Technical report on

Maps of paedophile activity

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

Matthieu Latapy[1], Clémence Magnien, Raphaël Fournier and Massoud Seifi

Abstract

As policy-making and law enforcement institutions generally operate at the national level, or at least at a regional level (Europe for instance), we studied geolocated recordings available in a large dataset obtained by a measurement of keyword-based queries submitted to a large P2P server. We observed that the fractions of paedophile queries may be orders of magnitude larger in some countries than in others. We also investigated the possibility of building topical maps of activity using *community detection* methods, which cluster paedophile files in a small number of communities. Obtained maps show that some communities do indeed have a high fraction of paedophile files (compared to others), but the relations between them are still unclear.

1 Introduction.

Paedophile activity may be observed by inspecting queries entered by users in a P2P system as a whole. However, in most cases, this means that observed users are worldwide distributed. This raises two key issues.

First, policy-making and law enforcement institutions generally operate at the national level, or at least at a regional level (Europe for instance). Therefore, information on the *location* of users is a key resource, and location-aware analysis is necessary: if an important paedophile activity is detected in a country which pays only little attention to this problem, then little may be done. On the contrary, one may consider this as a motivation for changing the policy in such countries. Quantifying paedophile activity at the country level may also help in assessing the effectiveness and impact of national law-enforcement initiatives, and thus identify best practices.

Another important motivation is more technical: as our ability to detect and quantify paedophile activity is directly related to our ability to understand observed queries, language issues may bias our estimations. Likewise, heterogeneity in character encodings makes the identification of paedophile activity geo-dependent. By restricting analysis to a set of target countries, one may make automatic detection of paedophile queries more reliable.

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

The first objective of this work is to explore geo-located recordings of P2P queries in order to gain insight on the geographical distribution of observed paedophile activity, with a focus on Europe. We present in Section 2 the data we use, in Section 3 the statistical analysis we conduct, and in Section 4 the maps we obtain at European level.

We also studied topical maps of activity using community detection methods, which group similar files together. The obtained maps are presented in Section 5.

2 Data.

For this study, we used the second server measurement performed in the project. It consists in a set of queries received by a large *eDonkey* server from August 29th 2009 to October 14th 2009. The administrator of this server activated the log feature of the server software, thus recording information on keyword-based queries it received. This information consists of lines containing a query each, under the form:
time-stamp anonymised-IP-address keyword-1 keyword-2 … keyword-n country-code.

In order to protect user privacy, and in conformance with French law, IP addresses were anonymised. Geolocation is not possible from this anonymised data, therefore we had to perform it on-the-fly [2], before storing the log data.

We finally obtained 54,274,002 queries from 214 different countries (see Figure 1 below for the distribution of queries among countries). This dataset is one of the largest ever collected on queries entered in a P2P system [1].

Among these queries, some were entered by users seeking files with paedophile content. Automatically detecting such queries is necessary because of the huge amount of data to process. This is however a challenge in itself. We used here the method described in [6], which is by far the most accurate currently available [3] Moreover, one may expect that its success rate is independent of the country. This is not always true, though, as we will see in Section 3 below.

Finally, we consider in this study a set of 54 274 002 queries collected during more than 6 weeks. In this set, we identify 77 548 paedophile queries.

3 Statistics.

Let us first notice that, because of local specificities like available technologies and languages, P2P users are not uniformly distributed among countries. Going further, it is known that the top P2P systems are not the same in regions like Asia, Europe and USA. In addition, the population size and the internet-enabled population size may vary a lot

[2] The geolocation was performed with the GeoIP database from MaxMind, `http://www.maxmind.com/app/country`

[3] It does not provide the actual number of paedophile queries, as it has false positives and negatives. In order to obtain a more accurate estimate of the actual number, one may apply techniques developed in [6, 5]. Since this would not change the *relative* ratios of paedophile queries in different countries, we kept here the original values.

[7]. As a consequence, our dataset may contain very heterogeneous numbers of queries captured for each country.

This is confirmed by Figure 1, which displays the distribution of the number of queries observed from each country. This distribution is heterogeneous, as the number of queries by country varies over several orders of magnitude.

