Exploring The Use of Tags for Georeplicated Content Placement

Stéphane Delbruel, Davide Frey, François Taïani

Exploring The Use of Tags for Georeplicated Content Placement

Stéphane Delbruel
University of Rennes 1
IRISA - Inria, Rennes, France
Email: stephane.delbruel@irisa.fr

Davide Frey
Inria
Rennes, France
Email: davide.frey@inria.fr

François Taïani
University of Rennes 1 - ESIR
IRISA - Inria, Rennes, France
Email: francois.taiani@irisa.fr

***Abstract*—A large portion of today's Internet traffic originates from streaming and video services. Such services rely on a combination of distributed datacenters, powerful content delivery networks (CDN), and multi-level caching . In spite of this infrastructure, storing, indexing, and serving these videos remains a daily engineering challenge that requires increasing efforts on the part of providers and ISPs. In this paper, we explore how the tags attached to videos by users could help improve this infrastructure, and lead to better performance on a global scale. Our analysis shows 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 demonstrate the potential of tags to help predict distribution of a video's views, and present results suggesting that tags can help place videos in globally distributed datacenters. We show in particular that even a simplistic approach based on tags can help predict a minimum of 65.9% of a video's views for a majority of videos, and that a simple tag-based placement strategy is able to improve the hit rate of a distributed on-line video service by up to 6.8% globally over a naive random allocation.**

***Index Terms*—User-generated content, YouTube, tag, prediction**

I. Introduction

Videos streaming 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 peak load [1]. 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 [2]. 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 [3], most of which need to be served to niche audiences, in limited geographic areas [4], [5], [6].

Better understanding what these niche audiences and geographic areas are is a 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 [7], the navigation behavior of users [8], [9], or the geographic diffusion of views triggered by social media [10]. Other studies have highlighted the potential of peer-assisted VoD systems [11], [12] to support the long tail of video popularity typically observed in UGC video services, or P2P architectures [13], [5] that exploit the relationship between viewing behavior and the graph of related videos [9].

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 [14], very few works have explored how tags relate to the viewing patterns of the videos they describe [15], [16], and to the best of our knowledge, none have considered how tags could help design better UGC video delivery systems. 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 [4], [6].

In this paper, we investigate how tags could help improve the design of UGC platforms by providing insights on the geographic distribution of video views. We first analyze an extensive Youtube dataset of 590,897 videos to substantiate our claim that tags can be used as markers of a video's geographic diffusion. Based on our findings, we then show that the geographic distribution of videos' views can be predicted from their tags. Our results demonstrate that a tag-based linear interpolation can predict more than 65.9% of a video's views for a majority of videos. Finally, we propose a novel tag-based placement strategy for a global video storage platform. In our evaluation, our approach improves the system's hit rate by 6.8% compared to a random placement strategy in the presence of an LRU cache.

In the following, we first present our analysis of tag and view distribution in Youtube (Sec. II), and then present how the geographic distribution of a video's views can be predicted from its tags (Sec. III). In Section IV we propose how new videos could be placed based on their tags, and present an experimental evaluation of our approach. Section V presents related approaches, and Section VI concludes.

II. Tags, Views, and Geodistribution in Youtube

Our study uses a Youtube data set collected by our research group in March 2011 [5]. The seeds of the data set are the 10 most popular videos in 25 different countries, obtained through Youtube's public API. The data set was then completed using a breadth-first snowball sampling of the graph of related videos,

TABLE I: 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	...

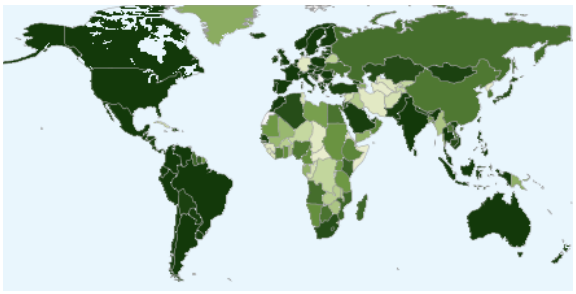


Fig. 1: Popularity map of the most viewed video of our data set *Justin Bieber - Baby ft. Ludacris*, as provided by Youtube.

as reported by Youtube. For each crawled video, the data set 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 [16], [15].

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 data set (*Justin Bieber - Baby ft. Ludacris*). Such maps were provided using Google's Map Chart service [17] 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 I).

The original data set 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 data set with 590,897 videos, associated with 705,415 unique tags, totaling 173,288,616,473 views.

In the following, we first present a number of notations and concepts we will use in the remainder of the paper (Sec. II-A), explain how we extracted views from popularity vectors (Sec. II-B), and discuss the metrics we are interested in (Sec. II-C). We then turn to our description and analysis of the data set in terms of views, tags, and geographic distribution (Sec. II-D and following).

A. Notation

$\mathcal{V}$ is the set of videos in our data set. 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.

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

B. 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 I, 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* [18], 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. [19] on July 2014, an authoritative source of 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 data set 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 data set:

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

[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.

