Classification of Content and Users in BitTorrent by Semi-supervised Learning Methods

Konstantin Avrachenkov
INRIA Sophia Antipolis, France
K.Avrachenkov@sophia.inria.fr

Paulo Gonçalves
INRIA Rhone-Alpes, France
paulo.goncalves@inria.fr

Arnaud Legout
INRIA Sophia Antipolis, France
arnaud.legout@inria.fr

Marina Sokol
INRIA Sophia Antipolis, France
marina.sokol@inria.sophia.fr

***Abstract*—P2P downloads still represent a large portion of today's Internet traffic. More than 100 million users operate BitTorrent and generate more than 30% of the total Internet traffic. Recently, a significant research effort has been done to develop tools for automatic classification of Internet traffic by application. The purpose of the present work is to provide a framework for subclassification of P2P traffic generated by the BitTorrent protocol. The general intuition is that the users with similar interests download similar contents. This intuition can be rigorously formalized with the help of graph based semi-supervised learning approach. We have chosen to work with PageRank based semi-supervised learning method, which scales well with very large volumes of data. We provide recommendations for the choice of parameters in the PageRank based semi-supervised learning method. In particular, we show that it is advantageous to choose labelled points with large PageRank score.**

I. Introduction

P2P downloads still represent a large portion of today's Internet traffic. More than 100 million users operate BitTorrent and generate more than 30% of the total Internet traffic [7]. According to the Wikipedia article about BitTorrent [2], the traffic generated by BitTorrent is greater than the traffic generated by Netflix and Hulu combined. Recently, a significant research effort has been done to develop tools for automatic classification of Internet traffic by application [9], [8], [11]. The purpose of the present work is to provide a framework for subclassification of P2P traffic generated by the BitTorrent protocol. Unlike previous works [9], [8], [11], we cannot rely on packet level characteristics (packet size, packet interarrival time, etc). Instead we make use of the bipartite user-content graph. This is a graph formed by two sets of nodes: the set of users (peers) and the set of contents (downloaded files). From this basic bipartite graph we also construct the user graph, where two users are connected if they download the same content, and the content graph, where two files are connected if they are both downloaded by at least one user. Using methodology developed in [7] we were able to use the snapshots of P2P downloads from the whole Internet. Even a snapshot corresponding to half an hour duration represent a huge among of data. Without some filtering technique, which will be explained in Section 3 we were even not able to operate with the user graph constructed from a single snapshot. The content graph is smaller and we were able to construct an aggregated content graph from several snapshots corresponding to the week-long observation.

The general intuition is that the users with similar interests download similar contents. This intuition can be rigorously formalized with the help of graph based semi-supervised learning approach [13]. In particular, we have chosen to work with PageRank based semi-supervised learning method [3], [4], [12]. It has been demonstrated in [4] that the PageRank based semi-supervised learning method has implementations with quasi-linear complexity and produces robust results with respect to the method's parameters.

We have three goals in the present work. The main goal is to provide a robust graph based semi-supervised learning approach for content and user classification of BitTorrent P2P transfers. The second goal is to demonstrate that the PageRank based semi-supervised learning method, thanks to its quasi-linear complexity, can deal with classification of very large datasets. Some datasets used in the present paper is several orders of magnitude larger than datasets typically used in the literature on graph based semi-supervised learning. The third goal is to test the impact of the choice of the labelled nodes on classification result. In particular, we test the following three options for the choice of the labbelled points: randomly chosen labelled points, labelled points with large PageRank values and labelled points with large degrees. We demonstrate that in the context of P2P classification the choice of labeled points with large PageRank values gives good results in the majority of classification tasks.

The article is organized as follows: In the next Section II we describe the PageRank based semi-supervised learning method. Then, in Section III we give detail description of our datasets and method implementation. In Section IV we perform topic based and language based classifications of the whole collection of the P2P traffic based on the content graph and user graph, respectively. Then, we examine our method on smaller subsets of content. In Section V we analyse the content classification of the video plus music subgraph. In Section VI we study the classification of the untagged content

in the "other video" category. Again the graph based semi-supervised learning method performed very well and provided good suggestions for finer subcategorization. In Section VII we give conclusions and provide directions for future research.

