This Account Doesn't Exist: Tweet Decay and the Politics of Deletion in the Brexit Debate

Marco Bastos
University College Dublin
City, University of London

Abstract
Literature on influence operations has identified metrics that are indicative of social media manipulation, but few studies have explored the lifecycle of low-quality information. We contribute to this literature by reconstructing nearly 3M messages posted by 1M users in the last days of the Brexit referendum campaign. While previous studies have found that on average only 4% of tweets disappear, we found that 33% of the tweets leading up to the referendum vote are no longer available. Only about half of the most active accounts that tweeted the referendum continue to operate publicly and 20% of all accounts are no longer active. We tested whether partisan content was more likely to disappear and found more messages from the Leave campaign that disappeared than the entire universe of tweets affiliated with the Remain campaign. We compare these results with an assorted set of 45 hashtags posted in the same period and find that political campaigns present much higher ratios of user and tweet decay. These results are validated by inspecting 2M Brexit-related tweets posted over a period of nearly 4 years. The article concludes with an overview of these findings and recommendations for future research.

Keywords
Manipulation; Disinformation; Misinformation; Twitter; Web archive; Brexit

Highlights
- 33% of the tweets leading up to the referendum vote have been removed
- Only about half of the most active accounts that tweeted the referendum continue to operate publicly
- 20% of the accounts are no longer active and 20% of these accounts were blocked by Twitter
- While Twitter suspended fewer than 5% of all accounts, they posted nearly 10% of the entire conversation about the referendum
- There are more messages from the Leave campaign that have disappeared than the entire universe of tweets affiliated with the Remain campaign

Introduction
The backdrop of influence operations and social media manipulation that emerged during key electoral events in 2016 constitute a challenge that policymakers and researchers continue to grapple (Bastos & Mercea, 2019; Bessi & Ferrara, 2016; Ferrara, 2017; Weedon et al., 2017). The open infrastructure of social platforms, whence the cornerstone of networked publics, has rapidly evolved into increasingly opaque social infrastructures largely unaccountable to the public. While the effectiveness of influence operations is crucially dependent on social tensions and biases that can be exploited with divisive information, the relatively sudden emergence of this ecosystem of information warfare is equally reliant on the attention economy and the social media supply chain underlying social platform's business model (Walker et al., 2019).

Social media platforms have ramped up efforts to flag false amplification (Weedon et al., 2017), remove "fake accounts" (Twitter, 2018b), and prevent the use of highly optimized and targeted political messages on users (Dorsey, 2019). These efforts sought to clear social platforms from "low-quality content," a broad definition including user accounts, posts, and weblinks to content selected for removal. The removal of user accounts and posts constitutes an important line of action social platforms undertake in their struggle against influence operations, misinformation, false or fabricated news items, spam, and user-generated hyperpartisan news, with Twitter being the most effective platform at countering manipulation by identifying and removing more inauthentic behavior (Bay & Fredheim, 2019). Along with Facebook, Twitter is also the only platform to have publicly released their community standards defining "problematic content" and the process through which such policies are enforced (Facebook, 2018a, 2018b; Twitter, 2019b).

The large-scale removal of social media content has the problematic drawback of altering the record of social interactions. Unlike traditional media used to distribute propaganda, such as newspapers and posters in the 20th century (Sanders & Taylor, 1982), or pamphlets and leaflets going back as far as the 16th century (Raymond, 2003), the removal of social media content eliminates any trace of the event, thereby preventing forensic analysis and academic research on influence operations targeting social media platforms. While there have been attempts to create public archives of social media posts, these institutional efforts faced considerable challenges and failed to come to fruition (Zimmer, 2015). Similarly, although social platforms have at times offered archives of disinformation campaigns identified and removed by the very platforms (Elections Integrity, 2018), such sanctioned archives offer only a partial glimpse into the extent of influence operations and may prevent researchers from examining organic contexts of manipulation (Acker & Donovan, 2019).

Indeed, influence operations on social platforms, particularly on Facebook and Twitter, continue to rely on coordinated and targeted attacks where the accounts and profiles sourcing the content disappear in the months following the campaign. Some accounts are suspended by the social platforms for violating standards and Terms of Service (Gleicher, 2019), such as posting inappropriate content or displaying bot-like activity patterns; others are deleted by the malicious account holders reportedly to cover their tracks (Owen, 2019). The modus operandi of influence operations often consists of amplifying original hyperpartisan content by large botnets that disappear after the campaign. The emerging thread is then picked up by high-profile partisan accounts that seed divisive rhetoric to larger networks of partisan users and automated accounts (Bastos & Mercea, 2019). These tactics suggest that Tweet Decay is a key metric to identify problematic content, including influence operation and false amplification on social media.

Unfortunately, no study has systematically analyzed the fraction of deleted tweets during political events. Similarly, no study has established whether partisan content is more likely to disappear compared with nonpartisan content in the period leading up to and following political campaigns. To address this gap in the literature, we revisit a large data set of three million tweets posted in the period leading up to the 2016 United Kingdom European Union membership referendum to quantify the extent to which partisan and non-partisan communication related to the referendum is still available. We compare these figures with a range of data sets also recorded in 2016 and to a larger database of Brexit-related tweets encompassing nearly four years of Twitter activity.

Previous work

There is a large body of scholarship exploring the transformative potential of social platforms to complement or substitute deliberative forums supporting civic participation, particularly with respect to access to information, reciprocity of communication, and commercialization of online space (Malina, 2005; Papacharissi, 2002). This scholarship pays tribute to the Habermasian concept of the public sphere, identifying the internet with domains of social life where public opinion could be expressed, and often praising digital communication as a force for democratization and deliberation (Davis et al., 2002). The underlying assumption was that social media would offer the public accord—and perhaps more critically, a public record—of the decision making process underpinning rational deliberation (Papacharissi, 2008). Yet, only limited empirical work has been carried out on the use of social platforms as archives supporting a healthy public sphere (Mylonas, 2017).

