Discussion Paper Series – CRC TR 224

Discussion Paper No. 580
Project B 05

Feed for Good? On the Effects of Personalization Algorithms in Social Platforms

Miguel Risco[1]
Manuel Lleonart-Anguix[2]

August 2024

[1]University of Bonn, Email: risco@uni-bonn.de
[2]The Autonomous University of Barcelona, Email: manuel.lleonart@bse.eu

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Collaborative Research Center Transregio 224 - www.crctr224.de
Rheinische Friedrich-Wilhelms-Universität Bonn - Universität Mannheim

Feed for good? On the effects of personalization algorithms in social platforms*

Miguel Risco† Manuel Lleonart-Anguix‡

August 6, 2024

Abstract

This paper builds a theoretical model of communication and learning on a social media platform, and describes the algorithm an engagement-maximizing platform implements in equilibrium. Such algorithm excessively exploits similarity, locking users in echo chambers. Moreover, learning vanishes as platform size grows large. As this is far from ideal, we explore alternatives. The reverse-chronological algorithm the DSA mandated to reincorporate turns out to be not good enough, so we build the "breaking echo chambers" algorithm, a modification of the platform-optimal algorithm that improves learning by promoting opposite thoughts. Additionally, we seek a natural implementation path for the utilitarian optimal algorithm. This is why we advocate for horizontal interoperability, which interoperability compels platforms to compete based on algorithms. In the absence of platform-specific network effects that entrench users within dominant platforms, the retention of user bases hinges on implementing algorithms that outperform those of competitors.

Keywords: personalized feed, social learning, network effects, interoperability

JEL Codes: D43, D85, L15, L86.

*Risco acknowledges financial support by the German Research Foundation (DFG) through CRC TR 224 (Project B05) and funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 949465). We thank Sven Rady, Pau Milán, Francesc Dilmé, Alexander Frug, Carl-Christian Groh, David Jiménez-Gómez, Daniel Krähmer, Philipp Strack, Marc Bourreau, Mikhail Drugov, Manuel Mueller-Frank, Robin Ng, Raquel Lorenzo, Malachy Gavan, Francisco Poggi, Fernando Payro, Jordi Caballé, Inés Macho-Stadler, Iván Rendo and Diego Fica and the participants at the BSE Summer Forum in Digital Platforms, the EAYE 2024, the Paris Digital Economics Conference, the MaCCI Annual 2024, the CRC TR 224 retreat, the 6th Workshop on The Economics of Digitalization, the EDP Jamboree, the 2nd ECONtribute YEP Workshop, the CoED, the JEI, various CRC TR 224 YRWs, the ENTER Jamboree, the BSE Jamboree, as well as seminars and workshops at the University of Bonn, UPF, TSE, UAB, UV, UA, EUI and UB for their helpful comments and discussions.

†Bonn Graduate School of Economics, University of Bonn. risco@uni-bonn.de

‡Universitat Autònoma de Barcelona. manuel.lleonart@bse.eu

1 Introduction

On May 25, 2024, a video went viral on TikTok after showing the stark difference in comments displayed on Instagram to Eli, the user who posted it, and her boyfriend when reading the same post.[1] In the post, which is public, we see a girl waiting for her boyfriend, who was supposed to meet her at 3 p.m. after playing golf. The post shows the girl recording herself after each extra half-hour she has to wait for him. When Eli read the comments below the post, which were displayed under the *most relevant* tag, they were along the lines of "oh this is so rude!" or "it is disregard of her time"—as she expected, as she literally says. Eli then sent the post to her boyfriend, who was sitting next to her. He opened it at the same time, but surprisingly, the *most relevant* comments were strikingly different: "or you could get your own hobby instead of waiting around for him", "he meant 3 a.m., he is ahead of schedule" or "God forbid he has a good time". People look at the comments on a video to gain perspective and see how others feel about it. However, now, with personalized feeds, each user gets tailor-made content based on her interactions and behavior in the app. "The only difference about our interactions with Instagram is that he is a guy and I am a girl", Eli says. She tried to find the comments that appeared to her boyfriend in her own list, but she could not.

