Quantifying Echo Chambers and Their Impact on News Engagement: Evidence from a Facebook Algorithm Update

By Callum Shaw*

Draft: November 5, 2025

I present evidence that social media platforms face incentives to make algorithmic design choices that spread unreliable news and encourage polarization in news consumption. I show that a 2018 Facebook algorithm update increased the homophily of the network, increased user engagement with pro-attitudinal news, and encouraged engagement with unreliable mainstream news outlets over reliable ones. I rationalize these findings with a framework which shows how increasing homophily on a social network can encourage engagement with pro-attitudinal, unreliable news, and estimate elasticities of these outcomes with respect to homophily in a two stage least squares procedure where I instrument for homophily with the timing of the algorithm update. The findings highlight a misalignment between platform incentives and the objectives of a well informed population, and consumer engagement with diverse viewpoints.

The rise of social media as a news source has raised concerns about information reliability and political polarization, particularly since the 2016 U.S. presidential election (Allcott and Gentzkow (2017)), and in response to the development of new social media technologies like TikTok and RedNote. A characteristic trait of social media networks is homophily - the extent to which individuals are more closely connected with those who share their existing views - creating environments often referred to as echo chambers[1]. While the phenomenon has been linked with negative outcomes (McPherson et al. (2001)), the exogenous sources of variation in homophily necessary for a systematic analysis are naturally rare, as its extent in any social network is the result of self-selection by individuals[2]. As an increasing proportion of news intake comes via social media ecosystems characterized by high homophily, there is a crucial need to better understand the consequences of this type of media consumption.

I leverage a January 2018 Facebook algorithm update along with a rich, granular Facebook dataset to evaluate how increased homophily affects user engagement with news on social media. The update - termed the 'Meaningful Social Interactions' update – prioritized re-shares from socially and ideologically closer connections on a user's Newsfeed, and therefore plausibly

* Shaw: London School of Economics, c.j.shaw1@lse.ac.uk; funded by an ESRC PhD Studentship

[1] A number of empirical studies confirm the existence of homophily in social networks (Bakshy et al. (2015), Cinelli et al. (2021))

[2] A defense proponents of social media often invoke, see, for example, Patel (2021).

constituted an exogenous increase in network homophily (Hagey and Horwitz (2021)). I use my data to measure the change in homophily at the time of the algorithm update[3], and then to measure the impact of the update on three news engagement outcomes. My first result is that the algorithm update increased network homophily[4]. I then show that the update increased user preferences to re-share pro-attitudinal news, encouraged engagement with low-reliability mainstream news outlets, and increased the divisiveness of news spread on the platform. I rationalize these results with a model of re-sharing on social media, where outcomes are driven by reputational concerns of users.

I use data from Meta (the parent company of Facebook) provided through the Social Science One initiative, which tracks user engagement on the platform with URLs which have been shared on the platform, including the URLs of news articles. This dataset includes the number of views[5] and shares each URL received each month, disaggregated by five political-leaning groups of users, covering the period from 2017 to 2019. This political leaning categorization of users is constant over time[6], and provides the crucial variation which facilitates my central results.

The first empirical contribution of the paper is the measurement of homophily itself. I develop an estimation approach which leverages the longitudinal nature of the data. The core idea of my estimation strategy here is to measure how strongly sharing activity by one political group predicts exposure among other groups, while using the timing of interactions to control for other factors which might be influencing the probability of someone seeing something in their newsfeed (such as additional factors the algorithm is taking into account about the articles headline and content, for example). I observe a statistically significant increase of 117% in the homophily of the network at the time of the algorithm update.

Next I measure the impact of the algorithm update on total engagement with a sample of the 35 most-shared news articles in the US. I find that the sign of the change in engagement for a news outlet due to the algorithm update can be well predicted by whether that outlet falls above or below a threshold reliability cutoff - this splits the outlets into 'reliable' and 'unreliable' groups. Using an event study approach, along with comparisons with off-platform measures of engagement, I show that this result is not driven by non-parallel pre-trends, external events, or

[3]Note there are two types of homophily at work here. On one hand, users themselves connect with users who share their beliefs, self-selecting into more homophilic networks. On the other hand, the platform can decide to expose a user to more activity from her close friends and groups, or to more activity from those more distant in the network. The former mechanism highlights that homophily on social media can be the result of existing tribal tendencies in society. I leverage an exogenous change in the platform-controlled homophily to demonstrate that causation can also occur in the reverse direction.

[4]Measured as the negative of the correlation between the following two variables: the amount by which an article being re-shared by one person i on Facebook increases the probability of it appearing on the Newsfeed of another person j; the distance in political ideology between person i and person j.

[5]Throughout the paper, by 'view' I mean 'instance of an article being seen by a user on their Newsfeed'. The Newsfeed is the central feature of the Facebook platform, which users scroll through to encounter posts.

[6]That is, for any interaction (view or share) of an article by a user, that interaction is tallied under the political leaning classification which has been assigned to that user by the platform. A user's classification is constant over the time period of the dataset.

substitution away from out-of-sample news outlets.

My next finding demonstrates that the algorithm update increased user preferences to engage with pro-attitudinal news. I measure the extent to which a user is more likely to engage with news from pro-attitudinal, rather than counter-attitudinal news outlets, conditional on seeing a piece of content from this outlet. To do this, I use variation in news outlet political leaning, and again leverage the disaggregation of engagement data by political leaning of users. I measure the extent to which the probability of a user re-sharing a piece of news content can be predicted by the closeness of the match between her political leaning and the political leaning of the news outlet which produced the news content. I find that this measure sharply increased at the time of the algorithm update.

I find, additionally, that the algorithm update causes an increase in engagement for more divisive news articles - an effect which is more intense for less reliable articles and so compounds the effect of the first result. This result relies on an imputation of article divisiveness scores (using a natural language processing model), and so should be interpreted with more caution.

I rationalize my results with a model that builds on Acemoglu et al. (2023). In this model, rational agents encounter news on a social media platform and decide whether to re-share or express their dislike for it. Agents' incentives are driven by subsequent users' expressions of approval for the content (by re-sharing it) and by a desire to avoid spreading misinformation. Homophily of the network is modeled explicitly with an island network structure, and news content varies in reliability, divisiveness and political message.

In this model, increasing network homophily increases aggregate engagement on the platform with unreliable news outlets, decreases aggregate engagement with reliable news outlets, and increases engagement with more divisive news articles. I extend the model's results to show that an increase in homophily also raises the preference of users to re-share pro-attitudinal news, conditional on seeing it, over re-sharing counter-attitudinal news, conditional on seeing it. This formalizes the intuition that tribalistic behaviour[7] can arise from reputational concerns - an effect closely related to that of 'group polarization' (Sunstein (2002)). In this framework, this effect occurs even as political beliefs are held fixed, consistent with results presented in Guess et al. (2023) suggesting that reshares on social media do not detectably affect political beliefs or opinions in the short run. This result instead arises because, in a more homophilic network, an agent is emboldened to re-share pro-attitudinal content with the assurance that this opinion is now less likely to be scrutinized and more likely to be applauded. This is a more subtle consequence of an echo chamber which I term an 'agitation bubble' effect. While related phenomena have been noted in some prior work (for example, Hampton et al. (2017)),

[7]Throughout the paper, I use the term 'tribalism' to refer to the tendency of individuals to engage more with pro-attitudinal news, conditional on seeing it, than with counter-attitudinal news, conditional on seeing it.

this aspect of echo chambers has received less attention than other forms of online polarization.

This result also highlights an important reverse causality of the impact made by social media platforms: while existing tribalism in society causes people to organize themselves into echo chambers (more homophilic networks) on platforms, homophily which is introduced exogenously can itself cause tribalism to emerge or escalate. This is indicative of a positive feedback mechanism which could accelerate shifts towards a more polarized society.

Treating the January 2018 algorithm update as an exogenous increase in the homophily of the Facebook network, I use it to overcome the reverse causality. I measure the elasticity of the three news engagement outcomes I mention above with respect to homophily using monthly time variation in these variables, in a Two-Stage Least Squares (2SLS) approach where I use the timing of the algorithm update to instrument for network homophily. The validity of this instrument is supported by institutional evidence on the nature and timing of the algorithm update, together with robustness analyses suggesting that no other contemporaneous developments can plausibly account for the observed changes in news engagement outcomes.

I find that tribalism on the network increases by 0.44%, that average divisiveness of a shared article increases by 0.05%, and that the amount by which unreliable mainstream news outlets outperform reliable mainstream news outlets increases by 0.39%, for every 1% increase in homophily.

In sum, my findings imply that the increase in homophily increased the virality of less reliable, more divisive news, and increased user preference to engage with pro-attitudinal news on the platform. The findings show that changing network algorithms can exacerbate polarized behaviour on platforms, and my theoretical framework rationalizes this in a way which is consistent with existing empirical findings showing no detectable effect on political beliefs of users. This shifts focus away from explanations for polarization and the spread of unreliable news based on changing beliefs or individual cognitive biases, and towards the structure of communication network structure as a driver of misinformation and polarization.

My analysis confirms anecdotal accounts[8] indicating that the algorithm update improved aggregate engagement with the Facebook platform, and thus was to the benefit of the company. The results thereby demonstrate that there is a potential misalignment between the incentives of social media platforms and the objectives of a well informed population, and exposure of consumers to diverse viewpoints.

A. Related Literature

This paper contributes to the growing empirical literature studying the interplay between social media, polarization, and news dissemination, which has been reviewed in Zhuravskaya et al.

[8] And some existing evidence - see, for example Fraxanet et al. (2024)

(2020) and Aridor et al. (2024). Its primary contribution is to demonstrate that an algorithmic change by a platform can increase users' propensity to engage in more tribal behavior, thereby contributing to the debate on the extent to which social media platforms are responsible for polarizing outcomes. A compelling recent result in this literature comes from Guess et al. (2023), who find that reshared news on social media does not detectably affect users' beliefs or opinions. My paper offers a fresh perspective through the framework I use to rationalize my results. In this framework, political beliefs are held fixed—consistent with the findings of Guess et al. (2023)—and the observed increase in tribal behavior arises from changes in the reputational payoffs of re-sharing pro-attitudinal news induced by greater network homophily. In this sense, the change in network structure reveals latent political attitudes rather than shifting them.

My results corroborate the predictions of the model of rational social media user behaviour developed in Acemoglu et al. (2023), one of which I derive independently. My results bear a significant empirical resemblance to the main predictions of this framework, and extend the literature on the adverse effects of social media by demonstrating that platform decisions about network homophily can plausibly be antithetical to the objectives of a well informed population (by increasing engagement for unreliable mainstream outlets), and engagement of users with diverse viewpoints (by increasing preferences for engagement with pro-attitudinal news).

I build on the theory in Acemoglu et al. (2023) by incorporating user tribalism in how individuals engage with and disseminate news, thereby linking issues regarding social learning with the polarization driven by echo chambers. The central role played by reputation in this framework is fundamental to this result, and is consistent with the findings of Guriev et al. (2023), which demonstrate the importance of reputational concerns in social media. A foundational discussion of echo chambers is included in Sunstein (2002); Seargeant and Tagg (2019) emphasize the active role of users in shaping the media ecosystem, and Levy and Razin (2019) offer a complementary theoretical perspective by highlighting the bidirectional relationship between segregation and beliefs.

Several other theoretical contributions have enriched our understanding of information dynamics on social media platforms. Gong and Yang (2024) offer a closely related perspective on the relationship between homophily and belief polarization, with findings that are broadly consistent with mine, though they focus more explicitly on downstream misinformation. Additional work in this area includes models that incorporate platform interventions: for instance, Papanastasiou (2020) highlight the role of fact-checking in shaping belief propagation, while Dasaratha and He (2023) examine the broader dynamics of virality. Although both approaches are relevant, their limited formal treatment of homophily renders them less applicable to the specific network-based mechanism I study.

On the empirical side, a paper of particular relevance to my work is Levy (2021) who uses a

large-scale Facebook experiment to contrast the effects of pro-attitudinal versus cross-cutting news exposure, providing direct evidence on how algorithmic curation influences user behavior. My work adds to this literature by explicitly measuring network structure to investigate its impacts on user behaviour. More broadly, this paper contributes to the empirical study of misinformation and its political effects, with comprehensive overviews provided in Allcott and Gentzkow (2017) and Muhammed T (2022). The theoretical foundation I build on assumes user behaviour which is consistent with the findings of Guriev et al. (2023), and another notable analysis which uses the same dataset I do is presented by Braghieri et al. (2024).

My analysis of tribal behaviour builds on Garz et al. (2020), who introduce a related measure of "congeniality". My results are also broadly aligned with Germano et al. (2022), who provide some survey-based evidence showing that political polarization in Italy rose after Facebook's 2018 algorithm change. While their framework assumes a correlation between prior beliefs and platform activity[9], I offer a distinct mechanism - the agitation bubble - where reputational concerns drive an increase in preference for ideologically aligned news *conditional* on exposure to this content. My approach is enabled by URL-level exposure data and allows me to establish a direct link between tribal behavior and network homophily.

Halberstam and Knight (2016) and Conover et al. (2011) document political homophily in social media networks and its reflection of preexisting polarization. By contrast, my analysis reverses the causal direction, examining how exogenous increases in homophily can drive greater tribalism and misinformation, even when political ideology is held fixed. This distinction is central to current debates: while prior work focuses on how polarization is mirrored in online spaces, I show how platform design itself can amplify polarization.

With its emphasis on social network architecture and a Bayesian model of online engagement, my contribution stands in contrast to approaches that attribute engagement with misinformation primarily to individual cognitive biases. This perspective aligns with rational-update accounts of tribal behavior, as discussed in Benjamin (2019) and Bénabou and Tirole (2016), and differs from models emphasizing irrational behaviour (further work which explores this distinction comes from Pennycook and Rand (2019), Mostagir and Siderius (2022), and Kahan (2017)).

Importantly, by adopting a rational framework and using a causal inference approach, I show that behaviors often interpreted as evidence of motivated reasoning - such as re-sharing ideologically aligned content - can also emerge naturally from rational responses to the structure of the communication network. In this way, the evidence I present challenges interpretations based solely on individual cognitive bias (e.g. identity-protective cognition), and instead places greater explanatory weight on the design of the social media environment itself. This has

[9]In contrast to Acemoglu et al. (2023), where sharing behaviour is determined in equilibrium.

implications for accountability, as it shifts responsibility toward the platform intermediaries, like Facebook, who control the architecture of the communication network.

Methodologically, I develop a new estimation technique for measuring network homophily using aggregated data, extending the theoretical model into an empirical framework. This approach requires a time series strategy to address omitted variable bias and contributes to the literature on the empirical identification of network structures, as reviewed in VanderWeele and An (2013) and de Paula (2020).

Finally, this paper contributes to a growing line of work in economics that applies natural language processing tools to operationalize theoretical concepts as inputs to empirical models. Gentzkow et al. (2019) provide a comprehensive overview of such methods.

The remainder of the paper is organized as follows. Section I provides a detailed description of the January 2018 Facebook algorithm update. In Section II I describe my data, and in Section III I summarize the theoretical framework and my extension of it. Sections IV and V detail my empirical approach to measuring homophily and news consumption outcomes, respectively. My results are summarized in Section VI, and Section VII concludes.

I. Facebook algorithm update

Facebook is today the most widely used social media platform in the world, boasting over 3 billion active users (We Are Social (2024)), and accounting for around 45% of all social media site visits in the US in the year to April 2024 (StatCounter (2024)). The Newsfeed is a central feature of the platform which displays a continuous stream of Facebook posts algorithmically curated to be relevant to each individual user. It exposes consumers to a mix of user-generated content, and professionally produced content, which could appear either having been shared to the platform from the publisher's own institutional Facebook account, or by being re-shared by another Facebook user. An important subset of this is news content; 54% of Americans use social media as a source of news, and Facebook has historically been the most news-focused social media platform (Center (2024)).

Facebook continuously updates the Newsfeed algorithm to better target the company's goals, which normally center user engagement as a key metric. Some updates are significant enough to warrant announcements by the company, and can have large impacts on the business of news outlets and the news diets of consumers. The Newsfeed algorithm has become an increasing concern for regulators as, since 2016 in particular, a debate has formed over the responsibility Facebook has for the content which is published on its site (Allcott and Gentzkow (2017)).

Two of these updates have garnered particular attention, both of which were part of a longer term push by Facebook to increase interactions between people who are more likely to be socially adjacent offline. The first, in June 2016, altered the algorithm to increase the weighting that

activities of a user's friends and family have in the Newsfeed. The second, in January 2018, took another step in a similar direction, with the aim of increasing what the company termed 'Meaningful Social Interactions' by increasing the prominence in one's Newsfeed of content shared by other users closer to one in the network. The update's stated aim was to use this new value (MSI) as a key metric for the Newsfeed performance, and thereby improve the user experience by increasing the quality of time spent on Facebook (Zuckerberg (2018)). There is also evidence from an article publicising the algorithm update by Head of News Feed for Facebook Adam Mosseri that the changes to the algorithm were introduced over the first few months of 2018 (Mosseri (2018)). This would be suggestive of a ramp up effect, a possibility which I discuss when discussing my results.

Since the update, Facebook employees have themselves documented that a less publicized motive was to reverse a downward trend in user engagement (Hagey and Horwitz (2021))[10]. As Figure 1 shows, and as has been documented in existing research (Fraxanet et al. (2024)), the update seemed to have been successful in this aim[11]. The update was not accompanied by a noticeable change in the trend of new users joining the platform (Platforms (2024)), indicating that the change in the trend in engagement was driven by a change in the behaviour of users, conditional on being on the platform. Alongside the boost to engagement, several commentators documented mainly anecdotal evidence of what seemed to be an uptick in the proliferation of divisive, outrageous and unreliable news (Hagey and Horwitz (2021), Tonkin (2021)) as a result of the change.

At the same time as the Meaningful Social Interactions update, Facebook also implemented an additional algorithm update whose purpose was to ensure news "only comes from trusted sources", as defined by consumer surveys (Media (2024)) - I term this the Trusted Source (TS) update. This additional result is relevant for my results regarding the reliability of news, and I take this into account when considering the mechanism driving this result.

Under the assumption that a user is more likely to share political beliefs with those who she is friends with, or shares group affiliation with, we can hypothesize that the algorithm update constituted a step increase in the homophily of the network - the extent to which the strength of the connection between any two nodes is correlated with the similarity of those nodes[12]. This is explicitly evidenced by the fact that, at the time of this update, < 1 multipliers were introduced on the 'Newsfeed score' given to content and activities of users who are more distant from a

[10]The fact that the update was part of a longer term strategy by Facebook supports my identification approach. This helps rule out that the timing of the algorithm update was itself a function of the changes to tribalism and reliability of viral news that occurred simultaneously, and which I treat as the result of the update. It also seems likely that the time needed to plan and execute an algorithm update of this scale is larger than would be necessary if it were in response to the changes I treat as outcomes.

[11]Some previous research has also found heterogeneity in the update's effects - Gruen (2018), for example, demonstrates that the update decreased engagement for smaller and non-profit news organizations.

[12]Barberá (2014) points out the importance of exposure to these diverse viewpoints to the potential benefits of social media. We can see this algorithm update as a reduction in the extent of the mechanism described in that paper.

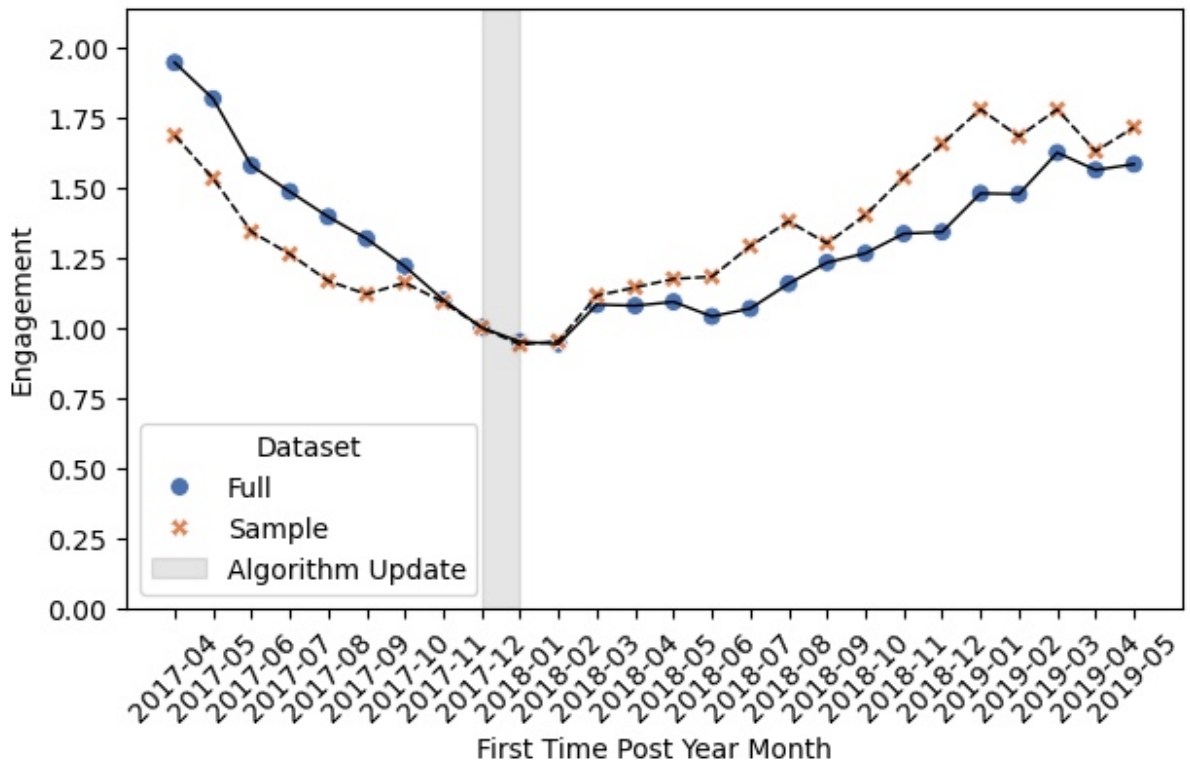


Figure 1. Change in aggregate engagement with news content at the time of the algorithm update.