The process of verifying whether a social media post remains available after being posted online can be made via http requests or programmatically using social platform's application programming interface (API). Twitter allows developers to retrieve programmatically and at scale (i.e., "rehydrate") the full tweet, user profile, or direct message content using their APIs (Twitter, 2019a), but their Terms of Service state that content deleted by a user or blocked by the platform due to infringements on the ToS ought to disappear from the platform altogether; similarly, deleting a tweet automatically triggers a cascade of deletions for all retweets of that tweet (Twitter, 2018a). This specific affordance of social platforms has of course facilitated the disappearance of posts, images, and weblinks from the public view, with important and negative effects to research on influence operations.

Social platforms rarely disclose content that was flagged for removal, and therefore studying the politics of deletion on social platforms is an exercise in reverse engineering. Conversely, content that has been deleted or blocked from social platforms is likely to fall within the broad category of "low-quality content," and we hypothesize that content decay—tweet and user decay in the context of Twitter—can be used as proxies to examine the extent to which deliberation on social media is hindered by influence operations, including disinformation and other forms of problematic content (Starbird et al., 2019; Starbird et al., 2014). Indeed, previous research has found that hyperpartisan content is marked by a short shelf life, disappearing or significantly changing shortly after being posted online (Bastos & Mercea, 2019), an indication of the perishable nature of digital content at the center of political deliberation (Walker, 2015).

There is surprisingly little research on how social media data sets change when observed at different points in time and how this may impact the results of the analysis. Walker (2015) contrasted data collected from social media platforms in real-time versus data collected minutes, hours, or days after the post went online. McCreadie et al. (2012) explored the effect of this collection decay on the Tweets2011 data, a set of 16 million tweets offered by Twitter and made available to participants of the TREC Conference. The data set, whose tweets cover the period of January 23 to February 8, 2011, was reconstructed using an asynchronous HTTP fetcher (as opposed to Twitter APIs) in 2012. Given the limitations and the unreliability of HTTP crawlers, no precise information was given regarding the ratio of tweets that disappeared within the one-year gap between the data being made available and the reconstruction of the data set.

Bagdouri and Oard (2015) also developed an evaluation design to predict whether a tweet will be deleted within the first 24 hours of being posted. The classifier, which takes into account the distribution of deleted tweets, along with tweet-based and user-based features, reported a very sharp skew in the data, with some users regularly deleting their tweets while others would rarely do so. Bagdouri and Oard (2015) also found that the most prolific deleters were automated systems engaging in advertisement, a known marker of spambots that push low-quality content and are ultimately purged from social platforms. More interestingly, Bagdouri and Oard (2015) reported a remarkably low ratio of tweet decay in the 24h period, with only 3.6% of the messages having been deleted, or 2.2% of the data once retweets were excluded. Finally, they reported that 2% of users were responsible for just above one third of all deletions that were not of retweets.

In related research, Xu et al. (2013) collected a corpus of over 300 thousand bullying-related tweets and estimated the survival rate by querying the URL of the tweet for around two months at regular intervals. The deletion rate found in the data, hypothesized to be a function of user's regret, allowed the construction of a "regrettable posts predictor." Deletion rate was found to decay over time, with a drop-off in deletion rate that was so extreme that the authors could safely exclude deletions occurring after two weeks from the filtered data set without significantly introducing any noise. In the end, the overall fraction of deleted tweets was rather low and similar to those reported by Bagdouri and Oard (2015), at 3.75%. This fraction of deleted tweets of 4% is considerably lower than what we have anecdotally observed in a range of political and protest data sets. In fact, from our experience monitoring Twitter Compliance Firehose we estimate the baseline of tweet deletion to currently stand at around 15%.

We seek to contribute to this literature by calculating the fraction of deleted tweets and accounts in a large data set of 3 million tweets posted in the period leading up to the 2016 United Kingdom European Union membership referendum. Given the polarized and highly partisan nature of the campaign, we hypothesize that (H1) the fraction of deleted tweets will be higher than the 4% reported in the literature. We also hypothesize that, *ceteris paribus*, (H2) partisan messages will present higher decay compared with neutral messages. We validate the tests of H1 and H2 against a range of hashtags that also emerged in 2016, and finally to a larger database of Brexit-related tweets encompassing nearly four years of Twitter activity. Given the relatively long period of time under analysis, and that deletions seem largely concentrated in the 24h of the post going online (Bagdouri & Oard, 2015; Xu et al., 2013), we expect the occurrence of significant political events to affect tweet decay and hypothesize that (H3) tweet decay will be associated with the emergence of significant political events.

Data and Methods

We relied on Twitter's Streaming and REST APIs, two of the endpoints offered by Twitter to programmatically collect data, to amass three key databases explored in this study: pre-Brexit, post-Brexit, and non-Brexit data sets. The pre-Brexit data set includes a total of 8,821,116 tweets collected using a set of keywords and hashtags, including relatively neutral tags (e.g., referendum, inorout, and euref), hashtags supporting the Leave campaign (e.g., voteleave, leaveeu, takecontrol, voteout, and beleave), and tags clearly aligned with the Remain campaign (e.g., strongerin, leadnotleave, voteremain, moreincommon, and lovenotleave). The hashtags were parsed across multiple querying pools to avoid API filtering. Queries that exceeded the one-percent threshold were parsed across separate queries, cumulatively requiring a combination

of twelve independent calls to the Streaming API. The pre-Brexit database was ultimately sampled to the 13 days leading up to the referendum vote, totaling 2,775,789 tweets—1,742,756 of which are retweets.

The partisan affiliation of users in this data set was identified based on the usage of hashtags clearly aligned to one of the referendum campaigns. The mean affiliation is only calculated for users that actively tweeted referendum hashtags. As such, @-mentioned or retweeted users are only identified as Leaver or Remainer if the user in question tweeted or retweeted a separate message with hashtags clearly aligned with one side of the campaign. This conservative approach renders most media outlets and high-profile accounts as neutral and has the added benefit of filtering out retweets or @-mentions intended as provocation or ironic remark; these messages are offset by the broader ideological orientation tweeted by the account, and users that have only sourced information or received @-mentions are classified as neutral for not having themselves tweeted any partisan hashtag.

