Technical report on

Database Specification and Access

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

Matthieu Latapy [1], Clémence Magnien, Fabien Tarissan and Guillaume Valadon

Abstract

In this report we present the web interface we designed and implemented to make our data on P2P activity more accessible. In particular, this interface makes data browsable from files to related users, and conversely. Moreover, this interface was enriched with precomputed information, 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 included our content rating and fake detection system in this interface, which indicates if a file should be considered as having pornographic or paedophile content, and/or as being a fake.

1 Introduction.

The data presented in this Web interface was collected during the project. It consists in recordings of the activity captured on the *eDonkey* system. The measurements are described in [1]. The data is delivered as a set of XML files [8], together with the corresponding XML grammar [6]. Captured data is duly anonymised, and the anonymisation procedure is described in the above measurements.

XML file format allows for rigorous description and specification of the data. It is a convenient way to store it in a normalised way. However, reading, searching, and more generally browsing 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, this interface provides the different filenames associated to each file, and the queries entered by each user. It also gives access to the content rating and fake detection system developed in the project, described in detail in [3]. This system indicates 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 report describes the final version of our interface in Section 2 and the richer information it provides. We implemented a first version of the web interface in 2008,

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

described in a previous report of the project [5]. Section 4 presents the modifications made to the web interface since this date. 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].

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

2 Browsing and inspecting the data.

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. 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 dataset (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 in this report; 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 used to navigate between files (identified by their *fids*) and users (identified by their *cids*) [2].

As explained in [7], 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, we created a public version and a version with restricted access to the web interface. 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 the internal dataset: words appearing in less than one hundred strings are replaced with an integer, and other words appear in clear. Finally, to help the evaluation of our data and results by external experts, including law-enforcement personnel, a third version of the web interface displays the *eDonkey* hash of the files, instead of their anonymised *fids*.

The main goal of the interface is to be able to know which files a specific 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 informations are displayed.

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

- the first and last times when this file was searched for or provided, as well as the interval elapsed in between: 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;

[2] See [1, 2] and [7] for more details about *fids*, *cids* and our anonymisation procedures.

- the number of users who have downloaded or provided this file, and the corresponding list of *cids*[3]; 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;

- the number of users who made at least one paedophile query, among those who have downloaded or provided this file. Paedophile queries are identified by the filter developed in the project for automatic paedophile query detection [4]. These users are identified in the interface as *paedophile users* and their *cid* appear in bold red in the *cid* list;

- the names of this file: we present the number of different filenames corresponding to this file in the system, and give the list of these (fully or partly) anonymised names; the filenames that are detected as paedophile by the filter are printed in bold red;

- a plot showing the evolution of the number of users who have downloaded or provided this file over time; this plot represents both the number of distinct users and the cumulated number of queries for this file and citations of providers for this file;

- the *fid* at the top of the page is printed with a colour depending on the file type, inferred from the extensions of its names; we identified seven categories: Archive, Audio, Documents, Games, Images, Softwares, Video; if the extensions are unknown or if they do not all belong to the same category, the file type is unknown and the *fid* is printed in black.

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

- the first and last time this user was seen on the system, as well as the interval elapsed in between, which indicates the period during which the user was active in the system.

- 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 *fids* in the lists are presented on a coloured background, the colour depending on the file type, inferred from the extensions of its names (see above);

- the keyword queries performed by this user: we present the number of distinct keyword queries sent by the user, as well as the number of distinct queries classified as paedophile by the automatic paedophile query detection filter [4]; we also give the list of these (fully or partly) anonymised queries; this makes it possible to study what a user's interests are; queries classified as paedophile are printed in bold red;

[3]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.

- a plot showing the evolution of the number of files that the corresponding user had downloaded or provided during time; this plot represents both the number of distinct files and the cumulated number of queries for files and citations of this user as a provider for a file;
- the *cid* at the top of the page is printed in bold red if the corresponding user has made at least one query classified as paedophile; it is printed in black otherwise.

