Temporal Patterns of Pedophile Activity in a P2P Network: First Insights about User Profiles from Big Data

Raphaël Fournier, Matthieu Latapy

International Journal of Internet Science **ARTICLE IN PRESS**
2015, 10 (1), in press ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

Temporal Patterns of Pedophile Activity in a P2P Network: First Insights about User Profiles from Big Data

Raphaël Fournier[1] , Matthieu Latapy[2]

[1]Laboratoire CEDRIC, CNAM Paris, France; [2]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris, France

Abstract: Recent studies have shown that child abuse material is shared through peer-to-peer (P2P) networks, which allow users to exchange files without a central server. Obtaining knowledge on the extent of this activity has major consequences for child protection, policy making and Internet regulation. Previous works have developed tools and analyses to provide overall figures in temporally-limited measurements. Offenders' behavior is mostly studied through small-scale interviews and there is few information on the times at which they engage in such activity. Here we show that the proportion of search-engine queries for pedophile content gradually has grown by a factor of almost 3 in three years. We also find that during the day, certain hours are, on average, privileged by seekers. Our results demonstrate that P2P networks are actively used to search for pedophile content and we find new and large-scale results on pedophile offenders' profile, indicating that a substantial proportion is well-integrated into family life and professional work activities.

Keywords: P2P networks, eDonkey, pedophile activity, user profile, Big Data

Introduction

Online pedophile activity is a major concern to the society, being a direct consequence of children-targeting crimes. Understanding this online phenomenon may have implications in designing effective methods to arrest child offenders (Gupta, Kumaraguru, & Sureka, 2012; Liberatore, Levine, & Shields, 2010; Penna, Clark, & Mohay, 2005). The presence of pedopornography in P2P networks is also used as an argument in prosecutions and in debates to regulate the Internet (Roubaty, 2007). However, current established knowledge regarding the extent of the activity remains limited (Latapy, Magnien, & Fournier, 2013; Rutgaizer, Shavitt, Vertman, & Zilberman, 2012).

After collecting search-engine queries in a peer-to-peer (P2P) network at an unprecedented scale, we perform a series of detailed analyses of the temporal patterns of pedophile activity. We study these queries for child abuse material following two different and complementary approaches. We first address the key question of whether pedophile activity increased, decreased or stayed stable over the course of three years from 2009 until 2012. We then turn to the study of daily dynamics of pedophile activity. We examine in particular hours in the day at which users seek child abuse material. These may or may not be compatible with a professional and/or family life, thus giving insight into pedophile users' social integration.

Address correspondence to Raphaël Fournier, Laboratoire CEDRIC, CNAM, 292 Rue St Martin FR-75141 Paris Cedex, France. Phone: (+33) 158808635, e-mail: fournier@cnam.fr.

Related Work

Analysing search queries is an active scientific topic for several communities, such as information retrieval, psychology, sociology and marketing (e.g., Klemm, Lindemann, Vernon, & Waldhorst, 2004; Nguyen, Jia, Yee, & Frieder, 2007; Spink, Koricich, Jansen, & Cole, 2004). The growing popularity of P2P networks since their outbreak in 1999 has brought some new opportunities to collect some data and obtain insight into specific activities through query analysis. A dedicated survey was presented in Gayo-Avello (2009) in 2009. Leveraging such analyses to increase knowledge on pedophile activity has been addressed during the last decade by a few authors.

In 2006, a team of computer scientists presented a first attempt to gauge the presence of sexual violence on some different P2P networks (Hughes, Walkerdine, Coulson, & Gibson, 2006). Their methodology was more active than ours: they constructed a list of specific queries and analysed the search engine answers. They obtain results much higher in magnitude (approximately 1.3% of "sexually violent" queries), but their scope was significantly broader than pedophile behavior. This work was pioneering on the subject, showing that there were some sub-communities of users dedicated to sharing "illegal pornography" (not only child abuse material), with probably little interaction between these communities and the rest of the network.

Another group of researchers presented in 2011 some adequate tools and metrics to accurately estimate the fraction of pedophile activity in P2P queries (Latapy et al., 2013). They concluded that there were 0.25% pedophile queries (0.2% users) in their datasets. They had two large-scale datasets (of 10 and 28 weeks), while here we use only one, but of unprecedented duration (144 weeks). They provided the research community with a precise global insight into the extent of the *demand* for such content (approximately one query every 33 seconds was "pedophile"). Their methodology and results were used as a reference by another team which performed an extensive study of a prominent BitTorrent portal (Rutgaizer et al., 2012). This latter work relied on a 453-million queries dataset, complemented by a file-based dataset to study the download activity (with 1.3 million files, but only 5 of them being pedophile material). A focus on the frequency of pedophile keywords was presented as well as a detailed study of the associations of such keywords (with other pedophile keywords or with common words). This deepens the previous analysis, aiming at a finer granularity, focusing on keywords associations rather than queries. Authors briefly evoke an analysis of the times at which pedophiles queries were submitted but do not detail nor their methods nor their conclusions. However, we think that such an indicator could improve knowledge on pedophile users' behavior, at a large scale, which is what we propose in the second part of our study.

Recently, the Gnutella network was studied extensively (in a one-year long experiment) and compared to *eDonkey* (Hurley et al., 2013). Authors partnered with law enforcement personnel to collect data on child pornography trafficking. They identified a list of 384,000 *files of interest* and performed a detailed analysis of their availability (89% of the files shared on *eDonkey* are also available on Gnutella, 30% of the files are available for more than 10 days). They mostly focused on finding optimal strategies to remove the minimal number of peers from the network to reduce the number of available pedophile-related files. They found some particularly active subgroups of users that should be targeted. They greatly helped to understand the *supply* of child abuse material: how many users are involved in the distribution of files, what is their importance in the network, etc.

These works have studied various aspects of pedophile activity in P2P networks: queries' global frequencies, keywords associations, file sharing, communities of users. The work by Hurley et al. (2013) had a clear law-enforcement goal ("which peers should be targeted to limit the distribution of files?"), while others' works were centered on computer networks' usage. They were executed at different time scales and in different years, which is why focusing on temporal patterns is relevant. Long-term trends may tie together those previous analyses. And studying daily patterns draws the attention on individual activity rhythms, which contributes to closing a knowledge gap about the profile of pedophile content searchers.