II. PageRank based classification

Let us present some basic facts about PageRank based semi-supervised learning method. An interested reader can find more theoretical results in [4] and in related works [3], [12].

Suppose we need to classify N data points into K classes and P data points are labelled. In particular, this means that for a labelled point $i = 1, ..., P$ the function $k(i) \in 1, ..., K$ is defined. Graph based semi-supervised learning approach uses a weighted graph connecting data points. The weight matrix, or similarity matrix, is denoted by W. Here we assume that the weight matrix W is symmetric. Each element $w_{i,j}$ represents a degree of similarity between data points i and j. Denote by D a diagonal matrix with its (i,i)-element equals to the sum of the i-th row of matrix W: $d_{i,i} = \sum_{j=1}^{N} w_{i,j}$.Define $N \times K$ matrix Y as

$$Y_{ik} = \begin{cases} 1, & \text{if } X_i \text{ is labeled as } k(i) = k, \\ 0, & \text{otherwise.} \end{cases}$$

We refer to each column $Y_{\cdot k}$ of matrix Y as labeling function. Also define $N \times K$ matrix F and call its columns $F_{\cdot k}$ classification functions. A general idea of the graph-based semi-supervised learning is to find classification functions so that on the one hand they will be close to the corresponding labeling function and on the other hand they will change smoothly over the graph associated with the similarity matrix. This general idea can be expressed by means of the optimization formulation

$$\underset{F}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| \frac{F_{i\cdot}}{d_{ii}} - \frac{F_{j\cdot}}{d_{jj}} \|^2 + \mu \sum_{i=1}^{N} \frac{1}{d_{ii}} \| F_{i\cdot} - Y_{i\cdot} \|^2 \quad (1)$$

where μ is a regularization parameter. In fact, the parameter μ represents a trade-off between the closeness of the classification function to the labeling function and its smoothness.

Proposition 1: The classification functions for the PageRank based semi-supervised learning are given by

$$F_{\cdot k} = \frac{\mu}{2+\mu} \left(I - \frac{2}{2+\mu} W D^{-1} \right)^{-1} Y_{\cdot k}, \quad (2)$$

for $k = 1, ..., K$.

Let us now explain why the following framework corresponds to the PageRank based clustering method. Denote $\alpha = 2/(2+\mu)$ and write $F_{\cdot k}$ in a transposed form

$$F_{\cdot k}^T = (1-\alpha) Y_{\cdot k}^T (I - \alpha D^{-1} W)^{-1}.$$

If the labeling functions are normalized, this is exactly an explicit expression for PageRank [10]. This expression was used in [3] but no optimization framework was provided.

Note that $D^{-1}W$ represents the transition probability matrix for the random walk on the similarity graph. Then, the (i,j)-th element of the matrix $(I - \alpha D^{-1} W)^{-1}$ gives the expected number of visits to node j starting from node i until the random walk restarts with probability $1 - \alpha$. This observation provides the following probabilistic interpretation for the PageRank based method. In the PageRank based method with normalized labeling functions, F_{ik} gives up to a multiplicative constant the expected number of visits to node i, if the random walk starts from a uniform distribution over the labeled nodes of class k.

The choice of the labelled points can potentially have a significant influence on classification results. Therefore, in the present work we study this influence. Specifically, we consider the following options for the choice of labelled points:

1) randomly chosen labelled points, that is, in each class we take several samples of random labelled points;
2) labelled points are chosen among points with large values of Standard PageRank; (with large values of π_i, $i = 1, ..., N$, where π_i are elements of a solution of the equation $\pi = \pi \alpha D^{-1} W + (1-\alpha)/N \underline{1}^T$);
3) labelled points are chosen among points with large degree (with large values of $d_{i,i}$).