We further probed whether ideologically motivated messages were more likely to disappear by classifying tweets that espoused nationalist versus cosmopolitan values, and populist versus economist values. This ideological value space leverages Inglehart and Norris (2016) thesis of economic insecurity versus cultural backlash which arguably accounts for the political realignment in Western political parties. To this end, we relied on a set of 10,000 tweets manually classified along the ideological polarities of Globalism vs Nationalism and Economism vs Populism to train a machine learning algorithm using text vectorization (Selivanov, 2016), an approach purposefully built for text analysis. For each ideological pair, the classifier returns a range of values from 0 (completely globalist and/or economist) to 1 (completely nationalist and/or populist), so that values from 0.45 to 0.55 are somewhere in the middle of this scale and assumed to be relatively neutral (Bastos & Mercea, 2018). We rely on the second data set—the non-Brexit data set—to compare tweet decay in the Brexit data versus non-political, generic hashtags that also trended in 2016. The non-Brexit data set includes 45 hashtag (see Figure 2 for details) and a random sample of one thousand tweets was rehydrated to verify if user accounts and tweets were still available in the platform.

The third and last is the post-Brexit data set which was derived by querying a database of 100 million Brexit-related tweets, starting in the period leading up to the referendum campaign and ending in October 2019. This database includes 43 months of Brexit-related messages. We take a sample without replacement of 50,000 tweets per month, an approach similar to the constructed week sampling employed in journalism studies that maximizes generalizability beyond consecutive days and is suitable for estimating content for a six-month period or longer (Riffe et al., 1993). The monthly sampling approach returned a data set of 2,150,000 tweets (of which 1,404,704 are retweets), subsequently rehydrated to estimate the fraction of deleted tweets and user accounts in the data for each of the 43 months that followed the referendum campaign.

A computer script was written to programmatically query Twitter REST API for user accounts and tweet IDs (rehydration). These steps allowed us to calculate the tweet decay coefficient for Twitter data sets at scale. We relied on this program to validate the results on a range of hashtags with several thousand tweets posted in the same period of the official campaign of the UK EU membership referendum, which took place in the first half of 2016. For each hashtag, we take a

random sample without replacement of one thousand tweets in the data set and query Twitter API to verify whether the message is still available and the account (by user ID) that sourced the content remains active.

Finally, identifying temporal trends in the presence of anomalies is a non-trivial task for anomaly detection, so we relied on the Seasonal Hybrid ESD (S-H-ESD) algorithm to discover statistically meaningful anomalies in the input time series of deleted and active tweets (Vallis et al., 2014). S-H-ESD employs time series decomposition and Generalized Extreme Studentized Deviate (ESD) to test for meaningful anomalies in temporal data with inherent seasonal and trend components, such as timestamped social network transaction data. This approach builds upon the Generalized ESD test for detecting outliers introduced by Rosner (1983) to identify global and local anomalies. The algorithm supports long time series such as the one explored in this study and employs piecewise approximation to identify both positive and negative anomalies in input time series (point-in-time increase versus decrease in tweets), which is important as we are interested both in the upsurge and decline of the fraction of tweets deleted in the observed time period.

Results

We approached H1 by exploring the period of 13 days leading up to the referendum vote in June 23, 2016. A total of 2,775,789 messages were tweeted by 792,663 users, thereby averaging around 3.5 messages per user. Social media activity presents a long-tailed distribution and Twitter is no exception to this: 169 users tweeted over 500 messages and 28 posted more than one thousand messages. The official campaign accounts @ivotestay and @ivoteleave posted 15,928 and 11,647 tweets in this 13-day period, respectively, and since the referendum both accounts have been suspended by Twitter. Indeed, we manually checked the 100 most active user accounts and 37 are no longer active: 16 have been suspended and 21 no longer exist (deleted account). From the remaining 63 active accounts, 3 were recreated after the referendum and 2 have been set to private. In short, only about half of the most active accounts in the referendum debate continue to operate publicly.

This trend is not restricted to hyperactive users. We queried Twitter REST API and the web interface to check which user accounts remained active in October 2019, three years after the vote. For the universe of 792,663 users that tweeted the referendum between 12-24 June, 20% or 155,157 accounts are no longer active. These numbers are consistent whether checking by username or user ID, with only 118 mismatches (false positive or negative) when querying Twitter REST API by usernames instead of user ID. Twitter REST API does not inform if the accounts have been deleted, blocked, suspended, removed, or set to private. We therefore checked the web interface of each of these accounts to identify accounts that have been suspended by Twitter. Twenty percent of the accounts that are no longer active have been blocked by Twitter (n=36,159).

The figures are yet more dramatic when we inspect the share of messages that are no longer available on Twitter, whether on Enterprise, Search, or REST APIs, or via web interface. Twenty-two percent or 631,700 tweets are no longer available due to the removal of the seeding user and the consequences of the Terms of Service governing content authored by deleted accounts. Posts from deleted accounts are retrospectively removed from Twitter and generate orphaned data, but this number only accounts for messages that are no longer available due to the

user account that posted the content no longer being active. We rehydrated the data to account for this difference, thereby identifying messages that were actively deleted by users, as well as retweet cascades that disappeared due to the original seeding post no longer being available in the platform.

One third of the near 3 million tweets posted in the period are no longer available (*n*=932,815), with only 1,842,974 tweets remaining available. In other words, 33% of the tweets that shaped the discussion about the referendum are no longer retrievable three years after the vote and nearly half of this universe of messages disappeared because the seeding account was removed, blocked, deleted, suspended, or set to private. One-fifth of the tweets that are no longer available disappeared because the seeding account was suspended from Twitter. This figure (*n*=203,681) includes not only tweets authored by these accounts, but also intermediary points in retweet cascades that vanished because the user account was no longer available. Suspended accounts were particularly prolific: while less than 5% of users were suspended, these accounts posted nearly 10% of the entire conversation about the referendum on Twitter.