Data and Methodology

We collected our data on a large *eDonkey* server from June 2009 to April 2012 (spanning approximately 147 weeks). *EDonkey* (Heckmann, Bock, Mauthe, & Steinmetz, 2004) is a P2P network relying on a handful of servers which index available files. Users send keyword-based queries and a server provides them with a list of files matching their keywords. Our server was of medium importance in the network and it was located in France. We activated the log capability embedded in the server software, which records timestamped keyword-based queries and the IP addresses of the senders. Geolocation of the IP address was performed at recording time, before anonymisation. A subset of this dataset was used by Latapy et al. (2013). The final dataset contains 1,290,377,956

queries, from 82,264,897 different IP addresses, in 24,544 hours. It constitutes, to the best of our knowledge, the largest set of P2P queries ever recorded.

In order to study pedophile activity, we need to spot pedophile-related queries among all the queries. Several languages are used, and the (pedophile-related) vocabulary contains various abbreviations, spelling mistakes or dedicated sequences of letters, making the task rather challenging. A pedophile query detection tool was developed by Latapy et al. (2013). After gathering knowledge of pedophile keywords with law enforcement personnel, they identified four categories of pedophile queries based on the keywords they include. The queries in the first category contains at least one domain-specific keyword, *i.e.* dedicated to pedophile related queries. Other categories contain various combinations of keywords, which belong to several semantically-formed groups such as *sexual acts*, *infancy*, *family positions* or *ages*. The authors of this work wrote a text-matching tool, available online (Latapy, Magnien, & Fournier, 2011). Also using large-scale eDonkey search-engine data (collected on another server), they estimated the volume and fraction of pedophile queries and users that were involved in searching for child abuse material. Their tool, validated by human experts in the field, has very good precision/recall (98.5% precision, 74% recall) results and our analyses rely on its output to decide which query is pedophile and which is not.

Long-Term Evolution

Figure 1 presents the number of all queries observed per week, not just the pedophile ones. There are on average 8.7 million queries each week, which is a little more than 14 per second. The maximum of 11,008,609 is reached during week 71 (November 2010) and, except for the first week when the server was starting, the minimum of 5,332,788 is obtained at week 54 (July 2010), during which the server has experienced a 20-hour service interruption. There were between 7.3 million and 10.0 million queries for 90% of all weeks. The rolling mean, presented with 5-week windows on Figure 1a, helps in obtaining a less noisy value. The autocorrelation plot on Figure 1b is used to check for the randomness of the time-series: if random, autocorrelation values should be near zero for all time-lag separations; if non-random, then one or more of the autocorrelations will be significantly non-zero.

There are small weekly variations, but the rolling mean shows a smoother plot, with trends spanning several months. The autocorrelation plot has statistically significant values for a lag close to 52, which shows that there is also a yearly (52 week-period) trend. From 2009 to 2012, the number of queries remains stable between mid-October and mid-May, then decreases until the end of July, before increasing between August and October. There are large differences between the weeks, with several million queries between summer 2010 (5.3 millions) and November 2010 (11.1 millions), most likely resulting from human rhythms variations, induced by work and/or school holidays. Yet, the weekly activity remains within the range of [8,500,000; 9,500,000] for more than 70% of the cases. Events like the French HADOPI law (Légifrance, 2012), adopted by the end of 2009, with the goal of ending exchanges of copyrighted contents on P2P systems, have had limited impact on P2P traffic. The shutdown at the beginning of 2012 of MegaUpload, a prominent website distributing copyrighted material, had no noticeable impact on the numbers presented in Figure 1.

The long-term evolution of the fraction of pedophile queries, displayed in Figure 2 has a very different shape. It starts below 0.2 percent mid-2009, then increases during the first months of 2010 until the first semester of 2011 and then remains at around 0.45 percent.

Figure 2 shows that the number of pedophile queries increased by a factor of almost 3 between 2009 and 2012. This large increase may be due to either an increase in the number of pedophile queries per user or to more users being interested in this topic. In order to know if there had been more users in the network, we plot in Figure 3 the evolution of the fraction of pedophile IP addresses, with the same time-scale as before (weekly aggregate). Notice that using IP addresses may lead to slightly overestimating the fraction of pedophile users. Queries are grouped by IP addresses and the group is said to correspond to a “pedophile user” if *at least one* of them is pedophile. An error on a single non-pedophile query results in a user entering the “pedophile” category. However, using one-week time windows reduces this phenomenon (compared to a larger time scale): since groups are weekly based, an erroneously detected pedophile user in week n is no longer a pedophile user at the beginning of week $n + 1$. A more complete discussion on this matter was developed in the study of Latapy et al. (2013). However, this bias should be the same every week and should not affect the *evolution* of the fraction, which is what we need.

