Does Content Determine Information Popularity in Social Media?

A Case Study of YouTube Videos' Content and their Popularity

Flavio Figueiredo[1], **Jussara M. Almeida**[1], **Fabrício Benevenuto**[1]
[1]Department of Computer Science, UFMG, Brazil
{flaviov,jussara,fabricio}@dcc.ufmg.br

Krishna P. Gummadi[2]
[2]MPI-SWS, Germany
gummadi@mpi-sws.org

ABSTRACT

We here investigate *what drives the popularity of information on social media platforms*. Focusing on YouTube, we seek to understand the extent to which content by itself determines a video's popularity. Using mechanical turk as experimental platform, we asked users to evaluate pairs of videos, and compared users' relative perception of the videos' content against their relative popularity reported by YouTube. We found that in most evaluations users could not reach consensus on which video had better content as their perceptions tend to be very subjective. Nevertheless, when consensus was reached, the video with preferred content almost always achieved greater popularity on YouTube, highlighting the importance of content in driving information popularity on social media.

Author Keywords

Content popularity; social media; user study

ACM Classification Keywords

H.5.4 Hypertext/Hypermedia: User issues.

INTRODUCTION

What drives the popularity of information in social media? Recently, this question has attracted a lot of research attention as social media sites become increasingly popular. An unresolved part of this question is about the relative roles of two primary forces that drive the popularity of a piece of information: (i) its content, i.e., the interestingness, topicality, or quality of the information *as perceived by users*, and (ii) its dissemination mechanisms, such as propagation by word-of-mouth, blogs or mass media channels. It stands to reason that both factors matter, but the extent to which they impact the popularity of a piece of information remains an open issue.

Many previous studies on how information becomes popular in social media sites focused on dissemination related factors (e.g, social influence, mechanisms that expose content to users, time of upload) [2, 4, 7, 9, 10], ignoring the role of the content itself. Other efforts, instead, analyzed social media content focusing on data mining tasks such as popularity prediction [11] and video classification [5], analyzing popularity differences in content duplicates [2], and exploring content importance as parameter of popularity evolution models [8]. In this paper we take a different and complementary approach, focusing on understanding the extent to which content matters for popularity of videos on YouTube.

Our methodology attempts to assess *users' relative perceptions* of the contents of pairs of videos through user surveys conducted over Amazon mechanical turk. Users in our experiments are exposed only to the video content, and are not subjected to other factors (inherent to the YouTube site) that may impact their perceptions of content (e.g., user comments, social links, appearance of content in external sites). Specifically, we present to users pairs of videos from the same major topic and uploaded around the same date, and ask them to choose which one: (1) *they enjoyed more*, (2) *they would be more willing to share with friends*, and (3) *they predicted would become more popular on YouTube*. These questions target the user's individual perception of content interestingness and of the interests of her social circle (and thus the chance of the content spreading through it), as well as the user's expectations on a global scale. Our goals are to assess, for each of these questions, whether users reach consensus, and, when there is consensus, whether user perceptions match the relative popularity achieved by the videos on YouTube.

We find that users could not reach consensus in many evaluations, even when the popularity (on YouTube) of the evaluated videos differs by orders of magnitude. The lack of consensus is more striking for sharing and liking choices. It also depends on the video topic. This suggests that users' perceptions about content are quite subjective and that content may not be the most important factor that drives popularity in many cases. However, whenever participants reached consensus, their choices mostly match the video with largest popularity on YouTube, suggesting that, in these cases, content has a significant impact and predictive power on video popularity.

The goals of our study complement previous work. In particular, Salganik *et al.* [9] also relied on a user study to understand popularity dynamics. However, they focused on the impact of social influence on popularity, whereas we focus on the role of content and rely on users to evaluate the content in a setup that is isolated (to the extent possible) from dissemination mechanisms that might influence popularity. To our knowledge, the human perceptions of content and how they correlate to popularity in a social media site have not been

Acknowledgments: Research supported by InWeb - Institute of Science and Technology for Web Research and by individual grants from CNPq, Capes and Fapemig.

