keywords' occurrences in *eDonkey* queries over the 102 days of the 2009 dataset's capture. The time unit for both figures (i.e. the x axis) is a day.

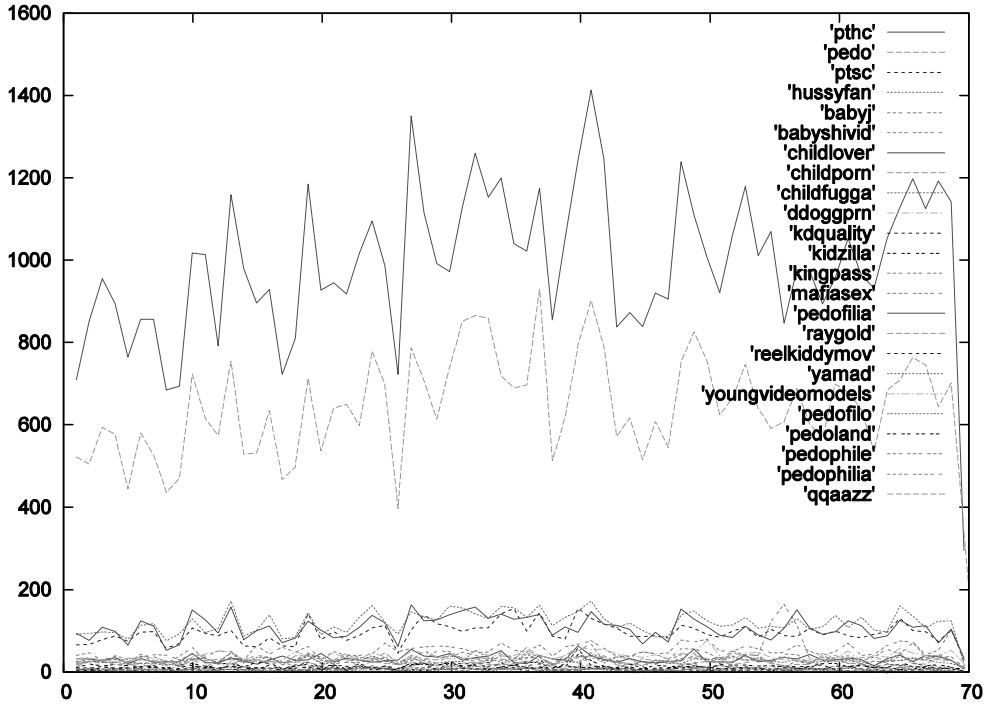

Figure 5. Evolution of paedophile keywords' occurrences during the whole 2007 dataset

The first observation confirms that *pthc* and *pedo* keywords are significantly more frequently used in *eDonkey* queries than the 24 other paedophile keywords of the list provided in Section 1.3. In the following, we will therefore focus on these two most significant keywords.

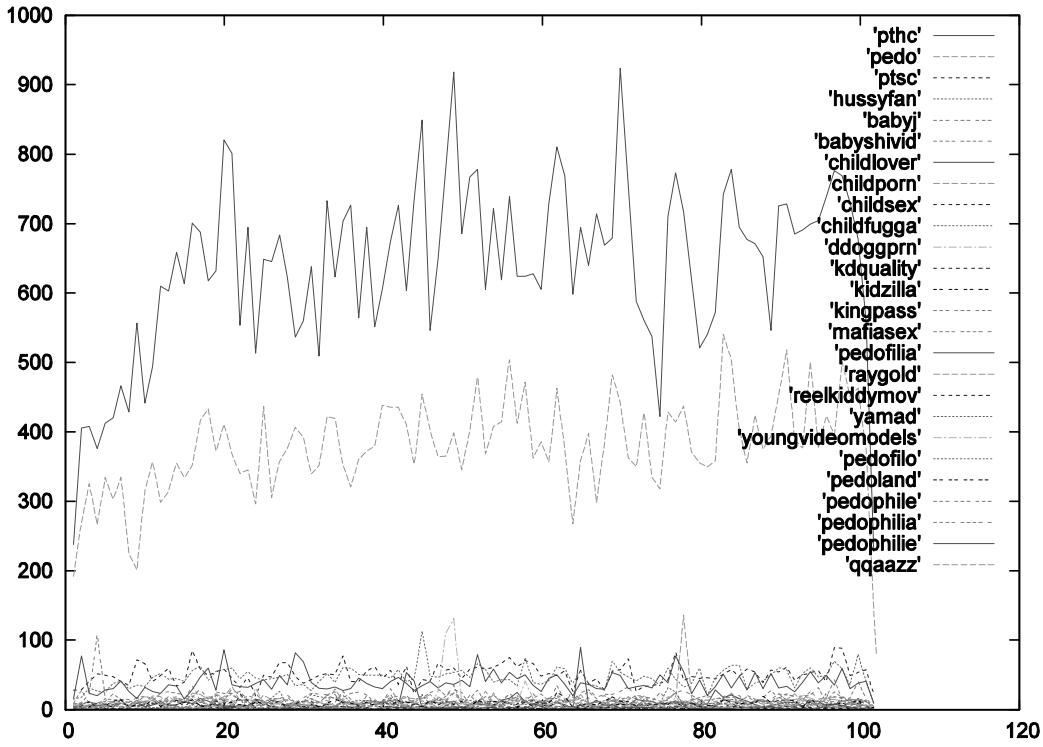

Figure 6. Evolution of paedophile keywords' occurrences during the whole 2009 dataset

Figure 5 shows that in the 2007 dataset the number of paedophile queries containing *pthc* varies between 700 and 1400 per day, and between 400 and 900 per day for *pedo*. It should be noted that the two corresponding curves seem "synchronized" as they increase and decrease at the same time.