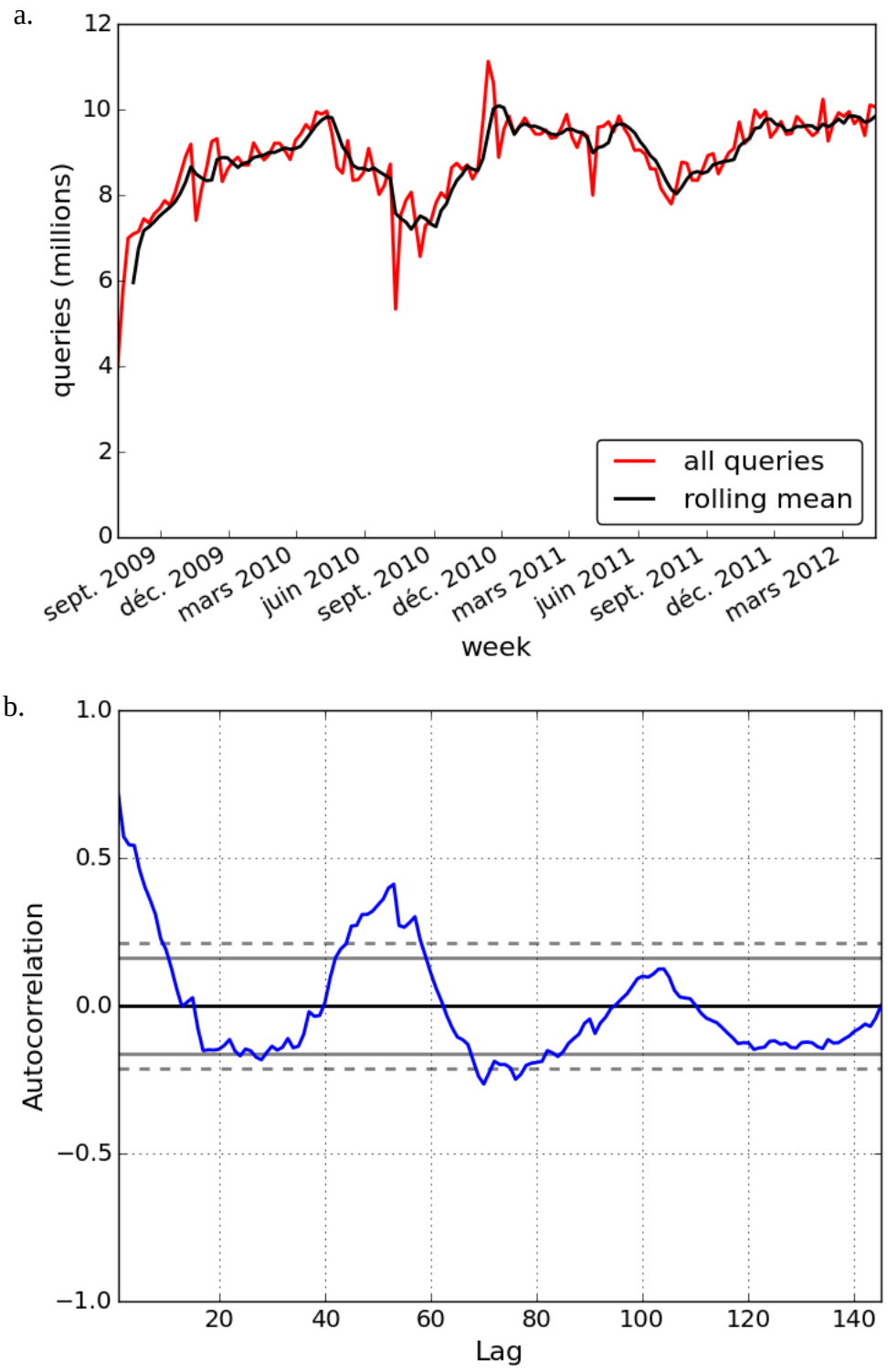

Figure 1. Number of queries per week from July 2009 to April 2012. (a) Total number of queries per week from July 2009 to April 2012, and rolling mean (five-week windows). (b) Autocorrelation plot for the number of queries per week. If a time series is non-random then one or more of the auto correlations will be significantly non-zero. The horizontal lines displayed in the plot correspond to 95% and 99% confidence bands, the dashed one is the 99% confidence band.

The overall number of IP addresses entering the system each week is rather stable, with approximately 15 million users. The fraction of IP addresses sending pedophile queries is close to 0.25% until the end of 2009, then increases slowly until mid-2011 up to 0.8%. It then slowly decreases towards 0.7% by the end of the measurement. The general trend is therefore an important growth between 2009 and 2012. The increasing fraction of pedophiles queries is due to more IP addresses submitting them, either because new pedophile users enter in the system or because former users become inclined to search more often on this topic.

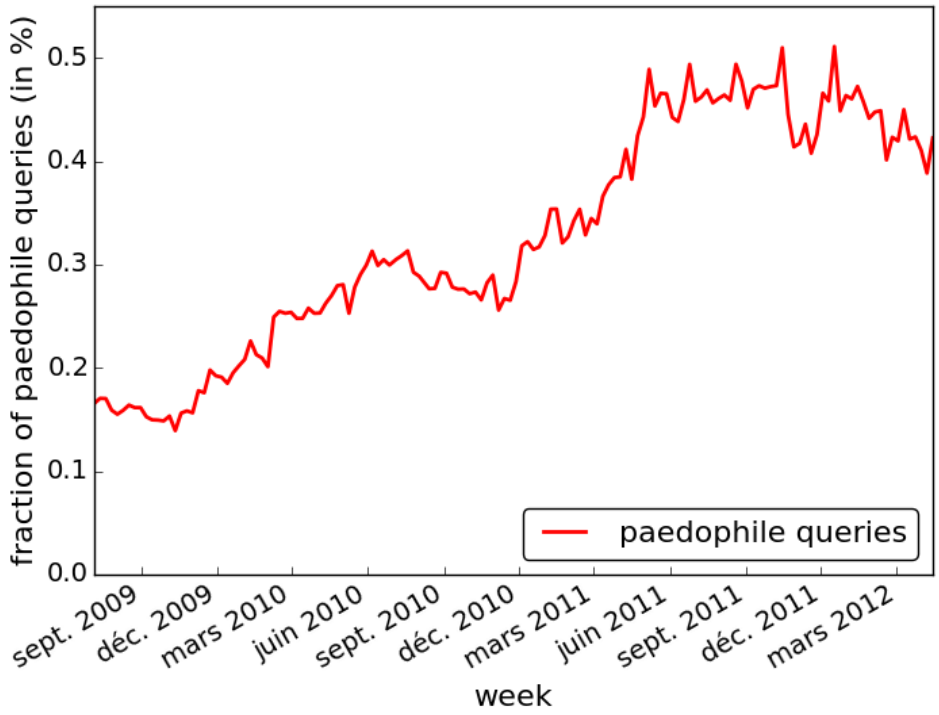

Figure 2. Fraction of pedophile queries per week, from 2009 to 2012.

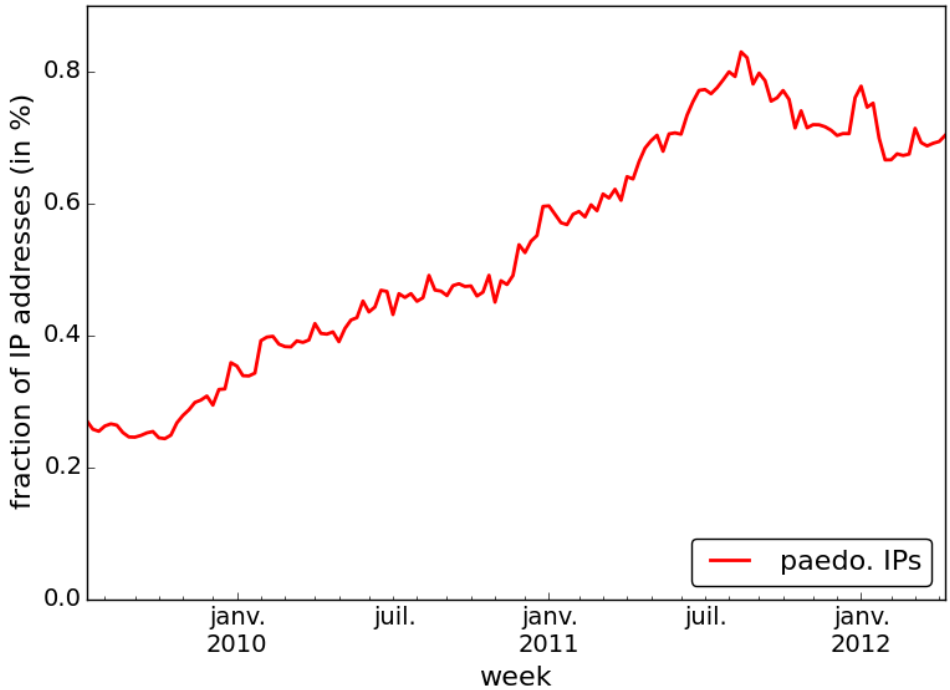

Figure 3. Fraction of pedophile IP addresses among all addresses, per week between 2009 and 2012.

