Evaluative Patterns and Incentives in YouTube

David Garcia(✉), Adiya Abisheva, and Frank Schweitzer

ETH Zurich, Weinbergstrasse 56/58, 8092 Zurich, Switzerland
dgarcia@ethz.ch

Abstract. Users of social media are not only producers and consumers of online content, they also evaluate each other's content. Some social media include the possibility to down vote or dislike the content posted by other users, posing the risk that users who receive dislikes might be more likely to become inactive, especially if the disliked content is about a person. We analyzed the data on more than 150,000 YouTube videos to understand how video impact and user incentives can be related to the possibility to dislike user content. We processed images related to videos to identify faces and quantify if evaluating content related to people is connected to disliking patterns. We found that videos with faces on their images tend to have less dislikes if they are posted by male users, but the effect is not present for female users. On the contrary, videos with faces and posted by female users attract more views and likes. Analyzing the probability of users to become inactive, we find that receiving dislikes is associated with users becoming inactive. This pattern is stronger when dislikes are given to videos with faces, showing that negative evaluations about people have a stronger association with user inactivity. Our results show that user evaluations in social media are a multi-faceted phenomenon that requires large-scale quantitative analyses, identifying under which conditions users disencourage other users from being active in social media.

Keywords: Social psychology · Incentives · YouTube

1 Introduction

The rise of the Social Web fundamentally changed the role of Information and Communication Technologies in the flow of information. The early platforms that connected mass media to wider audiences evolved into social media technologies that allow users to find content produced by other users [7]. While mass media focused on producing content of interest for their audience, social media became *participatory media* that encouraged users to produce content of interest for other users. Beyond this shift from audience to content producers, users of social media also became *evaluators* that can positively or negatively asses the content produced by others [24]. This new feature of social media is the source of one of the main challenges for the sustainability of online communities: the possibility of criticism and negative expression in social media can be a negative incentive

G.L. Ciampaglia et al. (Eds.): SocInfo 2017, Part II, LNCS 10540, pp. 301–315, 2017.
DOI: 10.1007/978-3-319-67256-4_24

for user activity. Just few down votes, dislikes, or salty comments can be the cause behind a user abandoning an online community.

The way users interact in online participatory media depends, among other factors, on the design of the online platform they use [4]. This raises various questions about the mechanism design of a website, which are usually aimed to optimize user participation and involvement. A common question is the influence of the dislike button on user participation, which is currently excluded from the design of some of the leading social networking sites, like `Instagram`. Other sites, like `Facebook`, include a wider variety of buttons to express emotional reactions, but leave out the *"dislike"* option from the ways users can evaluate each other's content [6]. The rationale behind this decision is often attributed to the assumption that, when users receive explicit dislikes by other users, their participation decreases and might opt out from an online community. This *risk of the dislike button* leads various social media to only allow positive evaluations through the user interface, leaving any kind of criticism for comments or other kinds of textual interaction.

The risk of the dislike button heavily depends on the purpose and functionality of an online platform. As opposed to the above argument to exclude the dislike button, negative incentives can be critically necessary for social media that generate content aggregates, such as featured lists and front pages. An example of the *necessity of the dislike button* is the "Digg collapse" [37], in which a massive amount of users stopped using `Digg` to start using `Reddit`, following a platform redesign that disabled the option to down vote content [34]. Without the possibility to negatively assess content, the quality of the front page heavily suffered and the main functionality of the site was damaged. The role of explicit negative evaluations in social media is thus a multi-faceted phenomenon that requires a research approach that can distinguish the risks associated with the possibility to negatively evaluate other users' posts.

The above difference between the role of disliking on `Facebook` and `Digg` lies on the nature of the content posted in the online medium. While the content shared in `Facebook` is very close to the identity of the user that posts it (e.g. profile pictures or pictures of family and friends), the content shared in `Digg` is usually composed of web links that might even not be authored by the posting user. To understand the effect of negative evaluations in social media, we differentiate two evaluation scenarios: (i) *subject evaluation* when a person or group of people are salient in the evaluated content, and (ii) *object evaluation* when people are not salient in the content that is evaluated and objects, concepts, or events are at the center of the posted content. As it is not the same to dislike *something* as to dislike *someone*, this differentiation between object and subject evaluations is a potentially pivotal point in the effect of the dislike button.

We hypothesize that the difference between subject and object evaluation contexts affects the role of negative evaluations in user interaction. Users who evaluate are protected by their anonymity and are free to negatively evaluate any content they want, but some social and psychological factors that affect face-to-face evaluations might also appear online. First, when content represents

a person in the subject evaluation context, the content has a closer resemblance to the evaluating user than in the object evaluation context. Implicit self-esteem [15] can generate biases towards positive evaluations when interacting with content that might resemble oneself, such as the name-letter effect [20]. In this case, users should be less likely to provide negative evaluations when people are salient (subject evaluations) than when people are not at the center of the evaluation (object evaluations). Inspired by this, we formulate the *negative subject evaluation avoidance hypothesis*: the tendency to receive dislikes in user content is lower when the content is about a person.