This is less true for the 2009 dataset: in Figure 6, the *pthc* and *pedo* keywords' occurrences globally evolve at the same time, but some peaks may be observed only on one of both. For instance, around day 20, the peak of *pthc* does not correspond to a peak for *pedo*. The popularity of these two keywords thus seems to be more independent in 2009 than it was in 2007.

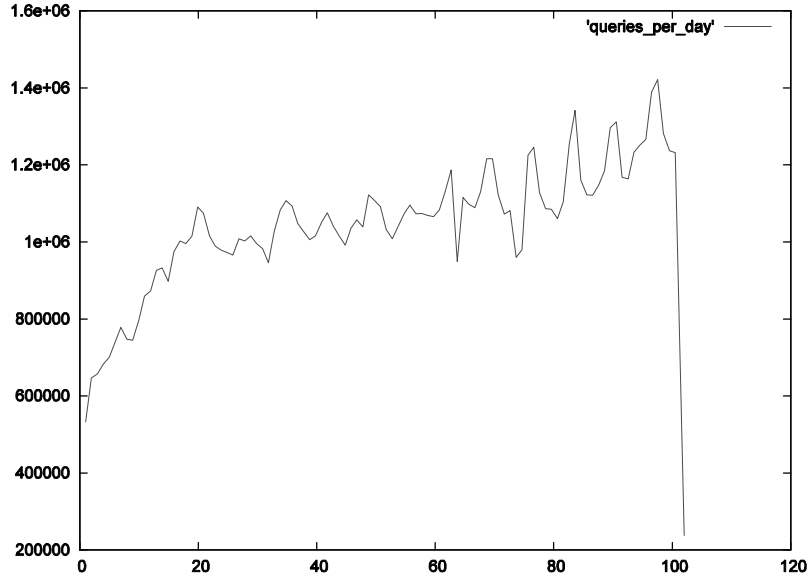

Figure 7. Evolution of the total number of queries during the whole 2009 dataset

Although the number of queries containing paedophile keywords is quite stable from days 20 to 102, it starts with a low value and keeps increasing from days 0 to 20. In order to know whether this phenomenon is specific to paedophile keywords' occurrences, the evolution of the total number of queries in the 2009 dataset has been computed. Figure 7 shows that the overall number of queries also grows during the 20 first days; this is probably due to an increasing activity of the *eDonkey* server, as this is the case after a reboot.

Moreover, Figure 6 shows that in the 2009 dataset the number of paedophile queries containing *pthc* varies between 250 and 900 per day, and between 200 and 550 per day for *pedo*. These values are lower than those of 2007. Indeed, Figure 8 confirms that *pthc* always appear in more queries (everyday) in 2007 than in 2009; this is also true for *pedo*. This could seem surprising as the proportion of paedophile keywords is identical in the both datasets (as seen previously).

However, the 2009 dataset contains 106 315 335 queries corresponding to 102 days, which represents an average of 1 042 207 queries per day. On the other hand, the 2007 dataset contains 127 316 861 queries corresponding to 70 days, which represents an average of 1 818 812 queries per day, i.e. 1.74 times more than in 2009. This explains why the number of occurrences of paedophile keywords is higher in 2007 (even if their proportion within the overall queries is identical in 2009).

This difference is due to the popularity of the *eDonkey* servers on which queries were captured: the server used for the captures in 2009 is less popular (i.e. it receives fewer queries per day) than the one used in 2007.

It should be noted that the decrease in the number of keywords' occurrences at the very end of the datasets (both in 2007 and 2009) is due to the fact that the captures did not finished at the end of a day (for instance, the last query in the 2009 dataset was done at 9.45 am).

