Crowding out the truth?
A simple model of misinformation, polarization, and meaningful social interactions[*]

Fabrizio Germano[†] Vicenç Gómez[‡] Francesco Sobbrio[§]

November 2022

Abstract

Social media are at the center of countless debates on polarization, misinformation, and even the state of democracy in various parts of the world. An essential feature of social media is the ranking algorithm that determines how content is presented to the users. This paper studies the dynamic feedback between a ranking algorithm and user behavior, and develops a theoretical framework to evaluate the effect of popularity and personalization parameters on measures of platform and user welfare. The model shows the presence of a fundamental trade-off between platform engagement and user welfare. A higher weight assigned to online social interactions such as likes and shares and to personalized content, increases engagement while having a detrimental effect in terms of misinformation—*crowding-out the truth*—and polarization. Besides increasing actual polarization, an increase in the weight assigned to social interactions may also increase perceived polarization, as it makes it more likely for individuals to see more extreme content—both like-minded and not—in higher-ranked positions. Finally, we provide empirical evidence in support of the main predictions of our model. By leveraging a rich survey dataset from Italy and exploiting Facebook's 2018 "Meaningful Social Interactions" update—which significantly boosted the weight given to social interaction in its ranking algorithm—we find an increase in political polarization and ideological extremism in Italy following the change in Facebook's algorithm.

Keywords: Social media, ranking algorithm, engagement, misinformation, polarization, popularity ranking, personalization, algorithmic gatekeepers.

[*] We thank Roland Bénabou, Alexander Frug, Gaël Le Mens, and audiences at the 2022 AI+Economics Workshop (ETH Zurich), the 8th International Conference on Computational Social Science in Chicago (IC^{2}S^2 2022) and at the 5th Economics of Media Bias Workshop (Berlin 2022) for helpful comments. Fabrizio Germano acknowledges financial support from Grant PID2020-115044GB-I00//AEI/10.13039/501100011033 and from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S). Francesco Sobbrio is grateful to Nando Pagnoncelli and IPSOS for allowing access to the data of the Polimetro.

[†] Department of Economics and Business, Universitat Pompeu Fabra, and BSE, fabrizio.germano@upf.edu

[‡] Department of Information and Communications Technologies, Universitat Pompeu Fabra, Barcelona, vicen.gomez@upf.edu

[§] Department of Economics and Finance, Tor Vergata University of Rome and CESifo, francesco.sobbrio@uniroma2.it

1 Introduction

Recent revelations by whistleblowers at Facebook have once again brought to the attention of the public the risks and dangers associated with the algorithms of digital platforms to manage their informational content.[1] The algorithms used by social media like Facebook, Twitter or Instagram, or by search engines like Google and Bing decide what information to show to users and, importantly, also in what order to show it. Indirectly, they determine what information is more or less relevant for any given user. A rapidly growing body of empirical research has documented how social media platforms may foster polarization and misinformation (Allcott et al., 2020; Di Tella et al., 2021; Levy, 2021) sometimes associated with a tangible impact (Bursztyn et al., 2019; Müller and Schwarz, 2020, 2021; Amnesty International, 2022; Ananthakrishnan and Tucker, 2022). There are journalistic and academic claims suggesting that such adverse effects may be a consequence of the way profit-maximizing social media platforms design their algorithms (CNN, 2021; Lauer, 2021), namely with the objective of ensuring a high level of engagement (Liao et al., 2017).

This paper contributes to the above discussion by providing a theoretical framework—and related empirical evidence— showing that algorithmic rules that tend to be desirable from the perspective of social media platforms may be detrimental to their users and, more broadly, to the health of democracy. We build on and extend the work of Germano et al. (2019) and Germano and Sobbrio (2020) to develop a model where a platform ranks news items (e.g., posts, tweets, etc.), while individuals sequentially access the platform to decide which news items to click on and possibly "highlight" (e.g., like, share, retweet). At the center of our model, there is a ranking algorithm that decides the order of news items to be displayed to any given user. The ranking evolves according to the *popularity* of news items, which is a weighted combination of the clicks and highlights received by a given news item. Simply put, the more people click on and the more people highlight a news item, the higher the probability that the news item will go up in the ranking and will be displayed in a higher-order position. The model also allows assessing the role of *personalization*, that is, when the platform provides different rankings to different individuals. The objective is to study the dynamic feedback between ranking algorithms and individual behavior.

To preserve tractability, individuals' choices over which news items to click on and highlight are modeled as driven by behavioral traits that are rooted in ample empirical evidence. For the clicking choices, we assume that individuals have a preference for choosing news items that are higher ranked (Pan et al., 2007; Novarese and Wilson, 2013; Glick et al., 2014; Epstein and Robertson, 2015) and with positive probability, they also have a preference for choosing news confirming their prior beliefs (Gentzkow and Shapiro, 2010; Yom-Tov et al., 2013; White and Horvitz, 2015; Flaxman et al., 2016). For the highlighting choices, we

[1] See, for example, `https://www.wsj.com/articles/the-facebook-files-11631713039`

assume that individuals highlight news items with some probability, provided they are sufficiently close to their prior beliefs (An et al., 2014; Pogorelskiy and Shum, 2019; Garz et al., 2020) and are more likely to do so when they hold more extreme prior beliefs (Bakshy et al., 2015; Grinberg et al., 2019; Pew, 2019; Hopp et al., 2020).

Armed with this theoretical framework, we assess the impact of popularity-driven and personalized rankings on platform engagement, misinformation, and polarization. We show that, when optimizing with respect to the ranking algorithm, social media platforms face a trade-off between platform profitability and user welfare: increasing the weight given to highlights in the popularity ranking increases engagement, which is desirable from the platform's perspective, yet may be detrimental from a public policy perspective, as it leads to higher levels of misinformation—*crowding-out the truth*—and polarization. Accordingly, our results suggest that ranking algorithms may have contributed to generating the "missing middle" (Bartels, 2016) and affective polarization (Iyengar et al., 2019) in the political realm. We also point out the presence of a related insight. Besides increasing actual polarization, a boost in the highlighting weight may also increase *perceived* polarization, since it makes it more likely that an individual will see more extreme content—both like-minded and not like-minded—in higher-ranked positions. This is very much in line with Bail (2021) who argues that social media act as a "prism" that conveys a distorted image of others and ends up muting moderates, while fueling actual and perceived polarization (see also Yang et al. 2016). The trade-off between social media engagement and user welfare also speaks directly to the empirical evidence by Beknazar-Yuzbashev et al. (2022) who show that curbing toxic content on social media platforms leads to a reduction in content consumption. Concerning personalization, the results also show the presence of a trade-off between engagement and polarization. That is, increasing the degree of personalization in the ranking algorithm is conducive to both a higher level of engagement and a higher degree of polarization. This is consistent with the empirical literature on the link between personalization and polarization (Levy, 2021; Dujeancourt and Garz, 2022; Huszár et al., 2022).

We further provide direct empirical evidence in support of our main theoretical insights. We leverage a rich survey dataset from Italy and exploit Facebook's "Meaningful Social Interaction" (MSI) algorithmic ranking update implemented in January 2018, which significantly increased the weights given to likes and shares in order to boost engagement.[2] We estimate a Differences-in-Differences empirical model comparing the ideological extremism and affective polarization of people interviewed after the Meaningful Social Interaction (MSI) algorithm was introduced (i.e., January-June 2018) and that use internet to form an opinion, relative to those of people also using internet to form an opinion who were interviewed before such a change (i.e., June-December 2017) and at the same time relative to people interviewed after such

[2] See https://www.facebook.com/business/news/news-feed-fyi-bringing-people-closer-together and https://edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math.

change in the algorithm who were not using internet as one of the main sources to form an opinion.[3] As predicted by our theoretical model, the results show that Facebook's 2018 MSI update led to an increase in ideological extremism and affective polarization in Italy.[4]

To the best of our knowledge, this is the first paper to explore both theoretically and empirically how an algorithmic boost given to highlighted content may affect platform engagement and social welfare. The model generalizes and extends Germano et al. (2019); Germano and Sobbrio (2020) essentially by giving individuals the option to highlight content, thereby affecting the ranking, besides allowing for a broader set of individual clicking behavior and also larger signal spaces. The model is complementary to the one of Acemoglu et al. (2022) who focus on endogenous social networks and fact-checking.[5] As in our model, Acemoglu et al. (2022) show that platforms have an incentive to increase personalization (more homophilic communication patterns) as this increases platform engagement. In their setting, this is detrimental in terms of social welfare as it increases the level of misinformation. Instead, in our case, more personalization increases polarization yet it does not affect the overall level of misinformation, since our model does not embed the issue of fact-checking and cannot, therefore, capture such an effect. At the same time, because we explicitly model the endogenous dynamic ranking used by social media platforms, we are instead able to provide insights on the incentives—and possible perverse effects on social welfare—of such platforms to boost the weight given to content highlighting in their ranking algorithm.

2 The Model

At the center of the model is a digital platform characterized by its ranking algorithm, which ranks and directs individuals to different news items (e.g., websites, Facebook posts, tweets), based on the popularity of individuals' choices. Such news items may be used by individuals to obtain information on an unknown, cardinal *state of the world* $\theta \in \mathbb{R}$ (e.g., net benefit of a vaccine, consequences of inaction on global warming, adequacy of a presidential candidate, etc.).

The ranking of each news item is inversely related to its popularity, where the popularity is determined by the number of clicks and the number of "highlights" received by a given item (e.g., likes received by a Facebook post/number of shares, like/retweets of a tweet, etc.). Each click has a weight of one, and each "highlight" has an additional weight of $\eta \geq 0$. We briefly illustrate the working of the model before

[3] Around the time of the MSI update, Facebook was by far the first social network in Italy with 34 million active users per month and a penetration rate of almost 80% with respect to the population of Italian internet users (We Are Social, 2018). As such, the use of internet to form a political opinion provides, in the context of Italy between 2017 and 2018, a meaningful proxy for exposure to Facebook content.

[4] The theoretical predictions of the model pointing out the role of social media algorithms in fostering misinformation, are also consistent with Vosoughi et al. (2018) providing evidence that false stories spread faster than true ones on Twitter. Similarly, Mosleh et al. (2020) points out the presence of a negative correlation between the veracity of a news item and its probability of being shared on Twitter.

[5] See also Azzimonti and Fernandes (2022) for a model of diffusion of misinformation on social media via internet bots.

entering into some more details regarding the ranking algorithm.

2.1 Individual Clicking and Highlighting Choices

There are N *individuals*, each of which receives a private informative signal on the state of the world, $x_n \in \mathbb{R}$, which is drawn randomly and independently from $N(\theta, \sigma_x^2)$ (we use $f(x; \sigma_x^2)$ to denote the corresponding density function). There are also $M > 2$ *news items* that individuals can click on, each of which carries an informative signal on the state of the world $y_m \in \mathbb{R}$, also drawn randomly and independently from $N(\theta, \sigma_y^2)$ (we use $f(y; \sigma_y^2)$ to denote the corresponding density function).

2.1.1 Individual clicking choices absent ranking

To model individuals' clicking choices, we assume there is a benchmark $\widehat{\theta} \in \mathbb{R}$—non-informative with respect to θ—which allows individuals to sort news items into "like-minded" or not. Specifically, we assume that, leaving aside the order of news items provided by the ranking algorithm, individuals are able to see whether a news item is reporting a "like-minded" information or not. Yet they need to click on the news item in order to see the actual signal y_m. This assumption is meant to capture a rather typical situation, where individuals observe the "coarse" information provided in the landing page by the platform (e.g., infer the basic stance of a news item, whether Left or Right, pro or anti something, from the website title, Facebook post intro, first tweet in a thread, etc.), yet, in order to learn the actual content of the news (i.e., the cardinal signal y_m) and update her beliefs, the individual has to click on the news item.

We formally translate this setting into assuming that an individual is able to observe whether her own signal x_n and the news items' signals y_m are above or below $\widehat{\theta}$. Accordingly, for each individual, the signal x_n has an associated binary signal indicating whether x_n is above or below $\widehat{\theta}$: $\text{sgn}(x_n) \in \{-1, 1\}$, where $\text{sgn}(x_n) = -1$ if $x_n < \widehat{\theta}$ and $\text{sgn}(x_n) = 1$ if $x_n \geq \widehat{\theta}$. Similarly, for each news item, the signal y_m has an associated binary signal $\text{sgn}(y_m) \in \{-1, 1\}$, where $\text{sgn}(y_m) = -1$ if $y_m < \widehat{\theta}$ and $\text{sgn}(y_m) = 1$ if $y_m \geq \widehat{\theta}$. Let M_- and M_+ denote the sets of news items with binary signal respectively -1 and 1 (by slight abuse of notation, we use $\#M_-$ and $\#M_+$ or sometimes directly M_- and M_+ to denote the number of news items in M_- and M_+ respectively). Thus, given the individual's signal x_n, the benchmark $\widehat{\theta}$ allows the individual to sort news items into "like-minded" or not, before actually clicking or reading. For most of the paper, we focus on the case where the benchmark separates signals in roughly symmetric groups: $\widehat{\theta} \approx \theta$.[6]

At the same time, as mentioned, individuals have to actually click on news item m in order to learn its cardinal signal y_m. In particular, we assume that, absent ranking effects, the individual's choice about

[6] Say, $|\widehat{\theta} - \theta| < \min\left\{\frac{\sigma_x}{4}, \frac{\sigma_y}{4}\right\}$. In Appendix.2.1 we discuss the case where individuals have heterogeneous benchmarks $\widehat{\theta}_n$; Appendix.2.2 discusses the case, where $\widehat{\theta}$ and θ are far apart.

which news item to click on depends on her *clicking type* ($k \in \{C, E, I\}$). To encompass all possible clicking behaviors, we consider three clicking types:

- *confirmatory type* ($k = C$): clicks with propensity γ_C on a news item with the same sign as her own signal $\operatorname{sgn}(x_n)$ and with propensity $1 - \gamma_C$ on one of opposite sign (where $1/2 < \gamma_C < 1$);
- *exploratory type* ($k = E$): clicks with propensity γ_E on a news item with the same sign as her own signal $\operatorname{sgn}(x_n)$ and with propensity $1 - \gamma_E$ on one of opposite sign (where $0 < \gamma_E < 1/2$);
- *indifferent (purely ranking-driven) type* ($k = I$): clicks with equal propensity $\gamma_I = 1/2 = 1 - \gamma_I$ on an item of either sign.