Figure 1 unpacks the fraction of deleted tweets and accounts: around 20% of users (or 155,157 out of 792,663) are no longer active and again 20% of this cohort of accounts have been blocked by Twitter (or 36,159 out of 155,157). It would be informative to know more about the remainder 118,998 accounts that are no longer active. These may be accounts that have been compromised, utilized in influence operations, or that violated the Twitter's Terms of Service, but unfortunately only the information available in the API can be made public due to the constraints of their Privacy Policy (Twitter, 2019c). This is a considerable limitation, as this small cohort of accounts posted nearly 10% of the entire conversation about the referendum on Twitter. As such, it is conceivable that this small cohort of removed accounts may have sourced low-quality content both at speed and scale. The results are nonetheless consistent with hypothesis H1, as the fraction of deleted tweets is considerably higher than the average of 4% reported in the literature.

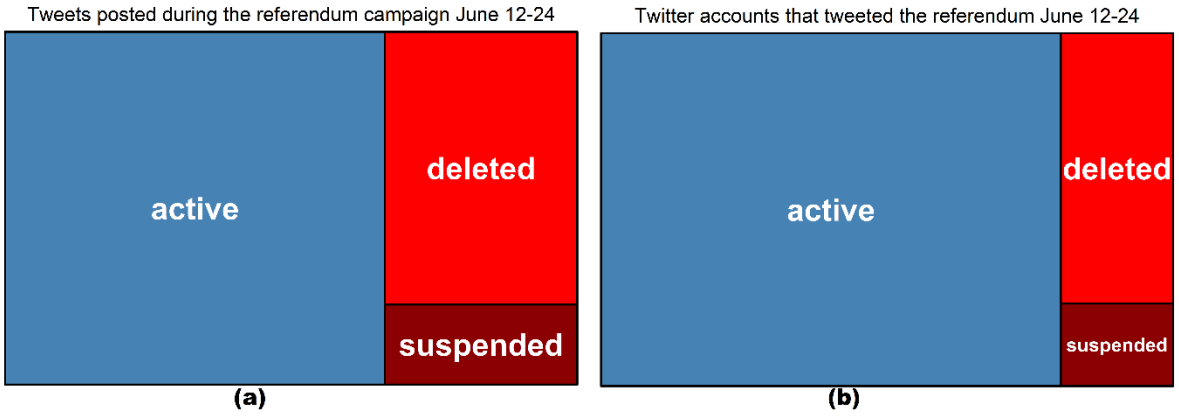

Figure 1: (a) 33% of the tweets leading up to the referendum vote are no longer retrievable; (b) 20% of the user accounts that tweeted the referendum are no longer active

Our second hypothesis states that partisan messages are likely to present higher decay compared with neutral messages. Both #voteleave and #voteremain account for a significant share of the hashtags in the data, with the former appearing in 15% of the tweets and the latter in just under

10%. The hashtag #voteleave is however significantly more likely to appear in tweets that are no longer available. Indeed, 20% of the deleted messages espoused this term compared with the regular 10% for #voteremain. No other hashtag presents such a large difference between the frequency observed in the population of deleted tweets compared with the frequency observed in the entire population of messages. The point difference observed for #voteleave is comparable to the point difference of the remainder 700 most popular hashtags combined (x=.064 for both groups).

One important caveat is that the vote Leave campaign accounts for a larger share of the tweets due to the prevalence of popular Leave hashtags. We therefore controlled for the higher number of tweets and users associated with Leave, but Leave users continued to be more likely to be deleted, blocked, or removed from Twitter. Similarly, the tweets posted by these accounts are considerably more likely to decay. Nearly 40% of the messages posted by Leavers are no longer available whereas the fraction of deleted tweets for Remainers is under 30%. If we assign the hashtag #brexit to the Leave campaign, we find that more messages from the Leave campaign disappeared from Twitter than the entire universe of messages affiliated with Remain: 468,419 tweets disappeared from a total of 1,224,568 messages posted by Leavers, and 130,245 posts are no longer available from a universe of 438,359 messages posted by Remainers.

As reported in previous research (Bastos & Mercea, 2018), the overall sentiment during most of the campaign was decidedly nationalistic, with three quarters of messages having a nationalistic sentiment. Most messages tweeted in the period were additionally preoccupied with economic implications of the decision to leave the E.U. There are however significant differences in the ideological orientation of the Brexit debate when controlling for users and tweets that have since disappeared. While the debate is predominantly focused on economic issues, most of messages that are no longer available (n= 932,815) largely appeal to populist slogans ($\bar{x}$=.52 and $\tilde{x}$=.50) compared with the subset of messages that remain active ($\bar{x}$=.44 and $\tilde{x}$=.43). This substantial difference is again observed in the ideological polarity Globalism vs Nationalism. The debate is indeed predominantly nationalistic, but it is significantly more so in the subset of messages that are no longer available: $\bar{x}$=.64 and $\tilde{x}$=.67, compared with $\bar{x}$=.59 and $\tilde{x}$=.63 for the subset of messages that remained active.

To further probe H2, we rehydrate a range of hashtags with several thousand tweets posted in the period of the official campaign of the UK EU Membership Referendum, which ran in the first half of 2016. For each hashtag, we take a random sample of one thousand tweets in the data set and query Twitter API to verify if the message is still available and whether the account (by user ID) that sourced the content remains active. A total of 45 hashtags were queried in 5 groups of 9 hashtags for the Remain[1] and the Leave[2] campaigns, 9 non-partisan hashtags discussing the referendum[3], 9 nonpolitical hashtags[4] that trended in the period, and 9 hashtags dedicated to protest activism[5] causes (Figure 2).