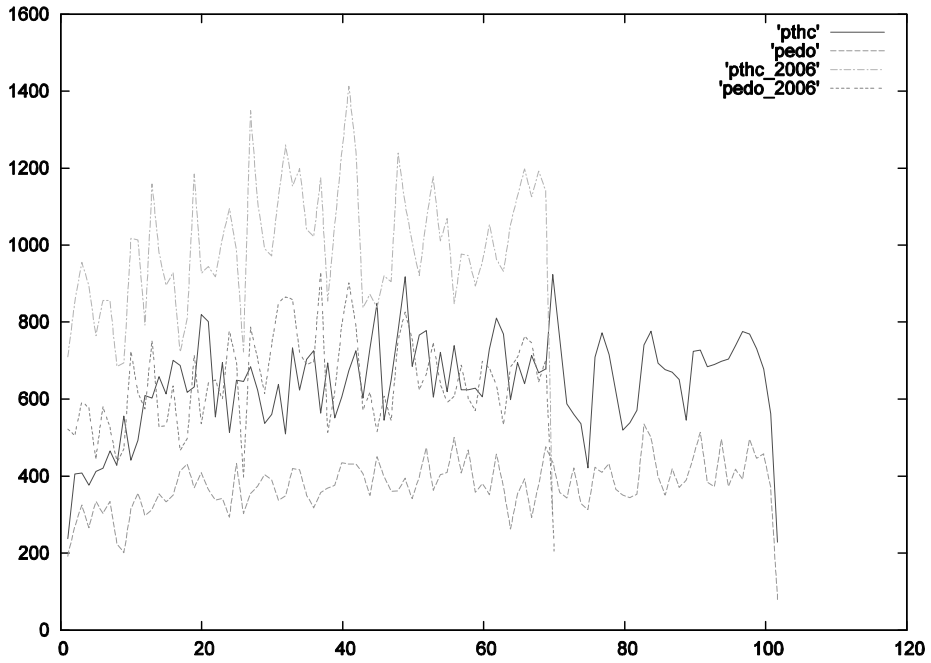

Figure 8. Comparison of *pthc* and *pedo* keywords' popularity between 2007 and 2009

Finally, although the popularity of paedophile keywords varies over the weeks of the captures, the average value remains quite stable; no significant long-term increase or decrease of popularity is observed.

3.2. Comparison with the evolution of non paedophile keywords' popularity during the whole captures

In this section, popularity of a selection of non paedophile keywords is studied in order to compare it to the results of previous section.

Figure 9 shows that in the 2007 dataset, there is no significant global increase or decrease of these keywords' popularity, whatever class they belong to (general theme, sex-oriented or file extension).

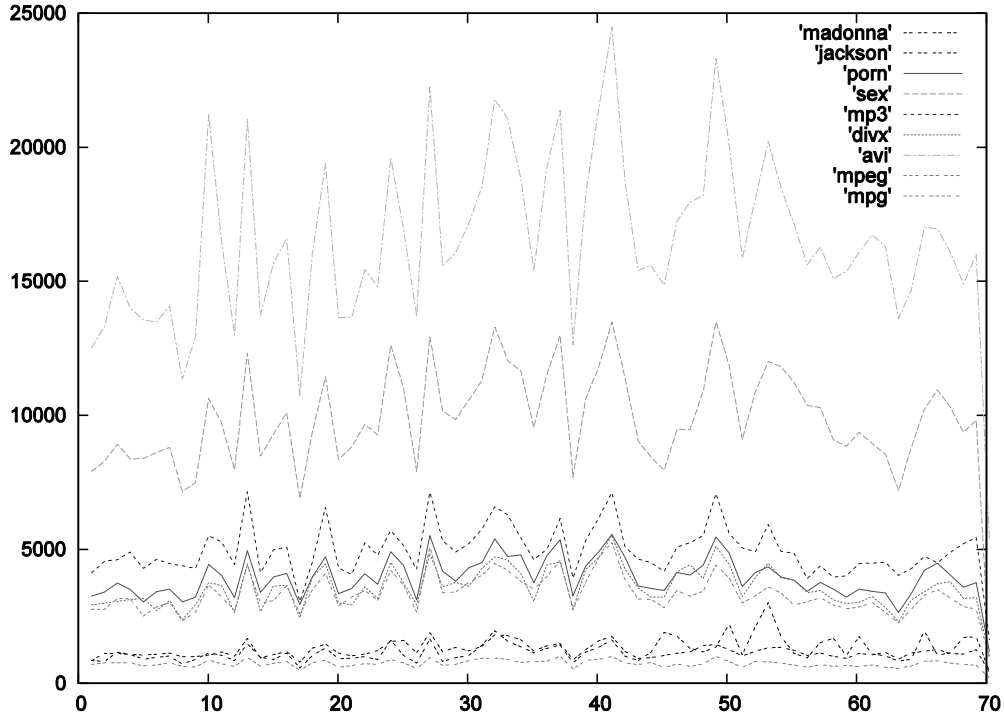

Figure 9. Evolution of non paedophile keywords' occurrences during the whole 2007 dataset

Figure 10 represents the evolution of *madonna* keyword, together with the two most popular paedophile keywords, *pthc* and *pedo* (in the 2007 dataset). The synchronization between the *pthc* and *pedo* plots had already been observed on Figure 6. The evolution of *madonna*'s popularity in *eDonkey* queries is somehow synchronized, but not as much as both paedophile plots (in particular in the second half of the capture).