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

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

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.

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

C. 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:

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

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 data set, *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.

D. Tag and view distributions

Our data set contains 7,717,815 tag occurrences, for an average of 11.18 tags per video, and a total of 705,415 unique tags. This large number of tags, in line with earlier findings [15], 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. 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.

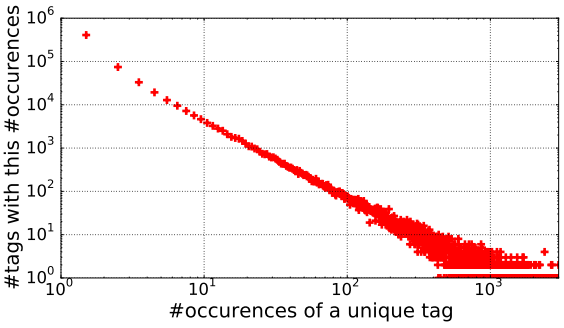


Fig. 2: The frequency distribution of tags follow a power law of the shape $y = K \times x^{-\alpha}$, as often observed in folksonomies and natural languages.

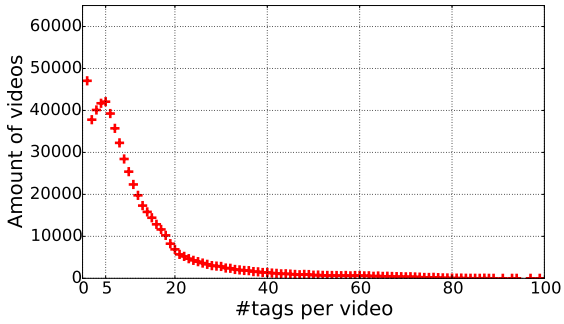


Fig. 3: Tags are widely used to describe videos, with 50% of videos showing a least 11 tags.

Tables II and III show, respectively, the 10 most frequent tags and the top 10 tags with the most views. Most tags describe content (*video*, *funny*) but some consist of grammatical words (*the*, *of*). The latter probably result from the former usage of spaces to separate tags (Youtube now uses commas), which caused compound terms such as *the␣rock* to be parsed into two tags (*the* and *rock*). The tables also show that the most viewed tags are not necessarily the most frequent. For example, *pop*, the second most viewed tag (Table II), only occurs 7877 times. The corresponding videos predominantly belong to the "Music" category, with a high average number of views per individual video (1,690,809 views, 2.7 times more than those of videos containing the tag *funny*).

As mentioned, videos have relatively rich tag descriptions (Figure 3) with 11.18 tags on average. One reason may be that users have an incentive to tag their videos to attract more views. 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). The median number of views of a video increases with up to 18 tags. But this relationship collapses beyond this value. For instance, the most tagged video in our data set possesses 102 tags, but only 1,220,496 views, which pales in comparison to the most viewed video—471,208,788 views for only 6 tags.

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

TABLE II: 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 III: 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

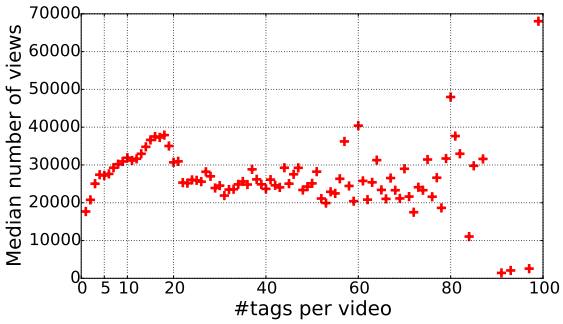


Fig. 4: Median number of views for the videos embedding a given number of tags. Views and size of the tag set seem only weakly correlated, with a clear growing trend limited to videos with less than 18 tags.

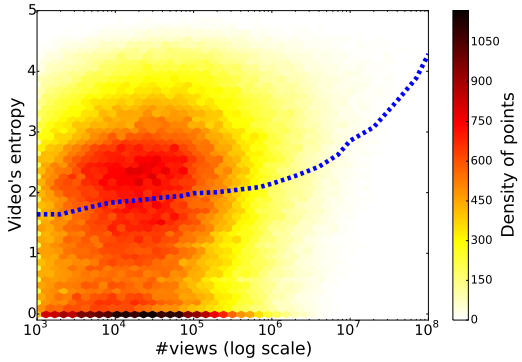


Fig. 5: Heatmap of each video's entropy vs. its number of views. Mean shown as a dashed line.

TABLE IV: 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%

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

E. From Videos to Tags and Back

To understand how tags can provide information to drive the storage of videos, we now analyze the geographic distributions of videos and tags in our data set. We start by considering videos, and analyzing the relationship between their popularity and their geographic distribution. Figure 5 depicts this relationship in the form of a heat map. The x axis represents the popularity of videos in terms of their number of views, the y axis measures the geographical distribution in terms of entropy, while colors indicate the density of videos with the corresponding entropy-popularity values.