Second, negative evaluations can have stronger effects on user incentives when they happen in the subject evaluation context than in the object evaluation context. This principle is the assumption behind the risk of the dislike button, as allowing the negative evaluation of people poses a risk for user integration, motivation, and future activity levels. We formulate the *negative incentives hypothesis*: the probability of a user becoming inactive grows faster with negative evaluations to the content posted by the user when such content is about a person than when the content is not about a person.

To test the above hypotheses, we need to control for various inter-individual effects, including user popularity and demographic factors. A demographic factor that plays key importance in online behavior is gender [16,22]. In our negative evaluation scenario, gender might have two effects. First, subjective gender biases are linked to the perception of risks in technology [28] and could affect evaluation patterns depending on the gender of the posting user. Second, social forces might have stronger effects for female users [30], strengthening the disincentives associated to negative subject evaluations to content posted by female users. Our analysis takes into account gender in the analysis of online evaluations, assessing whether the role of negative evaluations might differ between male and female users.

Testing the above hypothesis has been a challenging task in previous research due to the difficulty to compare across content and platforms. Comparing negative evaluations across platforms can reveal statistical differences, but a threat to validity lies in the difficulty to single out the effect of context in disliking when various other differences in platform designs are present. Furthermore, due to the risk of the dislike button, not many platforms allow negative subject evaluations (e.g. disliking a `Facebook` profile picture), to avoid the risk of creating negative incentives to user activity. A notable exception is the case of `YouTube`, where negative evaluations are possible through the dislike button and users upload content that brings both object evaluations (e.g. videos about events) and subject evaluations (e.g. videoblogger selfie-like videos). We apply image processing to the images related to `YouTube` videos to operationalize a metric that distinguishes object from subject evaluation, to statistically analyze the link between disliking and user activity. This way, we provide a novel analysis that bridges the research gap that, to date, has prevented the evaluation of the hypotheses explained above. In the following we briefly outline the research background on the topic, followed by a description of the data and methods used to analyze evaluative patterns and incentives in `YouTube`.

1.1 Research Background

Evaluative patterns in social media have been subject of previous research. The appraisal of online content leaves digital traces in the form of up and down votes, likes and dislikes, or numeric star-ratings. Extended research has analyzed ratings of products in reviews communities like `Amazon` [23], often related to recommender systems and sentiment analysis [36]. Beyond products, previous works analyze the relationship between up an down votes in `Reddit` [24], finding a scaling pattern that also appears in other media like `YouTube` and `Imgur` [1]. Collective evaluations are useful to understand the social factors of spreading misinformation [10] and to analyze natural experiments about the factors that influence the success of content [21]. The nature and volume of user evaluations and attention has been found to depend on user gender, from popularity levels in `Twitter` [22,26] to variability in worker ratings of gig economy platforms [16]. Certain user actions are strongly correlated with negative evaluations and have been shown useful to analyze human behavior. Edit conflicts in `Wikipedia` show the burstiness and memory of disagreement [38], and the creation of negative social links shows the existence of structural balance patterns that reduce cognitive dissonance [33].

User incentives and churn in social media have been subject of extensive research. From individual decisions to leave online communities [19] to models that aggregate such behavior at the level of complete websites [12,29]. The decision of users to become inactive in an online community is a multifaceted choice that can reflect nonlinear behavior. For example, the tendency of `Twitter` users to become inactive shows a nonlinear relationship to their amount of followers, such that more followers not always means lower chances to become inactive [11]. Incentives can explain other user decisions beyond churn, for example when psychological biases appear in the creation of social connections [20] and in the evaluation of online content [14], or when economic incentives explain the sharing of links to malware [17].

Research on YouTube has shed light on various aspects of human behavior. The analysis of viewing patterns shows how video impact can be predicted [32] as well as the relationship of video popularity and demographic factors visible in other social media [2]. The temporal information provided by `YouTube` has been analyzed to identify the classes of collective responses of a society [9]. The large size of `YouTube` data allows further research, leading to the identification of a new collective response class not observed before [31]. The data on likes and dislikes of `YouTube` has been applied to analyze general patterns of polarization linked to the filter bubble [2], and allow the understanding of polarization in various contexts, from political campaigns [13] to anorexia-related content [27]. `YouTube` data has been a good alternative to `Twitter` and `Wikipedia` data, alleviating the model organism bias suffered by research on social media [35] and posing an alternative data source to further validate the findings of research in Computational Social Science and Social Informatics.

2 Materials and Methods

2.1 Data on YouTube Channels

As part of a larger analysis of YouTube data [2], we extracted detailed information on a set of YouTube channels. Starting from a large sample of random channels, we identified channels owned by individual users through the data provided by the YouTube API in 2013. These channels could be identified thanks to various fields related to the profile of the owner, such as age and gender. After applying this filter, we count with a sample of 1,556 user channels that were active in 2013, from which we can identify their gender as self-reported in earlier versions of the YouTube platform. In 2016, we performed a retrieval of all publicly available videos on the channel of each user in the dataset. This way we gathered more than 150,000 videos, including their count of views, likes, and dislikes. A descriptive summary of the dataset is presented on Table 1 and further descriptive statistics are reported in the Appendix.