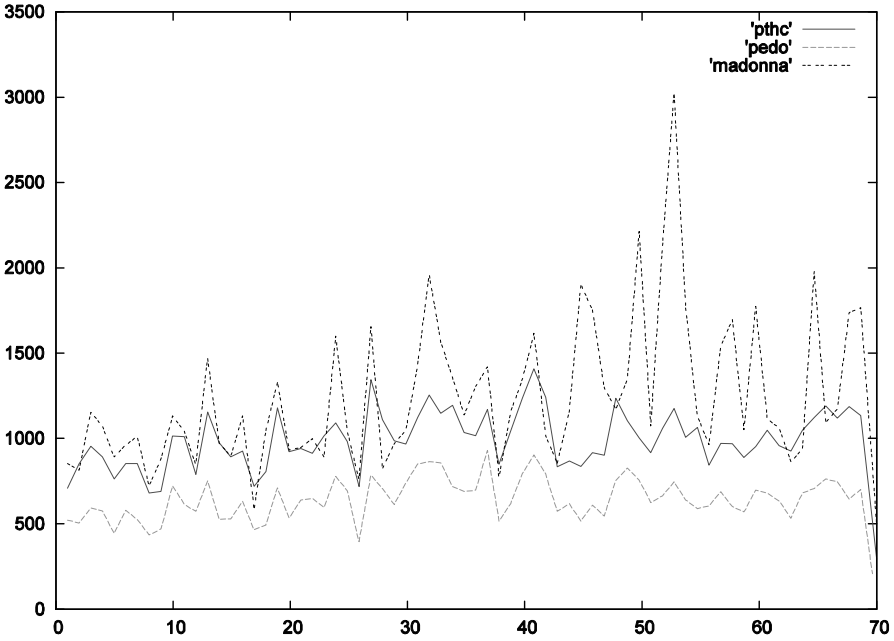

Figure 10. Evolution of the number of *madonna*, *mpg*, *pedo* and *pthc* keywords' occurrences during the whole 2007 dataset

This tends to show that the evolution of paedophile keywords' popularity over the whole 2007 dataset is quite close to the evolution of the selected non paedophile keywords' popularity. However a greater synchronization is observed among paedophile plots.

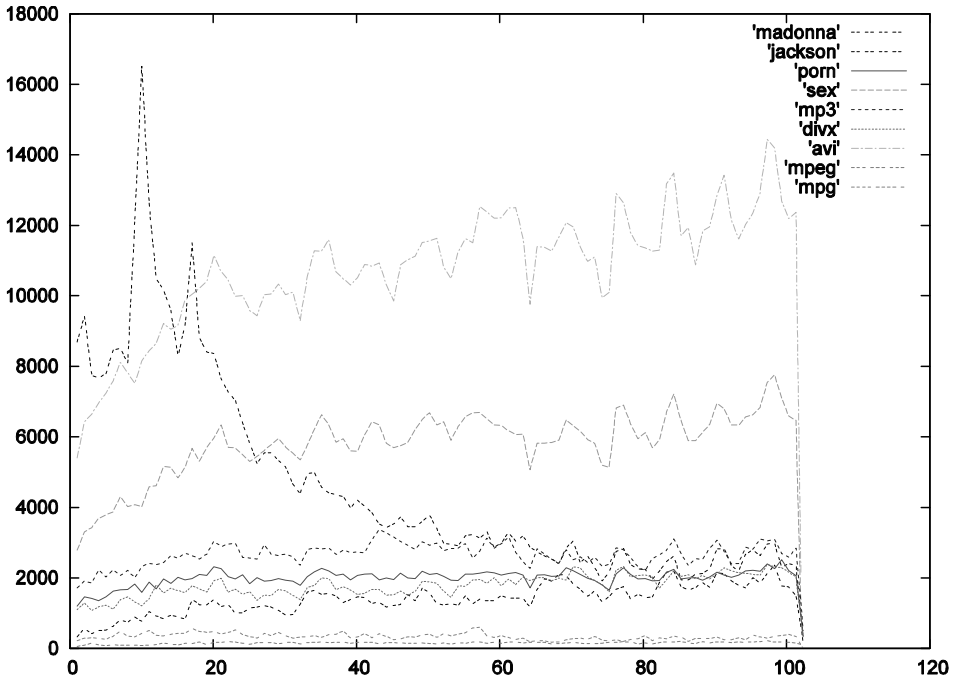

Figure 11. Evolution of the number of non paedophile keywords' occurrences during the whole 2009 dataset

However, the popularity of non paedophile keywords is not always stable as it appeared to be in the 2007 dataset. The evolution of the *jackson* keyword in the 2009 dataset (see Figure 11) is an example of keyword which becomes extremely popular during a few weeks and whose popularity rapidly decreases afterwards. Indeed, the 2009 capture started on June 29th, 2009, i.e. 4 days after Mickael Jackson's death. Many queries containing his name have therefore been observed the following month, before converging towards an average stable (lower) value later on.

On the other hand, the number of occurrences of *madonna* regularly increase during the 2009 dataset (see Figure 12), which is not true for the *mp3* keyword; this keyword has thus become more and more popular (and regularly) during the 102 days' capture. Figure 12 shows that this is not the case for paedophile keywords, as their popularity remains stable over the dataset. This therefore confirms the results of Section 3.1: no significant long-term evolution of paedophile keywords' popularity in *eDonkey* queries has been observed in the 2007 and the 2009 captures.