Eli's video went viral, reaching almost three million views and adding fuel to the social debate on the effects of personalized social media feeds on people's beliefs and perspectives. However, these concerns are not new, as platforms have been criticized for causing polarization and spreading misinformation, promoting echo chambers, and fueling hate speech (Silverman, 2016; Allcott and Gentzkow, 2017). The 2016 US presidential elections were likely the turning point where public opinion began to question the suitability of personalization and its consequences on beliefs and decisions; Facebook was accused of failing to combat fake news (Solon, 2016). Since then, much evidence on this has been collected at the academic, empirical level (Allcott et al., 2022; Bursztyn et al., 2023), showing how harmful engagement-maximizing platforms are and how they trap users in their echo-chambers (in particular, Bursztyn et al. (2023) show that users would be willing to pay to have others, including themselves, deactivating their TikTok and Instagram accounts). Moreover the journalistic investigations by Horwitz et al. (2021) called "The Facebook Files" revealed that Meta internally acknowledges the harmful effects of its algorithms on users, specifically on female teenagers ("[t]ime and again, the documents show, Facebook's researchers have identified the platform's ill effects"). However, platforms can create value through their superior level of information and,[2] then, it is essential to investigate how personalization algorithms affect social welfare, as their repercussions have

1 The video is public and can be accessed at https://www.tiktok.com/@elieli0000/video/7373012517016079649?lang=es.

2 Quoting Scott Morton et al. (2019): "The speed, scale, and scope of the internet, and of the ever-more powerful technologies it has spawned, have been of unprecedented value to human society."

emerged as a significant economic concern.

The *feed* is a customized scroll of friends' content and news stories that appears on most social media platforms. Until around 2015, it was reverse-chronological.[3] Now, a proprietary algorithm controls what appears on the screen, based on user behavior on the platform. Since platforms' revenues come from advertising, their primary goal is to maximize engagement, which may not align with promoting informative communication. If, as Eli's video shows, a biased set of comments will maximize the probability you stay on the platform longer, this is what you will receive. Personalized algorithms account for the increase in engagement and addictive behavior in social media platforms, regardless of the field (Guess et al., 2023).

The approval of the Digital Service Act (DSA) and the Digital Markets Act (DMA) by the European Commission in 2022 represents one of the first efforts to address the problems arising from algorithm personalization through regulation. In particular, the DSA requires platforms to reinstate the reverse-chronological algorithm as an option for their users, thereby providing an alternative to personalization. Some platforms, like X, were very compliant, while others, like Instagram, were less so: it is not only complicated to find the button reverting the feed to reverse-chronological, but the feed goes personalized again once you log out. Still, it does not seem that the availability of the reverse-chronological algorithm is really alleviating any of the urgent media-related problems society faces. Moreover, personalization need not be detrimental to social welfare; it could be used for an improvement, as the following quote illustrates (Lauer, 2021): "If Facebook employed a business model focused on efficiently providing accurate information and diverse news, rather than addicting users to highly engaging content within an echo chamber, the algorithmic outcomes would be very different". To achieve this, however, we would need to find a way to align the platform's incentives with social well-being, so that naturally, the optimal algorithm for the platform would also be optimal for the users.

With all this in mind, we claim that there is a need for theoretical research that guides the optimal regulatory approach, understanding the incentives of platforms in designing their optimal algorithm and how they would respond to regulation. It is crucial to understand the strategic interplay between an engagement-maximizing platform and users who value not just the instantaneous joy coming from scrolling down and posting their thoughts, but also the reward from learning. This is precisely what this paper intends to achieve.