[1] Group 1: *betteroffin, bremain, leadnotleave, lovenotleave, moreincommon, strongerin, votein, voteremain, yes2eu*
[2] Group 2: *beleave, betteroffout, britainout, leaveeu, loveeuropeleaveeu, no2eu, notoeu, voteleave, voteout*
[3] Group 3: *brexit, brexitornot, antibrexit, euref, eureform, eurefresults, jocoxrip, lovelikejo, referendum*
[4] Group 4: *mozfest, nsmnss, rstats, agchat, cadrought, foodsystem, homegrown, phdchat, usaid*
[5] Group 5: *15maydebout, 7mdebout, bruxellesdebout, globaldebout, nobillnobreak, nuitdebout, romadebout, blacklivesmatter,* and *lesvos*.

Group 4 identified 15% as the deletion baseline for hashtagged tweets in 2016 (Q_1=.10, $\bar{x}$=.15, $\tilde{x}$=.15, Q_3=.19), with this baseline sharply increasing as the topic of the conversation becomes more contentious. Protest activism hashtags presented a tweet decay of nearly 30% for 15maydebout and 44% for blacklivesmatter, a hashtag campaign notable for being targeted by the Russian IRA operations (Bastos & Farkas, 2019). Brexit related discussions also verge around 30% (Q_1=.24, $\bar{x}$=.28, $\tilde{x}$=.29, Q_3=.32), a figure that is not too different from what was observed in openly partisan hashtags associated with the Remain campaign (Q_1=.23, $\bar{x}$=.30, $\tilde{x}$=.26, Q_3=.32). Openly partisan hashtags associated with the Leave campaign, however, present a worrying and much higher coefficient of tweet decay (Q_1=.33, $\bar{x}$=.42, $\tilde{x}$=.42, Q_3=.50).

Indeed, three quarters of the content hashtagged with #betteroffout has been removed from Twitter. More than half of the tweets hashtagged with #voteleave, the official campaign to leave the EU, is no longer available. Figure 2 unpacks the differences across classes of hashtags. The size of the line indicates the discrepancy between account deletion rate and tweet deletion rate. In other words, the longer the line, the higher the fraction of deleted tweets relative to the fraction of deleted accounts. This can be caused by users regretting and deleting the original post, or else changing their account to private, in which case the original tweet is no longer available even though the account remains operational.

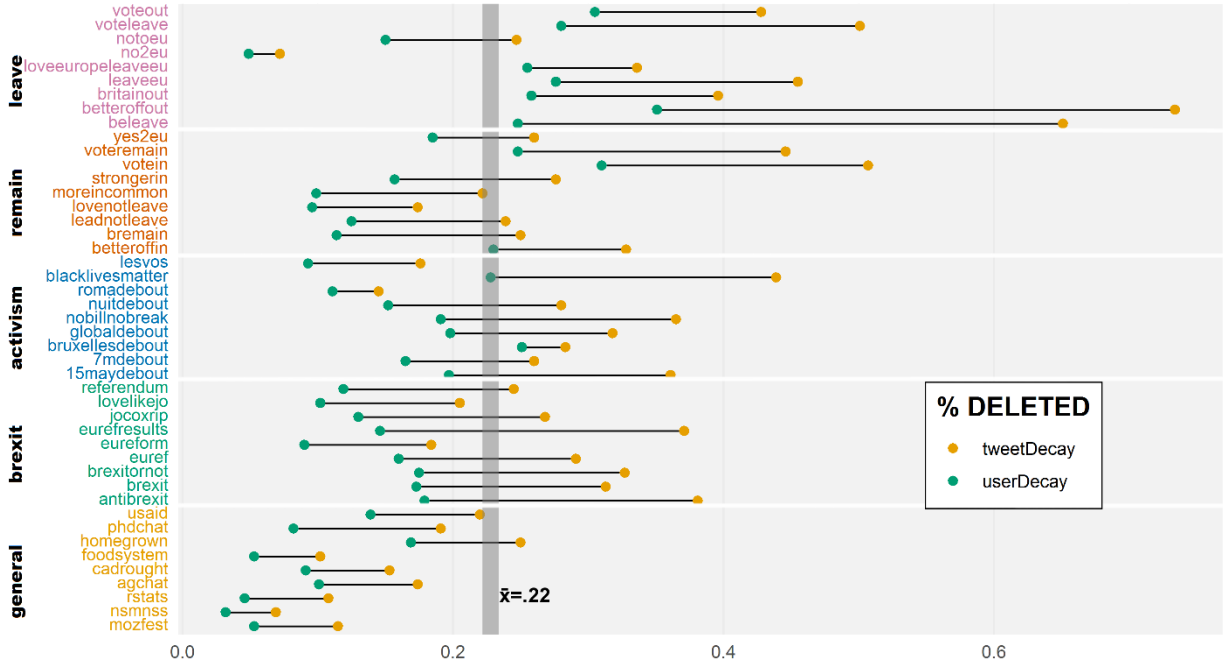

Figure 2: Green dots: fraction of deleted accounts (percentage points). Yellow dots: fraction of deleted tweets (percentage points). The tweet deletion rate for hashtags *betteroffout* and *beleave* are 74% and 65%, respectively. Vertical grey bar shows the mean deletion rate for this set of forty-five hashtags.

These results are consistent with hypothesis H2, which stated that partisan messages were to present higher decay compared with neutral messages. We validate this analysis by parsing the hashtags featured in the pre-Brexit data set and calculate the deletion rate per tag. Tweets hashtagged with terms associated with the Leave campaign are considerably more likely to have been deleted. In fact, the list of tags tweeted over 1 thousand times with a deletion rate of 40% or

higher includes mostly Leave hashtags, with online a few terms not clearly aligned with either side of the campaign: *voteleave*, *votein*, *leaveeu*, *ivoted*, *voteout*, *beleave*, *cameron*, *inorout*, *ukip*, and *eng*. For this set of hashtags, most of the messages tweeted in the period leading up to the vote are no longer available ($\bar{x}$=52%).

