# Model Draft

**Abstract**

## I. Introduction

Speaking of the Internet evolution for the last few years, an explosion of sociality-aware applications have entered the scenes, *e.g.*, Facebook, Twitter, Foursquare, YouTube *etc*. Benefitting from the power of the omnipotent social networks, an overwhelming trend of Internet multimedia applications have began to gain tremendously fast-growing popularity, *i.e.*, the User Generated Content (UGC). As a representative with over hundreds of millions of registered users, YouTube, reportedly absorbing 48 hours worth of videos uploaded every minute and hitting over 3 billions views per day [1], has been under intensive explorations and study since its debut in 2005 [2][3][4][5][6]. Compared with traditional Internet video services, UGC exhibits several unique features, which pose greater-than-ever challenges to Content Providers (CP). Among those challenges, "sensitivity to latency" is a key issue impacting the user experiences. Since most of the videos uploaded by regular users are usually short, *e.g.*, several minutes, a latency of more than 30 seconds, which is normal for traditional Internet video services, seems intolerable in UGC scenario. On the other hand, the sociality-awareness also brings potential opportunities. CPs are ready to facilitate users' willingness to participate in terms of creativity, collaboration and sharing, by exploiting the underlying social connections among both users and videos. Although the recommendation mechanism applied by most of the UGC systems lies mostly in video correlations while user recommendation contributes only to a limited extent[3] , we believe the contributions from the latter are seriously underestimated, since a great many accesses are incurred externally from the active users in other social platform, *e.g.*, Twitter and Facebook. Furthermore, people's habits of acquiring the information has shifted a lot as more and more people involve in the Internet social network. People are inclined to follow their friends and listen, though sometimes the sources are not trustworthy. *e.g.*, Bin Laden's death sparked a startling record of 12.4 million tweets per hour even before Obama's official announcement in the White House [7]. For similar reasons, people will likely click and view the videos posted or retweeted from their friends. Thus, designing a social-awareness UGC platform is appealing and promising for both the users and CPs.

The most straightforward solution to the "latency" issue is to put the videos as close as to the viewers. Traditional approaches, by deploying dedicated server resources across different geographical locations and caching the contents, always put CPs in the dilemma of either over provisioning resources which wastes extra investments or rejecting a portion of user requests which jeopardizes the enterprises' growth. What's more, to establish globally geographical server farms are over the top of most companies' capabilities, which forms an insurmountable barrier to the business. Under this circumstance, Cloud Computing is believed to be an ideal rescue to resort to.

Cloud Computing,though deemed as a hype in its early days, has beyond any doubt proven its power in various fields of both industry and academia nowadays. Despite not in its maturity yet, lots of institutes and companies have actively joined to the market, among which the leading giants include Amazon, Google, Facebook, Microsoft *etc*. Apple Inc. also just released its cloud product *i.e.*, *iCloud* in its WWDC held in San Francisco, which is a breaking event on June 6, 2011. As one of the driving forces, the large-scale Internet application ecosystem will surely leverage greatly from the the ideal inborn substrate, *i.e.*, Cloud Computing, which however has been explored to a limited extent.

## II. Work Progress

### *A. Goal*

In this paper, we will try to design a social-awareness cloud platform to facilitate CPs to provide UGC-like services to their users.

### B. Work Flow

1) The cloud-based hosting platform model will be firstly proposed. Then we will design an initial placement algorithm with a upper-bounded latency guarantee. Although the placement algorithm is a polynomial-time one, the computing effort is non-trivial since we have lots of channel to deploy. Thus, the algorithm will be executed only once when the system is launched.
2) Next, we will design our predicting algorithm for the user requests within each time interval, by leveraging the recommendation mechanism including both users' social connections and videos' correlations.
3) Lastly, a dynamic adjustment algorithm is proposed, with user requests derived from 2). The algorithm will be triggered every time interval and the content storage and request dispatching will vary over the time. In addition, the algorithm will be triggered whenever a new video is posted.