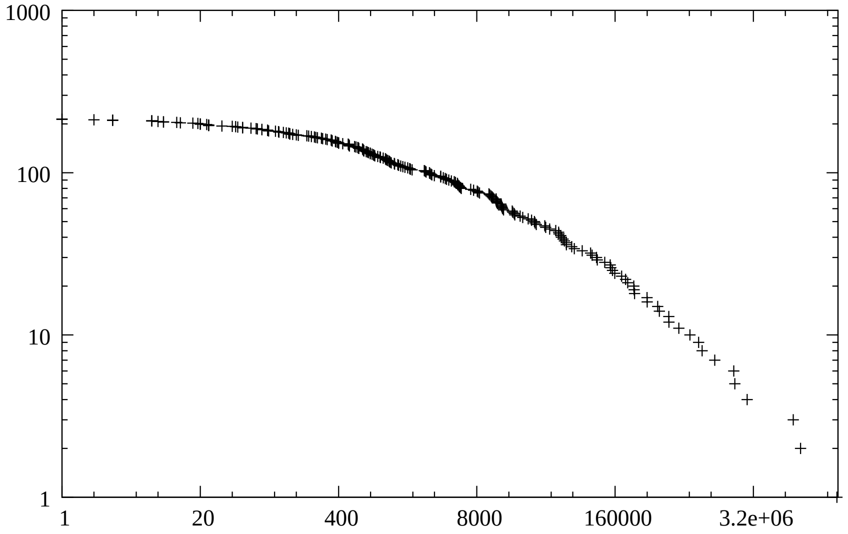

Figure 1: Complementary cumulative distribution of the number of queries observed from each country: for each value x on the horizontal axis, we plot the number y of countries for which we observed at least x queries. The plot is displayed in log-log scale in order to exhibit its heterogeneous nature (the observed values span several orders of magnitude).

In such a context, statistics on paedophile activity make sense only in countries from which enough queries are. For instance, Greenland, with few queries and a relatively high number of paedophile queries, has the overall highest paedophile ratio. This is however not significant, as the total number of queries from this country is not high enough to make relevant statistics. To avoid this kind of bias, we restrict ourselves to the 30 countries from which we observed at least 100 000 queries. This accounts for 97,05 % of our entire set of queries. We present the statistics for these countries in Table 1.

Notice that, as long as we capture many queries for a given country, even if a small number have a paedophile nature, the obtained fraction of paedophile activity is statistically significant (and it is low).

However, such cases may also occur when the automatic paedophile query detection system fails because of country-specific character encodings or language. This may for instance be the case of Korea (KR), for which we collect many queries but detect only a few as paedophile, see Table 1.

Finally, we provide in Table 2 the statistics ordered by decreasing fraction of paedophile queries observed, for countries in which this is statistically relevant.

country	# queries	# paedo	ratio
IT	19569361	15426	0.08 %
ES	8881405	5177	0.06 %
FR	7583815	8059	0.11 %
BR	2795090	4849	0.17 %
IL	2139697	2618	0.12 %
DE	2093106	11238	0.54 %
KR	1386799	336	0.02 %
US	1053183	6184	0.59 %
PL	975170	1178	0.12 %
AR	810466	1465	0.18 %
CN	635392	337	0.05 %
PT	513327	434	0.08 %
IE	511185	54	0.01 %
TW	417893	138	0.03 %
BE	402565	646	0.16 %
CH	320054	1710	0.53 %
GB	319386	1698	0.53 %
NL	243646	1131	0.46 %
CA	241460	1233	0.51 %
SI	239572	167	0.07 %
MX	210504	1098	0.52 %
RU	200958	2712	1.35 %
AT	184248	977	0.53 %
DK	159041	468	0.29 %
GR	150984	536	0.36 %
TR	145714	368	0.25 %
CL	143785	299	0.21 %
JP	127915	178	0.14 %
VE	108758	380	0.35 %
AU	106882	401	0.38 %

Table 1: Number of queries received, number of paedophile queries, and fraction of paedophile queries (*i.e.* queries detected by the filter), for each country for which at least 100 000 queries were observed in our dataset (ordered by decreasing number of queries by country).

