Twitter-driven YouTube Views: Beyond Individual Influencers

Honglin Yu, Lexing Xie, Scott Sanner
The Australian National University and NICTA, Canberra, Australia
honglin.yu@anu.edu.au, lexing.xie@anu.edu.au, ssanner@nicta.com.au

ABSTRACT

This paper proposes a novel method to predict increases in YouTube viewcount driven from the Twitter social network. Specifically, we aim to predict two types of viewcount increases: a sudden increase in viewcount (named as JUMP), and the viewcount shortly after the upload of a new video (named as EARLY). Experiments on hundreds of thousands of videos and millions of tweets show that Twitter-derived features alone can predict whether a video will be in the top 5% for EARLY popularity with 0.7 Precision@100. Furthermore, our results reveal that while individual influence is indeed important for predicting how Twitter drives YouTube views, it is a diversity of interest from the most active to the least active Twitter users mentioning a video (measured by the variation in their total activity) that is *most* informative for both JUMP and EARLY prediction. In summary, by going beyond features that quantify individual influence and additionally leveraging collective features of activity variation, we are able to obtain an effective cross-network predictor of Twitter-driven YouTube views.

Category and Subject Descriptors H.2.8 DATABASE MANAGEMENT Database Applications — data mining

Keywords Popularity prediction; social media; YouTube; Twitter

1. INTRODUCTION

Predicting item popularity in social media is a well-recognized open problem. Take YouTube videos, for example, where particular interesting and challenging questions include "will an obscure video suddenly become very popular, and when?", "which videos will be the top 5% popular ones in 1, 2, or 3 months?", and "when will the attention on a very popular video fade out, if ever?" Questions like these are inherently difficult, because determining popular items in social media not only relates to the attention dynamics of online videos, it is also affected by many complex factors from virality to news events to seasonality that make it difficult to quantify the interplay of such phenomena. Despite the mounting challenge, such prediction tasks are important both in their value for understanding collective online behavior, and in a wide range of applications from content discovery and recommendation to video distribution infrastructure to online advertisement.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

ACM http://dx.doi.org/10.1145/2647868.2655037.

This work is related to several areas in online video and social media popularity. Early studies included large-scale profiling of content [5] and their social networks [7]. Subsequent measurements provided statistics on video popularity [6] and its geographic distribution [4]. A number of recent works focused on forecasting popularity: YouTube video views are found to be predictable from recent viewcount history [13, 15], aggregate video sharing behavior is found to be related to geographical, social and temporal sharing trends [16], and the lifecycle of trending topics in the news are found to generalize from historically similar topics across multiple information platforms [2]. More recently, researchers started to examine video sharing behavior on microblogs. Abisheva *et al* [1] performed a large-scale profiling of the YouTube video audience by analyzing Twitter and Roy *et al* [14] used tweets to improve YouTube video recommendation. In video popularity prediction, viewcount-based methods tend to fit smooth trends well, but are inherently unable to predict sudden changes in popularity that may be driven by external sources such as Twitter. Cross-platform social media content has seen recent success for predicting individual consumption [19]. In this paper, we investigate whether Twitter feeds can also help predict YouTube video popularity.

This work investigates two particular challenging scenarios for YouTube video popularity prediction. The first is a sudden increase in viewcount, called "jumps", which are often triggered by external events and referrals. Fig 1(a) shows a short video having an abrupt increase of 1.5 million views after receiving little attention in the first 4 months after its upload – interestingly, we also note a few hundred tweets linking to this video right around the jump. The second scenario is predicting the popularity of newly uploaded videos, when no or very little viewcount history is available; this task can also be helped by information from external networks – as shown in Fig 1(b), a wave of tweets about this video started right around its upload, and viewcount continued to rise even after the tweet volume faded out. Our insight for tackling these two challenging scenarios is the effective use of content and user information from Twitter. Going beyond the conventional wisdom of social influencers [8], we design features to capture the volume of tweets, the network position and connectivity of users, and the dynamic interactions among tweets and users. We train SVM classifiers using these features for two prediction tasks, called JUMP and EARLY.