As pointed out in earlier work [5], the views of popular videos, in particular those with more than 10^7 views, tend to be widely distributed, with average entropy values between 3 and 4. These high entropy values mean that these videos need to be easily accessible from all over the world, which reduces the interest in predicting their geographical distribution [7]. However, the plot also shows that these popular videos constitute a minority. Most of the data points in the heat map represent videos with less than 10^6 views. For these videos, the average entropy remains around the value of 2, with a few high density points around entropy values of 2.5, 1.5 and 0.

These numbers show that a significant fraction of videos are geographically concentrated, For example, videos with an entropy below 1.5 constitute 40% of the data set with a mean number of views of 155,520, a mean number of tags of 9 (vs. 11.18 for the whole data set), and a mean entropy of 0.707. To get a feel of the meaning of these numbers, we observe that an entropy of 1.5 could, for example, correspond to a video that is present and uniformly distributed in only 4 countries. In general, such a low value corresponds to videos that are geographically concentrated and thus that constitute perfect candidates for proactive placement strategies.

We argue that tags can contribute to place these geographically concentrated videos close to their viewers. To verify this hypothesis, we start by analyzing the most popular tags. Table VI shows that the most viewed tag in in each of five western countries (France, Germany, Canada, Australia, and USA) is *music* (entropy of 3.80), *pop* (entropy of 4.27) or *funny* (entropy of 3.03). These tags also appear in Table III, which instead shows most viewed tags on global scale.

Based on this example, one might wonder if the popularity of tags correlates with that of the corresponding videos. But

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

tag	H(t)	#occurs	#views	average views
recovery	4.90	230	557,870,332	2,425,523
dominic	4.87	103	338,555,233	3,286,944
fifa	4.83	2722	690,092,931	253,524
passat	4.79	142	41,809,394	294,432
afraid	4.78	131	244,659,961	1,867,633

tag	H(t)	#occurs	#views	average views
piologo	0.04	101	3,985,341	39,458
mundo canibal	0.06	134	4,147,866	30,954
kvarteret	0.10	102	7,313,481	71,700
skatan	0.11	106	7,741,235	73,030
partoba	0.18	272	7,183,083	26,408

TABLE VI: 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

TABLE VII: 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%*

this is not necessarily the case. For example, the top three videos with the tag *pop* (Table VII) also turn out to be the most viewed in the entire data set. However, other tags, like *funny*, appear in a large number of possibly much less popular videos. To assess the potential of tags for predicting the consumption of videos we therefore seek for a correlation between their entropy values.

Figure 6 compares the cumulative distribution function (CDF) of the entropy of videos (solid line) with that of tags (dashes) in our data set. The two curves exhibit very similar trends: entropy values tend to be evenly spread for values below 3, which correspond to roughly 80% of all videos and tags. Table V complements this information by showing the tags with the highest and those with the lowest entropy.

Figure 7 depicts the relationship between the entropy and the popularity of tags in the form of heat map. Like for videos, popular tags constitute a minority: most tags have

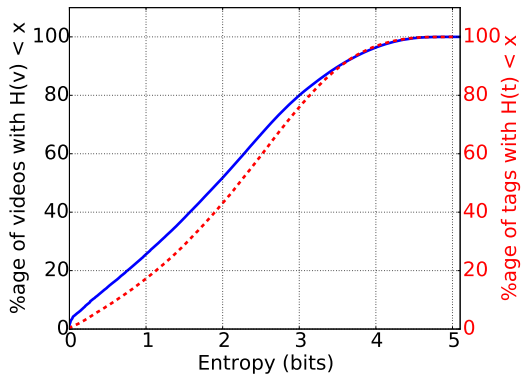


Fig. 6: CDF of videos (solid line) and tags (dashes) versus entropy

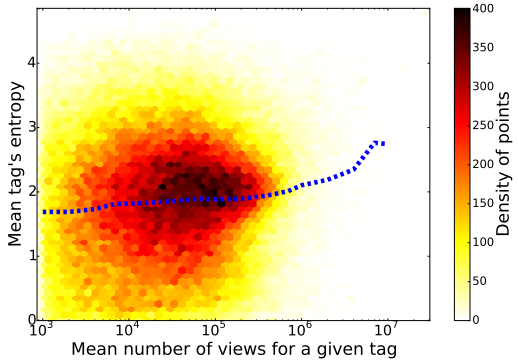


Fig. 7: Heatmap of the mean views for every occurrences of a given tag, versus the mean entropy of every occurrences of that tag. Mean showed as a dashed line.

entropy values around 2, and an average of $100,000$ views. We provide two examples of such specific tags in TableVIII and Figure 8, and in Table IX and Figure 9. The two tables and figures show the top-5 viewing countries and the viewership distribution for tags *bollywood* (entropy of 3.24) and *favela* (entropy of 2.22). In the figures, a higher color saturation indicates a higher proportion of views for the corresponding country. The views of *bollywood* mostly occur in India and the 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%. The views of *favela* are even more concentrated with Brazil responsible for almost 48% of all views, followed by the United-States with 34.9%, These figures suggest that caching or storing copies of videos containing these tag in the respective top countries would significantly benefit UGC video systems.

