Content and Geographical Locality in User-Generated Content Sharing Systems

Kévin Huguenin
EPFL
Lausanne, Switzerland
kevin.huguenin@epfl.ch

Anne-Marie Kermarrec
INRIA
Rennes, France
anne-marie.kermarrec@inria.fr

Konstantinos Kloudas
INRIA
Rennes, France
konstantinos.kloudas@inria.fr

François Taïani
Lancaster University
Lancaster, UK
f.taiani@lancs.ac.uk

ABSTRACT

User Generated Content (UGC), such as YouTube videos, accounts for a substantial fraction of the Internet traffic. To optimize their performance, UGC services usually rely on both proactive and reactive approaches that exploit spatial and temporal locality in access patterns. Alternative types of locality are also relevant and hardly ever considered together. In this paper, we show on a large (more than 650,000 videos) YouTube dataset that *content locality* (induced by the related videos feature) and *geographic locality*, are in fact correlated. More specifically, we show how the geographic view distribution of a video can be inferred to a large extent from that of its related videos. We leverage these findings to propose a UGC storage system that *proactively* places videos close to the *expected* requests. Compared to a caching-based solution, our system decreases by 16% the number of requests served from a different country than that of the requesting user, and even in this case, the distance between the user and the server is 29% shorter on average.

Categories and Subject Descriptors

H.3.2 [**Information Systems**]: Information Storage and Retrieval—*Information Storage*

General Terms

Measurement, Algorithm, Design

Keywords

User-generated content, content distribution

NOSSDAV'12, June 7–8, 2012, Toronto, Ontario, Canada.

1. INTRODUCTION

Over the last few years, users have become the most prolific content generators on the Web, prompting the phenomenal success of user-generated content (UGC) sharing sites such as YouTube [3, 6]. This rapid growth combined with the never fulfilled user demand for better quality, makes serving UGC to an increasing number of users all around the world a daily engineering feat [13]. To this end, UGC sharing sites rely on Content Delivery Networks (CDNs) to place content close to consumers. To achieve this goal, CDNs employ both *proactive* approaches that rely on *a priori* knowledge (e.g., it is likely that a French-speaking video will be accessed mostly from French-speaking countries) and *reactive* ones where the history of a given video is analyzed to predict its future requests [8, 4, 17].

Interestingly, beyond traditional forms of locality that account for each content item independently, recent studies have shown that UGC viewing patterns are significantly influenced by the fact that content in these sites is no longer independent but is organized in a *content graph*. In YouTube, this *content graph* is embodied by the lists of "related videos" present on each video's page, and its influence on the viewing behavior of users has been clearly documented [10, 18].

In this paper we study the geographic viewing patterns of UGC and how they are affected by the content graph. Our analysis on a novel YouTube dataset shows that (i) related videos tend to have correlated geographic viewing patterns, with most of their views coming from the same countries, and (ii) popular videos tend to have their views more uniformly spread across more countries than less popular ones. The latter category accounts for the vast majority of YouTube's content and have their views coming from a small number of countries.

Building on these insights, we propose DTube, a system that accurately predicts the origins of a video's *future* views by looking at its position in the *content graph* and *proactively* places its replicas close to its *expected* consumers. Although the impact of *content locality*, i.e. proximity in the *content graph*, on a video's views and *geographically concentrated* viewing patterns have been studied independently [14, 18, 2], to the best of our knowledge, this is the first work that considers both aspects to optimize the placement of UGC.

We show that our system manages to deliver videos over

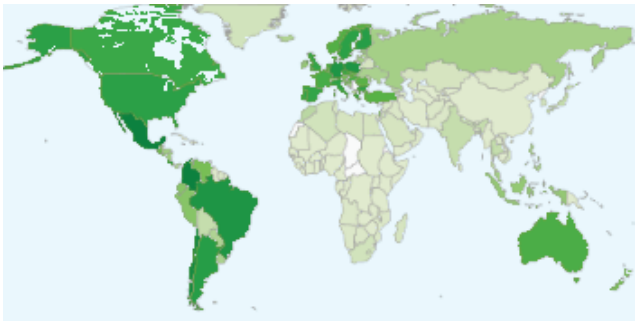

Figure 1: Geographic distribution of the origin of views for a sample YouTube video.

shorter distances than a standard caching-based solution, thus reducing network latencies and improving user experience. In particular, DTube can decrease the proportion of remote requests, i.e., that cannot be served by a node in the same country as the requesting users, by as much as 16%. We also show that DTube can decrease the average distance between users and stored videos by up to 29% for remote requests.