CHI'14, April 26–May 1, 2014, Toronto, Canada.

http://dx.doi.org/10.1145/2556288.2557285

previously analyzed. Thus, this work is a first step towards assessing the role of content in determining popularity of a piece of information, and our proposed experimental methodology, discussed next, is a key contribution towards that goal.

EXPERIMENTAL METHODOLOGY

Our study is guided by two questions: *[Q1] Given a pair of videos with similar topic, can users reach consensus on their relative popularity? [Q2] When users reach consensus, does the preferred video match the most popular one on YouTube?*

Question Q1 is focused on the collective notion of popularity reported by the users in our experiment, who are subject only to the content itself. This notion relates to whether a user likes and/or would be willing to share a video more than the other, and also whether a user, despite personal tastes, believes one video would become more popular than the other. Question Q2 aims at comparing this notion with the popularity achieved by the videos on YouTube, measured by the total number of views at the time we collected the videos, which can be affected by various factors, other than content alone.

Datasets

In order to identify videos with similar topic, we used Freebase[1], a collaborative semantic knowledge database covering over 30 million topics. We crawled YouTube for videos indexed under the same Freebase topic on its API, focusing on 2 topics that span diverse user interests and are neither too specific nor too broad: *major league baseball* and *music videos*.

We crawled YouTube on August 2013, focusing on videos that were uploaded from the US on April 2012. By studying videos of similar topic and uploaded from the same country, we factored out the notion of popularity due to latent social, cultural and psychological issues (e.g., soccer is less popular than baseball in the US). By focusing on videos uploaded around the same time, we factored out popularity variations due to first mover advantage [2] and upload date [7]. We also downloaded only videos considered safe by YouTube's safe search, limiting the chance of users finding a video offensive, and that could be embedded in external sites. The latter is to allow user evaluations to be done outside YouTube, and thus be unaffected by the other pieces of information (e.g., number of views, user comments) available on a video's page.

We also defined three ranges of YouTube popularity values: *low*, with number of views between 10 and 100; *medium*, from 1,000 to 10,000 views; and *high*, from 100,000 to 1,000,000 views. For each topic, we chose 3 videos of each range, each video having from 4 to 6 minutes of duration.

Human Intelligence Tasks

We ran our user experiments on Amazon mechanical turk (MT), recruiting as participants of our task only master workers (i.e., the best workers as ranked by MT) based on the US.

The first step was to build, for each topic, all 36 pairings for the 9 selected videos. These pairs were assigned to 9 *folds*, so as to have only unique videos in each fold, and deployed on a web application built by us, called YouRank. YouRank assigns one fold (4 video pairs) to each user following a round-robin schedule. It shows to the user only the embedded video streams, hiding any other video metadata kept by YouTube.

[1] `http://www.freebase.com`

S1. How old are you?
S2. Are you a male or a female?
S3. How often do you watch a video on YouTube?
S4. How often do you share YouTube videos with friends or colleagues?
S5. How often do you share any kind of online content with friends or colleagues?

(a) Demographic Survey

E1. Which video did you enjoy watching more?
E2. Which video you would be most willing to share with a friend or group of friends?
E3. Which video do you predict will be more popular on YouTube?
E4. Did you already know (watched previously) one of these videos?

(b) Video Evaluation Form

Figure 1: YouRank Forms

After logging in, each user was first asked to answer the demographic survey in Figure 1a. For questions S3 to S5, the possible answers were: 1) never; 2) rarely (few times a year); 3) occasionally (few times a month); 4) often (few times a week); 5) very often (once or more daily).

Next, the user was asked to watch 4 video pairs, and, for each pair, answer the form shown in Figure 1b picking one of four options: a) Video 1 (left); b) Video 2 (right); c) Both; d) Neither. Thus, two neutral options (c-d) were available in case the user could not choose one video. The user could also provide feedback in free text form for each pair. We asked users not to visit the video page on YouTube, and to indicate whether they had watched any of the videos in the past. To avoid bias due to user fatigue, the pairs of a fold were randomized whenever a new user was assigned to the fold. A user was expected to take roughly 45 minutes to evaluate 4 pairs of videos, each one from 4 to 6 minute long. Thus, upon task completion, each user was paid 4.50 US dollars, which is consistent with the MT suggested hourly rate of 6 US dollars. In practice, the users took on average 44.8 minutes to complete the task, although some evaluations were disregarded (see below).

