First report on
Database Specification and Access
including
Content Rating and Fake Detection system

Matthieu Latapy, Clémence Magnien and Guillaume Valadon

LIP6 – CNRS and Université Pierre et Marie Curie

Firstname.Lastname@lip6.fr

Abstract

We present here the web-interface we designed and implemented to make it easier to access our data on P2P activity. In particular, this interface makes it browsable by browsing from files to related users, and conversely. Moreover, we include in this interface richer information which we precompute, like lists of files of interest for a given user, list of users providing a given file, list of queries entered by a given user, list of names of a given file, etc. Finally, we include a first content rating and fake detection system to this interface, which automatically decides if a file should be considered as having pornographic or paedophile content, and/or is a fake.

1 Introduction.

The data collected in the project is delivered as a set of XML files available online (public versions). These files are recordings of the activity captured on *eDonkey* during our measurements. These measurements are described in separate papers [1, 2], as well as the XML grammar [4]. Captured data is duly anonymised, the anonymisation procedure being described with each measurement (as the anonymisation scheme depends on the measurement procedure).

XML file format allows for rigorous description and specification of the data. It is a convenient way to store it in a normalized fashion. However, reading, searching, and more generally browsing in the dataset is difficult in this format. Therefore, we designed a *web* interface in order to help investigators to gain more insight, and to get it more easily.

Going further towards this goal, the interface provides access to richer information obtained from the raw data. For instance, we give the different filenames associated to each file, and the queries entered by each user. We also implemented a first version of content rating and fake detection systems. Our goal here is to indicate for each file if it may contain porn and/or paedophile materiel (content rating) and if it may have a content significantly different from the description given by its name (fake detection). This information is included in the web interface for evaluation purposes.

This report describes the first version of our interface in Section 2 and the richer information it provides, with a special stress on content rating in Section 3 and on fake detection in Section 4. Perspectives on our future work are given in Section 5.

All numerical values and plots given in this report are obtained from the measurement described in [1]. Including the other measurements to the interface described here is in progress. The behavior of the system will be very similar on them.

2 Web interface.

The public version of the web interface is available from `http://antipaedo.lip6.fr/Data/` which puts together links to all versions of our datasets and their documentations.

The goals of our web interface to the dataset are twofold: to make it possible and easy to browse the data, and to provide additional information obtained from raw data, but richer than it. We therefore have to design both a simple and intuitive interface and to decide which information to precompute and display. Notice that, due to the huge size of the datasets (dozens of millions of users, hundreds of millions of files), we have to be very careful in the design and implementation of this interface: some statistics would be interesting to display but are too long to be computed and/or too large to be stored; others would slow down data access considerably. We do not detail the technical implementation here; it is however important to keep in mind that some features could not be added for performance reasons.

The main part of the web interface is a way to navigate between files (identified by their *fids*) and users (identified by their *cids*) [1].

As explained in [5], we created two versions of the dataset: a public, fully anonymised version, and a version internal to the project, which is less anonymised. To reflect this, there are two versions of the web interface, a public version and a version with restricted access. The difference between these two versions is that the public version only contains anonymised words in queries and filenames. The restricted access version has the same level of anonymisation as in the internal dataset: words appearing in less than one hundred strings are replaced with an integer, and other words appear in clear.

The main goal of the interface is to be able to know which files a user has, and which users have a given file. To achieve this, it is possible to enter a *cid* or *fid* in a form in the web interface. The web interface will then present a page corresponding to this *cid* or *fid*. Depending on the choice of the user, different information are displayed.

If it is a file (*fid*), the page provides:

- the number of users who have downloaded or provided this file, and the corresponding list of *cids* [2]; if the number of users is larger than one hundred, only the first hundred are presented in order to improve the readability of the page;

[1] See [1, 2] and [5] for more details about *fids*, *cids* and our anonymisation procedures.

[2] Technically, we considered that a user was linked to a file if the server listed this user as a provider for this file, or if this user sent a query for providers of this file.

- the names of this file: we present the number of different filenames that this file has in the system, and give the liste of these (fully or partly) anonymised names;

- the first and last times at which this file was searched for or provided: this makes it possible to see if there was some activity concerning this file throughout all ten weeks of measurements, or only during a more restricted period [3].

If it is a client (*cid*), the page provides:

- the number of files that the corresponding user has downloaded or provided throughout his/her use of the system, and the corresponding list of *fids*; if the number of files is larger than one hundred, only the first hundred are presented in order to improve the readability of the page;

- the keyword queries performed by this user: we present the number of distinct keyword queries that the user sent, and we give the list of these (fully or partly) anonymised queries; this makes it possible to study what a user's interests are;

- the first and last date this user was seen on the system, which indicates the period during which the user was active in the system [4].

Each *cid* or *fid* in the presented lists is associated with a hyperlink to its own page on the web interface, which in turn presents the list of associated *fids* or *cids*, and so on. Figure 1 to 4 present screenshots of both public and private web interface for *cids* and *fids*.

For files, the web interface gives additional information regarding content rating and fake detection. The content rating system tries to identify automatically pornographic or paedophile content, while the fake detection system tries to detect files whose names does not correspond to their real content. The web interface presents the ratings obtained by both these systems. We implemented two methods for content rating and four for fake detection. Each method gives a rating for a file. This rating is given on the page describing a file, associated with a color code for easy interpretation: a rating on a green background means that this file is not thought to have pornographic content (or paedophile content, or be a fake, depending on the case); a rating on an orange background means that there is a reasonable suspiscion that this file has pornographic content (or paedophile content, or be a fake, depending on the case), and a red background means that the system is almost certain that the file has pornographic content (or paedophile content, or be a fake, depending on the case). Sections 3 and 4 present the detailed workings of the content rating system and the fake detection system, respectively.