Notes: The figure shows the trend in total shares per month over time for the full dataset (blue) and the subsample of the 35 most shared news outlets which are the focus of my study (orange), and how each changed at the time of the algorithm update. Each time series is indexed to 1 at 2017-12.

user in the friend network (Hagey and Horwitz (2021)) - a detail which sets the 2018 update apart from the 2016 update[13].

A growing theoretical literature examines the behaviour of consumers who engage with and re-share news in social networks, and how this behaviour is impacted by the structure of the network in question (Acemoglu et al. (2023), Gong and Yang (2024), Dasaratha and He (2023), Papanastasiou (2020)) - in general, changes in network structure change the payoffs a consumer faces to taking particular actions on the network. In contrast with traditional media, the patterns of consumer engagement with news on a social media platform are crucial in determining which news proliferates across the social network. We should thereby expect these changes to consumer behaviour on the network to impact the composition of news which proliferates on the platform.

This points to a simultaneous causality relationship between two variables. On one hand, we can expect consumer behaviour to drive homophily (users self-select into networks on the platform via their explicit choice of friend connections and group affiliations); on the other hand,

[13]Note that both the platform (via its control of the Newsfeed algorithm) and individual users (via their control of who they are friends with, and which groups they join) both have the ability to vary network homophily. This paper leverages variation in homophily caused by the former factor.

network homophily can drive changes in consumer behaviour. Provided the algorithm update constituted an exogenous increase in network homophily, it can be used as an instrumental variable which isolates the second direction of causality. This then plausibly facilitates a means of estimating the extent to which social networks and echo chambers can drive tribal behaviour, rather than merely being a manifestation of it[14].

While I do not observe the full social network of Facebook itself, I do observe article-level engagement data, disaggregated by political affinity, before and after the January 2018 algorithm update. In section IV, I develop a method of inferring network homophily using this aggregate data, and use this to demonstrate that homophily increased at the time of the algorithm update.

II. Data

My dataset combines Facebook engagement data with measurements of news outlet political leaning and reliability, and data on the divisiveness of articles which has been generated using natural language processing methods.

My central dataset is the URL Shares dataset, provided by Social Science One (Messing et al. (2020)). This is proprietary data provided by Meta to accredited researchers and administered by the Social Science One organization. The dataset provides aggregated data on how users engage with URL posts which have been shared to Facebook at least 100 times between January 1, 2017 up to and including October 31, 2022. While the data is not separated out by each individual interaction, the engagement metrics are disaggregated by country, age, sex and political affinity (political leaning of user measured on an integer scale between -2 (very left wing) and 2 (very right wing)), and thus can be used to carry out analyses of how different demographic and political groups behave differently on the platform. The two metrics I am most interested in are a view (one instance of a URL appearing on a user's Newsfeed) and a share (one instance of a user sharing a URL). A unit of observation in my context is an article. I look at data between January 2017 and January 2019 inclusive. I use data on all URLs to measure network homophily in each month. To measure news consumption outcome variables, I focus on a subset of the 35 most-engaged-with news outlets, which matches very closely with the set of the most popular news outlets in the US measured by other means. The focus of this study is on mainstream news, and how the Facebook algorithm update disproportionately favors certain outlets as a function of their reliability, rather than on fringe and small news media outlets which have been the focus of existing work on misinformation and election interference by foreign actors[15]. The engagement data is filtered to only include engagement by users who

[14] I discuss the plausibility of the exclusion restriction of this instrument in greater detail in section 7

[15] Note that 'news outlet' is broadly defined to include outlets which have traditionally focused on any format including TV, radio or print. However, it is important to note that my data includes information only on the output of these news outlets which is formatted as written articles which appear on these outlets' websites. As such, it is a study of written-word news media.

reside within the US geographically[16].

The Social Science One dataset has been privacy protected with the addition of noise which effectively induces a measurement error that biases all statistical results in this paper towards 0. As such, we should consider all estimates conservative lower bounds on the true parameters. In addition to domain, URL, first post time, and engagement metrics, I also observe the headline for each article, which allows me to carry out the natural language processing necessary for the empirical analysis of divisiveness.

	Mean	Median	SD
Panel A. News Outlet level statistics			
Monthly outlet Views (m)	290.03	200.70	316.77
Monthly outlet Shares	1,424,376	985,711	1,403,295
Monthly outlet Articles	976	704	892
Panel B. Article level statistics			
Views per article	298,318	116,417	807,396
Shares per article	1,453	533	5,790
Panel C. Totals			
News Outlets		35	
Articles		811,104	
Views (m)		241,967	
Shares (m)		1,178,699	

Table 1—Descriptive Statistics

Note: Monthly mean outlet views, total views and total shares are measured in millions. The sample is 35 news outlets between January 2017 and January 2019.

Table 1 displays descriptive statistics for my data. As the table shows, at both levels of aggregation (that is, at both the month-news outlet level, and at the article level), the distributions of both views and shares are characterized by a right skew.

I combine the engagement data with expert data on the political leanings of news outlets, and a mix of expert and survey data on the reliability of news outlets.

The political leaning of news outlets are based on expert assessments carried out by Ad Fontes Media. Figure 2 displays the distribution of the political leanings of the news outlets in my sample, where lower numbers indicate more left leaning outlets. The political distribution of the full sample is shown overlayed with the distribution of the subsample used for the Reliability Result. The distributions are very similar to each other, both displaying the leftward weighting (and longer right tail) which has been extensively documented in the previous literature studying the political bias of US news outlets (Groseclose and Milyo (2005), Levy (2021)).

To measure reliability, I use the 2017 survey by Reynold's Journalism Institute (Kearney

[16]Table C2 in the data appendix provides a list of all domains in the sample.

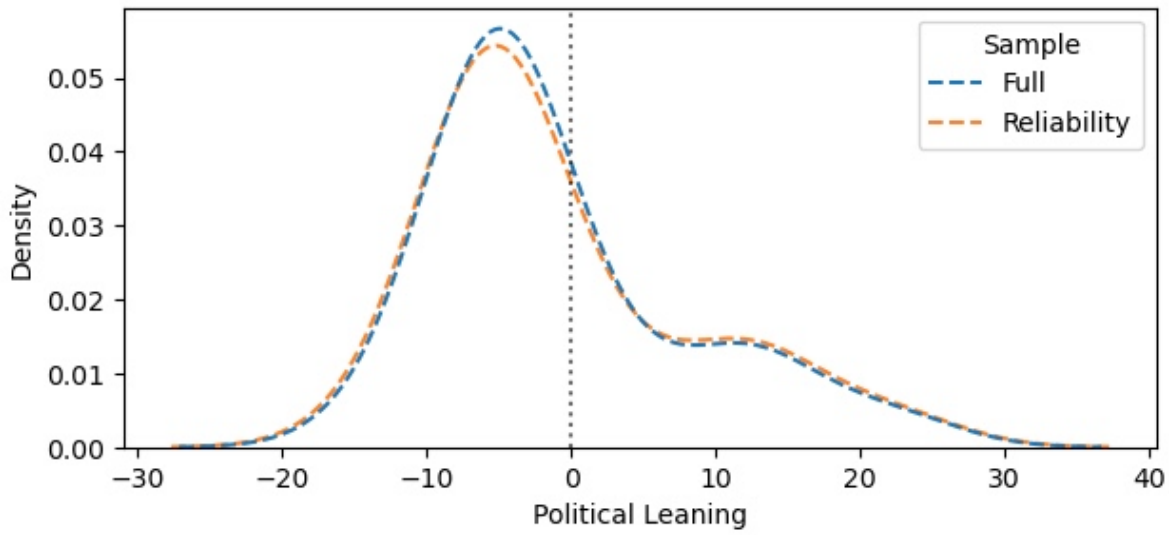


Figure 2. Distribution of Political Leaning of News Outlets.

Notes: The figure displays the distribution of news outlets along the political spectrum for the full sub sample of 35 news outlets (blue) and the truncated sample of news outlets for which I have reliability scores. Lower numbers correspond to more leftward leaning outlets and high numbers to more rightward leaning outlets. The distributions are each plotted separately using a Gaussian kernel, with an unadjusted bandwidth selection.

(2017)). This is a survey of a political, geographic and demographic cross section of US citizens which measures various attitudes towards a range of mainstream media sources, including their reliability. From this source I extract a numerical reliability score for each news outlet. Using a measure of reliability based on a consumer assessment may be susceptible to a bias in this measure towards one end or the other of the political spectrum. As such, I also robustness check results with measures from an expert assessment of reliability produced by Fontes (2018).

As a further robustness check of the Reliability Result, I use measures of engagement with news outlets which are not from the Facebook platform. For this purpose, I leverage two additional datasets. One uses data provided by the online marketing measurement firm Semrush, and provides estimates of the monthly traffic going to each news outlet's website domain. The other uses data from the Google search engine, and provides a proxy measure for traffic by measuring the number of times a news outlet was queried using Google in each month.

I use a natural language processing (NLP) approach to measure divisiveness of the articles in my dataset.

Initially, I produce divisiveness estimates for articles from each news outlet using a large language model[17]. This involves prompting the model to give a measure of the divisiveness of each article's headline between 1 and 10. Although cheap compared to crowd-workers, the expense of this labelling procedure limits its application to only a subset of articles. The resulting labelled data allows me to train a supervised machine learning model to classify the

[17]There is growing evidence of the superiority of LLMs to crowd-workers for text annotation tasks - see, for example, Gilardi et al. (2023).

remaining articles.

The training data consists of headlines and their respective divisiveness ratings. First, the model converts headlines into vector representations using a sentence embedding - this converts the dataset into a wide numerical dataset, where the number of features is equal to the dimension of the vectors produced by the sentence embedding. I then use a gradient boosted tree model to predict the divisiveness of each headline from this set of 'features'. This trained model can then be applied to predict the divisiveness of a previously unseen headline[18].

III. Theoretical Framework

I rationalize my empirical findings with a model which builds on Acemoglu et al. (2023) to study re-sharing behaviour of users on a social network, adding to their model the notion of tribalism. The model and detailed proofs are in the online appendix. Here I summarize the key model features and results to ground ideas before presenting my empirical analysis.

There is a binary true state of the world ($\theta \in \{L, R\}$), about which agents have a prior, $Pr(\theta = R) \in [0, 1]$, which is a random draw from a prior belief distribution specific to that person. When an agent encounters an article on a social media platform, she updates her prior to a posterior, taking into account the article's reliability ($r \in [0, 1]$) and its message $m \in \{L, R\}$. The agent then chooses whether to share, ignore or disapprove of the article; payoffs capture an agent's desire to share news which is correct and which will also be subsequently shared by more agents. Agents are arranged into a (weakly connected) island network, where the set of agents is partitioned into islands. A shared article has probability p_s of appearing on the newsfeed of one's co-islanders, and p_d of non-co-islanders, with $p_s \geq p_d$. An agent has a more similar prior belief distribution to her co-islanders than those who do not inhabit her island.

Homophily describes the extent to which a person is more closely connected to one's co-islanders than other islands[19]. **Divisiveness** is a measure of how tethered an article's message is to the true state of the world, such that more divisive content is content which is more likely to cause disagreement among people with contrasting political ideologies. I define as **tribalism** the extent to which a right (left) wing person is more likely to share right (left) wing content than a left (right) wing person, conditional on seeing a piece of right (left) wing content[20].

The equilibria are found using the properties of supermodular games. The key results are driven by strategic complementarities; if others who are closely connected to agent i are more

[18] More detail on the model is given in the online appendix, where table C3 gives examples of headlines and their predicted divisiveness ratings, taken from the dataset.

[19] Homophily is higher when the probability an agent's shared article appears on the Newsfeed of her co-islanders is relatively higher than the probability it appears on the Newsfeed of a non-co-islander - i.e. when $p_s - p_d$ is higher.

[20] Tribalism can equivalently be defined as the extent to which a right (left) wing person is more likely to share a piece of right (left) wing content than a piece of left (right) wing content, conditional on seeing such a piece of content. Tribalism is formally defined as $\mu = Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = R] - Pr[a_i = \mathcal{S} \mid b_i < 1/2, m = R]$. See the mathematical appendix for further discussion.

likely to share a piece of news, then that increases agent i's payoff from sharing the piece of news. The effects of an increase in homophily are driven by the changes in this payoff as the other agents most likely to see your re-shares become more similar to you in ideology.

I re-state two results, for which detailed proofs are in Acemoglu et al. (2023). The third result I prove in the mathematical appendix.

Reliability Result (Theorem 2 in Acemoglu et al (2023)). There exist $0 < \underline{r} < \bar{r} < 1$ such that, for any article:

(a) if $r < \underline{r}$, greater homophily increases user engagement;

(b) if $r > \bar{r}$, greater homophily decreases user engagement.

The theorem shows a non-monotonicity in the relationship between homophily and engagement. The intuition for this is that, after an increase in homophily, there are two mechanisms which work in opposite directions:

- "Discipline effect": the likelihood of being called out for spreading misinformation is lower when those you are sharing it with are more like-minded
- "Circulation effect": A shared article is now less likely to spread to those outside of a small group, lowering the potential benefits of sharing

The above result establishes that, for low reliability articles, the discipline effect dominates, and for high reliability articles, the circulation effect dominates. The result establishes that increasing homophily should lead to an increase in the proliferation ('virality') of low reliability content (content with $r < \underline{r}$).

Divisiveness Result (Proposition 1 in Acemoglu et al (2023)) There exist $r^* \in (0, 1)$ and $p^* \in (0, 1)$ such that:

(a) if $r < r^*$ and $p_s/p_d > p^*$, then greater divisiveness leads to greater user engagement;

(b) if $r > r^*$ and $p_s/p_d < p^*$, then greater divisiveness leads to less user engagement.

In words, an increase in homophily from below p^* to above p^* induces a positive correlation between divisiveness and engagement for low reliability articles, and removes a negative correlation between divisiveness and engagement for high reliability articles. In aggregate, this suggests increasing homophily can increase divisiveness of viral content.

This result is complementary to the authors' Theorem 2. The intuition is that more divisive content generates more divergent behaviour from individuals with different ideologies, so echo chambers matter especially for such divisive content. An increase in homophily gives the least

reliable articles greater virality, and the most divisive of the least reliable articles goes most viral.

Theorem 1. (Tribalism Result) *An increase in homophily increases the tribalism of sharing behaviour.*

This theorem establishes that an increase in the homophily of a network will increase the extent to which right wing people are more likely to share right wing content, and left wing people are more likely to share left wing content, conditional on seeing such a piece of content. The theorem, proved in the mathematical appendix, also demonstrates that we should expect positive tribalism in any network with $p_s > p_d$.

The intuition for this result is that, in a more homophilic network, the content shared by an agent is more likely to be received by those with whom that agent is more ideologically aligned. As such, the probability of content which aligns with an agent's prior being subsequently shared by further agents increases, and the payoff to sharing such content thus increases.

It's important to note that the tribalism discussed here is conditional on that article being viewed[21]. In other words, the effect described in Theorem 1 is not simply due to individuals encountering more like-minded content, but reflects behavior once content is presented. Much of the existing literature on echo chambers emphasizes filter bubbles, where users are disproportionately exposed to pro-attitudinal news. However, such mechanisms are not unique to social media—traditional media also fosters selective exposure through self-selection mechanisms like newspaper subscriptions. Theorem 1 goes beyond this exposure effect, showing that social media homophily produces dynamics better characterized as an *agitation bubble*, where agents feel emboldened to share more pro-attitudinal news as they anticipate a favorable response from their social network. This effect is akin to the group polarization effects discussed by Sunstein (2002).

The result underscores that homophily on social media — often intentionally amplified by platform design — can actively fuel tribalism, rather than merely mirroring pre-existing societal divisions. Both the theoretical model and my supporting empirical evidence demonstrate that social media, through the formation and intensification of echo chambers, plays a causal role in driving political polarization. Notably, this effect emerges even when individuals' prior beliefs remain fixed, as is assumed in the model; the rise in tribalism stems from changes in incentives induced by the altered network structure. This short-term dynamic operates independently of, and in addition to, any long-term shifts that might result from evolving beliefs.

[21] Conversely, the reliability result is driven by the sum of both user behaviour conditional on seeing an article (the discipline effect) *and* changes in the content presented to different groups of users (the circulation effect). I explore the distinction between these mechanisms in the context of the reliability result and the tribalism result in the appendix.

My empirical setting necessitates further elaboration of the model. First, I observe engagement metrics for many articles, whereas the model analyzes engagement with just one. I aggregate the model by allowing each article to be about a different story, with a message which is about a true state of the world $\theta_j \in \{R, L\}$[22], where there are as many stories as there are articles, and where agents draw a new prior belief from their fixed prior belief distribution for each story.

For my analysis of the reliability result, I aggregate article-level data to the news outlet level. To accommodate this, I extend the framework to formally define a news outlet and show in the theoretical appendix that the Reliability Result generalizes naturally to this level. In the extension, prior belief distributions remain fixed and consumers do not draw inferences about an article's reliability based on the political leaning of the outlet which produced it. This preserves an appealing feature of the model: the results are independent of any assumed correlation between political orientation and reliability, and further, are not driven by the political opinions of consumers changing as they consume more news[23]. This aligns with the empirical findings, which hold even when conditioning on political leaning.

I draw some welfare conclusions from my analysis of reliability. For this, I follow Acemoglu et al. (2023) in defining a regulator's objective as being to maximise the informedness of the population about political issues. While this is consistent with the concerns of many regulatory efforts globally, it is important to note that it is only a partial welfare analysis, as it omits the utility agents in the model gain from their reputational incentive. In my framework, I aggregate the results of Acemoglu et al. (2023) by allowing consumers' to form a posterior on each news story in response to reading it, and the intuitive result is drawn that proliferation of less reliable news sources (or suppression of more reliable news sources) has a negative impact on the regulator's objective.

In section 7 I connect my empirical analysis to the theory, treating the algorithm update as an exogenous shift to homophily which allows me to test the comparative statics of the model and measure elasticities of news engagement outcomes with respect to homophily. I measure these elasticities using monthly level time variation in each variable, and instrument for the endogenous variable (homophily) with the timing of the algorithm update.

My main empirical analysis is in the following sections. The measurement of homophily and its increase at the time of the algorithm update is described in section IV. Measurement of the changes in the news engagement outcomes at the time of the algorithm update are in

[22] As in the baseline model, each article has a message about story j which can either match or fail to match the true state of the world; veracity and reliability work in the same way they do in the baseline model.

[23] This is consistent with the evidence presented in Guess et al. (2023), where exposure to news on social media has no significant impact on political beliefs in the short run.

three subsections in section V. These sections go into greater detail on how the algorithm update impacted each outcome[24]. Throughout, I follow the theoretical framework by defining engagement for an article as its number of shares.

IV. Measuring Homophily

The dataset displays how many times each URL has been viewed and re-shared each month, by each of 5 different political affinity groups, $\{-2, -1, 0, 1, 2\}$, into which all users are placed (low numbers correspond to more left leaning political priors)[25]. I define the strength of the connection between any two groups i and j as the average increase in the probability of an article appearing on the Newsfeed of a person in group j as a result of a person in group i sharing this article. Measuring the strength of connections between different groups therefore requires an estimate of the relationship between shares and views for each group, which can be done by leveraging variation in these variables across URLs in each time period.

An identification issue arises in measuring homophily using this approach. The Facebook algorithm determines which articles appear on users' Newsfeeds for reasons other than shares by other people in the network. The algorithm's ranking decisions depend on features of each URL that may be correlated both with its likelihood of being shared and with its likelihood of being viewed by different user groups. As a result, the strength of the correlation between shares and views across political groups may in part reflect the platform's endogenous ranking behaviour rather than genuine information transmission across users[26].