4 Europe maps.

One may visualise the statistics presented in previous section in a more intuitive and appealing way by drawing maps in which the colour of each country reflects its statistics. We display in Figures 2 to 4 the maps of the Europe area coloured with respect to the number of observed queries, the number of observed paedophile queries, and the fraction

country	# queries	# paedo	ratio
RU	200958	2712	1.35 %
US	1053183	6184	0.59 %
DE	2093106	11238	0.54 %
CH	320054	1710	0.53 %
GB	319386	1698	0.53 %
AT	184248	977	0.53 %
MX	210504	1098	0.52 %
CA	241460	1233	0.51 %
NL	243646	1131	0.46 %
AU	106882	401	0.38 %
GR	150984	536	0.36 %
VE	108758	380	0.35 %
DK	159041	468	0.29 %
TR	145714	368	0.25 %
CL	143785	299	0.21 %
AR	810466	1465	0.18 %
BR	2795090	4849	0.17 %
BE	402565	646	0.16 %
JP	127915	178	0.14 %
IL	2139697	2618	0.12 %
PL	975170	1178	0.12 %
FR	7583815	8059	0.11 %
PT	513327	434	0.08 %
IT	19569361	15426	0.08 %
SI	239572	167	0.07 %
ES	8881405	5177	0.06 %
CN	635392	337	0.05 %
TW	417893	138	0.03 %
KR	1386799	336	0.02 %
IE	511185	54	0.01 %

Table 2: Number of queries received, number of paedophile queries, and fraction of paedophile queries (*i.e.* queries detected by the filter), for each country for which at least 100 000 queries were observed in our dataset (ordered by decreasing fraction of paedophile queries).

of observed paedophile queries, respectively. We represented the countries for which these numbers were not significant or not available in white.

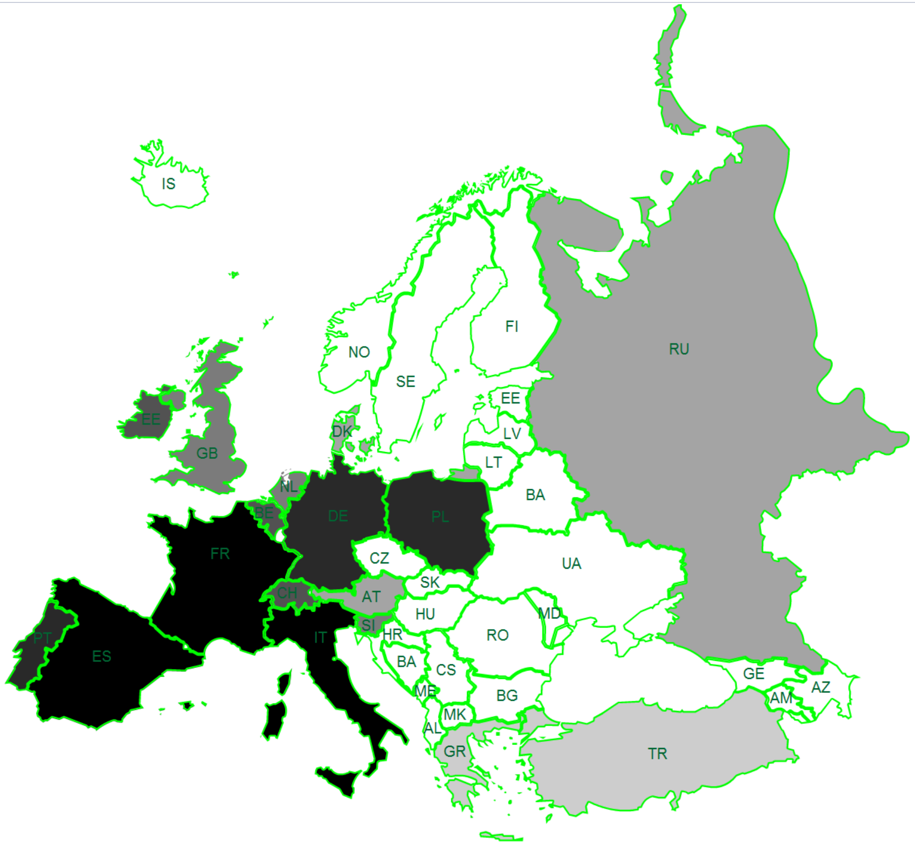

Figure 2: Map of the European area with each country coloured according to the number of captured queries entered by users in these countries. Dark grey indicates countries with many queries, light grey indicates countries with few queries.

