Fakes views, real trends. The crucial role of time in the authentication of online engagement metrics

Maria Castaldo, Paolo Frasca, Tommaso Venturini, Floriana Gargiulo

▶ **To cite this version:**

Maria Castaldo, Paolo Frasca, Tommaso Venturini, Floriana Gargiulo. Fakes views, real trends. The crucial role of time in the authentication of online engagement metrics. 2021. hal-03311188v1

HAL Id: hal-03311188

https://hal.science/hal-03311188v1

Preprint submitted on 30 Jul 2021 (v1), last revised 20 Jul 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FAKES VIEWS, REAL TRENDS. THE CRUCIAL ROLE OF TIME IN THE AUTHENTICATION OF ONLINE ENGAGEMENT METRICS

A PREPRINT

Maria Castaldo
Univ. Grenoble Alpes, CNRS
Inria, Grenoble INP, GIPSA-lab
F-38000 Grenoble, France
maria.castaldo@grenoble-inp.fr

Paolo Frasca
Univ. Grenoble Alpes, CNRS
Inria, Grenoble INP, GIPSA-lab
F-38000 Grenoble, France
paolo.frasca@grenoble-inp.fr

Tommaso Venturini
CIS, CNRS
75017 Paris, France
tommaso.venturini@cnrs.fr

Floriana Gargiulo
Gemass, CNRS
75017 Paris, France
floriana.gargiulo@cnrs.fr

July 30, 2021

ABSTRACT

In this work we dig into YouTube's policy to ensure the authenticity of engagement metrics: the platform decreases the views counts of videos when it deems some views to be produced not by humans but rather by automatic systems. Drawing on data that we collected during 18 months from 1300 French YouTube channels, we highlight the massive scale of YouTube views corrections, which affect the majority of the monitored channels and half the videos in our corpus. Given the extent of YouTube corrections, and bearing in mind the dynamics that regulate content diffusion on the platform, we investigate whether manipulating engagement metrics could produce some effect in boosting the popularity of videos. Our analysis shows that the number of illegitimate views removed from videos is not independent from their final popularity. Videos corrected later are on average more popular than those corrected earlier or the uncorrected ones and the median popularity of slowly corrected videos is almost 4 times bigger than the one of uncorrected videos. Our results bring forth the crucial role that time plays in the diffusion of online content: views that are corrected too late can interfere with the platform's trendiness dynamics and help a content reach a broader audience.

Keywords Social Bots · Manipulation of Online Information · YouTube

1 Introduction

"We want to make sure that videos are viewed by actual humans and not computer programs"(22). As clearly stated in its support page, YouTube has a policy on views counting and is not afraid of applying it. At the end of December 2012, the platform singlehandedly deleted 2-billion views from the channels of record companies such as Universal and Sony (19) (23) (2) (16) and over the years countless youtubers have suffered sudden and drastic cuts to their views' counts (and hundreds have loudly complained about it, often through videos posted on YouTube itself). The rationale for these interventions, explained in YouTube's official webpages (21) (31) (22), originates from the need to preserve "meaningful human interaction on the platform" as well as "the safety of [...] creators, viewers, and partners"(21) and to oppose "anything that artificially increases the number of views, likes, comments or other metric either through the use of automatic systems or by serving up videos to unsuspecting viewers" (21).

Unsurprisingly, not everyone agrees with this policy and numerous researchers have tried to test the fairness of its implementation (24) (34) (28) drawing on research carried out on ads fraud in other social media (12) (30). Marciel et al. (28), for instance, created a sample of YouTube channels and inflated their views through automated programs.

Strikingly, they found that "YouTube monetizes (almost) all the fake views" generated by the authors, while it "detects them more accurately when videos are not monetized".

No research, however, has so far investigated a seemingly marginal aspect of YouTube fake views policy: its timing. Because most existing research focused on monetization, the exact timing of fake views removal appeared irrelevant, as long as it happened before the settling of the advertising bills (on the 21st of the following month). This perspective is, however, without considering another crucial element of YouTube's infrastructure: its trendiness cycle. In YouTube, as in any other social media, future visibility (and therefore popularity) is highly dependent on past popularity, as trending contents are more likely to be picked up by humans and algorithms and be further recommended to platform users (20). This tendency is observed in human influencers, who are sensitive to vanity metrics of trendiness and early popularity (38) (45), and in the platform recommendation algorithm. The latter represents one of the most important (if not the most important) source of views on the platform (52) and, according to its own creators, "in addition to the first-order effect of simply recommending new videos that users want to watch, [has] a critical secondary phenomenon of bootstrapping and propagating viral content" (15), i.e. contents that are 'on the rise'. As most social media recommendation systems, YouTube recommending algorithm analyzes the engagement previously generated by content to anticipate (and promote) its interest for YouTube audiences, creating a positive feedback that skews YouTube popularity according to a rich-get-richer dynamic (5) (32) (42).YouTube engineers have acknowledged this dynamics, but not entirely defused it: "models trained using data generated from the current system will be biased, causing a feedback loop effect. How to effectively and efficiently learn to reduce such biases is an open question."(51). And this is where the exact timing of the fake views policing becomes crucial.

