Mining bipartite graphs to improve semantic pedophile activity detection

Raphaël Fournier*, Maximilien Danisch†‡
*L2TI, Université Paris Nord, 99 avenue JB Clément, 93430 Villetaneuse, France. raphael.fournier@univ-paris13.fr
†Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France. maximilien.danisch@gmail.com
‡CNRS, UMR 7606, LIP6, F-75005, Paris, France.

***Abstract*—Peer-to-peer (P2P) networks are popular to exchange large volumes of data through the Internet. Pedophile activity is a very important topic for our society and some works have recently attempted to gauge the extent of pedophile exchanges on P2P networks. A key issue is to obtain an efficient detection tool, which may decide if a sequence of keywords is related to the topic or not. We propose to use social network analysis in a large dataset from a P2P network to improve a state-of-the-art filter for pedophile queries. We obtain queries and thus combinations of words which are not tagged by the filter but should be. We also perform some experiments to explore if the original four categories of paedophile queries were to be found by topological measures only.**

I. Introduction

Over the last fifteen years, peer-to-peer (P2P) networks have enabled millions of users to easily exchange files between distant places in the world. With a dedicated software, each member of the system may share some content on his machine and obtain some from others by requesting a keyword-based search engine.

Pedopornography – the depiction of sexual acts involving minors – is a problem of crucial importance for our society. Criminal networks involved in the production and the diffusion of such content must be dismantled and child molesters must be arrested. Recently, a methodology has been developed to study the extent of child pornography exchanges in P2P networks. It leads to the collection of very large scale datasets and the design of a semantic filter, the performances of which were assessed by world-wide experts in the domain [LMF12]. These authors have then realized experiments to obtain reliable estimates of the fraction of pedophile queries, which were completed by another team on different networks [RSVZ12].

P2P networks may also be seen as social networks, in which users form communities of interest (around searched topics) and exchange files [BLT13], [IRSNF11]. Then, we may use the methods developed in recent works in graph metrology and social network analysis, to deal with the relations between the semantic categories defined by a query filter and the topology of the underlying user graph.

More specifically, we address the two following questions:

- may the analysis of the graph enable the discovery of new keywords and combinations of keywords related to a given topic?
- may the semantic categories of the filter be rediscovered thanks to the topological analysis of a graph?

In both cases, we want to confirm these results on classifying pedophile queries with some keywords split in adequate categories, with an independent methodology. We also aim at proposing some ways to improve such a filter. Especially with the topic of pedophile activity, keywords and combinations may significantly change over time and updating the detection tool is very important. However, human expertise to perform such a task remains limited, even at a world-wide level. We then hope that topological experiments provide additional input which may reduce the need for human expertise.

This paper is organized as follows: we first present, in section II, our P2P data and the semantic filter used to tag queries as pedophile. We then discuss, in section III, the state of the art of the graph analysis approachs to complete a community. In section IV, we detail our alternative framework to increase the precision of a semantic filter. In Section V, we present our results on pedophile queries along with some results on the segmentation of the queries. Finally, we conclude and present future work in section VI.

II. Data and semantic filter

The data we use are keyword-based queries submitted to the *eDonkey* search engine by the users of the network. It was collected during a 10-week long experiment in 2007, on one of the most prominent *eDonkey* server at that time. There are 107,226,021 queries, from 23,892,531 distinct IP addresses, anonymised at collection time. Each query contains also a timestamp, a connection port and a list of keywords.

In [LMF12], authors presented their methodology to detect queries submitted to obtain child abuse material (which we will call "*pedophile queries*" subsequently). We use their "automatic detection tool". Figure 1 illustrates the classifying process in four categories of pedophile queries, relying on detecting keywords (or combinations of keywords) belonging to specific lists to decide if a query is pedophile or not. On the dataset, the filter identifies 207,340 pedophile queries, 151,545 in category 1, 27,753 in category 2, 35,264 in category 3 and 4,299 in category 4 (a query may be labeled in several categories).

In our context, where users are only identified by a public IP address and a connection port, it is hardly tractable to use the classification of queries to classify the users. Several mechanisms may impact the IP address received by the server for a given user (e.g.: virtual private networks, ISP dynamic allocation of IP addresses, etc.). Thus, by aggregating the queries over time, we would mix the queries of several distinct

users or, conversely, consider that queries are from different users when they come from one single individual. However, we neglect this effect in the rest of the paper (since queries were collected in a rather limited timescale) and we consider that a user is directly identified by a pair $(IP, port)$.