III. Datasets and method implementation description

We have several snapshots of the Torrents collected from the whole Internet using methodology described in [7]. Each snapshot contains half an hour of P2P transfers. In total, we have about one week of observations. We have also an aggregate representing the transfers observed during the whole week. To test the effect of NATs, to save memory and to reduce information noise, the following filtering has been applied which we denote by $g(X,Y)$: we filter out all IP addresses with more than or equal to X ports (X = 0 means no filtering), and we filter out all contents with less than or equal to Y IP addresses seen downloading the content (Y = 0 means no filtering). Two users with the same IP addresses but with different ports could be the same user. So the filtering by ports helps us to reduce the influence of counting the same user as different ones. The second filter by IP address helps to remove unpopular contents which were downloaded less than or equals to Y times.

TABLE I: The content graphs after preprocessing.

Graph	# nodes	# edges
g(2,10)	200 413	50 726 946
g(0,10)	200 487	174 086 752
g(2,0)	624 552	92 399 318

We use the whole aggregate to create the content graph. Some files are tagged with information about name, language, topic, login of the person who inserted these files. Those tags correspond to the classification made by popular torrent sites like ThePirateBay [7]. If two files are downloaded by the same user, we create an edge between these two files. The weight of the edge shows how many users downloaded these two files. We filter out all links with the weight equal to one to reduce the noise and memory usage. Without this filtering even the

PageRank based method with quasi-linear complexity cannot be applied on a standard desktop computer.

We start with the smallest aggregated dataset $g(2, 10)$ which contain information with small noise. To evaluate the impact of the noise with respect to user identification we have also made experiments with datasets $g(0, 10)$ and $g(2, 0)$.

The graph for $g(2, 0)$ dataset after preprocessing contains three times more nodes and two times more edges than the dataset $g(2, 10)$. The graph for $g(0, 10)$ dataset after preprocessing contains two times more edges than the dataset $g(2, 10)$.

TABLE II: The quantity of language base line expert classifications.

Language	#content	#user
English	36465	57632
Spanish	2481	2856
French	1824	2021
Italian	2450	3694
Japanese	720	416

TABLE III: The quantity of topic base line expert classifications.

Topic	# content	# user
Audio Music	23639	13950
Video Movies	20686	43492
TV shows	12087	27260
Porn movies	8376	7082
App. Windows	4831	2874
Games PC	4527	8707
Books Ebooks	1185	281

Let us now describe how we construct the user graph. The user graph is constructed with the help of HADOOP realization of MapReduce technology [1] from the basic user-content bipartite graph from a single half an hour snapshot. The aggregated user graph is too large to work with.

The snapshot contains information on which content was downloaded by whom. In the user graph an edge with the weight M signifies that two users download M same files. The user graph has $3\,228\,410$ nodes and $3\,436\,442\,577$ edges. The number of edges with weight one is equal to $3\,309\,965\,972$. Also we have noticed that some users downloaded much more files than a normal user would do. One user who has downloaded $655\,727$ files for sure is a robot. Thus, we have decided remove all edges with weight one and the user-robot. The modified user graph has $1\,126\,670$ nodes and $124\,753\,790$ edges. This filtering significantly reduces required computing and memory resources. In fact, by doing this filtering we also remove some information noise. If two users download only one common item it could be by pure chance, if they both download more than two same files - it is more likely that they share same interests.

We classify contents and users by both language and topics. The considered languages and topics are given in Tables II and III.

Our base line expert classification is based on P2P content tags if they are available. For instance, in the case of classification by language we consider that the content is in English if it has only tag "English". And we consider a user to be an English language user, if he or she downloads only English language content.

We have implemented PageRank based classification method in the WebGraph framework [6]. The WebGraph framework has a very efficient graph compression technique which allows us to work with very large graphs.