These three types occur with probabilities, respectively, $p_C, p_E, p_I \geq 0$, such that $p_C + p_E + p_I = 1$.[7] While we allow for all three types, all the key results of the model hold even if we were to focus on only one or two of the types.[8]

The binary signal $\operatorname{sgn}(x_n)$ together with the individual's clicking type k determine the *propensity with which individual n of clicking type k will click on an item m, absent ranking*:

$$\varphi_{n,m}(k) = \begin{cases} \gamma_k/[m] & \text{if } \operatorname{sgn}(x_n) = \operatorname{sgn}(y_m) \\ (1-\gamma_k)/[m] & \text{if } \operatorname{sgn}(x_n) \neq \operatorname{sgn}(y_m), \end{cases} \tag{1}$$

for $k \in \{C, E, I\}$, and where $[m] = \#M_-$ if $\operatorname{sgn}(y_m) = -1$ and $[m] = \#M_+$ if $\operatorname{sgn}(y_m) = 1$.

2.1.2 Individual clicking choices with ranking

So far, the ranking of the items did not enter the clicking and highlighting choices. We now allow the individual choice function to take into account that individuals see the news items ordered by the ranking $r_n = (r_{n,m})_{m \in M}$, where $r_{n,m}$ is the rank of news item m as seen by individual n. We assume individuals have an *attention bias* calibrated by the parameter $\beta > 1$, with the interpretation that, a news item of equal sign but placed one position higher in the ranking, has a likelihood β times larger to be clicked on than the one in the lower position (Pan et al., 2007; Glick et al., 2014; Novarese and Wilson, 2013). Together with the propensity to click, absent ranking, these jointly determine the probability with which

[7] Similar to the literature on political economy that parametrizes the fraction of different types of voters (Krasa and Polborn, 2009; Krishna and Morgan, 2011; Galasso and Nannicini, 2011), the model does not micro-found the individuals' clicking choices. At the same time, it is easy to see that the confirmatory type might be driven by a preference for like-minded news (Mullainathan and Shleifer, 2005; Bernhardt et al., 2008; Gentzkow and Shapiro, 2010; Sobbrio, 2014; Gentzkow et al., 2015). Yom-Tov et al. (2013); Flaxman et al. (2016); White and Horvitz (2015) provide empirical evidence on confirmation bias by users of digital platforms. Similarly, the exploratory type might be the by-product of incentives to cross-check different information sources (Rudiger, 2013; Athey et al., 2018). Finally, the indifferent type allows us to consider the role of individuals with a high attention bias or search cost (Pan et al., 2007; Glick et al., 2014; Novarese and Wilson, 2013).

[8] Note that we assume clicking types to be independent of individuals' priors x_n, but when formalizing the "highlighting" behavior, we allow highlighting types to be correlated with the priors.

individuals click on news items. Define the *probability of individual n of clicking type k to click on news item m* as:

$$\rho_{n,m}(k) = \frac{\beta^{(M-r_{n,m})}\varphi_{n,m}}{\sum_{m' \in M} \beta^{(M-r_{n,m'})}\varphi_{n,m'}}. \tag{2}$$

2.1.3 Individual highlighting choices

After clicking on a given news item m, the individual sees the actual signal $y_m \in \mathbb{R}$ and then decides whether or not to *highlight* such a news item (e.g, like, share, retweet, etc.). This depends on the individual's *highlighting type* ($h \in \{A, P\}$). We consider two highlighting types:

- *passive type* ($h = P$): never highlights a news item regardless of her signal;
- *active type* ($h = A$): highlights a news item if and only if the news item's signal is sufficiently close to her own signal, $y_m \in H(x_n)$,

where $H(x_n) \equiv [x_n - \sigma_x/2, x_n + \sigma_x/2]$ and σ_x is the standard deviation of the individual's signal x_n. We assume the active and passive highlighting types occur with probabilities $p_A, p_P \geq 0$, respectively, where $p_A + p_P = 1$.

Importantly, we consider two alternative cases for the *probability of being an active type* (p_A):

- *flat case*: p_A is a constant in $(0, 1)$;
- *non-flat case:* p_A is a function of the signal received by the individual given by:

$$p_A(x_n) = 1 - e^{-\frac{1}{2\alpha}\left(\frac{x_n - \widehat{\theta}}{\sigma_x}\right)^{2\alpha}}, \quad \alpha \geq 1\,. \tag{3}$$

Thus, both in the flat and in the non-flat case, we always assume that an individual highlights only if the news item reports a signal sufficiently close to her prior ($y_m \in H(x_n)$) (An et al., 2014; Pogorelskiy and Shum, 2019; Garz et al., 2020). Moreover, in the flat case, the probability of highlighting a news item (p_A) is independent of the individual's signal x_n (again, provided that the news item reports a signal sufficiently close to individual's prior). By contrast, in the non-flat case, the highlighting probability is correlated with the individual's signal. Specifically, p_A increases with the (square of the) deviation of x_n from the benchmark $\widehat{\theta}$, normalized by σ_x. And, while the specific functional form assumed in Eq. (3) is not crucial for our results, what does matter is that individuals with more extreme priors are more likely to be active and hence to highlight a given news item y_m.[9]

Figure 1 shows the assumed signal distributions of individuals (black dashed line) and of news items (gray dashed line). It also shows, for the non-flat case, the probability of being an active type (orange dashed line) and the resulting highlighting propensity (red solid line).

[9] Results are also clearly robust to replacing θ for $\widehat{\theta}$ in Eq. (3).

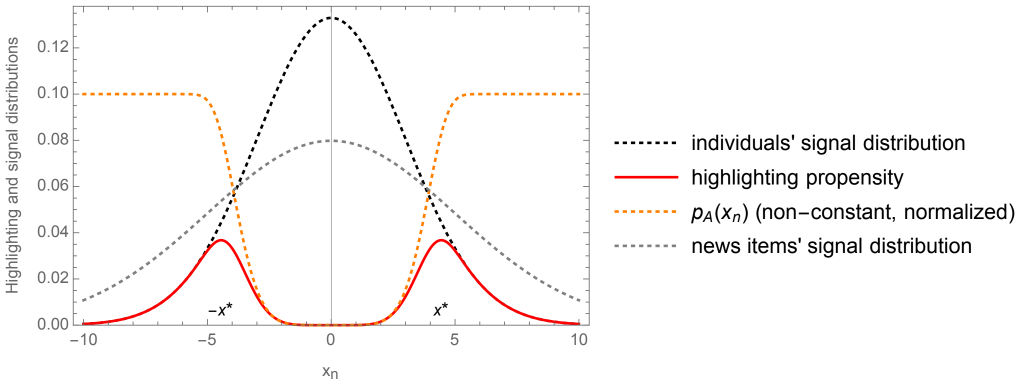

Figure 1: Individuals' signal distribution and highlighting propensity in the non-flat case, and items' signal distribution for $\widehat{\theta} = \theta = 0$ and $\sigma_x^2 < \sigma_y^2$; $-x^*, x^*$ denote the values of x_n where the highlighting propensity is locally maximal.

As with the clicking choice, individuals' highlighting choice is not derived from a maximization problem. Nevertheless, the correlation between extreme beliefs and propensity to highlight in the non-flat case is reminiscent of the link between overconfidence and ideological extremism modeled by Ortoleva and Snowberg (2015). Importantly, the non-flat case—which is our main focus—is rooted in observed empirical regularities on how individuals with more extreme ideological beliefs tend to be more actively engaged and also more likely to highlight political news items on social media platforms (Bakshy et al., 2015; Grinberg et al., 2019; Pew, 2019; Hopp et al., 2020). In particular, by using data on over 10 million Facebook users in the US, Bakshy et al. (2015) provide evidence on the ideological distribution of shared political news.[10] The data clearly show a bimodal distribution with large mass on the tails of the distribution, i.e., individuals with more extreme preferences account for a larger proportion of the overall shared content on Facebook. Remarkably, such a bimodal distribution is present only when looking at the distribution of shares related to contents defined by Bakshy et al. (2015) as "hard" information (e.g., national news, politics, world affairs). By contrast, no such bimodality is present when looking at "soft" information (e.g., sport, entertainment, travel). This suggests that the bimodal distribution observed in the shares of "hard" information is unlikely to be driven by large tails in the ideological distribution of Facebook's users (i.e., a bimodal distribution of Facebook users' ideology) or by a larger density of the network in such tails. Rather, such a bimodal distribution of shares is likely to be driven by users with more extreme ideological preferences having a higher propensity of sharing political contents.[11]

[10] More specifically, they measure the ideological alignment of content shared on Facebook as the average affiliation of sharers weighted by the total number of shares. Similar evidence is presented by the authors when weighting by the total number of distinct URL shared.

[11] See also Grinberg et al. (2019); Pew (2019); Hopp et al. (2020) for additional empirical evidence on the positive correlation between extremist political preferences and the propensity to share political news in social media.

2.2 Platform and Ranking Algorithm

We consider *popularity-based rankings* that evolve as a function of the clicking and highlighting behavior of the individuals, as well as *personalized rankings* that may depend on the identity of the individuals.

2.2.1 Popularity ranking

After each individual makes her choices, the algorithm updates the popularity of each news item such that a click has a weight of 1 and a highlight has a weight of $\eta \in \mathbb{R}_+$. That is, starting from $\kappa_{0,m} \in \mathbb{R}_+$, the *popularity* of each news m, $\kappa_{n,m}$, for $n \geq 1$, is updated according to:

$$\kappa_{n,m} = \kappa_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on and not highlighted by } n \\ 1+\eta & \text{if } m \text{ is clicked on and highlighted by } n. \end{cases} \tag{4}$$

The *ranking* of the news that individual n sees $(r_{n,m})_{m\in M}$ is inversely related to the popularity before she clicks:

$$r_{n,m} < r_{n,m'} \iff \kappa_{n-1,m} < \kappa_{n-1,m'}. \tag{5}$$

We also keep track of the traffic a news item receives without counting the highlights. Hence, starting from $\widehat{\kappa}_{0,m} = \kappa_{0,m} \in \mathbb{R}_+$, the *number of clicks* on news item m, $\widehat{\kappa}_{n,m}$, for $n \geq 1$, is updated according to:

$$\widehat{\kappa}_{n,m} = \widehat{\kappa}_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on by } n. \end{cases} \tag{6}$$

Similarly, we track keep of the number of highlights a news item receives, without counting the clicks. Thus again, defining $\widetilde{\kappa}_{0,m} = \kappa_{0,m} \in \mathbb{R}_+$, the *number of highlights* of news item m, $\widetilde{\kappa}_{n,m}$, for $n \geq 1$, are updated accordingly.

2.2.2 Popularity ranking with personalization

Suppose now the ranking, while still being based on popularity, weights differently the clicks from different groups. Because in this simple model individuals only enter once in the platform, the model does not allow the personalization to be based strictly speaking on previous clicks and highlights. Nevertheless, we assume the algorithm can somehow infer the sign of the individuals' signals (e.g., using the location of her IP address, cookies from past browsing history, etc.) so that individuals are naturally divided into two groups, $x_n \in L$ if their signal satisfies $\text{sgn}(x_n) = -1$, and $x_n \in R$ if $\text{sgn}(x_n) = 1$. Choices and highlights are determined as above, but the difference is that now there are two rankings, $r^L_{n,m}$ and $r^R_{n,m}$, whereby individuals in L see $r^L_{n,m}$ when doing their search, while individuals in R see $r^R_{n,m}$. Moreover, the ranking of group $g \in \{L, R\}$ depends only in part on the clicks and highlights of individuals from the opposite group. Specifically, starting from $\kappa^g_{0,m} \in \mathbb{R}_+$, the *popularity for group* g of each news item m, $\kappa^g_{n,m}$, for $n \geq 1$, is updated according to:

$$\kappa^g_{n,m} = \kappa^g_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ \lambda(n) & \text{if } m \text{ is clicked on and not highlighted by } n, \\ \lambda(n) \cdot (1+\eta) & \text{if } m \text{ is clicked on and highlighted by } n, \end{cases} \tag{7}$$

where $\lambda(n) = 1$ if $n \in g$ and $\lambda(n) = \lambda$ if $n \notin g$, and where λ, $0 \leq \lambda \leq 1$, is a parameter of the personalized ranking algorithm and determines how much clicks and highlights from the opposite group $g' \neq g$ count for the ranking seen by group g. When $\lambda = 0$ each group sees a fully personalized ranking, independent of the clicks and highlights of the other group. When $\lambda = 1$ clicks from both groups count the same, so that the two rankings are identical, and we get back the case of a single ranking as in the previous subsection.

As before, the *ranking* of news item m that individual $n \in g$ sees $(r^g_{n,m})_{m \in M}$ is inversely related to the popularity of m before n clicks:

$$r^g_{n,m} < r^g_{n,m'} \iff \kappa^g_{n-1,m} < \kappa^g_{n-1,m'}, \quad g \in \{L, R\}. \tag{8}$$

We also keep track of traffic and engagement of news items separately for each group. Starting again from $\widehat{\kappa}^g_{0,m} = \kappa_{0,m} \in \mathbb{R}_+$, the *number of clicks by group* g on website m, $\widehat{\kappa}^g_{n,m}$, for $n \geq 1$ and $g \in \{L, R\}$, is updated according to:

$$\widehat{\kappa}^g_{n,m} = \widehat{\kappa}^g_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on by } n, n \in g \\ 0 & \text{if } m \text{ is clicked on by } n, n \notin g, \end{cases} \tag{9}$$

The same can be done by counting the highlights of group g without counting the clicks, $\widetilde{\kappa}^g_{n,m}$, for $g \in \{L, R\}$.