To overcome this issue, I leverage the longitudinal nature of the dataset[27]. For each URL, I observe the interactions it received from each political leaning group in each month. Whilst the majority of interactions with a post occur in the same month that the URL was first published, a proportion of these interactions occur in the month afterwards. Using sharing activity in period t provides variation in the shares an article receives which cannot be the result of the reverse causal effect of views in period $t+1$. I formalise my approach in the following paragraphs.

Let $k \in \{1, \ldots, K\}$ index URLs (articles), $t \in \{1, \ldots, T\}$ index months, and $i, j \in \{-2, -1, 0, 1, 2\}$ index the five political affinity groups, ordered from most left-leaning to most right-leaning.

For each article k and time period t, the dataset reports:

[24]For the two stage least squares estimate, I summarize each news engagement outcome with a scalar variable

[25]The composition of these groups is invariant over time. That is, the data does not allow for users to shift between groups over time.

[26]I am interested in measuring only the extent to which a share by one user impacts the probability of a view by another user somewhere else in the network (and how this is impacted by the Meaningful Social Interaction algorithm update) as this is what is important for user behaviour in my framework. I want to *exclude* from this measure the extent to which shares and views are correlated just because the algorithm feeds people more of what it thinks they will like, based on information from the article itself.

[27]VanderWeele and An (2013) discuss other examples of the use of longitudinal observations in distinguishing various properties of networks.

- $shares_{k,i,t}$: the number of times users in political group i shared article k in month t;
- $views_{k,j,t}$: the number of times users in political group j viewed article k in month t.

The relationship between shares and views can be interpreted as a simplified representation of the underlying communication network. When a user of type i shares an article, it appears on the newsfeeds of users in group j with probability p_{ij}. This defines a complete directed graph over the five user groups, with edges parameterized by the transmission probabilities p_{ij} (This structure is illustrated in figure 3). The degree of homophily in the network can then be characterised by how rapidly p_{ij} decays with political distance.

Let $f(i,j)$ denote a function of political distance between groups i and j, such as

$$f(i,j) = \begin{cases} \mathbb{1}\{i=j\}, & \text{indicator function} \\ -|i-j|, & \text{linear distance specification} \\ -(i-j)^2, & \text{quadratic distance specification.} \end{cases} \tag{1}$$

Each choice of $f(i,j)$ corresponds to a different functional form for how transmission probabilities vary with ideological proximity. The coefficient on $f(i,j)$ in the estimating equation below thus measures the degree of homophily in the network.

An observation in the estimation dataset is defined at the tuple level (k,i,j,t): that is, one article k, one month t, a group i that shares, and a group j that views. For each observation, the outcome variable is the number of views received by group j in month $t+1$, while the key regressor is the the number of shares by group i in month t.

To illustrate the identification issue regarding the algorithmic intervention, suppose that each article k in month t has an unobserved component U_{kt}, capturing its "algorithmic appeal" - characteristics of the URL that the platform's ranking system uses to predict engagement (such as its topic, headline structure, or historical click-through rate). This component affects both sharing and viewing behaviour across groups:

$$U_{kt} \rightarrow shares_{k,i,t}, \qquad U_{kt} \rightarrow views_{k,j,t}, \qquad U_{kt} \rightarrow views_{k,j,t+1}.$$

Because U_{kt} simultaneously influences both shares and views, a regression of $views_{k,j,t+1}$ on $shares_{k,i,t}$ will be confounded if U_{kt} is omitted.

To mitigate this problem, I exploit the longitudinal structure of the data and control for lagged views of the same URL by the same user group, $views_{k,j,t}$. Conditional on $views_{k,j,t}$, any residual variation in $views_{k,j,t+1}$ cannot be attributed to what the algorithm "already knew" about the article at time t. Instead, this variation reflects new exposure and diffusion dynamics

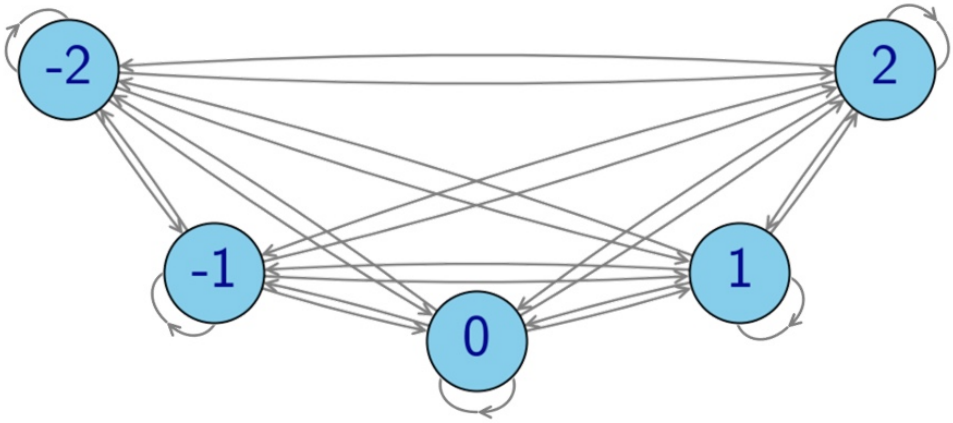


Figure 3. Graph representation of the network of political alignment blocks.

Notes: The directed graph illustrates the simplified structure of the Facebook network, with each node representing a consumer political leaning type. The nodes are horizontally arranged to reflect the political leaning that each represents (they are arranged as they are vertically to make the directed edges of the graph clearer), and as such the length of each directed edge reflects the political distance between users in each pair of nodes. A directed edge from i to j is associated with a probability of Newsfeed exposure p_{ij}, which gives the probability an article shared by a user with type i thereby appears on the Newsfeed of a user with type j. In this representation, a higher correlation between p_{ij} and the length of a directed edge from i to j corresponds to higher homophily.

in the network. The identifying assumption is therefore that, after conditioning on lagged views, residual shocks to $shares_{k,i,t}$ are uncorrelated with residual shocks to $views_{k,j,t+1}$:

$$\mathbb{E}[\varepsilon_{k,i,j,t} \,|\, shares_{k,i,t},\, views_{k,j,t}] = 0.$$

Intuitively, shocks that increase the number of shares by group i for article k in month t should not, after controlling for the above terms, directly affect the number of views by group j in month $t+1$ except through the causal channel of content transmission across the network[28].

I implement this approach to estimate the following specification separately in each time period:

$$views_{k,j,t+1} = \beta\, views_{k,j,t} + \Big[\theta_t + \chi_t f(i,j)\Big]\, shares_{k,i,t} + \varepsilon_{k,i,j,t}. \tag{2}$$

In this specification, χ_t measures the extent of network homophily in period t.

A. Homophily Estimates

I apply the method outlined above to measure the homophily of the Facebook network in each month for a period either side of the January 2018 algorithm update, which, as discussed in section I, I hypothesize constituted an exogenous increase in the homophily of the social

[28]If we allow that shares also affect the number of views for group j in the same month, the estimate will be a lower bound of the full causal effect of shares on views and hence so will our estimate of homophily. Provided that there is sufficient impact of shares from block i in period t on views for block j in period $t+1$, I am still able to provide an estimate of homophily and its change at the time of the algorithm update.

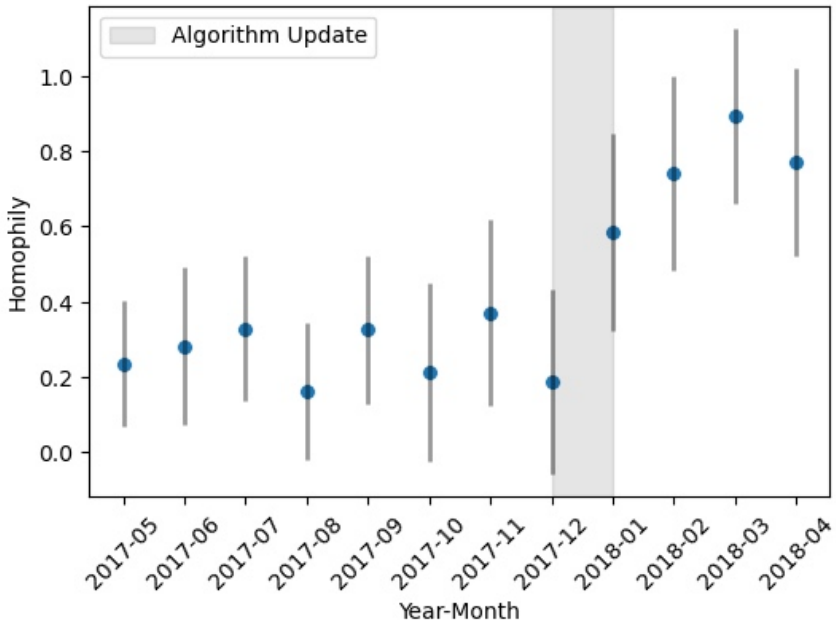


Figure 4. Time series showing the increase in homophily at the time of the January 2018 algorithm update.

Notes: The graph plots the homophily of the Facebook network over time, measured as the size of the estimate for the coefficient χ in specification 2 in the main body text of the paper. Error bars showing the 95% confidence interval for the coefficient estimate in each year-month are also shown.

network. Figure 4 displays these estimates, measured using the linear distance function in specification 1.

All estimates for homophily are positive and significant. This result is therefore consistent with existing empirical work documenting the presence of homophily in online social networks (Bakshy et al. (2015), Conover et al. (2011), Halberstam and Knight (2016))[29].

I observe a sharp increase in homophily at the time of the algorithm update that appears to ramp up over the 3 months following January, with all homophily measurements following the update being higher than all measurements before the update. This is consistent with what we'd expect given the institutional information available regarding the algorithm update. I also observe some volatility in homophily independent of the algorithm update. Homophily will vary for reasons besides algorithm updates - most importantly as a result of user self-selection into more or less homophilic friend networks. The fact that we observe an increase in the homophily measure which persists after the update indicates that users did not endogenously respond by changing their friendship networks to offset the increase in homophily. [30].

Using this approach, I can quantify the impact the algorithm update had on the homophily in the network by regressing homophily on a post-update dummy. I also estimate specifications with a 3-month window and a 9-month window either side of the algorithm update, to exclude

[29]This result also corroborates the assumption in the theoretical framework that $p_s \geq p_d$.

[30]As there is variation in homophily following the period of potential ramp up, in 2 I display results measuring the change with both a 9-month window and a 3-month window either side of the algorithm update.

further upward adjustments which may not be the direct result of the update.

Table 2—First Stage Estimates: Increase in homophily at time of algorithm update.

Specification	Estimate (SE)	Effect Size
Binary	0.88^{***} (0.20)	1.13^{*} (0.45)
Absolute Distance	1.17^{***} (0.15)	1.09^{**} (0.24)
Quadratic Distance	1.25^{***} (0.15)	1.12^{**} (0.21)
Width of window (months)	9	3
Observations	18	6

Source: Coefficient estimates are shown with their estimated standard errors in brackets. Stars display p value thresholds: $^{*} :< 0.1$, $^{**} :< 0.05$, $^{***} :< 0.01$. 9 monthly observations either side of the algorithm update are included. Log(homophily) is used as the outcome variable in order to approximate a percentage increase.

Table 2 displays the first stage estimates, which suggest that homophily of the network roughly doubled as a result of the algorithm update. When I link my empirical analysis to theory, I use this as a first stage estimate to generate an instrumented value for homophily, $\hat{\chi}$, which I use for second stage estimations of the effects of homophily on news engagement outcomes in section VI.

V. Measuring News Engagement Outcome Variables

The outcome variables of interest for this study map to the three theoretical results presented in section III regarding the reliability and divisiveness of news content, and the tribalism of user behaviour on the network.

A. Reliability

The average engagement per article for a newspaper n at a particular time t is written as a function of the reliability, r_n of the newspaper and the homophily of the network at time t, χ_t

$$S_{nt} = f(r_n, \chi_t)$$

with first derivative with respect to χ_t, $f_\chi(r_n, \chi_t)$. Provided the Reliability Result holds, the following is true: There exists $\underline{r}$ and $\bar{r}$ (with $\underline{r} < \bar{r}$), such that $f_\chi(r_n, \chi_t) > 0$ for $r_n < \underline{r}$, and $f_\chi(r_n, \chi_t) < 0$ for $r_n > \bar{r}$ [31].

The Reliability Result thus suggests that news outlets can be categorized based on their position either side of a reliability cutoff into outlets which benefit from increased homophily and those which are disadvantaged by it.

A natural empirical test of the theoretical result involves examining how engagement changes for all news outlets following the algorithm update, and leveraging cross-outlet variation in

[31] Theory does not provide us with theoretical guidance on the precise values of $\underline{r}$ and $\bar{r}$. Nor does it allow us to characterize the behaviour of the function on the interval $[\underline{r}, \bar{r}]$.

reliability to assess its relationship with these changes. A straightforward aggregate measure of network reliability is the difference in engagement between reliable and unreliable outlets.

Homophily χ is indexed by time period t, as it is a property of the social media network (**P** in the model), which undergoes changes over time; as we observe in section IV, it increases at the time of the algorithm update. For this analysis, I estimate f_χ for each newspaper n as

$$\widehat{f_\chi} = S_{n,\underline{t}>t>t^a} - S_{n,t^a>t>\bar{t}} \tag{3}$$

where t^a is the period of the algorithm update, and where $t^a - \underline{t} = \bar{t} - t^a$ is the size of the window I focus on either side of the algorithm update[32] The analysis proceeds by investigation of the relationship between $\widehat{f_\chi}$ and r_n.

Formally testing the Reliability Result presents the difficulty that theory does not provide guidance on the values of $\underline{r}$ and $\bar{r}$, or fully specify the $f_\chi(r_n)$ function. I address this issue with two different approaches.

Linear Approximation: The first approach is to estimate a straightforward linear specification. Political leaning of news outlet n (denoted pol_n) is also included as a control in this and all other specifications:

$$\widehat{f_\chi} = \alpha + \beta^l r_n + \beta^{pol}\text{pol}_n + \varepsilon_n. \tag{4}$$

This specification provides a test to rule out the null hypothesis that $\beta^l = 0$, which would indicate either that the Reliability Result does not hold or that the f_h function exhibits unusual behaviour on the interval $(\underline{r}, \bar{r})$ which would itself undermine the conclusions drawn by Acemoglu et al. (2023) regarding the Reliability Result.

Threshold Estimation: The second approach is to take a more direct interpretation of the Reliability Result and to assume that the function $\widehat{f_\chi}(r_n)$ can be characterized by direct estimation of some number of unknown thresholds in r_n which predict the value of $\widehat{f_\chi}$. The first threshold approach consists in estimating

$$\widehat{f_\chi} = \alpha + \tilde{\beta^{tl}}\mathbb{1}\{r_n < \underline{r}\} + \tilde{\beta^{th}}\mathbb{1}\{r_n > \bar{r}\} + \beta^{pol}\text{pol}_n + \varepsilon_n. \tag{5}$$

[32]The identifying assumption here is that changes that occur to the engagement for each outlet moving from before to after the algorithm update are solely due to the change in homophily at the time of the update. Concerns regarding the use of time variation to provide identifying variation in homophily are addressed during robustness checks. For this initial analysis, I estimate a specification with a 12 month window either side of the algorithm update, and a specification with a 3 month window either side of the algorithm update. Results are consistent across both specifications.

Following Hansen (2017), the unknown cutoff values $\bar{r}$ and $\underline{r}$ can be estimated using a grid search:

$$(\widehat{\underline{r}}, \widehat{\bar{r}}) = \underset{\underline{r},\bar{r}}{\operatorname{argmin}} \min_{\beta^{tl},\beta^{th}} \Omega(\underline{r}, \bar{r}, \beta^{tl}, \beta^{th}).$$

where $\Omega(\underline{r}, \bar{r}, \beta^{tl}, \beta^{th})$ is the sum of squared errors function. Given cutoff estimates, the Reliability Result can be tested by testing the hypothesis that $\beta^{tl} > 0$ and $\beta^{th} < 0$ against the respective nulls that $\beta^{tl} = 0$ and $\beta^{th} = 0$ in the regression

$$\widehat{f_\chi} = \beta^{tl}\mathbb{1}\{r_n < \widehat{\underline{r}}\} + \beta^{th}\mathbb{1}\{r_n > \widehat{\bar{r}}\} + \beta^{pol}\text{pol}_n + \varepsilon_n, \tag{6}$$

where the intercept has been removed to measure the average f_χ value in each relevant section[33].

This approach depends on the validity of the estimates which have been produced for the two cutoff values. To test the significance of these estimates, I first employ a likelihood ratio test similar to that discussed by Hansen (2000) (and by Bai (1999) in the context of structural changes in time series data). This test works iteratively by using the likelihood ratio to compare a model containing each successive threshold against the model in the absence of this new threshold. This allows a test of the hypothesis that one threshold exists rather than none, and that two thresholds exist rather than one.

The likelihood test depends on the asymptotic properties of the threshold estimators. As the size of the dataset may cast doubt over whether such asymptotic inferences are valid, I also test the reliability of the threshold estimates by generating bootstrapped standard errors.

An estimate of zero thresholds would cast doubt on the Reliability Result. A result where only one threshold is reliably identified is consistent with the Reliability Result, where the interval $(\underline{r}, \bar{r})$ is vanishingly small. In this case, rather than specification 6, I instead estimate the model:

$$\widehat{f_\chi} = \beta^{ol}\mathbb{1}\{r_n < r^R\} + \beta^{oh}\mathbb{1}\{r_n > r^R\} + \beta^{pol}\text{pol}_n + \varepsilon_n, \tag{7}$$

The Reliability Result is then tested by testing the hypotheses that $\beta^{oh} < 0$ and $\beta^{ol} > 0$.

Robustness Checks. — As my approach leverages the timing of the algorithm update to estimate f_χ, inference is dependent on a similar set of assumptions to those which are normally

[33]Including r_n also as a running variable in this specification does not have a significant effect on results, and so I omit it from the regression for clarity.

invoked in the case of a difference-in-difference design. Most importantly, I need to rule out the possibility that the observed changes are driven by different long-term trends in engagement between more reliable and less reliable outlets. To do so, I estimate the following event study specification

$$S_{nt} = \sum_{t \in \mathcal{T} \setminus \{t_0\}} \left(\phi_t \, D_t^{\tau} \cdot \mathbb{1}\{r_n > r^R\} \right) + \sum_{t \in \mathcal{T} \setminus \{t_0\}} \left(\beta_t^p \, D_t^{\tau} \cdot \text{pol}_n \right) + \alpha_n + \lambda_t + \varepsilon_{nt} \tag{8}$$

where S_{nt} is the number of shares news outlet n accumulates in time period t, D_{it}^{τ} is a dummy equal to 1 if $t = \tau$, $\mathbb{1}\{r_n > r^R\}$ is an indicator that news outlet n has a reliability above the threshold estimated in 7, pol_n is the political leaning of outlet n, α_n and λ_t are news outlet and month fixed effects. For identification, the dummy for time period December 2017 is omitted ($t_0 = $ Dec 2017.).

In this specification, ϕ_t measures the difference in shares between the reliable group of news outlets (those with $r_n > r^R$) and the unreliable group of news outlets (those with $r_n > r^R$) in each period t; we should expect to see a jump in the value of ϕ_t at the time of the algorithm update. I include an analogous time series term for the political leaning of outlets in order to control for the possibility that the results are driven by the political leaning of outlets, rather than reliability.

I run this specification first defining S_{nt} as the total shares received by outlet n in period t. I then estimate three alternative specifications: one where I define S_{nt} as the shares received from users of political leaning groups -2 and -1 by outlet n in period t; one where it is defined as the shares received by political leaning group 0; and one where it is defined as the shares received by political leaning groups 1 and 2. I do so to investigate the possibility that the results are driven by the behaviour of users at just one end of the political spectrum, or if the behavioural change is across all users.

I also carry out two further specifications where I define S_{nt} the engagement received by news outlet n in period t from two other off-platform (non-Facebook) sources. The first off-platform measure is the number of visits to a news outlet's homepage (data for which is gathered from the web traffic data platform Semrush). The second is the number of Google searches for the news outlet (data for which is gathered from the Google Trends platform). This checks whether results are driven by are due to time-varying factors unrelated to the algorithm change on the Facebook network.

For further transparency, I check against the possibility that results are driven by size difference between news outlets by plotting total shares for reliable and non-reliable groups of news outlets over time.

The Trusted Source algorithm update, announced at the same time as the Meaningful Social Interactions update, lowered the prevalence in the Newsfeed of news outlets deemed highly untrustworthy. This change was designed to remove fringe outlets, and so as my analysis focuses on mainstream news outlets, I do not expect those included in my sample to be directly affected by the algorithm update. However, this does raise the question of substitution between those outlets affected by the Trusted Source outlets and those outlets in my sample. I discuss this issue, and robustness checks against it, in the online appendix.