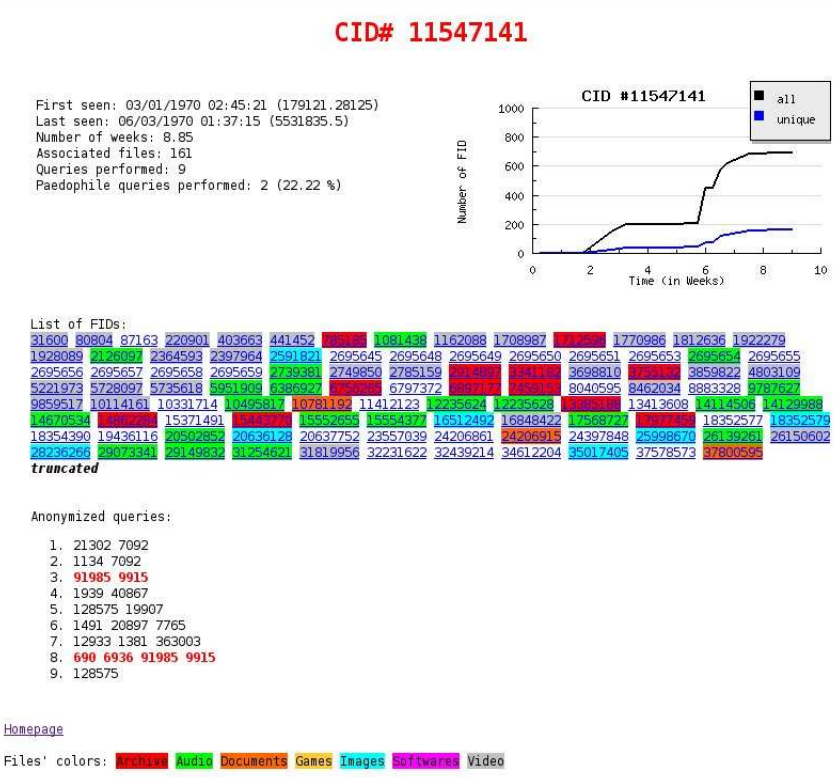

Figure 1: Web interface for clients. Public version where all keywords are anonymised.

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*. We do not present any screenshot of the version of the interface showing the unanonymised *eDonkey* hashes, since it is very similar to the private version.

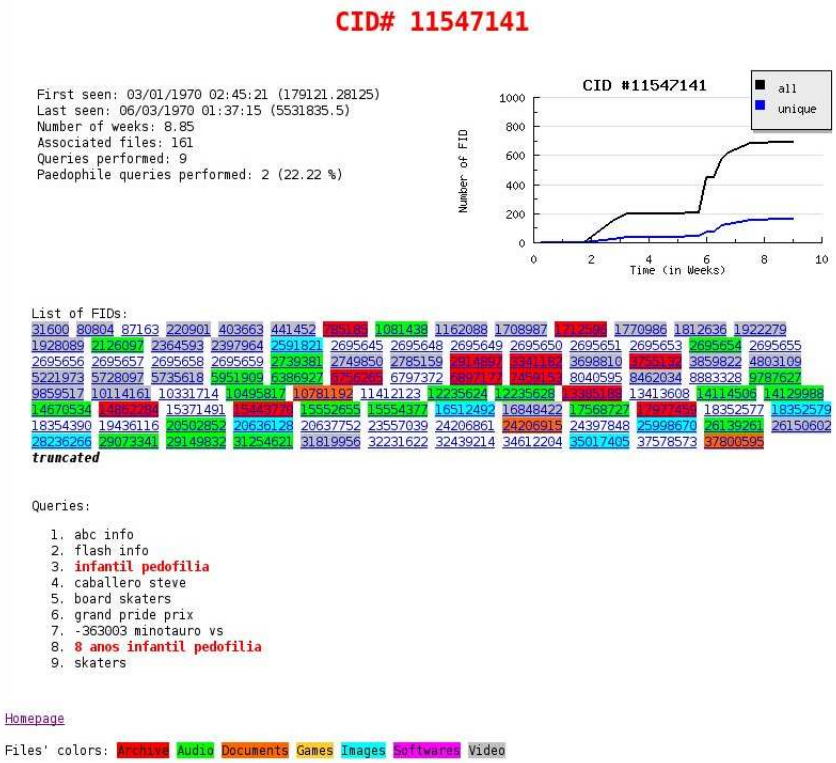

Figure 2: Web interface for clients. Private version where frequent keywords are displayed in clear.

Finally, a dedicated page [4] lists all the queries, made by all users, which were classified as paedophile. Each query is associated with a hyperlink to the page of the user who entered this query. This page makes it easier to study paedophile activity in the dataset.

For files, the web interface gives additional information regarding content rating and fake detection. The content rating system attempts to identify automatically pornographic or paedophile content, while the fake detection system tries to detect files whose names do not correspond to their real content [3]. The web interface presents the ratings obtained by both systems. The content detection system classifies files which have at least one provider [5] as *not paedophile*, *maybe paedophile*, and *probably paedophile*, as well as *not*

[4] Also available from `http://antipaedo.lip6.fr/Data/`

[5] We do not have enough information on other files to perform accurate classification.

