Exploring the Geography of Tags in Youtube Views

Stéphane Delbruel, Davide Frey, François Taïani

Exploring the Geography of Tags in Youtube Views

Stéphane Delbruel, Davide Frey, François Taïani

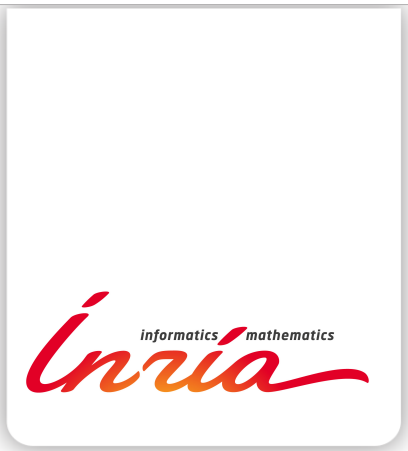
TECHNICAL REPORT

N° 461

May 2015

Project-Team ASAP

ISSN 0249-0803 ISRN INRIA/RT--461--FR+ENG



Exploring the Geography of Tags in Youtube Views

Stéphane Delbruel, Davide Frey, François Taïani

Project-Team ASAP

Technical Report n° 461 — May 2015 — 27 pages

Abstract: Although tags play a critical role in many social media, their link to the geographic distribution of user generated videos has been little investigated. In this paper, we analyze the correlation between the geographic distribution of a video's views and the tags attached to this video in a Youtube dataset. We show that tags can be interpreted as markers of a video's geographic diffusion, with some tags strongly linked to well identified geographic areas. Based on our findings, we explore whether the distribution of a video's views can be predicted from its tags. We demonstrate how this predictive power could help improve on-line video services by preferentially storing videos close to where they are likely to be viewed. Our results show that even with a simplistic approach we are able to predict a minimum of 65.9% of a video's views for a majority of videos, and that a tag-based placement strategy can improve the hit rate of a distributed on-line video service by up to 6.8% globally, with an improvement of up to 34% in the USA.

Key-words: User-generated content, YouTube, tag, prediction

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Corrélation entre tags et distribution géographique des vues dans YouTube

Résumé : Bien que les tags jouent un rôle critique au sein de nombreux médias sociaux, leur rapport à la distribution géographique des vidéos générées par les utilisateurs a très peu été évoqué. Dans ce papier, nous analysons la corrélation entre la distribution géographique des vues d'une vidéo et les tags relatifs à cette vidéo, au sein d?un dataset YouTube. Nous démontrons que les tags peuvent être interprétés comme des indice pour la diffusion géographique d'une vidéo, avec certains tags très fortement liés à des zones géographiques bien définies. En s'appuyant sur nos découvertes, nous explorons la possibilité de prédire la distribution des vues d'une vidéo à partir de ces tags. Nous démontrons que même avec une approche très simple, nous sommes capables de prédire correctement un minimum de 65.9% des vues pour la majorité des vidéos.

Mots-clés : YouTube, tag, prediction

1 Introduction

Videos streaming is currently reshaping the global Internet. It has grown to become one of the largest sources of worldwide Internet traffic, with reports of video content accounting for up to 60% of an ISP's load during peak periods [11]. A large proportion of this traffic is caused by User Generated Content (UGC) services such as Youtube, Dailymotion, or Vimeo: in 2013 for instance, Youtube accounted for 18.69% of the overall network traffic in North America, 28.73% in Europe, and up to 31.22% in Asia [3]. Storing, processing, and delivering this amount of data poses a constant engineering challenge to both UGC service providers and ISPs. One of the main difficulties is the sheer number of submissions these systems must process (300 hours of videos uploaded to YouTube every minute in April 2014 [25]) , most of which need to be served to niche audiences, in limited geographic areas [5, 16, 20].

Better understanding what these niche audiences and geographic areas are is a first critical step to improve the delivery infrastructure of UGC systems, and thus save bandwidth, electricity, and storage costs. Earlier studies have considered different facets of UGC video consumption, such as the popularity and temporal evolution of user generated videos [6], the navigation behavior of users [17, 26], or the geographic diffusion of views triggered by social media [19]. Other studies have highlighted the potential of peer-assisted VoD systems [14, 24] to support the long tail of video popularity typically observed in UGC video services, or P2P architectures [8, 16] that exploit the relationship between viewing behavior and the graph of related videos [26].

Although particularly useful, most of these works assume that UGC video demand is uniformly distributed, with few or no geographic differences that would need to be accounted for. Similarly, despite the critical role of tags in UGC online systems [12], very few works have explored how tags relate to the viewing patterns of the videos they describe [9, 10]. The lack of works in these areas is striking as tags and geographical areas seem to drive to a large extent the sharing and consumption of UGC videos [5, 20].

Because tags capture elements of a video's semantic, they provide a promising starting point to analyze how videos with related content may be viewed and distributed geographically. For instance, tags appealing to audiences in well delineated cultural areas are likely to mark videos primarily viewed in these areas, thus helping to predict where and when the videos they describe will be consumed. Such an ability holds the promise of better UGC delivery systems, for instance by informing cache policies of Content Delivery Networks (CDNs), or by refining content look-up in peer-to-peer and peer-assisted systems. Yet, to the best of our knowledge, the link between tags and a video's geographic distribution has never been investigated.

In this paper, we fill this gap and investigate the relationship between a video's tags and the geographic distribution of its views. Based on our findings, we explore whether the distribution of videos' views can be predicted from their tags, and whether tags can help improve the storage and delivery of UGC services. We show that even a simplistic approach can predict more than 65.9% of a video's views for a majority of videos. We also illustrate how a tag-based placement can improve the hit rate of a distributed on-line video service by up to 6.8% globally, with an improvement of up to 34% in the USA.

The remainder of this paper is organized as follows. We present our analysis of tag and view distribution in Youtube in Section 2. Building on this analysis, Section 3 explores how the geographic distribution of a video's views can be predicted from its tags. Section 4 investigates whether tags can improve UGC systems by storing videos preferentially in locations where they are likely to be most viewed. Section 5 discusses the limitations of our work, Section 6 presents related approaches, and Section 7 concludes the paper.

Table 1: Popularity vector of the map of Fig. 1 (excerpt)

US	SG	SE	RO	PT	PH	PE	NL	MY	MX	IL	...
61	61	61	61	61	61	61	61	61	61	61	...

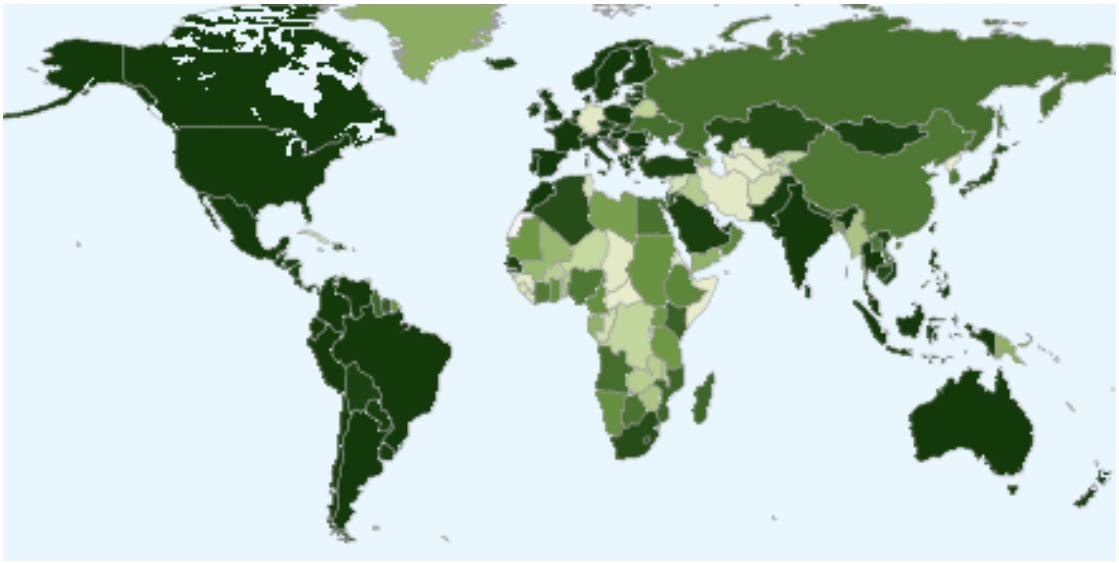


Figure 1: Popularity map of the most viewed video of our dataset *Justin Bieber - Baby ft. Ludacris*, as provided by Youtube.

2 Tags, Views, and Geodistribution in Youtube

Our study uses a Youtube dataset collected by our research group in March 2011 [16]. The seeds of the dataset are the 10 most popular videos in 25 different countries, obtained through Youtube's public API. The dataset was then completed using a breadth-first snowball sampling of the graph of related videos, as reported by Youtube. For each crawled video, the dataset contains, among others, the *video's id*, its *title*, its *total number of views*, its *popularity vector* (a vector of integers representing the video's popularity by country, more on this below), and a set of *descriptive tags* provided by the user who uploaded the video [10, 9].

The popularity vector of each video was obtained by crawling the world map which, at the time[1], was provided by Youtube to indicate in which country a video was most popular. Figure 1, for instance, shows the world map of the video with the most views in our dataset (*Justin Bieber - Baby ft. Ludacris*). Such maps were provided using Google's Map Chart service[2], making it possible to extract for each of the 235 countries of the ISO 3166-1-alpha-2 standard an integer—from 0 to 61—representing the video's popularity in this country (Table 1).

The original dataset contains 1,063,844 unique videos, but not all videos have a complete set of metadata. As a result, we filter out all videos containing no tags (6,736 videos), or with an incorrect or empty popularity vector. This filtering step results in a dataset with 590,897 videos, associated with 705,415 unique tags, totaling 173,288,616,473 views.

[1]This information is unfortunately no longer available since YouTube changed their API and graphical user interface in September 2013, and closed access to the geographic information regarding a video's views.