All *fid*s described above in the navigation part of the web interface are anonymised, and it is therefore not possible to know to which file in the *eDonkey* system a given *fid* in the interface corresponds. We want however to provide a content rating and fake detection

[3] This feature is not fully implemeted yet, all displayed times are currently 0.

[4] This feature is not fully implemeted yet, all displayed times are currently 0.

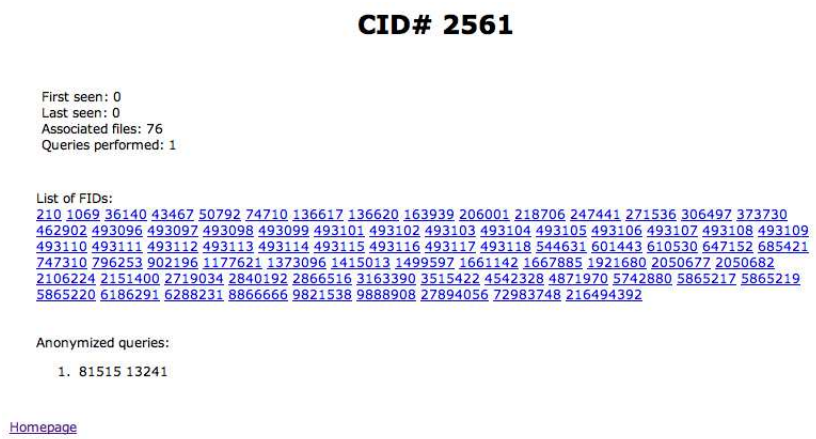

Figure 1: Web interface for clients. Public version where all keywords are anonymised.

service to end-users, allowing them to know the ratings for files the encounter on the system. We therefore set up a separate part of the web interface for this.

This part of the web interface will consists in a form, in which a user can enter the identifier of a file, as used in the *eDonkey* system [5]. The interface then provides the content and fake detection ratings for this file, in the same way as they are provided for anonymised *fid* in the navigation part of the interface, and they will be color coded in exactly the same way. No other information will be provided for this file (such as the number of users who have this file, or their list, or this file's names).

The idea for setting up a separate page is to not compromise the anonymisation of the dataset: we cannot allow users to make the correspondance between our *fid*s and the real identifiers used in the *eDonkey* system. THe navigation part of the web interface and the unanonymised files rating part of the interface raimain in this way totally separate from each other. This allows users to use both parts of the interface and benefit from the most services, without compromising the anonymisation of the dataset.

The content and fake detection part of the web interface is especially useful for providing an *automatic* connection to our system: developers of P2P applications can make the applications connect automatically to this interface, thus giving warnings about the content

[5] This is a `md4` hash of the file's content

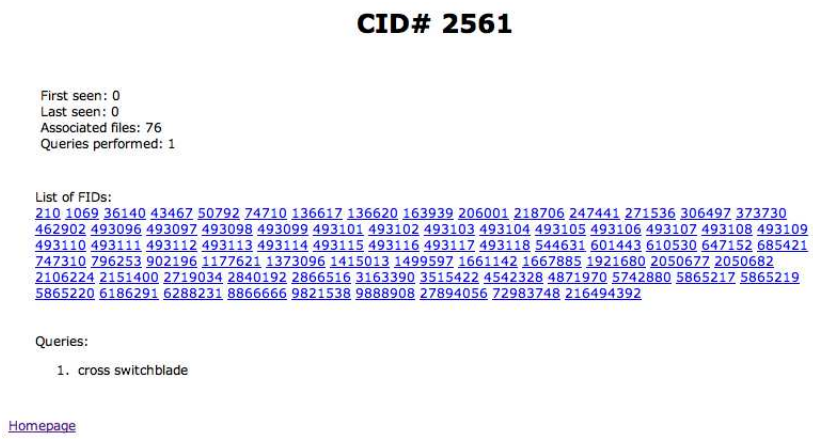

Figure 2: Web interface for clients. Private version where frequent keywords are displayed in clear. The data presented are fictitious.

of files to users.

3 Content rating.

Many files have a pornographic or paedophile content. The goal of the Content Rating System is to identify automatically these files. This version of the content rating system relies on filenames for doing so. Though it may seem easy to identify pornographic or paedophile files by reading their names, in practice this is not always the case: some words may have a pornographic or paedophile focus, but a user might not be aware of this; some files may have several names in the system, not all equally explicit, and the user may not have access to all these names; finally some files may be fakes, *i.e.*, have misleading names (we consider fakes in the following section).

The approach we used to detect if a filename has a pornographic (resp. paedophile) connotation first consists in defining a basic list of keywords that are known to have an explicit pornographic (resp. paedophile) meaning:

- the list of pornographic keywords we used here consists is based on the word *porn*: it contains this word and its derivatives and/or translations, such as *porno*;