To do so, we build a theoretical model of communication and learning on a social media platform, describing and characterizing the algorithm a platform implements in

3 Social media platforms began transitioning from reverse-chronological feeds to personalized feeds at different times. Facebook started implementing personalized feeds in 2009, while Twitter (now X) and Instagram transitioned between 2015 and 2016. Younger platforms, like TikTok, have provided curated content since their launch.

equilibrium. We find that such algorithm exploits similarity too much, locking users in echo chambers. Moreover, learning disappears as platform size grows large. As this is far from ideal, we explore alternatives. As we could expect, the reverse-chronological algorithm the DSA mandated to reincorporate is not that proficient: in general, it cannot compete against the platform-optimal algorithm. In light of recent efforts from platforms to *give context* or *promote fact-checked content*, we build the "breaking echo chambers" algorithm, a modification of the platform-optimal algorithm that improves learning significantly as platform size grows large. Still, we look for a natural way for platforms to implement the optimal algorithm for the users, the utilitarian optimal algorithm. This is why we propose horizontal interoperability. Under horizontal interoperability, platforms are forced to compete on algorithms because, absent platform-specific network effects that capture users in the dominant site, the only way to retain the user base is for platforms to implement an algorithm that is preferred over those implemented by other competing platforms. This simple argument à la Bertrand leads to platforms opting for the utilitarian optimal algorithm.

We highlight four main contributions of this paper. First, we build a model where users post messages and learn through a feed designed by the engagement-maximizing platform. We assume that users derive instantaneous utility from engaging in communication with peers about some underlying topic, and we call this utility stream *within-the-platform utility*. It has three channels: satisfaction is brought by reading a post written by a friend, expressing one's own views (in the sense of being loyal to own innate opinions; *sincerity*), and conforming with the rest (in the sense of matching the opinions that others have shared; *conformity*).[4] The strength of these incentives depends on model parameters. In particular, we encompass situations in which conformity is almost negligible. The second utility stream comes from gathering valuable information on the platform to improve a decision, termed *action utility*.[5] The effectiveness of the Covid-19 vaccine, which triggered significant public debate,[6] is our leading example, as each part of the previously described utility function can be easily identified. People, driven by both a desire for sincerity and conformity, used social media to express their views about the benefits and risks of vacci-

4 Conformity is a driving-force in social media behavior (Mosleh et al., 2021). It is defined as the act of matching attitudes, beliefs and behaviors to group norms (Cialdini and Goldstein, 2004). Here we treat conformity as a behavioral bias included at the outset, but it has been widely found as a product of rational models. See Bernheim (1994) for a theory of conformity and Chamley (2004) for an overview.

5 The first component of the utility function is similar to the payoffs in Galeotti et al. (2021), where agents prefer taking actions closer to those of their neighbors and to their own ideal points. Utility is given by a weighted average of two loss functions representing *miscoordination* and *distance from favourite action*, and the action is not necessarily a message, as it is in our within-the-platform utility. However, one of their motivating examples perfectly fits our model: "the action may be declaring political opinions or values in a setting where it is costly to disagree with friends, but also costly to distort one's true position from the ideal point of sincere opinion".

6 Loomba et al. (2021) find that the acceptance of the Covid-19 vaccine in US and UK declined an average of 6 percentage points due to misinformation.

nation. Note that individuals sought to communicate their personal viewpoints because vaccination was a pivotal societal concern, but at the same time expressing dissenting opinions proved to be socially taxing. Additionally, gathering information was crucial to deciding whether to get vaccinated or not.

Engagement is defined, in this paper, as the number of posts a user reads. It is equivalent to the time spent scrolling down before logging out. Crucially, we assume that users do not rationally decide how much to engage, but that their engagement is controlled by a stochastic process driven by the instantaneuos joy of consuming content. After reading each post, a user continues scrolling down with some probability depending on the instantaneous within-the-platform utility derived. Otherwise, she logs out. Scrolling is, then, seen as an addictive behavior the user does not rationally control (Allcott et al., 2022): it is a rather automatic process corresponding with the intrinsic happiness derived within the platform.[7] However, the explicit decision to post a message is seen as a rational move in which the user consciously acts to achieve a goal. Users post messages and then observe those which appear in their feeds until they log out. Afterwards, they take an action based on the information gathered. Feeds are the product of an algorithm designed by the platform which, as explained earlier, has no incentives to promote learning: engagement purely depends on within-the-platform utility. The platform, which does not read messages, designs the algorithm leveraging its information on users' similarities in views. We assume the platform knows perfectly how similar users are, and utilizes this information to maximize its profits, i.e., to maximize total engagement. We think of similarities being derived from past interactions and users' personal data by using sophisticated machine learning techniques.[8]