Lastly, we probe hypothesis H3 by testing the extent to which tweet decay is dependent on the timeframe of the Brexit negotiations and deliberation. To this end, we queried a database of 100 million Brexit-related tweets posted by users based in Britain. The database encompasses the official campaign period for the 2016 referendum campaign until October 2019 and includes 43 months of Brexit-related messages. We take a sample without replacement of 50,000 tweets per month, therefore generating a data set of 2,150,000 tweets. We proceed by querying Twitter REST API to check whether users and tweets are still active in the platform. This approach allowed us to calculate a coefficient for tweet decay that is representative for each of the 43 months that followed the referendum campaign.

Figure 3 shows that the fraction of messages that remained active is constant throughout the period, while the ratio of deleted messages varies considerably and seems to follow contentious moments of the Brexit saga. There is considerable increase in tweet decay in the weeks leading up to the referendum vote—from 19% in April up to 33%, as discussed above. Tweet decay recedes after the ballot and resumes as pressure mounts for triggering Article 50. This is when the monthly fraction of deleted tweets peaks from 27% to one-third in the last week. The fraction of deleted tweets stabilizes in the wake of triggering Article 50 and only escalates, albeit mildly, when Theresa May announces the Chequers Plan, after which the monthly fraction of deleted tweets is again around one-fifth. In the following months tweet decay decreases steadily. The fraction of deleted tweets only becomes similar to those reported in the literature after Boris Johnson becomes Prime Minister, when tweet decay is the lowest at 7%.

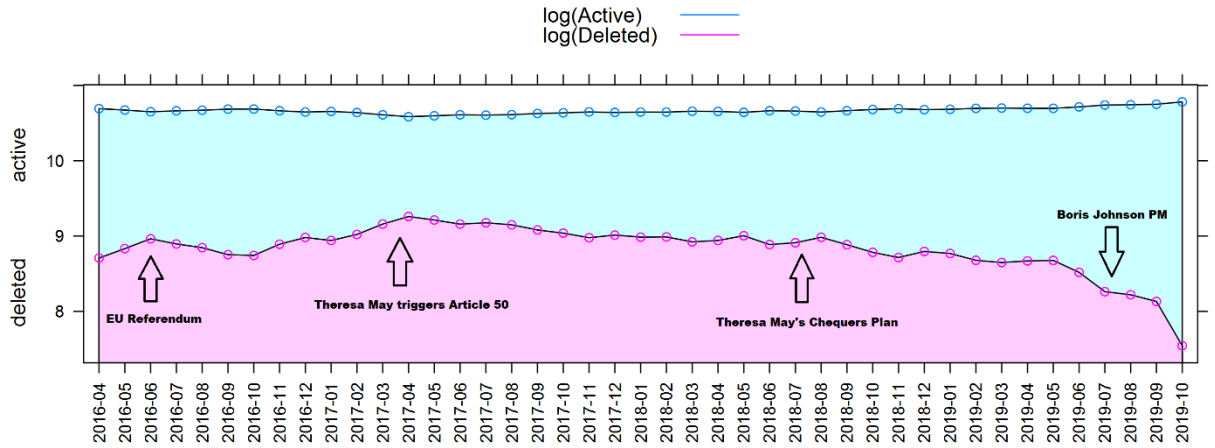

Figure 3: Monthly deleted vs active Brexit-related tweets (log) from April 2016-October 2019 (*N*=2,150,000)

These anomalies are promptly detected by the Seasonal Hybrid ESD (S-H-ESD) algorithm (Vallis et al., 2014). Anomalies are found in the series of active and deleted tweets on the date of the referendum and the triggering Article 50, which is unsurprising given the increase of activity reflecting these key dates in the Brexit calendar. There are nevertheless significant global

anomalies in the series of deleted tweets that are not present in the series of active tweets. These can be pinpointed to the large volumes of tweets that disappeared after key dates. Indeed, a significant number of messages disappear three weeks after Article 50 was triggered, and again another significant upsurge in tweet decay is registered the day after the snap general election (8 June 2017). None of above anomalies are registered in the regular series of tweets that remained active, which is consistent with the assumption that tweet decay is associated with politically contentious contexts (H3). In fact, most of the anomalies in the series of active tweets are registered in late 2019, when Boris Johnson becomes Prime Minister and Parliamentary Elections are called for 12 of December 2019. Figure 4 unpacks these differences, showing the temporal series of active and deleted tweets and the detected anomalies.