III. Community completion frameworks

Typically, users (or queries) detected by the semantic filter as being part of a category are nodes belonging to a community of interest in the P2P network, minus some mistakes of the semantic filter. While a community of interest is a group of nodes highly linked together (as they exchange files and search for similar files) and thus corresponds to the classical definition of a community in graph analysis [For09]. Specific framework designed to complete a community can thus be applied to unfold the whole community of interest (and thus add nodes that were tagged as not part of the community of interest, but should have been tagged so) and conversely remove nodes that were tagged as part of the community of interest by the semantic filter, but are actually not part of it. Such a completed community, centered around several nodes, is called a multi-ego-centered community [DGLG12], which is a generalisation of the ego-centered or local community notion [Cla05], [FCF11], [NTV12].

Given a quality function – taking as input nodes and giving as output a value quantifying to what extent a set of nodes is a good community of interest –, a possible approach to complete communities consists in starting with a set (community) composed only with the nodes tagged by the semantic filter and then optimizing greedily the quality function by adding or removing nodes from the set. The final set obtained would in the ideal case be the real community of interest.

A first step towards this goal is the work detailed in [SG10] where the quality function is defined as the minimum degree of the nodes in the subgraph induced by the community. For this specific quality function, an optimal, yet greedy, algorithm exists.

Other solutions are [KNV06] and [TF06] where a proximity measure approach is used rather than a quality function approach. In these articles, authors compute the proximity of every node in the graph to an input set of nodes, then they try to find the most relevant connected subgraph of size k (a parameter), i.e., the set of connected nodes which are globally the closest to the input set. With these approaches and in the ideal case with the correct k, the real community of interest would also be unfolded.

The methods previously cited are used on networks which are not bipartite, while the data we are dealing with are naturally modelled by a bipartite graph (from one side the queries and from the other the users). Thus methods dealing directly with bipartite graphs [GL04] should also be investigated.

While these techniques remain to be explored, we show here a methodology which is simpler yet leading to good results when applied to complete the community of interest of pedopornography in P2P network.

IV. Methodology

We propose to improve the filter by analysing a bipartite graph of queries for one part and users for another part (see

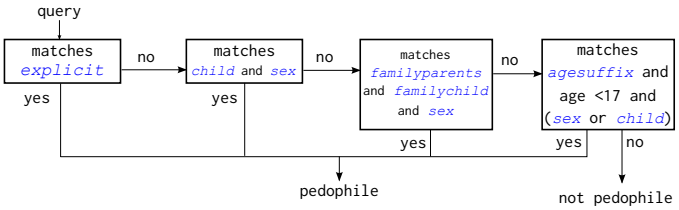

Figure 1: the pedophile detection tool presented in [LMF12] (in blue, groups of keywords). In the rest of the paper, we will denote by, from left to right, 1, 2, 3 and 4 those categories, and 0 for the non pedophile one.

Figure 2). Our aim is not to explore a completely new classification of queries but instead to lower the number of errors of the existing classification. Errors are of two categories: the false positives queries, which were labeled as pedophile but should not have been (i.e., they are not pedophile for a human expert), and the false negatives which were not detected but should have been.

Thus, by using the existing classification of queries, we define a score which rely on a topological similarity between the queries. We want that true pedophile queries receive a high score. We could then distinguish between true positive and false negative. Reducing this number may help in improving the recall measure for this filter. Conversely, we want queries that are not pedophile to have a low score, and aim at reducing the false positive errors of this filter.

We note U the set of users, R the set of queries. We define f which associates, to a given couple $(u, r) \in U \times R$, the number of times u submitted query r. For each query r, we define its neighborhood $V(r)$ such as:

$$V = \{u \in U | f(u, r) \geq 1\}.$$

For each user, we note $R(u)$ the set of its queries. For a given class C of queries, we define the score $s_C(r)$ with the following expression ((1)):

$$s_C(r) = \frac{\sum_{u \in V(r)} |C \cap R(u) \setminus \{r\}|}{\sum_{u \in V(r)} |R(u) \setminus \{r\}|} \quad (1)$$

Users with only one query do not contribute to making connections between queries and they are then removed. Similarly, the query whose score is being computed does not contribute either. The expression (1) takes into account the imbalance of number of queries per user. We also used the score (2) which normalise the contributions of each user, whatever her number of queries may be. However, the results are not significantly different than those provided by (1), we thus present only the results obtained with score (1) in the following.

$$s_C(r) = \sum_{u \in V(r)} \frac{|\{C \cap R(u) \setminus \{r\}|}{|\{R(u) \setminus \{r\}|} \quad (2)$$

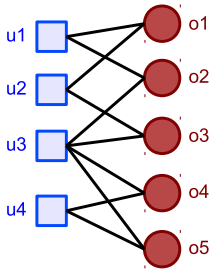

Figure 2: the bipartite graph *users-queries*. Users are represented by squares and they are linked to the queries they submitted. The numbers in each circle indicate the category of the query.