Our results demonstrate that Twitter information can be effectively used to predict YouTube viewcount, with a prediction system built from tens of thousands of videos mentioned by millions of tweets during a 3-month period in 2009. We observe that for viewcount JUMP, leveraging Twitter features nearly quadruples the performance from viewcount history, with a 0.46 Precision@100, compared to 0.12 using viewcount alone. When viewcount history is not available (EARLY), we can predict the top 5% popular videos

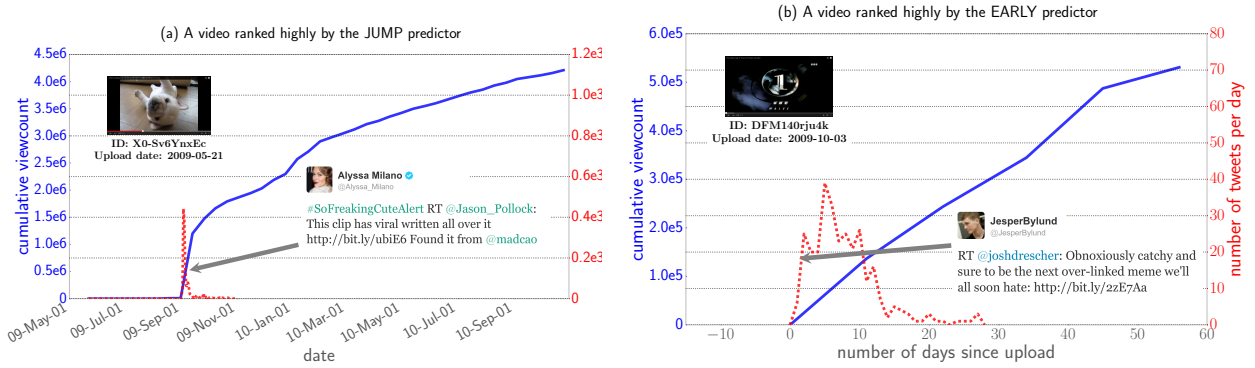

Figure 1: Examples of top predictions for JUMP and EARLY. (a) A video having less than 9000 views in its first 3 months, and then gaining 1.2 million views within 15 days (date format of x-axis: yy-mmm-dd). The insert shows a tweet linking to this video by celebrity user Alyssa Milano[2]. (b) A video with a few dozen Twitter mentions and nearly 2×10^5 views in its first 15 days. Note that the video popularity continues to rise even after the tweet volume has tapered off, illustrating the prediction power of "early tweets".

over their first 90 days with a 0.70 Precision@100, using only Twitter information from the first 15 days. Among many critical insights, we note that Twitter features describing user interactions are more predictive than features describing their social network derived influence; furthermore, aggregating these user interaction levels by standard deviation is most predictive of a video's popularity in both EARLY and JUMP, indicating that a diverse range of interactions is more important than average or total interactions.

2. DATA PROCESSING

We used three data sources: the YouTube video history providing total viewcounts received over time, a subset of tweets over a three-month period, and the Twitter user graph from the same period.

We obtain **YouTube viewcount history** from a video's webpage, when it's made available by the video owner. This history contains the number of views a video has received since its upload, in 100 evenly spaced time intervals, with daily viewcount obtained by temporal interpolation[3]. We use a collection of 467-million tweets of from 2009 [17]; this sample is estimated to contain about 20-30% of all posts published on Twitter during the 6-month period, and was authored by about 20 million users. Each tweet is represented as three fields: *author*, e.g., http://twitter.com/annieng ; *timestamp*, e.g., 2009-06-07 02:07:42; *tweet content*, e.g. "in LA now". We use a snapshot of the **Twitter user graph** from 2009 [11], with each user as a node, and each following relationship as a directed edge from a user to one of his/her followers.

We extracted URLs from all tweets and resolved shortened URLs, retaining references to YouTube videos. We found $1,624,274$ Twitter users tweeted YouTube video links at least once and that there are $2,350,881$ unique videos, of which $1,549,532$ (65.9%) are still online as of Oct 2013. Within this subset, $1,067,895$ (68.9%), have their viewcount history publicly available. We call the subset of tweets containing videos with available history *video tweets*. We match the tweets and the Twitter user graph dataset in order to extract the user graph information of the observed video tweets. About 80% of the users can be identified by matching the username directly, for 20% of the users we did not find a match. A tweet is dropped from the collection if its author cannot be identified.