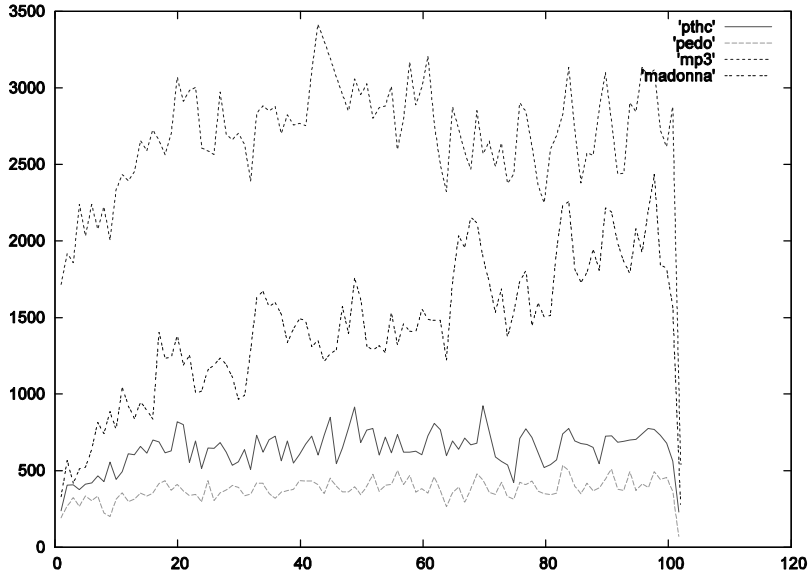

Figure 12. Evolution of the number of *madonna*, *mpg*, *pedo* and *pthc* keywords' occurrences during the whole 2009 dataset

4. Periodicity of paedophile keywords' dynamics in *eDonkey* queries

This section aims at studying specific periods of time (in particular weeks and days) in order to detect potential periodicity in paedophile keywords' dynamics in *eDonkey* queries. Section 4.1 deals with specific weeks whereas Section 4.2 focuses on given days. These results are finally compared to those of non paedophile keywords in Section 4.3.

4.1. Periodicity over one week

As explained above, specific weeks have been studied in order to see if the use of paedophile keywords varied according to days of week. Only the 2009 dataset has been used for this study as the days and months are known (which is not the case in the 2007 dataset).

Figure 13 shows the evolution of *pedo*, *pthc* and *hussyfan* keywords over the days of the week in the 2009 dataset. The values on the y axis represented the cumulated number of queries which took place on each Monday, Tuesday, etc.

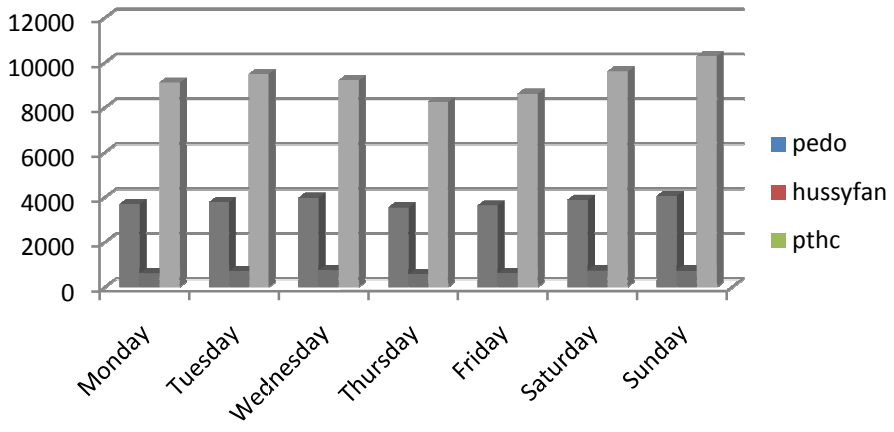

Figure 13. Evolution of pedo, pthc and hussyfan keywords' occurrences over the various days of the week (2009 dataset)

These results show that the day of week has limited impact on the number of occurrences of paedophile keywords; a slightly higher activity may however be observed during the weekends. This is confirmed by Figure 5 and Figure 6 which do not show any clear weekly periodicity (over several months).

4.2. Periodicity over one day

Previous section studied the dynamics of paedophile keywords according to days of week. In this section, the goal is to study the evolution over one day and the time unit is the hour.

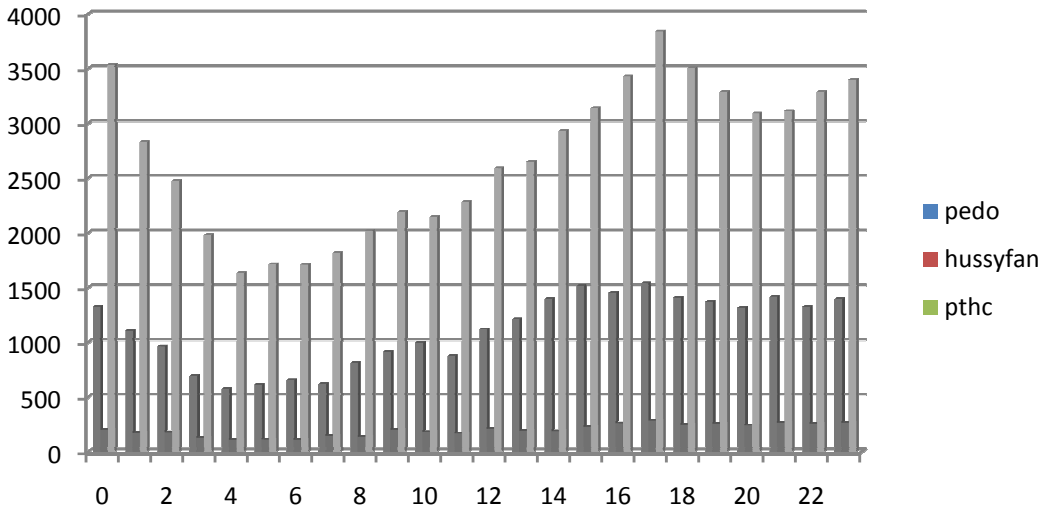

Figure 14. Evolution of *pedo*, *hussyfan* and *pthc* keywords' occurrences over 24 hours