IV. Results of classification of content and users

Using PageRank based classification method, we have performed four classification experiments. We have used the aggregated graph of content $g(2, 10)$ to classify the content into 5 classes according to the languages (see Table II) and into 7 classes according to the content type (see Table III). The classification of the aggregated content graph has taken approximately 15 minutes on a 64-bit computer with Intel-Core7i processor and 6GB RAM. The results of the classification evaluated in terms of accuracy are presented in Tables IV and V. Then, we have performed the classification of users also into 5 classes of the languages and into 7 classes of the content preferred by users (see Tables VI and VII). It has taken about 20 minutes on the same computer. However, the preprocessing of a single snapshop of the user graph was much more demanding than the preprocessing of the aggregated content graph. Our main conclusion is that the PageRank based classification method scales remarkably well with large volumes of data. Then, our second important observation is that by using a very little amount of information, we are able to classify the content and users with high accuracy. For instance, in the dataset of $1\,126\,670$ users, using only 50 labelled points for each language, we are able to classify the users according to their preferred language with 88% accuracy.

In all four classification experiment, we have tried three different options for the choice of the labelled points. We have chosen the labelled points: (a) with largest standard PageRank values; (b) with largest degree; and (c) randomly. When evaluating the performance with the randomly chosen labelled points we have averaged the accuracy over 10 random samples (because of the size of the data, making more than 10 samples for each of many experimental setups was very time demanding) and we have also reported the worst (rand min column) and the best (rand max column) accuracy. With respect to the choice of the labelled points, our conclusion is that in the majority of cases the labelled points with large values of the standard PageRank are the best picks (see topPR columns). In the case of classification with the aggregated content graph, the labelled points with large degrees give results comparable with the results obtained with the labelled points chosen according to PageRank. However, it was interesting to observe that in the case of the classification of users, the classification based on the labelled points with large degrees does not perform well at all. Our explanation is that in that dataset the nodes with very

large degrees are not representative. There is an independent confirmation of this idea given in [5].

We would like to note that there is not much difference if one considers weighted or unweighted graph for content classification. As one can see from Table IV, the accuracy of content classification by languages in the case of unweighted graph is 66.3% (choosing 50 labbelled points for each class according to top PageRank values). We have repeated the experiment with the weighted graph and have obtain 68.9% accuracy. We explain the relatively small difference in accuracies by the fact that 88.3% of edges have weight 1 and then 7.1% of edges have weight 2. So the majority of edges have weight 1 and the other edges have also small weight.

Finally, we have observed that the classification using $g(2, 10)$ filtering is one or two percent better in terms of accuracy than the classification using $g(0, 10)$ filtering. Thus, by doing the filtering we not only reduce the amount of data required for processing, but also we reduce the information noise.

To understand better how the graph based semi-supervised learning works let us consider in the next two sections smaller subsets of content.

TABLE IV: Accuracy of the classifications for the $g(2, 10)$ dataset by languages.

# seeds	topPR	topDeg	rand (10Exp)	rand min	rand max
5	0.579	0.573	0.51	0.44	0.578
50	0.663	0.647	0.634	0.614	0.649
500	0.688	0.676	0.658	0.653	0.663

TABLE V: Accuracy of the classifications for the $g(2, 10)$ dataset by topics.

# seeds	topPR	topDeg	rand(10Exp)	rand min	rand max
5	0.504	0.51	0.48	0.36	0.546
50	0.6344	0.6276	0.6278	0.604	0.645
500	0.7279	0.7182	0.6562	0.6525	0.6595

TABLE VI: Accuracy of the classifications for the user dataset by languages.

# seeds	topPR	topDeg	rand (10Exp)	rand min	rand max
5	0.788	0.765	0.732	0.613	0.817
50	0.88	0.78	0.834	0.82	0.85
500	0.853	0.535	0.901	0.896	0.907

TABLE VII: Accuracy of the classifications for the user dataset by topics.