[2]`https://developers.google.com/chart/image/docs/gallery/map_charts`

In the following, we first present a number of notations and concepts we will use in the remainder of the paper (Sec. 2.1), explain how we extracted views from popularity vectors (Sec. 2.2), and discuss the metrics we are interested in (Sec. 2.3). We then turn to our description and analysis of the dataset in terms of views, tags, and geographic distribution (Sec. 2.4 and following).

2.1 Notation

$\mathcal{V}$ is the set of videos in our dataset. For each video $v \in \mathcal{V}$ we use the following three pieces of information:

- $tags(v)$ is the set of tags attached to the video by the user who uploaded it. For instance, the most viewed video in our data set (Figure 1) is associated with the tags *Justin*, *Bieber*, *Island*, *Def*, *Jam* and *Pop*.
- $tot\_views(v)$ is the total number of views of the video;
- $\mathbf{pop}(v)$ is popularity vector of the video as provided by Youtube. $\mathbf{pop}(v)[c]$ is the integer representing the popularity of v in country c.

From this information, we compute for each tag t the following sets and statistics:

- $videos(t)$ is the set of videos containing t in their tag set.

$$videos(t) = \left\{v \in \mathcal{V} \mid t \in tags(v)\right\} = tags^{-1}(t)$$

- $freq(t)$ is the number of occurrences of t, i.e.

$$freq(t) = |videos(t)|$$

- $tot\_views(t)$ is the total number of views associated with t, i.e. the aggregated number of views of the videos containing t.

$$tot\_views(t) = \sum_{v \in videos(t)} tot\_views(v)$$

2.2 From popularity to number of views

The exact meaning of the popularity vector $\mathbf{pop}(v)$ is not documented by Youtube. This vector is however unlikely to capture the proportion of a video's views originating from individual countries: applied to Table 1, this assumption would imply that the video *Justin Bieber - Baby ft. Ludacris* has been viewed as many times in the USA (*US*, population 318.5M) as in Singapore (*SG*, population 5.4M).

Instead, taking cue from *Google Trends*[3], one of the analytics services provided by Youtube's parent company Google, we consider a video's popularity vector to represent the *intensity* of this video in individual countries, i.e. a number proportional to the share of this video's views in this country's Youtube traffic:

$$\mathbf{pop}(v)[c] = \frac{\mathbf{views}(v)[c]}{\mathbf{ytube}[c]} \times K(v) \tag{1}$$

where $\mathbf{views}(v)[c]$ is the number of views of v in country c, $\mathbf{ytube}[c]$ is the total number of Youtube views in country c, and $K(v)$ is a normalization factor, dependent of each video, to scale values in the range $[0 - 61]$.

Neither $\mathbf{ytube}[c]$ nor $K(v)$ are available to us. To estimate both, we use the distribution of Youtube traffic provided by Alexa Internet Inc.[4] on July 2014, an authoritative source of

[3] `http://www.google.com/trends/`

[4] `http://www.alexa.com/`

internet traffic, to approximate the distribution of Youtube views per country:

$$\mathbf{ytube}[c] = \mathbf{p}_{yt}[c] \times T_{yt} \simeq \widehat{\mathbf{p}}_{yt}[c] \times T_{yt} \tag{2}$$

where $\mathbf{p}_{yt}[c]$ is the proportion of Youtube view in country c at the time our dataset was collected, T_{yt} is the total number of Youtube views at the same time, and $\widehat{\mathbf{p}}_{yt}[c]$ is the Youtube traffic estimated by Alexa for country c.

We also use the fact that we know the total number of views of each video in our dataset:

$$tot\_views(v) = \sum_{c \in Wolrd} \mathbf{views}(v)[c] \tag{3}$$

Injecting (2) in (1), and (1) in (3) eliminates $\mathbf{ytube}[c]$, $K(v)$ and T_{yt}, and yields the following formula:

$$\mathbf{views}(v)[c] \simeq \frac{\widehat{\mathbf{p}}_{yt}[c] \times \mathbf{pop}(v)[c]}{\sum_{\gamma \in World} \left(\widehat{\mathbf{p}}_{yt}[\gamma] \times \mathbf{pop}(v)[\gamma]\right)} \times tot\_views(v) \tag{4}$$

This formula provides us with the geographic distribution of the views of each videos. For each tag t, we derive from these distributions the number of views associated with t in country c (noted $\mathbf{views}(t)[c]$), i.e. the aggregated number of views in country c of the videos containing t as tag.

$$\mathbf{views}(t)[c] = \sum_{v \in videos(t)} \mathbf{views}(v)[c] \tag{5}$$

2.3 Metrics

In this analysis, we are particularly interested in capturing a tag's geographic spread (resp. concentration), and in contrasting this spread to the videos associated with this tag. To this aim, we use Shannon's entropy $H(t)$ on the view distribution of a tag t (resp. video v) among countries:

$$H(x) = -\sum_{c \in World} \mathbf{p}_{\text{geo}}(x)[c] \times \log_2\left(\mathbf{p}_{\text{geo}}(x)[c]\right) \tag{6}$$

where x is either a video or a tag, and $\mathbf{p}_{\text{geo}}(x)[c]$ represents the proportion of views of this video or tag in country c:

$$\mathbf{p}_{\text{geo}}(x)[c] = \frac{\mathbf{views}(x)[c]}{tot\_views(x)}$$

A high entropy means a tag (or video) tends to be spread uniformly among many countries. By contrast, a low entropy denotes a tag (video) whose views are concentrated in a few countries. For instance, the video with the highest number of views in our dataset, *Justin Bieber - Baby ft. Ludacris* shown in Figure 1, has an entropy of 5.06. This value is close to the highest possible value of $log_2(235) = 7.87$, which would correspond to a video equally distributed among the 235 countries tracked by Youtube. By contrast, the lowest possible entropy value is $log_2(1) = 0$, corresponding to a tag (video) whose views originate from one single country.

Table 2: The 10 most frequent tags

tag	#occur	#views	average #views
the	30686	13,157,705,562	428,785
video	27239	12,898,383,171	473,526
music	23128	12,640,171,764	546,531
2010	22014	3,349,620,292	152,158
funny	21645	13,550,709,569	626,043
of	19820	5,940,302,641	299,712
new	17943	5,293,119,879	294,996
2011	14572	756,842,996	51,938
live	11614	3,196,117,558	275,195
de	11314	2,726,151,223	240,953

Table 3: The 10 most viewed tags (worldwide)

tag	#occur	#views	average #views
funny	21645	13,550,709,569	626,043
pop	7877	13,318,507,233	1,690,809
the	30686	13,157,705,562	428,785
video	27239	12,898,383,171	473,526
music	23128	12,640,171,764	546,531
of	19820	5,940,302,641	299,712
records	2478	5,920,162,042	2,389,088
hip	5085	5,615,505,842	1,104,327
hop	5047	5,615,431,517	1,112,627
comedy	9039	5,603,654,002	619,941

Table 4: The most viewed tags for various countries

country	tag	total views
United-States	funny	7,907,521,226
Germany	music	557,388,816
France	pop	536,096,206
Canada	funny	484,758,340
Australia	funny	236,812,186

2.4 Tag and view distributions

Our dataset contains 7,717,815 tag occurrences, yielding an average number of 11.18 tags per video. These tag occurrences encompass 705,415 unique tags, a large number in line with earlier findings [9]. This large number of unique tags can be explained by the presence of compound tags (e.g. "*korean pop*" is different from "*korean*" "*pop*", which counts as two tags), spelling mistakes ("*(music*" or "*music␣*" instead of "*music*"), and the use of multiple languages.

A sample of the 10 most frequent tags is shown in Table 2, and the top 10 tags with the most views in Table 3. The two tables highlight a few noteworthy features of tag usage in Youtube: although some grammatical words are present (*the*, *of*), most tags are about content (*video*, *funny*). Grammatical words can be explained by the former usage of spaces in Youtube to separate tags (commas are now used), which led tags intended as compound terms such as *the␣rock* to be parsed into two tags (*the* and *rock*). The most viewed tags are not necessarily the most frequent: this is in particular true of *pop*, the second most viewed tag (Table 2), with only 7877 occurrences. The corresponding videos are predominantly from the category "Music", with a high average number of views per individual video (1,690,809 views, 2.7 times more often than videos containing the tag *funny* for instance). The same comment applies to related tags such as *hip* and *records*.

The frequency distribution of individual tags (Figure 2) shows a typical power-law, which is commonly found in natural languages and folksonomies. About 462,549 tags (66%) only appear once.

The tag descriptions of individual videos are relatively rich (Figure 3), with an average of 11.18 tags per video, as mentioned earlier. One reason why videos usually possess a reasonable number of tags might be because users have an incentive to tag their videos to attract more views.

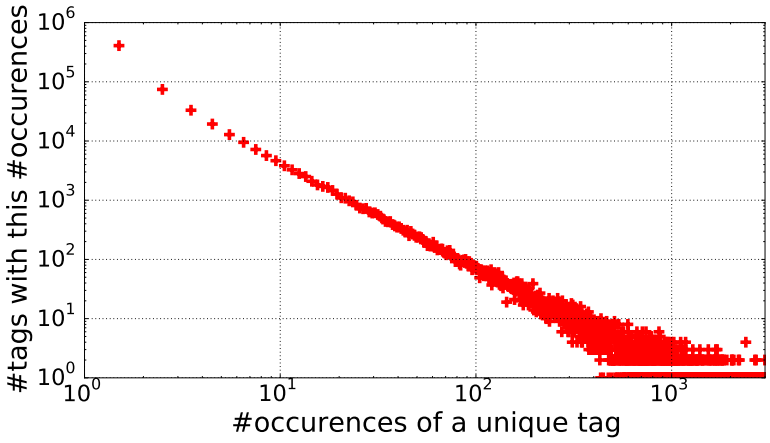


Figure 2: The frequency distribution of tags follow a power law of the shape $y = K \times x^{-\alpha}$, as often observed in folksonomies and natural languages.