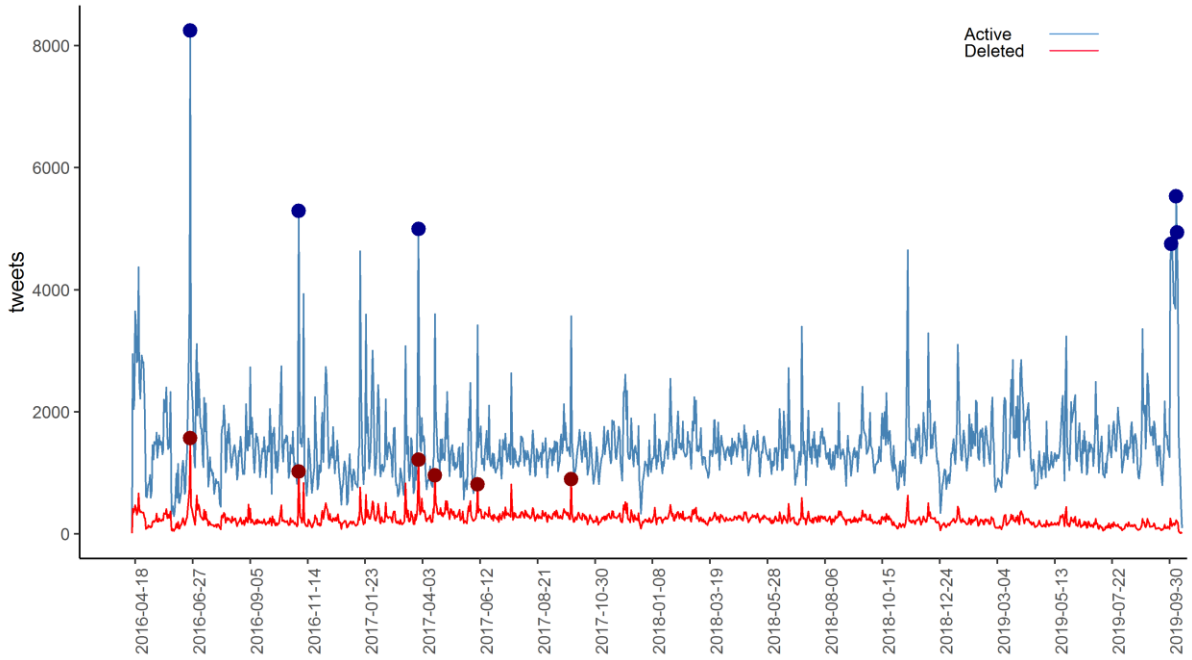

Figure 4: Time series of active and deleted tweets with anomalies identified with by the S-H-ESD algorithm

Conclusion

The results presented in this study caution against the use of social platforms as a complement or extension to deliberative forums, as the public record of social media interactions, or at least considerable portions of it, is subjected to being removed from public scrutiny. This is particularly problematic if one assumes social platforms may offer a substitute for deliberative forums of civic participation. The theoretical import of the Habermasian perspective to rational public deliberation assumes that the collective decision-making on issues of public concern resolves conflicts based on the quality of the argument and the evidence supporting it. The results reported in this study are largely at odds with these stipulations. While the volume of activity on social media may be directly linked to developments unfolding over time (Bastos et al., 2015), the results show that a significant portion of the political debate on Twitter is not designed for permanence.

Ephemerality is perhaps an expected affordance of social media communication, but it is not an expected design of political communication and deliberation across social platforms. The

disappearance of one third of the discussion underpinning a key event in contemporary politics indicates that the fraction of deleted tweets may be a proxy for manipulation and disinformation. As much of the deleted content resulted from Twitter actively blocking user accounts, thereby generating orphaned data, it is conceivable that the Brexit debate may have been subjected to considerable volumes of low-quality information whose distribution often resorts to artificial manipulation and false amplification (Walker et al., 2019). In other words, while the ephemerality of social media posts may be a reasonable expectation, this poses considerable challenges for informed public deliberation around matters where the issue being deliberated on is constantly disappearing from public scrutiny.

Unfortunately, the identification of removed content and user accounts entails computational routines that cannot be implemented in real time, as there are multiple triggers that may block or delete an account and post from social platforms. Influence operation may conceivably exploit these limitations by offloading problematic content that is removed from platforms before the relentless—though time consuming—news cycle has successfully corrected the narratives championed by highly volatile social media content. This process could be described as the involuntary but spontaneous gaslighting of social platforms: the low persistence and high ephemerality of social media posts are leveraged to transition from one contentious and unverified political frame to the next before mechanism for checking and correcting false information are in place.

Influence operations can daisy chain multiple disinformation campaigns that are phased out and disappear as soon as rectifying information or alternative frames starts to emerge. The politics of deletion can thus be leveraged in propaganda campaigns centered around the firehose of falsehood model (Paul & Matthews, 2016), where a large number of messages are broadcast rapidly, repetitively, and continuously over multiple channels without commitment to consistency or accuracy (Bertolin, 2015). The high volume posting of social media messages can be effective because individuals are more likely to be persuaded if a story, however confusing, appears to have been reported repetitively and by multiple sources. Traditional counterpropaganda methods tend to be infective against this technique. Similarly, fact-checking social media posts that have disappeared is not technically possible and perhaps not desirable either.

In other words, while social media content may be fundamentally ephemeral, the fraction of deleted tweets reported in this study was at times disturbingly high and prevented further analyses of the accounts that seeded the content, as once users are removed from the platform no further information can be gleaned from the account. The implications of observing such high decay in dynamic content at the center of political discussion certainly warrant further research. For one, it is important to establish the cutoff point after which political content is likely to disappear from social platforms. While previous studies show that decay is associated with time, our results show that political and especially contentious messages are more likely to disappear from the public record than nonpartisan conversations recorded in the same period.

Further studies should examine the extent to which different forms of political discussion are equally likely to disappear from social platforms, which temporal patterns are indicative of survival, and whether decay is caused by manipulation and influence operations detected by and

ultimately removed from social media by the platforms themselves. This may require research collaboration with social platforms, especially if the objective is to establish whether the decay in social media posts is associated with influence operations and low-quality content, which reportedly presents shorter shelf life compared with organic content.

Another important empirical question that could not be addressed in this study is determining the exact point in time when accounts and tweets sourcing political content are likely to start decaying. While Xu et al. (2013) have found a strong and inverse linear association between the fraction of deleted tweets and time (in one-minute increments), this may apply only to the corpus of bullying tweets studied by the authors. The extreme drop-off in deletion rate reported by the authors, along with the remarkable low ratio of deleted tweets at only 4%, are important indicators that social media decay likely differs across topics and may be substantively higher for conversations targeted by influence operations. In other words, one important empirical question that warrants further research is whether deletion rate is universally and inversely associated with time.