# seeds	topPR	topDeg	rand(10Exp)	rand min	rand max
5	0.683	0.399	0.631	0.563	0.678
50	0.752	0.477	0.767	0.752	0.777
500	0.789	0.52	0.86	0.858	0.865

V. Classification of Video plus Music subgraph

We have constructed a subgraph which consists of all files which have in their tags "video", "movie", "audio" or "music". In Table VIII we see the results of classification "music"+"audio" against "video"+"movie". The results are quite good (accuracy 90% against accuracy 63.4% in the case of 50 labelled points chosen according to the top PageRank values, see Tables V and VIII). The good classification is probably due to the fact that the dataset is smaller and the classes are balanced. In particular it is interesting to observe how the files tagged "Music Video Clips" are classified. 143 such files are classified into "music", and 45 files are classified into "video". This is quite in agreement with intuition that "music video clips" are better related to music than to video. On opposite only 20 "video movie clips" are classified as "music" and 125 "video movie clips" are classified as "video". This also agrees with our intuition since most of "video movie clip" files are short extracts from movies.

# seeds	Accuracy	CV matrix		
50 topPR	0.90	music video	30833 4775	4337 50862
500 topPR	0.938	music video	32008 2418	3162 53219
1000 topPR	0.946	music video	32705 2474	2465 53163
500m/1000v topPR	0.942	music video	32423 2510	2747 53127

TABLE VIII: Accuracy and Cross-Validation (CV) matrix for music&audio vs video&movies classification, $\alpha = 0.5$.

VI. Classification of untagged content

We have also created "other video" subgraph from the whole content graph. We have taken all nodes for which we have topic tags as "other video" and all edges induced by the supergraph. The subgraph contains 1189 nodes and 20702 edges. We made the expert evaluation manually by popular categories: "Sport Tutorials" [ST] (116), "Science Lectures" [SL] (127), "Japanese Cartoons" [JC] (93), "Porno" [P] (81), "Software Tutorials" [SFT] (113), "Movies" [M] (129).

The results of the semi-supervised classification are presented in Tables IX, X, XI. In Table IX we demonstrate the effect of the choice of the labelled points. As expected the more labelled points we take the better. In Table XII we compare in detail the random choice of labelled points with the labelled points chosen according to their PageRank value. Specifically we average the results over 100 experiments with random labelled points. We can see that the precision corresponding to the labelled points chosen by PageRank is better than the average precision corresponding to the random choice of the labelled points. The coefficient of variation (CoV) for the random choice of the labelled points is significant (around 20%), which means that if we choose labelled point randomly the result of the classification is much less reliable than the result of the classification according to the labelled points with large PageRank values. It was surprising to observe that choosing labelled points with large degree does not help much. May be here we also face the phenomenon described in [5].

In Tables X, XI we present the Cross-Validation matrices for experiments with 10 and 15 labeled points chosen according

to large PageRank values. In both tables we see strong diagonal domination. It is nice to observe that we have good classification despite the fact that nearly the half of the files do not belong to any of the mentioned above six classes. This can be interpreted as robustness of graph based semi-supervised learning approach with significant presence of noisy data.

Furthermore, it is interesting to observe that most of the "other video" files with the content as "Dance Tutorials" (21 from 27) are classified into "Sport Tutorials" [ST], which seems to be indeed related category. And all tutorials about gun shooting (13) are classified in "Sport Tutorials", even though they have not initially been classified as "Sport Tutorials". This automatic classification appears to be quite logical and suggests the possibility of application of graph based semi-supervised learning for refinement of P2P content categorization.

# seeds	TopPR	TopDeg	rand(100Exp)	rand min	rand max
1	0.56	0.519	0.45	0.21	0.64
5	0.66	0.53	0.62	0.53	0.7
10	0.70	0.66	0.685	0.623	0.73
15	0.731	0.68	0.72	0.66	0.75

TABLE IX: Accuracy for "Other Video" subgraph classification, $\alpha = 0.5$.

Classified as→	JC	M	P	SFT	SL	ST
JC	65	2	1	1	5	8
M	6	47	18	6	11	21
P	0	8	59	4	2	3
SFT	3	4	3	91	9	3
SL	5	5	3	10	85	19
ST	2	9	5	8	2	85

TABLE X: Cross-Validation matrix for "Other Video" subgraph classification, TopPR 10 labeled points, $\alpha = 0.5$.