2.3 Evaluation Indices

To evaluate the effect of the highlighting and personalization parameters of the ranking algorithm on various aspects of social welfare, we formally define a few indices. Let $y(n) \in M$ denote the signal of the news item clicked on by individual n, and let L (R) denote the individuals with signals x_n with $\text{sign}(x_n) = -1$ $(= +1)$. Then we can define the following indices:

- *Engagement* on item m by group g: $ENG^g_m = \widehat{\kappa}^g_{N,m} + \widetilde{\kappa}^g_{N,m}$ (clicking and highlighting by grop g);
- *Total Engagement*: $ENG = \sum_{m \in M} \left(ENG^L_m + ENG^R_m \right)$ (total clicking and highlighting);
- *Misinformation*: $MIS = \frac{1}{N} \sum_{n \in N} |y(n) - \theta|$;
- *Polarization*: $POL = \frac{1}{N} \left| \sum_{n \in R} y(n) - \sum_{n' \in L} y(n') \right|$;

The first two indices are meant to capture one the key dimension digital platforms care about: user engagement, or the expected amount of activity generated by the individuals.[12] The third index is a straightforward measure of misinformation capturing the average distance between the information carried by the news items chosen by individuals and the true state of the world. The fourth index measures polarization as the average distance between the information provided by the news items chosen by individuals in group R with the respect to the one provided by the news items chosen by individuals in group L.[13]

Finally, we aim to asses the impact of highlighting and personalization weights on the overall social welfare. In principle, such weights might have opposite effects on the welfare of the platform versus the one of the individuals. Accordingly, we consider two sources of social welfare, namely, one based on what the platform mainly cares about: generating high levels of engagement (ENG); and another based on what may concern the users of the platform: guaranteeing low levels of misinformation (MIS) and polarization (POL). For convenience, we capture all these aspects in a single measure of *welfare* of the form:

$$W_\psi(\eta, \lambda) = \psi \cdot ENG(\eta, \lambda) - (1 - \psi) \cdot MIS(\eta, \lambda) \cdot POL(\eta, \lambda), \tag{10}$$

where $0 \leq \psi \leq 1$ is a weight for the relative importance of the platform's likely goal (high ENG) relative to the users' welfare (low MIS and POL).

3 Engagement, Misinformation, Polarization and Social Welfare

In this section, we study the dynamic interplay between individuals' clicking and highlighting behavior and the platform's ranking algorithm, based on popularity and personalization. To simplify the discussion, throughout this section we assume that $\widehat{\theta} = \theta$. That is, we focus on the symmetric case where signals are symmetrically distributed around the benchmark.[14] We first present a preliminary discussion of the

[12] In particular, the willingness to increase engagement was behind the boost in η implemented in 2018 by Facebook with the stated objective of increasing *meaningful social interactions* (see Section 4 for a related discussion).

[13] Notice that we abstract from the specific belief updating of each individual. Our focus is on the comparative static effect of changes in the algorithm parameters (η and λ) on misinformation and polarization. Accordingly, the proposed misinformation and polarization indices will be informative on such effects as long as individuals update their beliefs in the direction of the signal carried by the news item they click on, $y(n)$.

[14] Allowing for heterogeneous benchmarks across individuals $\widehat{\theta}_n$ does not qualitatively affect the results; see Appendix.2.1. Similarly, allowing for small asymmetries in the distribution of signals with respect to the benchmark $\widehat{\theta}$ also does not affect the

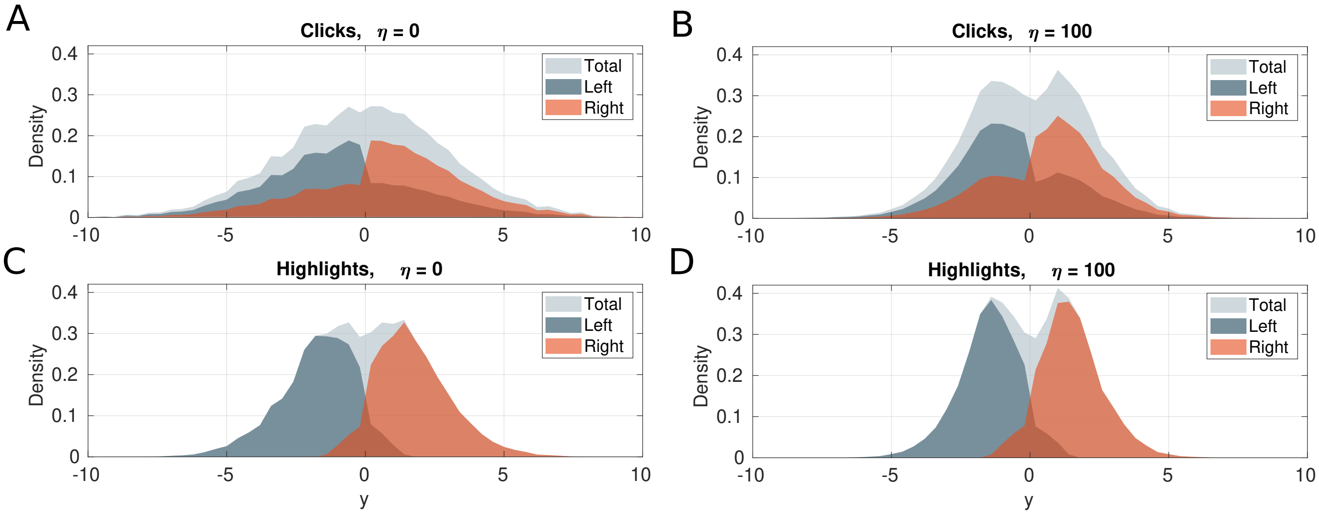

Figure 2: Users' clicking (top) and highlighting (bottom) behavior for $\eta = 0$ (left) and for $\eta = 100$ (right) in the *constant (flat)* highlighting case. For increasing values of η, both polarization and misinformation *decrease*. Polarization decreases from an average value of 1.8 (stev 0.5) to 1.3 (stdev 0.3), and the misinformation also decreases from an average value of 2.4 (stev 0.6) to 1.7 (stdev 0.3). See Section 3.5 for more details.

mechanism linking η and the dynamics of clicking and highlighting in the flat and non-flat cases. We then provide analytical results characterizing limit clicking and highlighting distributions and the impact of η and λ on the key indices introduced in Section 2.3.

3.1 Increasing Meaningful Social Interactions in the Flat and Non-Flat Cases

A key objective of our model is to understand the effect of the weight of a highlight (η) on engagement, misinformation, and polarization. The effect, especially for misinformation and polarization, crucially depends on the propensity with which individuals decide to highlight news items. In particular, it matters greatly whether individuals' highlighting behavior is what we call *flat* or *non-flat*. The reason is that, as the weight of a highlight increases, more items close to the truth (y's $\approx \theta$) are highlighted in the flat case and consequently also clicked on, whereas, in the non-flat case, it is items farther from the truth (y's $\approx -x^*, x^*$; see Fig. 1) that are highlighted and consequently also clicked on more frequently.

3.1.1 Flat case

Consider what we referred to as the *flat case* for the active types in the previous section, then increasing the weight on highlighting (η) can be shown to have some nice properties for both the platform and users welfare, namely, it increases engagement, while also reducing both polarization and misinformation. To

results qualitatively ($|\widehat{\theta} - \theta| < \min\left\{\frac{\sigma_x}{4}, \frac{\sigma_y}{4}\right\}$) . However, allowing for large asymmetries may change the results; Appendix.2.2 discusses the case where $\widehat{\theta}$ and θ are far apart.

see this, suppose then that a constant share ($p_A \in (0,1)$) of individuals who read a given article are also willing to highlight it, provided the news item's signal is sufficiently close to the individual's signal ($y_m \in H(x_n)$). Then as η increases, articles that get highlighted increase in total popularity and hence go up higher in the ranking, meaning that they are in turn also more likely to get clicked on. Since both the news items' and the individuals' signals are normally distributed, there is a relatively higher mass of individuals with signals around the truth (θ) and so, such individuals are more likely to read and highlight articles closer to the truth. This pushes them further up in the ranking. Hence, higher values of η will tend to concentrate clicking around the truth. This decreases polarization and misinformation, and, because there are relatively more individuals with signals around the truth, due to their normal distribution, it also increases engagement.

Thus increasing η in the flat case directly increases what Facebook calls *meaningful social interactions*, and at the same time concentrates clicking around the truth, thereby decreasing misinformation and polarization. This is illustrated by Figure 2, where Panel B shows the (simulated) clicking distribution for large η more concentrated around the truth than the corresponding Panel A with small η. Panels C and D show the highlighting distribution, which is a key part of engagement, and which is also more concentrated around the truth for large η (Panel D) than for small η (Panel C). Moreover, it can be checked that the overall number of highlights is greater for large η (Panel D) than for small η (Panel C).

3.1.2 Non-flat case: Crowding out the truth

Consider now our leading case, which is the *non-flat case*. Then increasing the weight on highlighting (η) can have desirable properties for the platform, namely, higher engagement, but not necessarily for users since it can result in higher misinformation and higher polarization. To see this, notice first that the combination of normally distributed priors and a propensity to highlight that increases in the extremeness of the prior leads to highlighting behavior that is bimodal (see Section 2.1.3, Figure 1). An important intermediary result shows that, as η becomes large and highlighting becomes relatively more important, the clicking distribution inherits the basic shape of the highlighting distribution. The reason is that, due to the ranking algorithm, when η is large, items that have a high propensity of being highlighted go higher up in the ranking, and highlighting behavior becomes more important as a driver of what individuals see as being ranked prominently and ultimately end up clicking on.

As a result, as η increases, the clicking distribution goes from being roughly unimodal and centered around the true state θ when η is small to being increasingly bimodal as η gets large. Since the clicking distribution reflects what people read, this shows that a higher η increases both misinformation and polarization. It can be checked that it also leads to higher engagement (measured as the sum of clicks and

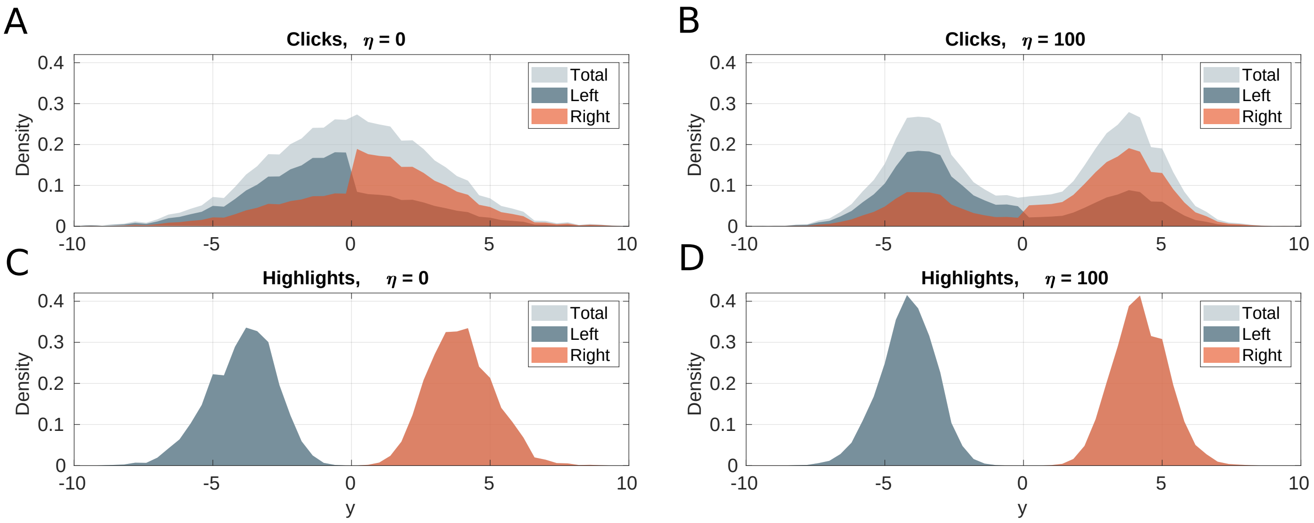

Figure 3: Polarization and misinformation *increase* for increasing values of η in the *non-flat* highlighting case. Panels (A) and (B) show users' clicking behavior for $\eta = 0$ and $\eta = 100$, respectively. Panels (C) and (D) show users' highlighting behavior for $\eta = 0$ and $\eta = 100$, respectively. Polarization increases from an average value of 1.8 (stev 0.5) to 2.8 (stdev 0.4), and misinformation increases from an average value of 2.4 (stev 0.6) to 3.6 (stdev 0.4). See Section 3.5 for more details.

highlights). This is illustrated in Figure 3, where Panels A and B show the clicking distribution for the cases of respectively small and large η. We refer to this phenomenon, whereby individuals are less likely to click on items close to the true state and more likely to click on ones further away, due to a higher parameter η, as *crowding out the truth.* As mentioned in the introduction, the case of large η seems to capture what Bail (2021) refers to as the social media "prism", which he argues besides fueling extremism and polarization, mutes moderates and gives a distorted image of others. To see this, consider Panel B, where we see that individuals on the left, for example, are more likely to click on extreme items from the left when clicking on the left (blue), but are also more likely to click on extreme items from the right when clicking on the right (shaded red on the right). This suggests that they will get a more extreme impression of individuals both on the left and on the right and, given that priors are centered around θ, this is also consistent with higher *perceived* polarization on social media (see also Yang et al. (2016)). In other words, the highlighting parameter η can be seen as directly contributing to the "prism" effect of Bail (2021). Panels C and D, finally, show the highlighting behavior in the case of small and large η respectively, and where it can be seen that there is more highlighting in D than in C, suggesting that engagement increases with η.

In the subsections that follow, we look at the above phenomena more formally and also connect them to welfare evaluations for the platform and the consumers.

3.2 Limit Clicking and Highlighting Behavior

In order to evaluate the impact of the parameters of the algorithm (η, λ), it is useful to obtain actual clicking and highlighting behavior of the individuals, while keeping track of the dynamic feedback between clicking, highlighting and ranking. We do this by characterizing the limit clicking and highlighting distributions, since these ultimately determine what to expect in terms of engagement and changes in posterior beliefs.

A central feature of the ranking algorithm is its dependence on the popularity of different items. Define the *expected popularity of an item with signal y, absent ranking*, as the sum of the expected clicking and highlighting propensities, absent ranking, where highlighting is weighted by η:

$$\pi(y) = \frac{1 + \eta \cdot \mu_H(y)}{M \cdot (1 + \eta \cdot \bar{\mu}_H) + \eta \cdot (\mu_H(y) - \bar{\mu}_H)}, \tag{11}$$

where $\mu_H(y) = \int_{x \in H^{-1}(y)} p_A(x) f(x; \sigma_x^2) dx$ and $\bar{\mu}_H = \int \mu_H(y) f(y; \sigma_y^2) dy$.[15]

An assumption that we maintain in this and the following section is that the expected rank of a given item with signal $y \in \mathbb{R}$ is approximated by a linear decreasing function of the expected popularity of that item, absent ranking:

$$r(y) \approx \zeta_0 - \zeta_1 \cdot \pi(y), \tag{12}$$

where $\zeta_0, \zeta_1 > 0$ are constants.