Indeed, if the correction of illegitimate views happens too late, these views have the potential to weight in the cycle of trendiness (11) and unfairly propel targeted contents. In other words, if YouTube fake views correction is significantly slower than its recommendation dynamics, then illicitly promoted videos risk to be favored by human and algorithmic recommendations, and thus reach larger audiences and collect extra real views. If fake views are able to generate a cascade effect that increases the popularity of some content, then they potentially constitute a tool to manipulate online debates. Not unlike the widely studied phenomenon of social bots that automatically produce content mimicking human behavior, fake views could give the false impression that some piece of information, regardless of its accuracy, is highly popular and endorsed by many (18). Such activities may result in endangering democratic processes (50), as we now know in the case of content-producing social bots in the 2010 U.S. midterm elections (36), the 2010 Massachusetts special election (29), the 2016 U.S. Presidential election (when one fifth of Twitter conversations was generated by bots (4)) or during many other political campaigns (27) (3) (37) (26).

While much research has been dedicated to identify social bots, both tracking their sources (35) (7) (40) (44) and analyzing artificially produced contents (4) (46) (17) (48) (8) (47) (13), (14), (41) (25), little attention has so far been devoted to the role that fake clicks or fake views might play in the diffusion of content. This work addresses this open and urgent question, thanks to a 18 month-long collection of the hourly views counts of over 350.000 videos published by more than a thousand French YouTube channels about politics and news. This dataset will allow us, for the first time, to examine in detail the timing of YouTube fake views corrections and its potentially unfortunate consequences.

2 Results

2.1 Scale of the phenomenon

Our first observation is the large scale of the phenomenon of views corrections. While YouTube does not provide precise information on views corrections (through its API or otherwise), we can obtain an estimation of this activity by observing the hours with a negative delta in the views count, i.e. those hours in which one video loses more views than it gains. Figure 1A shows some examples of hourly changes in terms of views in the first 170 hours since publication of 5 videos randomly chosen from our collection. In the following we will refer to such negative deltas as to *negative views*. Negative views are necessarily an underestimation of the platform real corrections because they result from the difference between views gained and views removed[1]. Yet, the scale of the YouTube policy is impressive: we detected negative views for almost all the monitored channels and half of the videos in our corpus. The extent of YouTube corrections by itself encourages to have a closer look at the correction activities of the platform.

2.2 Correction rhythms

YouTube correction policy presents some interesting recurrent patterns. First of all, the majority of the observed corrective *interventions*, namely the time-slots with negative views, takes place between 4 p.m. and 5 p.m. The analysis

[1]We present our main results using negative views as a measure of YouTube's corrections: all our conclusions remain valid (and actually become stronger) when estimating corrections in more sophisticated ways (See Material and Methods)