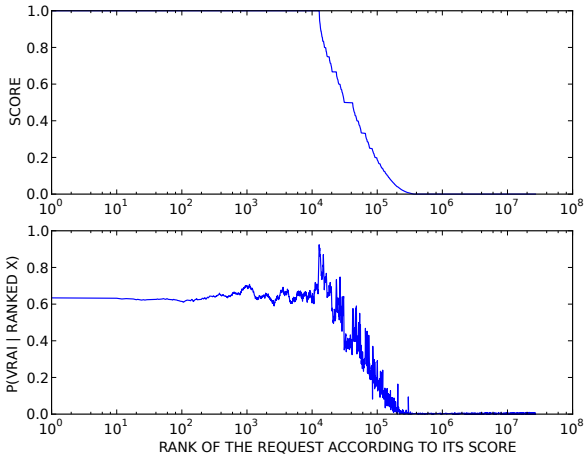

Figure 3: Score obtained with the class C of all queries labeled 1, 2, 3 or 4 (top) and probability for a query to be pedophile as a function of the ranking order (bottom)

V. Results

A. False positive and false negative queries

Figure 3 (top) presents the scores obtained by all the queries from the dataset, for the class C including all pedophile queries tagged by the semantic filter, ranked by their score. Figure 3 (bottom) presents a different view were queries are by sliding groups of k[1]. For a query at position X, the corresponding value is the fraction of pedophile queries detected by the original filter between position X and $X + k$.

We obtain 12,858 queries with a score of 1, i.e. submitted by users whose other queries were *all* pedophile (except the given query which can be not pedophile). Among those queries, 4,518 of them (slightly above 35%) were not detected by the filter. Most of them are very close from the topic, with either new specific keywords or combinations of relevant keywords. It thus seems relevant to perform an in-depth study of the keywords and their combination to include them in the filter. However, great attention should be paid in order to avoid a unwanted increase in the number of false positives associated with those new keywords. Indeed, other queries could be studied, such as those with a score close to 1, perhaps even until 0.75: since many users have submitted few queries, a score of 0.75 may be attributed to queries from people with 5 queries, 3 being paedophile, 1 not paedophile and the current query, thus 0.75 may still be a somewhat significant score. There are 5,361 queries with a score greater or equal to 0.75 (and not equal to 1) which could benefit from further investigation.

For the queries labeled pedophile with a score of 0, results are less conclusive: even though they were submitted by users with *no* other pedophile queries, most of them do not appear to be false positives.

[1]We used $k = 500$ as a trade-off to limit oscillations

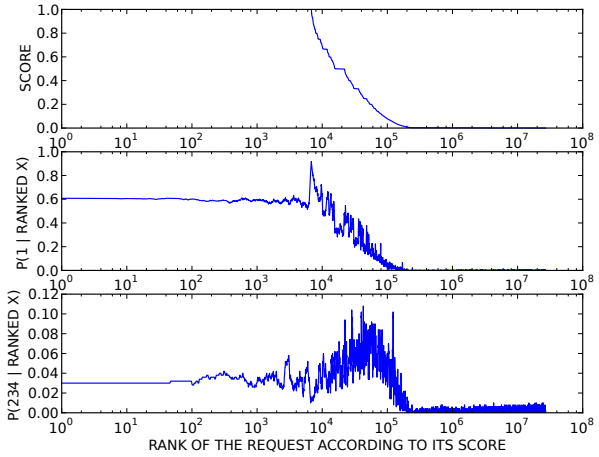

Figure 4: Score obtained with the class C of all queries labeled 1 (top), probability for a query to be labeled 1 (middle) and probability to be labeled 2, 3 or 4 (as a function of the ranking order (bottom)

B. Segmentation of pedophile activity

In this section, we want to examine if the original 4 categories of pedophile activities seems relevant, with regard to the topology. Especially, we want to know if we were to hide category 1 queries, composed of queries with specific keywords (most of them being unknown to the general public), would we be able to find them by computing only with categories 2, 3 and 4 queries (categories which are semantically more similar).

Figure 4 presents the results of our experiment: this time, the score is computed only with respect to the class C of queries from category 1 (4-top) , we see that category 1 queries are massively ranked at the top (4-middle), before observing that queries from categories 2, 3 and 4 have low scores (4-bottom). This last plot shows that category 1 is not much linked with others: users submitting queries from category 1 are less likely to also enter queries from categories 2, 3 or 4. This is a first step to show that semantic types of pedophile activity are also found in the topology. We also tried the other combinations, for instance finding queries 4 using only queries 2, and all other combinations lead to similar results.