The rest of this paper is organized as follows. In Section 2 we show that content and geographic locality are correlated in YouTube. In Section 3 we present DTube's *replica placement* algorithm and we report on its evaluation in Section 4. We survey related work in Section 5 and conclude in Section 6.

2. LOCALITY IN UGC

Using a YouTube dataset we crawled in March 2011, we show that there is a strong correlation between *the geographic distribution* of a video's views and that of its related videos. We then explore how this correlation can be used to predict the geographic distribution of a video's future views.

2.1 Dataset Description

We crawled our own dataset from YouTube, during the first three weeks of March 2011, using snowball sampling with an initial set consisting of the 10 most popular videos for 25 different countries. For each video, we collected three attributes: (i) its list of related videos as provided by YouTube, (ii) its total number of views, and (iii) its *View Source Vector* (VSV). The VSV of a video represents how many views this video received from each country in the world. For most of the videos, the VSV is available, on the statistics page, as a color map (Fig. 1) generated by a specific URL (`charts.apis.google.com`) containing (country, #views) couples encoded in Google's *Simple Encoding Format*. In our experiments, we extracted the actual VSVs from these URLs. Tab. 1 shows the distribution of the views (top 8 countries) at the granularity of a country, over the whole dataset. The original dataset contained $1,063,844$ videos. We removed the videos with no VSV, and filtered out non-crawled videos from the related video lists. This left us with $689,265$ videos, each having 8 related videos on average, for a maximum of 25 related videos allowed in YouTube.[1]

In the following, we use the view-per-country information

[1]Because the geographic information in our dataset is given at the granularity of a country, we conduct our analysis at the same granularity throughout the paper.

Country	US	CA	GB	BR	JP	DE	PL	AU
Prop. of views (%)	6.6	3.1	3.0	3.0	2.6	2.5	2.2	2.2

Table 1: Geographic distribution of views at the granularity of a country (top 8 countries).

contained in the VSVs to analyze the geographic distribution of views in our dataset. Our goal is to explore the feasibility of a geographically-driven *proactive* placement mechanism that places videos in countries where they are likely to be viewed.

Category	# views	% of videos	% of total views
C_1	$[0, 10^4]$	42.0	0.49
C_2	$(10^4, 10^5]$	33.5	5.10
C_3	$(10^5, 10^6]$	19.7	25.18
C_4	$(10^6, 10^7]$	4.4	45.02
C_5	$(10^7, 10^8]$	0.3	21.10
C_6	$(10^8, \infty)$	0.04	3.10

Table 2: Popularity categories and statistics

2.2 Popularity vs. Geographic Distribution

We first investigate the link between a video's overall popularity and the geographical distribution of its views. To this end, we partition our dataset into six categories based on a video's number of views. The distribution of videos shown on Tab. 2 confirms earlier analysis [2, 3], highlighting a long-tail distribution of views. Very popular videos (categories C_5 and C_6) represent less than 1% of videos while accounting for almost 25% of all views. Because of the long tail, the bandwidth cost of "unpopular" videos is however far from being negligible: The 3 least popular categories (C_1, C_2, C_3, or 95.2% of all videos) still represent more than 30% of the views.

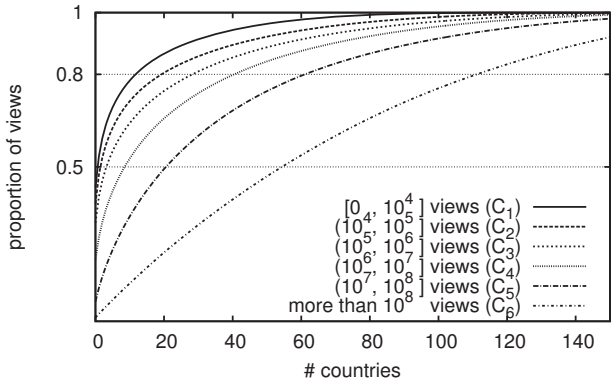

Figure 2: Geographic cumulative distribution of views for various popularity categories

To analyze the geographic distribution of views for each category, we compute the average geographic spread of video views as follows. For a given video, we sort the countries of its VSV in decreasing order according to the proportion of views originating from each country. We then compute the cumulative distribution of views of each video and plot

the average over each popularity category (Fig. 2). A point (n, m) on the graph means that the n top countries for videos in this category account for m% of the total number of views.