### C. Challenges

1) The initial placement algorithm is almost done. But in future, how to make the algorithm more feasible should also be considered, since the number of videos is huge.
2) The most difficult job is how to leverage the users' social connections and video correlations to predict the user requests in near future. We are considering to leverage a regression model. To derive a linear regression relationship among the viewers in the next time interval *i.e.*, $D(T+1)$, the number of friends of current viewers *i.e.*, $V(t)$, the upload time stamp, *i.e.*, $t_0$, and the video recommendation list. But the social effect brought by users' social network and videos' correlations are sort of correlated, how to separate them? Another question is, are there any linear relationship among those variables, which is a fundamental assumption?
   Besides, I am considering the possibility to use a epidemic model. Still investigating.
3) The dynamic adjustment algorithm lies mostly in how we will utilize the social connections in 2). Prefer a physical model to solve it, *i.e.*, spring model. Or we can derive a incremental pattern from the initial placement algorithm. I haven't considered too much on this.

## III. Model Overview

The model consists of a set of geographically diverse cloud clusters $F$, a set of videos $O$ and a set of clients $D$. ($D_f$ denotes the consolidated viewer group within domain of cloud cluster $f$ ($f \in F$).) Without loss of generality, I will first assume all the videos have unit length. (Extend this when the model is done)

### A. Alphabet Soup

1) Each cloud cluster is assigned a storage capacity, $S_f$.
2) Each cloud cluster has a bandwidth capacity, $\mu_f$. Here I plan to borrow the idea from our ICDCS paper's model *i.e.*, bandwidth will be abstracted into a VM instance, which will actually provide the bandwidth.
3) $x_{jf}^{(o)}$ denotes the variable indicating whether the request for video $o$ issued from viewer $j$ will be directed to cloud cluster $f$.
4) $y_f^{(o)}$ denotes the variable indicating whether to store a copy of video $o$ at the cloud cluster $f$.
5) $c_f$ denotes the storage cost of cloud cluster $f$.
6) $v_f$ denotes the transferring cost from viewer $j$'s location to cloud cluster $f$.
7) $R_{jf}$ denotes the transferring latency from viewer $j$'s location to cloud cluster $f$. It can be assigned with value of RTT between these two geographical regions.

There should be a mapping function to map user $j$ to a location $f$. For simplicity, we can denote it as $D^{-1}(j)$. In that way, we can represent $x_{jf}^{(o)}$, $c_{jf}$ and $R_{jf}$ as $x_{D^{-1}(j)f}^{(o)}$, $c_{D^{-1}(j)f}$ and $R_{D^{-1}(j)f}$ respectively. For clearness, I will use $j$ afterwards.

### B. Objective function

To minimize the operation cost, based on the premise that the expected average global latency should be bounded below some tolerant value.

$$min \sum_{o \in O} \sum_{f \in F} y_f^{(o)} \times c_f + \sum_{o \in O} \sum_{f \in F} \sum_{j \in D_f} x_{jf}^{(o)} \times v_f$$

### C. Constraints

1) *Storage*
$\sum_{o \in O} y_f^{(o)} \leq S_f$ , $\forall f \in F$

2) *VM capacity*
$\sum_{o \in O} \sum_{j \in D} x_{jf}^{(o)} \leq \mu_f$, $\forall f \in F$

3) *Placement*
$\sum_{f \in F} y_f^{(o)} \geq 1$, $\forall o \in O$

4) *Latency guarantee*
$\frac{\sum_{o \in O} \sum_{f \in F} \sum_{j \in D_f} x_{jf}^{(o)} \times R_{jf}}{|D|} \leq R_{threshold}$, where $R_{threshold}$ is an input into the system.

5) Variable constraint
$y_f^{(o)} \in \{0, 1\}, x_{jf}^{(o)} \in \{0, 1\}$

6) $x_{jf}^{(o)} \leq y_f^{(o)}$, $\forall j \in D$

## IV. Alternative LP (Relaxation)

Obviously, the optimization in Sec. III is an integer problem. Here we want to make an intuitive relaxation. The reason is two-fold. First, the number of users makes the optimization problem too large to solve. Second, we want to transform the original problem into a more tractable one.