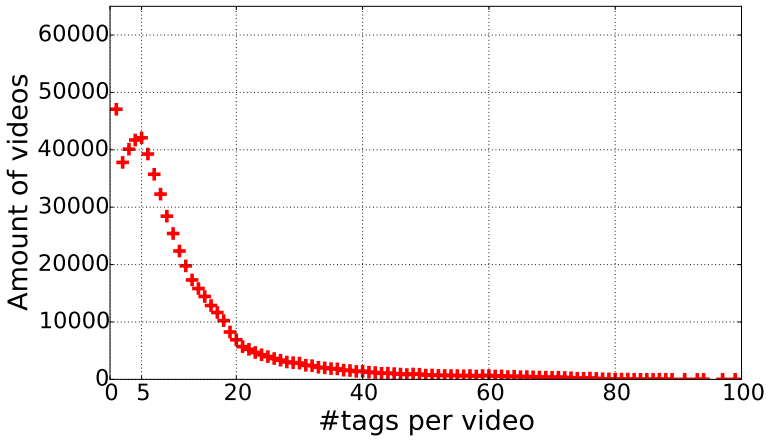


Figure 3: Tags are widely used to describe videos, with 50% of videos showing a least 11 tags.

However, and perhaps surprisingly, there seems to be only a weak link between the number of tags of a video and this video's viewership (Figure 4). Although up to 18 tags the median number of views of a video increases with its number of tags, this relationship collapses beyond 18 tags. For instance, the most tagged video in our dataset possesses 102 tags, but has only been viewed 1,220,496 times, which pales in comparison to the most viewed video (471,208,788 views), which only sports a modest 6 tags.

This weak or absent correlation between number of tags and number of views is also apparent in Figure 5, which shows the proportion of aggregated views, as a function of the proportion of videos categorized in different ranges of tag numbers. With small variations, all categories of videos show the same heavily skewed distribution, with 10% of the most viewed videos accounting for slightly over 80% of the views, a finding reported in other Youtube datasets [6].

In the following, in order to avoid artifacts caused by videos with very low numbers of views,

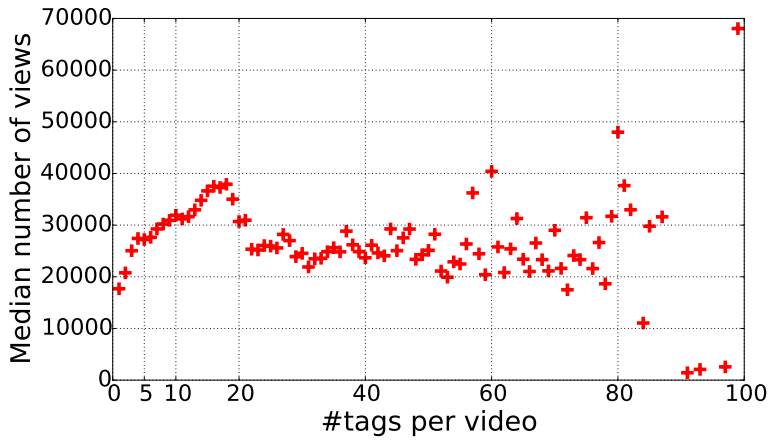


Figure 4: Median number of views for the videos embedding a given number of tags. Views and size of the tag set seem only weakly correlated, with a clear growing trend limited to videos with less than 18 tags.

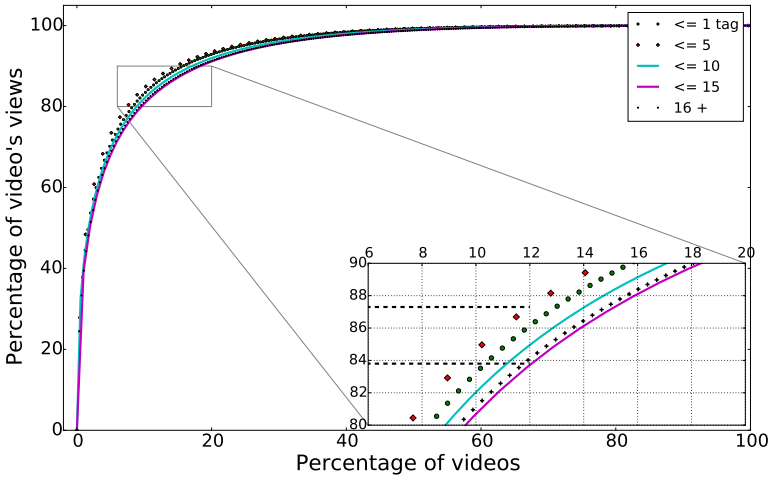


Figure 5: Cumulative distribution for video's views, by number of tags. As expected our data-set shows a long tail of videos with few views.

we only consider videos with at least 1000 views. We also limit our discussion to iso-latin1 tags (91.03% of all tag occurrences).

2.5 Geographical distribution of tags

In terms of geographic distribution, the tags most viewed globally (Table 3) also tend to be those most viewed in individual countries. For instance Table 4 shows the one most viewed tag in five western countries (France, Germany, Canada, Australia, and USA). This tag is either *music*, *pop* or *funny*, which all appear in Table 3.

Table 5: Top 5 countries (by views) for *pop*

country	#views	%age
United-States	4,700,159,350	35.2%
United-Kingdom	759,449,112	5.7%
Brazil	751,342,295	5.6%
Mexico	603,876,310	4.5%
India	586,339,771	4.4%

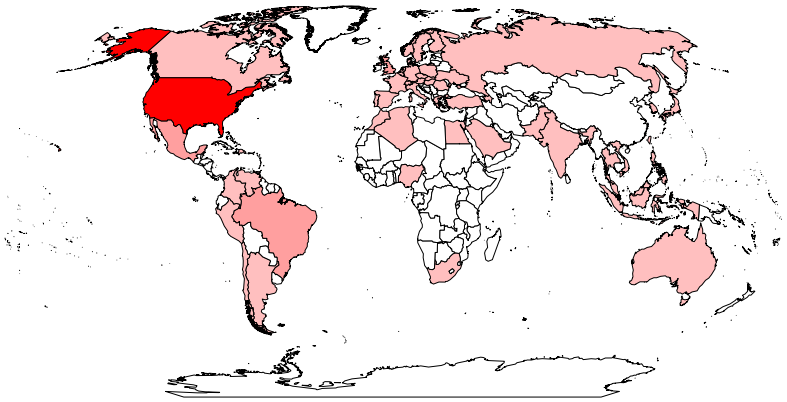


Figure 6: Videos associated with the tag 'pop' tend to be uniformly distributed over the globe, taking into account the number of YouTube users in a country.

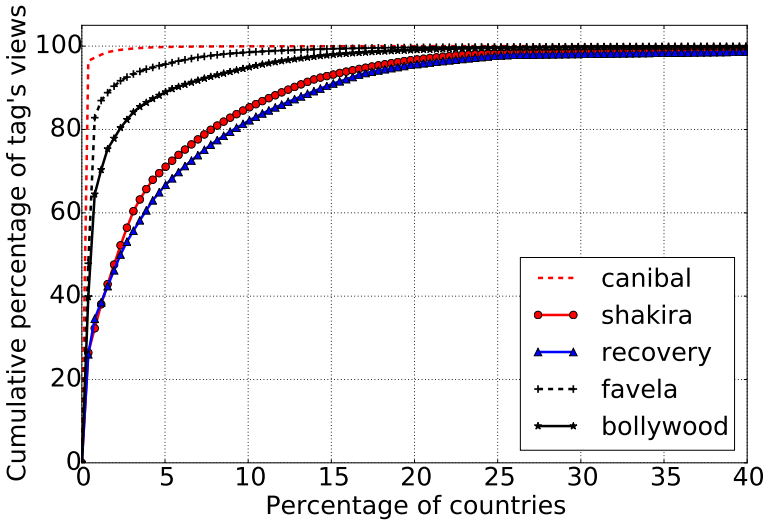


Figure 9: Whereas views associated with 'canibal' or 'favela' are concentrated in a very few countries, the other tags tends to be more uniformly spread. 'recovery' is the most equally spread tag in the dataset.

Individual tags, however, can present starkly different geographic distributions, with some

Table 6: Top 5 countries (views) for *bollywood*

country	#views	%age
India	200,956,055	39.8%
United-States	124,461,447	24.7%
United-Kingdom	29,506,586	5.8%
Pakistan	25,218,518	5.0%
Germany	12,842,983	2.5%

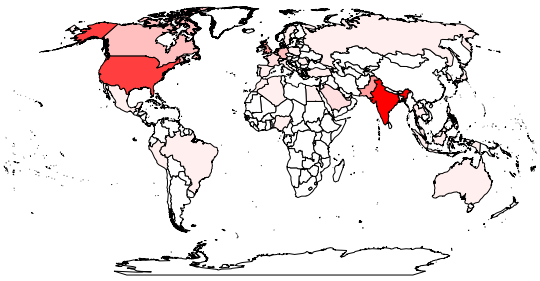


Figure 7: Videos associated with the tag 'bollywood' tend to be viewed mainly in India, USA and UK, with some secondary spots in some asian and european countries (Pakistan, Germany, ...)

Table 7: 5 top countries (views) for *favela*

country	#views	%age
Brazil	19,834,633	47.9%
United-States	14,468,608	34.9%
United-Kingdom	1,701,496	4.1%
Canada	785,725	1.9%
Mexico	639,375	1.5%

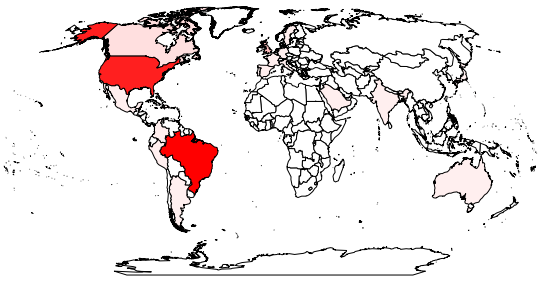


Figure 8: Videos associated with the tag 'favela' are mostly viewed in Brazil

Table 8: Top 3 Videos (views) containing *pop*

title	#views	%
Justin Bieber - Baby ft. Ludacris	471,208,788	3.54%
Lady Gaga - Bad Romance	348,924,582	2.62%
Shakira - Waka Waka ...	306,374,501	2.30%
total for top 3	*1,126,507,871*	*8.46%*

tags widely spread, and others concentrated in only a few countries. This is illustrated in Figure 9, which shows the cumulative views of a selection of tags when ranking countries in descending order according to the number of views obtained by each tag in each country (Figure 9). Curves close to the top left corner (*favela*, *canibal*) represent tags whose views are concentrated in a few countries. For example, 80% of views associated with *favela* originate from only 1% of the world's countries. By contrast, curves towards the diagonal (*shakira*, *recovery*) highlight more evenly spread tags.