Fig. 2 clearly shows that the views of niche videos (C_1, less than 10,000 views) are geographically highly concentrated: 80% of all views for videos in C_1 come from less than 15 countries. This phenomenon fades out as the popularity of videos increases, to reach an almost uniform global distribution for extremely popular videos (C_6).

Yet, this concentration effect remains relatively strong for all videos up to 100M views (categories C_1–C_5, 96.9% of all views). This shows that a proactive placement mechanism could gain from accurately predicting the top n countries from which a video's views originate. For instance, for videos in C_3 (between 100,000 and 1M views, 25.18% of all views), proactively placing video replicas in (or close to) the top 25 countries would cover 80% of all views.

We further study the geographic spread of the origins of the views by taking into account the distances between the main sources of views. The motivation behind this experiment is that it is easier to serve a video with low latency, with a single replica, for users in France and in Switzerland than for users in the UK and in India. For each video, we compute the average of the pairwise distance between the main sources of views (i.e., the top country covering 80% of the views), weighted by the proportion of views each country is responsible for. For instance, for a video viewed $1,000$ times, whose three main sources are US (500 views), UK (200 views), and Japan (100 views), our metric is:

$$\frac{(0.5+0.2)d(\mathsf{us},\mathsf{uk}) + (0.2+0.1)d(\mathsf{uk},\mathsf{jp}) + (0.1+0.5)d(\mathsf{jp},\mathsf{us})}{(0.5+0.2)+(0.2+0.1)+(0.1+0.5)},$$

In our dataset, we observed this average distance to be 26% less for unpopular videos ($\sim$5,200 km) than for popular ones ($\sim$7,000 km).

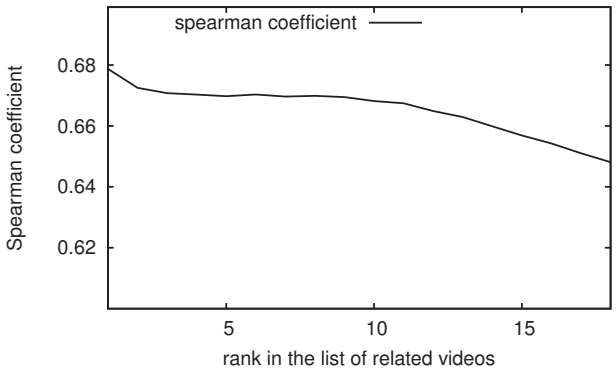

Figure 3: Spearman correlation coefficient between a video's geographic distribution of views and that of its related videos, as a function of their rank in the list of related videos

2.3 Geographic vs. Content Locality

To explore how the top n countries of a video might be predicted, we now turn to the relation between *content locality* and *geographic locality*. For each video, we compute the Spearman correlation coefficient between its sorted VSV (i.e., the list of all countries sorted by decreasing number of views) and that of each of its related videos, and plot it as function of the rank in the list of related videos (see Fig. 3). The Spearman coefficient captures the correlation between the rank of countries in two sorted VSVs, taking into account the permutations of ranks. The closer the absolute value of the coefficient to 1, the more correlated the lists. In our dataset, this correlation is relatively high for all related videos (in 0.64-0.68, Fig. 3) and decreases with the rank. This means that a video's VSV can be inferred from that of its related videos, and that the first related videos are the best candidates.

In order to see if the above finding can be translated into an efficient mechanism for proactively placing video replicas, we conduct the following experiment. For a given video V and its first related $Rel(V)[1]$, we compute for a given number m of replicas, the percentage of views covered by placing them on the first m countries of the VSV of $Rel(V)[1]$, normalized by the percentage of views covered by placing the replicas on the first m countries of the actual VSV of V. The later corresponds to an ideal case where the placement mechanism knows in advance where the views will come from. The results are presented in Fig. 4: even for a small number of replicas, this simple prediction mechanism can accurately follow the actual geographic distribution of the views of a given video. For instance, 85% of the views covered by the first 5 countries of V's VSV are covered by the first 5 countries of $Rel(V)[1]$'s VSV.