B. Tribalism

The Tribalism Result implies that, following the algorithm update and the resultant increase in homophily, we should observe an increase in the tribalism of sharing behaviour. This section aims to demonstrate that tribalism of sharing behaviour increased at the time of the algorithm update.

I define 'tribalism of sharing behaviour' (consistent with the definition used for theorem 1) as the extent to which right (left) wing people are more likely to share right (left) leaning content than left (right) leaning content, *conditional on seeing such a piece of content.* As such, this measure is not mechanically linked to the political leaning of the content to which consumers are exposed. This means that the phenomenon I am interested in here is one regarding user behaviour, rather than changes to the news to which consumers are exposed. The latter is relevant, but has been extensively documented in existing literature on filter bubbles (e.g. Levy (2021)). Importantly, I am able to distinguish consumer behaviour from the effects of exposure as I observe the number of times an article has been viewed, as well as the number of interactions it has received from each of 5 political affinity groups[34].

Another upshot of this definition of tribalism is that it isn't something that is mechanically linked to the increase in homophily which happens at the time of the algorithm update (i.e. via people seeing more pro-attitudinal content because the people sharing this content are more likely to be more ideologically aligned with them).

As I mention in section IV, in my data, engagement and view metrics are disaggregated by 'political affinity' - a measure of a user's political ideological leaning. Let the political affinity of consumer i be $\ell_i \in \{-2, -1, 0, 1, 2\}$ (as it appears in the data), where a higher (lower) number indicates a more right (left) wing political ideological leaning. Let a newspaper n's political ideology be denoted $\text{pol}_n \in \mathbb{R}$, where $\text{pol}_n > 0$ ($\text{pol}_n < 0$) indicates an ideologically right (left)

[34]The marginal (non-conditional) probability of sharing news content, as well as the political leaning of content to which consumers are exposed, will still be of interest in some settings. Importantly, we should expect the marginal probabilities to be a function of both the content of the Newsfeed and what I define as tribalism of sharing behaviour. I discuss these issues further in the appendix, but focus just on tribalism of sharing behaviour in the main body of the text for clarity of exposition.

wing newspaper[35].

I estimate the tribalism of sharing behaviour over time by estimating the following fixed effects specification with an interaction term:

$$S_{n\ell t} = \alpha_{nt} + \zeta_{\ell t} + \sum_{\tau} \left(\gamma_t D_t^{\tau} \times (\text{pol}_n \times \ell)\right) + \beta^v V_{n\ell t} + \varepsilon^v_{n\ell t} \tag{9}$$

Where $S_{n\ell t}$ is the number of shares news outlet n receives from political affinity group ℓ in period t, α_{nt} is an outlet $\times$ time fixed effect, μ_{lt} is a political affinity group $\times$ time fixed effect, D_t^{τ} is a dummy equal to 1 if $\tau = t$, and $(\text{pol}_n \times \ell)$ is an interaction term between the political leaning of outlet n ($\text{pol}_n \in [-40, 40]$) and the political leaning group $\ell \in \{-2, -1, 0, 1, 2\}$. $V_{n\ell t}$ is the number of views for outlet n on the Newsfeeds of users in group l in period t.

γ measures tribalism of sharing behaviour in period t. The intuition behind this procedure is that $\text{pol}_n \times \ell$ will be high when news outlet political leaning and user political leaning are a close match, and low when they are not. Using this multiplicative specification is a convenient and intuitive way to account for the fact that pol_n and ℓ have different scales (but are both centered on zero)[36].

Theory predicts that γ should all be positive in all time periods (as we assume $p_s \geq p_d$, which can be tested by rejecting the null of a 0 coefficient in any time period). Plotting γ over time also allows me to test the hypothesis that it increases at the time of the algorithm update, against the null of no change. A rejection of this null supports the Tribalism Result.

C. Divisiveness

Provided the algorithm update causes a sufficiently large increase in homophily, the Divisiveness Result predicts an increase in the correlation between divisiveness and engagement at the time of the update. Using article-level divisiveness scores imputed with the method outlined in section II, I am able to carry out a test of this prediction using an article-level event study regression. I also use this analysis to investigate heterogeneity in this effect across news outlets of different reliability levels. The specification is displayed in equation 10.

[35]Braghieri et al. (2024) find that within-outlet variation accounts for more of the total article-level slant than between-outlet variation. As I proxy for article slant with outlet slant, this makes my estimation procedure lower power than one which successfully measures article-level slant. I maintain the use of outlet-level slant due to the higher measurement error introduced by article-level slant measures.

[36]The results displayed in section VI use the specification $\gamma \equiv (\gamma^{ac} + \eta^{ac} V_{n\ell})$; I allow the coefficient on $(\text{pol}_n \times \ell)$ to vary with the number of views to allow for the possibility that this may alter the scale of the coefficient over time. I measure actual tribalism of sharing behaviour as $\bar{\gamma} = (\gamma^{ac} + \eta^{ac} \bar{V_{n\ell}})$, where $\bar{V_{n\ell}}$ is the mean number of views per news outlet per time period over the entire dataset. Robustness checks with the more simple specification $\gamma \equiv \gamma^{ac}$ produce similar results.

$$S_{nkt} = \alpha + \beta_1^{nd}\mathbb{1}\{t \geq t^a\} + \beta_2^{nd} r_n + \beta_3^{nd} D_{nkt} + \beta_4^{nd}\big(\mathbb{1}\{t \geq t^a\} \times D_{nkt}\big) + \beta^{ad}\big(r_n \times \mathbb{1}\{t \geq t^a\} \times D_{nkt}\big). \tag{10}$$

where S_{nkt} and D_{nkt} are, respectively, the number of shares and the divisiveness level for article k from news outlet n at time t. A rejection of the null that the coefficient on the first interaction term, $\beta^{nd} = 0$ in favour of $\beta^{nd} > 0$ indicates that the algorithm change increased the correlation between divisiveness and engagement, lending support to the Divisiveness Result. Estimating β^{ad} also allows me to identify heterogeneity in this result across outlets with different reliability levels.

An aggregate measure of divisiveness of shared content across the network can be generated by calculating the average divisiveness in each month weighted by the shares each article received. This can then act as an outcome variable in the second stage of the 2SLS IV estimation.

VI. Results

A. Impact on the reliability of viral content

Figure 5 shows how engagement with each news outlet changed following the algorithm update[37].

The figure shows a division between two groups of outlets: those with a reliability score below around 0.45, which mostly increase in engagement, and those with a reliability score above this level, which mostly decrease in engagement. The raw data thus presents an empirical analogue to the theoretical Reliability Result[38], and seems indicative of at least one clearly defined reliability threshold which separates the news outlets by their change in engagement. In the following paragraphs I formally test for the presence of this threshold using the methods outlined in section V.

The estimates for the empirical specification in equation 4 are presented in table 3. Consistent with the theoretical prediction, the estimated values for β^l is negative in all specifications[39].

We see the ipact of the MSI update was to increase engagement for the lower reliability outlets and either decrease or leave unchanged the engagement for the more reliable outlets.

[37]This graph thus gives a measure of how estimates f_χ vary across news outlets, measured as described in 3. This figure displays 1-year windows. I estimate specifications of my empirical tests with varying window widths.

[38]Figure C5 in the appendix, and the accompanying discussion, digs deeper into this result, showing that the patterns in the change in conditional engagement are in fact consistent with the mechanism for the Reliability Result in Acemoglu et al. (2023), where the drop in engagement for high reliability outlets is most likely the result of a circulation effect.

[39]We can also see that, when political leaning of outlet is included as a control in this specification, the estimated coefficients become more significantly negative. For the remaining specifications, $\widehat{r^f} = 0.125$ is the same level and maintains a similar level of significance; I do not report this in remaining results.

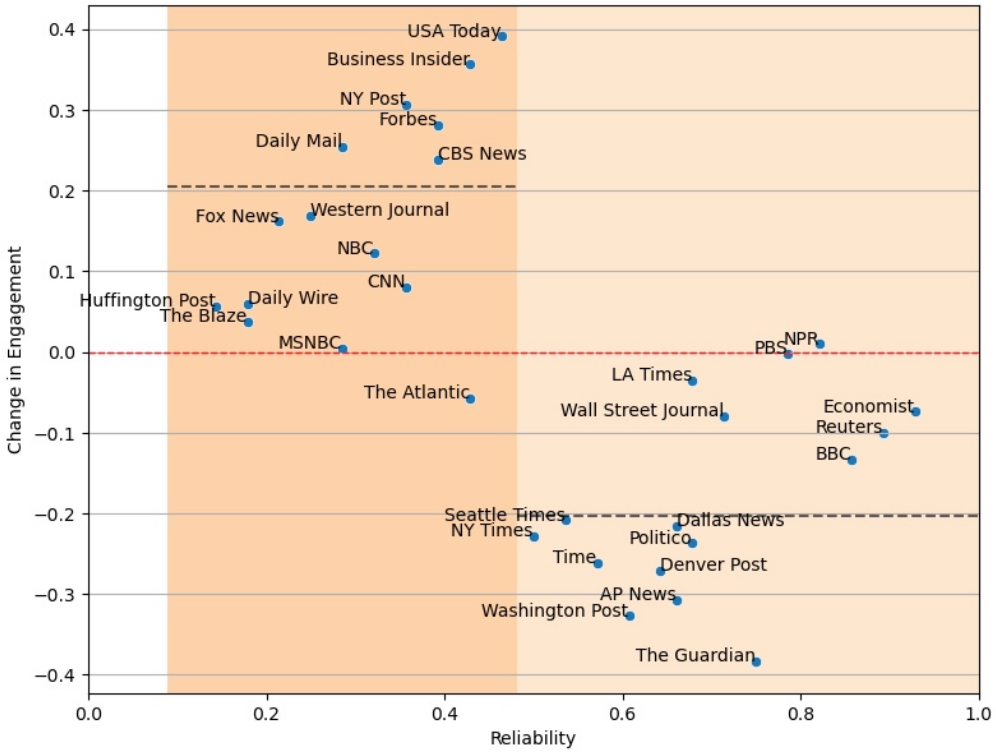


Figure 5. Change in Engagement at time of Algorithm Update, by Reliability of News Outlet

Notes: This figure shows a scatter diagram plotting the reliability of a news article against the log ratio change in engagement going from before to after the January 2018 algorithm update. For this figure and the main estimates displayed in the test of the paper, a 1 year window either side of the algorithm change is used. Yahoo (Change in Engagement 0.8154, Reliability Score 0.143) and the Kansas City Star (Change in engagement -0.61, Reliability score 0.518) are included in estimation but have been excluded from the scatter plot to make the graph clearer. Three outlets lower than the threshold r^f are also excluded from the plots. The colored regions display the fit for specification 7, which divides the outlets into low and high reliability groups; low reliability outlets almost entirely see increases in engagement, and high reliability outlets almost entirely see decreases.

Table 4 displays the results of the threshold estimation. The two thresholds estimated from specification 6 are 0.482 and 0.768[40]. We can also see that the two coefficient estimates for the two-threshold model align with theory - low reliability outlets see a positive change in engagement, and high reliability outlets see a negative change in engagement.

These results are, however, only weakly significant. This is due to the instability of the upper threshold estimate ($\bar{r}$). Both threshold estimates have low p-values derived using the likelihood ratio test; however, the low sample size may render this an unreliable test of the significance of the thresholds. Standard errors estimated by bootstrapping the data indicate that, while the lower threshold estimate seems stable, the upper estimate is far less so[41].

When we instead estimate the one-threshold specification described by equation 7, the $r = 0.482$ threshold is the only one estimated, and we gain significant estimates for β^{ol} and β^{oh}. The data is therefore consistent with the theoretical Reliability Result, supporting a specification

[40]These reliability scores correspond roughly to the reliability levels of *USA Today* and *PBS*, respectively.

[41]The appendix displays the bootstrapped sampling distribution for each estimator in figure C3. $\hat{\underline{r}}$ has a unimodal distribution narrowly centred around the estimate presented in table 4. $\hat{\bar{r}}$ has a bimodal distribution, indicating a stable estimate for this threshold cannot be obtained for a sample this size, and explaining the higher standard error.

Table 3—Reliability Result Linear Specification.

Coefficient	Estimate (SE)			
r_n	-0.662^{***}	-0.697^{***}	-0.393^{*}	-0.430^{**}
	(0.174)	(0.187)	(0.218)	(0.191)
Political Leaning Control	No	Yes	Yes	Yes
Width of time window (months)	12	12	3	1
N	35	35	35	35
R^2	0.395	0.401	0.104	0.163

Source: Coefficient estimates are shown with their estimated standard errors in parentheses. Stars display p value thresholds: * :< 0.1, ** :< 0.05, *** :< 0.01. The table shows the results of estimation of specification 4, where the dependent variable is the change in engagement going from before to after the algorithm update. 'Width of time window' gives the number of months over which average pre and average post engagement is calculated for each news outlet.

Table 4—Reliability Result Threshold Regression.

Parameter	Estimate	Standard Error	p-value
Threshold Estimates			
$\underline{r}$	0.482	$0.060^{\dagger}$	$0.001^{\ddagger}$
$\bar{r}$	0.768	$0.134^{\dagger}$	$0.027^{\ddagger}$
Coefficient Estimates			
Two-threshold model			
β^{tl}	0.205^{***}	0.065	0.076
β^{th}	-0.042	0.128	0.736
One-threshold model			
β^{ol}	0.205^{***}	0.047	0.001
β^{oh}	-0.219^{***}	0.048	0.001

Source: Coefficient estimates are shown with their estimated standard errors in brackets. Stars display p value thresholds on the coefficient estimates: * :< 0.1, ** :< 0.05, *** :< 0.01. The table shows the results of estimation of specification 6 and 7, where the dependent variable is the change in engagement going from before to after the algorithm update. All estimates include political leaning of outlet as a control.
† Standard errors displayed for threshold estimates are bootstrapped standard errors.
‡ p-values displayed for threshold estimates are derived from the likelihood ratio test outlined in V.

where the interval $(\underline{r}, \bar{r})$ is either vanishingly small or unrepresented in this sample of news outlets.

I also test the sensitivity of this estimate of the single threshold to measuring the change in engagement with different time windows around the time of the update. Table 5 displays estimates for the single threshold model as the measurement window either side of the algorithm update is varied.

I now proceed with robustness checks to check whether the results are driven by differences in trends between the two groups of outlets that have been partitioned by the algorithm update: unreliable and reliable. I also investigate whether the result is driven by the behaviour of users from one particular end of the political spectrum, and whether similar patterns are observable for measures of engagement with the news outlets off the Facebook platform.

Table 5—Further Single Threshold Estimates

	Estimate (SE)		
Parameter			
r^R	0.482	0.482	0.518
	(0.060)	(0.091)	(0.095)
Political Leaning Control	Yes	Yes	Yes
Width of time window (months)	12	3	1
N	35	35	35

Source: Coefficient estimates are shown with their estimated (bootstrapped) standard errors in parentheses. The table shows the results of estimation of specification 7, where the dependent variable is the change in engagement going from before to after the algorithm update, and the estimates are of a single threshold which separates the news outlets based on their reliability score. 'Width of time window' gives the number of months over which average pre and average post engagement is calculated for each news outlet.

Robustness Checks. — There are two main concerns that arise in interpreting the results as evidence supporting the theoretical finding. On one hand, the observed correlation between $\widehat{f_h}$ and r_n may be caused by trends in engagement over time which themselves depend on the reliability of a newspaper[42]. On the other hand, there may be some change that occurs to the news media industry more broadly at the time of the algorithm change which is causing the patterns we see in the Facebook data for reasons unrelated to the change to the Facebook network.

To test whether the differential change in engagement between the two groups around the time of the update reflects non-parallel pre-trends, I estimate the event study specification in 8. Figure 6 displays plots of the ϕ_t coefficient over time for each of a number of different variations of this specification.

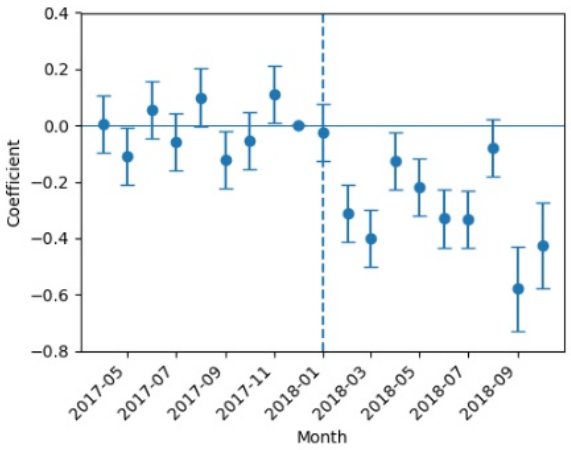
In the top four panels we see the specification for the full engagement data, and then for specifications where the data is filtered to only include engagement metrics of 3 different political leaning groups - moderate users, left leaning users and right leaning users. For all specifications we see a significant drop at the time of the algorithm update in the correlation between a news outlet being flagged as in the 'reliable' group, and the engagement for that news outlet[43].

Panel 6e displays the results for total shares where an alternative measure of news outlet reliability is used (expert assessment, rather than consumer survey based). In panel 6f, we see that a similar effect is not observed when instead we look at the off-platform engagement metric web page visits, indicating this phenomenon is isolated to the Facebook platform (Results for Google searches also show no significant change at the time of the algorithm update).

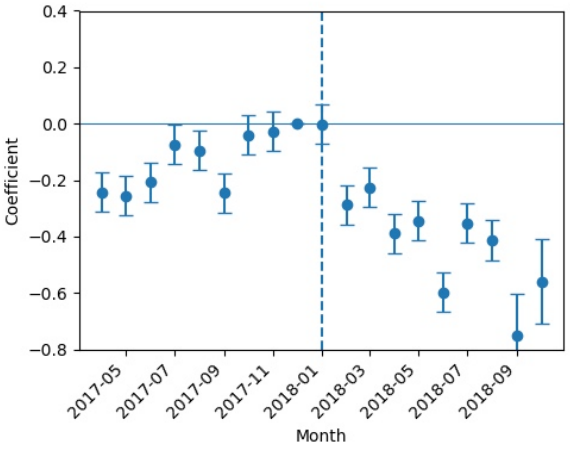
Section C.C2 of the online appendix details a robustness check where I compare total shares for each group of news outlets over time. In section C.C2 I present evidence addressing the concern that the results are driven by spillovers from out of sample news outlets.

[42]Such a possibility is plausible - imagine, for instance, that there are long term trends in media literacy of the US population.

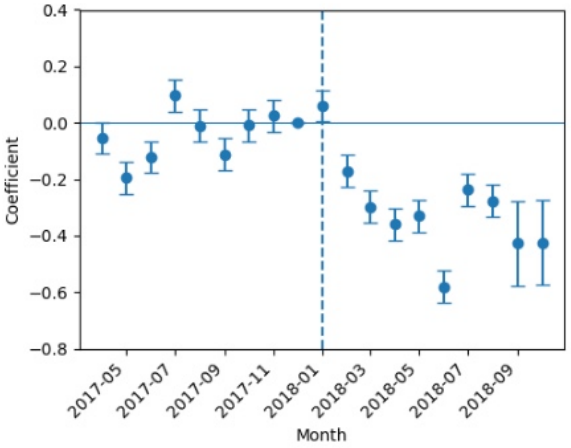
[43]This is persistent for all groups but the right leaning group, where there seems to be some adjustment following 3 months after the algorithm update.