We process tweets to extract tags and user interactions. We rely on text processing for this since our historical tweets collection does not contain the full Twitter API feed (where many interactions are already encoded). We extract **hashtags** and **mentions** by finding words prefixed with # and @ symbols. We also extract non-broadcasting tweets (**nbcTweet**) – when a tweet starts by mentioning a user, it is treated as a targeted interaction between the author the user being mentioned, and the followers of the author will not see this tweet in their timeline. We extract variations of retweet (**RT**). Symbols for retweeting have evolved since the early days of Twitter [10], and there are still a diverse set of symbols in use in the 2009 data. We extract ten major variants, i.e., *RT, R/T, via, HT, H/T, OH, retweet, ret*, plus two variants of the "recycle" symbol (♲).

[2] http://en.wikipedia.org/wiki/Alyssa_Milano

[3]The temporal granularity of viewcount history is about 12 days (with 1,100+ days between August 1, 2009 and data collection in Oct 2012) and varies depending on the video upload date.

3. PREDICTING VIDEO VIEWS

We begin describing the features and prediction tasks by defining units of data over time. We take a sliding time window of length τ as a unit for feature extraction and viewcount prediction. We define time index $t \in \{0, 1, \ldots, T\}$ with increments of τ. With slight overload of notation, index 0 may represent the real-world interval $[0, \tau)$, or time point $t = 0$, the interpretation should be clear from context. For a YouTube video v, denote its total viewcount (i.e. number of views received since upload) on time t as $c_v(t)$, and the viewcount increment between t and $t+1$ as $\Delta c_v(t)$. We use $U_v(t)$ to denote the set of Twitter users who tweeted video v in time interval t. In this work, the prediction targets are videos tweeted between August and November 2009 with $T = 93$ *days*, and $\tau = 15$ *days* due to viewcount data granularity. Tweets published before August are used to compute features.

3.1 Features from YouTube and Twitter

We extract one set of YouTube features and four sets of Twitter features for prediction. There are two general types of aggregated Twitter features we investigate: those directly involving TWEETS on the video and those involving the users who have tweeted on a video. Among the latter type, we further make distinctions among ACTIVE, PASSIVE, and social GRAPH features of those users.

YT-VIEWS is the number views a video v receives in two time intervals before time t on which we are making a prediction, i.e. $[\Delta c_v(t-2), \Delta c_v(t-1)]$. Historical viewcount is shown to highly correlate with future viewcounts [15], and using more than one historical interval is shown to further improve prediction [13]. We chose two intervals via cross-validation. This feature is comparable to those used in prior work [15, 13], and used as the baseline for predicting JUMP .

TWEET includes five counting metrics that describe the properties of video tweets about video v in interval t: $T.tweet(v,t)$ is the number of video tweets; $T.hashtag(v,t)$ is the number of times a hashtag is used; $T.mention(v,t)$ and $T.nbcTweet(v,t)$ are the numbers of broadcasting and non-broadcasting mentions, respec-

Table 1: Six summary statistics for user features. u: a Twitter user; U: a set of Twitter users; $f(u)$: a user feature.

Name	Description
sum-log	$\sum_{u \in U} \log(f(u)+1)$
log-sum	$\log(\sum_{u \in U}(f(u))+1)$
mean-log	$\frac{1}{\lvert U \rvert}(\sum_{u \in U} \log(f(u)+1))$
log-mean	$\log((\frac{1}{\lvert U \rvert}(\sum_{u \in U} f(u))+1))$
std-log	$std(\{\log(f(u)+1)\}_{u \in U})$
log-std	$\log(std(\{f(u)\}_{u \in U})+1)$

tively; $T.RT(v,t)$ is the number of retweets for each of the 10 variants. Intuitively, videos are likely to obtain more views when they are tweeted or are part of twitter interactions (via hashtags, mentions, or retweets).