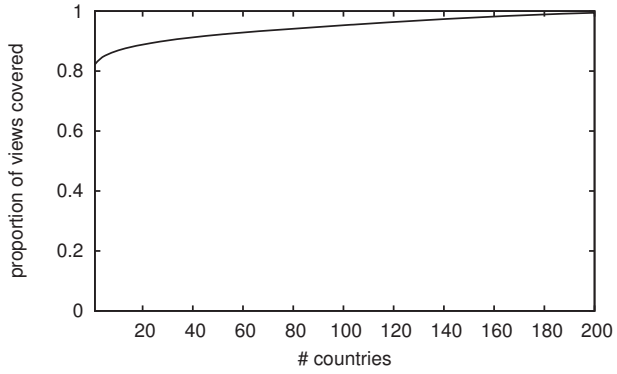

Figure 4: Views covered by placing replicas of a video V in top-countries of the VSV of V's first related video, normalized by the number of views covered when using V's actual VSV.

In summary, unpopular videos, which represent a large proportion of the YouTube dataset, have (i) most of their views originating from a few countries which (ii) spread in a limited region, thus foreseeing a great potential for geographic locality-aware data placement. Furthermore, the geographic distribution of views of a video is strongly correlated with that of its related videos. This implies that the geographic distribution of views of a video can be predicted, but most importantly, it makes the case for a placement mechanism in which videos close in the content graph are stored geographically close to one another.

3. DTUBE

Building on the insight from the previous section, we propose DTube, a proactive placement mechanism that places videos close to their future requests, extracting geographical patterns from the content graph.

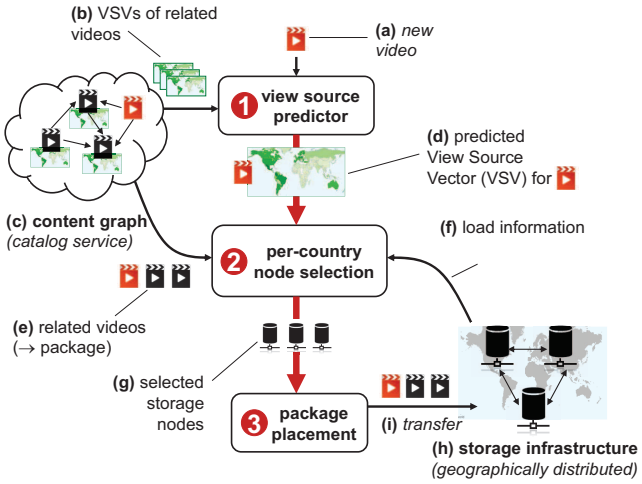

Figure 5: Overview of DTube's placement strategy.

3.1 System model

We consider a UGC sharing system that uses geographically-distributed nodes as its storage infrastructure. Our findings hold for both residential gateways [7, 11], and peer-assisted CDN systems [8].

DTube assumes the existence of a *catalog service* that holds, for each video, the location of its replicas, its current View Source Vector and its meta-data including the list of its related videos. This service is used to retrieve up-to-date meta-data about videos. We further assume the existence of a recommendation algorithm that dynamically computes videos' lists of related videos and updates the catalog accordingly.

3.2 Video placement

The key steps of DTube's placement mechanism are depicted in Fig. 5. When a new video V is uploaded to the system **(a)**, YouTube's recommendation mechanism computes the related videos list, thus making it part of the content graph **(c)**. DTube then estimates V's main future view sources, $\widehat{\mathsf{VSV}}(V)$ **(d)** which is later used to place replicas of V on $\Re$ storage nodes. $\Re$ is a system parameter called *replication factor* and corresponds to the minimum number of replicas a video must have to ensure *durability* in case of storage node failure. In Section 2 we showed that the higher the position of a video in V's related videos list, the higher the correlation between its VSV and the one of V. Applying this finding, we compute $\widehat{\mathsf{VSV}}(V) = \mathsf{VSV}(Rel(V)[1])$ **(d)** by obtaining the VSV of its first related video **(b)** from the catalog **(d)**. One may envision a more sophisticated prediction strategy that combines the VSV of several related videos. Yet simple, our strategy performs well (see Section 4).