Our second contribution is to identify the platform-optimal algorithm and study its properties. As expected, the platform-optimal algorithm is driven by the desire to maximize expected conformity, because it is the main force behind engagement. However, the fact that each user knows her own signal creates an information friction and the platform brings *too much* similarity to the feeds. The user would have preferred to have more diverse views, but is locked in an echo chamber, precisely as Eli shows in her video. This excess of similarity in the feeds becomes more pronounced as the platform size grows. Then, the feed becomes flooded with close *copies* of a user and consequently learning vanishes, contrasting with classical results where large societies learn better (Golub and Jackson, 2010).

7 This assumption could be also interpreted following the *Dual Process Theory* as in Benhabib and Bisin (2005). For an overview on Dual Process Theory, see Grayot (2020). The fact that engagement depends only on within-the-platform utility can be seen under the light of present bias: the user weights disproportionally low the benefits of learning when reading posts.

8 Facebook's FBLearner Flow, a machine learning platform, is able to predict user behavior through the use of personal information collected within the platform. See Biddle (2018) for a news piece on it. The early paper Kosinski et al. (2013) already showed that less sophisticated techniques could predict a wide range of personal attributes by just using data on "likes".

The third contribution consists in studying alternatives to the platform-optimal algorithm. We start with the reverse-chronological algorithm brought back by the DSA. This algorithm is generally not good enough to be considered a suitable alternative, so we analyze a variation of the platfom-optimal algorithm that maximizes learning when platform size is large. This is the breaking echo chambers algorithm, which adds a user with opposite views at the top of each feed induced by the platform-optimal algorithm. It improves learning but slightly decreases conformity and consequently engagement. While it still outperforms the platform-optimal algorithm for many types of users, it is plausible that real-world individuals may disregard information from a completely opposing source, complicating its practical implementation.

Regardless, the utilitarian optimal algorithm is the only one that maximizes social welfare, so then we must explore its implementation. This leads us to the fourth and last contribution of this paper, the discussion whether implementing horizontal interoperability would suffice for platforms to opt for the utilitarian optimal algorithm through competition. Without horizontal interoperability, the network effects that social media platforms feature (i.e., the fact that the more users join a platform, the more valuable its service becomes) create high barriers to entry and induce winner-takes-all (or most) market dynamics. Horizontal interoperability compels platforms to connect, so that users from different platforms can be linked. A user's feed would then be an ordered list of the posts coming from all her friends, regardless of which platform they are registered on, designed by the platform she joined. Crucially, network effects will be shared, and platforms will have to compete along the non-interoperable dimension, i.e., they will have to compete in algorithms. Each user will join the platform whose algorithm offers the highest expected utility, disregarding platform size. And this algorithm is, of course, the utilitarian optimal algorithm. Then, competing platforms would be *forced* to implement this algorithm; otherwise, they risk losing their user base. The pursuit of this goal aligns with the intentions of EU regulators, as reflected in the Digital Markets Act,[9] which mandates certain large social platforms to achieve interoperability in their messaging communications in the immediate future. Quoting Kades and Scott Morton (2020): "Interoperability eliminates or lowers the entry barrier, which is the anticompetitive advantage the platform has maintained and exploited. Users will not switch to a new social network until their friends and families have switched. [...] Interoperability causes network effects to occur at the market level—where they are available to nascent and potential competitors—instead of the firm level where they only advantage the incumbent."

The rest of the paper is organized as follows. After the literature review, each section corresponds to each of the contributions described above: Section 2 develops the model,

9 See regulation (EU) 2022/1925 of the European Parliament and of the council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828.