GRAPH consists of three features computed from the Twitter user graph. For a Twitter user u, $G.outdegree(u)$ is the number of followers he/she has. $G.pagerank(u)$ contains the pagerank score of a user, robust measure of a user's influence for hashtag adoption [18]. $G.hubauthority(u)$ contain a pair of hub and authority scores [9]. A Twitter user has high hub score if her followees have high authority scores; she has high authority scores if her followers have high hub scores.

ACTIVE consists of five types of behavior features for Twitter user u up to time $t-1$. They are denoted as $A.tweet(u)$, $A.hashtag(u)$, $A.mention(u)$, $A.nbcTweet(u)$, $A.RT(u)$ to capture the users' tweet volume, use of hashtags, sending of broadcasting and non-broadcasting mentions and retweet behaviors, respectively. Moreover, each feature type is represented with a number of metric variants. $A.tweet(u)$ includes four variants: the total and per-day average of all and unique tweets. Each interaction feature (hashtag, nbcTweet, mention, RT) includes four variants: the total number of interactions, its average per day; the number of unique user-to-user interactions, and its average per day.

PASSIVE consists of three behavior features that Twitter user u receives from other users up to time $t-1$. Denoted as $P.nbcTweet(u)$, $P.mention(u)$ and $P.RT(u)$, they represent interactions where user u is mentioned in broadcasting or non-broadcasting tweets, and retweeted, respectively. Each of these interactions features has the same metric variants as those for ACTIVE features.

Both active and passive features have been recognized as capturing user influence within Twitter [18], here we use them to infer YouTube popularity. Note that features from tweets are computed from dataset inception, i.e. 2009-05-31 [17]. Furthermore, we aggregate the GRAPH, ACTIVE and PASSIVE features from the set of users tweeting about the same video into six summary statistics. These statistics incorporate three kinds of aggregation – sum, mean, standard deviation (std); over two scaling variants – log-aggregate or aggregate-log. The method to compute them can be found in Table 1. This is to account for the variable number of users tweeting each video, and being able to generalize across users. An overview of all features is in Table 2. Note that the feature dimensionality includes summary statistics for all user features and all metric variants, e.g., A.RT (and P.RT) has 10 RT literals $\times$ 4 metric variants $\times$ 6 summary statistics, totaling 240 dimensions.

3.2 Two Prediction Tasks

JUMP captures cases when a video gains a large number of views in a relatively short period of time. For video v, denote the total viewcount gained between time 0 and T as $\Delta c_v(0,T)$; we compute the *normalized* gain during interval t as $r_v(t) = \Delta c_v(t)/\Delta c_v(0,T)$. For a video v, a *jump* is deemed to have occurred in time t if $\Delta c_v(t)$ has more than 50 views; $\Delta c_v(t-1)$ is not more than $\Delta c_v(0,T)/T$, the average gain over interval $[0,T)$; and $r_v(t) \geq \alpha$ with predefined threshold α. Defining jumps using such normalized increments allows us to compare videos that undergo popularity changes at very different levels, e.g., from hundreds to millions of views.

Table 2: Youtube and Twitter feature summary (Sec 3.1)

Feature group	Feature name	# of dimensions
YT-VIEWS	viewcount	2
TWEET	T.tweet	1
	T.hashtag	1
	T.mention	1
	T.nbcTweet	1
	T.RT	10
GRAPH	G.outdegree	6
	G.pagerank	6
	G.hubauthority	12
ACTIVE	A.tweet	24
	A.hashtag	24
	A.mention	24
	A.nbcTweet	24
	A.RT	240
PASSIVE	P.mention	24
	P.nbcTweet	24
	P.RT	240

EARLY captures cases when a video receives a significant number of views just after being uploaded. We take the most popular videos as prediction targets, i.e., those having the most viewcount in their first $\hat{\tau}$ days, denoted as $\Delta c_v(0,\hat{\tau}) > \beta$, with a pre-defined threshold β. Prior approaches that rely on historical viewcount [13, 15] cannot be used to analyze such phenomenon.