In the following, we discuss in more detail these two cases (widely spread vs. concentrated) by considering more closely the three tags *pop*, *bollywood*, and *favela*. The top 5 viewing countries for these tags are shown on Tables 5-7, whereas the distribution of their viewership is mapped in Figures 6-8. On these maps, a higher color saturation indicates a higher proportion of views for the corresponding country.

Views associated to the tag *pop* (entropy 4.25) tend to be broadly distributed over the world (Table 5 and Figure 6). The country with the most views associated with *pop* are the USA,

representing 35.2% of the total number of views for *pop*.

The case of *pop* is interesting on two further accounts. First, three videos among the 7877 ones associated with *pop* are responsible for almost 10% of the total amount of views for that tag (Table 8). It turns out that these three videos are also the three most viewed videos of the entire dataset. As a result, precisely predicting the actual distribution of these videos is less important than predicting early that they will be widely viewed, and will be viewed on a global scale [6].

By contrast tags such as *bollywood* (Table 6 and Figure 7) and *favela* (Table 7 and Figure 8), are much more concentrated on a few countries, which is reflected in their entropy scores: 3.24 for *bollywood*, and 2.22 for *favela*.

The views of the tag *bollywood* are mostly centered in India and United-States (64.5%), as expected for cultural and language reasons, with three additional countries with important South Asian minorities accounting for another 11.3% of all *bollywood* views. By preferably caching, or by placing proactive copies of videos containing *bollywood* close or in these five countries, a UGC video service would cover 75.8% of all views for these tags, a substantial share of their induced traffic.

The distribution of *favela* is even more concentrated than *bollywood*, with Brazil responsible for almost 48% of all views, followed by the United-States as a distant second with 34.9% (over a total of 41,417,318 views, Table 7). The tag *favela* is also concentrated on only a few videos: the three most viewed videos containing that tag account for 22.6% of the tag's total views (respectively 8.1%, 7.6% and 6.9%). In that case, placing or conserving copies of videos containing this tag in South America would seem particularly beneficial.

2.6 Entropy analysis

To investigate systematically and comprehensively how Youtube tags are distributed we now turn to entropy. We will apply this metric to tags and videos, in order to characterize their geographic distribution.

Figure 10 shows the cumulative distribution function (CDF) of the entropy of videos (solid line) and tags (dashes) in our data set. The two curves are similar: entropy values tend to be evenly spread for values below 3 (which corresponds to roughly 80% of all videos and tags), with a higher concentration in the range $[3, 4]$. Only a few percents of videos and tags have an entropy beyond that range, e.g. only 2.81% of all videos have an entropy higher than 4.

These numbers highlight the substantial share of videos with low entropy (i.e. whose views are geographically concentrated): 40% of all videos have an entropy lower than 1.5. As an intuitive reference point, this is the value a video would obtain if it were only present in 4 countries, and uniformly distributed in those 4 countries.

Figure 11a investigates the relation between a video's entropy and its number of views. As expected, and as pointed out in earlier work [16], popular videos, in particular beyond 10^6 views, tend to have a high entropy, meaning their views are widely distributed. This is also somewhat true for lower numbers of views, with a concentration of videos whose views are between 10,000 and 200,000 and whose entropy is around 2.5. Outside this region of high concentration, and for videos with less 10^6 views, entropy values tend to be equally distributed, with two smaller concentrations of videos around entropy values of 1.5 and 0.

These distributions mean that highly popular videos need on average to be accessible from all over the world, since their entropy is high. Less viewed videos are different: quite a few of them have low entropy values, and would benefit most from an accurate prediction of their geographic distribution.

Our argument in this paper is that this information can be contributed by tags, at least in part. The more tags a video possesses, the more likely these tags might be able to help predict

Table 9: The 5 tags with the most (left) resp. least (right) entropy (for #occurrences > 100)

tag	H(t)	#occ.	#views	av. views	tag	H(t)	#occ.	#views	av. views
recovery	4.90	230	557,870k	2,425,523	piologo	0.04	101	3,985k	39,458
dominic	4.87	103	338,555k	3,286,944	mundo	0.06	134	4,147k	30,954
fifa	4.83	2722	690,092k	253,524	kvarteret	0.10	102	7,313k	71,700
passat	4.79	142	41,809k	294,432	skatan	0.11	106	7,741k	73,030
afraid	4.78	131	244,659k	1,867,633	partoba	0.18	272	7,183k	26,408

its distribution. As shown in Figure 11b, a video's entropy (and hence geographic spread) is not strongly linked to the number of its tags $|tags(v)|$. Videos with less than 5 tags tend to have a low entropy, with a high density point below 0.25 of entropy and a lower density point between 2 and 2.5. This observation reveals that an important proportion of videos have a low entropy, and tend to be linked to a smaller number of tags. These videos represent an important part of the dataset: if we consider the lower quarter of the entropy values, videos with an entropy value in this range represent 38% of the dataset, with an average number of 155,520 views, a mean number of tags of 9 (vs. 11.18 for the whole dataset), and a mean entropy of 0.707.

Moving on to tags, Figure 12 shows the relationship between the mean number of views of a tag's videos and its entropy value, plotted on a density graph. Similarly to videos, the highest concentration of tags is found for entropy values around 2, and average views of $100,000$, but the spread of tags in the rest of the graph remains substantial, i.e. in the area with an entropy < 3 and average number of views between $10,000$ and $200,000$.

The fact that almost a third of videos have a low entropy, and are thus mainly distributed in only few countries, hints at the possibility and interest of predicting these videos' geographic distribution.

Figure 13 highlights the potential of tags in doing so: the figure plots the mean entropy of

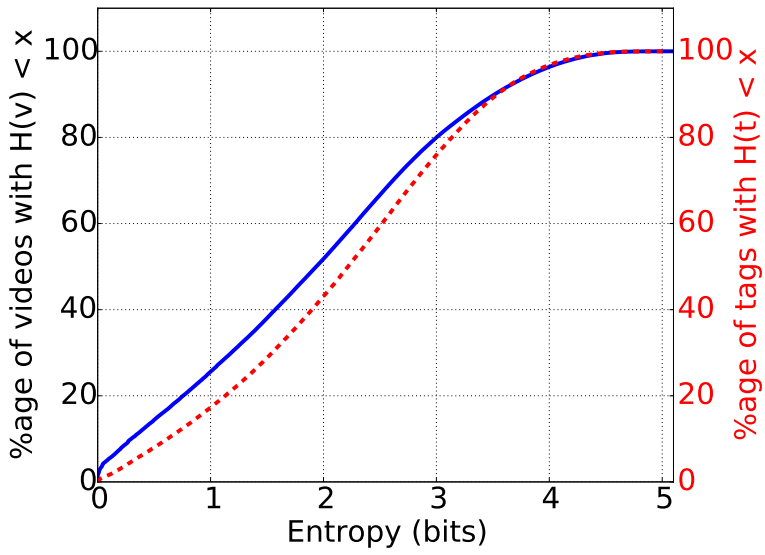


Figure 10: CDF of videos (solid line) and tags (dashes) versus entropy

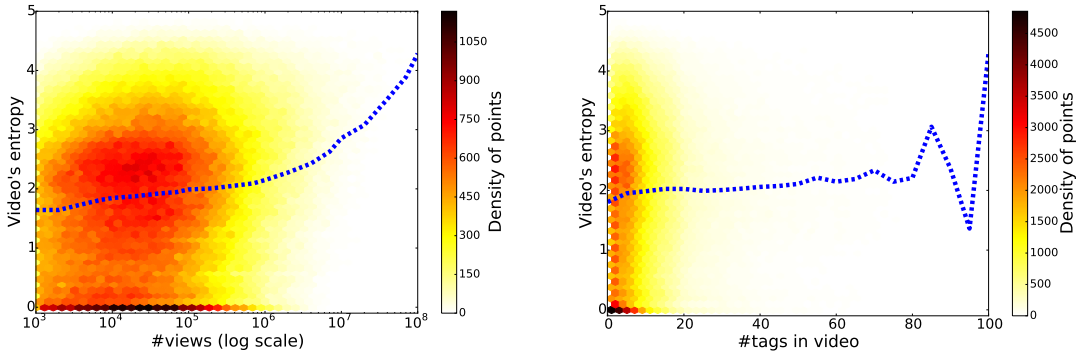


Figure 11: Heatmap of each video's entropy vs. its number of views (left), resp. its number of tags (right). Mean shown as a dashed line.