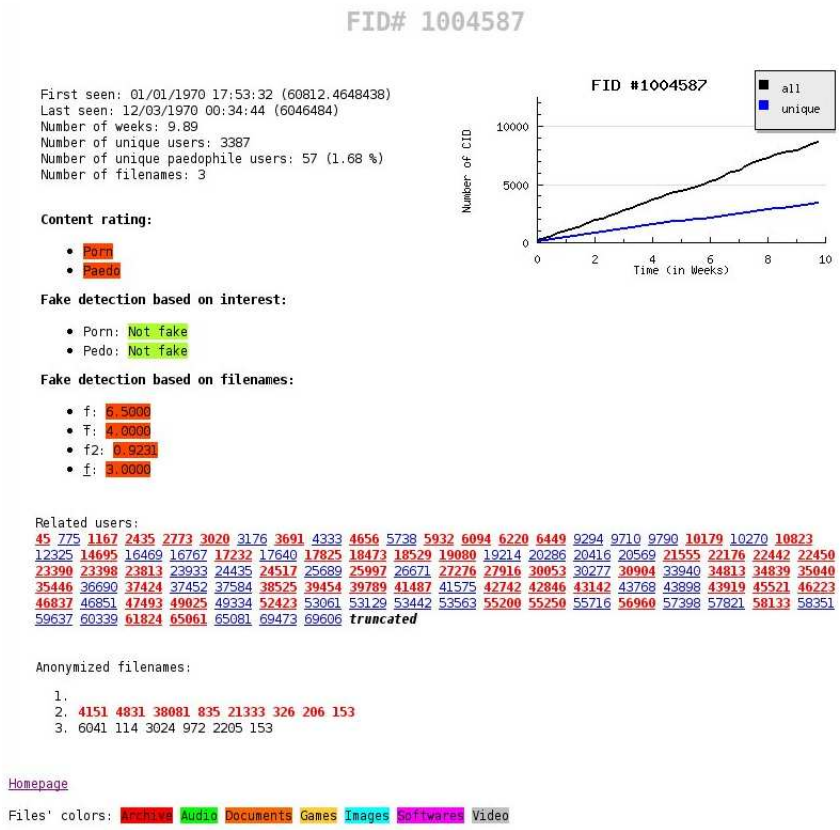

Figure 3: Web interface for files. Public version where all keywords are anonymised.

pornographic and *probably pornographic*. The fake detection system classifies files as *not a paedophile fake*, or *fake: file with a paedophile name but non paedophile content*, or *fake: file with a non paedophile name but with paedophile content*. It also classifies them as *not a pornographic fake*, or *fake: file with a pornographic name but non pornographic content*, or *fake: file with a non pornographic name but with pornographic content*. These classifications are associated to a colour code, composed of three colours: green, orange and red, for easy interpretation: a green background indicates that this file is not thought to have paedophile or pornographic content, or be a fake, depending on the case; an orange background indicates that this file may have paedophile or pornographic content, or be a fake, while a red background indicates that this file probably has paedophile or pornographic content, or is a fake. The web interface also presents the results of the previous fake detection system implemented in the project [5], which was applied to files having more than one name.

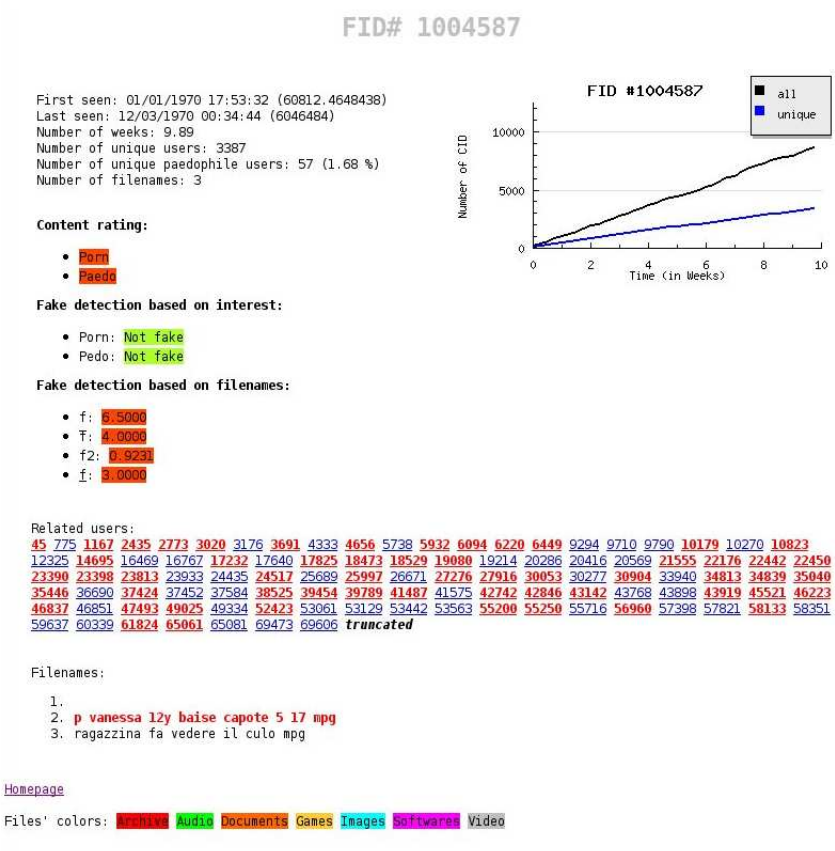

Figure 4: Web interface for files. Private version where frequent keywords are displayed in clear.