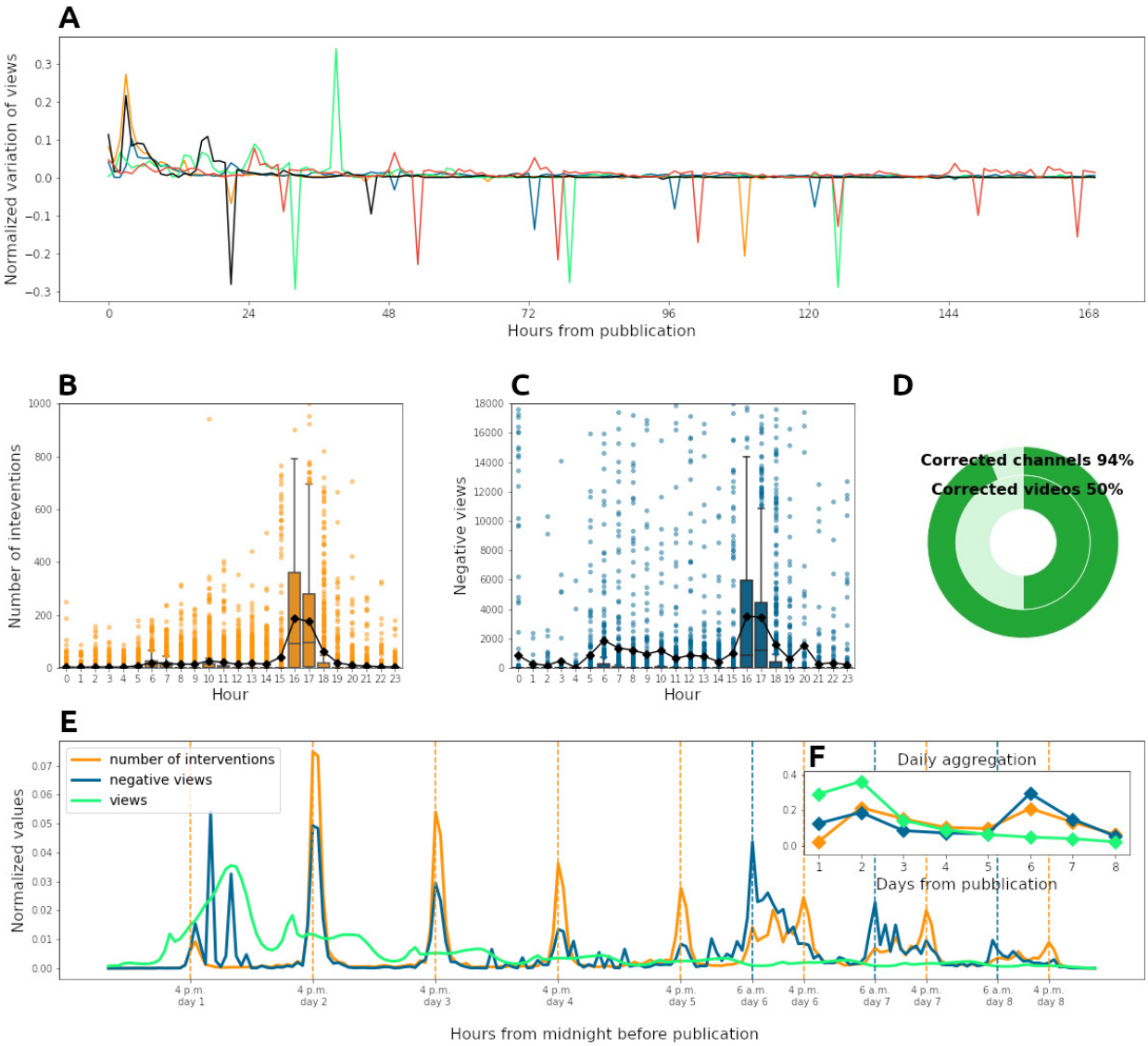

Figure 1: **Correction rhythms. A** Hourly change in the views counts of 5 sample videos (normalized by their final number of views). **B** Distribution of the number of observable interventions per hour of the day. **C** Distribution of negative views per hour of the day. **D** Scale of the phenomenon. **E** Distribution of views, number of interventions, and negative views along the life of a video. Time is counted in hours and it is set to start at 00:00 of the day of publication. **F** Daily aggregation of the same quantities presented in subplot E.

of the distribution of the number of interventions during the hours of the day (Figure 1B) reveals a substantial difference between the average number of intervention at 4-5 p.m. (around 190) and the remaining hours of the day (always close to 0). The same concentration between 4 and 5 p.m. can be observed when considering the number of negative views (rather than the correction events) (Figure 1C). In order get a deeper insight in the correction policy and investigate the possible interference with the recommendation system, we should consider the timing of the corrections not only in absolute, but also relatively to the moment of publication of each video. Figure 1E provides an overview of the distribution of the number of corrective interventions, negative views and views along the life of a video, starting from the midnight before publication (we set to zero the number of interventions, views and negative views for all the hours preceding the publication, in order to align the hours of the day for videos that are published at different times). Figure 1E allows us to better understand the mechanism of adjusting views. As we can see, in the first 5 days the number of the corrective interventions shows a precise periodical behavior, with peaks at 4 and 5 p.m. that decrease as days pass by. From the 6th day on (i.e. at hour 120) the pattern changes: even if peaks at 4 p.m. are preserved, other

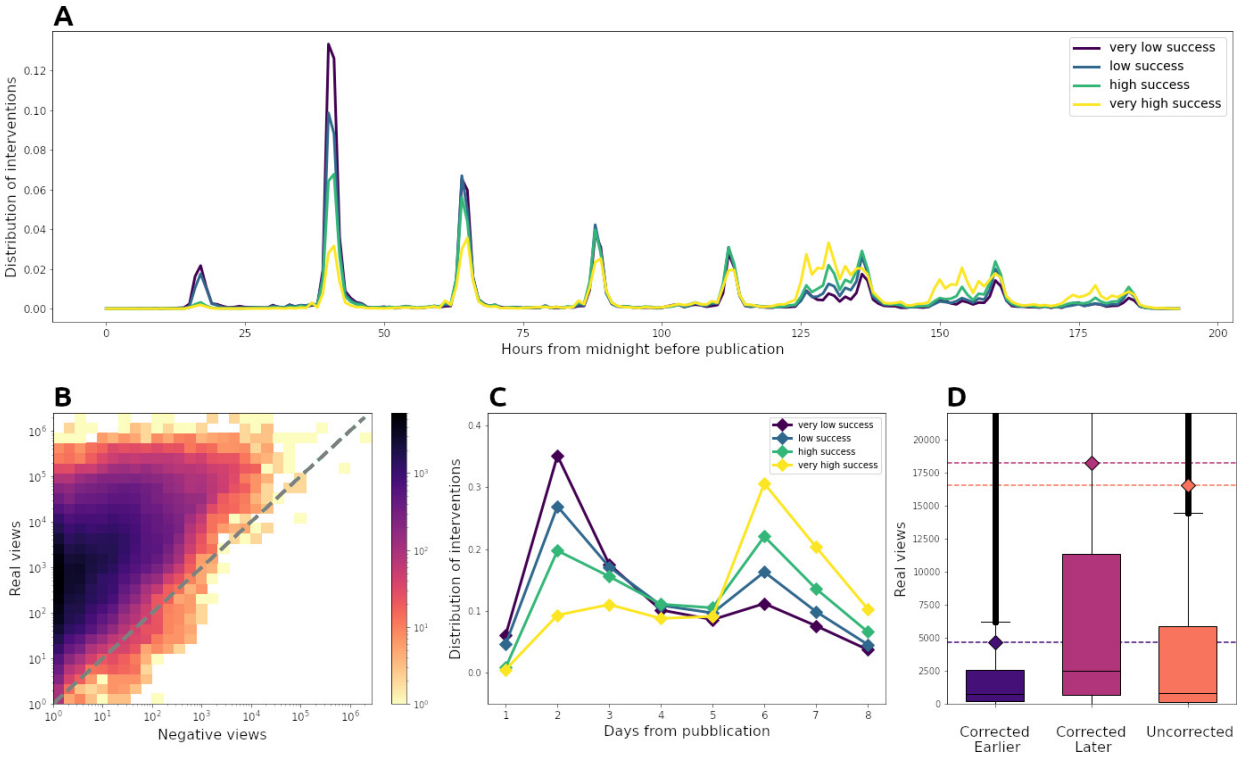

Figure 2: **Success and Negative Views. A** Distribution of interventions among the hours following the publication day midnight. **B** 2D histogram of negative views and real views after corrections. **C** Distribution of interventions among the days following the publication. **D** Distribution of real views for corrected earlier, corrected later and uncorrected videos.

time slots present significant amounts of interventions, especially between 6 a.m. and 4 p.m. Moreover, by looking at the number of negative views, from the 6th day the daily peak shifts at 6 a.m..