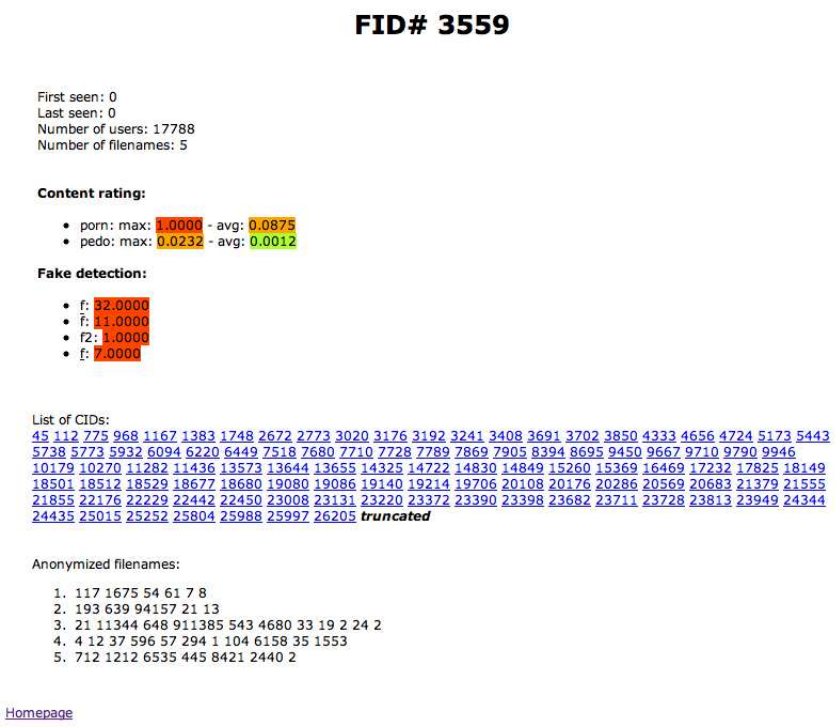

Figure 3: Web interface for files. Public version where all keywords are anonymised. The data presented are fictitious.

- the list of paedophile keywords is: *babyj, hussyfan, kidzilla, pthc, ptsc, raygold,* and *ygold*, a set of well known and explicit paedophile keywords.

The pornographic (resp. paedophile) nature of a filename cannot be established simply by asserting whether a filename contains a word from our list or not, for several reasons. First, it is impossible to put together an *exhaustive* list of words with a pornographic (resp. paedophile) connotation, and even if it was possible at a given time, this list would be very quickly outdated. Second, some words may have a weak pornographic or paedophile connotation: the fact that a filename contains a single such word may not be significant; however if a filename contains many such words, then it definitively has a pornographic or paedophile connotation (words such as *girl* or *teen* are good examples of this). Finally, some words can have a different connotation depending on the context: for the word *lolita* for instance, the filename `vladimir nabokov lolita.rar` is a writer's name and a book title and does not have a paedophile connotation, while the filename `lolita pics folder.rar` has a stronger paedophile connotation since it claims a set of pictures of a (young) girl.

We therefore assigned a rating to all words we observed in all filenames. The rating of a word depends on its tendency to appear in filenames that contain pornographic (resp. paedophile) words from our list. For each word w that appears in any filename, we selected all filenames containing w: $F(w)$. We then extracted the subset of $F(w)$ which also contains

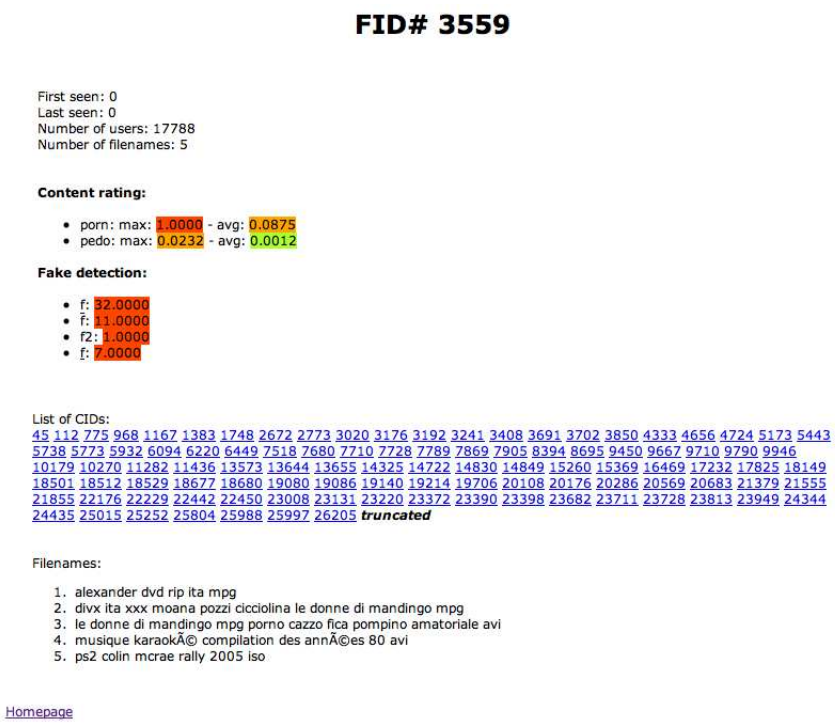

Figure 4: Web interface for files. Private version where frequent keywords are displayed in clear. The data presented are fictitious.

pornographic (resp. paedophile) words from our list: $F_{porn}(w)$ (resp. $F_{paedo}(w)$). The pornographic (resp. paedophile) rating of the considered word w is then the fraction of filenames containing words from our list, over the total number of filenames containing this word: $|F_{paedo}(w)|/|F(w)|$ [6]. For example, a word w has a paedophile rating of 50% if it appears half of the time in filenames which contain paedophile keywords from our list and half of the time in filenames which do not contain such words.

Using these words ratings, it is then possible to define a rating for each file, based on the ratings of the words composing its names. There are several ways to assign such a rating. For instance, if a file has several names, we can consider these names separately or join them in a single, larger, filename. We can decide that the presence of a single word with a high pornographic or paedophile rating in a filename is sufficient to assign a large pornographic or paedophile rating to the corresponding file, or we may choose to weaken this rating if this word appears together with words with very low ratings, etc.

All these solutions have their advantages and drawbacks, and we chose to implement two of them to evaluate their relevance. In both cases, we began by merging all filenames corresponding to a single file into a single, long name. We then considered the ratings of

[6] $|E|$ is the mathematical notation for the number of elements in E.

all words in this long name. The first method gives to the file the maximal rating between all words in its name. The second method computes the average of the ratings of all words in the long filename (if a word appears several times, it is counted several times in the average).