Figure 10 highlights the potential of tags in doing so: the figure plots the mean entropy of 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 reinforces our conjecture that tags can predict the geographic distribution of the associated videos. In the following sections, we first show that this is possible, and then apply this finding in a proactive video-placement strategy.

III. Predicting Views from Tags

We investigate in this section whether a video's views can be inferred from its tags using a simple interpolation approach. A positive answer would indicate that tags can indeed be used to

TABLE VIII: 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%

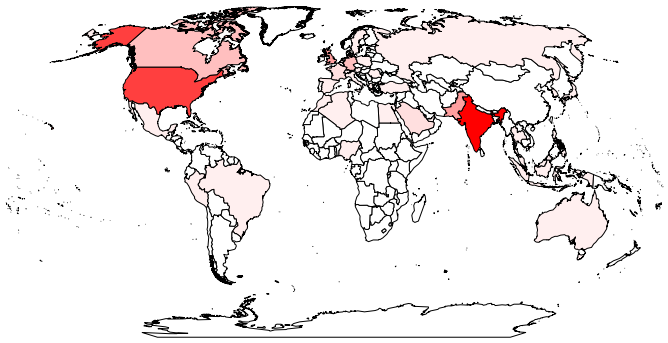


Fig. 8: Videos associated with the tag 'bollywood' tend to be viewed mainly in India, USA and UK.

TABLE IX: 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%

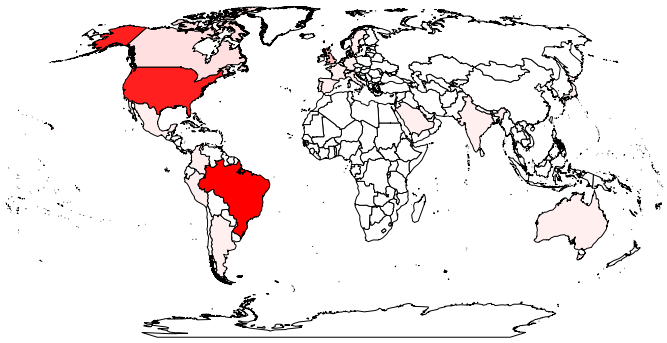


Fig. 9: Videos associated with the tag 'favela' are mostly viewed in Brazil

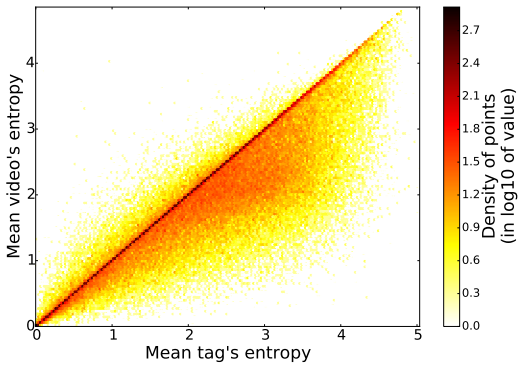


Fig. 10: Tag entropy versus video entropy

help place videos in a georeplicated storage system, a question we turn to in Section IV.

A. General approach

When a new video v is uploaded, 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 set of videos already known to the system $\mathcal{V}$:

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

where $\mathbf{p}^{\mathcal{V}}_{\text{geo}}(t)$ is the geographic distribution vector of tag t in the dataset. $\mathcal{V}$.

B. Baseline

As a baseline prediction, we use the average distribution of global Youtube views, estimated from the YouTube network traffic reported by Alexa Internet Inc. [3] (Table X). Alexa only covers the top 40 countries producing the most Youtube traffic,

TABLE X: Youtube traffic share according to Alexa

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

totaling 85.2% of global Youtube network usage. We apportion the remaining 14.8% to the 217 countries not covered by Alexa proportionally to their share of internet users, as reported by the International Telecommunication Union [20]. This process yields the same baseline view prediction for all videos, that we compare against the results returned by (7).

C. Evaluation and metric

To compare the two approaches, we divide our dataset into two equal parts: a training set $\mathcal{V}_{\text{train}}$ and a testing set $\mathcal{V}_{\text{test}}$, containing 295449 ($\pm 1$) videos each. We then use (7) to predict the view distribution of each video v in $\mathcal{V}_{\text{test}}$ from the tag distribution extracted from $\mathcal{V}_{\text{train}}$ (which plays the role of known videos in the formula).