Most importantly for our argument, the daily aggregation in Figure 1F reveals that, while the number of views decreases steadily from the day two, the number of interventions and of negative views reaches its peak at day 6. Such a delay in the corrections raises the question of whether illegitimate views are removed fast enough by the platform or if they survive long enough to interfere with trendiness cycle.

2.3 Corrections and success

To investigate the connection between illegitimate views and the popularity of videos, we examine the relation between the number of negative views and the total number of legitimate views, where the latter is the number views returned by YouTube API 170 hours after the publication of each video (hence corresponding to all the views collected by the videos in one week minus the views subtracted by YouTube). Figure 2B shows the density of observations for every tuple of negative and legitimate views. In principle, such two quantities should be independent: indeed, if the YouTube correction policy was rapid and efficient enough, illegitimate views should have no impact on real ones and, at the same time, there is no apparent reason why popular videos should attract more fake views then others. And yet, Figure 2B highlights a relationship of dependence between the two variables: the distribution of total real views changes for higher negative views and the correlation between the logarithms of the two variables is 0.48.

To better understand the mechanism of correction and its relation to popularity, we divided videos into four classes of popularity: *very low / low / high / very high success*, representing the four quartiles of the total legitimate views distribution. Figure 2A presents, for each class, the normalized number of YouTube interventions at every hour. The figure shows that the lowest popularity quartiles are corrected earlier, mainly around 4-5 p.m. of the second day after publication, while the most popular quartile tends to be corrected more uniformly throughout the whole week of monitoring. Moreover, *non-periodic corrections* seem to be more present for *very high popularity* videos that receive significant interventions outside the regular 4-5 p.m. slot. Figure 2C shows the aggregation of the same metrics by day, making even clearer that the more a video is successful the later we observe corrections on its views. In conclusion, highly popular videos have more negative views and appear to be corrected later.

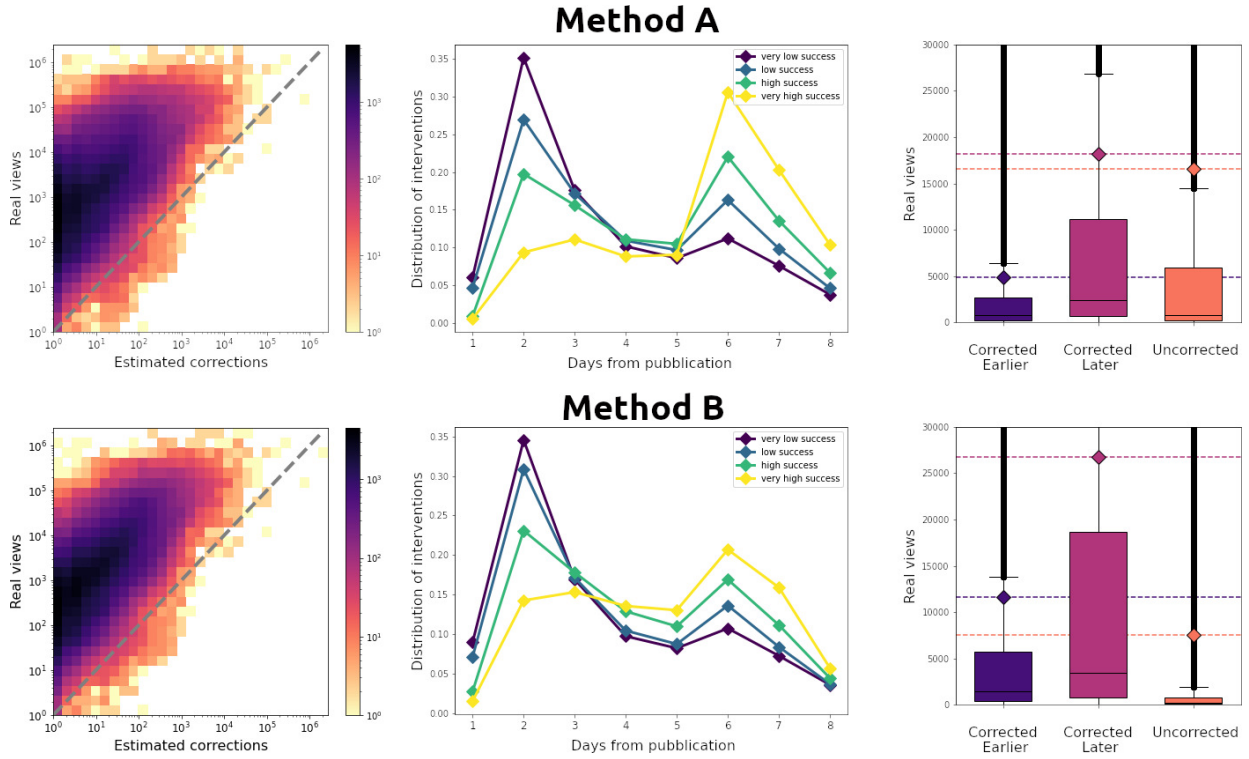

Figure 3: **Validation of the results: (Left):** 2D-histograms of the number of real views and corrections, after estimating corrections through Methods A and B. **(Middle)**: Distribution of intervention along the days following publication. **(Right)**: Distribution of total views for videos corrected earlier, videos corrected later, and uncorrected videos.