Each of the $\Re$ replicas of V attracts a replica of each of V's related videos $Rel(V)$ **(e)**. We call this bundle of $|Rel(V)|+1$ videos a *package*, with V being a *primary* replica and the others, *secondary* replicas. This *package* mechanism creates a coupling between *content* and *geographic* locality, as related videos are placed on the same node. In addition, a coupling is established between a video's *number of replicas* (primary and secondary) and its *in-degree* in the content graph. This is a desirable property as it is shown in [18] that there is a strong correlation between the view count of a video and those of its top referrer videos, i.e., its in-degree.

Each *package* of a video V is placed on a node in each of the $\Re$ first countries of $\widehat{\mathsf{VSV}}(V)$ **(d)** as most views are expected to come from these countries. To minimize transfer and storage costs, only replicas of the videos that *do not* exist in the country are transferred. In addition, to evenly balance the load among the nodes in the system, for a node to be eligible to store a new *package*, the number of videos it already stores must be lower than the average storage load over all nodes (this value can be computed with a standard averaging gossip protocol). Finally, copies of the package are transferred **(i)** to the selected nodes **(g)**.

4. EVALUATION

In this section, we evaluate through simulations the performance of DTube with respect to the geographic distance between users and the storage nodes serving the videos and compare it against a system that employs *reactive* caching on top of persistent storage.

4.1 Evaluation Setup

We distribute the storage nodes in countries according to the proportion of views originating from this country as observed in our dataset (see Tab. 1). We set the number of storage nodes to 10,000 and we consider the videos from our dataset, with the corresponding popularity and the content-graph induced by the related video feature.

We generate synthetic view traffic based on individual users' behavior, using the model proposed in [18], with the popularity values from our dataset: We consider a number of users (50,000 in our experiments), distributed across all countries as storage nodes according to the geographic distribution of views observed in our dataset (see Tab. 1). The number of videos a user watches during a session is picked at random, with an average value of 10. The first video V a user watches is selected from the whole set of videos according to the probability of this video being watched in her country, i.e., the number of views for V originating from her country divided by the total number of views originating from her country (for all videos). Each subsequent video she watches is selected among the related videos of the previous video, excluding already viewed ones. The probability of a video being picked is set to be inversely proportional to the video's rank in V's list of related videos, following a Zipf distribution.

4.2 DTube and Alternatives

We compare DTube against standard caching. For DTube, we use the placement algorithm as described in Section 3. In addition, we implement and evaluate several variations of DTube to identify the performance gains conveyed by the different mechanisms involved, namely without the use of packages (Partial DTube) and using the actual VSV of the video (Ideal DTube) instead of that of its first related video. Ideal DTube can be thought of as an upper bound on DTube's performance and reflects how the efficiency of the VSV prediction mechanism (evaluated in Section 2) translates in practice with respect to the viewer experience. In our experiments, videos are served from the node closest to the user.

As for caching, we consider a storage infrastructure composed of *persistent* storage nodes (e.g., YouTube servers)

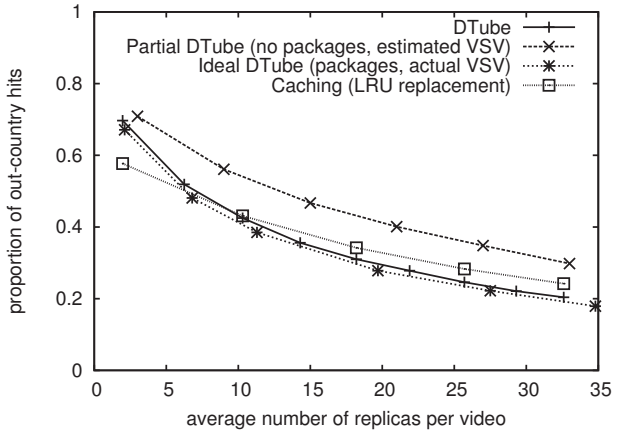

Figure 6: Proportion of out-country requests with DTube and caching.