In the first case, taking the maximal rating could artificially give a very high rating to the file. Consider the case of a fake, *i.e.*, a file who has completely different names for a same content. If one of the names contains a paedophile word then the file will be highly rated whatever the other names are. In the second case on the contrary, it would be possible to hide paedophile keywords with a high rating by adding generic, low rating, keywords to lower the average.

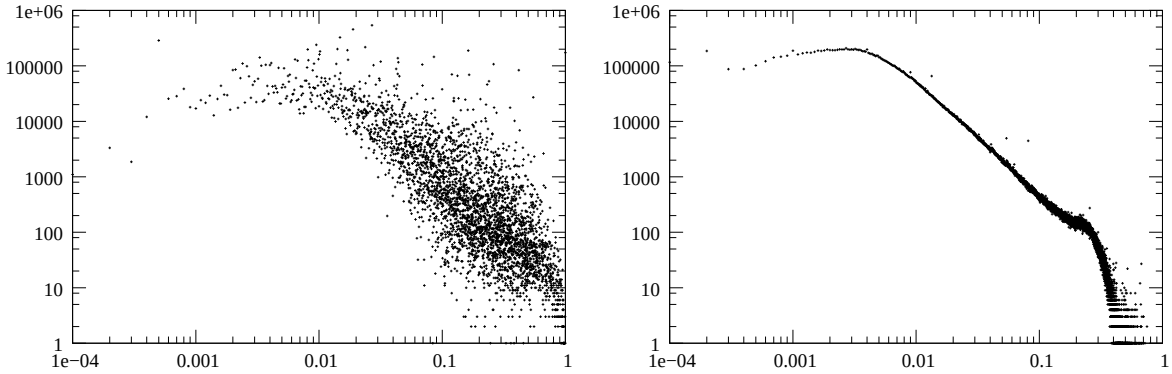

Figure 5: Distribution of the pornographic ratings of files. Left: maximum; Right: average.

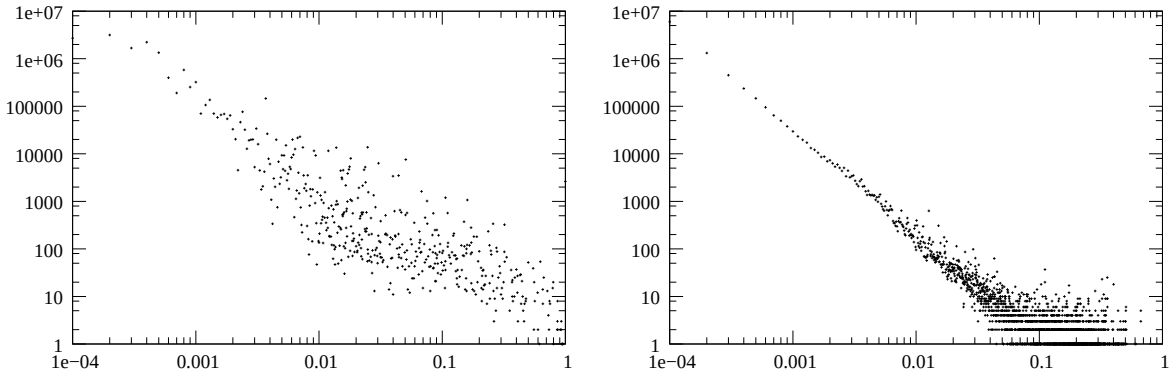

Figure 6: Distribution of the paedophile ratings of files. Left: maximum; Right: average.

Figure 5 (resp. Figure 6) presents the distribution of the pornographic (resp. paedophile) ratings obtained with these two methods for all files. As expected, in both cases the distributions for the maximum rating are more to the right-hand side of the plots than the distributions for the average rating: for a given set of words, the maximum rating is always larger than or equal to the average rating. However, though the distributions of the maximum ratings seem more scattered than the distributions of the average ratings, their shapes are similar. This indicates that both methods give the same type of classification.

While it is not visible on the Figure, some files have a paedophile or pornographic rating equal to 0. Among the 18 953 263 files that have a filename, 992 046 have a pornographic rating of 0, and 2 924 428 have a paedophile rating of 0 [7].

	porn.	paedo
maximum	2 360 109	12 242
average	369 884	2 460

Table 1: Number of files having a pornographic or paedophile rating larger than 0.1, for both rating methods.

We can see that all the distributions are very heterogeneous with a set of values going from 0 up to 1, with all intermediate values being well represented. Table 1 gives an indication of how heterogeneous these distributions are. It presents the number of files that have a rating larger than 0.1.

pornographic

	max.	avg.
low. threshold	0.05	0.01
high. threshold	0.15	0.1

paedophile

	max.	avg.
low. threshold	0.01	0.01
high. threshold	0.1	0.1

Table 2: Thresholds for the strengths of the ratings

This means that it is not easy find a clear threshold above which we have a high confidence that the corresponding file has a strong pornographic or paedophile connotation. After analysing the distributions and the data, we decided on a system of two thresholds of confidence for each rating. If the rating is below a lower threshold, this means that we are quite certain that it does not indicate a pornographic or paedophile connotation; if the rating is between a lower and a higher threshold, then we think there is a pornographic or paedophile connotation, but without any certainty; if the rating is above a higher threshold, then we are sure there is a pornographic or paedophile connotation. Table 2 presents the lower and higher thresholds for all our rating methods.

The ratings of a file are presented in the web interface, see 2. If a rating its below its lower threshold, it will be presented on a green background; if it is between its lower and higher thresholds, it will be presented on an orange background; finally, if it is higher than its higher threshold, it will be presented on a red background.