### A. Consolidate users

As what we have assumed, at any time, each user can at most view one video. So we can treat the users within one specific region $f$ ($f \in F$) as one, which will make our optimization much slimmer. Based on that, we are able to eliminate all the variables $x_{jf}^{(o)}$ and consolidates them as one viewer. Suppose the total user set at time slot $T$ is represented as $D(T)$, so the viewer set in region $f$ at that time slot is $D_f(T)$. We introduce a new variable $\alpha_{jf}^{(o)}$, which denote the portion of request for video $o$ issued from the aggregate user $j$ to cloud cluster $f$. To note that, $\alpha_{jf}^{(o)}$ is a fractional variable. Due to our charging mode, the storage deployment is scheduled at a larger time scale while the VM rental is done at a smaller one, i.e. T. From above, the original ILP has only one type of integer variable $y_f^{(o)}$ and the original optimization problem is changed to,

$$min \sum_{o \in O} \sum_{f \in F} y_f^{(o)} \times c_f + \sum_{o \in O} \sum_{f \in F} \sum_{j \in F} \alpha_{jf}^{(o)} \times D_f(T) \times v_f \quad (1)$$

*B. Constraints*

1) *Storage*
$\sum_{o\in O} y_f^{(o)} \leq S_f$ , $\forall f \in F$

2) *VM capacity*
$\sum_{o\in O}\sum_{j\in F} \alpha_{jf}^{(o)} \times D_f(T) \leq \mu_f$, $\forall f \in F$

3) *Placement*
$\sum_{f\in F} y_f^{(o)} \geq 1$, $\forall o \in O$

4) *Latency guarantee*
$\frac{\sum_{o\in O}\sum_{f\in F}\sum_{j\in F}\alpha_{jf}^{(o)}\times D_f(T)\times R_{jf}}{|D|} \leq R_{threshold}$, where $R_{threshold}$ is an input into the system.

5) Variable constraint
$y_f^{(o)} \in \{0,1\}, \alpha_{jf}^{(o)} \in [0,1]$
6) $\alpha_{jf}^{(o)} \leq y_f^{(o)}$, $\forall j, f \in F$
7) $\sum_{f\in F} \alpha_{jf}^{(o)} = 1$

*C. How to solve?*

If we relax the variable $y_f^{(o)}$, the optimization problem in Sec. IV-A is a problem with complicating constraints and we can utilize an efficient dual decomposition to solve it. More specifically, the original constraint (5) is relaxed into $y_f^{(o)} \in [0,1], \alpha^{(o)jf} \in [0,1]$. The original problem can be formulated as,

$$\begin{aligned} &min \sum_{o\in O}\sum_{f\in F} y_f^{(o)} \times c_f + \sum_{o\in O}\sum_{f\in F}\sum_{j\in F}\alpha_{jf}^{(o)} \times D_f(T) \times v_f \\ &s.t. \begin{cases} y_f^{(o)} \in \mathbb{C}_1 (\forall o \in O, f \in F) \\ \alpha_{jf}^{(o)} \in \mathbb{C}_2 (\forall o \in O, j, f \in F) \\ \alpha_{jf}^{(o)} - y_f^{(o)} \leq o (\forall o \in O, j, f \in F) \end{cases} \end{aligned} \tag{2}$$

$\mathbb{C}_1$ is the convex set defined by linear constraints (1), (3) and (5), while $\mathbb{C}_2$ is the convex set defined by linear constraints (2), (4), (5) and (7).