Section 3 finds the platform-optimal algorithm and characterizes it, Section 4 analyzes alternative algorithms, and Section 5 discusses horizontal interoperability. Finally, Section 6 concludes.

1.1 Related literature

The effects of personalized feeds on social welfare have not, to the best of our knowledge, been studied from a theoretical perspective. However, a recent paper by Guess et al. (2023) examines the empirical effects of Facebook's and Instagram's feed algorithms. The study reveals that transitioning users back to chronological feeds decreases the time they spend on the platforms as well as their overall activity (i.e., engagement). Additionally, it leads to a reduction in the proportion of content derived from ideologically like-minded sources, thereby diminishing the impact of the echo-chamber effect.

In broad terms, our paper is related to two areas of literature. The first area studies the impact of revenue-maximizing platforms on social learning. This is a growing field, and we highlight two papers for their similarities to our work. Mueller-Frank et al. (2022) build a model of network communication and advertising where the platform controls the flow of information. In equilibrium, the platform may manipulate or even suppress information to increase revenue, even though this ultimately decreases social welfare. In a model where agents decide whether or not to pass on (mis)information, Acemoglu et al. (2023) study the algorithm choice of an engagement-maximizing platform. They show that when the platform has the ability to shape the network, it will design algorithms that create more homophilic communication patterns. Thus, in line with our results, both papers find that platforms' incentives are not aligned with users' preferences and that engagement-maximizing behavior harms social welfare. Homophilic communication patterns, commonly known as echo chambers or "filter bubbles", also appear in Pariser (2011): to increase metrics like engagement and ad revenue, recommendation systems connect users with information already similar to their current beliefs. This hypothesis is further discussed in Sunstein (2017), while Chitra and Musco (2020) experimentally analyze the effects of filter bubbles on polarization and show the large impact of minor algorithm changes. Relatedly, Demange (2023) shows that platforms promote the visibility of their most influential individuals. Additional research on media platforms providing distorted content for economic reasons can be found in Reuter and Zitzewitz (2006), Ellman and Germano (2009), Abreu and Jeon (2019), and Kranton and McAdams (2022). Hu et al. (2021) shows that rational, inattentive users prefer to learn from like-minded neighbors, while Törnberg (2018) shows that echo chambers harm social welfare by increasing the spread of misinformation.

Not just the mentioned literature, but also empirical work (Sagioglou and Greitemeyer, 2014; Levy, 2021) reveals the need for further intervention or regulation on social media platforms. This topic constitutes the second strand to which our paper is closely

related. Franck and Peitz (2023) study competition between social media platforms, claiming that market power (mainly represented by the network effects) leads to sub-optimal outcomes for society. In particular, it may not be the platform with the best offer that dominates the market. Biglaiser et al. (2022) offer a micro-foundation for incumbent advantage. Essentially, network effects prevent users from migrating to even Pareto-superior equilibria when they receive stochastic opportunities to migrate to an entrant. Kades and Scott Morton (2020) also examine network effects in digital platforms and offer an overview of interoperability. Popiel (2020) and Evens et al. (2020) assert that regulations to manage digital platform markets in the US and EU, respectively, are inadequate in addressing their negative effects. In response to this need, there has been a surge of recent papers examining interventions. Regarding structural interventions, Jackson et al. (2022) examine how limiting the breadth and/or depth of a social network improves message accuracy. The work of Benzell and Collis (2022) aligns with our own, as they analyze the optimal strategy of a monopolistic social media platform and evaluate the impact of taxation and regulatory policies on both platform profits and social welfare. However, in their paper, the platform chooses net revenue per user rather than shaping communication among users. The authors apply their model to Facebook and find that a successful regulatory intervention to achieve perfect competition would increase social welfare by 4.8%. Finally, Agarwal et al. (2022) provide empirical evidence of the negative consequences of deplatforming (shutting down a community on a platform), mainly due to migration effects, which supports a call for globally applicable regulations.