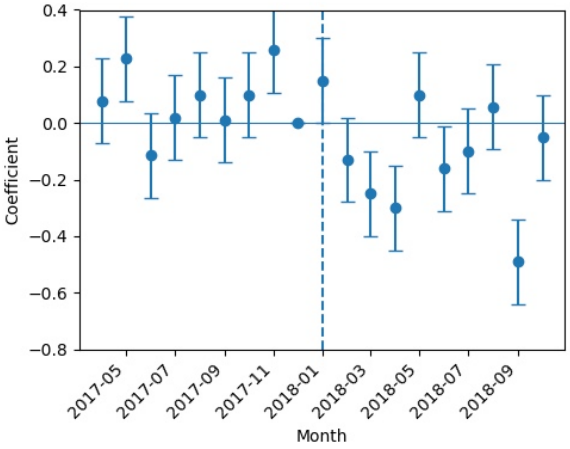
(a) All Users



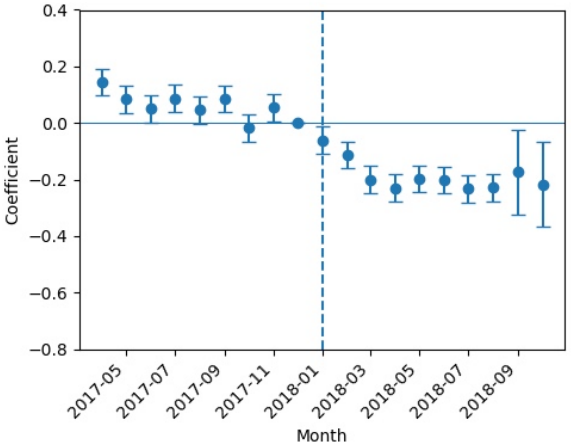
(b) Moderate Users



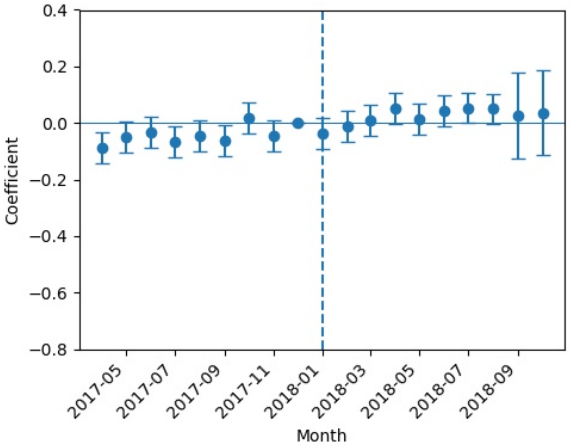
(c) Left Leaning Users



(d) Right Leaning Users



(e) Alternative Reliability Measure

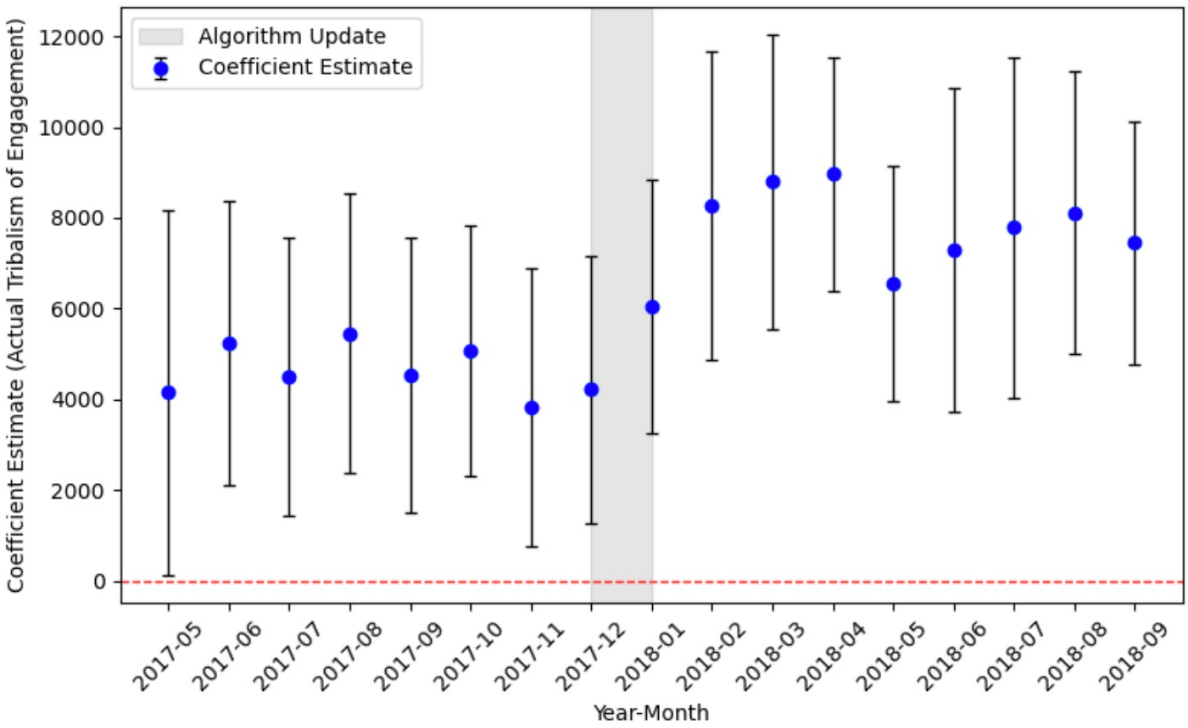


(f) Semrush Data

Figure 6. Results of Event Study Estimates

Notes: Each figure plots the estimated value of ϕ_t from specification 8 for each time period t. Panel 6a plots this using engagement data for all users. Panels 6b, 6c and 6d plot the estimates when the data is filtered to only include engagement metrics for moderate users, left leaning users and right leaning users, respectively. Panel 6e plots the results where an alternative measure of news outlet reliability is used. Panel 6f plots the estimates using the off-platform measures of engagement of home web page visits. 95% confidence intervals are displayed for each coefficient estimate. In each case, the omitted dummy is the December 2017 dummy, and engagement metrics are rescaled for each specification estimate.

B. Impact on the Tribalism of User Behaviour



(a) Change in Actual tribalism of sharing behaviour

Figure 7. Structural change in actual tribalism of sharing behaviour

Notes: This figure plots how estimates for the coefficient γ in regression equation C5 changes over time, along with the coefficient estimate's 95% confidence interval, and displays the structural increase at the time of the algorithm update.

Figure 7 shows how the tribalism of sharing behaviour behaviour changes over time, plotting the coefficient estimates γ_t in each time period t. The figure, first, shows that the tribalism of sharing behaviour is positive and statistically significant in all time periods, consistent with the corollary to the proof of Theorem 1 which shows that we should expect positive tribalism in any network with $p_s > p_d$[44].

The figure also shows a clear, persistent increase in the tribalism of sharing behaviour behaviour at the time of the algorithm update. This is consistent with the theoretical predictions derived above, and is empirical evidence of an 'agitation bubble' phenomenon. Framed in the context of the theoretical framework: when people know their expressed or re-shared opinions are less likely to be scrutinized by people of opposing views, they become more tribal with their expressed views.

The evidence demonstrates that homophily which is exogenously introduced by a social network can itself drive tribalism on the platform, indicating that tribal behaviour on social media

[44]This finding is also consistent with existing empirical evidence of user tendencies to share like-minded news, such as that presented in Pogorelskiy and Shum (2019).

goes beyond being a mere reflection of existing tribal inclinations in wider society. As discussed above, this finding also demonstrates that an important mechanism underlying the polarization and tribalism caused by homophily on social behaviour is a behavioural one, alongside a mechanical one driven by the content people are exposed to through filter bubbles[45].

C. Impact on the divisiveness of viral content

The estimates for the parameters of equation 10 are displayed in Table 6.

Table 6—Divisiveness Event Study Regression Results.

Variable	Estimate (SE)
Reliability × Post × Divisiveness	-197.59^{***} (13.08)
Post × Divisiveness	348.87^{***} (23.61)
Post	-760.96^{***} (88.22)
Divisiveness	862.45^{***} (12.01)
Reliability	-645.62^{***} (50.12)
Observations	563,342

Source: Coefficient estimates are shown with their estimated standard errors in brackets. Stars display p value thresholds: * :< 0.1, ** :< 0.05, *** :< 0.01. The table displays the coefficient estimates from the regression specified by equation 10.

As the table shows, the estimate for the coefficient on the Post × Divisiveness term is positive and significant, consistent with the theoretical prediction that higher homophily increases the virality of divisive content. The negative, significant result on the 3 way interaction term Reliability × Post × Divisiveness further indicates that the increase in divisiveness is highest for the least reliable outlets. This shows how the association between homophily and unreliable news is compounded - when homophily increases, not only do less reliable news sources gain more virality, but it is the most divisive news stories that are gaining the most in virality.

This is, further, consistent with anecdotal accounts from publishers on the impact of the Facebook algorithm update. While consistent with theory, these results warrant more caution than those for reliability and tribalism, due to potential misclassification error in the divisiveness model.

D. Connecting Results to Theory

This section discusses the identifying assumptions under which the empirical estimates can be interpreted as evidence consistent with the comparative statics implied by the theoretical framework.

[45] As I point out when discussing Theorem 1, the theoretical framework attributes this observation to a strategic response by social media users to the change in the social network, and rules out the explanation which appeals to user beliefs changing. In the online appendix I discuss this interpretation in greater detail, and present evidence in its favor. To summarize, as is shown in figure C7 in the appendix, the content to which consumers are exposed did not become more like-minded at the time of the algorithm change (despite the increase in tribal behaviour itself), indicating that the effect observable in 7 is unlikely to be a result of user belief's changing in response to changes in news exposure.

Table 7—Second Stage Estimates.

Outcome Variable	9-month window	3-month window
Reliable Engagement Gap	0.387***	0.338**
	(0.087)	(0.166)
Divisiveness	0.051***	0.035
	(0.010)	(0.022)
Tribalism of Sharing Behaviour	0.436***	0.515**
	(0.056)	(0.131)
Observations	18	6

Source: Coefficient estimates are shown with their estimated standard errors in brackets. Stars display p value thresholds: * :< 0.1, ** :< 0.05, *** :< 0.01. Each row displays the estimated coefficient on the dependent variable log(homophily) for a different regression with a different outcome variable. Homophily is defined as the predicted value for homophily from the first stage estimate using the absolute distance measure (displayed in section IV). Nine year-month observations either side of the algorithm update are included. Each outcome variable is also log-transformed, meaning the displayed coefficient estimates can be interpreted as elasticities.

As shown in Section IV, the algorithmic update led to a plausibly exogenous increase in network homophily on Facebook. In the empirical analysis, I document that, coinciding with this update, (i) the tribalism of sharing behavior increased, (ii) engagement with unreliable news outlets rose, and (iii) engagement with divisive content intensified. The theoretical model predicts that each of these outcomes can arise endogenously from an increase in network homophily, and the observed empirical patterns are consistent with these comparative statics.

To formally test the model's predictions, I implement a two-stage least squares (2SLS) estimation strategy, exploiting the algorithm update as an instrument for homophily. Specifically, I estimate the causal effect of homophily on three dimensions of news engagement—tribalism, divisiveness, and reliability—using monthly variation in these aggregate measures. The results of this estimation are reported in Table 7. The second stage is estimated in log–log form, allowing the coefficients to be interpreted as elasticities of each outcome with respect to network homophily.[46]

The key identifying assumption underlying this approach is an exclusion restriction stating that the algorithm update affects these engagement outcomes solely through its effect on network homophily. This assumption is supported by institutional evidence on the nature of the update (see Section I), which indicates that its primary mechanism was to increase the relative prominence of content shared by close friends and community members—effectively strengthening within-group connections in the Newsfeed ranking algorithm[47].

A potential concern for this identification strategy arises from the contemporaneous Trusted Source update, which could have directly influenced engagement with reliable outlets. I address this issue in detail in the Online Appendix. In brief, because the empirical analysis focuses

[46]Each news engagement outcome is measured as a monthly scalar variable. Divisiveness is defined as the average divisiveness of shared articles in a given month; reliability corresponds to the estimated coefficient ϕ_t; and tribalism corresponds to the coefficient estimate γ_t.

[47]Another assumption underlying this identification strategy is that there was no trend in either homophily or each outcome variable at the time of the algorithm update - time series plots of these variables provided in previous sections provide evidence in favour of this.

on mainstream news outlets, and there is no evidence of substitution from non-mainstream to mainstream sources around the time of the update, it appears unlikely that this concurrent change materially biases the estimated effects. Overall, the available evidence supports the validity of the exclusion restriction and the interpretation of the 2SLS results as causal effects of homophily on news engagement outcomes.

VII. Conclusion

As a higher and higher proportion of news diets become dominated by news which has been accessed via social media, it becomes ever more important to understand the mechanisms which drive outcomes in this media ecosystem.

Using a Facebook algorithm update as an instrument, I demonstrate that a key driver of these outcomes is the homophily which characterizes the social networks of platforms. My IV estimates indicate that an increase in the homophily of a network drives higher engagement for the least reliable news outlets, whilst decreasing or leaving unchanged engagement with more reliable outlets. An analysis of the distribution of engagement across articles within each outlet indicates, additionally, that the increase in homophily favours the most divisive articles published by these unreliable outlets. These findings constitute a striking empirical analogue to the results of the theoretical model of news consumption on social media presented in Acemoglu et al. (2023).

Within the same theoretical framework, I derive an additional comparative static showing that agents exhibit more tribal behavior on more homophilic networks, as the reputational benefit of re-sharing pro-attitudinal news has now increased. This prediction is supported by further IV estimates. The result helps address a simultaneity bias between homophily and tribalism, suggesting that homophily can drive tribalism rather than merely reflecting pre-existing divisions in news engagement. It highlights that echo chambers can give rise to an "agitation bubble" dynamic - distinct from, yet potentially reinforcing, the filter bubble effect emphasized in much of the existing literature. This finding connects the theoretical and empirical results to broader work on group polarization and the emergence of extreme behavior in highly homophilic environments. It also posits a novel mechanism connecting news consumption on social media with polarization. As the mechanism I suggest to rationalize this result does not depend on political attitudes of users shifting, my result is consistent with and complementary to existing empirical evidence showing no significant short term impact of social media news exposure on political polarization.

The likelihood that Facebook's January 2018 algorithm change was a profitable decision further highlights the potential misalignment between platform incentives and the objective of a well-informed population. While the theoretical framework does not explicitly address the

welfare implications of divisiveness and tribalism, both are also a pressing concern for media regulators. Demonstrating that a social media platform can be incentivized to amplify these phenomena reinforces the concerns related to news reliability.

The findings have important implications for the regulation of today's media markets. Social media continues to be a dominant—and rapidly expanding—news source, particularly among young people, who are increasingly turning to social media platforms (recently including the likes of TikTok and RedNote) as a source of news. Recognizing homophily as a key driver of negative outcomes offers a framework for evaluating the potential harms of new social media innovations. It also provides a fresh regulatory perspective for assessing shifts in company policies, such as Facebook's removal of fact-checkers, the change in ownership of X (formerly Twitter), and LinkedIn's most recent algorithm updates.

A promising avenue for future research is to extend the theoretical framework to capture emerging mechanisms of misinformation spread. While this study focuses on mainstream or 'legacy' media outlets, an increasing share of consumers now turn to non-institutional sources - such as podcasts, influencers, and small independent outlets - for information. Understanding how these actors shape and respond to network dynamics will be vital moving forward. Additionally, this paper does not address how news outlets themselves adapt to changes in social media structures; exploring this feedback mechanism would be a key step toward a general market equilibrium of news outcomes in online environments, and allow for a fuller picture of the welfare impacts of social media innovations.

Mathematical Appendix

Below I prove Theorem 1 using the theoretical framework first presented in Acemoglu et al. (2023). In the online appendix to this paper, I give a recapitulation of the framework, which presents a thorough introduction to the theory, its solution concepts and notation. I reference equations in this appendix at points during the following proof.

Proof of Theorem 1 Holding fixed reliability of an article, I demonstrate that increasing homophily increases tribalism μ by showing that, when homophily increases, $Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = R]$ increases and $Pr[a_i = \mathcal{S} \mid b_i < 1/2, m = R]$ decreases. I first show that $p_s \geq p_d$[48] implies that agents in the model will exhibit actual tribalism of sharing behaviour (that is, $\mu > 0$). I then demonstrate that an increase in homophily increases the severity of this tribalism.

Consider a right leaning agent i (with $b_i > 1/2$) and a left leaning agent j (with $b_j < 1/2$) encountering an article with message $m = R$ and some given reliability r. As is pointed out in Acemoglu et al. (2023), it is clear from inspection of equation C2 that:

[48] as is stipulated in the model setup

- π is increasing in b_i
- The payoff to sharing is increasing in π_i, since the first component of utility from sharing $U_i^{(1)}$ is increasing in π_i (as the agent would like to share truthful articles), while $U_i^{(2)}$ is independent of π_i.

These points together imply that the $U_m^{(1)}$ is increasing in b_m. Thus, $U_i^{(1)} > U_j^{(1)}$. Note, further, that, as $p_s \geq p_d$, $S_i \geq S_j$ and $D_i \geq D_j$, and so $U_i^{(2)} \geq U_j^{(2)}$. Therefore, $U_i(a_i = \mathcal{S}) > U_j(a_j = \mathcal{S})$ and $U_i(a_i = \mathcal{D}) < U_j(a_j = \mathcal{D})$, establishing that $Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = R] > Pr[a_i = \mathcal{S} \mid b_i < 1/2, m = R]$ and therefore that $\mu > 0$.

To demonstrate that μ increases as homophily increases, note first that $Pr[b^{\ell_i} > 1/2 \mid b_i > 1/2] > Pr[b^{\ell_i} < 1/2 \mid b_i > 1/2]$; that is, it is more likely that the right leaning agent came from a right leaning island than from a left leaning island. Second, note that $\mathbb{E}[b_m|\ell_m = \ell_i] > \mathbb{E}[b_m|\ell_m \neq \ell_i]$ for any given agent $m \neq i$, meaning a randomly drawn agent from island ℓ_i has a more right wing prior in expectation than an agent drawn from the population excluding those on ℓ_i. Combining this with the same logic from the paragraph above, this establishes that i's co-islanders are on average more likely to share an $m = R$ article than her non-co-islanders.

Now consider two different networks $\mathbf{P'} = (p'_s, p'_d)$ and $\mathbf{P} = (p_s, p_d)$ with $p'_s > p_s$ $p'_d < p_d$. As i's co-islanders are more likely to re-share an $m = R$ article than non-co-islanders, S_i is higher, D_i is lower, and therefore $U_i^{(2)}$ is higher on $\mathbf{P'}$ than on $\mathbf{P}$ (while $U_i^{(1)}$ is constant), and therefore $Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = R]$ is also higher. An analogous argument can be used to demonstrate that $Pr[a_i = \mathcal{S} \mid b_i < 1/2, m = R]$ is also lower on $\mathbf{P'}$ than on $\mathbf{P}$, establishing that $\mu(\mathbf{P'}) > \mu(\mathbf{P})$.

*

REFERENCES

Acemoglu, D., Ozdaglar, A., and Siderius, J. (2023). A Model of Online Misinformation. *The Review of Economic Studies*, page rdad111.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–76.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Aridor, G., Jiménez-Durán, R., Levy, R., and Song, L. (2024). The economics of social media. *Journal of Economic Literature*, 62(4):1422–74.

Bai, J. (1999). Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, 91(2):299–323.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Barberá, P. (2014). How social media reduces mass political polarization. evidence from germany, spain, and the u.s. *https://api.semanticscholar.org/CorpusID:1345712*.

Benjamin, D. J. (2019). Chapter 2 - errors in probabilistic reasoning and judgment biases. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics - Foundations and Applications 2*, volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pages 69–186. North-Holland.

Braghieri, L., Eichmeyer, S., Levy, R., Mobius, M., Steinhardt, J., and Zhong, R. (2024). Article level slant and polarization of news consumption on social media. *Available at SSRN*, 4932600.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Center, P. R. (2024). Social media and news fact sheet. *Pew Research Center*. Accessed online at [https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/].

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. *Fifth International AAAI Conference on Weblogs and Social Media.*

Dasaratha, K. and He, K. (2023). Learning from viral content. *arXiv*.

de Paula, A. (2020). Econometric models of network formation.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fontes, A. (2018). Ad fontes media.

Fraxanet, E., Germano, F., Kaltenbrunner, A., and Gómez, V. (2024). Engagement, content quality and ideology over time on the facebook url dataset.

Garz, M., Sörensen, J., and Stone, D. (2020). Partisan selective engagement: Evidence from facebook. *Journal of Economic Behavior Organization*, 177(C):91–108.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Germano, F., Gómez, V., and Sobbrio, F. (2022). Crowding out the truth? a simple model of misinformation, polarization and meaningful social interactions.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Gong, Q. and Yang, H. (2024). Homophily and spread of misinformation in random networks. *Economic Theory*.

Groeling, T. (2013). Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16(Volume 16, 2013):129–151.

Groseclose, T. and Milyo, J. (2005). A measure of media bias*. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Gruen, A. (2018). Facebook friends? the impact of facebook's news feed algorithm changes on nonprofit publishers. *Media Business Publications, Shorenstein Center*.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., et al. (2023). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656):404–408.

Guriev, S., Henry, E., Marquis, T., and Zhuravskaya, E. (2023). Curtailing false news, amplifying truth. *Amplifying Truth (October 29, 2023)*.

Hagey, K. and Horwitz, J. (2021). Facebook tried to make its platform a healthier place. it got angrier instead. *Wall Street Journal*. Accessed online at [https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215].

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of Public Economics*, 143:73–88.

Hampton, K., Inyoung, S., and Lu, W. (2017). Social media and political discussion: when online presence silences offline conversation. *Information, Communication & Society*, 20(7):1090–1107.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business & Economic Statistics*, 35(2):228–240.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kearney, M. W. (2017). Trusting news project report 2017. *Reynolds Journalism Institute*.

Levy, G. and Razin, R. (2019). Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, 11(Volume 11, 2019):303–328.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(Volume 27, 2001):415–444.

Media, W. (2024). Wallaroom media. *Wallaroom Media*.

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, B., and Wilkins, A. (2020). Facebook Privacy-Protected Full URLs Data Set.