Another point concerns the shapes for the pornographic distributions and the paedophile distributions which differ significantly. In particular the distributions for pornographic ratings are much more to the right-hand than the corresponding distributions for

[7] When one of the ratings, maximum or average, is equal to 0, the other one is also necessarily equal to 0.

paedophile ratings. Therefore, in general, paedophile ratings are lower than pornographic ones, which was expected. This is confirmed by Table 1, which shows the numbers of files with ratings larger than 0.1. These numbers are much smaller for paedophile ratings than for pornographic ratings.

It is however interesting to note that some files have a paedophile rating larger than their corresponding pornographic rating: 63 056 files are in this case for the maximum rating, and 30 619 for the average rating. These are files with names having a strong paedophile connotation but no explicit sexual references. An example of such a file is `japan lolita kayo shiina 16yo ptsc zip` which contains 3 words with a strong paedophile connotation (lolita, 16yo and ptsc). The maximum paedophile rating of this file is more than 39 times larger than its maximum pornographic rating.

As mentioned earlier maximum and average ratings can be very different, indeed we can observe that some files have a high (pornographic or paedophile) maximum rating, while their corresponding average rating is low. This could indicate that, at least for these files, one method is largely more accurate than the other. A quick look at some of these files however shows that most of them seem to be fakes, *i.e.*, files with misleading names. For instance, one file with a maximum pornographic rating of 1 and an average pornographic rating of 0.043 has in fact 7 different names, all indicating different contents: 6 indicate different kinds of music, while the last one contains various keywords including pornographic ones (porno, sex, sexy, hentai, xxx):

```
06 scissor sisters i don t feel like dancing paper faces remix mp3
abba -165993 mp3
blink 182 all the small -1556559 mp3
christina aguilera aint no other man mp3
high school musical breaking free karaoke instrumental mp3
singoli house febbraio 2006 pryda friends 123 mp3
guadagna in un mese 5000 euro su ebay testato e funzionante divx ita soldi
gratis msn porno sex sexy pc game hentai manga xxx xbox ps2 psp ebay doc
```

It is not surprising that our methods do not give consistent results in such cases: the file should be first rated as fake before going further and trying to evaluate its content ratings. We actually study the case of fake files in the next section.

In some specific cases, the difference between the maximum and the average rating is significant while the filenames are consistent with each other. For instance, one file having a maximum pornographic rating of 1 and an average pornographic rating of 0.042 corresponds in fact to the movie *Romanzo Criminale*. All seven names of this file are very similar, and one of them is `divx ita romanzo criminale finalmente non porno avi`. Though this name explicitely indicates non-pornographic content, the presence of the word *porno* in it gave the file a maximum pornographic rating of 1. Using both indicators together might in this case give more insight on the real content of the file.

In summary, we presented here a first content rating system which gives an indication of whether a file has a pornographic/paedophile content or not. We showed that this system,

though very simple, already gives relevant results. A key point is that in many cases fake files lead to very different results for maximum and average ratings, which prevents from using these ratings advisedly. This emphasizes the need for accurate fake detection methods, which we address in the next section.

The methods we presented, tough relevant, are still very basic; we plan to improve them in future versions. For instance, it would be possible to use conjointly both maximum and average ratings to define a third one, or to define more precise ratings based on filenames. Other approaches would consist in studying proximity relations between files: if a user is offering two different files then the content of these might be related in some way, in particular if one of these files is paedophile then the other one might also be paedophile, whatever its name. We will detail these perspectives in Section 5.

4 Fake detection.

A fake file is a file that has a misleading name; in other words, its content differs significantly from the description given by its filename. Consequently, users may download it and be exposed to unwanted, possibly harmful, content. For instance, it is well known that some pornographic content may have innocent names (in particular, names of cinema movies). Fake files may be introduced by malicious users who want to disturb the system, to expose other users (in particular children) to harmful content, etc. By extension, if a file has different names including misleading ones, it can be considered as a fake.

Detecting fakes at a large scale is a subtle task, as in principle one would have to inspect the content of many files. Designing automatic detection methods (or tools to help manual inspection) therefore is of great interest, in particular for user protection. We propose here a first approach based on filenames addressing this goal. More advanced versions with much more subtle methods will be designed later within our project, see Section 5.

First notice that filename-based fake detection can work only on files which have several names which we observed in our measurements. Among the 18 953 264 files for which we observed at least one filename in our trace, 15 849 844 have only one. However, some have many filenames, up to 82 (see Figure 7). This means that files with exactly the same content have different filenames (see for instance the example in Section 3). However, according to our definition of filenames, the difference may be tiny: change in separators (spaces replaced with comas, for instance); replacement of some upper case letters by their lower case version, or conversely; switched words; etc. More subtle changes may occur, like translations of filenames, addition or deletion of details, etc. Normalised filenames can help for some of these problems, but not all of them.
For instance, the file with 82 filenames in our dataset actually has only 6 different names after normalization (see [5], Section 2). Moreover, its names are permutations of the same words. This should not lead us to conclude that this file is a fake:

```
jojo the high road too little too late wma (19 times)
jojo too little too late the high road wma (9 times)
the high road jojo too little too late wma (13 times)
```