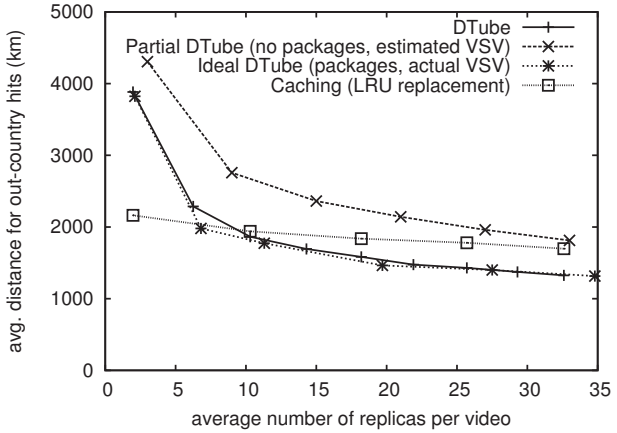

Figure 7: Average distance to storage node for out-country requests with DTube and caching.

and CDN *caching* nodes (e.g., Akamai servers). The persistent storage nodes hold a complete copy of the YouTube dataset, thus ensuring durability. Videos are only served by CDN nodes, the caches of which are populated in a reactive fashion, based on the users' view traffic: Consider a user located in a given country who requests a given video. If the video is stored on a CDN node in this country, it is served from this node to the user. If not, the video is first fetched from a persistent node to a random CDN node (with free storage space) in the country and then served. If none of the CDN nodes in the country has sufficient free storage space to store the video, we apply LRU cache replacement: the least recently used video is replaced by the new entry.

4.3 Evaluation Results

We evaluate and compare the performance of all placement strategies with respect to the geographic distance between the user and the node serving the video. More specifically, we look at (i) the out-country hit-rate, that is the proportion of requests that are served from a storage node (i.e., a gateway or a CDN node) located in a different country than the user, and (ii) the distance between the user and the storage node when the video is served from a different country. We assume that networking infrastructure is usually well integrated in each country, thus in-country hits are likely to encounter better network quality and that geographic distance is a good indicator of transfer latency.

In order for DTube and caching to be comparable, we use the same storage space in both. More specifically, for a given replication factor $\Re$, we first run simulations with DTube. Because it makes use of packages, the average number of replicas R per video is larger than $\Re$. We therefore run simulations with caching, for a system composed of $\Re$ persistent storage nodes and CDN nodes with a storage space of $(R-\Re)\times$(total number of videos)/(number of CDN nodes), which corresponds to the same total storage space as for DTube. We evaluate our metrics at steady state, i.e., when all the caches of all CDN nodes are full.

Fig. 6 depicts the out-country hit-rate for the different versions of DTube and caching. It can be observed that for larger values of the average number of replicas per videos, DTube outperforms the caching-based solution. For instance, for an average number of 30 replicas per video, DTube decreases the proportion of out-country request from 0.25 to 0.21, that is a 16% improvement. By comparing DTube to Ideal DTube and Partial DTube, we observe that (i) the use of packages accounts for a significant part of DTube's performance, (ii) the estimation of a video's VSV from that of its related videos (i.e., based on the content graph) incurs only a little decrease in performance compared to an omniscient solution in which the actual VSV is known in advance. This illustrates the synergy between our geographic prediction and the use of packages. Similar results can be observed in Fig. 7, which depicts the average distance between the user and the storage node serving the video for out-country requests. For instance, with an average of 30 replicas per videos, DTube reduces the average distance by 29%. Note that the distance remains relatively high as the distance to the closest country can be large, e.g., ∼2000km between the US and Canada which are the two main sources of views.

5. RELATED WORK

YouTube has generated numerous studies on user behavior and video characteristics: Cha *et al.* propose to use a video's history to predict its future demand [3], while Zhou *et al.* in [18] study the impact of a video's position in the content graph on its popularity. In [2], Brodersen *et al.* analyze the correlation between popularity and geographical locality. Finally, Torres *et al.* [16] evaluated YouTube's CDN performance.

In [11], the authors show the feasibility of building a system like YouTube, using residential gateways. The concept of CDNs composed of gateways has been investigated for decentralized video storage and delivery in [7]. Peer-assisted CDNs also received a great deal of attention lately [8, 4, 17]. However, none of the proposed systems leverages the content graph.

In [14], monitoring social cascades in online social networks is proposed to predict a video's future view pattern. Kangasharju *et al.* [9] investigate optimal placement strategies in P2P content networks to maximize availability in content communities. Tan *et al.* [15] investigate the same problem for VoD but focus on upload bandwidth.