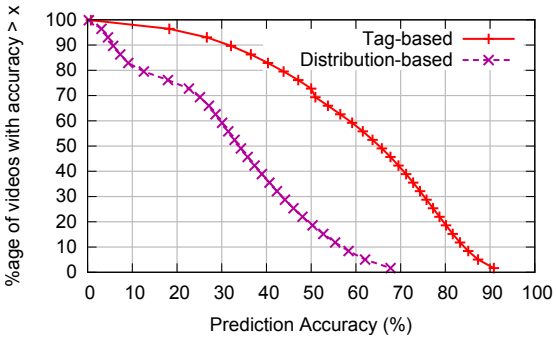
To evaluate the quality of a prediction $\widehat{\mathbf{p}_{\text{geo}}}(v)$, we compute the proportion of views correctly placed by the prediction (what we term the prediction's *accuracy*):

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

where $\mathbf{p}_{\text{geo}}(v)$ is the actual geographic distribution of video v, and the division by 2 normalizes the result between 0 (all views were misplaced) and 1 (no misplaced views).

D. Results and discussion

Figure 11 plots the cumulative distribution of prediction accuracy obtained by our approach (*Tag-based prediction*)



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

Fig. 11: 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.

and by the baseline (*Distribution-based prediction*) with the corresponding mean and median values indicated below. Our approach clearly outperforms the baseline, yielding a median accuracy (65.9%) that is almost twice that of its competitor (33.9%). This confirms that tags hold the promise of predicting the geographic distribution of UGC videos.

Figures 12 and 13 delve deeper into the results and show the effect of the number of views and of the entropy of a video, respectively, on the accuracy of prediction for both approaches. The heat maps show the distribution of individual videos, while the dashed lines indicate the average accuracy obtained for a given number of views, resp. entropy.

Figure 12 indicates that tag-based prediction significantly outperforms the baseline regardless of a video's popularity. Both plots further show a weak positive correlation between the number of views of a video and its accuracy. This correlation probably might stem from the link between popularity and entropy. Highly popular videos tend to be scattered all over the world (high entropy), and are therefore easier to predict.

By contrast, Figure 13 shows that tag-based prediction works best for video with an average entropy (between 2 and 3, accuracy above 70%), with lower results both for both highly concentrated and widely distributed videos (corresponding to low resp. high entropy values). This behavior is in stark contrast to that of the baseline, whose performance is directly linked to that of entropy, indicating that the predicting value of tags is particularly interesting for videos with low to medium entropy, which tend to diverge from the average behavior.

IV. Using tags for proactive video placement

Building on the previous results, we now explore whether tags can help design better UGC systems by determining where to place new videos. This ability will become increasingly important as more and more applications manage short-lived content.

A. System model

Our scenario considers a company that must decide where to store the primary copies of a set of new videos $\mathcal{V}_{\text{new}}$ on its global storage infrastructure using tag information extracted from videos already served by the service $\mathcal{V}_{\text{known}}$.

In terms of infrastructure, we consider an extreme case, in which each country has some storage capacity available for new videos (a datacenter, or share of datacenter for small countries). We assume the system's overall available capacity ($\mathcal{S}_{world}$) is able to store R copies of each new video. $R = 3$ for instance is a typical value for R used in cloud storage systems (e.g. GFS, HFS). For simplicity's sake, we also consider that all videos have the same size.

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

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

where $\mathbf{p}_{global}[c]$ is the proportion of views in country c. UGC providers typically rely on multiple layers of caches (within browsers, at Internet Points-of-Presence, within datacenters), in addition to their primary storage system [21]. In our model, we aggregate all these caches in one single layer located in each country, set to an LRU eviction policy. We set the capacity of this caching layer to 10% of each country's primary storage. This value is relatively low on purpose in order to better analyze the effect of tags on the system.

B. Placement mechanism

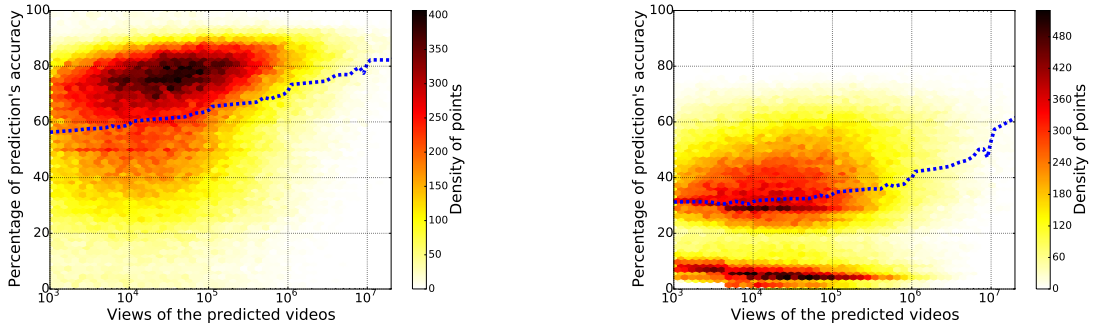
We place each new video v according to an estimation of its per-country viewing vector $(\widehat{\mathbf{views}}(v)[c])_{c \in World}$. This estimation uses the geographic distribution of tags observed in the videos already served by the service $\mathcal{V}_{\text{known}}$. More precisely, we compute for each tag t an average per-video and per-country "contribution" of this tag to the views of the known videos in which t appears:

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

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

$$\widehat{\mathbf{views}}(v)[c] = \mathop{\mathbb{E}}_{t \in tags(v)} \left(\mathbf{views\_p\_vid}^{\mathcal{V}_{\text{known}}}(t)[c]\right) \quad (9)$$

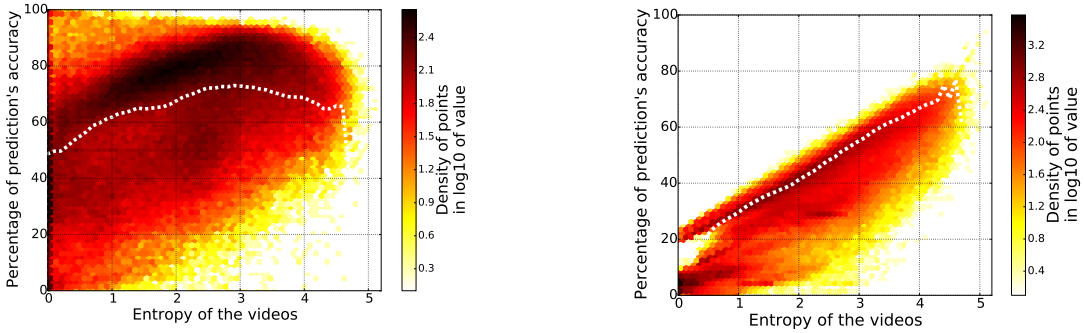
The placement works then as follows: we iterate over the videos of $\mathcal{V}_{\text{new}}$, and place R copies of each video v in the first R countries in which v is predicted to get most of its views, among the countries with some remaining storage.



(a) Tag-based prediction

(b) Distribution-based prediction

Fig. 12: Prediction accuracy vs video views. The dashed lines show the average accuracy. The tag-based approach outperforms the baseline across the range of video views.



(a) Tag-based prediction

(b) Distribution-based prediction

Fig. 13: Prediction accuracy vs video entropy. The dashed lines show the average accuracy. The benefit of tags is particularly strong for low entropy values.

C. Experiment, metrics, baseline

As in Section III, we split our dataset in two, using the same reference ($\mathcal{V}_{\text{train}}$), and testing sets ($\mathcal{V}_{\text{test}}$). $\mathcal{V}_{\text{train}}$ plays the role of known videos $\mathcal{V}_{\text{known}}$. Because $\mathcal{V}_{\text{test}}$ remains particularly large (295448 videos, and 86,624,310,171 views), we sample it down: We first generate a trace $\mathcal{T}$ of 10 millions requests for the videos of $\mathcal{V}_{\text{test}}$ that respects the distribution of views between videos and countries. We then choose $\mathcal{V}_{\text{new}}$ as the set of unique videos present in the trace $\mathcal{T}$.

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

We evaluate the quality of a placement by replaying the trace $\mathcal{T}$, and counting how often a request can be served from the country it originates from (a *hit*). In case of a miss, we store the video in the local country cache for future use.

D. Results

Results are shown in Figures 14–16. Figure 14 plots the average hit ratio obtained by each approach for different replication factors ($R \in [1, 5]$). It shows that a tag-based placement clearly outperforms the baseline with an improvement that oscillates between 5.6% ($R = 1$) and 6.8% ($R = 5$).

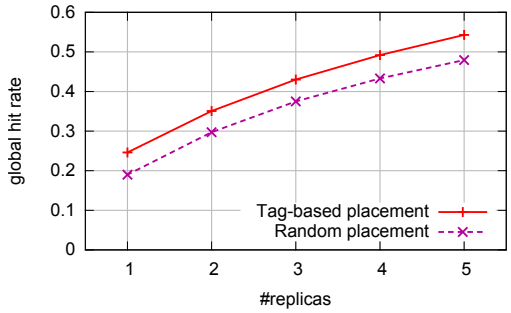
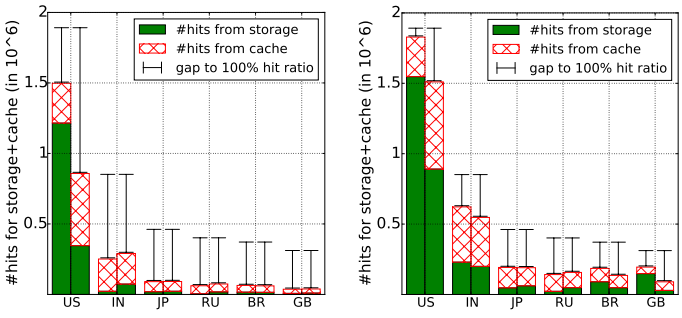


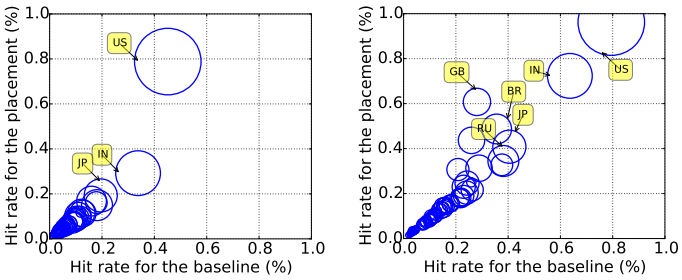
Fig. 14: A tag-based placement strategy consistently improves the system's global hit rate by about 6%, independently of the number of copies per video.