Classified as→	JC	M	P	SFT	SL	ST
JC	77	3	1	0	4	3
M	9	54	13	4	6	25
P	0	8	57	5	3	3
SFT	4	4	0	98	5	2
SL	5	7	2	9	92	12
ST	10	9	5	7	1	82

TABLE XI: Cross-Validation matrix for "Other Video" subgraph classification, TopPR 15 labelled points, $\alpha = 0.5$

Seeds	Average	Variance	min	max	CoV
rand10	0.684	0.022	0.622	0.725	0.217
rand15	0.726	0.018	0.682	0.773	0.185

TABLE XII: Statistics for accuracy for "Other Video" subgraph classification, $\alpha = 0.5$, random labeled points, 100 experiments

VII. Conclusions and future research

We have proposed to apply the PageRank graph-based semi-supervised learning method to classify P2P content and users. The proposed method have appeared to be highly scalable. We were able to deal with all world-wide torrents active in some point in time. With very few labelled points we have achieved very high precision. One of our principal recommendations is to choose labelled points with large values of PageRank. We have also demonstrated that the graph-based semi-supervised method is very robust with respect to various types of noise in the data. As a future research direction we suggest to consider a combination of graph-based unsupervised and semi-supervised methods to produce an automatic or computer-aided categorization of P2P traffic.

Acknowledgement

The work has been supported by the joint INRIA Alcatel-Lucent Laboratory. A part of this work has been presented at NIPS BigLearning workshop with no copyright binding.

References

[1] Hadoop mapreduce software framework, http://hadoop.apache.org/mapreduce/. 2011.

[2] Wikipedia article "bittorrent (protocol)", http://en.wikipedia.org/wiki/bittorrent_(protocol). 2011.

[3] Konstantin Avrachenkov, Vladimir Dobrynin, Danil Nemirovsky, Son Kim Pham, and Elena Smirnova. Pagerank based clustering of hypertext document collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 873–874. ACM, 2008.

[4] Konstantin Avrachenkov, Paulo Gonçalves, Alexey Mishenin, and Marina Sokol. Generalized optimization framework for graph-based semi-supervised learning. *Accepted to SIAM Conference on Data Mining, also available as INRIA Research Report at http://hal.inria.fr/inria-00633818/en/*, 2012.

[5] Brian Ball, Brian Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Phys. Rev. E*, 84:036103, Sep 2011.

[6] Paolo Boldi and Sebastiano Vigna. The webgraph framework i: compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 595–602, New York, NY, USA, 2004. ACM.

[7] Stevens Le Blond, Arnaud Legout, Fabrice Lefessant, Walid Dabbous, and Mohamed Ali Kaafar. Spying the world from your laptop: identifying and profiling content providers and big downloaders in bittorrent. In *Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*, LEET'10, pages 4–4, Berkeley, CA, USA, 2010. USENIX Association.

[8] Wei Li, Marco Canini, Andrew W. Moore, and Raffaele Bolla. Efficient application identification and the temporal and spatial stability of classification schema. *Comput. Netw.*, 53:790–809, April 2009.

[9] Wei Li and Andrew W. Moore. A machine learning approach for efficient traffic classification. In *Proceedings of the 2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 310–317, Washington, DC, USA, 2007. IEEE Computer Society.

[10] Cleve B. Moler. *Numerical Computing with MATLAB*. 2004.

[11] Marcin Pietrzyk, Jean-Laurent Costeux, Guillaume Urvoy-Keller, and Taoufik En-Najjary. Challenging statistical classification for operational usage: the adsl case. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 122–135, New York, NY, USA, 2009. ACM.

[12] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004.

[13] Xiaojin Zhu. Semi-supervised learning literature survey, technical report 1530, department of computer sciences, university of wisconsin, madison, 2005.