5 Community maps.

We also investigated maps of activity using *community detection* methods. The idea was to study observed user interests, with the intuition that if a given user provides two different files, then these files are somewhat related. They are even more related if a large number of users provide them both. Using the large amount of data collected during the project, it is then possible to use this approach to build a graph representing similarities between files, according to user interests. A *community detection* method is then able to extract groups of files which are similar to each other.

In particular, known paedophile (resp. pornographic) files should be grouped in a small number of communities, and it should then be possible to draw a map of these communities, and their links to other communities.

We computed such a community structure, and used it for creating our content rating and fake detection system. The data used was the information about which users provided which files in the first server measurement performed in the project. The detail of the

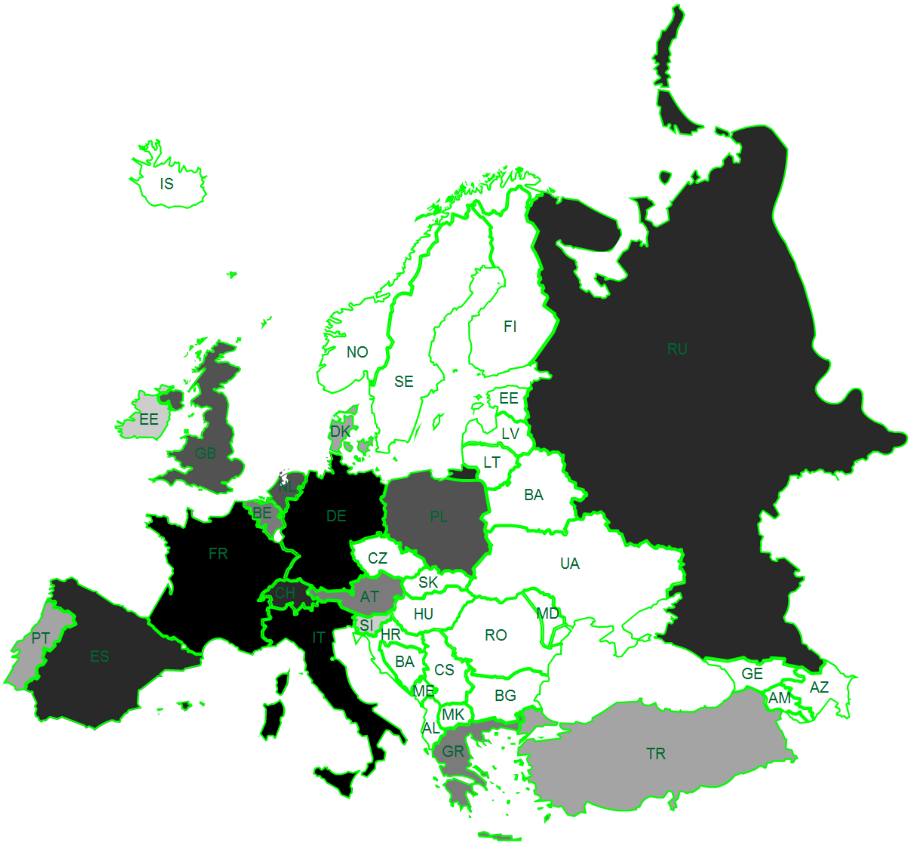

Figure 3: Map of the European area with each country coloured according to the number of detected paedophile queries entered by users in these countries. Dark grey indicates countries with many paedophile queries, light grey indicates countries with few paedophile queries.