Figure 15 charts the cache performance of the 6 countries receiving the most views for two values of R (1 and 3). The left bar above each country corresponds to the performance of the tag-based placement, and the right bar to that of the random placement. Each bar shows the absolute number of misses (top black line), of hits served by the LRU cache (middle red hatched section), and of hits served by the primary storage (bottom green solid section).



(a) $R = 1$ (b) $R = 3$

Fig. 15: Hits and misses for the top 6 countries for $R \in \{1, 3\}$, for the tag-based (left bars) and random placement (right bars).



(a) $R = 1$ (b) $R = 3$

Fig. 16: Hit ratios obtained through random placement (x-axis) vs. tag-based placement (y-axis) for all country (individual bubbles) for $R \in \{1, 3\}$. The bubbles' area shows the number of views of individual countries.

Figure 16 shows the same data as a function of the number of views. Each bubble corresponds to a country: its area is proportional to the country's number of views, its x-coordinate represents the hit ratio of random placement, while its y-coordinate represents that of tag-based placement. Countries above the diagonal thus obtain an improvement in hit ratio.

Both Figures 15 and 16 show that tag-based placement works best for countries with many views, and that the number of benefiting countries increases with R (from one—the USA—with $R = 1$, to more than ten with $R = 3$). This phenomenon directly results from our greedy placement algorithm: countries with many views are predicted more often as a top country and thus attract more "good" primary copies. With a single copy per video ($R = 1$), runner-up countries (such as India, or Japan) are thus prevented from storing videos that would be good matches for their viewership, because these have been preferentially attracted to the US. When R increases this phenomenon moves down the list of countries. The overall effect remains an average increase in hit ratio (Figure 14).

E. Discussion

The above results show that tags can help predict the distribution of future video views, and that this predictive power can be exploited to design better UGC services. These results are all the more encouraging considering they are based on simple linear interpolations. More advanced techniques from the area of machine learning are likely to deliver higher payoffs, for instance by distinguishing between tags with no or low prediction power and those with a high potential. They could thus limit the cost of computing predictions (by pruning inefficient tags), and produce confidence intervals which could guide placement or caching strategies. The spectrum of potential techniques to explore is large, ranging from linear regressions and Bayesian inference techniques, to principal component analysis and random forests.

These techniques would also be able to exploit additional information, such as the time-stamps of video views, and real-time information on the dynamics of view consumption [4], [6]. This temporal information would make it possible to predict the viewing "trajectory" of a video, and help predict for instance viewing cascades driven by social media [6].

In terms of external validity, i.e. the extent to which our experiments may translate to deployed systems, our model comprises a number of simplifications that are likely to warrant further work. For example, we assumed that all videos had the same size. Taking into account video sizes would impact the behavior of the cache and storage layers, but also provide an additional property from which to predict a video's views: long videos might for instance present different geographic distributions than short ones with the same set of tags. We also took an extreme and hypothetical view of a primary storage layer present in all countries: In practice, large UGC providers only manage about a dozen datacenters (14 in 2015 for Google for instance [22]), mostly concentrated in the USA and Europe (11 out of 14 in the case of Google). Our model is therefore much more likely to make sense for a federation of globally distributed small providers, rather than for a major player with an already well established infrastructure. Finally, the ability to place videos proactively before their first views should prove particularly useful for applications that host short-lived user-generated content.

V. Related work

We are not aware of other works exploiting the geographic distribution of tags in a UGC video service. In the following, we review some related works on the use of tags in on-line services [14], [15], [16]; on the use of geographic information in UGC and VoD systems [6], [10]; and we finally discuss implications for actual deployments [7], [11], [12], [23].

A. Tags & folksonomies in UGC systems

Geisler and Bruns report an average number of 7.86 tags per video in a Youtube dataset of a size similar to ours collected in 2007 [15]. Although from a similar order of magnitude, this number diverges from our measurement (11.56). This might be explained by the distance in time between the two datasets (2007 and 2011). This might also be due to our different methods of sampling (a snow-ball approach in our case, vs. a search on random words in [15]).

Heckner, Neubauer and Wolff have compared how tagging is used across different on-line media [14]. They highlight interesting features of Youtube tags: Youtube users tend to use the tag field as a general free text description of a video, rather than as a organizational means. Some users simply repeat a video's title, while others "overtag" their videos in an attempt to attract more views. These characteristics point at refinements we could take to further improve predictions, such as including title words, or detecting overtagging.

B. 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 [24], [25], [26]. Others have used regression techniques to discriminate tags capturing geographic positions in Flickr from other tags [27].