Each YouTube video has an associated image that is used as a thumbnail to summarize the content of the video. We applied face recognition through the face++ API[1] to identify which videos contain a face and which ones do not. The face++ API is a tool that has been shown useful to detect faces in previous research [18,26] and is accurate enough [39] to have a valid approximation to the measurement of whether at least one person is salient in a video. We use the output of face++ to operationalize a variable that captures subject versus object evaluations. Subject evaluations are those directed to content where a person is salient ($Face = 1$), while object evaluations are given to content where people can be present but are not salient enough to be detectable in the image summarizing the video ($Face = 0$). Note that subject evaluations do not need to be evaluations directed to the user who posted the video, they are evaluations to videos in which at least one person is relevant, as opposed to videos not centered around people.

Table 1. Dataset summary. Total sample size of users and videos, counts of views, likes, and dislikes, and means and medians over the set of videos.

Users	1,556	Female	377	% Female	24.2%
Videos	157,661	With Face	48,366	% With Face	30.68%
Views	67,974,981,442	Mean	431,146.5	Median	30,129
Likes	666,168,168	Mean	4,225.3	Median	385
Dislikes	33,496,449	Mean	212.5	Median	24

[1] https://www.faceplusplus.com/.

2.2 Video Impact Models

To understand the interplay between video impact, evaluation context, and the gender of YouTube users, we apply regression models of the impact that videos have in terms of three variables: views, likes, and dislikes. More precisely, we define regression models for three dependent variables measured over each video: (i) the log-transformed amount of views of the video $log(views)$, (ii) the logarithm of the ratio of likes per view $log(L_R) = log(likes/views)$, and (iii) the logarithm of the ratio of dislikes per view $log(D_R) = log(dislikes/views)$. Log-transformations are applied to reduce skewness, as explained more in detail in the Appendix.

We analyze the views of a video through a mixed-effects regression model [5]:

$$\begin{aligned} log(views) &= a_v + b_v \cdot Face + c_v \cdot Female + d_v \cdot Face \cdot Female \\ &+ f_v \cdot log(D_R) + g_v \cdot log(L_R) + Z_v * u + \epsilon_v \end{aligned} \quad (1)$$

where $Face = 1$ if a face was detected on the image of the video and 0 otherwise, $Female = 1$ if the user that posted the video is female and 0 if male, and u is a categorical variable that identifies each user. The fixed-effects parameter a_v measures the intercept of the model, while b_v measures the increase in views that can be attributed to subject evaluation, c_v to the gender of the posting user, and d_v to the statistical interaction between the $Face$ and $Female$ variables, i.e. the additional effect of subject evaluation for videos posted by female users. The fixed effect terms $f_v \cdot log(D_R)$ and $g_v \cdot log(L_R)$ are statistical controls to remove possible confounds with the likes and dislikes ratios. The vector Z_v contains the random effects of the model as an intercept per user, to correct for any inter-individual differences that can explain views. This way we solve a possible Simpson's paradox effect stemming from different popularity and activity levels of the users. The term ϵ_v is the residuals of the model, which are assumed to be normally distributed with zero mean and no relevant correlations to other terms of the model.

In a similar fashion, we model the logarithm of the likes ratio:

$$\begin{aligned} log(L_R) &= a_l + b_l \cdot Face + c_l \cdot Female + d_l \cdot Face \cdot Female \\ &+ f_l \cdot log(views) + g_l \cdot log(D_R) + Z_l * u + \epsilon_l \end{aligned} \quad (2)$$

and of the dislikes ratio:

$$\begin{aligned} log(D_R) &= a_d + b_d \cdot Face + c_d \cdot Female + d_d \cdot Face \cdot Female \\ &+ f_d \cdot log(views) + g_d \cdot log(L_R) + Z_d * u + \epsilon_d \end{aligned} \quad (3)$$

where the control terms have been set up to capture possible confounds with the other two impact variables. This last model of the dislikes ratio is of special interest, as the negative subject evaluation avoidance hypothesis implies that $b_d < 0$, with the parameter d_d quantifying the case of a difference on the effect between genders.

2.3 User Incentives Model

We analyze user incentives through an inactivity model that relates the probability of a user to become inactive with the dislikes received by the last video of the user, including the interplay between the content of the video and the gender of the user who posted it. We operationalize the inactivity of a user through the video variable I, which takes value 1 if the user did not post any videos for a period of two months after the video, and 0 otherwise.[2]

$$\begin{aligned} logit(P(I)) &= (\beta + \beta_F \cdot Face + \beta_I \cdot Face \cdot Female + \beta_G \cdot Female) \cdot log(D_R) \\ &+ \alpha + \gamma \cdot log(views) + \delta \cdot log(L_R) + Z_I * u + \epsilon_I \end{aligned} \tag{4}$$