This assumption significantly simplifies the analysis. It allows us to directly derive the effects of the popularity and personalization parameters and to characterize both the limit clicking and highlighting distributions, which represent expectations of T repetitions (for $T \to \infty$) of the process described in Sections 2.1 and 2.2.[16] Importantly, Figure 4 shows that the assumption is quite accurate for simulations with different values of η in the flat (Panels A-C) and non-flat cases (Panels D-F). Moreover, all the simulations in Section 3.5 are obtained without imposing the assumptions and the results are fully consistent with the analytical results obtained assuming Eq. (12). This suggests that the linearity assumption on the expected rank is relatively innocuous within our overall framework.

Lemma 1 (Limit Clicking and Highlighting Distributions)**.** *Assume Eq. (12), then the limit clicking*

[15] Eq. (11) is derived from:

$$\pi(y_m) = \frac{\int \sum_{k \in \{C,E,I\}} p_k \cdot \varphi_{n,m} \cdot \left(1 + \eta \cdot p_A(x_n) \cdot 1_{\{x_n \in H(y_m)\}}\right) f(x; \sigma_x^2) dx}{\sum_{m' \in M} \int \sum_{k \in \{C,E,I\}} p_k \cdot \varphi_{n,m'} \cdot \left(1 + \eta \cdot p_A(x_n) \cdot 1_{\{x_n \in H(y_{m'})\}}\right) f(x; \sigma_x^2) dx},$$

taking averages over T repetitions for $T \to \infty$.

[16] The analysis would otherwise require computing distribution over limit rankings, which is an open problem already without highlighting and with $M > 2$ items (Analytis et al., 2022). However, it can be verified that in the case of $M = 2$, using a result from Analytis et al. (2022), there is an exact linear relationship between the expected rank and the clicking propensity as assumed in Eq. (12).

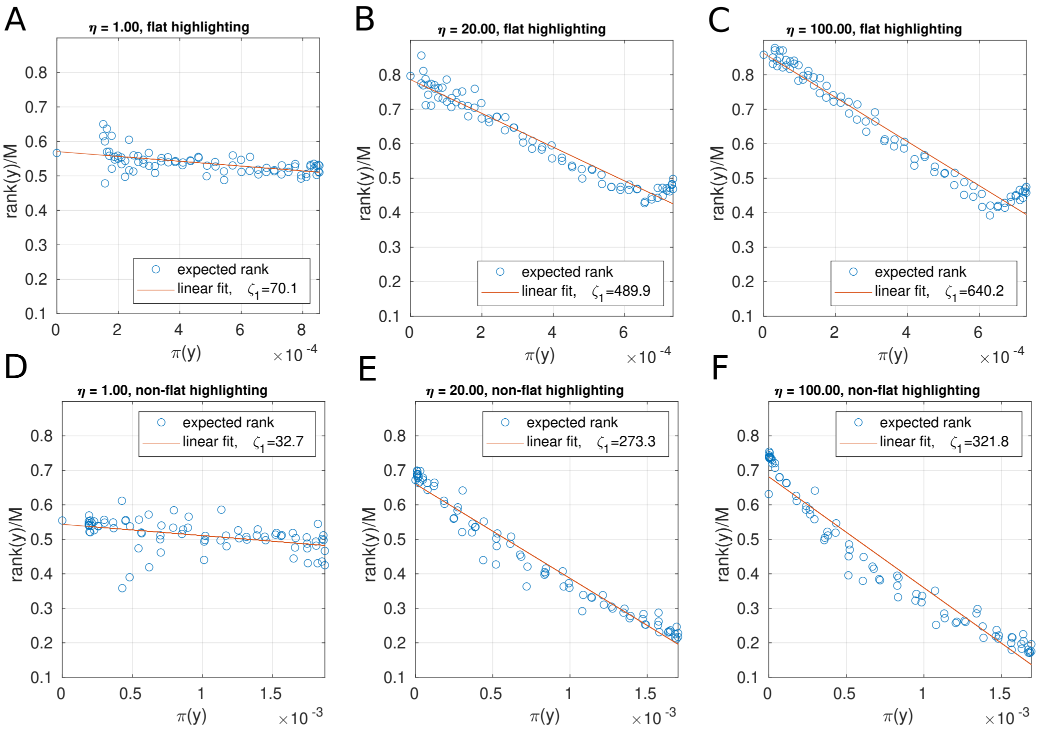

Figure 4: Linear dependence between the expected rank (blue circles, obtained from simulations) and the expected popularity $\pi(y)$ as assumed in Eq. (12). Panels (A)-(C) depict the flat case highlighting case, and Panels (D)-(F) the non-flat case. Red line denotes the best linear fit. To compute each blue dot, we binned the item's signals into 81 bins (bin-size = 0.2) and compute the mean popularity and the corresponding mean rank from $T = 10^3$ experiments, each of them with different $M = 20$ item's signals and different $N = 5 \cdot 10^3$ individual's signals. See Section 3.5 in the main text for more details.

distribution can be approximated as:

$$LCD(y) \approx \Lambda_\beta(\pi(y)) \cdot f(y; \sigma_y^2), \tag{13}$$

where, Λ_β is a linear function with $\Lambda'_\beta > 0$ for $\beta > 1$, π is defined in Eq. (11). Accordingly, the limit highlighting distribution can be approximated as $LHD(y) \approx \mu_H(y) \cdot LCD(y)$.

The first result shows a basic feature of our ranking-based dynamics, namely, that the expected traffic on a given item is driven by its expected popularity, absent ranking. Attention bias enters through Λ_β as it's a strictly increasing function of $\pi(y)$ provided $\beta > 1$. A higher $\pi(y)$ increases the rank of an item with signal y, thereby increasing its traffic, the more so, the greater β is, for $\beta > 1$. With Eq.(11) we can write:

$$LCD(y) \propto (1 + \eta \cdot \mu_H(y)) \cdot f(y; \sigma_y^2), \tag{14}$$

meaning that, for any y, $LCD(y)$ is proportional to the expression on the right and directly shows how η affects limit clicking behavior. Both results follow from the linearity assumption in Eq. (12) combined with the functional form of the clicking probabilities assumed in Eq. (2).

3.3 Engagement, Misinformation and Polarization

3.3.1 Effect of the highlighting parameter η

We here focus on the comparative statics of the popularity parameter for highlighting (η) and assume there is no personalization ($\lambda = 1$).[17] As is clear from Section 3.2, as η increases, the expected popularity of an item and hence its expected traffic is increasingly driven by the highlighting propensity. This observation is a central message of the paper and has important consequences for how the parameter η affects engagement, misinformation and polarization.

Proposition 1 (Effect of η on Engagement, Misinformation and Polarization)**.** *Assume Eq. (12). If individuals' highlighting behavior is non-flat (p_A as in Eq. (3)), then, increasing the weight on highlighting (higher η) increases user engagement, misinformation and polarization. If, instead, the highlighting behavior is flat (p_A constant), then the above results need not hold; a higher η increases user engagement and decreases misinformation and polarization.*

Thus, in the non-flat highlighting case, increasing η, increases engagement, but also has the adverse effect of increasing misinformation and polarization. This is not the case when highlighting behavior is flat, where a higher η increases engagement, while decreasing misinformation and polarization, albeit slightly.

It is possible to interpret the results in light of the evidence provided by Bakshy et al. (2015). As discussed above, Bakshy et al. (2015) point out that in the case of "hard" news (e.g, national, political), the propensity to highlight content is indeed higher for individuals with a more extreme prior, whereas the same does not apply to "soft" news (e.g., entertainment). The results suggest that social media platforms have an incentive to choose a high level of η as this results in a high level of engagement across all types of news contents. Yet, while this is not so much a concern for users' welfare in the case of "soft" news, we show it might have detrimental effects on misinformation and polarization when it comes to political news contents.

3.3.2 Effect of the personalization parameter λ

Suppose that the ranking can be personalized, based on the sign of the individuals' signals as in Section 2.2.2. It is not difficult to see that the above results on the effect of η continue to hold for any degree of personalization ($\lambda \in [0, 1]$). The next result concerns the effect of the personalization parameter λ.

[17] The willingness to increase engagement was behind the boost in η implemented in 2018 by Facebook with the stated objective of increasing *meaningful social interactions*; see Section 4 for a related discussion.

Proposition 2 (Effect of λ on Engagement and Polarization)**.** *Assume Eq. (12) and fix $\eta \geq 0$ arbitrarily. Then increasing personalization (lower λ) increases user engagement and polarization, both when individuals' highlighting behavior is flat and when it is non-flat (p_A as in Eq. (3)).*

The fact that more personalization increases polarization is straightforward. Decreasing λ makes the rankings of the two groups increasingly less correlated, which in turn makes users in each group more likely to click on items carrying a signal of the same sign as their own. This directly increases the polarization measure POL. To see the effect on engagement, note first that users that are active types only share items that are close enough to their own signal ($y_m \in H(x_n)$). As λ decreases and the rankings become less correlated, users are more likely to see items that have signals closer to their own more prominently ranked, and are in turn also more likely to click on them. But since items that are more prominently ranked are more likely to be in the set $H(x_n)$, they are also more likely to be highlighted. Overall, whether highlighting behavior is flat or non-flat, a lower λ (more personalization) contributes to an increase in ENG.

One effect that the personalization parameter λ does not have in our model, differently from the highlighting parameter η, is that it does not significantly impact misinformation. This is due to the fact that it mainly contributes towards interchanging clicks made from one group on items with signals of the opposite sign with clicks made by individuals from the other group, who have signals of the same sign. While this contributes to increasing polarization it does not really affect misinformation.

3.4 Towards a socially efficient ranking

Consider the welfare index (W_ψ) defined in Section 2.3, Eq (10). From the analysis of the previous sections, we can show:

Proposition 3. *Assume Eq. (12). If individuals' highlighting behavior is non-flat (p_A as in Eq. (3)), then, for small values of ψ, ($\psi \approx 0$), social welfare (W_ψ) is maximized at $(\eta, \lambda) \approx (0, 1)$, while for large values of ψ, ($\psi \approx 1$), social welfare is maximized at $(\eta, \lambda) \approx (\infty, 0)$.*

If instead individuals' highlighting behavior is flat (p_A constant), then, for small values of ψ, ($\psi \approx 0$), social welfare is maximized at $(\eta, \lambda) \approx (\infty, 1)$, while for large values of ψ, ($\psi \approx 1$), social welfare is maximized at $(\eta, \lambda) \approx (\infty, 0)$.

This presents a key result of the paper. Namely, in the empirically relevant case of non-flat propensity to highlight, a clear dichotomy arises between the perspective of the platform and the one of the users with respect to the desirable weight to be assigned to the highlights. This resonates with the reports leaked by Facebook's whistle-blowers, who underscored the conflicting welfare effects created by the platform's 2018 "Meaningful social interactions" update, which boosted the weights given to content sharing and highlighting in the ranking algorithms. Indeed, while this change increased users' overall engagement on Facebook, it appears to also have led to an increase in misinformation and polarization and hence decreased user welfare, as predicted by Proposition 3 for the non-flat case.[18] Section 4 presents direct empirical evidence in this regard.

3.5 Numerical Simulations

We here provide simulation results for the more general case, where no restriction on the linearity of the expected ranking is imposed. We run $T = 4,000$ independent simulations of the basic model with both non-flat and flat highlighting propensities. Each run corresponds to $M = 20$ different news items signals $y_m \sim N(\theta = 0, \sigma_y^2 = 9)$ and to $N = 10^5$ individuals signals $x_n \sim N(\theta = 0, \sigma_x^2 = 9)$. When reporting the key evaluation indices, we only consider the last 2,000 clicks. This avoids dependence on the initialization of the ranking. We also set $\widehat{\theta} = \theta$. The proportions of confirmatory, exploratory, and indifferent clicking types are set to $p_C = 0.7$, $p_E = 0.15$, and $p_I = 0.15$, respectively, and their corresponding propensities to $\gamma_C = 0.8$, $\gamma_E = 0.4$, and $\gamma_I = 0.5$, respectively. The value of β determining the attention bias is set to $\beta = 1.25$ and the value of α for the non-flat highlighting probability p_A is 4. Results are robust to choosing different values of the parameters.

Figure 5 summarizes the simulated effects of η and λ on the key evaluation indices for the non-flat case (Panels A-C) and the flat case (Panels D-F), where the reported measure is an average among all T independent simulations. Panels A and D show total user engagement (ENG). We observe that increasing η and increasing personalization (decreasing λ) results in an increase of engagement in both non-flat and flat cases. The dependence on λ is more pronounced in the flat case. Panels B and E show results of misinformation (MIS). Again, increasing η has opposite effects in non-flat and flat scenarios, resulting in an increase of misinformation in the non-flat case, and a decrease the flat case. In contrast to the polarization results, we only observe a weak dependence of misinformation on the degree of personalization λ, which is only noticeable for higher values of η. Finally, Panels C and F show results of user polarization (POL). In agreement with the analytical results, we observe that the effect of increasing η is different depending on the highlighting propensities. In the non-flat case, increasing η results in an increase of polarization, whereas in the flat case, it has the opposite effect. However, increasing personalization (decreasing λ) results in an increase in polarization, for both types of highlighting propensities.

Finally, we show results of the welfare index $W_\psi(\eta, \lambda)$. Figure 6 shows $W_\psi(\eta, \lambda)$ for different values of

[18] An aspect that may play a role in moderating the platform-optimal level of η is how it impacts advertising through its effect on content, besides the effect on engagement. For example, if advertisers were to dislike a too high level of misinformation/polarization, this could be reflected in W_ψ and, consequently, could affect the platform-optimal level of η. Nevertheless, for $\psi > 0$, i.e., as long as the platform cares about engagement, the level of η chosen by the platform will be higher than the one minimizing misinformation and/or polarization. For theoretical models on the role of advertising in affecting content moderation in social media, see Madio and Quinn (2021); Liu et al. (2021); Jiménez Durán (2022).