Binary classifiers are trained with linear support vector machines for each task. We use $\alpha = 0.5$, and $\beta = 10^4$, high thresholds that yield popular videos which are likely to be mentioned in tweets. The empirical YouTube viewcount distributions are long-tailed, and do not show a clear separation around these (or any other) values. To this end, thresholds separating the top few percent videos are equally valid conceptually. Fig 1 (a) and (b) contain example videos in the respective JUMP and EARLY classes, which are ranked highly by our algorithm.

4. EXPERIMENTS

We evaluate JUMP and EARLY prediction with the following settings. For JUMP, each time interval t with at least one tweet about video v becomes an instance, there are 6,156 positive JUMP instances with a random guess prior of 1.2%. The five feature groups are specified in Section 3, and ALL is a result of concatenating features from all available groups. For EARLY, each video (in its first $\hat{\tau}$ days) becomes an instance. Results are reported on 29,998 videos, out of which about 5.3%, or 1,591 are positive examples. We report average precision (AP) [12] and Precision@100, with the average and the 95% confidence interval over 5-fold cross-validation with stratified sampling (preserving the random guess probability).

Table 3 summarizes the performance of different features for JUMP . We can see that among the four types of Twitter features, each improves upon results using viewcount history only. The best predictor doubles the AP and nearly quadruples the Precision@100 vs. viewcounts, and with Precision@100 at 0.46, almost half of the top-ranked videos actually contain a JUMP . In addition, differentiating users and taking into account user history (with GRAPH, ACTIVE and PASSIVE) perform significantly better than viewcounts or tweet properties.

Table 4 summarizes the prediction performance of EARLY , with the same feature groups as JUMP except that YT-VIEWS is unavailable for newly uploaded videos. The prediction is done for the first 15 days, and then $\hat{\tau} = 30, 60$ and 90 days. Longer term predictions are done with ACTIVE features, because (1) it is best-performing feature group – only 0.05 away from ALL in prec@100; (2) these features only need user history for the video tweets, and do not

Table 3: Performance for JUMP prediction. See Sec 4.

Features	Avg Prec	Prec@100
Random	0.012	0.012
YT-VIEWS	0.056 ± 0.006	0.125 ± 0.028
YT-VIEWS+TWEET	0.058 ± 0.002	0.204 ± 0.041
YT-VIEWS+GRAPH	0.097 ± 0.007	0.406 ± 0.023
YT-VIEWS+ACTIVE	0.105 ± 0.003	0.432 ± 0.057
YT-VIEWS+PASSIVE	0.104 ± 0.005	0.444 ± 0.044
ALL	0.113 ± 0.008	0.460 ± 0.053

Table 4: Performance for EARLY prediction. See Sec 4.

$\hat{\tau}$	Feature	Avg Prec	Prec@100
all	Random	0.053	0.053
15-d	TWEET	0.248 ± 0.142	0.450 ± 0.229
15-d	GRAPH	0.382 ± 0.030	0.646 ± 0.044
15-d	ACTIVE	0.441 ± 0.027	0.702 ± 0.058
15-d	PASSIVE	0.375 ± 0.055	0.656 ± 0.088
15-d	ALL	0.463 ± 0.029	0.750 ± 0.045
30-d	ACTIVE	0.421 ± 0.023	0.686 ± 0.060
60-d	ACTIVE	0.435 ± 0.024	0.722 ± 0.018
90-d	ACTIVE	0.424 ± 0.026	0.720 ± 0.043

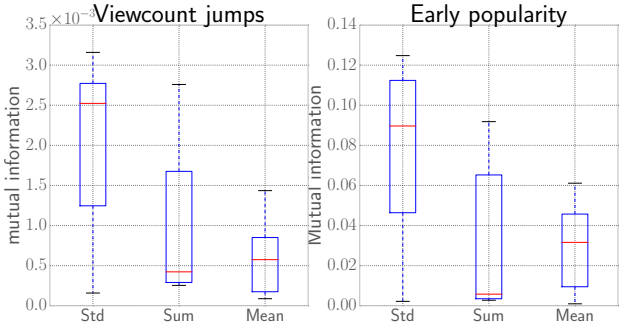

Figure 2: Box plots of mutual information grouped by feature aggregates. The most informative features are generated by std aggregation for both JUMP and EARLY .

need Twitter GRAPH or PASSIVE interactions, which are expensive to obtain. It is encouraging to see that the top 5% most popular videos can be predicted with an AP of more than 0.40, and there are 70+ correct entries in the top 100. Moreover, this accuracy is maintained from 15 to 90 days.