The above equation models a relationship between a logit transformation ($logit(x) = log(x/(1-x))$) of the probability of a user becoming inactive after posting the video $P(I)$ with its dislike ratio for different $Face$ and $Female$ conditions. The parameter α quantifies the baseline tendency to become inactive independently of any video or user variable. The parameters β quantifies how inactivity depends on the dislike ratio, which we can expect to be positive if users respond to negative evaluations of others with higher inactivity tendencies. The parameter β_F quantifies how the role of the dislikes ratio depends on the video having a face and β_G and β_I how this depends on the gender of the user. The negative incentives hypothesis implies a value $\beta_F > 0$, which quantifies the increase in the relationship between the dislikes ratio and the tendency of users to become inactive for videos in the subject evaluation context. The terms $\gamma \cdot log(views)$ and $\delta \cdot log(L_R)$ quantify controls for other properties of the video, and we can expect that likes in particular, as positive evaluation signals, should have a negative effect on the probability to become inactive. The term $Z_I * u$ accounts for random effects of user levels in inactivity tendencies, and ϵ_I measures the model residuals.

Note that the models formalized in the above equations are designed to test the hypotheses explained in the introduction, not to serve as predictors for video impact or user churn. All variables, including views, likes, and dislikes, are measuring a long period after the video has been posted, and thus the models are a way to test association between variables rather than to formulate predictive methods. Our analysis focuses on robustly testing the hypotheses that motivate our research, and thus formulating accurate predictors for user activity or video impact is out of the scope of this research. We fit video impact models and the user incentives model with the lme4 R package [5]. We assess the validity of model assumptions through regression diagnostics on the distribution of residuals and their possible correlations with other model terms. To understand interaction effects, we analyze the *statistical effect* of independent variables on dependent variables by holding all controls constant to their average value. We assess the variance in these predictions by repeating the model fits on 1,000 bootstrap samples of the empirical data, as shown in the Results section.

[2] We replicated the analysis with alternative intervals of one and three months to determine inactivity, and regression models were qualitatively unchanged.

3 Results

3.1 Video Impact Analysis

The fit results of video impact models are shown on Table 2. Videos with a face and by female users receive more views and more likes. There is no significant interaction between $Face$ and $Female$ for the case of likes, but it is significant and positive for views. This indicates that the statistical effect of $Face$ on views is higher for female users. The dislikes ratio model suggests that female users receive less dislikes per view than male users, as c_d is negative and significant. The dislikes ratio model supports the negative subject evaluation avoidance hypothesis, with an estimate of b_d significantly below zero. Nevertheless, the positive interaction term with $Female$ shows that this is the case only for male users, as the terms b_d and d_d cancel out to a slightly positive value.

Table 2. Regression results of impact models. Videos with faces and posted by female users get more views and likes. Videos with faces get less dislikes for male users.

Term	Views model		Likes model		Dislikes model	
Intercept	a_v	$\mathbf{2.128}^{***}$	a_l	$\mathbf{-2.525}^{***}$	a_d	$\mathbf{-2.828}^{***}$
$Face$	b_v	$\mathbf{0.064}^{***}$	b_l	$\mathbf{0.090}^{***}$	b_d	$\mathbf{-0.058}^{***}$
$Female$	c_v	$\mathbf{0.436}^{***}$	c_l	$\mathbf{0.267}^{***}$	c_d	$\mathbf{-0.205}^{***}$
$Face \cdot Female$	d_v	$\mathbf{0.043}^{*}$	d_l	-0.000	d_d	$\mathbf{0.072}^{***}$
$log(D_R)$	f_v	$\mathbf{-0.542}^{***}$	g_l	$\mathbf{0.151}^{***}$		
$log(L_R)$	g_v	$\mathbf{-0.574}^{***}$			g_d	$\mathbf{0.302}^{***}$
$log(views)$			f_l	$\mathbf{-0.166}^{***}$	f_d	$\mathbf{-0.315}^{***}$
AIC	501059.929		305887.837		414279.269	
R^2	0.763		0.617		0.604	
Num. obs	157661		157661		157661	
Num. groups	1556		1556		1556	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

To better understand the various interaction terms of the models, we performed an effect analysis of bootstrap samples, shown on Fig. 1. The statistical effects of $Face$ and $Female$ in the views and likes ratio models discussed above can be observed in the analysis of their respective models. The negative subject evaluation avoidance towards male users can be seen on the right panel, as the estimates of the dislikes ratio are lower for male users when the videos contain a face. The slightly opposite effect for female users confirms our observation that the negative subject evaluation avoidance effect does not exist for female users.

Model control terms reported on Table 2 show that, after accounting for other factors, views are negatively correlated with dislike and like ratios, and that the ratio of both evaluation metrics are positively correlated with each other.