Daily Dynamics

In this section, we study the daily dynamics of pedophile activity in order to shed light on the social status of involved users. We assume that users try to hide their involvement in child-abuse material seeking and so pedophile queries sent during work hours or the typical family time may indicate users with no job and/or family. We aggregated data by hour and computed some statistics, such as: average total number of queries, average fraction of pedophile queries, and average fraction of pedophile IP addresses. Our dataset spans 24,544 hours so for each hour, the average is computed on more than 1,020 data points. To the best of our knowledge, this is the first time such a study is performed for analysing pedophile behavior.

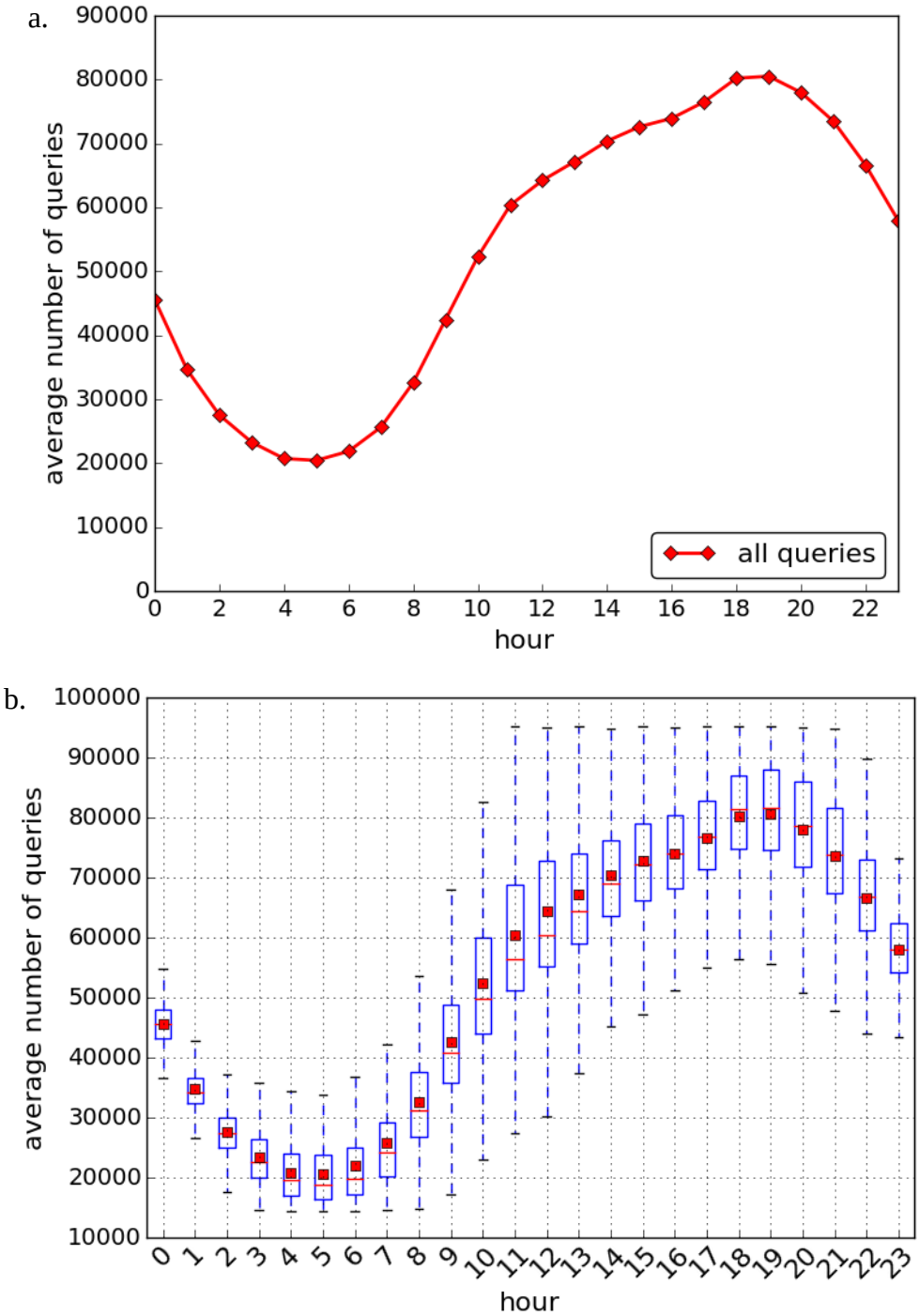

Figure 4. Average total number of queries per hour. (a) Average total number of all queries per hour. (b) Box-plot of the same data, with 25% and 75% quartile, median (horizontal red line), mean (red square) and 1.5 interquartile range.

Figure 4a presents the mean number of all (pedophile and non-pedophile) queries received by the server per hour, and the box-plot on Figure 4b gives more information on the distribution for each aggregated hour. The server time is the reference, and it is based on the French time zone. In order to avoid statistical bias, we removed the 2% hours with the least queries (mostly due to temporary interruptions of the server).

The average number of queries varies much during the day: a minimum of 21,851 is observed at 5 AM and the number then increases until 7 PM, where it is at a maximum (with a mean of 80,518 queries). The pace of the growth is important from 5 AM to mid-day, lower afterwards. The decrease from 7 PM to 5 AM is rather constant.

The pattern we observe follows the circadian rhythm of Western Europe. Peak hours are between 5 PM and 10 PM, after the workday. Low-traffic hours are between 12 AM and 6 AM, corresponding to the night. Geolocation information indicates that most queries on the server come from Western Europe (Italy, France, Spain).

