The MAPAP project - Measurement and Analysis of P2P activity Against Paedophile content

The MAPAP project studies paedophile exchanges on P2P systems, with two goals in mind:

- designing an automatic content rating system allowing users to be protected from this type of content; and
- help law enforcement institutions and child protection organizations in their task.

We aim in particular at designing tools for the automatic detection of new paedophile keywords. More generally, our work aims at providing an improved knowledge on paedophiles activity on P2P systems.

Many studies and independent contributions show that a huge amount of paedophile and harmful contents are distributed using p2p file exchange systems, and that the volume of such exchanges is increasing. The presence of such content, and its very easy access, make the current situation particularly worrying for p2p users, in particular children. This is even more alarming if one considers the fact that many fakes, i.e. files with contents that differ significantly from their names, are present in these systems. Because of this, all users, including children, face a high risk of downloading and visualising unwanted content.

Despite the fact that this situation is nowadays widely acknowledged, there is still no available filtering technique or content rating system to protect p2p users. Similarly, only few tools exist to help law enforcement authorities and other child protection organisations in fighting p2p paedophile exchanges.

The objective of the MAPAP project is to tackle these issues by implementing key software, setting up reference databases and conducting leading studies, both to protect p2p users, in particular children, and help law enforcement authorities and other child protection organisations in their task. More precisely, we will focus on the following three areas, each with its own objectives.

- Content rating and fake detection system

Our core objective is the design and implementation of a service able to give, for any file encountered in our measurements, a rating of its content as paedophile and/or pornographic, as well as an indication of the fact that it may be a fake or not. A confidence ratio will be associated to each of these indications. This service will be available on-demand to end-users through a web page form, but its use will be limited to avoid abuses (typically, we will limit the number of queries per user and per time unit in order to prevent users from searching paedophile content with it). A full unrestricted version will be provided to relevant institutions, with additional information like the date of first appearance of the content, the number of peers providing/downloading it during time, etc.

Such a tool would be a first step towards the possibility for ISP to filter p2p content, and for end-users to have indications on the content of a file they are interested in, before downloading it4. It may also be included in parental control systems and in p2p clients, which may send automatic queries to our system when needed. This would allow a significant reduction of exposure of p2p users, in particular children, to harmful content.

- Paedophile keywords

One may identify three different kinds of paedophile keywords: the basic ones that anyone would think of to find paedophile content, more specific ones known mainly by people with experience in handling paedophile content (like paedophiles themselves and law enforcement personnel), and hidden, short-term keywords known only by small groups of people (who exchange these keywords in chat systems or other interpersonal communications). Identifying paedophile keywords therefore is a key issue for filtering, as well as law enforcement. It is also necessary to send appropriate queries to p2p systems for the measurement of paedophile activity. An objective of the project therefore is to use huge amount of recorded queries and file names to uncover such keywords, including hidden ones that serve only for short periods of time.

This will result in a dynamic list of paedophile keywords, that will evolve during time, which we plan to send to law enforcement authorities and a restricted set of other relevant institutions5. This list will contain detailed information on the keywords, like their frequency during time, the other keywords with which they appear, their date of first appearance, etc.

- Improved knowledge of paedophile activity

Our objective here is to give an accurate and detailed view of what is going on concerning paedophile activity in currently running p2p systems. This includes the evaluation of the number of files/users involved, the identification of various kinds of files/users, and several other basic statistics, together with their evolution during time. We also seek more subtle information, like studies of how users develop an interest in paedophile content, global maps of paedophile contents, including their nested community structures, and methods to make the difference between people that probably download paedophile content accidentally and people that focus on such contents.

The objective here therefore is to obtain rigorous and deep enlightenment on p2p paedophile activity, which will lead to the publication of detailed reports on each aspect, as well as both technical and general public synthesis reports at the end of the project. We want to change the current situation into a situation in which we have a precise knowledge of paedophile activity in p2p systems.

Project website:
http://antipaedo.lip6.fr/

Contact

Mr Matthieu Latapy
Centre National de la Recherche Scientifique (CNRS)
3, rue Michel-Ange
75794 Paris Cedex 16
France

Tel.: +33 1 4427 5617

Fax: +33 1 4427 6849
email: latapy@liafa.jussieu.fr