Mosseri, A. (2018). Bringing people closer together. Accessed on [04-12-2024].

Mostagir, M. and Siderius, J. (2022). Learning in a post-truth world. *Management Science*, 68(4):2860–2868.

Muhammed T, S., M. S. (2022). The disaster of misinformation: a review of research in social media. *International Journal of Data Science and Analytics*, 13:271–285.

Olden, A. and Møen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3):531–553.

Papanastasiou, Y. (2020). Fake news propagation and detection: A sequential model. *Management Science*, 66(5):1826–1846.

Patel, N. (2021). Nick clegg doesn't think facebook is polarizing. *The Verge*. Accessed online at [https://www.theverge.com/2021/3/31/22359026/facebook-nick-clegg-newsfeed-medium-decoder].

Pennycook, G. and Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50. The Cognitive Science of Political Thought.

Platforms, M. (2024). Facebook monthly active users (mau) in the united states and canada as of 4th quarter 2023. *Meta Platforms*. Accessed online at [https://www.statista.com/statistics/247614/number-of-monthly-active-facebook-users-worldwide/: :text=As

Pogorelskiy, K. and Shum, M. (2019). Partisan selective engagement: Evidence from facebook. *Available at SSRN: https://ssrn.com/abstract=2972231 or http://dx.doi.org/10.2139/ssrn.2972231*.

Seargeant, P. and Tagg, C. (2019). Social media and the future of open debate: A user-oriented approach to facebook's filter bubble conundrum. *Discourse, Context Media*, 27:41–48. Post-truth and the political: Constructions and distortions in representing political facts.

StatCounter (2024). Leading social media websites in the united states as of march 2024, based on share of visits.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195.

Tonkin, S. (2021). Facebook quietly changed its algorithm in 2018 to prioritise reshared material - then kept it in place despite realising it encouraged the spread of toxicity, misinformation, and violent content, leaked internal documents reveal. *Daily Mail*. Accessed online at [https://www.dailymail.co.uk/sciencetech/article-9997467/Facebook-quietly-changed-algorithm-2018-prioritise-reshared-material.html].

VanderWeele, T. J. and An, W. (2013). Social networks and causal inference. *Handbook of causal analysis for social research*, pages 353–374.

We Are Social, DataReportal, M. (2024). Most popular social networks worldwide as of april 2024, by number of monthly active users (in millions).

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual review of economics*, 12(1):415–438.

Zuckerberg, M. (2018). Jan 2018 algorithm change announcement. Facebook post, accessed on [04-12-2024].

ONLINE APPENDIX

This appendix is intended to be an online appendix to the main text of the paper. It contains additional robustness checks and extensions to the central empirical results which are presented in the paper, as well as a more thorough introduction to the theoretical framework of Acemoglu et al. (2023), which may be of interest to some readers. This appendix will, in any published version of the paper, be available on the author's website. I include it in this version of the paper for the benefit of reviewers and editors.

C1. Online Mathematical Appendix

Here I give a full introduction to the theoretical framework first presented in Acemoglu et al. (2023). Note that this is effectively a recapitulation of the main points of that paper; my theoretical contribution (the proof of the Tribalism Result) is summarized above in the primary appendix. To begin with, I introduce the theoretical framework for the setting of one article about one state of the world θ, as it is applied in Acemoglu et al. (2023). Following this, I discuss the translation of the framework to my setting, where my data covers the consumption of many articles over a period of time.

CONSUMER SHARING BEHAVIOUR. — The true state of the world is $\theta \in \{L, R\}$ and there are N agents. Each agent i has a prior belief $b_i \in [0, 1]$ that $\theta = R$, drawn from distribution with cdf $H_i()$.

Each article has a 3-dimensional type (r, m, ν). Upon seeing an article, each agent observes its reliability $r \in [0, 1]$ and its message $m \in \{L, R\}$, but not its veracity $\nu \in \{\mathcal{T}, \mathcal{M}\}$. The type vector of the article is sampled from the following process:

1) The article has some given reliability score $r \in [0, 1]$.

2) The veracity of the article is drawn as $\nu = \mathcal{T}$ (contains truthful content) with probability $\phi(r)$ or as $\nu = \mathcal{M}$ (contains misinformation) with probability $1 - \phi(r)$. We assume that $\phi(r)$ is increasing and differentiable in r, and satisfies $\phi(0) = 0$ and $\phi(1) = 1$, so that the least reliable article always contains misinformation, and as the degree of reliability increases, the likelihood of misinformation monotonically declines and reaches zero.

3) If $\nu = \mathcal{T}$ (the article is truthful), then its message is generated as $m = \theta$ with probability $p > 1/2$. Conversely, if $\nu = \mathcal{M}$ (the article contains misinformation), then its message is generated as $m = \theta$ with probability $q \leq 1/2$ and is weakly anti-correlated with the truth.

Assume that agents update their beliefs about ν using Bayes' rule given their prior about θ and the observables (r, m) of the article.

Agent chooses one of three actions $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{D}\}$ standing for share, ignore and dislike. Sharing passes the article onto others after her (with probability of the article reaching another agent i determined by the network matrix P). If she ignores it, it is not passed on. If she dislikes it, she expresses disagreement with the content in some way which feeds back to the person who shared it such that it arrived on her newsfeed.

Payoffs are as follows:

$$\text{(C1)} \qquad U_i = \begin{cases} 0, & \text{if } a_i = \mathcal{I} \\ \tilde{u} \cdot \mathbf{1}_{\nu=\mathcal{M}} - \tilde{c}, & \text{if } a_i = \mathcal{D} \\ u \cdot \mathbf{1}_{\nu=\mathcal{T}} - c \cdot \mathbf{1}_{\nu=\mathcal{M}} + \kappa \cdot S_i - d \cdot D_i, & \text{if } a_i = \mathcal{S} \end{cases}$$

Following a decision to share, S, an agent receives utility from two sources. First, agents receive utility from sharing truthful content, but incur a cost from sharing misinformation. This explains the first component of utility following S, $U_i^{(1)} = u \cdot \mathbf{1}_{\nu=\mathcal{T}} - c \cdot \mathbf{1}_{\nu=\mathcal{M}}$.

Second, agents enjoy positive feedback from their peers (further re-shares) but are negatively affected by dislikes. This is represented by the second component of utility $U_i^{(2)} = \kappa \cdot S_i - d \cdot D_i$.

The total utility for agent i's sharing action is the sum of these two components, $U_i^{(1)} + U_i^{(2)}$.

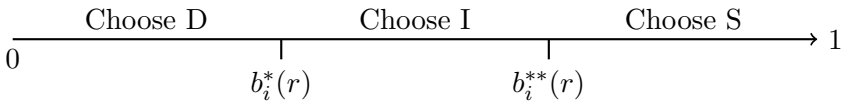
When agent i receives an article with reliability r and message $m = R$ (we normalise the message of the article for exposition), she updates her (ex post) belief, π_i, that the article is truthful according to Bayes' rule:

$$\text{(C2)} \qquad \pi_i = \frac{(pb_i + (1-p)(1-b_i))\phi(r)}{(qb_i + (1-q)(1-b_i))(1-\phi(r)) + (pb_i + (1-p)(1-b_i))\phi(r)}.$$

Clearly, π_i is increasing in b_i since an agent is more likely to believe in an article's veracity when its message agrees with her prior. Moreover, π_i is increasing in r, as the agent updates more on the basis of more reliable articles.

We can also see that the payoff to sharing (S) increases in π_i, since the first component of utility, $U_i^{(1)}$, is increasing in π_i (as the individual would like to share truthful articles), while $U_i^{(2)}$ is independent of π_i. With a similar reasoning, the payoff to disliking (D) is decreasing in π_i, whereas the payoff to ignoring (I) is independent of π_i. This monotone behaviour of payoffs will lead to simple best-response decision rules, as we explain next.

The equilibria of the game are analysed by first proving that all agents employ cutoff strategies whereby they condition their decision of whether to Share, Ignore or Dislike on which of three convex, disjoint subsets their prior belief b_i falls in in the partition of $[0, 1]$.



The authors model homophily and analyse its effects by restricting their attention to a subset of possible network structures, defined as 'island network structures'.

Namely, in an island network, agents are partitioned into k blocks of size $N_1, N_2, \ldots, N_k$, called *islands* each with some constant (but not necessarily equal) share of the population N. Each agent i has a type $\ell_i \in \{1, \ldots, k\}$ corresponding to which block (or "island") she is in. Link probabilities are then given as:

$$p_{ij} = \begin{cases} p_s, & \text{if } \ell_i = \ell_j \\ p_d, & \text{if } \ell_i \neq \ell_j \end{cases}$$

where $p_s \geq p_d$. Without loss, we assume each of the islands is weakly connected.

Second, we assume the prior distribution for agents on the same island ℓ is the same, and is denoted by H_ℓ. We also assume that islands are ranked according to their belief distributions. In particular, each island ℓ has distribution H_ℓ with support on $[b^{(\ell)}, b^{(\ell+1)}]$, where $1 \geq b^{(1)} > b^{(2)} > \cdots > b^{(k)} > b^{(k+1)} \geq 0$. This implies that lower-indexed islands have stronger right-wing beliefs.

Homophily. An important advantage of island networks, in addition to their lower-dimensional representation, is that, combined with this ranking assumption, they enable us to model the degree of *homophily*—the extent to which an individual interacts with others that have common characteristics as herself. Common characteristics for us are those that are relevant for prior beliefs, and therefore, by construction, individuals have more in common with those on the same island as themselves. As a result, homophily will be higher when most links are within islands and links between islands are sparse (high p_s and low p_d).

Divisiveness of Content. We say content with parameters (p', q') is *more divisive* than content with parameters (p, q) if $p \geq p'$ and $q \leq q'$. Divisive content has a message that is more tethered to the true state θ when it is truthful (and more likely to argue against θ if it is misinformation). In our case, we think of state θ as related to political ideology. Therefore, non-political content, such as wedding photos or cat videos, has little divisiveness relative to more political ones.

Tribalism. I characterize a feature of the equilibrium as tribalism - namely, the correlation between the probability of sharing an article and the extent to which the article aligns with the prior belief of the agent. This is defined precisely as

$$\mu = Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = R] - Pr[a_i = \mathcal{S} \mid b_i < 1/2, m = R]$$
$$= Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = R] - Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = L]$$

Tribalism is defined equivalently as either: how much more likely a right wing person is to share right wing content than a left wing person; or how much more likely a right wing person is to share right wing content than left wing content (the equivalence of the definitions is given by the symmetry of the model, which indicates $Pr[a_i = \mathcal{S} \mid b_i < 1/2, m = R] = Pr[a_i = \mathcal{S} \mid b_i > 1/2, m = L]$

News Outlets. — For parts of my analysis, I aggregate article-level data to the news outlet-level. For this reason, I extend the framework to formalize the concept of a news outlet. Although it does not change the intuition regarding any main results (and hence is omitted from the main text of the paper), this technical definition helps to resolve some minor ambiguities regarding the inferences users in the model can make upon observing an article's source.

Assume the state of the world is in fact a vector $\boldsymbol{\theta}$, with elements $\theta_j \in \{L, R\}$, which can be thought of as 'stories'. Each article an agent encounters is written on a particular story, and there are as many stories as there are articles. I redefine an agent i's prior belief as $b_{ij} = Pr(\theta_j = R)$. For each new article, the agent has a new draw from her prior belief distribution H_i, which is constant across all articles and hence not indexed by j. Encountering an article about θ_j causes her to update only her belief about θ_j, ruling out that the agent's prior distribution changes over time. This formulation is consistent with evidence presented in Allcott et al. (2020). There, the authors find that, while exposure to news on Facebook alters an agent's ability to correctly answer factual questions about recent news events, it does not statistically significantly alter affective polarization (negative feelings about the other political party) or polarization in factual beliefs about current events. This evidence measures the short run effects of Facebook news exposure, and hence these findings are appropriate to discipline the model for my setting, where I'm interested in the short run effects of Facebook deactivation [49].

I assume each newspaper n has an exogenously determined target political leaning, defined as the proportion of articles in the newspaper with $m = R$, and denoted pol_n. The editor assigns journalists to research as many news stories as are needed, instructing each to either find an L story or an R story, such that the political leaning target is met. Each journalist researches

[49] The question of whether prior belief distributions may be affected by persistent exposure to pro or counter-attitudinal in the long run is one which requires further research, but goes beyond the scope of this paper.

stories - this consists in an observed draw from the vector $\boldsymbol{\theta}$ which has a constant $\tilde{\phi}_n$ probability of having a message (true state of the world) which matches the instruction the journalist was given[50]. Any journalist who finds a story which does not match the instruction she was given 'spins' the story by writing an article which does not match the true state of the world.

The probability of an article from newspaper n being misinformation is thus $\tilde{\phi}_n$, and defining $\tilde{\phi}_n = \phi(r_n)$ allows us to model reliability r as a property of the newspaper (which is observed by consumers on the demand side of the model).

The purpose of this extension to the model is to provide a precise notion of a newspaper's political leaning in such a way that it can be considered exogenous from a newspaper's reliability. Under this model, a newspaper's reliability is separated from its political leaning by allowing newspapers to have political leaning both via story selection and via 'spin'. This is consistent with research into political media bias which demonstrates that such bias arises via both these mechanisms (Groeling (2013)). In this model, the extent to which a newspaper uses each mechanism is determined by its research capability $\tilde{\phi}_n$.

The Reliability Result generalizes to describe its effects in terms of engagement with particular news outlets[51].

Reliability Result (news outlet level): There exist $0 < \underline{r} < \bar{r} < 1$ such that:

(a) if newspaper n has $r_n < \underline{r}$, greater homophily increases user engagement per article for n.

(b) if newspaper n has $r_n > \bar{r}$, greater homophily decreases user engagement per article for n.

Welfare. — I follow Acemoglu et al. (2023) by assuming that the regulator's welfare objective is related to misinformation and learning, taking account only of the updating of users' beliefs about the true states of the world θ_j. This omits the utility that accrues to social media users via their reputational concerns, potentially omitting the benefits of any increase in 'meaningful social interactions' that were the ostensible target of Facebook's algorithm update. While I concede this renders this only a partial welfare analysis, it is nevertheless consistent with the primary concerns of regulators of social media, and yields important insights with regards to the incentives of social media companies (which plausibly fail to internalize the damages highlighted below).

Let us suppose that users who encounter an article about story θ_j update their prior belief b_ij

[50]Note I also rule out that two journalists from the same newspaper draw the same story

[51]I observe divisiveness at the article level in my data, and the Tribalism Result has already been stated in general enough terms that I do not need to do the same for those results

about θ_j to $\hat{b_{ij}}$ using the same Bayesian updating procedure which creates their ex-post belief π_{ij}. Users who do not encounter this article instead receive an i.i.d (across i) binary signal $s_{ij} \in \{L, R\}$ where $s_{ij} = \theta_j$ with probability $z \in (1/2, 1)$, and update their belief accordingly. The regulator's welfare objective is to minimize the expected average difference between of agents' posteriors from the true state, $-\sum_j \frac{1}{N} \sum_i^N |\hat{b_{ij}} - \mathbf{1}_{\theta=R}|$. I'm mainly concerned with ex-post evaluation of platform policy, and so I assume the regulator has full knowledge of all states of the world.

Theorem 2. *An increase in the reliability gap lowers the regulator's objective.*

Proof of Theorem 2 First, direct attention to the central welfare result (Lemma 2) in Acemoglu et al. (2023):

Lemma 2 in Acemoglu et al. (2023) *There exists $r_{\mathcal{R}} > 0$ such that:*

(i) *if $r < r_{\mathcal{R}}$, then welfare decreases whenever content virality increases;*

(ii) *if $r > r_{\mathcal{R}}$, then welfare increases whenever content virality increases.*

Lemma 2, which applies in the context of a single article, establishes that there is a threshold reliability level below which virality for that article is damaging to the informedness of the user base. Intuitively, the lemma follows from the fact that the 'effective signal strength' of an article (the probability that the article argues for the true state of the world) is monotonically increasing in r, and so there must be a threshold level of r above which the article is more informative than the non-article signal, and below which the article is less informative than the non-article signal.

Theorem 2 then follows from the observation that an increase in the reliability gap can only increase the probability that an article with $r < r_{\mathcal{R}}$ is encountered by any given user.

Intuitively, when less reliable content goes more viral on the platform (that is, when the reliability gap becomes larger) it is more likely that agents encounter lower reliability news, which is in turn more likely to have a lower signal strength than the non-article option z, and hence is more likely to leave agents worse informed than if the article had not been read. This thereby negatively impacts the regulator's objective.

Platform Decision. — The empirical content of this paper analyses the effect of a social media platform increasing homophily. It is not necessary for this analysis for me to expound the full equilibrium of the model where we also endogenise the platform's decision of the level of homophily on its platform. However, as Figure 1 shows that the increase in the homophily of its network that Facebook implemented in January 2018 does appear to have been beneficial to its profits, it is relevant to make a note of Theorem 3 from Acemoglu et al. (2023):

Theorem 3 in Acemoglu et al. (2023) *There exists $\bar{\varepsilon} > 0$ such that if $\varepsilon < \bar{\varepsilon}$, the platform's profit-maximizing sharing network has $k = 2$ islands and is determined by a reliability threshold $r_P \in (0, 1)$ such that:*

(i) *if $r < r_P$, the platform's profit-maximizing sharing network has maximal homophily;*

(ii) *if $r > r_P$, the platform's profit-maximizing sharing network has maximal connectivity;*

Moreover, the reliability threshold r_P increases as divisiveness and/or polarization increases.

As the authors discuss, Theorem 3 highlights an important perversion of a platform's incentives - that it is precisely when articles are likely to contain misinformation that a platform will seek to maximise engagement by increasing homophily and thereby creating filter bubbles and echo chambers. This fundamental misalignment between platform incentives and social welfare, demonstrated theoretically by Acemoglu et al. (2023), provides an additional contextual lens through which to view the empirical results of this paper.

C2. Online Empirical Appendix

Divisiveness classification model. — In this subsection I provide further details on the pipeline used to predict the divisiveness scores of headlines for which scores were not produced using the LLM approach.

The dataset used for training is the one built using the LLM labelling, and thus consists of the article headlines from this subsample, each paired with its divisiveness score. To transform the textual data into a numerical format suitable for machine learning, I employ a sentence embedding approach using the `BertTokenizer` Python package.

Sentence embeddings are dense vector representations of textual data derived from transformer-based models like BERT (Devlin et al. (2019)). The process involves tokenizing each headline into subword units, adding special tokens (`[CLS]` and `[SEP]`) to denote the start and end of a sequence, and passing these tokens through the BERT model. The output is a high-dimensional vector of fixed size (768 dimensions in this case), which captures the semantic and syntactic properties of the input text.

The resulting dataset consists of a matrix where each headline is represented by a 768-dimensional feature vector and a corresponding divisiveness label.

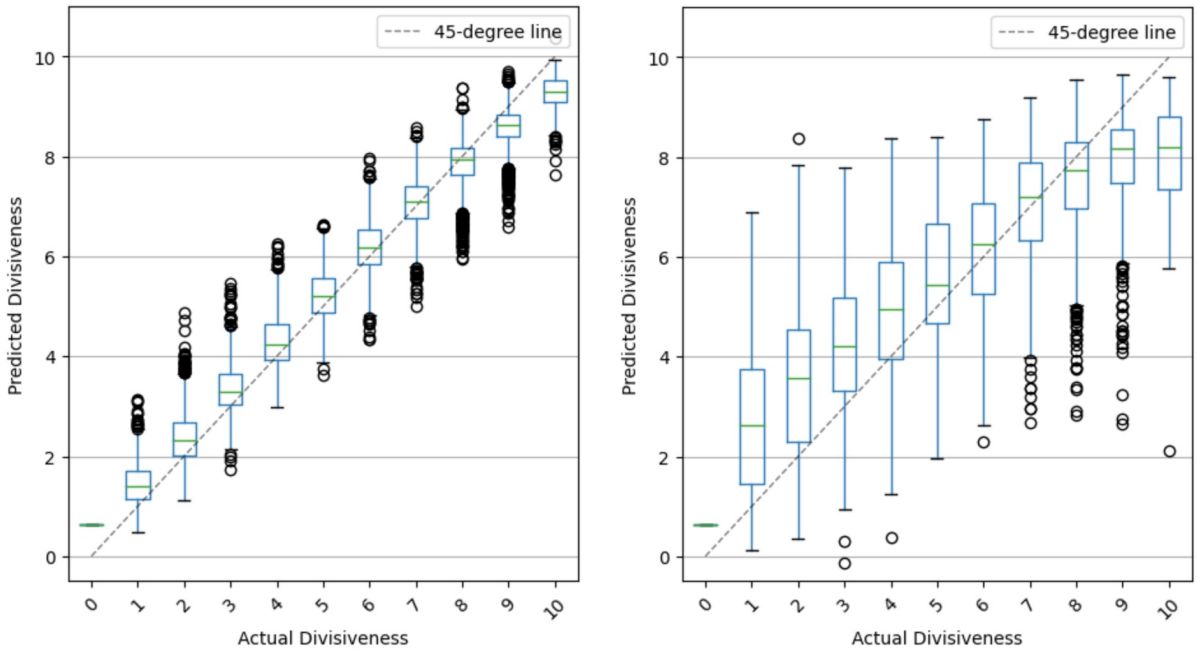
To predict divisiveness scores from the sentence embeddings, I use a gradient-boosted regression tree model implemented with LightGBM (Ke et al. (2017)). Gradient-boosted regression trees are an ensemble learning technique that iteratively combines weak learners (individual

regression trees, restricted to be small in size) to optimize a given objective function. In this case, the objective function minimizes the mean squared error between the predicted and true divisiveness scores. (give a citation for more info in case the reader wants it).