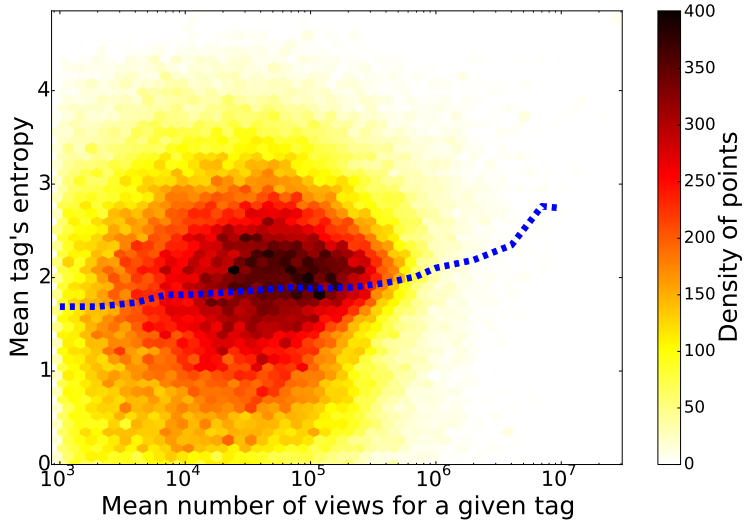


Figure 12: Heatmap of the mean views for every occurrences of a given tag, versus the mean entropy of every occurrences of that tag. Mean showed as a dashed line.

each unique tag versus the mean entropy of all the videos this tag appears in. The plot exhibits mainly a linear shape. For most pair (tag, video), the tag's entropy and the video's entropy are strongly correlated. This strong link leads us to conjecture that the geographic distribution of a video's views might be predictable from that of its associated tags. This is the very problem we turn to in the next section.

3 Predicting Views from Tags

The analysis in the previous section highlights a strong correlation between the distribution of tags and that of videos. In this section, we go one step further, and explore the potential of tags in predicting a video's view distribution.

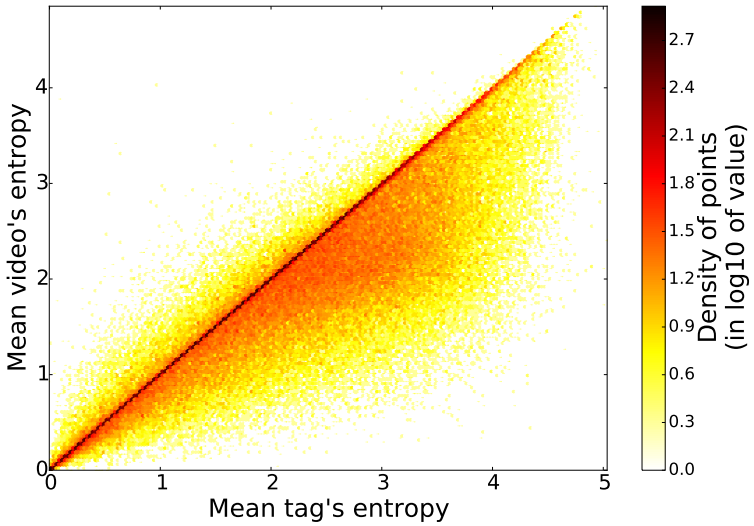


Figure 13: Tag entropy versus video entropy

We use a basic additive prediction technique that exploits the tags associated with videos, and compare it with a baseline prediction mechanism. To evaluate both approaches, we use a cross-validation technique. We split the dataset into a testing set and a training set. We then process the information (views and tags) in the training set, and use it to guess the view distributions of the videos in the testing set. In the following, we first detail our tag-based prediction approach, then present our experimental methodology, and finally, we present our results.

3.1 Tag-Based Prediction

For a video $v \in \mathcal{V}_{\text{test}}$ associated with a set of tags $tags(v)$, we predict the geographic distribution of v's views $\widehat{\mathbf{p}_{\text{geo}}}(v)$ as the average of the geographic distribution of v's tags in the training set $\mathcal{V}_{\text{train}}$:

$$\widehat{\mathbf{p}_{\text{geo}}}(v) = \mathbb{E}_{t \in tags(v)} \left(\mathbf{p}_{\text{geo}}^{\mathcal{V}_{\text{train}}}(t) \right) \tag{7}$$

where $\mathbf{p}_{\text{geo}}^{\mathcal{V}_{\text{train}}}(t)$ is the geographic distribution vector of tag t in the dataset $\mathcal{V}_{\text{train}}$ (which does not take into account the videos of $\mathcal{V}_{\text{test}}$). Our aim is for $\widehat{\mathbf{p}_{\text{geo}}}(v)$ to be as close as possible to $\mathbf{p}_{\text{geo}}(v)$, v's actual view distribution vector.

3.2 Evaluation Methodology

Baseline Prediction We compare our tag-based prediction technique with a simple baseline approach inspired by the data provided by Alexa Internet Inc. [25]. Alexa provides a list of the top 40 countries that view the most YouTube videos along with their percentage share of YouTube videos—Table 10 lists the data for the top 10 countries. We use this data as viewing probabilities: with reference to Table 10, a video has a 19% chance to be viewed in the USA, an 8.6% chance to be viewed in India, and so on. This allows us to assign probabilities to 40 of the 257 countries. For the remaining ones, we use the data about the number of internet users per country made available by the International Telecommunication Union [2]. We first observe that the probabilities in the top-40 list sum up to 85.2%. We then distribute the remaining 14.8%

Table 10: The 10 countries viewing the most videos according to Alexa

country	share
United-States	19.0%
India	8.6%
Japan	4.7%
Russia	4.1%
Brazil	3.8%
United-Kingdom	3.2%
Mexico	3.0%
Germany	3.0%
France	2.5%
Spain	2.3%

to the unassigned countries proportionally to their share of internet users. This process yields a baseline view prediction that is independent of the dataset and of the particular video.

Cross validation We evaluate the tag-based prediction strategy and compare it with the baseline using a cross-validation mechanism. We divide our dataset into two equal parts: a training set and a testing set. We first order the videos in the dataset by number of views. We then go through the sorted dataset starting from the most seen video, and alternatively assign one video to the training set and the next to the testing set. This process yields a training and a testing set consisting respectively of 295449 and 295448 videos, and accounting each for approximately 50% of the views.

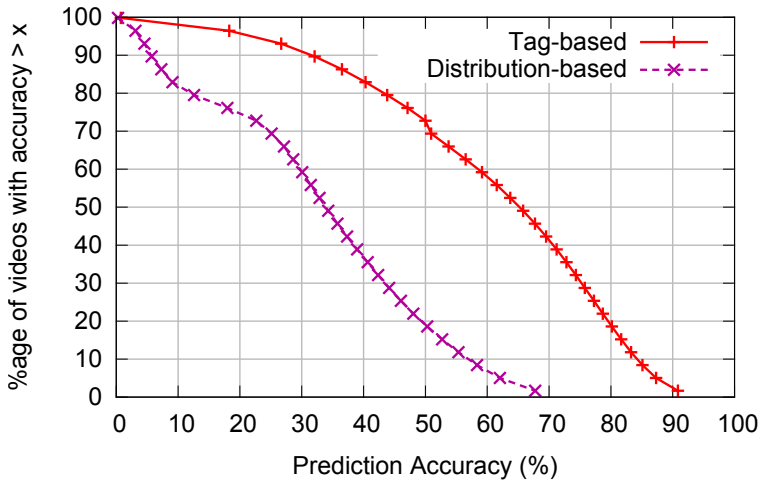
According to cross-validation, we use the training set to *train* the prediction approaches. In the case of our tag-based prediction, we go through the videos in the training set to compute the geographical distribution of tags: $\mathbf{p}_{\text{geo}}^{\mathcal{V}_{\text{train}}}(t)$. For the baseline approach, we instead simply ignore the training set since the baseline prediction does not depend on the dataset or on the specific video.

After the training phase, we test both prediction approaches on the testing set. For each video in this set, we take the total number of views from the dataset as an input and distribute it across the countries as dictated by the prediction approach—in other words we multiply the view probability of each country by the total number of views of the video. For our tag-based approach, we obtain a probability distribution from the geographical distribution of tags as in Equation (7); for the baseline, we use the video-independent probability distribution described in Section 3.2. Note that our tag-based prediction approach yields a specific prediction for each video, while the baseline applies the same prediction to all videos.

Metrics To evaluate the effectiveness of prediction, we measure the divergence between a prediction $\widehat{\mathbf{p}_{\text{geo}}}(v)$ and the geographic distribution of a video $\mathbf{p}_{\text{geo}}(v)$ (for $v \in \mathcal{V}_{\text{test}}$). Then, we compute the proportion of views misplaced by the prediction, $\mathbf{p}_{\text{wrong}}(v)$.

$$\mathbf{p}_{\text{wrong}}(v) = \frac{1}{2} \times \sum_{c \in World} \left| \mathbf{p}_{\text{geo}}(v)[c] - \widehat{\mathbf{p}_{\text{geo}}}(v)[c] \right| \tag{8}$$

We divide the sum in Equation (8) by 2 to avoid counting misplaced views twice (once where the views should have been, and another time where they have been wrongly placed). We then



	accuracy	
method	*mean*	*median*
Tag-based prediction	61.3%	65.9%
Distribution-based prediction	32.9%	33.9%

Figure 14: CDF of prediction accuracy (top) and mean and median (bottom) for the tag-based and distribution-based approaches for view prediction (higher is better). Tags clearly yield better predictions over a simple average distribution vector.

define our final metric, *prediction accuracy*, as the complement of the proportion of misplaced views.

$$\mathbf{p}_{\text{accurate}}(v) = 1 - \mathbf{p}_{\text{wrong}}(v) \tag{9}$$

An accuracy of 1 means that the prediction and the actual distributions match; a value of 0 instead indicates there is no overlap in terms of countries between the predicted and actual views.

3.3 Results

We start presenting our results by comparing the distributions of prediction accuracy for our tag-based approach and for the baseline view prediction. Figure 14 plots the cumulative distribution functions and shows the corresponding mean and median values. Our tag-based approach clearly outperforms the baseline view prediction. The plot shows that the baseline yields a prediction accuracy above 60% for only 7% of the videos, compared to nearly 60% for our tag-based approach. The table complements this result by conveying a mean and an average accuracy for the tag-based approach respectively of 61.3% and 65.9%. In other words, our approach can predict at least 65.9% of the view locations for a majority of videos, while this number drops to 33.9% with the baseline.

These results confirm our original assumption: tags hold the promise of predicting accurately the geographic distribution of UGC videos. This is particularly encouraging if we consider the simplicity of our technique, which does not attempt to distinguish between tags, or perform any kind of regression. In the following we continue our analysis by exploring the correlation between accuracy and parameters such as entropy, popularity, and the number of tags.