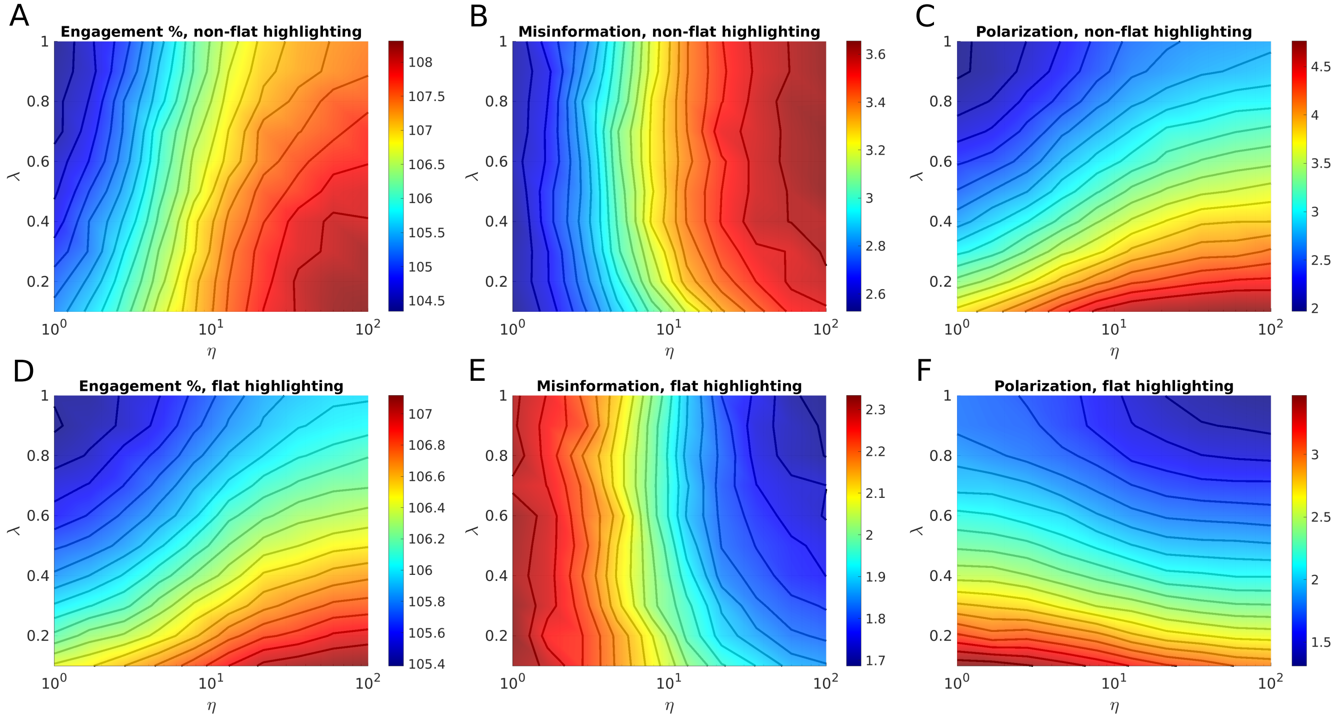

Figure 5: Effect of varying highlighting weight (η) and personalization (λ) in the simulations. For non-flat individuals' highlighting behavior: (A) user engagement, (B) misinformation, and (C) polarization. For flat individuals' highlighting behavior: (D) user engagement, (E) misinformation, and (F) polarization.

the weight ψ which controls for the relative importance of the platform's welfare (high ENG) relative to the users' welfare (low MIS and POL). We observe that the values of η and λ for which the welfare index is maximized coincide with those stated in Proposition 3 for both flat and non-flat cases.

4 Empirical Evidence: Meaningful Social Interactions and Political Polarization

The theoretical predictions of our model discussed in Section 3.3 suggest that an increase in the weight given by the ranking algorithm to the "highlights" (an increase in η) will result in individuals being more exposed to extremist political content and, in turn, in a higher level of political polarization. To connect this prediction to observational data, we exploit Facebook's "Meaningful Social Interaction" (MSI) update implemented in January 2018, which—with the goal of increasing platform engagement—boosted the weight given to likes and shares in Facebook's ranking algorithm.[19] In particular, our theoretical framework suggests that—if the propensity to highlight contents is higher for people with more extreme priors (non-flat case)—we should observe an increase in extremism and political polarization following such

[19] See `https://www.facebook.com/business/news/news-feed-fyi-bringing-people-closer-together` and `www.edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math`.

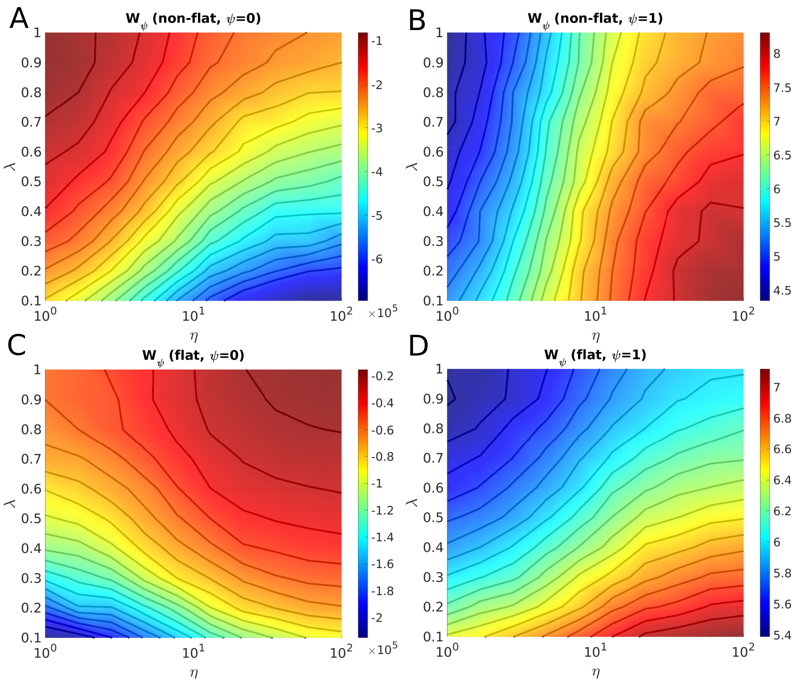

Figure 6: Effect of highlighting weight (η) and personalization (λ) on social welfare in the simulations. For non-flat individual highlighting behavior: (A) user welfare ($\psi = 0$), (B) platform welfare ($\psi = 1$). Similarly for flat individuals' highlighting behavior: (C) user welfare, (D) platform welfare.

a change in the algorithm. In what follows, we provide empirical evidence in support of such predictions by leveraging a survey dataset from Italy containing rich information on political preferences around the time of the change in Facebook's algorithm.

4.1 Data

The dataset comes from the *Polimetro* (i.e., Political meter) surveys run by the leading Italian public opinion polling company *Ipsos*. The *Polimetro* contains weekly/monthly interviews on a representative sample of the Italian voting population (i.e., aged 18 or above).

In particular, for the purpose of our analysis, the survey asks questions on the main sources of information used by an individual to form a political opinion (i.e., newspapers, radio news, tv news, friends, internet, etc). It is important to notice that around the time of its MSI update, Facebook was by far the first social network in Italy with 34 million active users per month and a 60% penetration rate in the overall population (corresponding to a penetration rate of almost 80% with respect to the population of Italian internet users), compared with the 33% and 23% penetration rates of Instagram and Twitter, respectively (We Are Social, 2018). Hence, while the Ipsos survey does not directly ask questions about Facebook use, it is possible to proxy the exposure to Facebook content with the use of internet to form a political opinion.

Furthermore, besides providing information on the socio-demographic characteristics of the respondents, the *Polimetro* asks questions regarding the ideological position of the respondent on the left-right scale and on the probability of voting for each party. Accordingly, we make use of these questions to construct two main outcome variables. The first one is a dummy variable taking value zero if a respondent self-identifies with a moderate political position (center, center-left or center-right) and one if she instead identifies with a more extremist position (left, right, extreme-left, extreme-right). This variable is thus meant to capture a simple measure of political extremism. The second one is a measure of affective polarization capturing "the extent to which citizens feel sympathy towards partisan in-groups and antagonism towards partisan out-groups" (Wagner 2021, page 1), see Appendix.3.1 for further details.

4.2 Empirical strategy

We implement a Differences-in-Differences empirical model to assess whether the change in the Facebook algorithm implemented in January 2018 via the introduction of the "Meaningful Social Interaction" weights had a causal impact on self-declared ideological extremism and affective polarization. Specifically, we look at such outcomes in the group of people living in a given municipality who use internet to form a political opinion and who were interviewed after the Meaningful Social Interaction (MSI) algorithm was introduced (i.e., January-June 2018) and then compare it with the ones of the group of people also using internet to form an opinion who were interviewed before such a change (i.e., June-December 2017) and at the same time with the group of people interviewed after such change in the algorithm that were not using internet as one of the main sources to form an opinion. Accordingly, we estimate the following econometric specification:

$$\begin{aligned}\mathtt{Y_{i,m,t}} &= \alpha + \beta_1(\mathtt{Opinion\ via\ internet}_{i,m,t} \times \mathtt{Post\ MSI}) \\ &+ \beta_2\ \mathtt{Opinion\ via\ internet}_{i,m,t} + \beta_3\ \mathtt{Post\ MSI} + \alpha_m + \mathtt{X_{i,t}} + \varepsilon_{i,m,t} \qquad (15)\end{aligned}$$

where $Y_{i,m,t}$ represents the outcome of interest relative to individual i, leaving in municipality m interviewed in the survey wave t (i.e., probability of declaring a non-moderate political ideology or affective polarization). α_m captures municipality fixed effects. β_1 is the parameter of interest. In more demanding specifications, we also include either time fixed effects (i.e., survey-wave fixed effects) or province-by-time fixed (which account for any unobservable shock at the province-time level). $X_{i,t}$ represents a vector of socio-demographic control variables including the respondent's age (and age squared), gender, number of resident family members, level of education, type of occupation and religiosity. Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population.

4.3 Results

Table 1 shows our baseline results on the effect of the introduction of Facebook's MSI update on the probability that an individual using internet to form an opinion holds a non-moderate political position. Column 1 provides estimates when including municipality fixed effects only (besides individual level controls). Column 2 includes also date-of-interview fixed effects accounting for possible overall time-varying patterns in ideological positions. Column 3 includes fixed effects both at the municipality and at the province-by-date-of-interview level, thus accounting for any province-time variation in political preferences. Columns 1-3 present estimates when clustering standard error at the regional levels (which in Italy correspond to electoral districts for the upper chamber). Column 4 provides evidence that results are robust when clustering standard error at a finer geographical level (provinces). The results suggest that in the period after the MSI implementation, individuals using internet to form a political opinion had a higher probability of holding a non-moderate ideology. The effect is sizeable accounting—in the most demanding specifications, columns 3 and 4—for around one standard deviation increase in such probability.

Table 1: MSI and non-moderate ideological position

	(1) Non-moderate Ideology	(2) Non-moderate Ideology	(3) Non-moderate Ideology	(4) Non-moderate Ideology
Opinion via internet × Post MSI	0.062***	0.058***	0.051***	0.051***
	(0.016)	(0.015)	(0.014)	(0.018)
Opinion via internet	-0.012	-0.006	-0.012	-0.012
	(0.020)	(0.020)	(0.024)	(0.022)
Post MSI	-0.017*			
	(0.009)			
Observations	25,690	25,690	25,690	25,690
Mean outcome	0.36	0.36	0.36	0.36
SD outcome	0.48	0.48	0.48	0.48
Municipality FE	YES	YES	YES	YES
Date of interview FE	NO	YES	NO	NO
Province-Date of interview FE	NO	NO	YES	YES
Cluster SE	Region	Region	Region	Province

Note: Time horizon: June 2017-June 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation and religiosity of the respondent. Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

We now turn to the analysis on affective polarization. Table 2 presents our results.[20] The results present in Columns 1-4 show a positive, statistically significant and robust effect. That is, in the period after the MSI algorithmic update, individuals using internet to form a political opinion had a higher level of affective polarization. Also in this case, the effect is sizeable accounting for around 1.2 of a standard

[20] The lower number of observations relative to Table 1 is due to the fact that the questions used as proxies of sympathy score for the different parties are asked less frequently (i.e., in fewer surveys) with respect to the one on the self-decleared ideological position.

deviation increase in affective polarization. All in all, Tables 1 and 2 provide evidence in support of the key theoretical predictions of our model.[21]

Table 2: MSI and Affective Polarization

	(1) Affective Polarization	(2) Affective Polarization	(3) Affective Polarization	(4) Affective Polarization
Opinion via internet × Post MSI	0.054**	0.055**	0.073***	0.073***
	(0.024)	(0.024)	(0.019)	(0.025)
Opinion via internet	-0.012	-0.011	-0.006	-0.006
	(0.023)	(0.022)	(0.023)	(0.025)
Post MSI	0.118***			
	(0.020)			
Observations	14,499	14,499	14,499	14,499
Mean outcome	1.29	1.29	1.29	1.29
SD outcome	0.61	0.61	0.61	0.61
Municipality FE	YES	YES	YES	YES
Date of interview FE	NO	YES	NO	NO
Province-Date of interview FE	NO	NO	YES	YES
Cluster SE	Region	Region	Region	Province

Note: Time horizon: June 2017-June 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation and religiosity of the respondent. Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

5 Conclusion

Ranking algorithms are central to the working of social media platforms. This paper provides a theoretical framework to study the dynamic feedback between social media users and a platform's ranking algorithm. The model shows how polarization and misinformation may naturally emerge from simple changes in the algorithm's parameters, combined with basic well-documented behavioral traits of the users. When optimizing its algorithm, the platform may face a trade-off between its own welfare and that of its users. We show that changes in popularity and personalization parameters that increase platform engagement may have detrimental effects in terms of misinformation—*crowding-out the truth*—and/or polarization, but may also distort the way individuals perceive each other on the platform (Bail, 2021; Yang et al., 2016). Our results are also consistent with the evidence provided by the empirical literature assessing the impact of personalization on political polarization (Levy, 2021; Huszár et al., 2022). Importantly, by exploiting the 2018 Facebook MSI algorithmic ranking update and leveraging a rich survey dataset from Italy, we provide direct empirical evidence corroborating the detrimental impact on political polarization created by a boost in the weight given by the ranking algorithm to "highlighted" contents (e.g., shares), as predicted

[21] Appendix Tables A.1 and A.2 show that the results are robust to excluding observations in the pre-electoral period (January-March 2018).

by our model. The paper thus provides academic guidance to the public debate on the potential desirable and undesirable consequences of social media and other "algorithmic gatekeepers" on social welfare.

We conclude by acknowledging that the model does not embed other important features of social media such as endogenous networks or fact-checking. Complementary research (Acemoglu et al., 2022) points out that these additional features may lead to further reinforcing the trade-off between platform engagement and social welfare outlined here. All in all, the insights from this line of research provide a "theory of harm," indirectly endorsing the recent attempt by the European Union to regulate digital platforms.[22] Future research combining endogenous dynamic algorithmic ranking and endogenous belief and network formation may provide additional insights to guide public regulators and social media platforms in their efforts to reduce the negative impact of ranking dynamics on social media users and on democratic society at large.