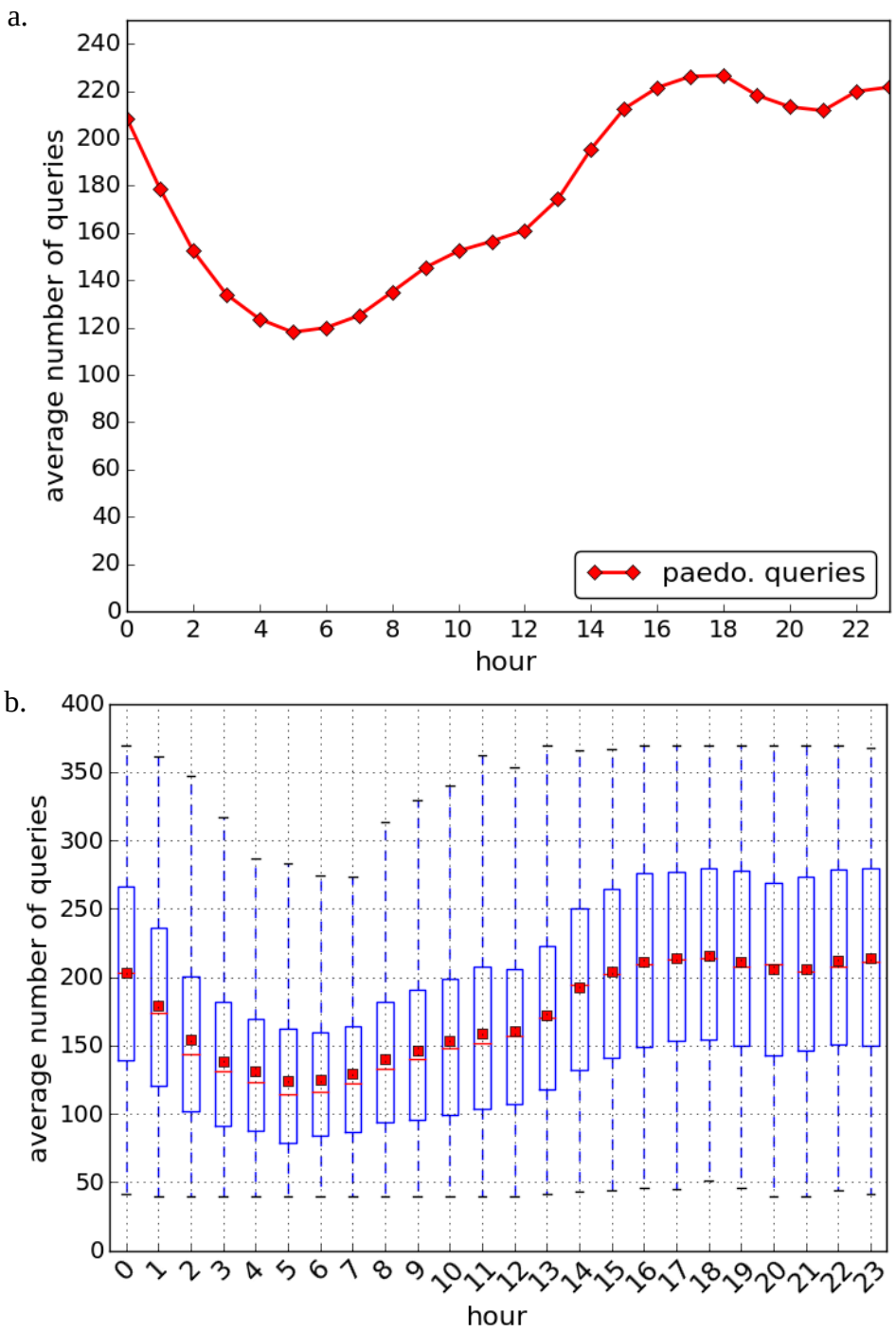

Figure 5. Average number of pedophile queries per hour. (a) Average number of pedophile queries per hour.(b) Box-plot of the same data, with 25% and 75% quartile, median (horizontal red line), mean (red square) and 1.5 inter-quartile range.

Figure 5a presents the number of pedophile queries per hour, and the associated box-plot is presented in Figure 5b. This plot is similar to the overall number of queries presented in Figure 4, with a pronounced day/night effect. The minimum value is reached at 5 AM (118.0 queries on average), the maximum value, of 226.7 queries, is reached at 6 PM. Much more importantly, amplitudes are different, with maximum/minimum ratio of 3.68 for the overall traffic, and only 1.92 for pedophile queries.