Figure 14 shows the cumulated number of queries containing the keywords *pedo*, *hussyfan* and *pthc* for each hour of the day over the 2009 dataset. This clearly shows an impact of the hour on paedophile queries: the activity decreases regularly from midnight to 4 am and then increases until 5 pm. A new decrease may be observed between 5 pm and 8 pm and the activity grows again afterwards from 9 pm to midnight.

4.3. Comparison with periodicity of non paedophile keywords

Same statistics as those of the two previous sections have been computed for selected non paedophile keywords.

4.3.1. Selection of specific weeks

The cumulated number of queries involving four non paedophile keywords (divx, Madonna, mpg and mp3) has been computed for each day of the week in order to detect a potential weekly periodicity; these results are presented on Figure 15. This is the equivalent of what has been done for paedophile keywords in Section 4.1.

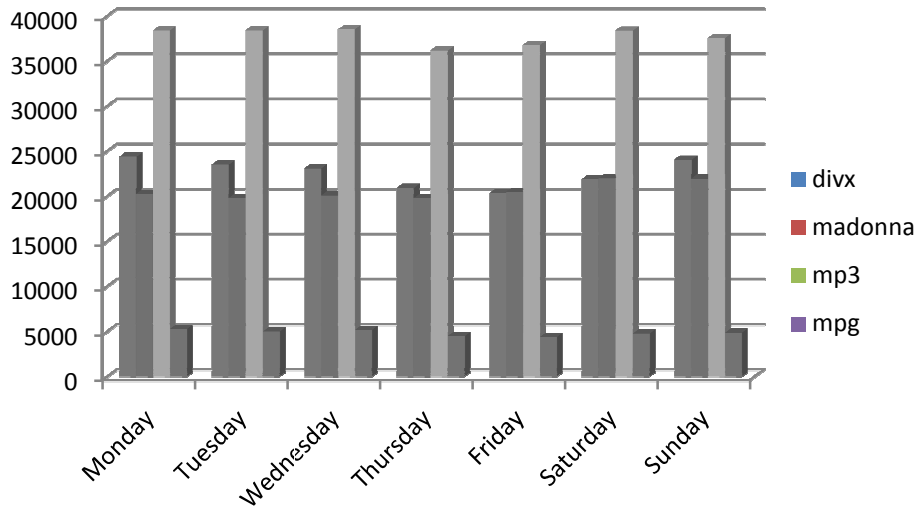

Figure 15. Evolution of 4 non paedophile keywords' occurrences over the days of the week

Although the number of occurrences of non paedophile keywords in *eDonkey* queries slightly decreases on Thursdays and Fridays for most keywords, Figure 15 does not show any significant impact of the day of the week on non paedophile queries, whatever their type is (general theme or file extension). This result is confirmed by Figure 16, which shows the evolution of non paedophile keywords over one week in 2007. Seven days can be identified, with no specific peak for any of those

in particular (except for the plot related to *madonna* keyword). This is confirmed by Figure 10 and Figure 11 which do not show any weekly period (over several months).

We also see on Figure 16 that non paedophile keywords' popularity seems periodic over one day, as it will be shown in the following section.

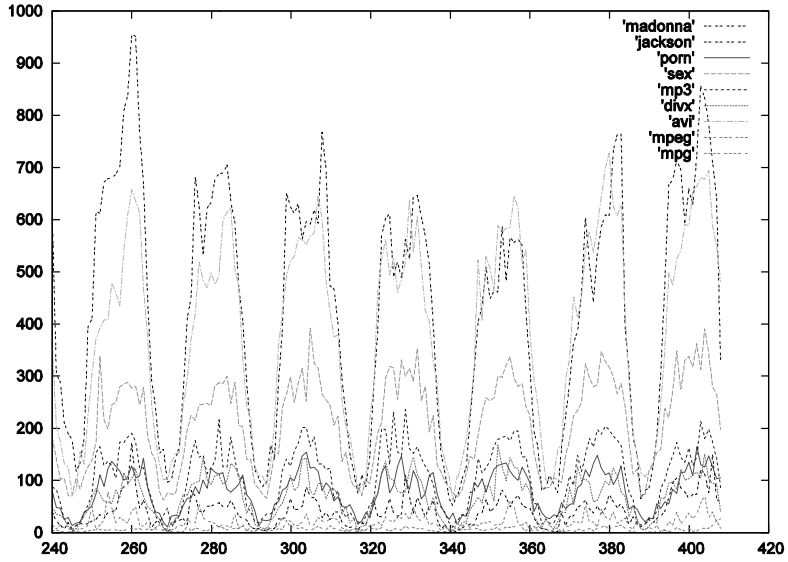

Figure 16. Evolution of non paedophile keywords during one week in 2009

4.3.2. Selection of specific days of the week

Figure 17 shows the cumulated number of queries containing keywords *divx*, *Madonna, mp3* and *mpg* for each hour of day over the 2009 dataset. This is the equivalent of what has been done for paedophile keywords in Section 4.2.