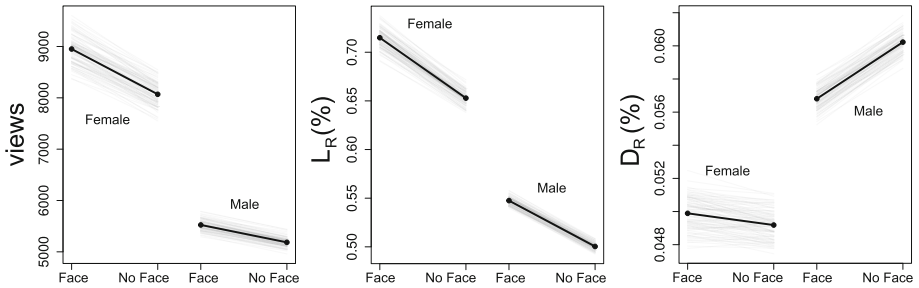

Fig. 1. User gender and video face statistical effect analysis. Fit estimated amount of views, L_R, and D_R as a function of the values of the *Face* and *Gender* variables when controls are set to their average value. Dark lines show the average estimates over 1,000 bootstrap samples of the empirical data, with the shaded lines showing all bootstrap results.

This points to the sublinear scaling of likes and dislikes with views previously reported in [1], and that polarization exists when positive and negative evaluations tend to coexist. The above inferences are coherent with the model assumptions explained in the Materials and Methods section, with normally distributed residuals and no relevant signs of heteroscedasticity.

3.2 Subject Dislikes and Inactivity

The results of the fit of the user incentives model are shown on Table 3. The first column shows the full model as expressed on Eq. 4, including controls and interaction terms. The negative incentives hypothesis is supported, as the estimate of β_F is positive and significant, i.e. the marginal effect of dislikes on the probability to become inactive is higher when the disliked video contains a face. In contrast to the dislikes model of the previous section, gender does not have any significant interaction and the hypothesis holds for both genders.

Regression results are qualitatively unchanged when fitting subsets of the variables, either ignoring controls or interaction terms with gender. Furthermore, a replication of the model with the value of I defined by one and three months instead of two months shed similar results, with β_F positive and significant. The positive and significant estimate of β in all models shows that higher dislikes ratios lead to higher chances for users to become inactive. The controls of the user incentives model show that the likes ratio is negatively associated with the probability of users becoming inactive, while the logarithm of the amount of views has a positive relationship with $P(I)$ when other variables are taken into account too. This suggests that likes incentivize users to stay active, and that views without likes are not the leading incentive for users to keep posting videos.

To understand interaction terms and incentives better, we fitted the model subset with the lowest Bayesian Information Criterion, reported on the last column of Table 3. We analyzed the estimate of $P(I)$ in the model as a function of *Face* and D_R in 1,000 bootstrap samples of the data. The results of this analysis

Table 3. Regression results of inactivity models. The dislike ratio is positively associated with the probability of users to become inactive, with a stronger association when the disliked videos contain a face. Gender has no significant effect in the model.

Term	Parameter	Full model	Subset 1	Subset 2	Best Model
Intercept	α	$\mathbf{-1.513}^{***}$	$\mathbf{-1.155}^{***}$	$\mathbf{-1.157}^{***}$	$\mathbf{-1.508}^{***}$
$log(D_R)$	β	$\mathbf{0.143}^{***}$	$\mathbf{0.058}^{***}$	$\mathbf{0.059}^{***}$	$\mathbf{0.141}^{***}$
$Face \cdot log(D_R)$	β_F	$\mathbf{0.027}^{***}$	$\mathbf{0.025}^{***}$	$\mathbf{0.027}^{***}$	$\mathbf{0.025}^{***}$
$Female \cdot log(D_R)$	β_G	-0.010		-0.007	
$Face \cdot Female \cdot log(D_R)$	β_I	-0.016		-0.014	
$log(views)$	γ	$\mathbf{0.045}^{***}$			$\mathbf{0.045}^{***}$
$log(L_R)$	δ	$\mathbf{-0.109}^{***}$			$\mathbf{-0.108}^{***}$
BIC		29884.423	29884.150	29907.510	29861.222
Cond. McFadden's R^2		0.34047	0.3394163	0.339429	0.3404538
Num. obs		157443	157443	157443	157443
Num. groups		1556	1556	1556	1556

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

are shown on Fig. 2, where the trend of the probability of inactivity estimated by the model is shown in relationship to the dislike ratio for the cases of videos with faces and without faces. The negative incentives effect is present, as the trend for videos with faces grows faster with the dislikes ratio than for videos without faces. To formalize the test of the negative incentives hypothesis over the bootstrap samples, the right panel of Fig. 2 shows the histogram of the estimate of β_F over the 1,000 bootstrap samples. From all samples, only one had a

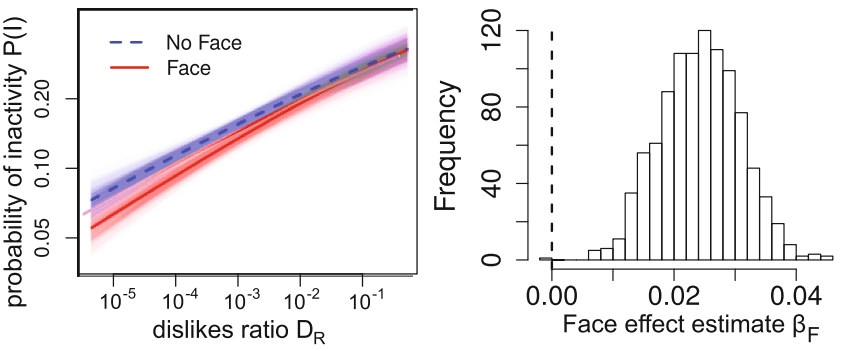

Fig. 2. Statistical effect of dislikes on inactivity. Left: Predicted probability of becoming inactive (I) as a function of the dislikes per view ratio (D_R) for videos with a face and without a face. Shaded lines show the results over 1,000 bootstrap samples of the dataset. Right: Histogram of estimates of the face effect parameter β_F in the 1000 bootstrap samples. The probability of becoming inactive grows faster with the dislike ratio for videos with faces than for videos without faces.

value slightly below zero, illustrating that the null hypothesis that $\beta_F = 0$ can be rejected with $p < 0.05$.