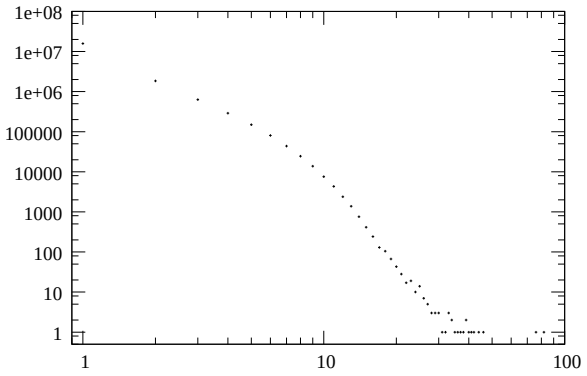

Figure 7: Distribution of the number of names per file.

```
the high road too little too late jojo wma (15 times)
too little too late jojo the high road wma (10 times)
too little too late the high road jojo wma (16 times)
```

In other cases, the fake nature of the file is clear. For instance, the following file is clearly a fake, since the first filename describes porn content while the second describes a music file:

```
porno paris hilton original private fuck video avi
rap sinik la main sur le coeur 2005 rar
```

These examples clearly show that the raw number of filenames (even normalized) is not sufficient to conclude that a file is a fake: we have to design more subtle computations to capture this.

A very natural idea is to compare the set of words in the various filenames of a same file. In the examples above, this would give a relevant indication, and one may expect that this is true in general. There are however many ways to compare these sets: one may count the number of words in a filename but not in others; or compute the number of words in common between filenames; these values may be normalized or not; etc. Choosing between such alternatives is a well-known problem, studied in various contexts. See for instance [3]. All solutions have their own advantages and drawbacks, and we decided to implement several of them in order to evaluate their relevance. We present them below.

Let us first introduce some notations useful in the following. We consider a specific file F for which we have several filenames. We denote by N the set of all these filenames. A filename is a set of words. In addition, we will use a few classical mathematical notations. We denote by $|S|$ the size (number of elements) of any set S. The difference between two sets S and T, *i.e.*, the elements in S but not in T, is denoted by $S \setminus T$.

We propose below 4 distinct ratings for fake files. They are all based on the idea that if filenames have many words in common then they may be considered as similar and thus do not indicate that the content under concern is a fake. On the countrary, if a filename has many words which do not appear in other filenames, then probably the file is a fake.

We will use in the following a simple example of file with 3 filenames to illustrate the ratings. The filenames are: (1) `jojo the high avi`, (2) `jojo high avi` and (3) `porn sex avi`. Two filenames are very similar and the third one is completely different; this file is thus a fake.

The first rating f is evaluated as:

$$f = \frac{|\cup_{n\in N} n|}{|\cap_{n\in N} n| + 1},$$

and represents the number of distinct words divided by the number of words in common in all filenames [8]. With our example, there is 6 distinct words and only 1 of them (*avi*) belongs to all filenames. The fake ratio therefore is $f = 6/2 = 3$. More generally this approach will tend to give high ratings if there are many words, therefore files with many different names will have high ratings.

To avoid this, one may consider that the content is a fake if *one* of its names is very different from the others. This is captured by the following:

$$\overline{f} = \max_{n\in N} \frac{|n|}{|\cap_{m\in N} m| + 1} = \frac{\max_{n\in N} |n|}{|\cap_{m\in N} m| + 1}.$$

In this case, we compare each name to the intersection and we take the worst case. In our example, the longest word is of size 4, the intersection is still 1, the rating is therefore $4/2 = 2$. With this rating, the number of distinct filenames in not taken into account but it is rather the length of the filenames which has some influence. In particular, if one filename is much longer than the others, this can give a high rating.

Both previous approaches were trying to compare all filenames simultaneously. In this third approach, we consider the largest difference between any two names:

$$f_2 = \max_{n,m\in N} \frac{|n \setminus m| + |m \setminus n|}{|n \cup m|}.$$

Here we compare all pairs of filenames, looking for the worst case, *i.e.*, a pair of filenames which are very dissimilar. Using our example, we obtain ratings of $(1 + 0)/4 = 1/4$ for (1) and (2), $(3 + 2)/6 = 5/6$ for (1) and (3) and $(2 + 2)/3 = 4/5$ for (2) and (3). The maximum is $5/6$ for (1) and (3) which corresponds to the intuition. Note that with this definition, the rating is always lower than or equal to 1, which is an advantage as it helps interpretation. As soon as two filenames are very different, the rating will be near 1 and exactly equal to 1 if two filenames have no word in common.

In a similar way, we consider the smallest intersection between any two names:

$$\underline{f} = \max_{n,m\in N} \frac{\min(|n|, |m|)}{|n \cap m| + 1}.$$

[8] The +1 at the denominator is used to avoid dividing by 0 if there is no word common to all filenames.

This approach also consists in comparing pairs of filenames. For our example it gives the following ratings: 3/4 for (1) and (2), 3/2 for (1) and (3) and 3/2 for (2) and (3). The maximum is 3/2.

Figures 8 and 9 give the cumulative distributions of the four ratings: for each value r of the rating on the x-axis, the y-axis gives the number of files which have a rating lower than or equal to r. Two distinct behaviour are observed, the first concerns ratings f, $\overline{f}$ and $\underline{f}$, the second concerns f_2.

The first behaviour concerns plots where there is a sharp transition (a nearly vertical increase of the cumulative distribution). This means that there are a lot of files with a very similar rating. For f, $\overline{f}$ and $\underline{f}$ there are many ratings around 1. However, it must be noticed that the ratings go far beyond these values, up to 210 for f, 37 for $\overline{f}$ and 29 for $\underline{f}$. Files with such high ratings are definitely way more fake than the average. To be more precise, only 10% of the files have a f (resp. $\overline{f}$ and $\underline{f}$) ratio over 3.5 (resp. 2.6 and 1.5) and only 1% of the files have a f (resp. $\overline{f}$ and $\underline{f}$) ratio over 17 (resp. 9 and 5). These results are summarised in Table 3.