graph and the community detection method are provided in the technical report on the content rating and fake detection system [3, 2]. We showed that this structure is relevant for grouping paedophile (resp. pornographic) files together. However, to correctly assess the nature of a file, we had to take into account the whole *hierarchical* structure of the communities: large communities made of smaller sub-communities, themselves containing even smaller sub-sub-communities, and so on. Indeed, if we study the largest communities, which are the final result of the community detection method, and therefore are the most relevant, we find that they are often very large, and that a small number of files (such as paedophile files) is diluted in these communities. Conversely, the smallest communities are the ones that are the most similar to the nodes they contain, and therefore contain nodes which are very similar to each other. However, there is a very large number of such communities, which is not appropriate for a visual representation; moreover, they are often very small and therefore not very relevant: it is not relevant for instance to include a community containing only two paedophile files in a map.

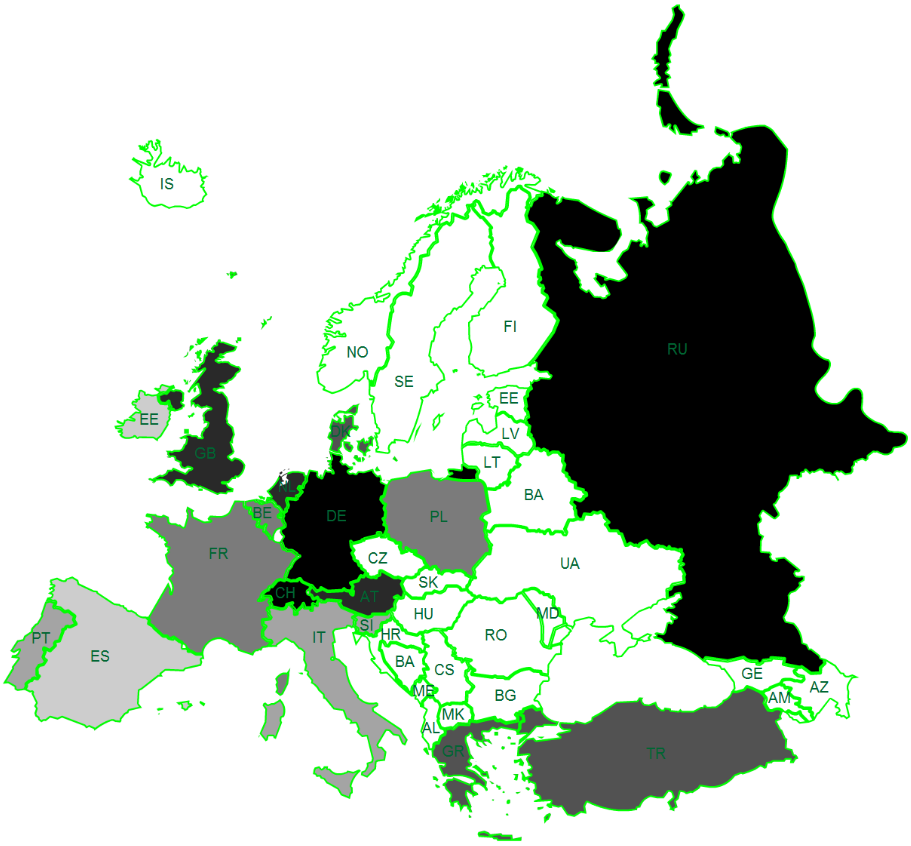

Figure 4: Map of the European area with each country coloured according to the fraction of detected paedophile queries with respect to the total number of queries entered by users in these countries. Dark grey indicates countries with a high fraction, light grey indicates countries with a low fraction.

Though they did not prove as enlightening as we had hoped, the maps built using community structures still are an interesting way to examine the data. Figure 5 (resp. Figure 6) presents the obtained map coloured as a function of paedophile (resp. pornographic) content: each community is coloured according to the ratio of the number of paedophile (resp. pornographic) files it contains, with respect to the total number of files with a name it contains.

The maps presented in these figures are built after a manual inspection of the community structure, which allowed to keep the most relevant communities and links between them.