*1) Dual Decompostion:* Let $\mathbb{L}(y, \alpha, L)$ be the Lagrangian problem of Eqn. 2, where $y$ and $\alpha$ is column-formatted primal variables,
*i.e.,* $y = (y_1^1, y_1^2, \ldots, y_1^{|O|}, y_2^1, y_2^2, \ldots, y_{|F|}^{|O|})^T$.

$$\begin{aligned} &\mathbb{L}(y, \alpha, L) \\ &= \sum_{o\in O, f\in F} y_f^{(o)} \times c_f + \sum_{o\in O, f\in F, j\in F} \alpha_{jf}^{(o)} \times D_f(T) \times v_f + \sum_{o\in O, f\in F, j\in F} \lambda_{jf}^{(o)} \times (\alpha_{jf}^{(o)} - y_f^{(o)}) \\ &= [\sum_{o\in O, f\in F} y_f^{(o)} \times c_f - \sum_{o\in O, f\in F, j\in F} \lambda_{jf}^{(o)} \times y_f^{(o)}] + [\sum_{o\in O, f\in F, j\in F} \alpha_{jf}^{(o)} \times (D_f(T) \times v_f + \lambda_{jf}^{(o)})] \end{aligned} \tag{3}$$

Obviously, Eqn. 3 can be easily decomposed into two sub problems, *i.e.*, shown as Eqn. 4

$$\begin{aligned} &g_1\lambda = \sum_{o\in O, f\in F} y_f^{(o)} \times c_f - \sum_{o\in O, f\in F, j\in F} \lambda_{jf}^{(o)} \times y_f^{(o)} \quad (A) \\ &s.t. y_f^{(o)} \in \mathbb{C}_1 (\forall o \in O, f \in F) \\ &g_2(\lambda) = \sum_{o\in O, f\in F, j\in F} \alpha_{jf}^{(o)} \times (D_f(T) \times v_f + \lambda_{jf}^{(o)}) \quad (B) \\ &s.t. \alpha_{jf}^{(o)} \in \mathbb{C}_2 (\forall o \in O, j \in F, f \in F) \end{aligned} \tag{4}$$

Both sub problems are standard linear optimizations with efficient polynomial-time solutions. However, a subsequent rounding for sub problem (A) seems essential since only integrals (*0, 1*) are feasible in the

original problem (Eqn. 1). We will prove that the optimal solution of (A) are integrals.

**Theorem 1.** *The optimal solutions for sub problem (A) are integrals.*

*Proof:* Since (A) is LP, the optimal solution should be a vertex. We have known that If $A$ is *totally unimodular*, then every vertex solution of $Ax \leq b$ is integral. So we prove the Theorem. 1 iff we can show the constraint matrix is *totally unimodular*. As mentioned, all the constraints of sub problem (A) is,

$$\begin{cases} \sum_{o \in O} y_f^{(o)} \leq S_f, \forall f \in F \\ -\sum_{f \in F} y_f^{(o)} \leq -1, \forall o \in O \\ y^{(o)} \leq 1, \forall o \in O, f \in F \\ -y^{(o)} \leq 0, \forall o \in O, f \in F \end{cases} \tag{5}$$

Thus, if we present Eqn. 7 into $Ax \leq b$. A =

$$\begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ & \vdots & & \vdots & & \vdots & \vdots & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \\ -1 & \cdots & 0 & -1 & \cdots & 0 & \cdots & -1 & \cdots & 0 \\ & \ddots & \vdots & & \ddots & \vdots & \ddots \; \vdots & & \ddots & \vdots \\ & & -1 & & & -1 & \quad -1 & & & -1 \\ \mathrm{E} & & & 0 & & & \cdots & 0 & & \\ 0 & & & \mathrm{E} & & & \cdots & 0 & & \\ & & & & & & \ddots & \vdots & & \\ & & & & & & & \mathrm{E} & & \\ \text{-E} & & & 0 & & & \cdots & 0 & & \\ 0 & & & \text{-E} & & & \cdots & 0 & & \\ & & & & & & \ddots & \vdots & & \\ & & & & & & & \text{-E} & & \end{pmatrix}$$