There is a plethora of recent empirical contributions regarding informational interventions: Habib et al. (2019), Hwang and Lee (2021) or Mudambi and Viswanathan (2022). Mostagir and Siderius (2023b) model community formation and show that the effect of interventions is non-monotonic over time. Additionally, there is another important aspect to consider when analyzing informational policies: Mostagir and Siderius (2022) demonstrate that cognitive sophistication matters when faced with misinformation, and Mostagir and Siderius (2023a) find that different populations (Bayesian and DeGrootians) react differently to certain interventions. While some papers, such as Mostagir and Siderius (2023a), include cases where sophisticated users are outperformed by their naive counterparts, Pennycook and Rand (2019, 2021) show that higher cognitive ability is associated with better ability to discern fake content. In our model, the results hold for both Bayesian and DeGrootian users, but the sophisticated agents always learn better. Finally, we also relate to the literature on learning in networks, for both naive and sophisticated users: DeMarzo et al. (2003), Acemoglu and Ozdaglar (2011), Jadbabaie et al. (2012), Molavi et al. (2018) or Mueller-Frank and Neri (2021).

2 A model of communication and learning through personalized feeds

Here we present the baseline model of the paper. We start by providing an overview, then delve into the formal details, and finally discuss some of the assumptions made along the way.

There is an underlying state of the world that users aim to discover in view of a subsequent action. Joining a social media platform offers users the benefit of accessing information, as fellow users share messages related to that state of the world. However, beyond mere information retrieval, users also derive utility from engaging in non-informative interactions within the platform. Expressing personal opinions and reading others' posts brings satisfaction, yet encountering disagreement imposes a burden. We define user engagement as the measure of messages read, representing the time spent on the platform until the user discontinues browsing and exits.

Users' utility comprises two components: the *within-the-platform utility*, influenced by engagement, conformity, and sincerity, and the *action utility*, which depends on learning, i.e., how close users can get to the state of the world after communicating on the platform. The platform's revenues, in turn, are contingent upon user engagement. Hence, the platform designs an algorithm seeking to maximize such engagement by leveraging information on similarities between users' worldviews. This algorithm curates a personalized feed for each user, determining the order in which messages appear on the scrolling screen.

In our baseline model, we assume a monopolistic platform with all users already on board. Once a user logs in, she decides on which message to post. Engagement, however, is not the product of a rational decision but follows an addictive process: after reading each message, with some probability depending on the amount of within-the-platform utility experienced so far, the user continues scrolling down, while she logs out otherwise.

Now, let us describe the model in detail. There is a set, $\mathcal{U}$, of n users aboard a social media platform. We assume that every user is a friend of all others, and hence her neighborhood is the whole user base (in network terms, we are working with the complete network). Users receive information on the state of the world θ in the form of a private signal $\theta_i \in \mathbb{R}$. Conditional on θ, signals $\{\theta_1, ..., \theta_n\}$ are jointly normal and their structure is given by

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\theta} = (\theta, ..., \theta)$ and $\boldsymbol{\Sigma} = (\sigma_{ij})$ is an $n \times n$ symmetric and positive definite matrix. The signal θ_i is interpreted as the information the user has about the state of the world prior to her entry on the social platform. It might be based on inherent personal characteristics as well as on information collected privately. As information sources, as well as ideology, might be similar, different users' private information might be correlated. This is captured by the matrix $\boldsymbol{\Sigma}$.

Users know their private signals, the distribution of all signals, the covariance matrix $\boldsymbol{\Sigma}$, and the distribution of the state of the world, for which we crucially assume improper priors.[10] Thus, conditional on θ_i, the posterior distributions of θ_j and θ are normal and centered on θ_i, namely $\theta_j|\theta_i \sim \mathcal{N}\left(\theta_i, \sigma_{jj} - \frac{\sigma_{ij}^2}{\sigma_{ii}}\right)$ for all $j \in N$ and $\theta|\theta_i \sim \mathcal{N}(\theta_i, \sigma_{ii})$.