They visually illustrate the fact that our community detection algorithm succeeds in identifying groups of files which contain significantly more paedophile files than others (red nodes). Limited relationships exist between these communities, arguing for the existence of quite separate topics, see also [4].

Conversely, we observe relationships between paedophile and pornographic communi-

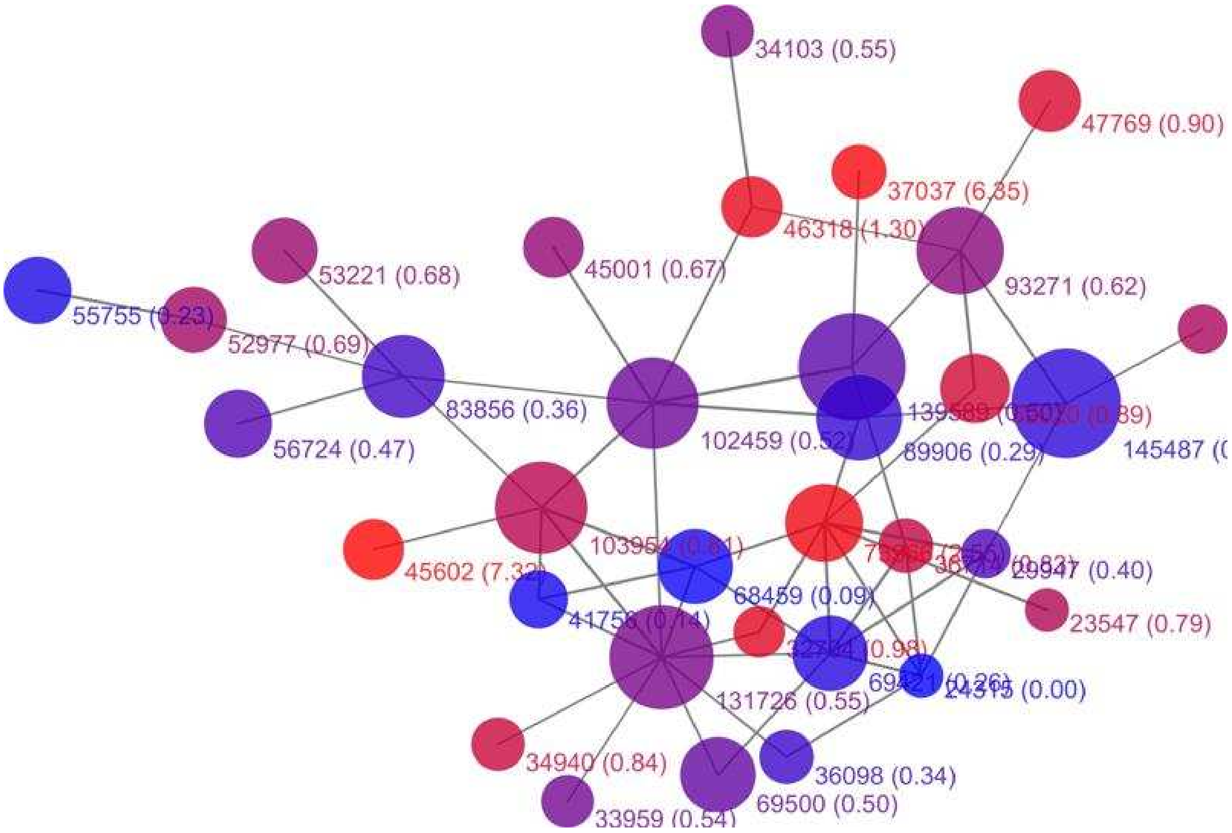

Figure 5: Community map of paedophile activity. Each community is labelled with the number of nodes it contains and the fraction of paedophile files it contains divided by the overall average fraction of paedophile files.

ties. Interestingly, although most paedophile communities are also strong pornographic communities, others are not. This confirms that paedophile activity is quite distinct from general pornographic activity in such systems.

Acknowledgements. We warmly thank the administrator of the `peerates.net` *eDonkey* server for his help in collecting data. 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] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Journal of Statstical Mechanics: Theory and Experiment*, page P10008, 2008.