Evaluation Metrics

To tackle question Q1, we measured consensus for each video pair using the Fleiss' Kappa (κ) score of agreement [6]. This score varies from -1 to 1, while values above .4 are interpreted as fair to good agreements, and above 0.75 as very good agreements [6]. We determined that consensus was reached if the null hypothesis of negative or no agreement ($\kappa \leq 0$) could be rejected. The same score was achieved regardless of whether the neutral responses c, d, were included. Thus, we computed it over all responses. We also applied Bonferroni correction to rule out significance due to random chance [3].

To answer Q2, we focused only on pairs of videos for which consensus was reached and computed the fraction $\hat{p}$ of those pairs for which the preferred video matches the one with larger popularity on YouTube. We then used an exact binominal sign test based on Clopper-Pearson confidence interval [6], which is suitable to small samples (as our case), to test whether $\hat{p}$ is above random chance (i.e., $\hat{p} > 0.5$).

RESULTS

We now discuss the results of two rounds of MT experiments, one for each chosen topic. We ran each round until 72 users

	Major League Baseball			Music Videos		
	S3	S4	S5	S3	S4	S5
Never	0%	4%	1%	0%	1%	0%
Rarely	0%	18%	12.5%	0%	22%	13%
Occasionally	8%	39%	28%	21%	45%	32%
Often	48%	29%	37.5%	39.5%	28%	37%
Very Often	44%	10%	21%	39.5%	4%	18%

Table 1: Answers to S3, S4 and S5 in the demographic survey (Fig. 1a).

p-value		Major League Baseball			Music Videos		
		E1	E2	E3	E1	E2	E3
.05	$\%(\kappa > 0)$	25%	13%	52%	11%	2.7%	13%
	Avg. κ	.68	.64	.74	.63	.53	.65
.01	$\%(\kappa > 0)$	19%	8%	41%	8%	2.7%	11%
	Avg. κ	.75	.76	.78	.63	.53	.69
.001	$\%(\kappa > 0)$	16%	5%	36%	5%	2.7%	8%
	Avg. κ	.79	.76	.83	.65	.62	.86

Table 2: Percentage of video pairs that rejected the Fleiss' Kappa null hypothesis of $\kappa \leq 0$ and the average level of agreement κ for those cases. The columns correspond to the questions in Figure 1b.

had finished their tasks. However, in both rounds, some users refused the task after logging in. Also, we disregarded evaluations in which the users reported they: (1) were unable to watch a video, 2 cases, or (2) had watched at least one of the videos before (5% and 8% of the cases for the major league baseball and music video experiments, respectively). The latter was done to avoid a bias due to prior knowledge. We were then left with 6 to 10 evaluations per pair (8 on average). We summarize the answers to the demographic survey next, and afterwards we discuss the results for our two driving questions based on the answers to the form in Figure 1b.

Demographic Survey

On MT, we required all users to be from the US. Moreover, 53% and 42% of the them (of 72 per round) were males in the baseball and music experiments, respectively, whereas in both rounds, most (57%) had from 20 to 45 years of age, and only 5% were under 20 years old. The answers of users regarding their viewing and sharing habits (S3-S5 in Figure 1a) are summarized in Table 1. Note that participants of both rounds of experiments are avid YouTube viewers: they watch videos at least occasionally, and most of them do it often (39.5-48%) or very often (39.5-44%). Also, most users share YouTube content occasionally (39-45%) or often (28-29%), whereas only 22% of the users in both rounds share YouTube videos only rarely or never. Finally, in both rounds, users tend to share online content in general more often, as expected.

Q1: Can users reach consensus?

Table 2 shows, for both video topics, the percentage of pairs in which users reached consensus, (i.e., pairs for which the null hypothesis of $\kappa \leq 0$ can be rejected). It also shows the average κ scores for *those pairs that rejected the null hypothesis*. Results are shown separately for each question in Figure 1b, and for different significance levels (p-values).