$|F|$ (rows 1–3), $|O|$ (rows 4–6), $|F| \times |O|$ (rows 7–10), $|F| \times |O|$ (rows 11–14)

$|O|$ (col 1–3), $|O|$ (col 4–6), $|O| \times (|F|-3)$ (col 7–8), $|O|$ (col 9–11)

$x = (y_1^1, y_1^2, \ldots, y_1^{|O|}, y_2^1, y_2^2, \ldots, y_2^{|O|}, \ldots, y_{|F|}^{|O|})^T$

$b = (S_1, S_2, \ldots, S_F, -1, -1, \ldots, -1, 1, 1, \ldots, 1, 0, 0, \ldots, 0)^T$

For any 1-by-1 sub matrix, we know the determinant can be {-1, +1, 0} since the entry of the constraint matrix is {+1, -1, 0}. By inductive hypothesis, we assume the determinant of any sub matrix $M'$ with a dimension of no greater than N is {-1, +1, 0}. We will try to prove the determinant of any sub matrix $M_{(N+1)\times(N+1)}$with a dimension of N+1 is {-1, +1, 0}.

For $M_{(N+1)\times(N+1)}$,

(i) If there is a row $r$ picked from $[|F| + |O| + 1, |F| + |O| + |F| \times |O|]$

We can easily see that, the corresponding row of $M_{(N+1)\times(N+1)}$ has at most one non-zero entry, *i.e.*, 1. Denote $r'$ be the corresponding row in $M_{(N+1)\times(N+1)}$. Thus,

$$det(M_{(N+1)\times(N+1)}) \;=\; \begin{cases} 0 & M_{r'} = \vec{0} \\ \{+1,-1\} \times det(M'_{N\times N}) = \{+1,-1,0\} \end{cases}$$

(ii) If there is a row $r$ picked from $[|F|+|O|+|F|\times|O|+1, |F|+|O|+2\times|F|\times|O|]$
Similar proof as (i).

(iii) If the submatrix is totally picked from the upper rows $[1, |F|+|O|]$
From the famous *Ghouila-Houri's characterization* [8], we can see that

A' =

$$\left(\begin{array}{ccc:ccc:ccc:ccc} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & & 0 & \cdots & 0 \\ & \vdots & & \vdots & & \vdots & \vdots & & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & & 1 & \cdots & 1 \\ \hdashline -1 & \cdots & 0 & -1 & \cdots & 0 & \cdots & & -1 & \cdots & 0 \\ & \ddots & \vdots & & \ddots & \vdots & \ddots & \vdots & & \ddots & \vdots \\ & & -1 & & & -1 & & -1 & & & -1 \end{array}\right)$$

is totally unimodular, since each column of $A'$ has exactly one $1$ and $-1$. So the determinant of any sub matrix is {+1, -1, 0}.

■

*2) Master Algorithm:* Like ordinary sub-gradient algorithms, an iterative algorithm (Table. I) is applied. In each cycle, with fixed Lagrangian variables $\lambda$, we solve those two subproblems derived in Sec. IV-C1, together with sub gradients $\partial g_1(\lambda)$ and $\partial g_2(\lambda)$. ($\partial g_1(\lambda_{jf}^{(o)}) = -y_f^{(o)}$, $\partial g_2(\lambda_{jf}^{(o)}) = \alpha_{jf}^{(o)}$)

TABLE I
MASTER ALGORITHM

| **Repeat** |
|---|
| Solve subproblem $g_1(\lambda)$ over $y$ |
| Solve subproblem $g_2(\lambda)$ over $\alpha$ |
| Update dual variables $\lambda_{jf}^{(o)} := \lambda_{jf}^{(o)} + \beta_k \times (\alpha_{jf}^{(o)} - y_f^{(o)})$ |

## V. SOCIALITY

As mentioned, $D(t)$ is the current viewers at time $t$. Let $V$ denote the potential viewers in future. Intuitively, $V$ is composed of the uploader's direct friends, the friends' friends and so on.