Volley [1] leverages the content graph to place the data so that the perceived latency is decreased. But contrary to

DTube, there is only one copy of each content item as data durability is not considered. NetTube [5] is a peer-assisted VoD system that leverages the content graph of UGC videos through social-aware pre-fetching and overlays to optimize swarming and decrease start-up delays. Finally, SPAR [12] is a social partitioning and replication system that achieves *one-hop replication* of user profiles in social networks.

6. CONCLUSION

In this paper we have highlighted the correlation between content locality and geographic locality in User Generated Content (UGC). More precisely, we have shown using a large YouTube dataset that related videos present similar geographic viewing patterns and that, except for extremely popular videos, video views are concentrated in a limited number of countries.

This coupling between content and geographic locality in UGC system has led us to propose DTube, a decentralized storage infrastructure which *proactively* places content close to their future requests leveraging on the videos' positions in the content graph.

In the future, we plan to investigate the serving part of our UGC system: more specifically, how to adapt the number of replicas and bandwidth allocation to the popularity of the videos and how to efficiently prefetch and serve videos from multiple storage nodes. We also plan to consider how caching and geographic view prediction can be combined, e.g., by exploring how predicted views might be used in the cache's replacement strategy.

Acknowledgment

This work has been partially funded by the ERC Starting Grant GOSSPLE number 204742.

7. REFERENCES

[1] Agarwal, S., Dunagan, J., Jain, N., Saroiu, S., Wolman, A., and Bhogan, H. Volley: Automated Data Placement for Geo-Distributed Cloud Services. In *NSDI* (2010).

[2] Brodersen, A., Scellato, S., and Wattenhofer, M. YouTube Around the World: Geographic Popularity of Videos. In *WWW* (2012).

[3] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *IMC* (2007).

[4] Chen, Z., Lin, C., Yin, H., and Li, B. On the Server Placement Problem of P2P Live Media Streaming System. In *PCM* (2008).

[5] Cheng, X., and Liu, J. NetTube: Exploring Social Networks for Peer-to-Peer Short Video Sharing. In *INFOCOM* (2009).

[6] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. YouTube Traffic Characterization: A View From The Edge. In *IMC* (2007).

[7] He, J., Chaintreau, A., and Diot, C. A Performance Evaluation of Scalable Live Video Streaming with Nano Data Centers. *Computer Networks 53* (2009), 153–167.

[8] Huang, C., Wang, A., Li, J., and Ross, K. W. Understanding Hybrid CDN-P2P: Why Limelight Needs its Own Red Swoosh. In *NOSSDAV* (2008).

[9] Kangasharju, J., Ross, K. W., and Turner, D. A. Optimizing File Availability in Peer-to-Peer Content Distribution. In *INFOCOM* (2007).

[10] Khemmarat, S., Zhou, R., Gao, L., and Zink, M. Watching User Generated Videos with Prefetching. In *MMSys* (2011).

[11] Marcon, M., Viswanath, B., Cha, M., and Gummadi, K. P. Sharing Social Content from Home: A Measurement-driven Feasibility Study. In *NOSSDAV* (2011).

[12] Pujol, J. M., Erramilli, V., Siganos, G., Yang, X., Laoutaris, N., Chhabra, P., and Rodriguez, P. The Little Engine(s) That Could: Scaling Online Social Networks. In *SIGCOMM* (2010).

[13] Saxena, M., Sharan, U., and Fahmy, S. Analyzing Video Services in Web 2.0: A Global Perspective. In *NOSSDAV* (2008).

[14] Scellato, S., Mascolo, C., Musolesi, M., and Crowcroft, J. Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades. In *WWW* (2011).

[15] Tan, B. R., and Massoulié, L. Adaptive Content Placement for Peer-to-Peer Video-on-Demand Systems. *CoRR abs/1004.4709* (2010).

[16] Torres, R., Finamore, A., Kim, J. R., Mellia, M., Munafò, M., and Rao, S. Dissecting Video Server Selection Strategies in the YouTube CDN. In *ICDCS* (2011).

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

[18] Zhou, R., Khemmarat, S., and Gao, L. The Impact of YouTube Recommendation System on Video Views. In *IMC* (2010).