4 Discussion

To analyze the role of the dislike button in participatory media, we generated and analyzed a dataset with more than 150,000 videos from YouTube. We processed their related images with face detection to identify when people are salient in video content. Our views and likes models showed that videos being posted by women receive more likes and more views. This effect had an interaction with a video having a face in the views model, suggesting that YouTube users are more likely to watch videos about people if they have been uploaded by a woman. While it stays as an open question to study the gender of the faces related to the videos, this interaction between gender, faces, and views suggests that female-related images might be used to attract the attention of YouTube users.

Our analyses only focused on faces detected in images related to the videos, and we did not include a nuanced analysis of the full content of the video. While our results show a signal in the noise in the hypothesized directions, advanced video processing techniques offer the opportunity to extend and improve our work. Quantifying the amount of persons appearing in a video or the amount of time devoted to people is a promising avenue to have more precise measurements of the subject evaluation context. Furthermore, identifying the individuals depicted on each video can reveal which videos are centered around the user owning the channel, in which the negative incentive of receiving dislikes might have the strongest effect.

Our regression models show that videos with faces hinder the reception of negative evaluations (dislikes), but only for male users. Analyzing the probability of inactivity of users, we found that videos with high ratios of dislikes per view are associated with users becoming inactive, and that this effect is stronger when videos contain faces, as hypothesized. In our inactivity analysis we found no effect of gender, but our controls with other signals show that likes are negatively associated with users becoming inactive. Our results showed a surprising interaction between faces and gender in the amount of dislikes received by videos, which calls for further research to identify the reasons that drive users away from disliking videos posted by male users in the subject evaluation context.

While we identified post hoc correlations in our analysis, our conclusions are not directly applicable yet to the design of social media. Real-time analyses can shed light on whether the patterns that we identified are predictive of the inactivity of users. In the case of being consistent with our findings, future designs of social media interfaces should consider the risks of giving the possibility to dislike user-centered content. To ensure the sustainability and inclusivity of social media in the future, we need to further study which platform designs, conditions, and contexts lead to users disencouraging other users from being active in social media, as we found here in the case of dislikes in YouTube.

Acknowledgements. This research was funded by the Swiss NSF (Grant number: CR21I1_146499)

Appendix

As a preliminary step to fitting models and testing hypotheses, we survey descriptive statistics to guide the models explained in the previous section. The distributions of views, likes, and dislikes per video are shown on Fig. 3. The histogram of the left panel confirms our observations over the mean and median values of Table 1: all variables are right skewed. This skewness presents heavy right tails that, when eyeballing the plots, suggest the possibility that views, likes, and dislikes follow power-law distributions. Nevertheless, this possibility seems less plausible on the Complementary Cumulative Density Function (CCDF) shown on the right panel of Fig. 3, where the right tails decay faster than it would be expected for a power-law.

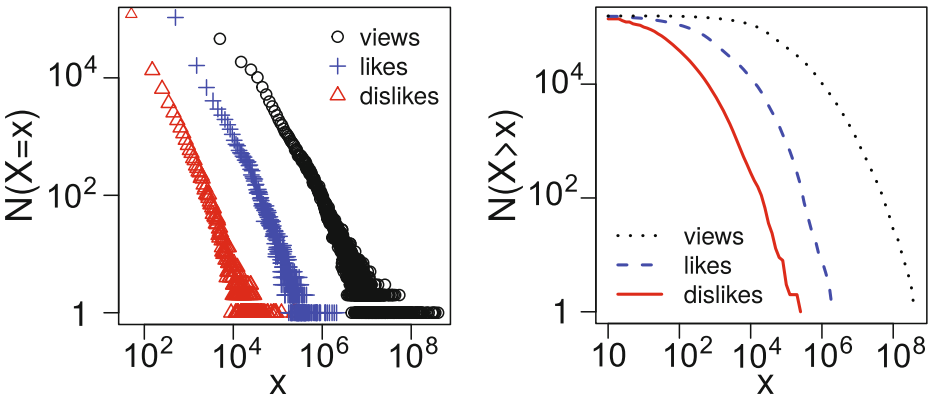

Fig. 3. Video impact distributions. Left: histogram of amount of views, likes, and dislikes over the videos of the dataset. Right: Complementary Cumulative Density Function (CCDF) of counts of videos with more than a certain amount of views, likes, and dislikes. While the histograms show right-skewness, the CCDF of counts show a decay faster than a power-law.