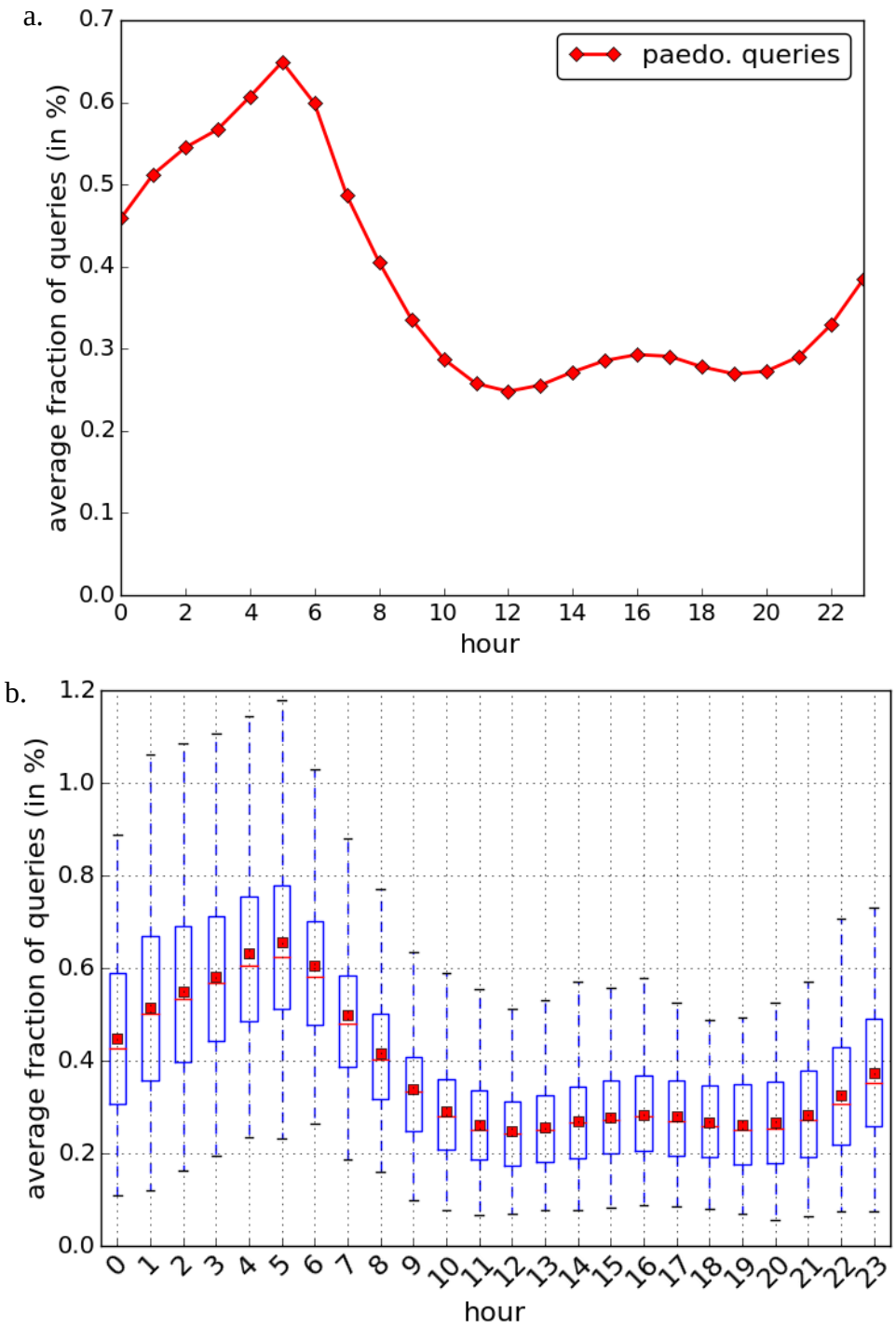

Figure 6. Average fraction of pedophile queries, out of all queries, per hour. (a) Average fraction of pedophile queries, out of all queries, per hour. (b) Box-plot of the same data, with 25% and 75% quartile, median (horizontal red line), mean (red square) and 1.5 inter-quartile range.

Figure 6a shows the average fraction of pedophile queries per hour and Figure 6b presents the associated box-plot. The value varies significantly; it peaks at 0.649% at 5 AM and reaches a minimum of 0.248% at 12 PM. It increases steadily from 10 PM to 5 AM, before decreasing from 5 AM to 12 PM. Then, it increases again, to a local maximum of 0.293% at 4 PM.

The notable peak around 5 AM and the global shape of the plot indicate that, even if pedophile users follow general rhythms of day/night activity, pedophile queries are over-represented by the end of the night. We also notice a local peak around 4 PM.

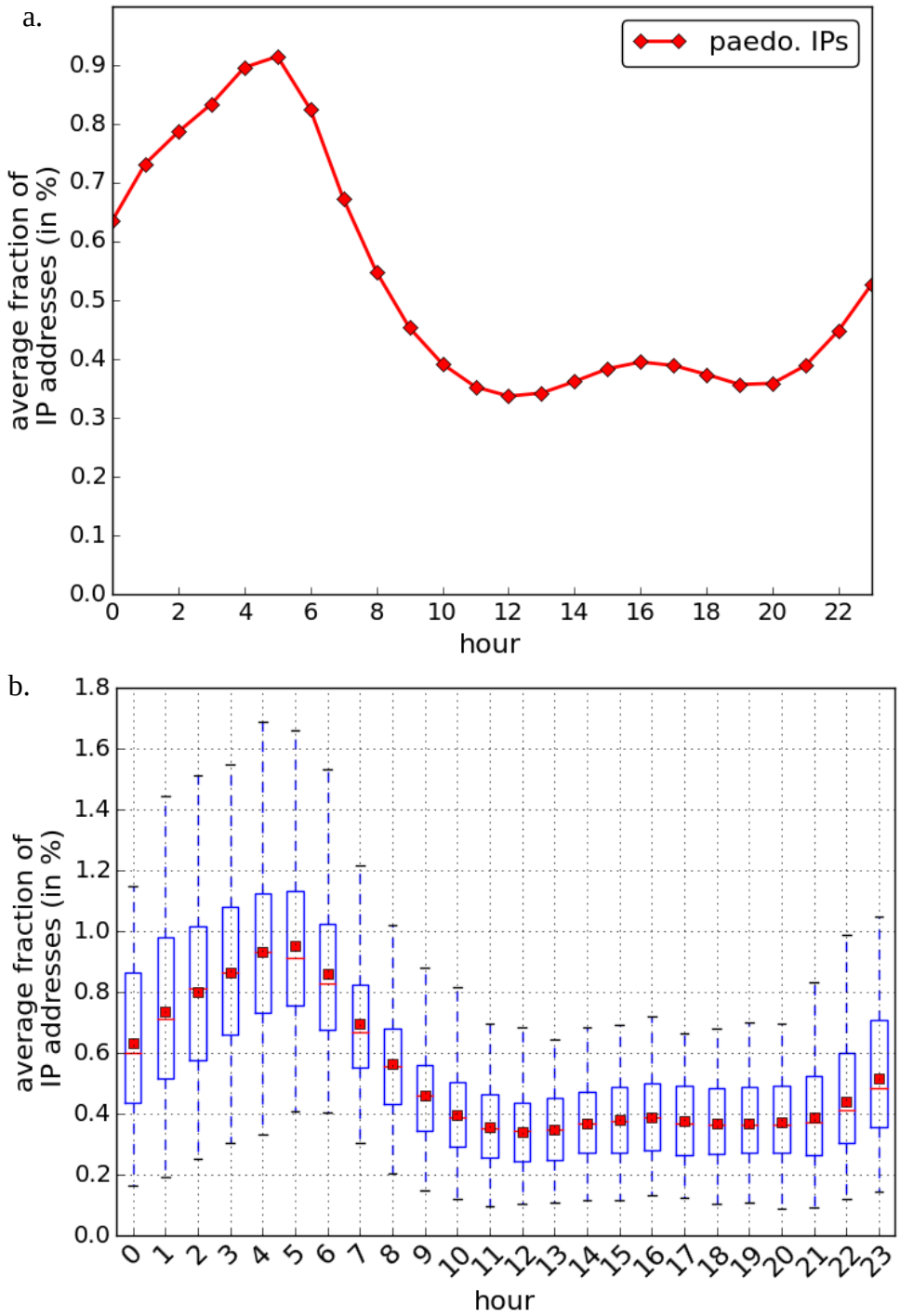

Figure 7. Average fraction of IP addresses sending pedophile queries, out of all IP addresses, per hour. (a) Average fraction of IP addresses sending pedophile queries, out of all IP addresses, per hour. (b) Box-plot of the same data, with 25% and 75% quartile, median (horizontal red line), mean (red square) and 1.5 inter-quartile range.