Some researchers have sought to exploit *social cascades* (the viral process by which users point each other to on-line content) to predict where videos would be consumed [6], [10]. Social cascades present a strong geographic component (users preferably forward resources to geographically close friends), and can be exploited to improve CDN cache policies [6].

These works are orthogonal to our approach, and combining their techniques with ours is likely to yield interesting results.

C. Implications for delivery platforms

The prediction of the geographic distributions of UGC video can help design better distribution platforms, not only with traditional, but also with peer-to-peer (P2P) and peer-assisted architectures [7], [11], [12], [23]. One key difficulty in such architectures consists in appropriately placing content to best exploit the limited outbound capacity of home networks, a task to which the analysis in this paper could contribute.

VI. Conclusion

In this paper we have proposed to use the tags attached to videos to improve the design of User Generated Content (UGC) video services. To inform our work, we have first presented an analysis of the geographic distribution of tags in Youtube, using of an original dataset of 590,897 videos. This analysis shows that the geographic distribution of tags is strongly correlated to that of the videos they are attached to, hinting at the potential of tags to design better UGC services.

We have confirmed this potential by demonstrating how the tags attached to a video could be exploited to predict a video's geographic distribution. More specifically, we have shown that we were able to predict a minimum of 65.9% of a video's views for a majority of videos using tags. Building on these results, we have proposed a novel tag-based placement strategy. Our evaluation shows our approach is able to improve the hit rate of a global video distribution infrastructure by 6.8% compared to a random placement strategy.

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.

Acknowledgments

This work was partially funded by the French ANR project SocioPlug (ANR-13-INFR-0003), by the DeSceNt project of the Labex CominLabs excellence laboratory (ANR- 10-LABX-07-01), and by the "Politique doctorale 2013" grant of the University of Rennes 1 "*Towards a Decentralized Embryomorphic Storage System*".

References

[1] 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.
[2] "Global internet phenomena report: 2H 2013," Sandvine Incorporated, Tech. Rep., 2013.
[3] http://www.youtube.com/yt/press/statistics.html, (accessed 2/5/2014).
[4] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: Geographic popularity of videos," in *WWW*, 2012.
[5] K. Huguenin, A.-M. Kermarrec, K. Kloudas, and F. Taïani, "Content and geographical locality in user-generated content sharing systems," in *NOSSDAV*, 2012.
[6] 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.
[7] 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.
[8] S. Khemmarat, R. Zhou, L. Gao, and M. Zink, "Watching user generated videos with prefetching," in *MMSys*, 2011.
[9] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *IMC*, 2010.
[10] N. Sastry, E. Yoneki, and J. Crowcroft, "Buzztraq: Predicting geographical access patterns of social cascades using social networks," in *SNS@EuroSys Workshop*. ACM, 2009.
[11] 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/*, vol. 25, no. 1.
[12] H. Yin, X. Liu, T. Zhan, V. Sekar, F. Qiu, C. Lin, H. Zhang, and B. Li, "LiveSky: Enhancing CDN with P2P," *ACM TOMCCAP*, vol. 6, pp. 16:1–16:19, 2010.
[13] X. Cheng and J. Liu, "NetTube: Exploring social networks for peer-to-peer short video sharing," in *INFOCOM*, 2009.
[14] 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.
[15] G. Geisler and S. Burns, "Tagging video: conventions and strategies of the youtube community," in *ACM/IEEE-CS Joint Conf. on Digital Libraries*, 2007.
[16] S. Greenaway, M. Thelwall, and Y. Ding, "Tagging youtube - a classification of tagging practice on youtube," in *Int. Conf. on Scientometrics and Informetrics*, 2009.
[17] https://developers.google.com/chart/image/docs/gallery/map_charts.
[18] http://www.google.com/trends/.
[19] http://www.alexa.com/.
[20] "International telecommunication union," http://www.itu.int.
[21] Q. Huang, K. Birman, R. van Renesse, W. Lloyd, S. Kumar, and H. C. Li, "An analysis of facebook photo caching," in *SOSP*, 2013.
[22] "Google data centers, data center locations," https://www.google.com/about/datacenters/inside/locations/index.html, accessed Sep. 10 2015.
[23] Y. Huang, T. Z. J. Fu†, D.-M. Chiu, J. C. S. Lui, and C. Huang, "Challenges, design and analysis of a large-scale P2P-VoD system," in *SIGCOMM*, 2008.
[24] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier, "Retrieving geolocation of videos with a divide & conquer hierarchical multimodal approach," in *ACM Conference on Multimedia Retrieval*, 2013.
[25] Y. Song, Y. Zhang, J. Cao, J. Tang, X. Gao, and J. Li, "A unified geolocation framework for web videos," *ACM TIST*, vol. 5, no. 3, 2014.
[26] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using flickr tags to describe city cores," *J. of Spatial Inf. Sc.*, no. 1, 2015.
[27] O. Chaudhry and W. Mackaness, "Automated extraction and geographical structuring of flickr tags," in *GEOProcessing 2012*.