While Figures 2A-C indicate that popular videos tend to be corrected later than less popular ones, Figure 2D indicates that the converse is also true: videos that are corrected later tend to be more popular. The plot compares the distribution of popularity of the videos that display the majority of the negative views in the first 5 days since their publication (i.e. *corrected earlier*) with those who get corrected mainly after the 6th day (i.e. *corrected later*). As revealed by the plot, videos corrected later are more popular than videos corrected earlier, as well as of videos for which we observe no corrections (i.e. *uncorrected*). The difference between the distributions of popularity of the three classes of videos has been tested through a two-sample Kolmogorov Smirnov test and all p-values were found to be less than 10^{-315}, confirming the significance of the differences. Such evidence supports our claim and confirms that videos corrected later have more chances to become popular and reach a wider public. Uncorrected videos present a lower expected popularity than videos corrected late in their life and the median number of legitimate views gained by slowly corrected video is almost 4 times greater than the median of videos on which the platform did not intervene.

3 Discussion

3.1 Limitations

As we already acknowledged, we have no access to the actual number of views corrections performed by YouTube on its videos. In fact, YouTube API's restrictions, described in the Materials and Methods section, prevent us from collecting the exact evolution in time of the engagement metrics. In order to monitor our full list of channels and still meet YouTube's constraints on the number of daily queries that can be performed through the API, we need our time series to be sampled every hour. In principle, this relatively coarse aggregation could obfuscate the real size of the platform corrections and could even hide some platform interventions when the size of the correction is smaller that the views collected in an hour. In the Results section, we have presented our findings by taking negative views (in the hourly views count) as a proxy of corrections. In the next paragraph, we will show that our findings are robust when using more elaborate estimates of the corrections.

3.2 Validating the results

Even though exact data about the YouTube correction policy are not disclosed, our results can be validated by estimating the platform corrections from the positive and negative views observed in our collection. Method A (see the Materials and Methods section for a full description) aims at reconstructing the real size of corrections in the hours in which we register negative views. Method B refines method A by detecting hours with a exceedingly high difference of views when compared to adjacent hours and attributes such a difference to the presence of a correction intervention carried out by the platform. Figure 3 displays the same analysis done in Figure 2B and Figure 2D, with the expected corrections reconstructed through Methods A and B. In both cases, videos corrected later remain more popular than videos correct earlier or uncorrected. All difference in distributions have been tested with Kolmogorov-Smirnov tests and result to be significant (p-values lower than 10^{-245}). Method B even accentuates such a difference in the distributions of popularity. At the same time, the 2D-histogram not only still shows a linear correlation between the logarithm of corrections and real views, but such linear correlation becomes even more important. In fact, the correlation coefficient grows from 0.48 to 0.55 when applying Method A and to 0.65 with Method B. In the absence of methods able to perfectly reconstruct the exact time evolution of the views count, we believe Methods A and B to be reasonable attempts to validate the findings presented so far.

3.3 From fake views to real views

As we pointed out, many videos only show corrections late in their life, sometimes after 6 days since publication. This evidence rises concerns in itself when acknowledging the importance that earlier popularity has in determining future visibility. "Rich gets richer" behaviors have been repeatedly observed in the evolution of views counts on the platform (6), and the total number of previous views has been credited as the most important predictor of future popularity (6) (43) (33).

Rewarding trendy contents with more visibility is a feature of the recommendation system, which –according to its own developers– is designed to bootstrap and propagate viral content. While encouraging diffusion trends, recommendation systems also constitute the number one source of views for the majority of YouTube videos (52) (53) and hence regulate the bulk of the attention economy on the platform. Its suggestions are updated on a relatively fast basis: according to Roth et al., for instance, two thirds of the suggestions given in a certain moment are not anymore associated with the same video after 2 days (39).

Such a fast refreshing of suggested videos, along with the massive impact recommendation system has in the diffusion of content, makes the likelihood of a delay in correcting views count even more alarming. With a 6 days delay in the correction policy a video would handsomely have enough time to be recognized as viral and hence be suggested to a wider audience then it would have reached without illegitimate boosting.

3.4 Conclusion: a call for transparency

Despite the absence of first-hand data on fake views corrections, we believe that our results robustly indicate the massive scale of views corrections and suggest the existence of a positive correlation between the illicit boosting and the popularity of YouTube videos. Our results highlight the crucial role of time in the diffusion of online content and, for the first time, reveal the role that even a relatively short delay in the correction of views frauds can play in the manipulation of public debate. Given the importance of the subject and the potential harm from the malfunctioning of the correction policy, our findings should –at the very least– encourage YouTube to include in its API the number of corrected views for each videos. As we have shown, this information is crucial to investigate the alarming possibility that relatively easy techniques of views inflation, if used early enough in the life of a video, could effectively manipulate the visibility by triggering the trending dynamics sustained by manual and algorithmic recommendation.

4 Material and Methods

4.1 Data

Starting from December 2019, through a collaboration with the Qatar Computing Research Institute (QCRI), we collected the time evolution of views of any video published by a list of over a thousand French YouTube channels dealing with politics and news disclosure. This dataset is particularly interesting as it cannot be straightforwardly obtained through the official YouTube application programming interface (API) (1). Indeed for the latest several years the YouTube API (1) has restricted the accessible engagement metrics of a video to the current values only, without providing their temporal evolution any longer: this restriction complicates temporal studies on development and prediction of online attention and indeed studies on the temporal evolution of engagement metrics have not been

	Hidden interventions	Hidden corrections	Mean distance error	Misclassification	Wrongly introduced corrections	Wrongly introduced interventions
Hourly Aggregation	65%	66 %	0.24	46%	0%	0%
Method A	65%	58 %	0.21	45%	0.6%	0%
Method B	54%	44 %	0.17	32%	2%	9%

Table 1: **Validation of Methods A and B on a 10-minute views series.** The table shows the loss of information with hourly aggregation and with the estimates by Methods A and B.