The second behaviour concerns f_2 where the distribution of the rank values is more linear. This means that there is no typical value of this ranking function, therefore it is less easy to find a threshold above which we can say for sure that a file is a fake.

function	10%	1%
$\underline{f}$	3.5	17.0
$\overline{f}$	2.6	9.0
f_2	0.77	1.0
$\underline{f}$	1.5	1.5

Table 3: Extremal ranks for the different ranking function. The table reads as follows: the value 3.5 for f and 10% means than only 10% of files have a f ranking over 3.5.

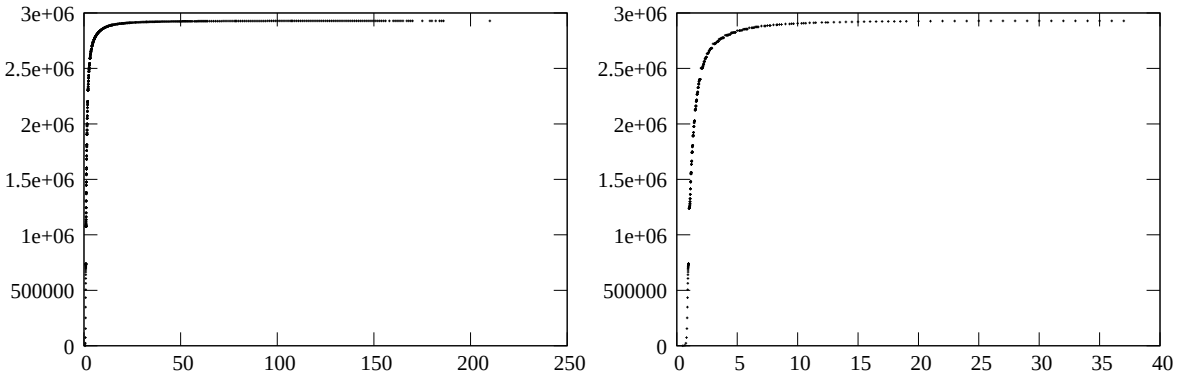

Figure 8: Cumulative distribution of the fake indicators for all files with at least two filenames. Left: f rating ; Right: $\overline{f}$.

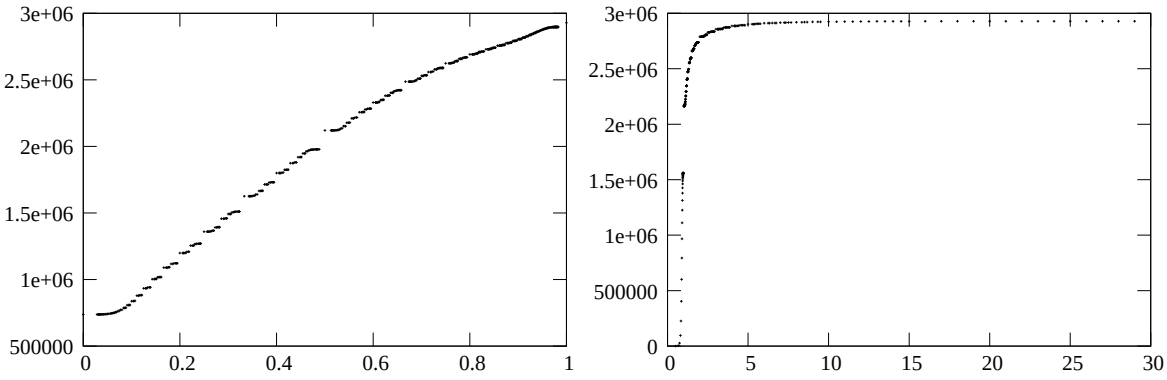

Figure 9: Cumulative distribution of the fake indicators for all files with at least two filenames. Left: f_2; Right: $\underline{f}$.

We give in Tables 4, 5 and 6 the files with the top ten values of each of the rankings f, $\overline{f}$ and $\underline{f}$. We omitted f_2 since there is 31 707 files with a f_2 ranking of 1, the maximal value. Choosing arbitrarily 10 out of 31 707 would have made no sense.

Using results from Table 3, we can see that the top 10 fake files according to f are also very high ranked for all other functions, these files are always in the top 1% for the other ranks. The same applies for $\underline{f}$ and $\overline{f}$, except for three files of the top ten of $\underline{f}$ which are only in the top 10% of $\overline{f}$ (in the top 3% to be precise). In particular, it must be noticed 18 040 are in the top 1% of all rankings.

rank	fid	nb names	f	$\overline{f}$	f_2	$\underline{f}$
1	953327	24	210	28	1	18
2	5248002	21	186	28	1	16
3	35889	22	186	26	1	19
4	1571381	24	185	19	1	17
5	256151	20	184	22	1	17
6	1402860	22	182	21	1	17
7	166248	20	180	26	1	15
8	4556264	21	179	19	1	17
9	4070223	17	175	26	1	19
10	975597	16	170	24	1	24

Table 4: Top ten words with the highest f ranking.

Finally, similarly to what we have done for the content ratings, we have analysed the plots and the data to obtain thresholds of confidence for the four ratings. If a rating is below its lower threshold, this means that we are quite certain that it does not indicate a fake; if the rating is between a lower and a higher threshold, then we think the corresponding file is a fake, but without any certainty; if the rating is above a higher threshold, then we are

rank	fid	nb names	f	$\overline{f}$	f_2	$\underline{f}$
1	12051325	8	87	37	1	15
2	1389059	7	61	36	1	11
3	7028825	5	65	35	1	13
4	5674953	4	40	35	1	4
5	5494832	4	47	35	1	7
6	3816768	4	47	35	1	7
7	24217102	3	44	35	1	4
8	199112674	9	90	35	1	12
9	19510077	2	38	35	1	3
10	17880046	2	44	35	1	9

Table 5: Top ten words with the highest $\overline{f}$ ranking.

rank	fid	nb names	f	$\overline{f}$	f_2	$\underline{f}$
1	12538484	5	73	29	1	29
2	1792881	4	60	30	1	28
3	3175307	5	57	28	1	27
4	2923788	7	89	32	1	26
5	1924501	11	148	27	1	26
6	1874216	7	88	26	1	26
7	12323381	14	152	26	1	26
8	960369	14	139	26	1	25
9	65900	13	126	28	1	25
10	2521164	5	72	25	1	25

Table 6: Top ten words with the highest $\underline{f}$ ranking.

sure that the corresponding file is a fake. Table 7 presents the lower and higher thresholds for all our rating methods.