To have a better idea on whether the distributions of amount of views, likes, and dislikes might have scaling properties or diverging moments [25], we applied the method explained in [3] to verify that they do not follow a power-law distribution. We fitted power-law and log-normal distributions to the empirical data, comparing the fits in a log-likelihood ratio test. The results lend very strong evidence favoring the log-normal distribution over the power-law in all three cases: views ($LLR = 449.97, p < 0.01$), likes ($LLR = 275.7, p < 0.01$), and dislikes ($LLR = 159.99, p < 0.01$). This is an example of how informal statistics can be

misleading in deciding whether distributions follow a power-law [3,8], suggesting that we should assume the distributions as log-normally distributed instead.

To ensure that we analyze the evaluative tendencies of videos and not their intrinsic correlation with video popularity, we divide likes and dislikes by the amount of views in the variables $L_R = likes/views$ and $D_R = dislikes/views$. These two variables and the amount of views are all roughly log-normally distributed, as it can be appreciated on the histograms of log-transformed values shown on the upper panels of Fig. 4. Some minor skewness can be attributed to integer approximations and boundary values. To cope with these possible deviations from normality in our models, we perform regression diagnostics to model fits to check that residuals are approximately normally distributed.

Figure 4D shows the distribution of the logarithm of the time between videos of the same user $log(\Delta t)$. A clear bimodality is present, but it disappears when normalizing over the average time between videos of each user $\langle t \rangle$. Figure 4E shows the distribution of $log(\Delta t/\langle t \rangle)$, where no bimodality can be observed. This points to the source of bimodality being a variable at the user level, i.e. the activity rate of each user, as the distribution of time intervals collapses to a unimodal distribution after normalization. In our mixed effects regression models of $P(I)$, we include random effects in the form of an intercept for each user that correct for this pattern, ensuring that our results are not a confound with idiographic properties of the users.

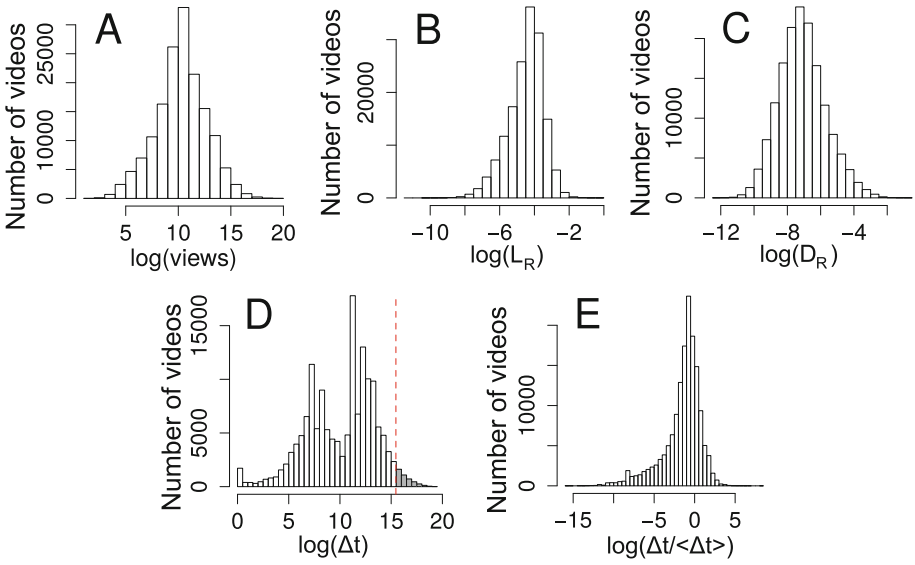

Fig. 4. Histograms of log-transformed video metrics. The upper panels (A,B,C) show the histogram of log-transformed amount of views, likes ratio (L_R), and dislikes ratio (D_R) over the videos of the dataset. Panel D shows the histograms of log-transformed time intervals between videos of the same user (Δt), in seconds. The vertical red line shows the threshold of inactivity of 2 months. Panel E shows the histogram of time intervals normalized over the average time between videos of the user.