An explanation to the 5 AM peak could be the presence of some users submitting many more pedophile queries than others. However, Figure 7a shows that the fraction of IP addresses sending pedophile queries has an evolution similar to the one of the fraction of pedophile queries, which discards this hypothesis. This peak may therefore indicate a rather strong social integration of pedophile users, because it is compatible with typical workday hours (it would have been very different if the peak was located around 10 AM, 2 PM or 6 PM).

We aim at generating new hypotheses concerning pedophile users' behavior. In the light of Figure 6a, we propose the following:

H1. *A substantial proportion of pedophile content searchers is socially integrated in professional activities and seeks for content outside of typical working hours.*

Geolocalized Queries

We investigate this further below. We first use the geolocation of IP addresses to check if we observe the same results when restricting to a country (or, better, different countries with different time zones). To achieve this, we must choose countries for which the number of queries is high enough to ensure statistically significant results (many countries send only few queries per hour) and for which the detection tool is the most reliable (typically countries speaking English, French or other European languages). The countries should also present a significant time difference. For these reasons, we chose France (time zone UTC+1, UTC+2 during summer French language) and an aggregate of Argentina and Brazil (time zone UTC-3 for both and Latin languages). There are 5 hours of time difference between the chosen countries.

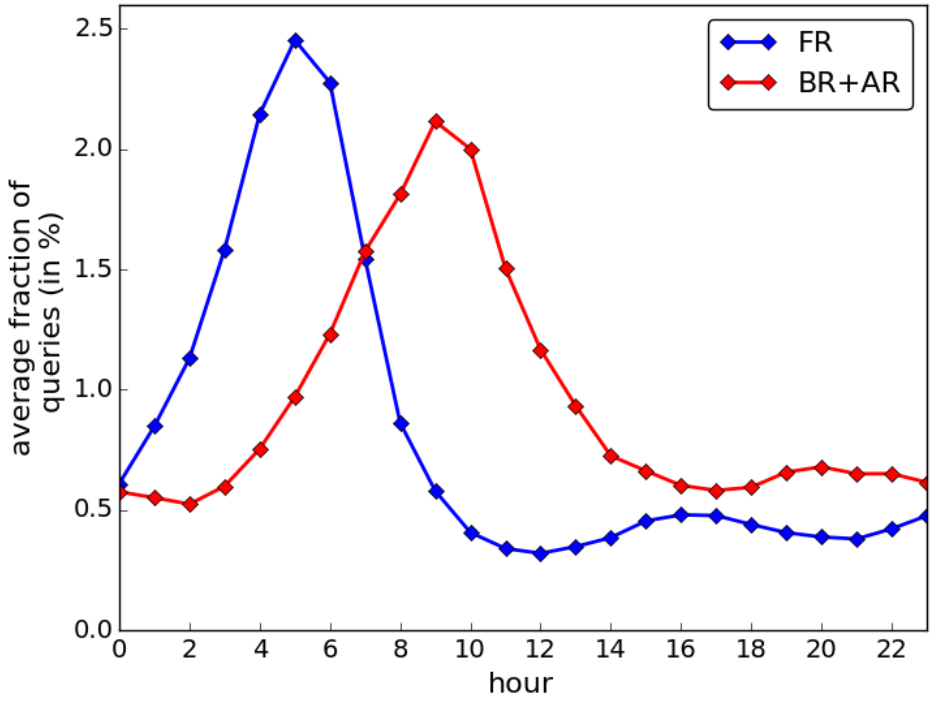

Figure 8. Average fractions of pedophile queries per hour, among queries from France or from Argentina and Brazil.

Figure 8 shows the average fraction of pedophile queries per hour for France and for Argentina+Brazil. The y-scale is higher than the value for the previous figure, mostly due to an increased amplitude for the peak around 5 PM. There are country variations for the fractions of pedophile queries, sensible to the performance of the tool on specific languages. Here, we are interested in the temporal evolution during the day, which is not affected by this language performance issue. The shapes of the curves are very similar with only a horizontal shift due to time zone differences, as expected. This confirms our previous observation: pedophile queries are over-represented in the traffic at the end of the night, between 4 and 6 AM (local time).

Comparison with Pornographic Queries

Now let us notice that users seeking pornographic content (not pedophile content) may also avoid doing so during work hours and typical family time. The behavior we observe for pedophile users may then be just a consequence of this more general behavior. To investigate this, we plot in Figure 9 the fraction of queries containing terms among “sex”, “porno”, “anal”, “xxx”, or “porn”, which are considered to be typical porn keywords. We use them as a proxy to observe the variations of pornographic activity in our dataset.

The fact that both fractions have the same order of magnitude is not significant because we limited the number of pornographic keywords. However, the shape of the plot for pornographic queries is similar to the one presented in Beitzel, Jensen, Chowdhury, Grossman, and Frieder (2004), which confirms that the chosen keywords provide relevant indications.

The bounds within which both fractions evolve are interesting. For pedophile activity, as already said, they are between 0.35% and 0.89%, whereas for pornography it is between 0.62% and 0.85%. This low amplitude indicates that pornography is well-distributed over the day, very different from pedopornography and its peak around 5 AM.