carried out since 2014 (42) (32).
The dataset covers 1300 popular French channels that are particularly influential in the French public debate. These channels, with their description, are available at (10). The channels have been selected through a qualitative analysis of the French YouTube, aiming to identify relevant actors that diffuse political opinions through the platform. The selected channels belong to the following categories: local and national media; influential YouTubers discussing political topics; militant associations; politicians; candidates for the 2019 European elections; political parties; Yellow Vests groups; associations devoted to public causes; large public or private institutions. YouTube provides no information about the location from which videos are viewed, but since the channels of our corpus focus on French public debate, we can assume most of their viewers to be based in France.
For these channels, we recorded hour by hour the evolution of views of each video published after the 9 of December 2019, for an entire week after publication (170 hours). Between the 9 of December 2019 and the 20 of May 2021, we collected the views time series of 385.151 videos. For every video, we also collected its title, its description and other metadata available through the official YouTube API. The dataset is available at (9) (resource IDs have been mapped to anonymize values consistently along the dataset). The choice of collecting only one week of views evolution is justified by noticing that news channels often collect the majority of their views in few days after publication, presenting a strong initial burst followed by a power-law decay (49). Indeed our data confirm this fast decrease of users engagement as only 3% of views is obtained in the last 24 hours and this percentage decreases to 1.5% in the last 12 hours.

4.2 Methods to estimate corrections

As the collection of data from YouTube is restricted, we propose here some methods to make our conclusions more robust. To reduce the chance for our findings to be the result of some bias, we need to infer the corrections c_h^i made by the platform at hour h on video i from the compound quantity $\tilde{v}_h^i = v_h^i - c_h^i$, where v_h^i are the effective non-observable views collected by video i during hour h, and v_h^i are the observed variation in the cumulative views counts. As we already argued, a basic way to (under)estimate corrections is to simply assume them equal to negative views. Additionally, more sophisticated methods can be used to rebuild the time series c_h^i starting from $\tilde{v}_h^i$. In general, we would like the estimated values $\hat{c}_h^i$ to:

- reduce the (absolute) distance to real corrections $|\hat{c}_h^i - c_h^i|$ (*distance error*);
- correctly label videos belonging to the classes of *corrected earlier, corrected later* and *uncorrected* videos, hence to minimize a *misclassification error* defined as the percentage of mislabeled videos.

To best develop methods able to reconstruct the original time series, we downloaded the evolution of views every 10 minutes in the first 7 days since publication for 126 random videos published by our monitored channels. Using this 10-minute-frequency timeseries allows us to understand how much information is lost through hourly aggregation. Such loss turns out to be significant: aggregating values by the hour hides 65% of interventions done by the platform and 66% of the corrections visible in the 10-minute-frequency timeseries. Moreover, 46% of videos is misclassified in Figure 2D. It is hence important to develop some methods to reduce this errors metrics.

1. **Method A: from negative views to corrections.** Our first attempt to reconstruct the real corrections c_h^i focuses on the hours that present negative views. In those hours we approximate the real corrections done by YouTube with the negative views plus a quantity minimizing the distance with real corrections. Among different tested quantities (such as the average and the minimum number of views over a time window with width varying from 1 to 12 hours), the best one turned out to be the minimum number of views in the 5 hours preceding and following hour h:
$$m_h^i = \min_{k\in\{h-5,\dots,h+5\};k\neq h}(\hat{v}_k^i),$$
Hence the estimated corrections become
$$c_h^i \sim \left(-\hat{v}_h^i + m_h^i\right)\delta_{(\hat{v}_h^i<0)\&(m_h^i>0)}. \quad (1)$$

Method A reduces the mean distance error from 0.24 to 0.21 and the fraction of *hidden corrections* (i.e. $\sum_{h,i}(c_h^i - \hat{c}_h^i)\delta_{c_h^i > \hat{c}_h^i} / \sum_{h,i}(c_h^i)$) from 66% to 58% with only 0.6% of *wrongly introduced corrections* (i.e. $\sum_{h,i}(c_h^i - \hat{c}_h^i)\delta_{c_h^i < \hat{c}_h^i} / \sum_{h,i}(c_h^i)$).

2. **Method B: detection of anomalous hours.** Although Method A adjusts negative views to estimate the platform corrections, it does not attempt to detect correction events that may have occurred when the observed views $\hat{v}_h^i$ are nonnegative. Hence we developed a method meant to detect anomalies in the views evolution and attribute them to concealed corrections.

   For each video, we define a set of anomalous hours as those hours between 3 and 6 p.m. in which the number of views is the minimum in a time-window starting at midday and ending at midnight. The choice of looking for anomalous hours only within a predefined interval, as well as the choice of the time-window, have been optimized to minimize the misclassification error. These optimal choices are consistent with the observation that the majority of interventions takes place at 3, 4, 5 and 6 p.m.: this fact is true both in the main dataset (Fig. 1B) and in the 10-minute-frequency collection. Over these anomalous hours we approximate the corrections as the difference between m_h^i, as defined in Method A, and $\hat{v}_h^i$, if $m_h^i > \hat{v}_h^i$. Method B reduces the misclassification from 46% to 32% and the number of hidden interventions from 65% to 53%.

5 Acknowledgments

This research has been supported by CNRS through the 80 PRIME MITI project "Disorders of Online Media" (DOOM) and by ANR through grant 19-P3IA-0003. A major acknowledgment goes to Yoan Dinkov and Preslav Nakov of the Qatar Computing Research Institute (QCRI) for their help in collecting YouTube data. A special thanks goes to Bilel Benbouzid for the precious discussions on the French YouTube landscape.