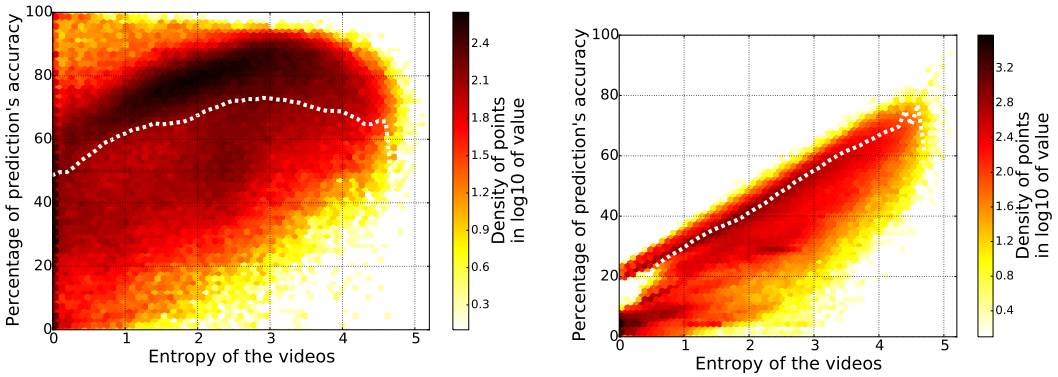


Figure 15: Prediction accuracy vs video entropy for the tag-based approach (left) and for the baseline (Right). Dashed lines depict the average accuracy for a given entropy value.

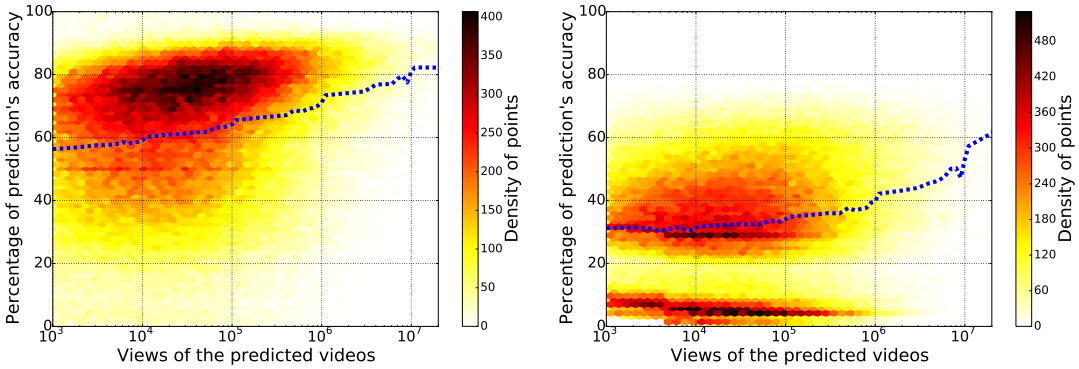


Figure 16: Views of predicted videos versus prediction's accuracy for the tag-based approach (left) and for the baseline (Right). Dashed lines depict the average accuracy for a given number of views.

Accuracy vs entropy We now start dissecting our results by examining the correlation between the accuracy of predictions (with either approach) and the entropy of the videos being predicted. To compute the entropy of a video, we apply Equation (6) while using the real number of views from the dataset, not the number of predicted views. Figure 15 shows the results for our tag-based mechanism and for the baseline view prediction, respectively on the left and on the right plot.

A quick comparison between the plots shows that our tag-based predictor consistently achieves better performance over all the entropy spectrum, except for a small decrease in accuracy for very high entropy values at the right end of the plot. The views of high-entropy videos tend to be rather uniformly spread across all countries, which corresponds to the assumptions implicitly made by the simple baseline approach. For all other entropy values however, the baseline cannot follow the non-uniform distribution of videos across countries.

The tag-based predictor, on the other hand is able to achieve good accuracy even for videos that are viewed in only a few countries. For example, it reaches an average accuracy of nearly 50% for an entropy value of 0, which corresponds to videos viewed in a single country. For entropy

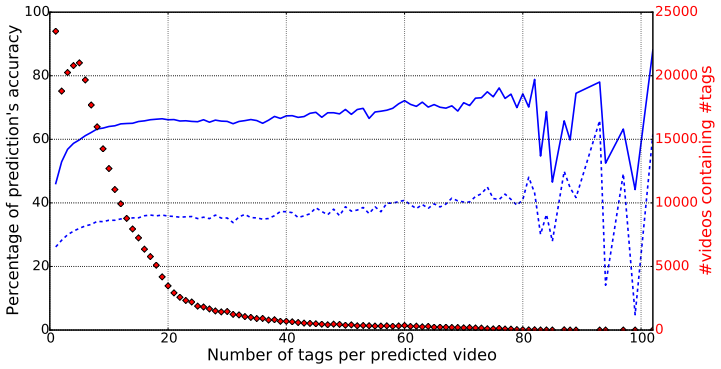


Figure 17: Accuracy of prediction vs. nb. of tags associated with each video (solid line for our tag-based, dashed for the baseline), overlapped with distribution of video with a given number of tags (points)(50% of videos removed)

values around 3, corresponding to videos viewed in only 8 countries, the average prediction accuracy reaches instead peaks of 73%, with a large number of videos faring over 80%. Finally, the tag-based plot shows that some videos reach accuracy values above 90% over almost all the entropy spectrum. The baseline obtains instead much lower maximum accuracy values.

Accuracy vs popularity Next, we study the correlation between prediction accuracy and the real number of views (popularity) of predicted videos. Figure 16 shows the results for our tag-based approach (left) and for the baseline view prediction (right). Tag-based prediction significantly outperforms the baseline view prediction regardless of popularity. The absolute difference between the accuracy values of the two approaches remains at about 30% over all popularity values

While the advantage of our tag-based prediction approach appears largely independent of popularity, both plots show a positive correlation between the number of views of a video and the accuracy of the predictions about it. Yet this correlation is weaker than in the accuracy-vs-entropy case, and appears almost identical for both prediction approaches. The reason lies in the correlation between popularity and entropy. Highly popular videos—those with a large number of views—tend to be scattered all over the world (high entropy), which makes them much easier to predict than low-entropy videos.

Accuracy vs number of tags Since we are proposing a tag-based prediction approach, it appears natural to study the relationship between the number of tags and the accuracy of predictions. Figure 17 groups videos by number of tags and compares the average prediction accuracy in each group of our tag-based approach and the baseline view prediction. Our approach shows a strong correlation between the number of tags and accuracy for up to 10 tags. Accuracy varies from 45% with one tag to over 66% with 10 tags. The curve flattens out for higher numbers of tags suggesting that a relatively small number of tags suffice to perform reliable predictions.

Albeit for lower accuracy values, the curve for the baseline also shows a slight increase in accuracy when moving from 1 to 10 tags. However, this smaller increase results from the fact that more popular videos, tend to be associated with more tags, and popularity positively correlates with prediction accuracy as shown in Figure 16.

Finally, both curves present an erratic behavior for videos with very high numbers of tags. The reason lies in the low number of videos corresponding to each of the points to the right of the figure. As shown by the decreasing dotted line in Figure 17, the vast majority of videos have less than 20 tags, and each of the points to the right of the plot corresponds to a small quantity of videos, with 185 videos embedding between 80 and 85 tags, 53 between 85 and 90 and only 12 videos with more than 90 tags..

4 Using tags for proactive video placement

The previous experiments clearly show that tags convey information about the location of a video's views. In this section, we explore whether this information can improve UGC systems by storing videos preferentially in locations where they are likely to be most viewed.

4.1 System model and storage capacity

Our scenario assumes that a company, e.g. YouTube, manages a set of datacenters located in different countries, and must decide where to store the primary copies of individual videos (i.e. these copies form the reference storage of the UGC service, in contrast to caching copies, which might be evicted). To test the usefulness of tags, we consider a somewhat extreme case, in which each country possesses its own storage infrastructure (a datacenter, or share of datacenter for small countries).

As in the previous section, we split our dataset in two, using the same reference ($\mathcal{V}_{\text{train}}$), and testing sets ($\mathcal{V}_{\text{test}}$). Due to the size of $\mathcal{V}_{\text{test}}$ (295448 videos, and 86,624,310,171 views), we sampled down $\mathcal{V}_{\text{test}}$ while conserving the distribution of views across countries and tags. This works as follows: we first generate a trace $\mathcal{T}$ of 10 millions video requests for the videos of $\mathcal{V}_{\text{test}}$ that respect the distribution of views between videos and countries. In other words, the probability to generate a request for video v in a country c in $\mathcal{T}$ is proportional to the number of views of v in c:

$$P(\text{generate } request(v,c)) = \frac{\mathbf{views}(v)[c]}{\sum_{v' \in \mathcal{V}_{\text{test}}} tot\_views(v')}$$

We then choose $\mathcal{V}_{\text{expe}}$ as the set of unique videos present in the trace $\mathcal{T}$.

Our goal consists in finding a good placement for the copies of the videos in $\mathcal{V}_{\text{expe}}$ by using the tag information contained in $\mathcal{V}_{\text{train}}$. A good placement is one that maximizes the number of video requests that are served from a copy stored in the country's local storage infrastructure.

We dimension the system's overall storage capacity ($\mathcal{S}_{world}$) so as to allow a total of R copies of every video from $\mathcal{V}_{\text{expe}}$ to be stored globally. $R = 3$ for instance is a typical value for R used in cloud storage systems (e.g. GFS, HFS). For simplicity's sake, we assume each video has the same size (an obvious simplification), and measure our storage capacity in number of videos.

We assume that the service's revenues, and hence its investment, will be roughly proportional to the number of views in one country. As a result, we make the storage capacity $\mathcal{S}_c$ of each country c proportional to the country's view shares:

$$\mathcal{S}_c = \mathcal{S}_{world} \times \mathbf{p}_{yt}[c]$$

where $\mathbf{p}_{yt}[c]$ is the proportion of Youtube views in country c, which we estimate using Alexa's estimations (as discussed in Section 2.2). As a result, we have

$$\mathcal{S}_{world} = \sum_{c \in World} \mathcal{S}_c = R \times |\mathcal{V}_{\text{expe}}|$$

$|\mathcal{V}_{\text{expe}}|$ is the number of videos we want to store and serve.