For a deeper analysis, we compare, for each category, the

number of pairs of queries submitted by a single user, to see whether users tend to stick to a particular category.

i	j	V_{ij}	$E(V_{ij})$	$V_{ij}/E(V_{ij})$
1	1	205,986	177,595.62	1.15
1	2	21,296	35,266.72	0.60
1	3	33,239	73,718.54	0.45
1	4	1,905	4,235.47	0.44
2	2	8,525	1,752.94	4.86
2	3	7,432	7,315.32	1.01
2	4	731	421.04	1.73
3	3	27,995	7,649.96	3.65
3	4	552	879.19	0.62
4	4	1,199	25.14	47.69

Table I: Statistic of the number of pairs of queries submitted by a single user and comparison to a *null model*

Table I presents those results: V_{ij} is the number of pairs from category i and j submitted by a single user. $E(V_{ij})$ is an expected value of V_{ij} by assuming that a query of category i is on average in $\sum_{k=1}^{4} V_{ik}$ pairs of queries, this corresponds to a *null model* where each pedophile query would be made by the same number of peers (as it is made in the real dataset) and each peer would make the same number of pedophile queries (as it had in the real dataset). The ratio of the two previous values is roughly above 1 for $i = j$ and below for $i \neq j$. The comparison to this null model shows that two queries from the same categories have a higher probability to come from a single user than two queries from two different categories. The semantic categorisation is thus also found at a topological level and sounds relevant.

VI. Conclusion and perspectives

Given a semantic filter to categorize queries related to a specific topic on a P2P network and the $(queries, peers)$-bipartite graph of this P2P network, we presented a framework that scores queries from the most similar to the less similar to the queries detected by the semantic filter. The aim of the framework is to help a human expert in finding keywords and new combination of keywords in order to refine the semantic filter and thus limit the number of false negatives and the number of false positives.

We applied the framework to the semantic filter presented in [LMF12] and on a large dataset from *eDonkey* and obtained qualitatively good results. Thus, the topological study of a bipartite graph between users and queries from a P2P network can help to improve and update a semantic filter.

We also presented some statistics in order to further validate the similarity between semantic and topological proximities of queries. We found that, for the four categories of the original filter, a given peer is more likely do make queries belonging to a same category than queries belonging to different categories.

Our work opens the way to several avenues of future work. A reasonable perspective with an applicative aim is to develop the complete methodology to perform the update of the original filter. The protocol of evaluation to include new keywords and combinations should be defined and tried.

A deeper study through more sophisticated tools from graph analysis will also be investigated, such as clustering algorithms. It could help in segmenting more subtly pedophile queries and users, for example to subdivide the existent categories or find some new ones, which we did not investigate here. These methods may also be fruitful to continue our preliminary analysis of the correlations between categories. Eventually, we hope that our methodology could be applied in other contexts of text categorization.

Acknowledgment

This work is partially funded by French National Research Agency (ANR CODDDE, ANR-13-CORD-0017-01) and the Paris Region (FUI AMMICO project).

References

[BLT13] Daniel F Bernardes, Matthieu Latapy, and Fabien Tarissan. Inadequacy of sir model to reproduce key properties of real-world spreading cascades: experiments on a large-scale p2p system. *Social Network Analysis and Mining*, 3(4):1195–1208, 2013.

[Cla05] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72(2):026132, 2005.

[DGLG12] M. Danisch, J.L. Guillaume, and B. Le Grand. Towards multi-ego-centered communities: a node similarity approach. *Int. J. of Web Based Communities*, 2012.

[FCF11] A. Friggeri, G. Chelius, and E. Fleury. Triangles to capture social cohesion. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 258–265. IEEE, 2011.

[For09] Santo Fortunato. Community detection in graphs. *Physics Reports*, 2009.

[GL04] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Inf. Process. Lett.*, 90(5):215–221, 2004.

[IRSNF11] Adriana Iamnitchi, Matei Ripeanu, Elizeu Santos-Neto, and Ian Foster. The small world of file sharing. *IEEE Trans. Parallel Distrib. Syst.*, 22(7):1120–1134, July 2011.

[KNV06] Yehuda Koren, Stephen C North, and Chris Volinsky. Measuring and extracting proximity in networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–255. ACM, 2006.

[LMF12] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, 2012.

[NTV12] B. Ngonmang, M. Tchuente, and E. Viennet. Local community identification in social networks. *Parallel Processing Letters*, 22(01), 2012.

[RSVZ12] Moshe Rutgaizer, Yuval Shavitt, Omer Vertman, and Noa Zilberman. Detecting pedophile activity in BitTorrent networks. In Nina Taft and Fabio Ricciato, editors, *PAM*, volume 7192 of *Lecture Notes in Computer Science*, pages 106–115. Springer, 2012.

[SG10] Mauro Sozio and Aristides Gionis. The community-search problem and how to plan a successful cocktail party. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 939–948. ACM, 2010.

[TF06] Hanghang Tong and Christos Faloutsos. Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 404–413. ACM, 2006.