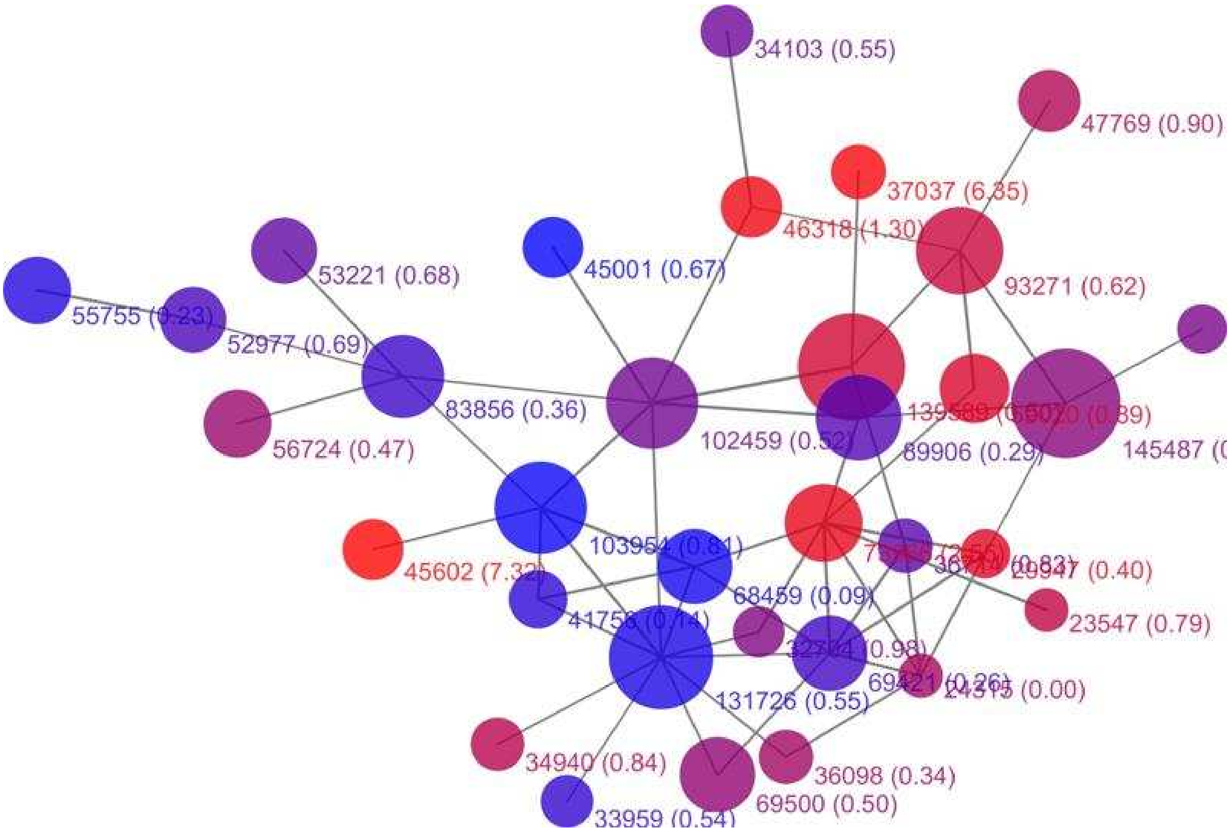

Figure 6: Community map of pornographic activity. Each community is labelled with the number of nodes it contains and the fraction of paedophile files it contains divided by the overall average fraction of paedophile files.

[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] Sean Hammond, Ethel Quayle, Jurek Kirakowski, Elaine O'Halloran, and Freda Wynne. Technical report on *An Examination of Problematic Paraphilic use of Peer to Peer Facilites*. In *International Conference on Advances in the Analysis of Online Paedophile Activity*, 2009. Measurement and Analysis of P2P Activity Against Paedophile Content Project.

[5] 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.

[6] 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.

[7] International Telecommunication Union. The world in 2009: Ict facts and figures. http://www.itu.int/ITU-D/ict/material/Telecom09_flyer.pdf.

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