Depending on the differences between these names, the content rating system computed four ratings (corresponding to four different backgrounds). Each rating is printed on a coloured background for easy interpretation: green means that the corresponding method does not consider the file to be a fake, orange that there is a reasonable suspicion that this file is a fake, and red that the system is almost certain that the file is a fake. The detailed operations of the content rating and the fake detection systems are presented in [3].

3 Rating an *eDonkey fid*.

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 service to end-users, allowing them to know the ratings of files they encounter on the system. We therefore set up a separate part of the web interface for this. It is available from `http://antipaedo.lip6.fr/Data/`.

This part of the web interface consists in a form, in which a user can enter the identifier of a file, as used in the *eDonkey* system. The interface then provides the content and fake detection ratings for this file, as it does for anonymised *fids* in the navigation part of the interface, with the same colour code. No other information is provided for this file (such as the number of users who have this file, or their list, or this file's names).

Setting up a separate page avoids compromising the anonymisation of the dataset: we cannot allow users to make the correspondence between our *fid*s and the real identifiers used in the *eDonkey* system. By doing this, the navigation part of the web interface and the unanonymised files rating part of the interface remain totally separate from each other. This allows users to use both parts of the interface and to benefit from the highest number of 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 peer-to-peer applications may connect the applications automatically to this interface, thus providing users with warnings about the content of files.

4 Changes

In this section we give a quick summary of the changes that have been implemented in the web interface since the first version, delivered in October, 2008 [5].

- a version allowing navigation with the unanonymised file hashes has been implemented;
- the dates at which users and files have been observed the first and last times, as well as the elapsed interval in between, have been added;
- *fids* are coloured according to file extension (both in title of *fid* pages and in *fid* lists);
- paedophile queries are printed in bold red;
- *cids* of users who entered at least one paedophile query are printed in bold red (both in title of *cid* pages and in *cid* lists);
- for all *fids*, the number of paedophile users (i.e. having entered at least one paedophile query) who have downloaded or provided this file is given (together with the corresponding percentage);
- for all *fids*, the filenames which are classified as paedophile by our filter are printed in bold red;

- for all *cids*, the number of paedophile queries entered is provided (together with the corresponding percentage);
- for each *fid*, a plot representing the evolution of the number of users having downloaded or provided this file is given;
- for each *cid*, a plot representing the evolution of the number of files this user has provided or downloaded is given;
- a dedicated web page listing all paedophile queries has been provided; each query has a link to the *cid* who entered this query;
- a form has been added to the unanonymised hash search page for the content rating and fake detection systems.

5 Conclusion and perspectives.

In this report we first presented the web interface we implemented to browse the data. This interface allows to get information on *fids* and *cids* and to navigate easily between both. This interface is available in a public version where everything is anonymised, in a restricted version where frequent keywords are displayed in clear, and a more restricted version in which the *eDonkey* hashes are not anonymised.

The web interface also provides access to the results of the content rating and fake detection system. Two accesses have been implemented: one is available on the page of each *anonymised fid*, when browsing through the data; the other is on a dedicated web page and can be accessed publicly. It provides the ratings for *unanonymised eDonkey* hashes, which is useful for end-users.

This interface may be improved further in several ways. In particular it would be interesting to have an access through keywords which would allow to directly know all files containing a given word in their filename, or all the clients who have typed this word.

In order to refine our methods, it would also be interesting to include a possibility for users to give feedback on our content and fake detection ratings. Feedback from users, and in particular law-enforcement personnel, consisting in saying if they think a file has a pornographic or paedophile content, or if this file is a fake, would be a great help for this.

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] Oussama Allali, Matthieu Latapy, and Clémence Magnien. Measurements of *eDonkey* activity with distributed honeypots. In *Proceedings of HotP2P'09*, 2009.

[3] Jean-Loup Guillaume, Matthieu Latapy, Clémence Magnien, and Guillaume Valadon. Technical report on the *Content Rating and Fake Detection System*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[4] 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.

[5] Matthieu Latapy, Clémence magnien, and Guillaume Valadon. First report on *Database Specification and Access including Content Rating and Fake Detection system*, 2008. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

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

[7] Clémence Magnien, Matthieu Latapy, Jean-Loup Guillaume, and Bénédicte Le Grand. First report on *Paedophile Keywords Observed in eDonkey*, 2008. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[8] Ten weeks measurement on an edonkey server. `http://content.lip6.fr/latapy/edonkey/weeks-1.0.1/`.

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