To improve the overall system, We add an LRU cache C_c to each country, representing 10% of the country's primary storage capacity S_c:

$$C_c = 0.1 \times S_c$$

4.2 Placement heuristics

To demonstrate the potential of tags to help organize the video storage of a UGC service, we propose to use the following simplistic approach. We first estimate, for each video $v \in \mathcal{V}_{\text{expe}}$, its per-country viewing vector $(\widehat{\mathbf{views}}(v)[c])_{c \in World}$.

For this estimation, we use the training set to compute $\mathbf{views}^{\mathcal{V}_{\text{train}}}(t)[c]$, the aggregated number of views in country c of the videos of $\mathcal{V}_{\text{train}}$ containing t as tag (Equation (5) from Section 2.2). From $\mathbf{views}^{\mathcal{V}_{\text{train}}}(t)[c]$, we then compute the average number of views in country c of the videos containing t:

$$\begin{aligned}\mathbf{views\_p\_vid}^{\mathcal{V}_{\text{train}}}(t)[c] &= \frac{\mathbf{views}^{\mathcal{V}_{\text{train}}}(t)[c]}{|\{v \in \mathcal{V}_{\text{train}} : t \in tags(v)\}|} \\ &= \mathop{\mathbb{E}}_{\substack{v \in \mathcal{V}_{\text{train}}: \\ t \in tags(v)}} \left(\mathbf{views}^{\mathcal{V}_{\text{train}}}(v)[c]\right)\end{aligned}$$

We then estimate $\widehat{\mathbf{views}}(v)$ for $v \in \mathcal{V}_{\text{test}}$as:

$$\widehat{\mathbf{views}}(v)[c] = \mathop{\mathbb{E}}_{t \in tags(v)} \left(\mathbf{views\_p\_vid}^{\mathcal{V}_{\text{train}}}(t)[c]\right) \tag{10}$$

The placement algorithm is then as follows: we simulate the uploading of videos by randomly iterating through all the videos of $\mathcal{V}_{\text{expe}}$. We then place R copies of v in the first R countries in which v is predicted to get most of its views, among the countries whose local storage infrastructure is not full yet.

4.3 Experiment, metrics, baseline

As our baseline, we use a random placement policy (noted *random placement*), which randomly allocates each of the R replicas of a video in $\mathcal{V}_{\text{expe}}$ to any country with some remaining storage capacity.

To evaluate the quality of a placement we replay the trace $\mathcal{T}$ that we used in Sec. 4.1 to generate our experimental set $\mathcal{V}_{\text{expe}}$. For each request originating from a country c, if the corresponding video is found in the primary storage S_c or in the country's cache C_c, the request counts as a hit, otherwise as a miss. In the case of a miss, the video is stored in the country's cache C_c, and the least recently used video of C_c is evicted if the cache is full. We use the hit ratio $(\#hits/(\#hits + \#misses))$ as our quality metric.

4.4 Results

We start by comparing the average hit ratios obtained by our placement approach and by the baseline across all countries for different values of R. Figure 18 shows that tag-based placement clearly outperforms the baseline with an absolute accuracy improvement that oscillates between 5.3% and 6.8%. This advantage remains fairly constant as R increases, although for very large

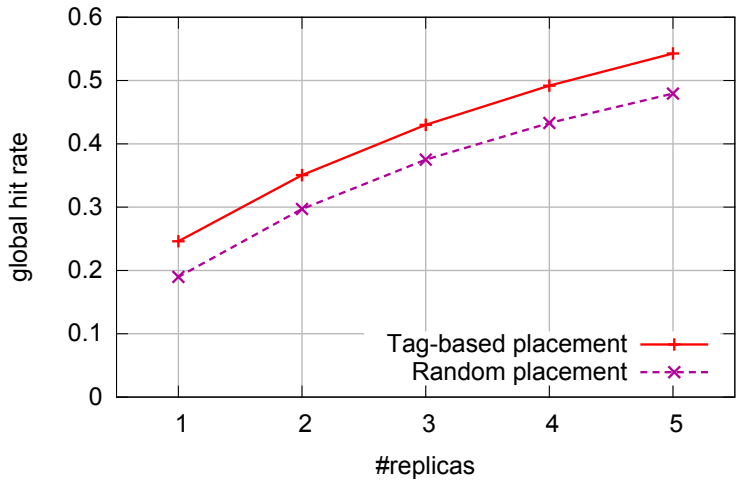


Figure 18: A basic placement strategy based on tags consistently improves the system's global hit rate by about 6% in our experiments, independently of the number of copies per video.

values of R, the difference between them will decrease. The two approaches will achieve the same 100% hit ratio when R is so large that all videos are stored in every country.

To understand how tag-based placement works, we plot, in Figure 19, the per-country hit ratio for the 6 countries that view the most videos. The solid green part of each bar shows the proportion of hits obtained through the primary storage, while the red hatched part shows those obtained via the LRU cache. The black line at the top of each bar marks the point corresponding to a hit rate of 100%. We reproduce this graph successively for R=1, R=3 and R=5. The results show that tag-based prediction provides the most advantage for countries that view the most videos. For $R = 1$, the US obtain a hit ratio of 79% with our model and only 45% with the baseline. The composition of this hit ratio also changes: our approach achieves 64% of hit ratio through the primary storage and only 15% through the LRU cache; the baseline achieves only

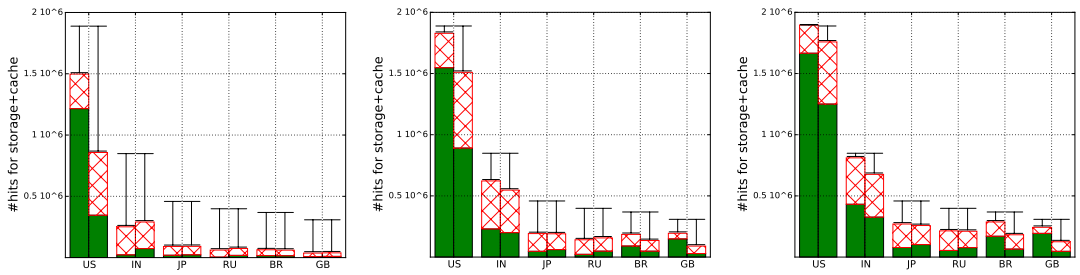


Figure 19: Number of hits for the 6 countries with the most views, for $R = 1$ (left), $R = 3$ (center) and $R = 5$ (right) replicas for each videos. For each country, the left bar shows the number of hits obtained with a tag-based placement, and the right bar with the baseline (random placement). The green and red portions denotes the contribution of the permanent storage S_c resp. of the cache C_c to hits.

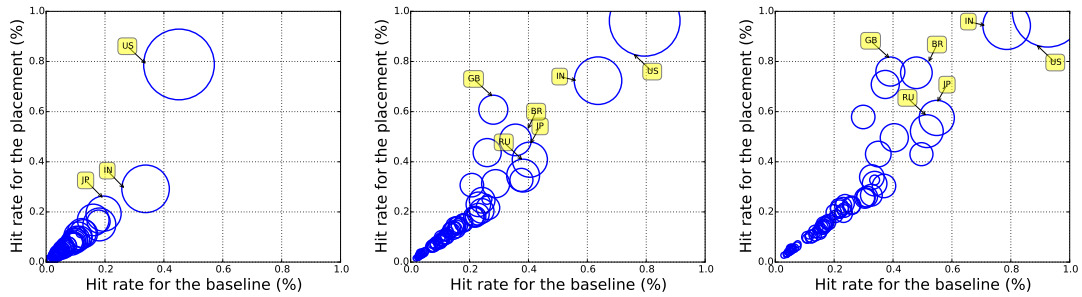


Figure 20: Each circle represents the hit rate of a country, placed in x according to its score with a random placement (baseline) and in y with a tag-based placement. The areas of the circles are proportional to the number of views received by each country. The three plots corresponds to $R = 1, 3$ and 5.

18% through the primary storage and as much as 27% through the cache.

Increasing the number of replicas, R, yields an improvement for both approaches in every country. However, our tag-based placement ends up providing better results in all countries except Russia. This likely results from the fact that we had to ignore a large number of Cyrillic tags from our dataset.

Figure 20 provides a different perspective on the results over the entire set of countries. Each circle in the figure represents a country and its surface is proportional to the number of requests from that country. The x axis represents the accuracy of the baseline approach and the y axis represents that of the tag-based approach. The figure shows that the countries with the biggest share of views benefit the most from our solution. With $R = 1$ the improvement remains limited to the US, but as R increases more and more countries see an improvement in their results. For $R = 5$, we also observe that the US reaches almost 100% with both approaches. This clearly limits the improvement that can be provided by any protocol.

5 Discussion and perspectives

Although our results are very encouraging, the preliminary techniques we have used for view prediction and video placement exhibit clear limitations. For example, they do not distinguish between "good" or "bad" tags in terms of predictive power, and do not provide any measure of prediction confidence.

Measuring the predictive power of tags could make computation faster (by excluding tags with no or low prediction power), and predictions more precise. Having a measure of prediction confidence could help refine an optimal number of copies required for a video in our placement strategy. To address these needs, we could incorporate several machine-learning techniques: these range from simple linear regressions and Bayesian inference techniques, to principal component analysis and random forest approaches.

Other limitations come from our assumptions. For example, in the case of placement, we allocated to every country a storage capacity proportional the country's global share of views, and assume all videos have the same size. However, it would be interesting to take into account the size of videos, and investigate storage allocation strategies that are not directly proportional, as our preliminary results show a tag-based placement seem to benefit countries with a big enough share of total views, but not the others.