[22] See `https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package`.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius, "A Model of Online Misinformation," CEPR Discussion Papers 16932, C.E.P.R. Discussion Papers January 2022.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The welfare effects of social media," *American Economic Review*, 2020, *110* (3), 629–76.

Amnesty International, "The Social Atrocity. Meta and the Right to Remedy for the Rohingya," `www.amnesty.org/en/documents/ASA16/5933/2022/en/` 2022. [Online; accessed 01-October-2022].

An, Jisun, Daniele Quercia, and Jon Crowcroft, "Partisan Sharing: Facebook Evidence and Societal Consequences," in "Proceedings of the Second ACM Conference on Online Social Networks" COSN '14 Association for Computing Machinery New York, NY, USA 2014, p. 13–24.

Analytis, Pantelis P., Francesco Cerigioni, Alexandros Gelastopoulos, and Hrvoje Stojic, "Sequential choice and selfreinforcing rankings," Economics Working Papers 1819, Department of Economics and Business, Universitat Pompeu Fabra February 2022.

Ananthakrishnan, Uttara M and Catherine E Tucker, "The drivers and virality of hate speech online," *Available at SSRN 3793801*, 2022.

Athey, Susan, Emilio Calvano, and Joshua S Gans, "The impact of consumer multi-homing on advertising markets and media competition," *Management science*, 2018, *64* (4), 1574–1590.

Azzimonti, Marina and Marcos Fernandes, "Social media networks, fake news, and polarization," *European Journal of Political Economy*, 2022, p. 102256.

Bail, Chris, "Breaking the social media prism," in "Breaking the Social Media Prism," Princeton University Press, 2021.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, 2015, *348* (6239), 1130–1132.

Bartels, Larry M, "Failure to converge: Presidential candidates, core partisans, and the missing middle in American electoral politics," *The ANNALS of the American Academy of Political and Social Science*, 2016, *667* (1), 143–165.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski, "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," *Working Paper*, 2022.

Bernhardt, Dan, Stefan Krasa, and Mattias Polborn, "Political polarization and the electoral effects of media bias," *Journal of Public Economics*, 2008, *92* (5-6), 1092–1104.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova, "Social media and xenophobia: evidence from Russia," Technical Report, National Bureau of Economic Research 2019.

CNN, "Likes, anger emojis and RSVPs: the math behind Facebook's News Feed — and how it backfired," `www.edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math` 2021. [Online; accessed 01-July-2022].

Dujeancourt, Erwan and Marcel Garz, "The Effects of Algorithmic Content Selection on User Engagement with News on Twitter," Technical Report, Jönköping International Business School March 2022.

Epstein, Robert and Ronald E Robertson, "The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections," *Proceedings of the National Academy of Sciences*, 2015, *112* (33), E4512—-E4521.

Flaxman, Seth, Sharad Goel, and Justin M Rao, "Filter bubbles, echo chambers, and online news consumption," *Public opinion quarterly*, 2016, *80* (S1), 298–320.

Galasso, Vincenzo and Tommaso Nannicini, "Competing on good politicians," *American political science review*, 2011, *105* (1), 79–99.

Garz, Marcel, Jil Sörensen, and Daniel F Stone, "Partisan selective engagement: Evidence from Facebook," *Journal of Economic Behavior & Organization*, 2020, *177*, 91–108.

Gentzkow, Matthew and Jesse M Shapiro, "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 2010, *78* (1), 35–71.

_ , Jesse M. Shapiro, and Daniel F. Stone, "Chapter 14 - Media Bias in the Marketplace: Theory," in Simon P. Anderson, Joel Waldfogel, and David Strömberg, eds., *Handbook of Media Economics*, Vol. 1 of *Handbook of Media Economics*, North-Holland, 2015, pp. 623–645.

Germano, Fabrizio and Francesco Sobbrio, "Opinion dynamics via search engines (and other algorithmic gatekeepers)," *Journal of Public Economics*, 2020, *187*, 104188.

_ , Vicenç Gómez, and Gaël Le Mens, "The few-get-richer: a surprising consequence of popularity-based rankings?," in "The World Wide Web Conference" 2019, pp. 2764–2770.

Glick, Mark, Greg Richards, Margarita Sapozhnikov, and Paul Seabright, "How does ranking affect user choice in online search?," *Review of Industrial Organization*, 2014, *45* (2), 99–119.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer, "Fake news on Twitter during the 2016 US presidential election," *Science*, 2019, *363* (6425), 374–378.

Hopp, Toby, Patrick Ferrucci, and Chris J Vargo, "Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data–based analysis of countermedia content dissemination on Facebook and Twitter," *Human Communication Research*, 2020, *46* (4), 357–384.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt, "Algorithmic amplification of politics on Twitter," *Proceedings of the National Academy of Sciences*, 2022, *119* (1).

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood, "The origins and consequences of affective polarization in the United States," *Annual Review of Political Science*, 2019, *22* (1), 129–146.

Jiménez Durán, Rafael, "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter," *Available at SSRN*, 2022.

Krasa, Stefan and Mattias K Polborn, "Is mandatory voting better than voluntary voting?," *Games and Economic Behavior*, 2009, *66* (1), 275–291.

Krishna, Vijay and John Morgan, "Overcoming ideological bias in elections," *Journal of Political Economy*, 2011, *119* (2), 183–211.

Lauer, David, "Facebook's ethical failures are not accidental; they are part of the business model," *AI and Ethics*, 2021, *1* (4), 395–403.

Levy, Ro'ee, "Social media, news consumption, and polarization: Evidence from a field experiment," *American economic review*, 2021, *111* (3), 831–70.

Liao, Hao, Manuel Sebastian Mariani, Matús Medo, Yi-Cheng Zhang, and Ming-Yang Zhou, "Ranking in evolving complex networks," *Physics Reports*, 2017, *689*, 1–54.

Liu, Yi, Pinar Yildirim, and Z. John Zhang, "Social Media, Content Moderation, and Technology," 2021.

Madio, Leonardo and Martin Quinn, "Content moderation and advertising in social media platforms," *Available at SSRN 3551103*, 2021.

Mosleh, Mohsen, Gordon Pennycook, and David G Rand, "Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter," *Plos one*, 2020, *15* (2), e0228882.

Mullainathan, Sendhil and Andrei Shleifer, "The market for news," *American economic review*, 2005, *95* (4), 1031–1053.

Müller, Karsten and Carlo Schwarz, "From hashtag to hate crime: Twitter and anti-minority sentiment," *Available at SSRN 3149103*, 2020.

_ **and** _ , "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, 2021, *19* (4), 2131–2167.

Novarese, Marco and Chris Wilson, "Being in the Right Place: A Natural Field Experiment on List Position and Consumer Choice," *Working Paper*, 2013.

Ortoleva, Pietro and Erik Snowberg, "Overconfidence in political behavior," *American Economic Review*, 2015, *105* (2), 504–35.

Pan, Bing, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka, "In Google We Trust: Users' Decisions on Rank, Position, and Relevance," *Journal of Computer-Mediated Communication*, 2007, *12*, 801–823.

Pew, "National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweet," Technical Report, Pew Research Center 2019.

Pogorelskiy, Kirill and Matthew Shum, "News we like to share: How news sharing on social networks influences voting outcomes," *Available at SSRN 2972231*, 2019.

Rudiger, Jesper, "Cross-Checking the Media," MPRA Paper 51786, University Library of Munich, Germany November 2013.

Sobbrio, Francesco, "Citizen-editors' endogenous information acquisition and news accuracy," *Journal of Public Economics*, 2014, *113*, 43–53.

Tella, Rafael Di, Ramiro Gálvez, and Ernesto Schargrodsky, "Does Social Media Cause Polarization? Evidence from Access to Twitter Echo Chambers During the 2019 Argentine Presidential Debate," *NBER Working Paper*, 2021, (w29458).

Vosoughi, Soroush, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science*, 2018, *359* (6380), 1146–1151.

We Are Social, "Digital in 2018 Report," `https://wearesocial.com/it/blog/2018/01/global-digital- report-2018/` 2018. [Online; accessed 01-October-2022].

White, Ryen W and Eric Horvitz, "Belief Dynamics and Biases in Web Search," *ACM Transactions on Information Systems (TOIS)*, 2015, *33* (4), 18.

Yang, JungHwan, Hernando Rojas, Magdalena Wojcieszak, Toril Aalberg, Sharon Coen, James Curran, Kaori Hayashi, Shanto Iyengar, Paul K Jones, Gianpietro Mazzoleni et al., "Why are "others" so polarized? Perceived political polarization and media use in 10 countries," *Journal of Computer-Mediated Communication*, 2016, *21* (5), 349–367.

Yom-Tov, Elad, Susan Dumais, and Qi Guo, "Promoting Civil Discourse Through Search Engine Diversity," *Social Science Computer Review*, 2013, pp. 1–10.

APPENDICES

Appendix.1 Proofs

Proof of Lemma 1. From Eqs. (2) and (4), we have that clicks on item m get updated according to:

$$\widehat{\kappa}_{n,m} - \widehat{\kappa}_{n-1,m} = \frac{\beta^{M-r_{n,m}}\varphi_{n,m}}{\sum_{m' \in M} \beta^{M-r_{n,m'}\varphi_{n,m'}}}.$$

Taking expectations of an average run out of T runs, we have $\varphi_{n,m} \approx 1/(2[m])$ and can write:

$$\begin{aligned}
\mathbf{E}[\widehat{\kappa}_{n,m} - \widehat{\kappa}_{n-1,m}] &= \begin{cases} \frac{\beta^{M-r_{n,m}}/(2M_+)}{\sum_{m' \in M_+} \beta^{M-r_{n,m}}/(2M_+) + \sum_{m' \in M_-} \beta^{M-r_{n,m}}/(2M_-)} & \text{if } m \in M_+ \\ \\ \frac{\beta^{M-r_{n,m}}/(2M_-)}{\sum_{m' \in M_-} \beta^{M-r_{n,m}}/(2M_+) + \sum_{m' \in M_+} \beta^{M-r_{n,m}}/(2M_-)} & \text{if } m \in M_- \end{cases} \\
&\overset{(1)}{\approx} \frac{\beta^{M-r_{n,m}}}{\sum_{m' \in M_+} \beta^{M-r_{n,m'}} + \sum_{m' \in M_-} \beta^{M-r_{n,m'}}} \qquad \text{for all } m \in M \\
&= \frac{\beta^{M-r_{n,m}}}{\sum_{m' \in M} \beta^{M-m'}} = \frac{\beta^M \beta^{\frac{M-r_{n,m}}{M}}}{\beta^M \sum_{m' \in M} \beta^{\frac{M-m'}{M}}} = \frac{\beta^{\frac{M-r_{n,m}}{M}}}{\sum_{m' \in M} \beta^{\frac{M-m'}{M}}},
\end{aligned}$$

where (1) follows from $\mathbb{P}(\text{sgn}(x_n) = \text{sgn}(y_m)) \approx \mathbb{P}(\text{sgn}(x_n) \neq \text{sgn}(y_m)) \approx 1/2$ and from $[m] \approx M_+ \approx M_- \approx M/2$, for T sufficiently large (in the limit as $T \to \infty$).

From this we can write the expected probability of an item with signal $y_m = y$ being clicked in an average run out of T runs (again as $T \to \infty$) as:

$$\mathbf{E}[\hat{\kappa}(y)] = \frac{\beta^{\frac{M-r(y)}{M}}}{\sum_{m' \in M} \beta^{\frac{M-m'}{M}}} \overset{(1)}{=} \frac{\beta^{\frac{M-\zeta_0+\zeta_1 \cdot \pi(y)}{M}}}{\sum_{m' \in M} \beta^{\frac{M-m'}{M}}} \approx \frac{1 + \log\beta \cdot \frac{M-\zeta_0+\zeta_1 \cdot \pi(y)}{M}}{1 + \log\beta \cdot \sum_{m' \in M} \frac{M-m'}{M}},$$

where (1) follows from applying Eq. (12) to the expected rank $r(y)$. Taking into account the distribution of the items' signals g, this readily implies as the limit clicking distribution:[1]

$$LCD(y) = \Lambda_\beta(\pi(y)) \cdot g(y), \tag{A.1}$$

where, for $z \geq 0$:

$$\Lambda_\beta(z) = \frac{M + \log\beta \cdot (M - \zeta_0 + \zeta_1 \cdot z)}{M + \log\beta \cdot \sum_{m' \in M}(M - m')} = \frac{M + \log\beta \cdot (M - \zeta_0 + \zeta_1 \cdot z))}{M + \log\beta \cdot M(M-1)/2}, \tag{A.2}$$

so that $\Lambda'_\beta(z)$ is a constant and $\Lambda'_\beta(z) \equiv \Lambda'_\beta = \frac{\zeta_1 \cdot \log\beta}{M + \log\beta \cdot M(M-1)/2} > 0$ since $\beta > 1, \zeta_1 > 0$.

Similarly, we can write the limit highlighting distribution as:

$$LHD(y) = \mu_H(y) \cdot LCD(y). \tag{A.3}$$

These distributions do not integrate to 1 but rather give a per capita probability of clicking and highlighting a given item with signal $y_m = y$. □

Proof of Proposition 1. Set $\theta = \widehat{\theta} = 0$ and fix $\beta > 1$. To simplify notation we drop the subscript β from the function Λ_β and write just Λ. Given Eq. (12), we can apply Lemma 1 and write engagement (ENG),

[1] To cut on notation, throughout the proofs, we write $f(x)$ for the density of the individuals signals ($f(x;\sigma_x^2)$ in the main text) and $g(y)$ for the density of the items' signals ($f(y;\sigma_y^2)$ in the main text). Note that $g(y)$ is not to be confused with the symbol g also used to denote a group of individuals.

polarization (POL) and misinformation (MIS), respectively as:

$$\begin{aligned}
ENG &= \int \left(LCD(y) + LHD(y)\right) dy = \int \left(1+\mu_H(y)\right) LCD(y) dy = \int \left(1+\mu_H(y)\right) \Lambda(\pi(y)) f(y) dy \\
MIS &= \int |y-0|\, LCD(y) dy = \int |y|\, \Lambda(\pi(y)) f(y) dy \\
POL &= \int \left|yLCD^R(y) - yLCD^L(y)\right| dy = \int \left|y\Lambda^R(\pi(y)) - y\Lambda^L(\pi(y))\right| f(y) dy,
\end{aligned}$$

where for $g \in \{L, R\}$, LCD^g is the limit clicking distribution of individuals from group g and (in the non-personalized case with $\lambda = 1$) can be written as:

$$LCD^g(y) = \Lambda^g(\pi(y))g(y),$$

where $\Lambda^g(\pi(y))$ is now the expected probability an item with signal $y_m = y$ will be clicked on by an individual in group g. Note that while clicking and highlighting by a given group is heavily dependent on the sign of the signal of the item, (that is, whether $m \in M_+$ or $m \in M_-$), both groups share the same ranking which depends on the total clicking and highlighting propensities of the two groups ($\pi(y)$).