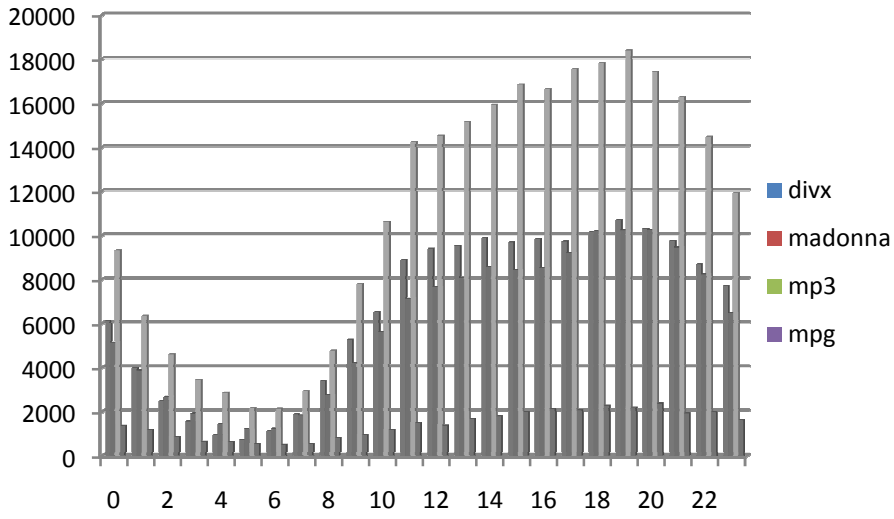

Figure 17. Evolution of non paedophile keywords' occurrences over the hours of the day

Figure 17 shows an impact of the hour on paedophile queries as they increase very rapidly from 6 am to 11 am; then they keep growing at a slower pace until 7 pm (however with a slight decrease around 5 pm). Finally they decrease regularly from 7 pm to 5 am.

In order to deepen the analysis, specific days have been studied for non paedophile keywords in the 2009 dataset: Monday, July 13th (see Figures 18) and Saturday, July 18th 2009 (see Figure 19).

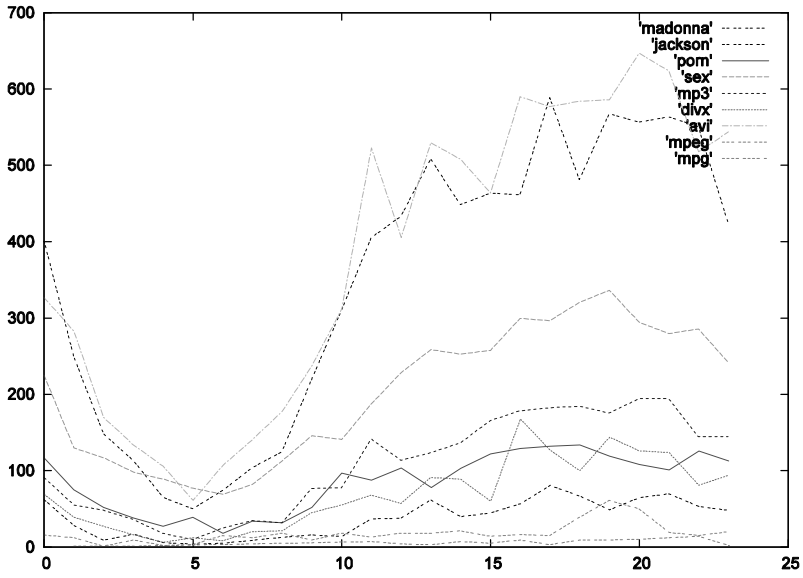

Figure 18. Evolution of non paedophile keywords' occurrences on Monday, July 13th 2009

Figures 17 and 18 confirm that the frequency of non paedophile keywords' in *eDonkey* queries is periodic over one day.

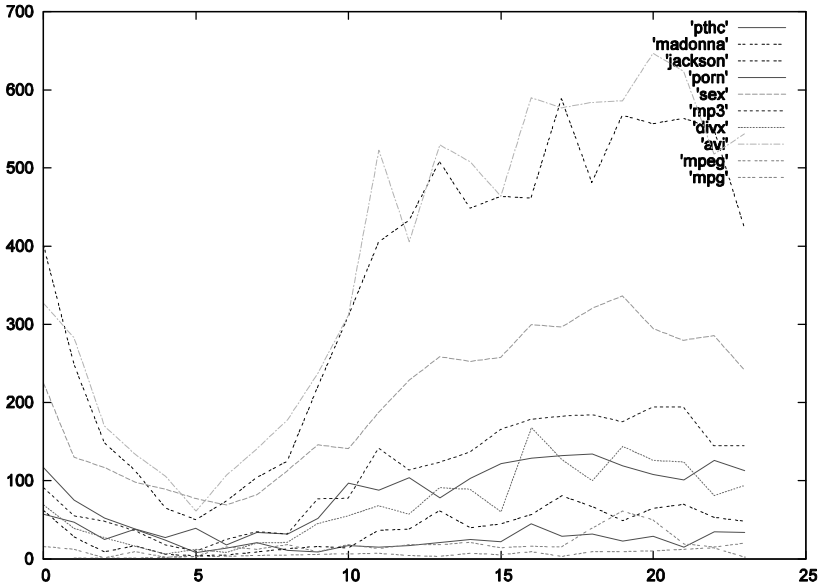

Figure 19. Evolution of non paedophile (+ pthc) keywords' occurrences on Saturday, July 18th 2009