The ratings of a file are presented in the web interface, see 2. If a rating its below its lower threshold, it will be presented on a green background; if it is between its lower and higher thresholds, it will be presented on an orange background; finally, if it is higher than its higher threshold, it will be presented on a red background.

We presented here a fake detection system based on four distinct rankings which tells whether a file is suspected to be a fake or not. Manual inspection will help enhancing these results. Moreover, fake detection is very useful for content rating system as stressed before and these results have to be used conjointly with content rating.

The methods we presented are still very basic and we will improve them in the future. Taking all or some rankings into account could help in enhancing the system. More advanced techniques could also be used to avoid problem such as the following two names for

	f	$\overline{f}$	f_2	$\underline{f}$
low. threshold	0.8	0.9	0.5	1
high. threshod	3.5	2.5	0.8	2

Table 7: Thresholds for the strengths of the ratings.

the same file, which is ranked in the top 1% for all 4 rankings:

`red head gang bang very good cum shot -1168109 avi`

`tyra olsen gangbang shaved big black white cocks anal double penetration bukkake mpg`

This file has two completely distinct names (even the extension is not similar), however both descriptions are very similar. Similarly, if a file has two names, one of which is the translation of the other in another language, then this file would have a very high fake ranking.

Among more advanced techniques, it would be possible to detect users who often share fake files and who could be at the origin of these fakes. Detecting such users would help in detecting more fake files, or to do it more accurately. Conversely, the number of users which are sharing a given file under a given name might help to find automatically the real content of the file; a trusting factor could therefore be defined based on the number of users. We will detail these perspectives in Section 5.

5 Conclusion and future work.

In this report we first presented the web interface to access the data. This interface allows to get information on *fids* and *cids* and to navigate easily between both. This interface is available both in a public version where everything is anonymised, and in a restricted version where frequent keywords are displayed in clear.

We plan in the future to improve this interface. In particular it would be interesting to have an access through keywords which would allow to directly know all the files containing a given word in their filename, or all the clients who have typed this word.

We also plan to include a possiblity for the users to give feedback on our content and fake detection ratings. It must be clear that these ratings are only prototypes at this stage, and need to be refined. Feedback from the users, saying if they think a file has a pornographic or paedophile content, of if this file is a fake, would be a great help for this.

We will also add new statistics to the web interface: currently the computation of the first and last date a file or client is seen is not fully implemented; this feature will be completed in the future. We will also present, for each file, the plot of the number of clients having this file, as a function of time: this plot will make it possible to see if this file was more popular (acquired by more clients) at some periods than others.

We also presented a content rating system which allows to detect automatically pornographic and/or paedophile files and a fake detection system which detects files whose names

are misleading. We implemented several methods for this, which give different ratings which are all included on the web interface for their evaluation.

We defined thresholds for the strength of each rating: above a certain threshold, a rating indicates a high confidence for the corresponding characteristic (pornographic or paedophile content, or fake). Defining such thresholds is however a difficult tasks, and the thresholds we defined will probably need to be refined in the future.

One other direction for improvement is that our content rating systems is based on lists of pornographic and paedophile keywords. It would be interesting to see if different lists would give the same results or not.

The current fake detection system relies on 4 distinct ratings which globally give similar results. Contrary to content rating, the results of these 4 functions are much less heterogeneous and it is easier to find extremal values.

Much work remains to be done to increase the potential of these systems. First, a deeper evaluation of the current ratings must be done. Then, many new approaches have to be tested, among which ratings based on graphs and communities.

This approach consists in constructing a graph from the raw data: for instance two files can be connected if a user in interested in both files. Indeed, if a user is interested in two files then it is more likely that these files have something in common. Using all the information contained in the raw data allows to build a very large network of similarities between files.

Given such graphs, algorithms are available to detect groups of very similar nodes. In this context, we hope to detect groups of pornographic or paedophile files even if the files have misleading names (*i.e.* are fakes), or don't have names. This could also help to study some other directions for fake detection, like for instace the fact that some users may be creating many fakes. Detecting such users could help in detecting fakes, even for files with only one filename.

Acknowledgements. All participants to the project helped in conducting the work presented in this report. It is supported in part by the MAPAP SIP-2006-PP-221003 and ANR MAPE projects.

References

[1] Frederic Aidouni, Matthieu Latapy, and Clemence Magnien. Ten weeks in the life of an edonkey server. Submitted, 2008.

[2] Oussama Allali, Matthieu Latapy, and Clémence Magnien. Measurement of edonkey activity with honeypots. Submitted, 2008.

[3] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. To appear in Social Networks, 2007.

[4] Clémence Magnien and Frédéric Aidouni. XML grammar for encoding of edonkey traces. `http://antipaedo.lip6.fr/Data/`.

[5] Clémence Magnien, Matthieu Latapy, Jean-Loup Guillaume, and Bénédicte Le Grand. First report on paedophile keywords observed in edonkey. http://antipaedo.lip6.fr/.