From this we can compute the effect of a change in η on he three variables. Suppose that highlighting behavior is non-flat. It suffices to compute:

$$\begin{aligned}
\frac{\partial ENG}{\partial \eta} &= \frac{\partial}{\partial \eta} \int \left(1+\mu_H(y)\right) \Lambda(\pi(y)) g(y) dy \\
&= \int \frac{\partial (1+\mu_H(y))}{\partial \eta} \Lambda(\pi(y)) g(y) dy + \int \left(1+\mu_H(y)\right) \frac{\partial \Lambda(\pi(y))}{\partial \eta} g(y) dy \\
&\overset{(1)}{=} \int \left(1+\mu_H(y)\right) \Lambda' \frac{\partial \pi(y)}{\partial \eta} g(y) dy \\
&\overset{(2)}{=} \int \left(1+\mu_H(y)\right) \Lambda' \frac{(M-1)(\mu_H(y) - \bar{\mu}_H)}{\left(M(1+\eta\bar{\mu}_H) + \eta(\mu_H(y) - \bar{\mu}_H)\right)^2} g(y) dy \\
&\overset{(3)}{>} 0,
\end{aligned}$$

where (1) follows because $\frac{\partial \mu_H(y)}{\partial \eta} = 0$ and $\frac{\partial g(y)}{\partial \eta} = 0$, (2) follows from Eq. (11), and (3) follows since $\Lambda'(\pi(y)) = \Lambda' > 0$ and $\mu_H(y) \geq 0$ for all y and, moreover, for $|y| > |y'|$ we have $\mu_H(y) > \mu_H(y')$ on a large enough mass of signals y (strictly speaking until $-x^*, x^*$), and hence also for $\frac{1+\mu_H(y)}{(M(1+\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^2}$. The latter implies $\int \frac{(1+\mu_H(y))(\mu_H(y)-\bar{\mu}_H)}{(M(1+\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^2} g(y) dy > 0$. Since $\int (\mu_H(y) - \bar{\mu}_H) g(y) dy = 0$, increasing η corresponds to shifting mass towards signals with higher absolute value, thereby strictly increasing engagement. Similarly:

$$\begin{aligned}
\frac{\partial MIS}{\partial \eta} &= \frac{\partial}{\partial \eta} \int |y| \Lambda(\pi(y)) g(y) dy = \int |y| \frac{\partial \Lambda(\pi(y))}{\partial \eta} g(y) dy \\
&= \int |y| \Lambda' \frac{(M-1)(\mu_H(y) - \bar{\mu}_H)}{\left(M(1+\eta\bar{\mu}_H) + \eta(\mu_H(y) - \bar{\mu}_H)\right)^2} g(y) dy \\
&\overset{(1)}{>} 0,
\end{aligned}$$

where (1) follows again because $\Lambda'(\pi(y)) > 0$ and $\mu_H(y) \geq 0$, and for $|y| > |y'|$ we have $\mu_H(y) > \mu_H(y')$ on a large enough mass of signals y, ensuring that $\int \frac{|y|(\mu_H(y)-\bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^2} g(y) dy > 0$.

Finally, to compute the effect on polarization, we need to keep track of clicking in the two groups. While there is a unique ranking (since $\lambda = 1$), individuals in the different groups nonetheless behave differently.

$$\begin{aligned}
\frac{\partial POL}{\partial \eta} &= \frac{\partial}{\partial \eta}\left|\int y\Lambda^R(\pi(y)) - y\Lambda^L(\pi(y))g(y)dy\right| \\
&= \frac{\partial}{\partial \eta}\left|\int y\left(\Lambda^R(\pi(y)) - \Lambda^L(\pi(y))\right)g(y)dy\right| \\
&\overset{(1)}{=} \frac{\partial}{\partial \eta}\left(\int_{y\leq 0}(-y)\left(\Lambda^L(\pi(y)) - \Lambda^R(\pi(y))\right)g(y)dy + \int_{y>0} y(\Lambda^R(\pi(y)) - \Lambda^L(\pi(y)))g(y)dy\right) \\
&\overset{(2)}{=} \frac{\partial}{\partial \eta}\left(2\int_{y>0} y\left(\Lambda^R(\pi(y)) - \Lambda^L(\pi(y))\right)g(y)dy\right) \\
&= 2\int_{y>0} y\left(\frac{\partial\Lambda^R(\pi(y))}{\partial \eta} - \frac{\partial\Lambda^L(\pi(y))}{\partial \eta}\right)g(y)dy \\
&\overset{(3)}{=} 2\int_{y>0} y\left(\Lambda^R_+\pi(y)\frac{\partial\pi(y)}{\partial \eta} - \Lambda^L_+\pi(y)\frac{\partial\pi(y)}{\partial \eta}\right)g(y)dy \\
&= 2\int_{y>0} y\left(\Lambda^R_+ - \Lambda^L_+\right)\pi(y)\frac{\partial\pi(y)}{\partial \eta}g(y)dy \\
&= 2\int_{y>0} y\left(\Lambda^R_+ - \Lambda^L_+\right)(1+\eta\mu_H(y))\frac{(M-1)(\mu_H(y)-\bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^3}g(y)dy \\
&\overset{(4)}{>} 0,
\end{aligned}$$

where (1) follows because of $\mathbb{R}_-$ we have $\Lambda^L(\pi(y)) > \Lambda^R(\pi(y))$, and on $\mathbb{R}_+$ we have $\Lambda^R(\pi(y)) > \Lambda^L(\pi(y))$, (2) follows by symmetry of the limit clicking distribution, (3) follows since Λ^R, Λ^L are linear and hence, for y on $\mathbb{R}_+$, $\Lambda^{R'}(y) \equiv \Lambda^R_+$ and $\Lambda^{L'}(y) \equiv \Lambda^L_+$ are positive constants with $\Lambda^R_+ > \Lambda^L_+$, and finally (4) follows for the same reasons as with the previous cases (ENG and MIS) since $\mu_H(y)$ and y are increasing in y and $\Lambda^R_+ > \Lambda^L_+$ on $\mathbb{R}_+$, ensuring $\int_{y>0}\frac{(1+\eta\mu_H(y))(\mu_H(y)-\bar{\mu}_H)}{(M(1+\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^3}g(y)dy > 0$. Note also that due to symmetry, $\int_{y>0}(\mu_H(y)-\bar{\mu}_H)g(y)dy = 0$.

Suppose now that highlighting behavior is flat. The above expressions continue to hold:

$$\begin{aligned}
\frac{\partial ENG}{\partial \eta} &= \Lambda' \cdot \int (1+\mu_H(y))\frac{(M-1)(\mu_H(y)-\bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^3}g(y)dy \\
&> 0.
\end{aligned}$$

However, in the flat case, as $|y|$ increases $\mu_H(y)$ decreases, or equivalently, $\mu_H(y)$ increases as $|y|$ decreases, but $\mu_H(y) - \bar{\mu}_H$ is positive for smaller values of y and hence the above integral is positive.

In the case of misinformation, the argument is reversed, since it is now $\frac{|y|}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^2}$ that multiplies $\mu_H(y) - \bar{\mu}_H$ and hence the fact that as $|y|$ increases the whole fraction increases, while $\mu_H(y)$ decreases, which means that the overall integral is now negative.

$$\begin{aligned}
\frac{\partial MIS}{\partial \eta} &= \Lambda' \cdot \int |y|\frac{(M-1)(\mu_H(y)-\bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^2}g(y)dy \\
&< 0.
\end{aligned}$$

A similar argument applies for polarization:

$$\begin{aligned}
\frac{\partial POL}{\partial \eta} &= 2\left(\Lambda^R_+ - \Lambda^L_+\right)\int_{y>0} y(1+\eta\mu_H(y))\frac{(M-1)(\mu_H(y)-\bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^3}g(y)dy \\
&< 0.
\end{aligned}$$

Here the inequality follows since the integral is on $\mathbb{R}_+$ so that it is $\frac{y(1+\eta\mu_H(y))}{(M(1+\eta\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^3}$ that multiplies the expression $\mu_H(y) - \bar{\mu}_H$, where it can be checked that the former is increasing in y, while $\mu_H(y)$ is decreasing in y, making the overall integral negative. Note that as before, due to symmetry, $\int_{y>0}(\mu_H(y) - \bar{\mu}_H)g(y)dy = 0$. □

Proof of Proposition 2. Set again $\theta = \widehat{\theta} = 0$ and fix $\beta > 1$, and write Λ for the function Λ_β, thus dropping the subscript β. Applying Eq. (12) to the personalized algorithm Eq. (7), we obtain for the expected rank:

$$r^g(y) \approx \zeta_0 - \zeta_1 \cdot \frac{\pi^g_g(y_m) + \lambda\pi^{\neg g}_g(y_m)}{1+\lambda}, \;\; g \in \{L, R\}, \tag{A.4}$$

where $\zeta_0, \zeta_1 > 0$ are constants and $\neg g$ denotes the group in $\{L, R\}$ other than g. Here the expressions $\pi^g_g(y)$ and $\pi^{\neg g}_g(y)$ denote respectively the popularity from individuals in g and in $\neg g$ in the ranking seen by group g:[2]

$$\pi^g_g(y) = \frac{1+\eta\cdot\mu^g_H(y)}{M(1+\eta\cdot\bar{\mu}^g_H) + \eta(\mu^g_H(y) - \bar{\mu}^g_H) + \lambda\left(M(1+\eta\cdot\bar{\mu}^{\neg g}_H) + \eta(\mu^{\neg g}_H(y) - \bar{\mu}^{\neg g}_H)\right)}$$

and

$$\pi^{\neg g}_g(y) = \frac{\lambda(1+\eta\cdot\mu^{\neg g}_H(y))}{M(1+\eta\cdot\bar{\mu}^g_H) + \eta(\mu^g_H(y) - \bar{\mu}^g_H) + \lambda\left(M(1+\eta\cdot\bar{\mu}^{\neg g}_H) + \eta(\mu^{\neg g}_H(y) - \bar{\mu}^{\neg g}_H)\right)},$$

where $\mu^g_H(y)$ is the propensity to highlight by individuals in g:

$$\mu^g_H(y) = \int_{x\in H^{-1}(y)} \mathbb{I}_{\{x\in g\}} p_A(x) f(x) dx.$$

We can apply Lemma 1 and write engagement as:

$$\begin{aligned} ENG &= \sum_{g=L,R} \int \left(LCD^g(y) + LHD^g(y)\right) dy = \sum_{g=L,R} \int \left(1 + \mu^g_H(y)\right) LCD^g(y) dy \\ &= \sum_{g=L,R} \int \left(1 + \mu^g_H(y)\right) \Lambda^g\left(\frac{\pi^g_g(y) + \lambda\pi^{\neg g}_g(y)}{1+\lambda}\right) g(y) dy, \end{aligned}$$

where as in the proof of Proposition 1, $\Lambda^g\left(\frac{\pi^g_g(y)+\lambda\pi^{\neg g}_g(y)}{1+\lambda}\right)$ is the probability of being clicked by an individual in g:

$$\Lambda^g\left(\frac{\pi^g_g(y) + \lambda\pi^{\neg g}_g(y)}{1+\lambda}\right) = \frac{M + \log\beta\cdot\left(M - \zeta_0 + \zeta_1\frac{\pi^g_g(y)+\lambda\pi^{\neg g}_g(y)}{1+\lambda}\right)}{M + \log\beta\cdot\sum_{m'\in M}(M - m')}, \tag{A.5}$$

[2] The distinction is necessary because of the normalizations that get applied to the two different rankings and that therefore change the denominators in the two cases.