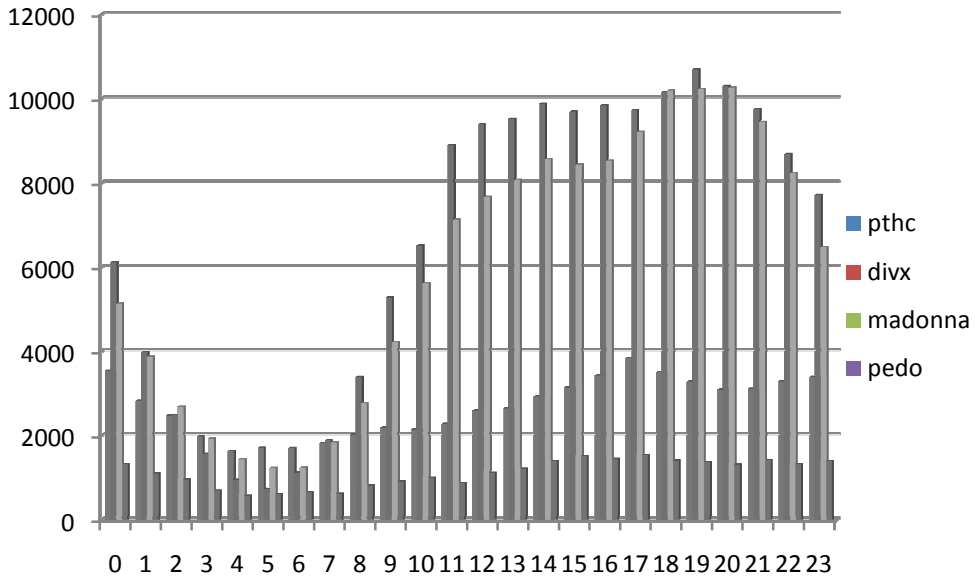

Figure 20. Comparison of *pthc, pedo, divx* and *madonna*'s popularity evolution over the hours of the day

We have seen in Section 4.2 that queries containing paedophile keywords also depended on the hour of the day. Figure 20 compares the evolutions of queries containing *pthc, pedo, divx* and *madonna* keywords. Although all plots are periodic over one day as already seen in previous sections, several differences may be observed:

- although both paedophile keywords-related plots are periodic, their amplitude is much smaller than the amplitude of plots related to non paedophile keywords. This is true for the maximum and interestingly also for the minimum values as there are more occurrences of *pthc* in *eDonkey* queries between 3 and 6 am than occurrences of *divx* and *madonna*.
- Another significant difference is the absence of peak (and even the presence of an off-peak) between 7 and 10 pm in the number of paedophile keywords occurrences.

5. Conclusion

Various aspects of paedophile keywords' dynamics have been explored in this report and a number of conclusions have been derived:

- The proportion of paedophile queries is identical in the datasets collected in 2007 and in the dataset from 2009 (although the actual number of paedophile queries is higher in the 2007 dataset, which is due to a greater popularity of the corresponding *eDonkey* server).
- The 2 significantly most frequent paedophile keywords (among a list of 26 explicitly paedophile terms) are *pthc* and *pedo*.
- Within each dataset, the overall popularity of paedophile keywords has remained quite stable. This is not always the case as demonstrated by some non paedophile keywords in the 2009 dataset (e.g. *jackson, madonna*).

- A potential periodicity of paedophile keywords' dynamics has been investigated, over two periods' lengths: week and day.
    - The week of the day does not have a significant impact on the frequency of paedophile queries, which is also the case for the selected non paedophile keywords.
    - The hour of the day does have a critical impact on paedophile queries, and this is also similar to non paedophile keywords which show a clear periodicity over one day. However, two notable differences have been observed: the plots representing the evolution of paedophile keywords' popularity over one day have a much lower amplitude than the plots related to non paedophile keywords, a striking consequence being that there are more occurrences of *pthc* keyword in *eDonkey* queries than *madonna* and *divx* keywords between 3 and 5 am. Another interesting difference between the two classes of plots is the presence of a peak between 7 and 10 pm in the plots displaying the number of occurrences of non paedophile keywords whereas there is an off-peak for paedophile keywords at the same time of the day.

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*. Sixth International Workshop on Hot Topics in Peer-to-Peer Systems HotP2P'09. May 29, 2009, Rome, Italy.

[2] Jean-Loup Guillaume, Bénédicte Le Grand, Matthieu Latapy, and Clémence Magnien. Technical Report on *Behaviours of Users Entering Paedophile Queries*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content project.

[3] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Technical Report on *Automatic Detection of Paedophile Queries*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content project.

[4] 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.

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