References

1. Abisheva, A., Garcia, D., Schweitzer, F.: When the filter bubble bursts: collective evaluation dynamics in online communities. In: Proceedings of the 8th ACM Conference on Web Science, pp. 307–308 (2016)
2. Abisheva, A., Garimella, V.R.K., Garcia, D., Weber, I.: Who watches (and shares) what on YouTube? and when? using Twitter to understand YouTube viewership. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 593–602 (2014)
3. Alstott, J., Bullmore, E., Plenz, D.: powerlaw: a python package for analysis of heavy-tailed distributions. PLoS ONE **9**(1), e85777 (2014)
4. Aragón, P., Gómez, V., Kaltenbrunner, A.: To thread or not to thread: the impact of conversation threading on online discussion. In: ICWSM, pp. 12–21 (2017)
5. Bates, D., Mchler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1), 1–48 (2015)
6. Cashmore, P.: Should Facebook add a dislike button? In: CNN articles (2010). http://cnn.it/2t7tu2h
7. Castells, M.: The Rise of the Network Society, vol. 1. Wiley, Hoboken (1996)
8. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. **51**(4), 661 (2009)
9. Crane, R., Sornette, D.: Robust dynamic classes revealed by measuring the response function of a social system. Proc. Natl. Acad. Sci. **105**(41), 15649–15653 (2008)
10. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proc. Natl. Acad. Sci. **113**(3), 554–559 (2016)
11. Garcia, D., Mavrodiev, P., Casati, D., Schweitzer, F.: Understanding popularity, reputation, and social influence in the Twitter society. Policy Internet (2017)
12. Garcia, D., Mavrodiev, P., Schweitzer, F.: Social resilience in online communities: the autopsy of friendster. In: Proceedings of the 1st ACM Conference in Online Social Networks (COSN 2013), pp. 39–50 (2013)
13. Garcia, D., Mendez, F., Serdult, U., Schweitzer, F.: Political polarization and popularity in online participatory media: an integrated approach. In: Proceedings of the 1st Workshop on Politics, Elections and Data - PLEAD 2012, pp. 3–10 (2012)
14. Garcia, D., Strohmaier, M.: The qwerty effect on the web: how typing shapes the meaning of words in online human-computer interaction. In: Proceedings of the 25th International Conference on World Wide Web, pp. 661–670 (2016)
15. Greenwald, A.G., Banaji, M.R.: Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol. Rev. **102**(1), 4 (1995)
16. Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., Wilson, C.: Bias in online freelance marketplaces: evidence from taskrabbit and fiverr. In: CSCW, pp. 1914–1933 (2017)
17. Huang, T.K., Ribeiro, B., Madhyastha, H.V., Faloutsos, M.: The socio-monetary incentives of online social network malware campaigns. In: Proceedings of the Second ACM Conference on Online Social Networks, pp. 259–270 (2014)
18. Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., Strohmaier, M.: Inferring gender from names on the web: a comparative evaluation of gender detection methods. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 53–54 (2016)

19. Karnstedt, M., Hennessy, T., Chan, J., Hayes, C.: Churn in social networks: a discussion boards case study. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 233–240. IEEE (2010)
20. Kooti, F., Magno, G., Weber, I.: The social name-letter effect on online social networks. In: International Conference on Social Informatics, pp. 216–227 (2014)
21. Lakkaraju, H., McAuley, J.J., Leskovec, J.: What's in a name? understanding the interplay between titles, content, and communities in social media. ICWSM **1**(2), 3 (2013)
22. Magno, G., Weber, I.: International gender differences and gaps in online social networks. In: International Conference on Social Informatics, pp. 121–138 (2014)
23. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys 2013, pp. 165–172. ACM (2013)
24. Mieghem, P.V.: Human psychology of common appraisal: the reddit score. IEEE Trans. Multimed. **13**(6), 1404–1406 (2011)
25. Newman, M.E.: Power laws, pareto distributions and zipf's law. Contemp. Phys. **46**(5), 323–351 (2005)
26. Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.Y., Kapadia, A., Rojas, F.: Twitter's glass ceiling: the effect of perceived gender on online visibility. In: ICWSM, pp. 289–298 (2016)
27. Oksanen, A., Garcia, D., Sirola, A., Nsi, M., Kaakinen, M., Keipi, T., Rsnen, P.: Pro-anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. J. Med. Internet Res. **11**(17), e2560 (2015)
28. Palmer, C.: Risk perception: another look at the 'white male' effect. Health, Risk Soc. **5**(1), 71–83 (2003)
29. Ribeiro, B.: Modeling and predicting the growth and death of membership-based websites. In: Proceedings of the 23rd International Conference on World Wide Web, WWW 2014, pp. 653–664. ACM (2014)
30. Simmel, G.: Fashion. Am. J. Sociol. **62**(6), 541–558 (1957)
31. Stommel, S., Garcia, D., Abisheva, A., Schweitzer, F.: Anticipated shocks in online activity: response functions of attention and word-of-mouth processes. In: Proceedings of the 8th ACM Conference on Web Science, pp. 274–275 (2016)
32. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Commun. ACM **53**(8), 80 (2010)
33. Szell, M., Thurner, S.: Measuring social dynamics in a massive multiplayer online game. Soc. Netw. **32**(4), 313–329 (2010)
34. Tassi, P.: Facebook Didn't Kill Digg, Reddit Did. In: Forbes (2012). http://bit.ly/2tx8e5C
35. Tufekci, Z.: Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: ICWSM (2014)
36. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424 (2002)
37. Walker, J., Ante, S.E.: Once a social media star, Digg sells for $500,000. The Wall Street J. (2012). http://on.wsj.com/2uv1AAS
38. Yasseri, T., Sumi, R., Rung, A., Kornai, A., Kertsz, J.: Dynamics of conflicts in Wikipedia. PLOS ONE **7**(6), 1–12 (2012)
39. Zhou, E., Cao, Z., Yin, Q.: Naive-deep face recognition: touching the limit of LFW benchmark or not? arXiv preprint arXiv:1501.04690 (2015)