We can compute

$$\begin{aligned}
\frac{\partial ENG}{\partial \lambda} &= \sum_{g=L,R} \frac{\partial}{\partial \lambda} \int \left(1+\mu_H^g(y)\right) \Lambda^g \left(\frac{\pi_g^g(y)+\lambda \pi_g^{\neg g}(y)}{1+\lambda}\right) g(y)dy \\
&= \sum_{g=L,R} \int \left(1+\mu_H^g(y)\right) \frac{\partial \Lambda^g \left(\frac{\pi_g^g(y)+\lambda \pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial \lambda} g(y)dy \\
&= \sum_{g=L,R} \left(\int_{y\leq 0} \left(1+\mu_H^g(y)\right) {\Lambda_-^g}' \frac{\partial \left(\frac{\pi_g^g(y)+\lambda \pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial \lambda} g(y)dy \right. \\
&\qquad\qquad \left. + \int_{y>0} \left(1+\mu_H^g(y)\right) {\Lambda_+^g}' \frac{\partial \left(\frac{\pi_g^g(y)+\lambda \pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial \lambda} g(y)dy \right) \\
&= 2 \sum_{g=L,R} \int_{y>0} \left(1+\mu_H^g(y)\right) {\Lambda_+^g}' \frac{\partial \left(\frac{\pi_g^g(y)+\lambda \pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial \lambda} g(y)dy \\
&= 2 \sum_{g=L,R} \int_{y>0} \frac{\left(1+\mu_H^g(y)\right) {\Lambda_+^g}'}{1+\lambda} \left(\frac{\partial \pi_g^g(y)}{\partial \lambda} + \lambda \frac{\partial \pi_g^{\neg g}(y)}{\partial \lambda} - \frac{\pi_g^g(y)-\pi_g^{\neg g}(y)}{1+\lambda} \right) g(y)dy
\end{aligned}$$

Now,

$$\frac{\partial \pi_g^g(y)}{\partial \lambda} = \frac{-(1+\eta\cdot\mu_H^g(y))\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g})+\eta(\mu_H^{\neg g}(y)-\bar{\mu}_H^{\neg g})\right)}{\left(M(1+\eta\cdot\bar{\mu}_H^g)+\eta(\mu_H^g(y)-\bar{\mu}_H^g)+\lambda\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g})+\eta(\mu_H^{\neg g}(y)-\bar{\mu}_H^{\neg g})\right)\right)^2},$$

and

$$\frac{\partial \pi_g^{\neg g}(y)}{\partial \lambda} = \frac{(1+\eta\cdot\mu_H^{\neg g}(y))\left(M(1+\eta\cdot\bar{\mu}_H^{g})+\eta(\mu_H^{g}(y)-\bar{\mu}_H^{g})\right)}{\left(M(1+\eta\cdot\bar{\mu}_H^g)+\eta(\mu_H^g(y)-\bar{\mu}_H^g)+\lambda\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g})+\eta(\mu_H^{\neg g}(y)-\bar{\mu}_H^{\neg g})\right)\right)^2},$$

where it can be checked that:

$$\frac{\partial\left(\pi_g^g(y)+\pi_g^{\neg g}(y)\right)}{\partial \lambda} = \frac{(1+\eta\cdot\mu_H^{\neg g}(y))\eta(\mu_H^g(y)-\bar{\mu}_H^g)-(1+\eta\cdot\mu_H^g(y))\eta(\mu_H^{\neg g}(y)-\bar{\mu}_H^{\neg g})}{\left(M(1+\eta\cdot\bar{\mu}_H^g)+\eta(\mu_H^g(y)-\bar{\mu}_H^g)+\lambda\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g})+\eta(\mu_H^{\neg g}(y)-\bar{\mu}_H^{\neg g})\right)\right)^2},$$

which, integrated on $\mathbb{R}_+$, is strictly negative for $g = R$ and outweighs (in absolute value) the case for $g = L$, such that for any $\lambda \in [0, 1]$ and whether or not $\mu_H^g(y)$ increases in y:

$$\sum_{g=L,R} \int_{y>0} \frac{\left(1+\mu_H^g(y)\right){\Lambda_+^g}'}{1+\lambda} \frac{\partial\left(\pi^g(y)+\lambda\pi^{\neg g}(y)\right)}{\partial \lambda} g(y)dy < 0.$$

Moreover, a similar argument applies for $-\frac{\pi_g^g(y)-\pi_g^{\neg g}(y)}{1+\lambda}$, such that also:

$$\sum_{g=L,R} \int_{y>0} \frac{\left(1+\mu_H^g(y)\right){\Lambda_+^g}'}{1+\lambda} \frac{(-1)(\pi_g^g(y)-\pi_g^{\neg g}(y))}{1+\lambda} g(y)dy < 0.$$

This shows that $\frac{\partial ENG}{\partial \lambda} < 0$ so that less personalization (larger λ) decreases engagement both with flat and non-flat highlighting.

Finally,

$$POL = \left| \int yLCD^R(y)dy - \int yLCD^L(y)dy \right|,$$

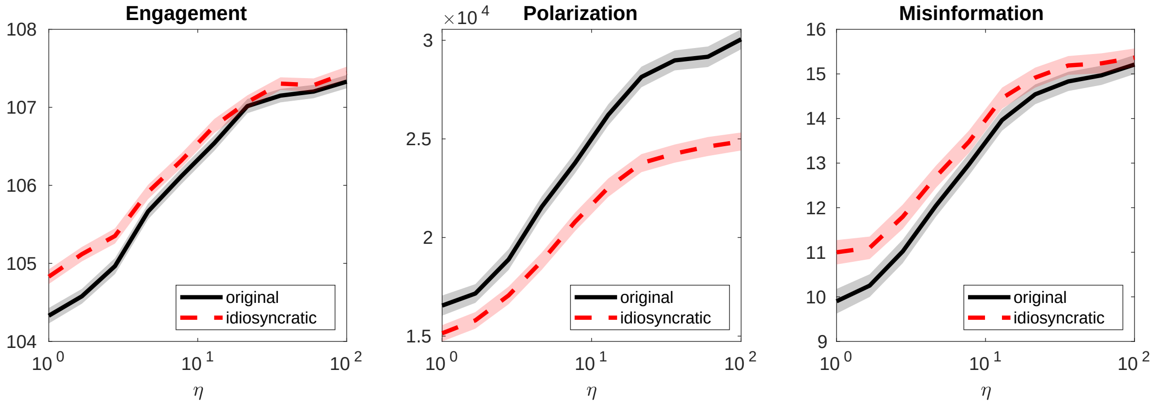

Figure A.1: Engagement, polarization, and misinformation as a function of the highlighting parameter η (non-flat case) with a common benchmark $\widehat{\theta}$ (solid line) and heterogeneous benchmarks $\widehat{\theta}_n$ (dotted line). The shaded areas represent the 95% confidence intervals.

so that, using the same reasoning as in the proof of Proposition 1, we can write:

$$\begin{aligned}\frac{\partial POL}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left(2\int_{y>0} y\left(\Lambda^R\left(\frac{\pi_R^R(y)+\lambda\pi_R^L(y)}{1+\lambda}\right)-\Lambda^L\left(\frac{\pi_L^L(y)+\lambda\pi_L^R(y)}{1+\lambda}\right)\right)g(y)dy\right)\\ &= 2\int_{y>0} y\Lambda_+^{R\prime}\left(\frac{\partial\pi_R^R(y)}{\partial\lambda}+\lambda\frac{\partial\pi_R^L(y)}{\partial\lambda}-\frac{\pi_R^R(y)-\pi_R^L(y)}{1+\lambda}\right)g(y)dy\\ &\qquad -2\int_{y>0} y\Lambda_+^{L\prime}\left(\frac{\partial\pi_L^L(y)}{\partial\lambda}+\lambda\frac{\partial\pi_L^R(y)}{\partial\lambda}-\frac{\pi_L^L(y)-\pi_L^R(y)}{1+\lambda}\right)g(y)dy.\end{aligned}$$

Moreover, similar calculations as above show that the first integral is negative and dominates in absolute value the second one, showing that overall $\frac{\partial POL}{\partial \lambda} < 0$ so that again less personalization (larger λ) decreases polarization both with flat and non-flat highlighting. □

Proof of Proposition 3. Recall from Eq. (10):

$$W_\psi(\eta,\lambda) = \psi \cdot ENG(\eta,\lambda) - (1-\psi)\cdot MIS(\eta,\lambda)\cdot POL(\eta,\lambda).$$

Hence, for $\psi = 0$, we have $W_0 = MIS \cdot POL$, while, for $\psi = 1$, we have $W_1 = ENG$. The results then follow directly from Propositions 1 and 2.

Consider the non-flat case. It follows immediately that W_0 is maximized at a smallest possible value of η, since $-MIS \cdot POL$ is maximized when $MIS \cdot POL$ is minimized and $\frac{MIS}{\partial \eta} > 0$, $\frac{POL}{\partial \eta} > 0$. Also, W_0 is maxized at a largest possible value of λ again since $\frac{POL}{\partial \lambda} < 0$ (less personalization decreases POL) while $\frac{MIS}{\partial \lambda} \approx 0$. The contrary is true for $\psi = 1$.

By contrast, by analogous argument, in the flat case, W_0 is maximized at a largest possible value of η and at a largest possible value of λ, while for $\psi = 1$, W_1 is maximized at a largest possible value of η and a smallest possible value of λ. □

Appendix.2 Additional Results

We here discuss some additional results that were not discussed or only very briefly mentioned in the main text of the paper.

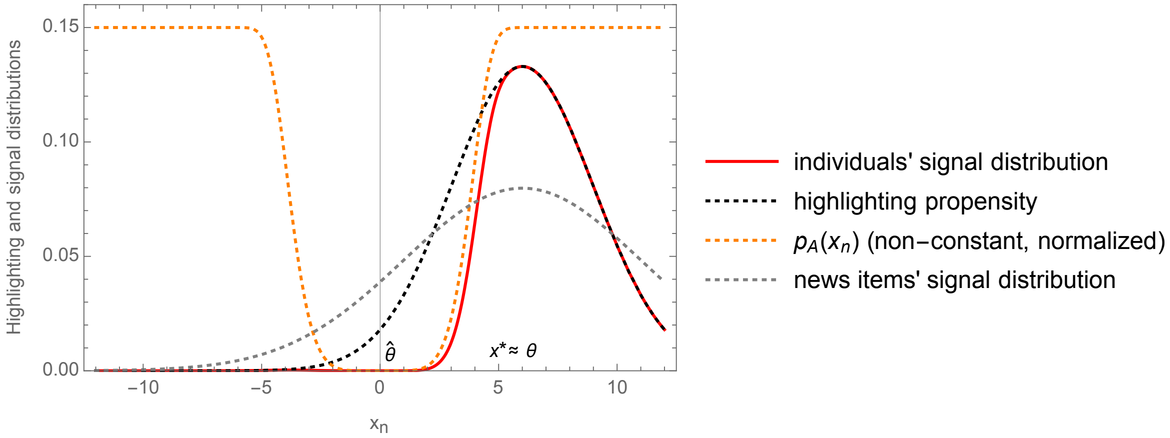

Figure A.2: Individuals' signal distribution and highlighting propensity in the non-flat case with non-centered $\widehat{\theta}$, (with $\theta = 6$ and $\widehat{\theta} = 0$); x^* denotes the value of x_n where the highlighting propensity is locally maximal.

Appendix.2.1 Heterogeneous benchmark

Consider the case where individuals can have idiosyncratic benchmarks $\widehat{\theta}_n$. The results do not change qualitatively. In fact, the simulations suggest that, for $\widehat{\theta}_n$'s centered around θ and not too dispersed ($\widehat{\theta}_n \sim N(\theta, \sigma^2_{\widehat{\theta}})$ with $\sigma_{\widehat{\theta}} \leq \min\{\frac{\sigma_x}{4}, \frac{\sigma_y}{4}\}$), our main results on engagement, popularity and misinformation are rather close to the cases where individuals have a common benchmark, $\widehat{\theta}_n = \widehat{\theta}$ for all n. This is illustrated in Figure A.1 that shows the effect of η on the variables ENG, POL, MIS for the non-flat case.

Appendix.2.2 Non-centered benchmark

While it is natural to assume that the benchmark $\widehat{\theta}$ splits the signals roughly in half in symmetric environments, so that $\widehat{\theta} \approx \theta$, it may occur occasionally that the two are far apart. In such a situation, individuals' and news items' signals are shifted away from the benchmark $\widehat{\theta}$. This means that a potentially large mass of individuals have a prior belief far from $\widehat{\theta}$ and are hence likely to highlight news items far from it but potentially close to θ. In such a case, an increase in η can contribute to both higher engagement and at the same time lower misinformation in the non-flat case. To see this consider Figure A.2 that illustrates a situation where clearly $\widehat{\theta} \neq \theta$. Here $x^* \approx \theta$ so that a large mass of individuals with a signal close to the truth has a large highlighting propensity. An increase in η leads to a more prominent ranking for items around $x^* \approx \theta$, which in turn, through the effect on the clicking distribution, leads to a lower level of misinformation as measured by MIS. Increasing the weight on highlights here actually accelerates individuals clicking on news items carrying truthful signals.

Appendix.3 Empirical Evidence on Meaningful Social Interactions and Political Polarization: Additional Information

Appendix.3.1 Affective Polarization

Since Italy is a multi-party political system, we follow Alvarez and Nagler (2004) and Wagner (2021) and define a measure of Weighted Affective Polarization (WAP) for individual i as:

$$WAP = \sqrt{\sum_{p=1}^{P} v_p * \mid symp_{ip} - \overline{symp_i} \mid}, \qquad \text{(A.6)}$$

where v_p is the vote share of party p (measured as a proportion ranging from 0 to 1), $symp_{ip}$ is measured with the probability attached by individual i to voting for party p (ranging from 0 to 10), and $\overline{symp_i}$ is individual i's weighted average party sympathy score. That is:

$$\overline{symp_i} = \sum_{p=1}^{P} v_p * symp_{ip}\,. \tag{A.7}$$

Appendix.3.2 Robustness

One possible concern regarding the causal interpretation of our results linking Facebook's MSI update and political polarization is due to the concurring general elections in Italy in March 2018. With respect to this issue we notice that, by including the date of interview fixed-effect, our empirical strategy takes into account and controls for any general trend in political polarization over time. At the same time, one might argue that the presence of elections might have led to a differential trend in political polarization between individuals that used internet to form an opinion and the ones who did not which was not due to the MSI algorithm *per se* (e.g., increase in online fake news before elections). In response to this argument, we first point out that the MSI algorithm might have further amplified the diffusion of fake-news as predicted by our model. Second, we provide below evidence suggesting a polarization effect even when dropping the months immediately after the MSI update and before the elections (i.e., January-March 2018). Specifically, Tables A.1 and A.2 present results when comparing the period June-December 2017 (pre-MSI) with April-December 2018 (post-MSI and post-elections).

Table A.1: MSI and non-moderate ideological position: Robustness

	(1)	(2)	(3)
	Non-moderate	Non-moderate	Non-moderate
	Ideology	Ideology	Ideology
Opinion via internet websites	0.047**	0.048**	0.048**
× Post MSI	(0.018)	(0.020)	(0.021)
Opinion via internet websites	-0.011	-0.013	-0.013
	(0.018)	(0.022)	(0.022)
Observations	29,570	29,570	29,570
Mean outcome	0.37	0.37	0.37
SD outcome	0.48	0.48	0.48
Municipality FE	YES	YES	YES
Date FE	YES	NO	NO
Province-Date FE	NO	YES	YES
Cluster SE	Region	Region	Province

Note: Time horizon: June 2017-December 2017 and April-December 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation and religiosity of the respondent. Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis.
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table A.2: MSI and Affective Polarization

	(1) Affective Polarization	(2) Affective Polarization	(3) Affective Polarization
Opinion via internet websites	0.051	0.064**	0.064*
× Post MSI	(0.030)	(0.031)	(0.034)
Opinion via internet websites	-0.007	-0.008	-0.008
	(0.021)	(0.023)	(0.022)
Observations	17,317	17,317	17,317
Mean outcome	1.38	1.38	1.38
SD outcome	0.67	0.67	0.67
Municipality FE	YES	YES	YES
Date FE	YES	NO	NO
Province-Date FE	NO	YES	YES
Cluster SE	Region	Region	Province

Note: Time horizon: June 2017-December 2017 and April-December 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation and religiosity of the respondent. Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis.

*** p<0.01, ** p<0.05, * p<0.1