The model was trained on the 768-dimensional sentence embedding vectors as input features and divisiveness scores as the target variable. The training data was split into a test set (a random sample of 1/5 of the data) and a train set. A grid search hyperparameter tuning exercise was carried out using cross validation on the training set as a performance metric to tune the hyperparameters of the gradient boosted tree model: the learning rate, number of trees, and maximum depth of the trees.

To estimate the divisiveness of a new headline, the headline is first converted into a sentence embedding using the same `BertTokenizer` pipeline applied during training. The resulting 768-dimensional feature vector is then fed into the trained LightGBM model to generate a predicted divisiveness score.

The entire pipeline, including data preprocessing, sentence embedding generation, model training, and prediction, was implemented in Python using the `transformers` library for BERT-based embeddings and the `lightgbm` package for regression modeling.



(a) Within-sample performance (b) Out of sample performance

Figure C1. Performance of the divisiveness classifier model

Notes: The figures display the performance of the divisiveness classifier model when it is used to estimate divisiveness scores for headlines from the training sample (Panel (a)) and the holdout (test) sample (Panel (b)). Each chart displays the distribution (with a box plot) of the divisiveness scores that are estimated for headlines with an actual divisiveness of 1, 2, and so on. The 45 degree line is also plotted on each diagram.

Figure C1 displays the performance of the model. Panel (a) illustrates the model's perfor-

mance when it is used to classify the divisiveness of those same headlines on which it was trained (making it a representation of within-sample performance), while Panel (b) displays the model's performance on the hold-out test set that was not included in the model's training - this provides the true check of how well the model performs. As we can see, the model performs fairly well, but clearly introduces further measurement error which we should expect to further bias any estimates based on this data to zero[52].

Total Engagement with news outlet groups over time. — In the main text of the paper, I present event study results testing the reliability result where each news outlet is weighted equally. Here, I perform a similar analysis where I compare total shares over time for the low reliability group of news outlets and the high reliability group of news outlets, to show that the equal weighting of news outlets is not masking large differences in sized between different news outlets in each group.

Panel (a) of Figure C2 displays the plot of total engagement over time for each group of newspapers. The time series plot indicates that the two groups had parallel (slightly downward) trends up until the time of the algorithm update, and then diverged at this point, with the unreliable news outlets (reliable flag = 0) displaying a sharp upward trend, while the reliable outlets group trended downwards in total engagement and then appears to level off. This analysis rules out that the results regarding the Reliability Result are an artifact of non-parallel pre-trends. The graph also shows that the change brought about by the algorithm change is not immediate, but follows a pattern indicative of an uptake effect[53].

While I have ruled out non-parallel pre-trends, there is still the possible (although more remote) concern that some change at the same time as the increase in homophily occurred which could be confounding the estimates. In particular, changes in engagement on the Facebook platform at the time of the algorithm change may be driven by factors external to the social media mechanisms under investigation in this paper, which could themselves be correlated with the reliability of a news source.

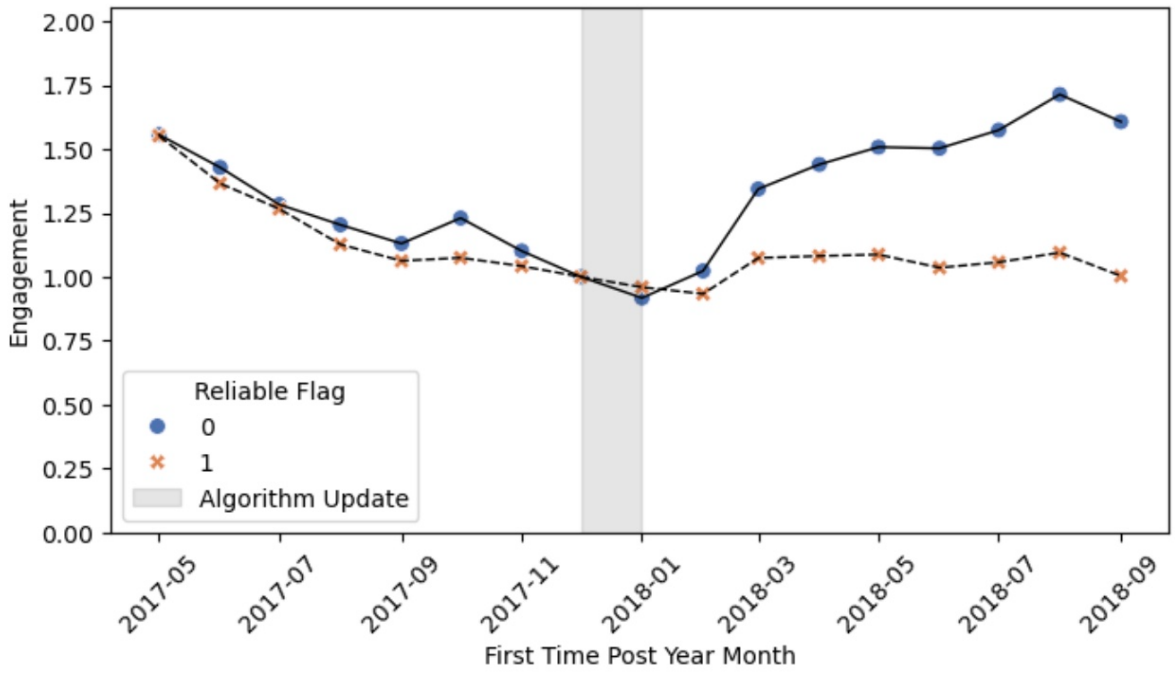
To address this concern, I leverage a triple difference approach, which involves measuring the divergence in engagement between the reliable and non-reliable group on not only the Facebook platform, but also using two other (non-Facebook) measures of user engagement[54]. I observe two other proxy measures of traffic at the news outlet-year month level: web traffic data (from Semrush) and organic search traffic (from Google)[55]. For each alternative data source, I

[52]Interestingly, the model appears to perform particularly badly when rating headlines with either very high or very low divisiveness scores

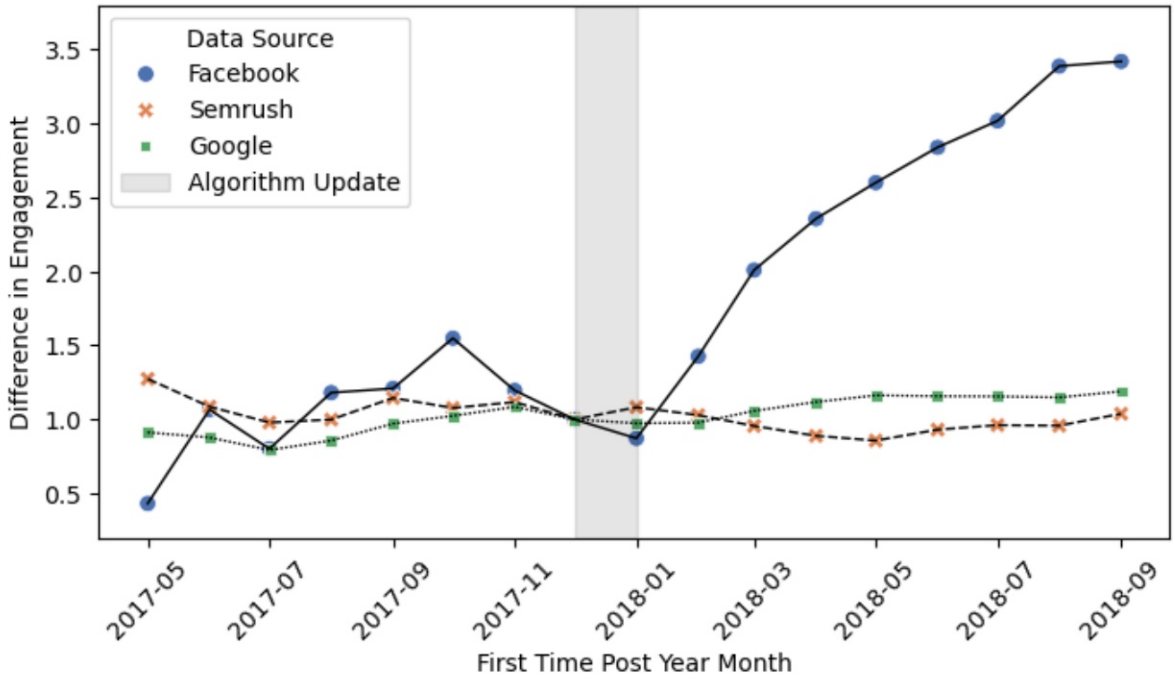
[53]Such a phenomenon is rationalizable when we consider that users on the platform over time learn about the features of the algorithm change via their Newsfeeds, rather than immediately changing their behaviour after the change takes place.

[54]As I present the estimate in the form of a time series plot, whether or not it meets the necessary conditions for its causal interpretation should be transparent. For a thorough treatment of the properties of triple difference estimators, see Olden and Møen (2022).

[55]These provide measures, respectively, of total visits to each domain and total visits which have come via the Google



(a) Engagement over time by Reliability (difference in difference approach)



(b) Difference in Engagement between reliable and unreliable sources, by measure of engagement (triple difference approach)

Figure C2. Time series analysis of the Reliability Result

Notes: Panel (a) displays the trend in total engagement over time for each different group of news outlets: unreliable (reliable flag = 0) and reliable (reliable flag = 1). Total engagement is used to appropriately weight contributions from outlets with different levels of shares long-term. The blue time series in Panel (b) displays the difference between the two lines displayed in Panel (a); that is, the indexed difference in engagement between the reliable and unreliable groups. The orange and green trends on Panel (b) display the equivalent measure (difference in engagement between reliable and unreliable sources), but with engagement measured using two alternative sources (Semrush web traffic data, and Google trends search engine data). All time trends use three month moving averages and are indexed to 1 at 2017-12.

measure the monthly difference in engagement between reliable and non-reliable news outlets; data source then acts as the third dimension for the triple difference estimation.

Panel (b) of Figure C2 plots the indexed difference for each data source over time. The figure shows clearly that the divergence between reliable and unreliable newspapers observed on the Facebook platform is not evident when using other forms of traffic to measure user engagement, indicating this result is unique to the Facebook social network and lending further support to the notion that the change was brought about by the algorithm update to that network.

Both panels in this figure illustrate that the 'reliability gap' was stable up until the algorithm update, but seems to have jumped up in magnitude at exactly the moment the update occurred, just as one would expect given the theoretical predictions of Acemoglu et al. (2023). I will quantify precisely the relationship between this gap in engagement and homophily in the second stage estimates of the 2SLS procedure.

This time series analysis corroborates the Reliability Result, demonstrating that the algorithm update increased engagement with less reliable news outlets, decreasing or leaving constant engagement with more reliable outlets. When we frame this in the light of the welfare results of the theoretical framework, it constitutes strong evidence that the algorithm update had a negative impact on welfare despite having a positive impact on overall engagement (and profits) for Facebook.

The results presented control for the political leaning of outlets. In the online appendix (figure C4) I show that there is no relationship between change in engagement due to the algorithm update, and political leaning of an outlet.

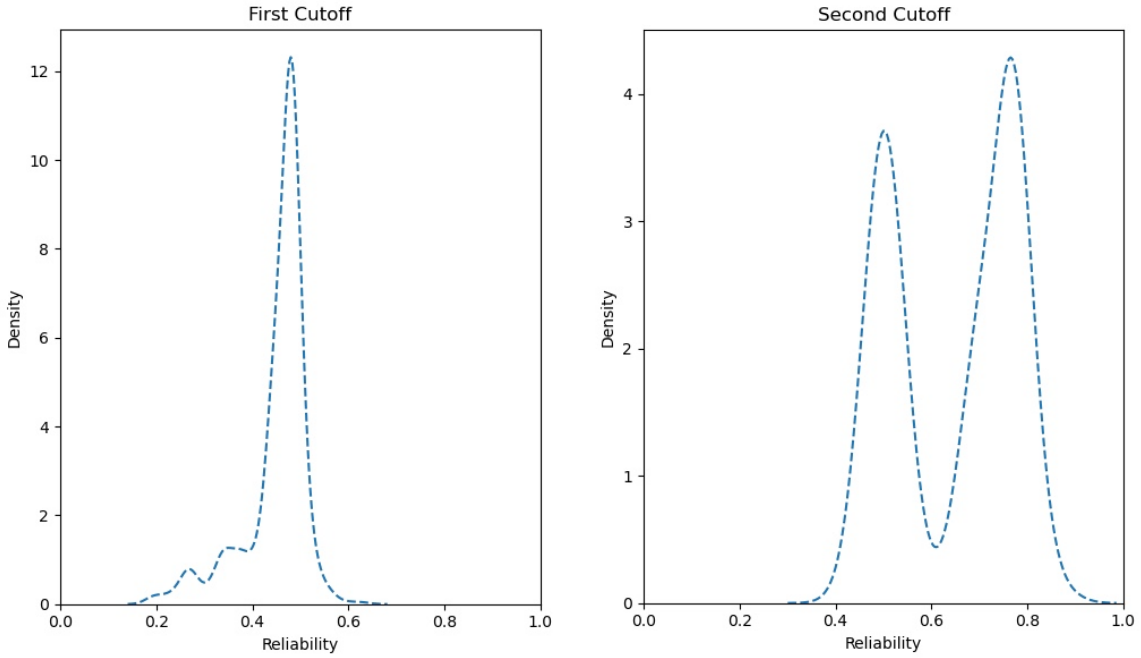
Spillovers from Out of Sample Outlets. — An additional concern which may arise in interpretation of the reliability result is that it is driven by spillovers from out of sample news outlets. A possible explanation of the dynamics following the January 2018 algorithm update is that changes occurred which caused substitution of Newsfeed exposure away from news outlets out of my sample toward outlets within my sample. If this occurred in a way which is correlated with the reliability of in-sample news outlets, it may provide an alternative explanation for the observations. While I cannot, with my data, analyze individual impacts for the entire tail of news outlets on the platform, I can provide indicative evidence based on how market share of the reliable group of news outlets, the unreliable group of news outlets and the out of sample outlets changed at the time of the update. If the increase in engagement for unreliable outlets were driven by spillovers from out of sample outlets, we should expect to see a drop in the market share of engagement in out of sample outlets commensurate with the increase in market share for unreliable outlets. Table C1 presents evidence to the contrary.

search engine.

Table C1—Share of Total News Engagement by Outlet Group, Jan–Mar 2018.

	Share of Total Engagement (%)		
Outlet Group	Jan 2018	Feb 2018	Mar 2018
Unreliable	7.7	8.2	8.6
Reliable	4.7	4.0	3.6
Out of sample	87.6	87.8	87.8

Source: Values represent the proportion of total Facebook news shares accruing to each outlet group in January, February, and March 2018.



(a) First threshold estimate (b) Second Threshold Estimate

Figure C3. Reliability Result analysis

Notes: The two panels in this figure show the distribution of the estimate for each threshold estimated by the process described in section V. Panel (a) displays the distribution of the lower threshold estimate, and Panel (b) of the second estimate. The figures are generated with a 1000-fold bootstrap process, where the full estimation procedure is repeated for each bootstrapped sample.

Bootstrapped Threshold Estimate Distributions. — Figure C3 displays the empirical bootstrapped distributions for the estimator of the lower reliability threshold (Panel (a)) and the higher reliability threshold (Panel (b)). The lower threshold estimator distribution is unimodal at a value between 10 and 15, with a tight distribution and a standard error of 1.676, indicating this threshold estimate (13.5, as is reported in the main body of the paper) is stable across repeated bootstrapped samples. The second threshold estimate is bimodal, sometimes taking a value very close to the first threshold estimate and sometimes a value between 20 and 25. This indicates the estimate for the second threshold estimate (21.5, as is reported in the main body of the paper) is far less stable (standard error 3.744), owing either to there being no true second threshold or low sample size.

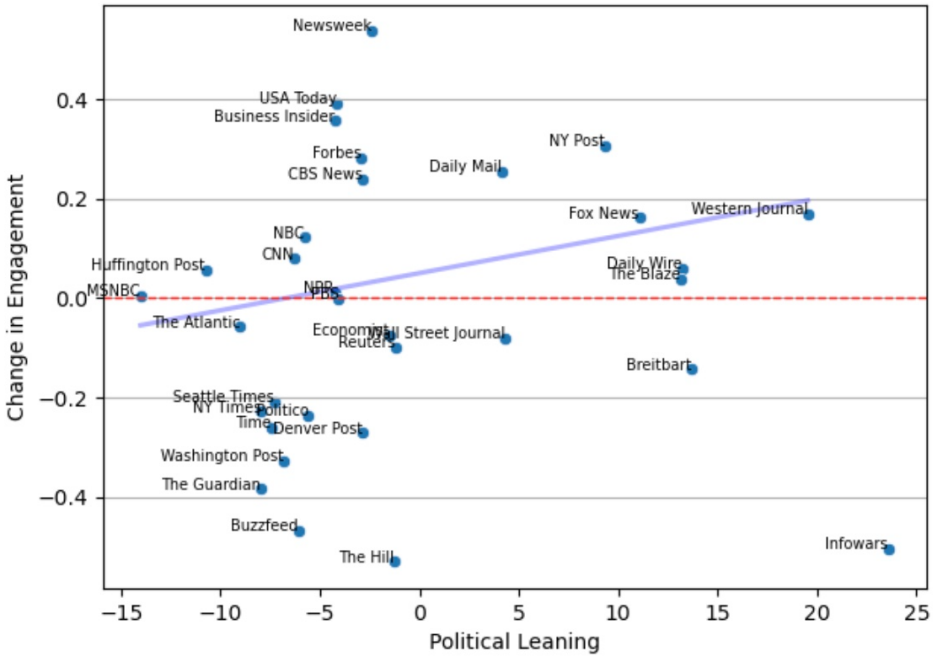


Figure C4. Plotting political leaning against change in engagement.

Notes: The figure displays a plot similar to those displayed in Figure 5, but where instead of plotting reliability on the horizontal axis, I plot political leaning. The figure is intended as a placebo test to check the possibility that results are not driven by the political leaning of news outlets. Yahoo (Change in Engagement 0.8154, Political leaning -5.69) has been included in estimation but excluded from the scatter plot to make it clearer. The fit line excludes those outlets with $r < \widehat{r^f}$ (as estimated by the previous estimation), for consistent comparison with the main analysis of reliability.

Reliability Analysis Political Leaning Robustness Check. — By including a control for political leaning in table 3, I have already shown that the empirical counterpart to the Reliability Result holds, conditional on political leaning of a news outlet. Figure C4 provides further transparency by plotting the change in engagement for each outlet against the political leaning for that outlet.

The estimated relationship between the two variables (excluding those outlets with $r < \widehat{r^f}$ (as estimated by the previous estimation), for consistent comparison with the main analysis of reliability) is mildly positive but insignificant (p-value 0.279), indicating that there is no strong relationship between political leaning and change in engagement.

This is consistent with the theory, whose results are driven entirely by reputational concerns which are rational and symmetric across the political spectrum, and which do not depend on partisan differences in assessment of news outlet reliability.

Mechanism Underlying the Reliability Result. — Figure C5 plots the reliability of each news outlet against the change in *conditional* engagement. That is, the change in shares per view going from before to after the algorithm update. This gives a measure of how much the likelihood of someone sharing a news source from this news outlet conditional on seeing it in their newsfeed changed.