The nature of our dataset, a punctual snapshot of some of Youtube's videos, also has important implications. First, our analysis considered all videos, independently of their age. However, some of the videos of our dataset were very recent at the time the dataset was crawled. This means that their number of views and their geographic distribution might not be representative of their future evolution. Taking these aspects into account could improve the predictive power of tags.

A running UGC systems would also possess information such as the time-stamps of video views, which would provide real-time information on the dynamics of view consumption [5, 20]. Considering these dynamics, as proposed for instance in [20], would make it possible to predict where a video's consumption might move, based on its consumption so far, and on the past behavior of similar videos as captured by their tags. For example, Brodersen *et al.* have observed in [5] that after a video has peaked, its views tend to reflush towards its region of production. Since geographical expansion is a temporary phenomenon, this knowledge could make it possible to vary the number of copies of a video dynamically depending on its "view trajectory".

Time-aware prediction engines would also need to address the problem of the size of the metadata associated with videos. The dataset we have used encompasses more than 173 billion views, and yet it only represents a small fraction of Youtube's overall traffic. Applying our approach to the size of modern UGC services would require techniques to speed up the extraction, and processing of all this information. A possibility could consist in sampling some of the data as in some epidemic distributed protocols [22], or as proposed in recent research on NoSQL databases [4]. Another, could rely on the results of Brodersen *et al.* [5], who showed that 50% of the videos have up to 70% of their views coming from the same geographic region. These and similar techniques could make it possible to fit batch computations within a one-day cycle of 24 hours. However, time-aware prediction systems should ideally follow an incremental rather than batch-based model. For example, they could rely on streaming-event platforms such as Storm [1] or S4 [18].

6 Related work

We are not aware of any other study on the link between the geographic distribution of tags and views in a UGC video service. In the following, we review some related works on the tagging practices of UGC users [12, 9, 10]; on the use of geographic information in UGC and VoD systems [20, 19]; and we finally discuss implications for actual deployments [6, 14, 24, 15].

6.1 Tags & folksonomies in UGC systems

In [9], Geisler and Bruns report that they found 517,008 unique tags in 898,282 Youtube videos collected in 2007. Although the orders of magnitude of their findings are in line with ours, the average number of tags per video they report (7.86) diverges from our measurement (11.56). This might be explained by the distance in time between the two data-sets (2007 and 2011), during which Youtube's GUI and user practices have evolved. This might also be due to our different methods of sampling: a snow-ball approach from most popular videos in our case, vs. a search on random words, followed by tag-based sampling in [9].

In [12], Heckner, Neubauer and Wolff, compare how tagging is used across different online media (Connotea, Del.ico.us, Flickr, and Youtube). They highlight interesting features of Youtube tags: Youtube users tend to use the tag field as a general free text description of a video's content, rather than as a organization mechanism. They also note that some users simply repeat a video's title, while others "overtag" their videos in an attempt to attract more views. These characteristics are particularly interesting in the context of our work, and point at refinements

we could take to further improve predictions, such as including title words, or detecting and dampening the effect of overtagging.

6.2 Tags & geolocation in UGC services

Quite a few works have been seeking to exploit the geolocation information embedded in Flickr pictures. Some have investigated the relation between tags associated with pictures and the position where the picture was taken. For instance, Hollenstein and Purves have used Flickr to investigate the meaning of specific geographic terms (such as "downtown" in US cities) [13]. In a related area, some researchers have used regression and optimization techniques (linked to machine learning) to discriminate tags capturing geographic positions in Flickr from other tags [7]. In a similar vein, the correlation between tags and location can be exploited to predict where a video was taken [23, 21].

Some researchers have sought to exploit *social cascades* (the viral process by which users point each other to on-line content) to predict where UGC videos would be consumed [20, 19]. Social cascades tend to show a strong geographic component (with users preferably forwarding resources to geographically close friends), and can be exploited to improve CDN cache policies [20].

These works, and the predictions they allow are orthogonal to the use of tags advocated in this paper. It would therefore be quite interesting to explore how they could be combined with our work.

6.3 Implications for delivery platforms

The prediction of the geographic distributions of UGC video views has obvious applications to CDNs and georeplicated storage systems, but also to peer-to-peer (P2P) implementations, and in particular peer-assisted streaming platforms, as already stressed for instance in [6]. Some providers (such as AT&T [14] or ChinaCache [24]—one of China's biggest CDN) have already experimented with peer-assisted approaches [15], in which a traditional data-center solution is extended with a P2P support. One key difficulty is however the need to appropriately place content to best exploit the limited outbound capacity of home networks, a task to which the analysis we have presented in this paper could contribute.

7 Conclusion

In this paper we have proposed an analysis of the geographic distribution of tags in Youtube, using of an original dataset of 691,349 videos, 7,717,815 tag occurrences, 705,415 unique tags, and 173,288,616,473 views. Our analysis shows that, as for videos, tags show a wide spectrum of distributions, with some tags concentrated in a few countries (low entropy) and others spread all other world (high entropy).

Comparing videos to the tags they contain in term of entropy, our analysis has highlighted that the geographic distribution of a video is strongly linked to that of its tags, with videos concentrated in a few countries (low entropy) typically linked to tags with the same behavior. We have shown, using a simple prediction technique, that this link could be exploited to predict a video's geographic distribution of views, a particularly interesting insight to improve current UGC video services. Our results (a minimum of 65.9% of views accurately predicted for a majority of videos) demonstrate the strong potential of tags to inform placement and caching policies, in particular when coupled with more advanced machine learning and regression techniques.

We think this work opens exciting perspectives to exploit tags and generally content-related data to improve the implementation of large-scale geo-replicated storage and delivery systems, an avenue which we plan to pursue in the future.

References

[1] Apache storm, distributed and fault-tolerant realtime computation. `http://storm.incubator.apache.org/`. accessed 7 May 2014.

[2] International telecommunication union. `http://www.itu.int`.

[3] Global internet phenomena report: 2h 2013. Technical report, Sandvine Incorporated, 2013.

[4] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. Blinkdb: queries with bounded errors and bounded response times on very large data. In *EuroSys*, 2013.

[5] A. Brodersen, S. Scellato, and M. Wattenhofer. YouTube around the world: Geographic popularity of videos. In *WWW*, 2012.

[6] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *IMC*, 2007.

[7] O. Chaudhry and W. Mackaness. Automated extraction and geographical structuring of flickr tags. In *GEOProcessing 2012*.

[8] X. Cheng and J. Liu. NetTube: Exploring social networks for peer-to-peer short video sharing. In *INFOCOM*, 2009.

[9] G. Geisler and S. Burns. Tagging video: conventions and strategies of the youtube community. In *ACM/IEEE-CS Joint Conf. on Digital Libraries*, 2007.

[10] S. Greenaway, M. Thelwall, and Y. Ding. Tagging youtube - a classification of tagging practice on youtube. In *Int. Conf. on Scientometrics and Informetrics*, 2009.

[11] F. Guillemin, B. Kauffmann, S. Moteau, and A. Simonian. Experimental analysis of caching efficiency for youtube traffic in an isp network. In *Int. Teletraffic Congress*, 2013.

[12] M. Heckner, T. Neubauer, and C. Wolff. Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types. In *2008 Workshop on Search in Social Media*. ACM.

[13] L. Hollenstein and R. Purves. Exploring place through user-generated content: Using flickr tags to describe city cores. *J. of Spatial Inf. Sc.*, (1), 2015.

[14] Y. Huang, Y.-F. Chen, R. Jana, H. Jiang, M. Rabinovich, A. Reibman, B. Wei, and Z. Xiao. Capacity analysis of mediagrid: a p2p iptv platform for fiber to the node (fttn) networks. *IEEE J. on Sel. Areas in Comm/*, 25(1).

[15] Y. Huang, T. Z. J. Fut, D.-M. Chiu, J. C. S. Lui, and C. Huang. Challenges, design and analysis of a large-scale P2P-VoD system. In *SIGCOMM*, 2008.

[16] K. Huguenin, A.-M. Kermarrec, K. Kloudas, and F. Taïani. Content and geographical locality in user-generated content sharing systems. In *Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video*, NOSSDAV '12.

[17] S. Khemmarat, R. Zhou, L. Gao, and M. Zink. Watching user generated videos with prefetching. In *MMSys*, 2011.

[18] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 2010.

[19] N. Sastry, E. Yoneki, and J. Crowcroft. Buzztraq: Predicting geographical access patterns of social cascades using social networks. In *SNSEuroSys Workshop*. ACM, 2009.

[20] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades. In *WWW*, 2011.

[21] Y. Song, Y. Zhang, J. Cao, J. Tang, X. Gao, and J. Li. A unified geolocation framework for web videos. *ACM TIST*, 5(3), 2014.

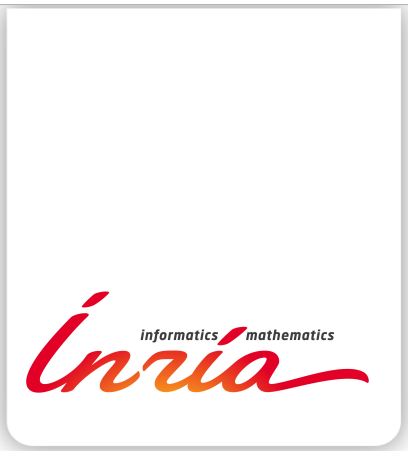
[22] F. Taiani, S. Lin, and G. Blair. Gossipkit: A unified componentframework for gossip. *Software Engineering, IEEE Transactions on*, 40(2), 2014.

[23] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In *ACM Conference on Multimedia Retrieval*, 2013.

[24] H. Yin, X. Liu, T. Zhan, V. Sekar, F. Qiu, C. Lin, H. Zhang, and B. Li. LiveSky: Enhancing CDN with P2P. *ACM TOMCCAP*, 6:16:1–16:19, 2010.

[25] L. Youtube. Statistics, viewership . http://www.youtube.com/yt/press/statistics.html. (accessed 2/5/2014).

[26] R. Zhou, S. Khemmarat, and L. Gao. The impact of youtube recommendation system on video views. In *IMC*, 2010.



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803