Once logged in the platform, each user i posts a message $m_i \in \mathbb{R}$ and then observes $e_i \in \mathbb{N}$ messages that appear in her personalized feed, which is provided by the platform. The number $e_i \leq n$ represents her engagement, and platform profits depend precisely on the sum of all users' engagement, $\sum_{i=1}^{n} e_i$. In order to maximize user engagement, the platform designs an algorithm consisting of an assigment that, given a pair of users i, j, tells which position user j occupies in user i's feed. Given engagement e_i, the feed is the set of users from whom messages will be observed. Formally, an algorithm $\mathscr{F}$ is a collection $(\mathscr{F}_i)_{i \in \mathcal{U}}$ where $\mathscr{F}_i \in \text{Bij}(\{1, ..., n-1\}, \mathcal{U} \backslash \{i\})$. Given $k \leq n$, $\mathscr{F}_i(k)$ is the k-th user in i's ranking induced by $\mathscr{F}$, so i's feed for engagement e_i is precisely

$$\mathscr{F}_i^{e_i} = \{\mathscr{F}_i(1), ..., \mathscr{F}_i(e_i)\}.$$

Users derive utility from two streams: within-the-platform utility and action utility. Their within-the-platform utility has three components: (i) a positive linear payoff coming from reading messages; (ii) *sincerity*: agents dislike deviating from their own signals,[11] and (iii) *conformity*: disagreeing with others' opinions is taxing. Formally, user i's realized within-the-platform utility is

$$u_i(e_i, m_i, m_{-i}, \mathscr{F}_i, \theta_i) = \alpha e_i - \beta \overbrace{(\theta_i - m_i)^2}^{\text{Sincerity}} - (1-\beta) \overbrace{\sum_{j \in \mathscr{F}_i^{e_i}} \frac{(m_i - m_j)^2}{e_i}}^{\text{Conformity}}, \tag{1}$$

where $\alpha > 0$ and $\beta \in (0, 1)$ represents how much sincerity is weighted with respect to conformity. Within-the-platform utility is not the only source of utility for users, as they are also concerned about taking an action $a_i \in \mathbb{R}$ that matches the state of the world. Total realized utility is the weighted average of within-the-platform utility and action utility (the squared distance of the action from the state of the world):

10 For a discussion of improper priors, see Hartigan (1983).

11 Due to improper priors, sincerity would yield the same results if, instead of being punished for deviating with her message m_i from θ_i, the user were penalized for deviating from θ.

$$U_i(e_i, m_i, m_{-i}, a_i, \mathscr{F}_i, \theta_i, \theta) = \lambda\, u_i(e_i, m_i, m_{-i}, \mathscr{F}_i, \theta_i) - (1-\lambda)\overbrace{(a_i-\theta)^2}^{\text{Action utility}}, \qquad (2)$$

where $\lambda \in (0,1)$ weights the relative importance of within-the-platform and action utilities. Summarizing, user i observes θ_i, chooses a message m_i and, after learning messages $\{m_j\}_{j\in\mathscr{F}_i^{e_i}}$, chooses an action a_i to maximize the conditional expectation of U_i.

Along the lines of digital addiction theory, we assume that the user does not rationally control her scrolling time but, after reading k posts, reads the next message with probability $g(u_i(k-1, m_i, m_{-i}, \mathscr{F}_i, \theta_i))$, where $g : \mathbb{R} \to [0,1]$ is some continuous and increasing function. With probability $1 - g(u_i(k-1, m_i, m_{-i}, \mathscr{F}_i, \theta_i))$, the user discontinues scrolling down and exits. Hence, user i features engagement e_i with probability $(1 - g(u_i(e_i, m_i, m_{-i}, \mathscr{F}, \theta_i)))\prod_{r=1}^{e_i-1} g(u_i(r, m_i, m_{-i}, \mathscr{F}_i, \theta_i))$. In particular, we assume that $\forall\, x \in \mathbb{R}$, $g(x) \in (0,1)$, i.e., no feed guarantees either continuation or abandonement. Note that because of the addictive nature of e_i, the user sees it as something exogenous and given.