References

Acker, A., & Donovan, J. (2019, 2019/09/19). Data craft: a theory/methods package for critical internet studies. *Information, Communication & Society*, *22*(11), 1590-1609. https://doi.org/10.1080/1369118X.2019.1645194

Bagdouri, M., & Oard, D. W. (2015). On predicting deletions of microblog posts. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management,

Bastos, M. T., & Farkas, J. (2019). "Donald Trump Is My President!": The Internet Research Agency Propaganda Machine. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119865466

Bastos, M. T., & Mercea, D. (2018). Parametrizing Brexit: Mapping Twitter Political Space to Parliamentary Constituencies. *Information, Communication & Society*, *21*(7), 921-939. https://doi.org/10.1080/1369118X.2018.1433224

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38-54. https://doi.org/10.1177/0894439317734157

Bastos, M. T., Mercea, D., & Charpentier, A. (2015, 2015). Tents, Tweets, and Events: The Interplay Between Ongoing Protests and Social Media. *Journal of Communication*, *65*(2), 320-350. https://doi.org/10.1111/jcom.12145

Bay, S., & Fredheim, R. (2019). *Falling Behind: How Social Media Companies Are Failing to Combat Inauthentic Behaviour Online* (NATO StratCom COE, Issue. https://www.stratcomcoe.org/download/file/fid/81308

Bertolin, G. (2015). Conceptualizing Russian Information Operations: Info-War and Infiltration in the Context of Hybrid Warfare. *IO Sphere*, 10-11.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First Monday*, *21*(11).

Davis, S., Elin, L., & Reeher, G. (2002). *Click on democracy: The Internet's power to change political apathy into civic action*. Routledge.

Dorsey, J. P. (2019, 30 October 2019). We've made the decision to stop all political advertising on Twitter globally. *Twitter*. https://twitter.com/jack/status/1189634360472829952

Elections Integrity. (2018). *Data archive* https://about.twitter.com/en_us/values/elections-integrity.html

Community Standards, (2018a). https://www.facebook.com/help/975828035803295

Understanding the Facebook: Community Standards Enforcement Report, (2018b). https://fbnewsroomus.files.wordpress.com/2018/05/understanding_the_community_standards_enforcement_report.pdf

Ferrara, E. (2017, 2017-07-31). Disinformation and social bot operations in the run up to the 2017 French presidential election. *2017*. https://doi.org/10.5210/fm.v22i8.8005

Gleicher, N. (2019, April 1, 2019). *Removing Coordinated Inauthentic Behavior and Spam From India and Pakistan* https://newsroom.fb.com/news/2019/04/cib-and-spam-from-india-pakistan/

Inglehart, R. F., & Norris, P. (2016, 1-4 September 2016). *Trump, Brexit, and the Rise of Populism: Economic Have-nots and Cultural Backlash* American Political Science Association Annual Meeting, Philadelphia, USA. https://research.hks.harvard.edu/publications/getFile.aspx?Id=1401

Malina, A. (2005). Perspectives on citizen democratisation and alienation in the virtual public sphere. In B. Hague & B. Loader (Eds.), *Digital Democracy: Discourse and decision making in the Information Age* (pp. 23-38). Routledge.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012). On building a reusable Twitter corpus. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,

Mylonas, Y. (2017, 2017/05/04). Witnessing absences: social media as archives and public spheres. *Social Identities, 23*(3), 271-288. https://doi.org/10.1080/13504630.2016.1225495

Owen, L. H. (2019, November 8, 2019). It is still incredibly easy to share (and see) known fake news about politics on Facebook. *NiemanLab*. https://www.niemanlab.org/2019/11/it-is-still-incredibly-easy-to-share-and-see-known-fake-news-about-politics-on-facebook/

Papacharissi, Z. (2002, February 1, 2002). The virtual sphere: The internet as a public sphere. *New Media & Society, 4*(1), 9-27. https://doi.org/10.1177/14614440222226244

Papacharissi, Z. (2008). The virtual sphere 2.0: The Internet, the public sphere, and beyond. In A. Chadwick (Ed.), *Routledge handbook of Internet politics* (pp. 246-261). Routledge.

Paul, C., & Matthews, M. (2016). The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It. *Rand Corporation*, 2-7.

Raymond, J. (2003). *Pamphlets and pamphleteering in early modern Britain*. Cambridge University Press. http://www.loc.gov/catdir/description/cam022/2002023373.html

http://www.loc.gov/catdir/samples/cam033/2002023373.html

http://www.loc.gov/catdir/toc/cam021/2002023373.html

Riffe, D., Aust, C. F., & Lacy, S. R. (1993, March 1, 1993). The Effectiveness of Random, Consecutive Day and Constructed Week Sampling in Newspaper Content Analysis. *Journalism & Mass Communication Quarterly, 70*(1), 133-139. https://doi.org/10.1177/107769909307000115

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics, 25*(2), 165-172.

Sanders, M. L., & Taylor, P. M. (1982). *British Propaganda during the First World War, 1914–18*. Macmillan International Higher Education.

Selivanov, D. (2016). *text2vec: Modern Text Mining Framework for R.* In (Version 0.4.0) CRAN. https://CRAN.R-project.org/package=text2vec

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact., 3*(CSCW), 1-26. https://doi.org/10.1145/3359229

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. IConference 2014, Berlin, Germany.

Twitter. (2018a). *Retweet FAQs* https://help.twitter.com/en/using-twitter/retweet-faqs

Twitter. (2018b). *Update on Twitter's Review of the 2016 U.S. Election* (Global Public Policy, Issue. https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html

Twitter. (2019a). *More about restricted uses of the Twitter APIs* https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html

Political Content, (2019b). https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html

Vallis, O., Hochenbaum, J., Kejariwal, A., Rudis, B., & Tang, Y. (2014). *AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test.* In *R Package Version*

Walker, S. (2015). *The Complexity of Collecting Social Media Data in Ephemeral Contexts* Internet Research 16, Phoenix, AZ.

Walker, S., Mercea, D., & Bastos, M. T. (2019). The Disinformation Landscape and the Lockdown of Social Platforms. *Information, Communication and Society, 22*(11), 1531–1543. https://doi.org/10.1080/1369118X.2019.1648536

Weedon, J., Nuland, W., & Stamos, A. (2017). *Information Operations and Facebook*.

Xu, J.-M., Burchfiel, B., Zhu, X., & Bellmore, A. (2013). An examination of regret in bullying tweets. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

Zimmer, M. (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday, 20*(7).