In general, for any considered significance level, and for both topics, the fraction of pairs that passed the test tends to be very small (with few exceptions). The fraction is larger when users were asked which video they predicted would be more popular (E3). Thus, user agreement is easier when it comes

	Baseball	Music
E1	100%**	75%
E2	100%*	100%
E3	84%**	100%*

Table 3: Percentage of pairs (with consensus) that match YouTube popularity. * (**) indicate above random chance p-val=.05 (.01)

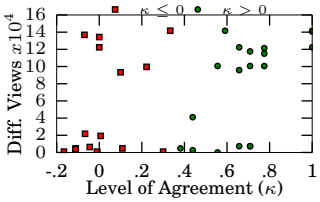

Figure 2: κ vs popularity gap

to the collective knowledge of popularity. However, this happened in at most 52% of the pairs (p-value=0.05). Users agreed less often when asked which video they enjoyed the most (E1), reflecting a natural heterogeneity of user interests. The consensus was even rarer for sharing patterns (E2), possibly reflecting heterogeneity in terms of social activities and users' perceptions of the interests of their social networks. This is reflected by the number of times users chose a neutral answer. For E2 this occurred from 40% (baseball) to 52% (music) of the evaluations. For both topics and the other questions, less than 20% of the evaluations used a neutral choice. Finally, we show that when consensus was reached, the agreements were on average good (κ>0.4) or very good (κ>0.75).

To illustrate these findings, note the divergence in opinions in the following feedbacks on the same music video:

U1: the girl in the video 2 was stunningly beautiful so i would share that one

U2: Video 2 was sad and dark and I didn't like the girl's voice.

U3: I secretly like Evanescence but I would never let my friends know.

U4: The video on the left was much better music for my tastes

The divergence in opinions reveals the more egocentric notions of liking/sharing content. U2 dislikes the video because it was sad, and U4 likes it because of her personal musical taste. U1 would share the video because of the girl in it, while U3 would not share it because she secretly likes the band.

Recall that users evaluated pairs of videos that covered a wide range of popularity values on YouTube. Thus, one may ask whether users could reach consensus more often for pairs of videos with a larger gap in their relative popularity. Surprisingly, we found no strong trend towards that, as illustrated in Figure 2 for E3 in the major league baseball experiments (p-val $= .05$). Very low κ values were obtained even for videos which differ in popularity by hundreds of thousands views.

Table 2 also shows that the agreements are more common for major league baseball videos than for music videos. While this may be related to a more diverse range of personal interests for music videos (e.g., U4's feedback), it may also relate to promotional campaigns for this kind of content. Such campaigns may cause videos to be popular for a short while, regardless of user tastes. As an example, we could note case of music videos that experienced a burst in popularity, possibly caused by promotion (professional or amateur such as in a blog) but was unable to remain popular over time.

Nevertheless, there are many cases of lack of consensus. One example is a pair of baseball videos, where one of them has over 100 times more views than the other, and remains more popular throughout the monitored period. Yet, the users of our experiment could not reach consensus in none of the questions. Further investigating these videos, we noted that the

most popular video has a watermark that affected user opinions in our experiment, as indicated by the feedback: "The watermark on video 2 ruins it" . This suggests that other latent factors may play a role on driving the popularity of social media.

Q2: Does consensus match the popularity on YouTube?
Considering only pairs for which consensus was reached (p-value=0.05), Table 3 shows the fraction of pairs in which the video preferred by MT users has higher popularity on YouTube. Note that, whenever consensus is reached, user preferences match YouTube's popularity in almost all cases. This result is above random chance (p-value=0.05) in most cases, except when the number of pairs with consensus is too small. Thus, if users can reach consensus on their opinions, the preferred video is likely to become more popular.