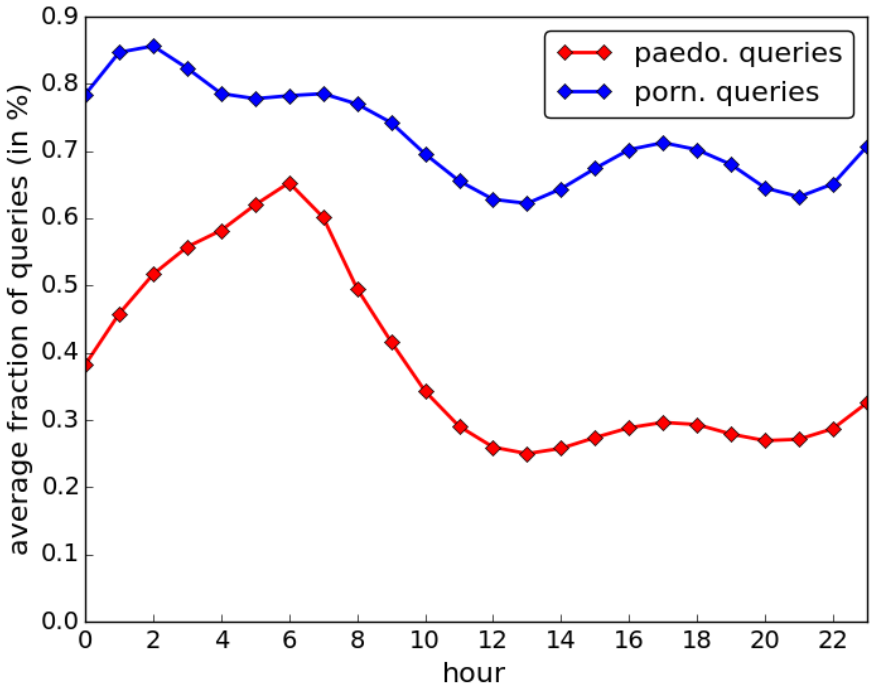

Figure 9. Average fraction of pedophile or pornographic queries, out of all queries, per hour.

However, the plots have very similar shapes from 9 AM to 11 PM: the values decrease until 12 PM, reach a local maximum around 5 PM, decrease until 8 PM and increase again. This indicates that users interested by the two topics have partially similar behavior for seeking content, and that the 5 PM peaks on each plot probably have similar causes. The local maximum at around 5 PM may indicate that at this time some users start searching for these kinds of content before family hours. After 5 PM, in many households, family life starts. We can then propose another hypothesis about pedophile users' behavior:

H2. *A substantial proportion of pedophile users is socially integrated in family activities and avoids family hours to search for such content.*

Hypotheses *H1* and *H2* are complementary and point in the same direction of a strong social integration for most pedophile users, regarding work and family life.

Conclusion

We have studied the temporal patterns of pedophile activity in a P2P system, relying on one of the largest datasets ever collected. We detected pedophile queries using the tool presented in the study of Latapy et al. (2013) and extended their analysis. We have obtained results with important consequences for the fight against online distribution of child abuse material. We observe that, from 2009 to 2012, there was a significant increase of the demand of pedophile material in *eDonkey*. Then, using hourly traffic aggregates, we show that a strong increase of pedophile activity is observed at 5 AM and a smaller one around 5 PM. This leads us to propose two hypotheses concerning pedophile users' social integration into family and work life, which appears to be rather strong for many of them.

Our contribution opens various directions for future work. The large dataset we use may be very valuable for research on query behavior in P2P networks, since it covers almost three whole years and contains more than a billion queries. The long term trends suggest that pedophile demand is increasing: it would be interesting to compare with other P2P networks (and other protocols), to check whether this is a general phenomenon or specific to *eDonkey*. The second part of our analysis, on the daily activity, calls for further sociological and clinical research to deepen the results.

Acknowledgements

We would like to thank Uwe Matzat and two anonymous reviewers for their valuable comments on earlier drafts of this paper.

References

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. A., & Frieder, O. (2004). *Hourly analysis of a very large topically categorized Web query log*. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), SIGIR (pp. 321–328). ACM.

Gayo-Avello, D. (2009, May). *A survey on session detection methods in query logs and a proposal for future evaluation*. Inf. Sci., *179*, 1822–1843. doi:10.1016/j.ins.2009.01.026

Gupta, A., Kumaraguru, P., & Sureka, A. (2012). *Characterizing pedophile conversations on the Internet using online grooming*. CoRR, abs/1208.4324

Heckmann, O., Bock, A., Mauthe, A., & Steinmetz, R. (2004). *The eDonkey file-sharing network*. In P. Dadam & M. Reichert (Eds.), Gi jahrestagung (2) (Vol. 51, pp. 224–228). GI.

Hughes, D., Walkerdine, J., Coulson, G., & Gibson, S. (2006). *Is deviant behavior the norm on P2P file-sharing networks?* IEEE Distributed Systems Online, *7*(2).

Hurley, R., Prusty, S., Soroush, H., Walls, R. J., Albrecht, J., Cecchet, E., ... & Wolak, J. (2013, May). Measurement and analysis of child pornography trafficking on P2P networks. *Proceedings of the 22nd International Conference on World Wide Web* (pp. 631–642). International World Wide Web Conferences Steering Committee.

Klemm, A., Lindemann, C., Vernon, M. K., & Waldhorst, O. P. (2004). *Characterizing the query behavior in peer-to-peer filesharing systems*. In ACM SIGCOMM Internet Measurement Conference (IMC).

Latapy, M., Magnien, C., & Fournier, R. (2011). *Automatic paedophile query detection tool.* Retrieved from https://www-complexnetworks.lip6.fr/~latapy/antipaedo/

Latapy, M., Magnien, C., & Fournier, R. (2013). *Quantifying paedophile activity in a large P2P system*. Information Processing and Management, *49*, 248–263.

Liberatore, M., Levine, B. N., & Shields, C. (2010). Strengthening forensic investigations of child pornography on P2P networks. *Proceedings of the 6th International Conference on Emerging Networking Experiments and Technologies* (pp. 1–12). New York, NY, USA: ACM.

Légifrance. (2012). *LOI n° 2009-1311 du 28 octobre 2009 relative à la protection pénale de la propriété littéraire et artistique sur Internet.* http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000021208046&categorieLien=id.

Nguyen, L. T., Jia, D., Yee, W. G., & Frieder, O. (2007). *An analysis of peer-to-peer file-sharing system queries*. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, & N. Kando (Eds.), SIGIR (pp. 855–856). ACM.

Penna, L., Clark, A., & Mohay, G. (2005). Challenges of automating the detection of paedophile activity on the Internet. *Proceedings of the First International Workshop on Systematic Approaches to Digital Forensic Engineering* (pp. 206–220).

Roubaty, R. (2007, July). *Rapport technique d'appui aux services de police dans le cadre de l'affaire dite Razorback*. Haute École Valaisanne, Laboratoire de Sécurité. Retrieved from http://www.numerama.com/media/pdf/Ratiatum-Razorback_expertise.pdf

Rutgaizer, M., Shavitt, Y., Vertman, O., & Zilberman, N. (2012). Detecting pedophile activity in BitTorrent networks. In N. Taft & F. Ricciato (Eds.), *Passive and Active Measurement*, 7192, 106–115. Springer.

Spink, A., Koricich, A., Jansen, B. J., & Cole, C. (2004). *Sexual information seeking on web search engines*. CyberPsychology & Behavior, *7*, 65–72.