### *A. The model of users' social connections*

It's difficult to give an accurate model for a random social network. In contrast, we only try to capture the soul. So a series of assumption will be made beforehand.

1) The social connection in our paper is directed. *i.e.*, the action by a user will only impact his followers. If we denote $A$ as the adjacent matrix among users, its entry $a_{ij}$ indicates the connection between user $i$ and $j$. More specifically, "$a_{ij} = 1$" means user $j$ is following $i$.
2) The social effect of an user is decided by the number of his followers. The more followers, the more influencing the user is.

3) Different videos have different lifespans. News and other timeliness-oriented videos usually have shorter lifespan, while other videos last longer. In our paper, we will assume all the videos have an exponentially decreasing *influence index*, denoted by $\gamma^t$ ($\gamma < 1$). For those videos which have a short lifespan, a smaller $\gamma$ is chosen. For those which a long lifespan, a bigger $\gamma$ is chosen. The physical meaning is the probability that a follower will click and view a tweeted or re-tweeted video.

### *B. How to estimate $V$*

Generally, the poster of a video has a great impact on the popularity of the video. The more followers he has, the wider the video is likely spread. Besides, after the poster's friends see the video, they may "like" it and re-tweet (recommend) the video to their own friends. In this case, the average followers for an individual user *i.e.*, $\bar{a}$, is needed for our estimation. Formally, it can be derived as $\bar{a} = \frac{\sum_{ij} a_{ij}}{2 \times N}$. But in practice, it is usually derived in a statistical fashion due to the huge scale of the network.

Based on the assumptions above, if a poster has $a_0$ direct followers, the video has an influence index of $\gamma^t$. We can estimate

$$V = a_0 \times \gamma + a_0 \times \gamma \times \bar{a} \times \gamma^2 + \ldots + a_0 \times \gamma \times \bar{a}^{t-1} \times \gamma^{\frac{(t-1)\times t}{2}} \tag{6}$$

, where $a_0 \times \gamma \times \bar{a}^{t-1} \times \gamma^{\frac{(t-1)\times t}{2}} < 1$.

### *C. How to denote $D(t)$*

By utilizing a SIR epidemic model, we can get a differential equation as

$$\begin{cases} \frac{dV}{dt} = -\gamma^t \times V \times D \\ \frac{dD}{dt} = \gamma^t \times V \times D - \rho \times D \end{cases} \tag{7}$$

, where $\rho$ denotes the interest lost rate. The bigger $\rho$ is , the faster a viewer who has seen the video will lose interest.

## REFERENCES


[1] *YouTube Blog*, http://youtube-global.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html.

[2] V. K. Adhikari, S. Jain, and Z.-L. Zhang, "YouTube Traffic Dynamics and Its Interplay with a Tier-1 ISP: An ISP Perspective," UMN, Tech. Rep., November 2010.

[3] R. Zhou, S. Khemmarat, and L. Gao, "The Impact of YouTube Recommendation System on Video Views," in *Proc. of ACM IMC*, November 2010.

[4] X.Cheng, J.Liu, and C.Dale, "Understanding the Characteristics of Internet Short Video Sharing: A YouTube-based Measurement Study," *IEEE Transactions on Multimedia*, 2010.

[5] R. Torres, A. Finanmore, J. Kim, M. Mellia, Maurizio, M.Munafo, and S. Rao, "Dissecting Video Server Selection Strategies in the YouTube CDN," in *Proc. of IEEE ICDCS*, June 2011.

[6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Characterizing User Sessions on YouTube," in *Proc. of ACM/SPIE. MMCN*, Jan 2008.

[7] *How Fast the News Spreads Through Social Media*, http://blog.sysomos.com/2011/05/02/how-fast-the-news-spreads-through-social-media.

[8] A. Ghouila-Houri, "Charactrisations des Matrices Totalement Unimodulaires," *Comptes Rendus de I' Acadmie des Sciences*, pp. 1192 – 1193, 1962.