References

[1] API reference | YouTube data API. Available at: https://developers.google.com/YouTube/v3/docs [Accessed July 12, 2021].

[2] Fake YouTube views cut by 2 billion as google audits record companies' video channels, 2012. Available at: https://www.huffpost.com/entry/fake-YouTube-views-cut-google-audit_n_2380848 [Accessed July 12, 2021].

[3] BASTOS, M. T., AND MERCEA, D. The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review 37*, 1 (2019), 38–54.

[4] BESSI, A., AND FERRARA, E. Social bots distort the 2016 U.S. presidential election online discussion. *First Monday 21*, 11 (2016).

[5] 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 *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12* (2012), ACM Press, p. 1186.

[6] 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 *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2012), KDD '12, Association for Computing Machinery, p. 1186–1194.

[7] BOSHMAF, Y., MUSLUKHOV, I., BEZNOSOV, K., AND RIPEANU, M. Key challenges in defending against malicious socialbots. In *Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats* (2012), LEET'12, USENIX Association, p. 12.

[8] BOSHMAF, Y., MUSLUKHOV, I., BEZNOSOV, K., AND RIPEANU, M. Design and analysis of a social botnet. *Comput. Netw. 57*, 2 (2013), 556–578.

[9] CASTALDO, M. anonymizedTimeseries.csv. *Figshare Dataset* (2021). https://doi.org/10.6084/m9.figshare.15073410.v1.

[10] CASTALDO, M. channelList.csv. *Figshare Dataset* (2021). https://doi.org/10.6084/m9.figshare.14994777.v2.

[11] CASTALDO, M., VENTURINI, T., FRASCA, P., AND GARGIULO, F. Junk news bubbles: Modelling the rise and fall of attention in online arenas. *New Media & Society*, forthcoming (2021).

[12] CHEN, L., ZHOU, Y., AND CHIU, D. M. Fake view analytics in online video services. In *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop* (New York, NY, USA, 2014), Association for Computing Machinery, p. 1–6.

[13] CHEN, Z., JI, C., AND BARFORD, P. Spatial-temporal characteristics of internet malicious sources. In *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications* (Apr. 2008).

[14] CORREIA, P., ROCHA, E., NOGUEIRA, A., AND SALVADOR, P. Statistical characterization of the botnets C&C traffic. *Procedia Technology 1* (Jan. 2012), 158–166.

[15] COVINGTON, P., ADAMS, J., AND SARGIN, E. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (New York, NY, USA, 2016), Association for Computing Machinery, p. 191–198.

[16] DREDGE, S. Google goes to war on 'fraudulent' YouTube video views., 2014. Available at: http://www.theguardian.com/technology/2014/feb/05/YouTube-fake-views-counts-google [Accessed July 12, 2021].

[17] FERRARA, E. Bots, elections, and social media: a brief overview, 2019. Available on https://arxiv.org/abs/1910.01720.

[18] FERRARA, E., VAROL, O., DAVIS, C., MENCZER, F., AND FLAMMINI, A. The rise of social bots. *Communications of the ACM 59*, 7 (2016), 96–104.

[19] GAYLE, D. YouTube cancels billions of music industry video views after finding they were fake or 'dead', 2012. Available at: https://www.dailymail.co.uk/sciencetech/article-2254181/YouTube-wipes-billions-video-views-finding-faked-music-industry.html [Accessed July 12, 2021].

[20] GILLESPIE, T. #trendingistrending. When algorithms become culture. In *Algorithmic Cultures: Essays on Meaning, Performance and New Technologies*, R. Seyfert and J. Roberge, Eds. Routledge, London and New York, 2016, pp. 52–75.

[21] GOOGLE - YOUTUBE TERMS OF SERVICE. Fake engagement policy YouTube help. Available at: https://support.google.com/YouTube/answer/3399767 [Accessed July 12, 2021].

[22] GOOGLE - YOUTUBE TERMS OF SERVICE. How engagement metrics are counted. Available at: https://support.google.com/YouTube/answer/2991785?hl=en%E2%80%8B [Accessed July 12, 2021].

[23] HOFFBERGER, C. YouTube strips universal and sony of 2 billion fake views, 2012. Available at: https://www.dailydot.com/unclick/YouTube-universal-sony-fake-views-black-hat/ [Accessed July 12, 2021].

[24] KAMINSKA, I. The real-world cost of YouTube's fake viewers, 2015. Available at: https://www.ft.com/content/7a5d4b84-62af-11e5-9846-de406ccb37f2 [Accessed July 12, 2021].

[25] LEE, K., CAVERLEE, J., AND WEBB, S. Uncovering social spammers: social honeypots + machine learning. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (Geneva, Switzerland, 2010), ACM Press, p. 435.

[26] LIPTON, E., SANGER, D. E., AND SHANE, S. The perfect weapon: How Russian cyberpower invaded the U.S., 2016.

[27] LLEWELLYN, C., CRAM, L., FAVERO, A., AND HILL, R. L. For whom the bell trolls: Troll behaviour in the Twitter Brexit debate. *JCMS: Journal of Common Market Studies 57*, 5 (2019), 1148–1164.

[28] MARCIEL, M., CUEVAS, R., BANCHS, A., GONZÁLEZ, R., TRAVERSO, S., AHMED, M., AND AZCORRA, A. Understanding the detection of view fraud in video content portals. In *Proceedings of the 25th International Conference on World Wide Web* (Republic and Canton of Geneva, CHE, 2016), International World Wide Web Conferences Steering Committee, p. 357–368.