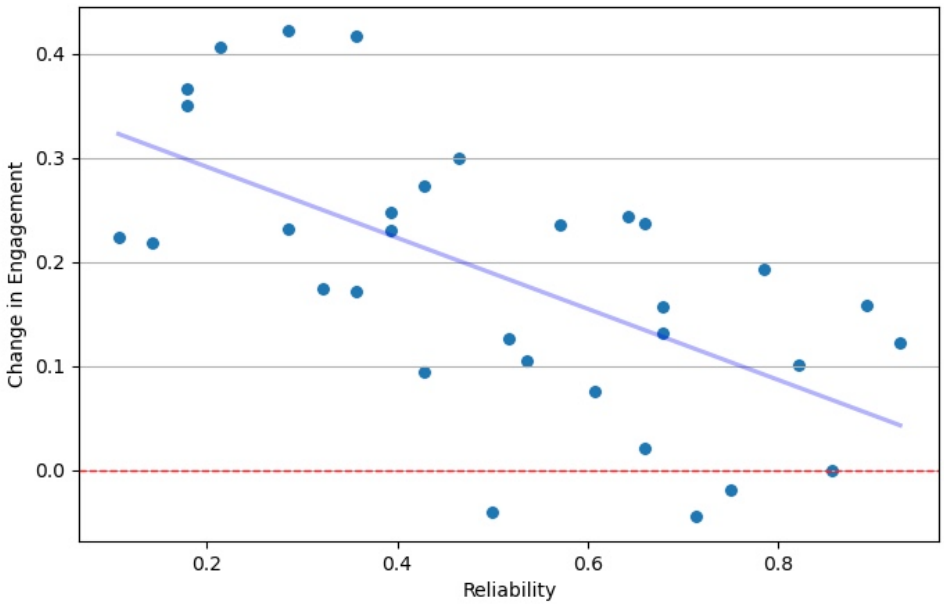


Figure C5. Relationship between reliability and conditional engagement.

Notes: The figure plots the reliability of a news outlet against the amount by which the conditional engagement for that outlet changed (that is, the engagement per view on the Newsfeed). News outlets with $r < \widehat{r^f}$ have been omitted.

As we can see, while this measure increased by more for the less reliable outlets, it did in fact increase for all nearly outlets. This shows that the drop in engagement we see in Figures 5 for the more reliable news outlets is not due to consumers being less likely to share articles conditioning on seeing them, but because they are seeing less of them.

This is in fact consistent with the mechanism underlying the Reliability Result in Acemoglu et al. (2023). In the proof for this result, the authors point out that increasing homophily increases engagement with articles by decreasing the 'discipline effect' on agents who encounter these articles; they are more likely to share a low reliability article, conditional on seeing it, because they know those who see their share of the article are more likely to share their biases. That is, conditional on seeing an article, a consumer becomes more likely to share that article, *regardless of the reliability.* This is consistent with the empirical finding presented in C5.

The reason that the engagement of high reliability articles can go down is the 'circulation effect' - even though consumers haven't decreased their proclivity to share any articles conditional on seeing them, in a more homophilic network, articles can end up getting 'stuck' being circulated around an island with a low probability of 'escaping' that island. Because high reliability articles cause a more dramatic updating of the prior, the discipline effect is not enough to outweigh the circulation effect, and engagement can decrease. The results presented in section VI and here are thus consistent with the underlying theoretical mechanism of the Reliability Result.

C3. Further Discussion of Tribalism

In the main body of the paper, I define tribalism of sharing behaviour as the extent to which right (left) wing people are more likely to share right (left) leaning content on the platform, *conditional on seeing such a piece of content.* To discuss the tribalism of sharing behaviour in more detail, I distinguish it from what I refer to as *apparent* tribalism of sharing behaviour, which is the unconditional counterpart to tribalism of sharing behaviour. Along with this, I define tribalism of the *Newsfeed* as the extent to which right (left) wing people are more likely to be exposed to right (left) leaning content on their Newsfeed.

Apparent tribalism of sharing behaviour can be high either because actual (conditional) tribalism of sharing behaviour is high, or because tribalism of the Newsfeed is higher. That is, at the aggregate, 'apparent' level, we might observe users engaging more with like minded content either because of a behavioural change (where they are actually more likely to share it, conditional on seeing it), or just because they are being shown more of it [56].

Apparent tribalism of sharing behaviour and tribalism of the Newsfeed are both objects of interest in their own right; they are often the focus of regulators and political commentators. Understanding how (conditional) tribalism of the Newsfeed has changed helps us to better understand the mechanisms underlying the apparent tribal behaviour on social media - it is also the phenomenon which maps most naturally to the theoretical framework, and so is the focus of attention in the main body of the paper. Here, I produce empirical estimates of all three phenomena using my data.

The clearest theoretical prediction regarding the algorithm update and tribalism is that the increase in homophily should drive the (conditional) tribalism of sharing behaviour up. This aims to demonstrate that homophily (echo chambers) drives tribal behaviour, rather than the causation only going in the other direction.

Conversely, the update's effect on apparent tribalism of sharing behaviour is ambiguous. We should expect the increase in actual tribalism of sharing behaviour to increase apparent tribalism directly as well as have an upward effect on the tribalism of the Newsfeed. However, at the same time as increasing homophily, the MSI algorithm update lowered the prevelance of publisher shared content (that is, content which is shared directly by a news outlet and effectively 'broadcast' to users via their newsfeed). We should expect the drop in publisher-shared content to lower the tribalism of the Newsfeed, as users will see less content from publishers they have themselves subscribed to. As such, the aggregate effect on the tribalism of the Newsfeed and therefore the apparent tribalism of sharing behaviour is theoretically ambiguous, and depends

[56] While more work has been done on the ways in which social network algorithms change outcomes via the latter, mechanical route (see, for example, Germano et al. (2022)), the focus of my contribution is on the behavioural changes that can be induced by network structure.

on the balance of these effects.

Apparent tribalism of sharing behaviour of the network can be measured in any particular time period t by estimating γ^{ap} in equation C3

$$S_{n\ell} = \xi_n + \zeta_\ell + \gamma^{ap}(\text{pol}_n \times \ell) + \varepsilon^{ap}_{np} \tag{C3}$$

where $S_{n\ell}$ is the number of shares newspaper n receives from political affinity group ℓ in period t; ξ_n and ζ_p are fixed effects for newspaper and political affinity group. The intuition for this measure of tribalism is that the term $\text{pol}_n \times \ell$ will be high when the newspaper's ideology and the agent's ideology are a close match, and low when they are a poor match. A positive γ^{ap} indicates that there is apparent tribalism of sharing behaviour.

Allowing $V_{n\ell}$ to be the number of times a newspaper n's articles are viewed on the Newsfeed by consumers with political affinity ℓ, I can construct a similar measure of the tribalism of the newsfeed by estimating γ^{nf} in

$$V_{n\ell} = \xi_n + \zeta_\ell + \gamma^{nf}(\text{pol}_n \times \ell) + \varepsilon^{nf}_{n\ell}. \tag{C4}$$

The measurement of conditional tribalism of sharing behaviour is described in the main body of the paper, and repeated here. It can be measured by altering C3 to condition on the number of views. I do so by estimating $\tilde{\gamma}$

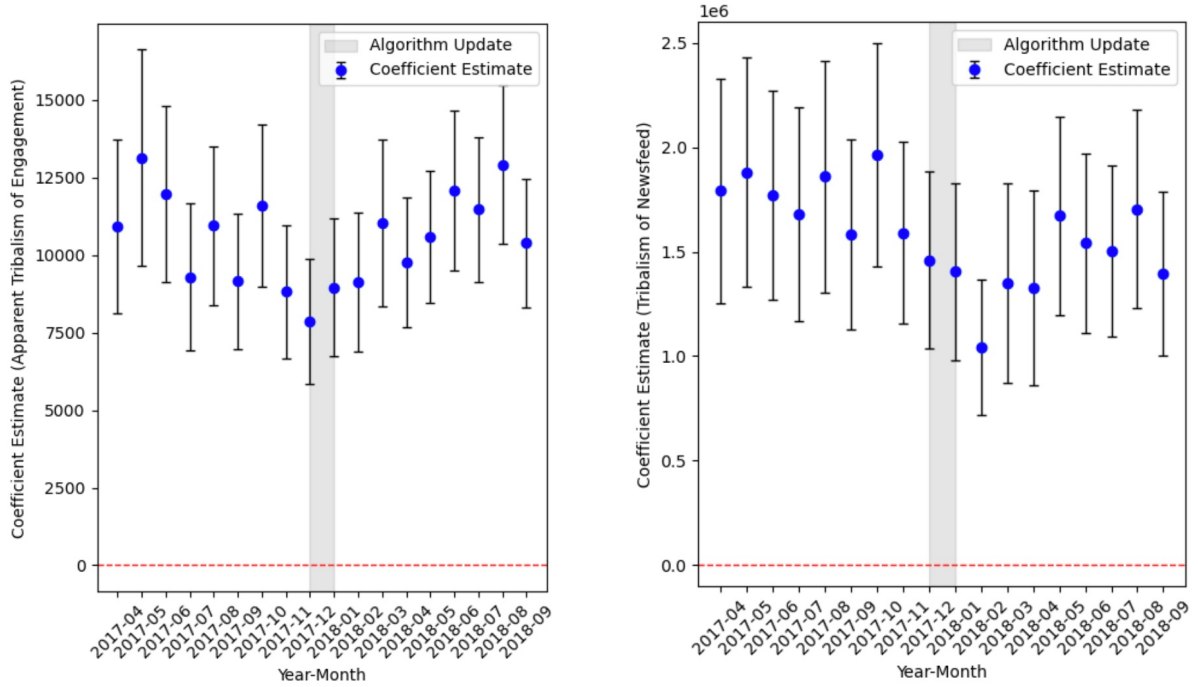
$$S_{n\ell} = \xi_n + \zeta_\ell + \tilde{\gamma}(\text{pol}_n \times \ell) + \beta^v V_{n\ell} + \varepsilon^v_{n\ell} \tag{C5}$$

Where $\tilde{\gamma} \equiv (\gamma^{ac} + \eta^{ac} V_{n\ell})$; I allow the coefficient on $(\text{pol}_n \times \ell)$ to vary with the number of views to allow for the possibility that this may alter the scale of the coefficient over time. I measure actual tribalism of sharing behaviour as $\bar{\tilde{\gamma}} = (\gamma^{ac} + \eta^{ac} \bar{V_{n\ell}})$, where $\bar{V_{n\ell}}$ is the mean number of views per news outlet per time period over the entire dataset.

The estimate of γ^{ap} in equation C3 is by construction subject to bias due to the failure to condition on views. However, the apparent tribalism of sharing behaviour (exhibited as an output of the combination of consumer behaviour and the network algorithm) is of interest in and of itself as it is a measure of the aggregate extent to which agents on social media engage with like minded content, so I include results related to this measure.

Theory predicts that γ^{nf}, γ^{ap} and $\tilde{\gamma}$ should all be positive in all time periods, which can be

tested by rejecting the null of 0 coefficients in any time period. Plotting $\tilde{\gamma}$ over time also allows me to test the hypothesis that it increases at the time of the algorithm update, against the null of no change. A rejection of this null supports the Tribalism Result.



(a) Apparent tribalism of sharing behaviour

(b) Tribalism of newsfeed

Figure C6. Main caption for the figure, describing both panels.

Source: Panel (a) plots how estimates for the coefficient γ^{ap} in regression equation C3 changes over time, along with the coefficient estimate's 95% confidence interval, displayed by the vertical bars. Panel (b) plots the same information for coefficient γ^{nf} in regression equation C4.

Figure C6 displays how the apparent tribalism of sharing behaviour (γ^{ap}, Panel (a)) and the tribalism of the newsfeed (γ^{nf}, Panel (b)) have changed over time. We see positive, significant coefficients across all time periods in both panels. This demonstrates, consistent with existing literature, that we observe both phenomena of apparent tribalism of sharing behaviour and tribalism of the newsfeed, corroborating existing empirical literature related to these phenomena.

At the time of the algorithm change, we see apparent tribalism of sharing behaviour increase even though, at the same time, we see no increase in the tribalism of the Newsfeed. Panel (a) of Figure C7 shows the two time trends on the same graph, with the error bars removed to make the figure clearer.

As I pointed out above, theory does not provide a guide to how we should expect these measures to move in response to the algorithm change. We can posit that the failure of Tribalism of the Newsfeed to increase is most likely due to the suppression of publisher-shared content on the platform, as posited in section I. The increase we see in the apparent tribalism of sharing

behaviour despite this change is thus all the more surprising.

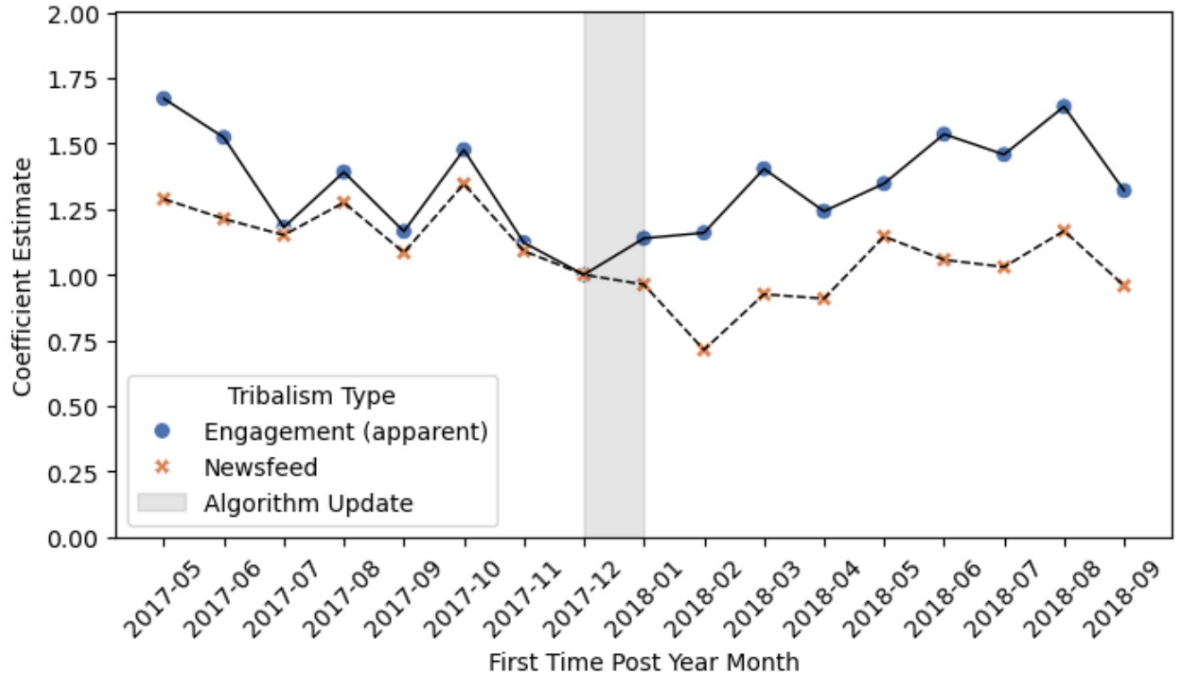
Figure C7 demonstrates that, at the time of the algorithm update, the trends in apparent tribalism of sharing behaviour and tribalism of the Newsfeed cease to track each other. This demonstrates that the increase in (apparent) tribalism that we see at the time of the algorithm update is driven by a change in the sharing behaviour of Facebook users, and happens despite the fact that the news being shown to consumers in fact became less tribally targeted as a more mechanical result of the drop in publisher-shared content.

The fact that the tribalism of the Newsfeed does not increase indicates that the increase in (conditional) tribalism of sharing behaviour cannot have happened via a change in user beliefs, which is consistent with the model, where prior belief distributions for any particular user are fixed. The change to the tribalism of sharing behaviour happens purely through the strategic behavioural change, as posited by the theoretical framework.

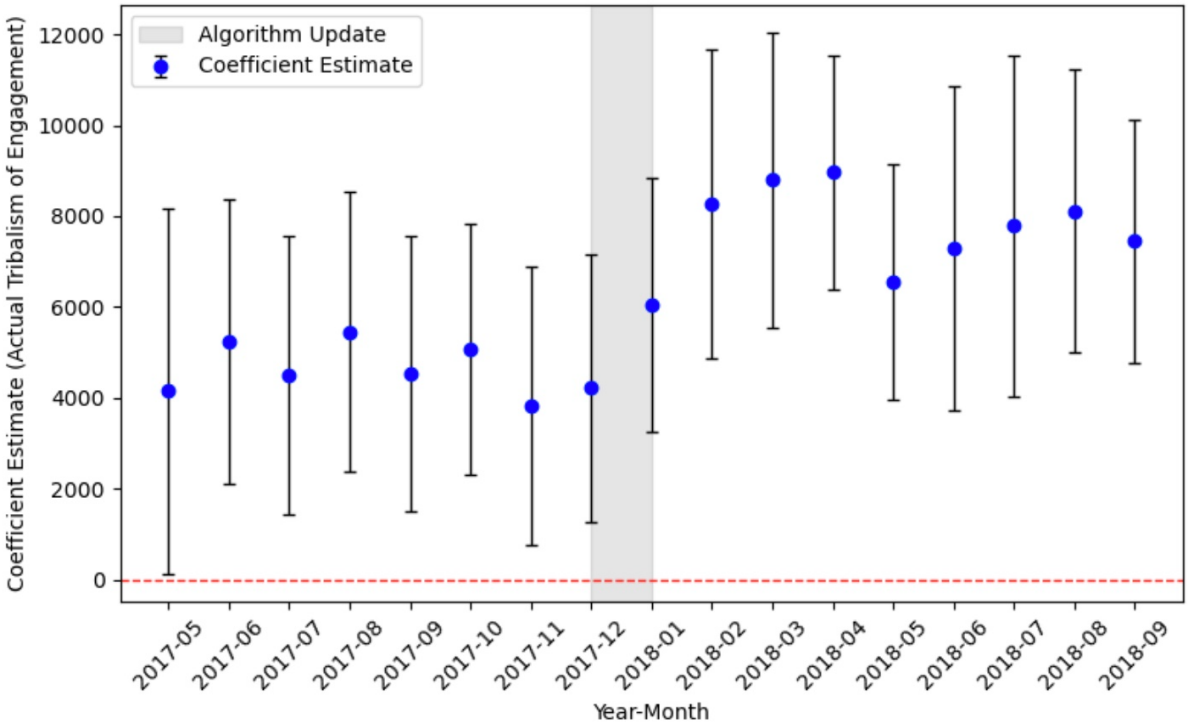
C4. Data Summaries

Domain List. — Table C2 provides a list of all news outlets used in the sample.

Example Headline Divisiveness Scores. — Table C3 gives an example headline for each possible divisiveness rating, as it was rated by the LLM.



(a) Change in Apparent tribalism of sharing behaviour and of the Newsfeed



(b) Change in Actual tribalism of sharing behaviour

Figure C7. Structural change in actual (conditional) tribalism of sharing behaviour

Notes: Panel (b) plots how estimates for the coefficient γ^{ac} in regression equation C5 changes over time, along with the coefficient estimate's 95% confidence interval, and displays the structural increase at the time of the algorithm update. As is pointed out in the main text, this measures the tribalism of sharing behaviour on the platform. The term 'actual tribalism of sharing behaviour' refers to tribalism engagement conditioning on a view (this distinguishes it from a different concept which is introduced and discussed in the online appendix: apparent tribalism of sharing behaviour).

Domain	Reliability Score	Political Leaning
BBC	0.89	-1
Breitbart	0.07	14
Business Insider	0.46	-4
Buzzfeed	0.04	-6
CBS News	0.43	-3
CNN	0.39	-6
Daily Mail	0.32	4
Daily Wire	0.21	13
Denver Post	0.67	-3
Economist	0.96	-1
Forbes	0.42	-3
Fox News	0.25	11
Huffington Post	0.18	-11
Infowars	0.11	24
LA Times	0.71	-6
MSNBC	0.32	-14
NBC	0.35	-6
NPR	0.86	-4
NY Post	0.39	9
NY Times	0.54	-8
Newsweek	n/a	-2
PBS	0.82	-4
Politico	0.71	-6
Reuters	0.92	-1
Seattle Times	0.57	-7
The Atlantic	0.46	-9
The Blaze	0.21	13
The Guardian	0.79	-8
The Hill	n/a	-1
Time	0.61	-7
USA Today	0.50	-4
Wall Street Journal	0.75	4
Washington Post	0.64	-7
Western Journal	0.29	20
Yahoo	0.14	-6

Table C2—News Outlet List

Headline	Divisiveness Rating
M&M's Has A New Nutella-Esque Flavor And I'd Like 20 King-Size Bags Please	1
This Waterpark Campground In Minnesota Belongs At The Top Of Your Summer Bucket List	2
NBC Is Saving Brooklyn Nine-Nine So Maybe Not Everything Is Garbage	3
Sperm count drop 'may lead to human extinction'	4
San Juan National Forest closes for the first time in 113 years as 416 fire continues to grow	5
ABC Hit With Boycott For Canceling 'Last Man Standing'	6
Canada Now Wants U.S. To Enforce Its Immigration Laws – To Protect Canada	7
We've fallen off a cliff': Scientists have never seen so little ice in the Bering Sea in spring	8
Anti-gun student walkout included stomping on American flag and jumping on cop car	9
Trump gives Liberal Snowflakes a new thing to bitch about. Withdraws U.S. from Paris climate accord.	10

Table C3—Example Headline Divisiveness Ratings