We perform an analysis on the informativeness of individual feature dimensions described in Sec 3. We compute the mutual information between the target class $Y \in 0, 1$ and each feature X, as $I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$. The larger the mutual information, the more informative a feature is towards predicting the target. Fig 2 contains box plots of such mutual information on user features (GRAPH, ACTIVE and PASSIVE) grouped by the three feature aggregation methods: std, sum, and mean. We can see that the majority of most informative features (e.g. top 1/6 above the median for std) are std-features, with sum features moderately informative and mean features the least informative. A high standard deviation for a feature implies that there is broad interest across a spectrum of users with a range of activity and interaction levels. This concurs with a recent observation on hyperlinks in Twitter [3] – that having a diverse set of users (std) mentioning an item is helpful for improving its popularity.

5. CONCLUSION

In this work, we showed user and content information from Twitter can be effectively used to predict content popularity on YouTube, as shown in two challenging tasks – predicting viewcount JUMP and EARLY popularity – both with significant and quantifiable performance gains. These results are encouraging in that they show viewcounts are predictable from one external source alone, without taking into account the influence from many other social and traditional media sources (Reddit, Tumblr, Pinterest, Facebook, ...). Furthermore, the results show the predictive power of different features and aggregation methods and reveal that having a diverse range of users and associated tweeting activities is more informative than the total or average volume of activity of these users and also more informative than other features – including those based on social network derived measures of influence. This work raises many interesting avenues of future work, such as leveraging diffusion patterns on Twitter to further improve popularity prediction and quantifying the roles of "influencers" vs. "grassroots" users.

Acknowledgments NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.This work was supported in part by the US Air Force Research Laboratory, under agreement number FA2386-12-1-4041 and Australian Research Council under project DP140102185.

6. REFERENCES

[1] A. Abisheva, V. R. K. Garimella, D. Garcia, and I. Weber. Who watches (and shares) what on youtube? and when?: using twitter to understand youtube viewership. In *WSDM*, pages 593–602, 2014.

[2] T. Althoff, D. Borth, J. Hees, and A. Dengel. Analysis and forecasting of trending topics in online media streams. In *ACM Multimedia*, 2013.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. WSDM '11, pages 65–74, 2011.

[4] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. WWW '12.

[5] 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. IMC '07, pages 1–14, 2007.

[6] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A first step towards understanding popularity in youtube. In *INFOCOM Workshops*, pages 1–6, 2010.

[7] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238, 2008.

[8] M. Gladwell. *The tipping point: How little things can make a big difference*. Hachette Digital, Inc., 2006.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[10] F. Kooti, H. Yang, M. Cha, K. Gummadi, and W. A. Mason. The emergence of conventions in online social networks. In *AAAI ICWSM*, 2012.

[11] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.

[12] C. D. Manning, P. Raghavan, and H. Schütze. chapter Evaluation in information retrieval. Cambridge University Press, 2008.

[13] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. WSDM '13.

[14] S. D. Roy, T. Mei, W. Zeng, and S. Li. Socialtransfer: cross-domain transfer learning from social streams for media applications. In *ACM Multimedia*, pages 649–658. ACM, 2012.

[15] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.

[16] Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang. Propagation-based social-aware replication for social video contents. MM '12, pages 29–38, 2012.

[17] J. Yang and J. Leskovec. Patterns of temporal variation in online media. WSDM, pages 177–186, 2011.

[18] L. Yang, T. Sun, M. Zhang, and Q. Mei. We know what @you #tag: does the dual role affect hashtag adoption? WWW '12, pages 261–270, 2012.

[19] Y. Zhang and M. Pennacchiotti. Predicting purchase behaviors from social media. In *WWW '13*, pages 1521–1532, 2013.