[29] METAXAS, P. T., AND MUSTAFARAJ, E. From obscurity to prominence in minutes: Political speech and real-time search. In *Proceedings of the 2nd international Web Science Conferences* (2010).

[30] NAGARAJA, S., AND SHAH, R. Clicktok: Click fraud detection using traffic analysis. In *Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks* (New York, NY, USA, 2019), WiSec '19, Association for Computing Machinery, p. 105–116.

[31] PFEIFFENBERGER, P. Keeping YouTube views authentic., 2014. Available at: https://security.googleblog.com/2014/02/keeping-YouTube-views-authentic.html [Accessed July 12, 2021].

[32] PINTO, H., ALMEIDA, J. M., AND GONÇALVES, M. A. Using early view patterns to predict the popularity of YouTube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2013), WSDM '13, Association for Computing Machinery, p. 365–374.

[33] PINTO, H., ALMEIDA, J. M., AND GONÇALVES, M. A. Using early view patterns to predict the popularity of YouTube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2013), WSDM '13, Association for Computing Machinery, p. 365–374.

[34] QUINN, B. Google charges advertisers for fake YouTube video views, say researchers., 2015. Available at: https://www.theguardian.com/technology/2015/sep/23/google-advertisers-fake-YouTube-video-views-adwords-bot [Accessed July 12, 2021].

[35] RATKIEWICZ, J., CONOVER, M., MEISS, M., GONÇALVES, B., PATIL, S., FLAMMINI, A., AND MENCZER, F. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web* (2011), Association for Computing Machinery, pp. 249–252.

[36] RATKIEWICZ, J., CONOVER, M. D., MEISS, M., FLAMMINI, A., AND MENCZER, F. Detecting and tracking political abuse in social media. In *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media (ICWSM'11)* (2011).

[37] RHEAULT, L., AND MUSULAN, A. Efficient detection of online communities and social bot activity during electoral campaigns. *Journal of Information Technology & Politics 0*, 0 (2021), 1–14.

[38] ROGERS, R. Otherwise engaged: Social media from vanity metrics to critical analytics. *International Journal of Communication 12*, 732942 (2018), 450–472.

[39] ROTH, C., MAZIERES, A., AND MENEZES, T. Tubes & bubbles. Topological confinement of YouTube recommendations. *PLoS ONE 15*, 4 (2020).

[40] SOLDO, F., ARGYRAKI, K., AND MARKOPOULOU, A. Optimal source-based filtering of malicious traffic. *IEEE/ACM Transactions on Networking 20*, 2 (Apr. 2012).

[41] SUBRAHMANIAN, V., AZARIA, A., DURST, S., KAGAN, V., GALSTYAN, A., LERMAN, K., ZHU, L., FERRARA, E., FLAMMINI, A., AND MENCZER, F. The DARPA Twitter bot challenge. *Computer 49*, 06 (jun 2016), 38–46.

[42] SZABO, G., AND HUBERMAN, B. A. Predicting the popularity of online content. *Commun. ACM 53*, 8 (2010), 80–88.

[43] SZABO, G., AND HUBERMAN, B. A. Predicting the popularity of online content. *Commun. ACM 53*, 8 (Aug. 2010), 80–88.

[44] VENKATARAMAN, S., BLUM, A., SONG, D., SEN, S., AND SPATSCHECK, O. Tracking dynamic sources of malicious activity at internet-scale. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2009), NIPS'09, Curran Associates Inc., p. 1946–1954.

[45] VENTURINI, T. From fake to junk news, the data politics of online virality. In *Data Politics: Worlds, Subjects, Rights*, D. Bigo, E. Isin, and E. Ruppert, Eds. Routledge, London, 2019, pp. 123–144.

[46] XIE, Y., YU, F., ACHAN, K., PANIGRAHY, R., HULTEN, G., AND OSIPKOV, I. Spamming botnets: Signatures and characteristics. *SIGCOMM Comput. Commun. Rev. 38*, 4 (Aug. 2008), 171–182.

[47] XIE, Y., YU, F., ACHAN, K., PANIGRAHY, R., HULTEN, G., AND OSIPKOV, I. Spamming botnets: Signatures and characteristics. In *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication* (New York, NY, USA, 2008), Association for Computing Machinery, p. 171–182.

[48] YANG, J., KELLER, F., SCHOCH, D., AND STIER, S. How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. In *Proceedings of the International AAAI Conference on Web and Social Media* (2017).

[49] YU, H., XIE, L., AND SANNER, S. The lifecyle of a YouTube video: Phases, content and popularity. In *Ninth International AAAI Conference on Web and Social Media* (2015).

[50] ZHANG, J., ZHANG, R., ZHANG, Y., AND YAN, G. On the impact of social botnets for spam distribution and digital-influence manipulation. In *Proceedings of the 2013 IEEE Conference on Communications and Network Security, CNS 2013* (2013), IEEE Computer Society, pp. 46–54.

[51] ZHAO, Z., HONG, L., WEI, L., CHEN, J., NATH, A., ANDREWS, S., KUMTHEKAR, A., SATHIAMOORTHY, M., YI, X., AND CHI, E. Recommending what video to watch next: A multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems* (New York, NY, USA, 2019), Association for Computing Machinery, p. 43–51.

[52] ZHOU, R., KHEMMARAT, S., AND GAO, L. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (2010), Association for Computing Machinery, p. 404–410.

[53] ZHOU, R., KHEMMARAT, S., GAO, L., WAN, J., AND ZHANG, J. How YouTube videos are discovered and its impact on video views. *Multimedia Tools Appl. 75*, 10 (May 2016), 6035–6058.