DISCUSSION

In the traditional media (e.g., newspapers, TV), dissemination mechanisms are closely tied to the content generators. Content is traditionally generated or selected by professionals on behalf of organizations that have vested interest and ability to promote their content (e.g., ad campaigns). Differently, social media is dominated by content generated by ordinary users, and the key dissemination mechanisms are (i) crowd-endorsements: information "liked" by crowds is promoted in search results and recommendation tools, and (ii) viral propagation over a social network: anyone who finds the information content interesting can "share" it with friends. Thus, the dissemination mechanisms are democratized and only loosely coupled with the content generators. This democratization offers the hope that information popularity would be driven to a larger extent by its content (i.e., how users perceive or like it) than it is in traditional media. In this paper we give the first step towards understanding the extent to which this is true.

To that end, we relied on user evaluations of pairs of YouTube videos of similar topic, factoring out dissemination related factors. We found that users' perception of content is very subjective, as users often could not reach consensus at which video they liked or would share more, or predicted would become more popular. This result indicates the difficulty in determining the role of content in driving popularity, and complements previous observations that users cannot estimate the extent of visibility of their content [1]. However, whenever there was consensus, the preferred video almost always matched the one with higher popularity on YouTube, highlighting the key role played by content in those cases.

From a social media research perspective, this finding emphasizes the need to account for content in studies of popularity. From a media site operator's or viral marketer's perspective, it has implications for popularity prediction. For example, it can be leveraged by marketers or advertisers to compare new videos against old ones with known popularities to define which one has more chance of attracting viewers. It also motivates future studies on how a site operator can design a scalable way for gathering users' feedback to predict which of newly uploaded videos are more likely to become popular.

We note that representativeness is an important but challenging issue in any empirical study, as ours. We here designed an experimental methodology that is as thorough as possible, given our practical constraints. We chose one of the most popular social media sites, YouTube, and recruited only master MT workers, who are known to perform better tasks. Given our budget, we carefully chose the videos of our dataset, covering three popularity levels, with multiple videos per level. To avoid extraneous factors, we only compared videos of the same topic and similar age, and only used evaluations of users who had not seen the video before. To ensure that our sample sizes are not too small to draw conclusions, we applied conservative and exact statistical tests, adequate for them, presenting results for various significance levels. Thus, our setup was designed to yield the most accurate and representative results, within our constraints[2].

However, we acknowledge that it is impossible to generalize our findings without future studies. We hope that this work will encourage future efforts to apply our methodology across various applications and over more content instances.

REFERENCES

1. Bernstein, M. S., Bakshy, E., Burke, M., and Karrer, B. Quantifying the Invisible Audience in Social Networks. In *Proc. CHI* (2013).
2. Borghol, Y., Ardon, S., Carlsson, N., Eager, D., and Mahanti, A. The Untold Story of the Clones: Content-Agnostic Factors that Impact YouTube Video Popularity. In *Proc. KDD* (2012).
3. Bretz, F., Hothorn, T., and Westfall, P. *Multiple Comparisons Using R*, 1 ed. CRC Press, 2010.
4. Figueiredo, F., Benevenuto, F., and Almeida, J. The Tube Over Time: Characterizing Popularity Growth of YouTube Videos. In *Proc. WSDM* (2011).
5. Filippova, K., and Hall, K. B. Improved Video Categorization from Text Metadata and User Comments. In *Proc. SIGIR* (2011).
6. Fleiss, J. L., and Levin, B. *Statistical Methods for Rates and Proportions*, 3 ed. Wiley-Interscience, 2003.
7. Lakkaraju, H., McAuley, J., and Leskovec, J. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *Proc. ICWSM* (2013).
8. Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. Rise and Fall Patterns of Information Diffusion. In *Proc. KDD.* (2012).
9. Salganik, M., Dodds, P., and Watts, D. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science 311*, 5762 (2006), 854–856.
10. Shamma, D. A., Kennedy, L., and Churchill, E. F. Peaks and Persistence: Modeling the Shape of Microblog Conversations. In *Proc. CSCW* (2011).
11. Yano, T., and Smith, N. A. Whats Worthy of Comment? Content and Comment Volume in Political Blogs. In *Proc. ICWSM* (2010).

[2]In favor of reproducibility, our source code and our datasets are available at: `http://github.com/flaviovdf/yourank`