The platform knows the distributions and $\boldsymbol{\Sigma}$, but not θ nor $\{\theta_i\}_{i=1}^n$. It builds the algorithm $\mathscr{F}$ based on $\boldsymbol{\Sigma}$ to maximize $\sum_{i=1}^n \mathbb{E}_p[e_i]$ (where $\mathbb{E}_p$ stands for the platform's expectations), the sum of the expected engagement of all users. In summary, the game of *communication and learning through personalized feeds* described above consists of the following sequence of events:

1. The platform chooses an algorithm $\mathscr{F}$ and (publicly) commits to it.
2. Each user observes her private signal θ_i.
3. Each user i posts a message $m_i \in \mathbb{R}$.
4. Each user i observes e_i messages in her feed $\mathscr{F}_i^{e_i}$ and chooses an action a_i.
5. The state of the world is revealed and payoffs are realized.

We devote the last part of this section to a discussion on some of the assumptions that build the model:

Complete network. This model could be extended to any network given by some undirected graph $\mathcal{G}$. In such a case, each user i belongs to a neighborhood n_i and hence $e_i \leq |n_i|$. All the results presented below hold. Thus, we prefer to work with the complete network for ease of notation and exposition.

Monopolistic platform, all users on board. In this baseline model, we assume there is only one platform, and all users are already on board. Hence, the platform does not need to care about capturing users, but only about their engagement. This is, of course, a simplifying assumption, but the main social media platforms (Facebook, Instagram, TikTok or X) are monopolists of their fields:[12] even though they can be broadly

12 Regarding monopoly structures in the social media platform market, the Bundeskartellamt (the German competition protection authority) states in its case against Facebook (B6-22/16, "Facebook",

described as social media platforms that enable public posting and private communication, they differ in their core functionality. Each site dominates a specific field: photography (Instagram), short videos (TikTok), reciprocal communication with friends (Facebook), and micro-blogging (X). In most cases, there is no realistic alternative for the average user but to stay out, and then, as Bursztyn et al. (2023) show, fear of missing out makes users join even when they would prefer the platform not to exist.[13]

Improper priors. Users' prior distribution is uniform along $\mathbb{R}$. Intuitively, this means that none of them understands whether her signal is extreme. Indeed, every user believes her opinion is central (Greene, 2004). This assumption is made for the sake of model tractability. Under normal priors, we can only determine the users' optimal linear messaging strategies, but we cannot derive an explicit expression for the platform-optimal algorithm.

Non-rational engagement. Following the literature on digital addiction (Allcott et al., 2022), we dismiss a rational framework for engagement, and opt for a simplified setting in which digital addiction is captured as a by-product of habit formation and self-control problems. The user irrationally continues scrolling down depending on the instantaneous within-the-platform utility experienced so far. Our configuration encapsulates addictive behavior in a reduced form, capturing some essential features: the probability of engaging for k periods is always higher than the probability of engaging for k' periods if $k < k'$, a higher utility derived from reading a message implies a greater probability of staying, and engagement does not depend on action utility. This last feature could be intuitively conceptualized as an extreme form of present bias: when scrolling down, the user heavily discounts the long-run reward from learning (Guriev et al., 2023). This is also in line with the main case in Bonatti and Cisternas (2020), where consumers ignore the link between their current actions and the future consequences.

Platform's profits as a function of total engagement. Social media platforms are generally free to access, and their revenues come from advertisers' payments for product placement. These payments depend on user engagement: the larger the engagement, and hence the greater the exposure to their content, the more an advertiser is willing to pay. For simplicity, in this model the platform objective is to maximize total engagement, so its profit function is $\Pi_p(\mathscr{F}, \Sigma) = \sum_{i=1}^{n} \mathbb{E}_p[e_i]$.

p. 6): "The facts that competitors are exiting the market and there is a downward trend in the user-based market shares of remaining competitors indicate a market tipping process that will result in Facebook becoming a monopolist." (Franck and Peitz, 2023).

13 As already commented above, Bursztyn et al. (2023) show that users would be willing to pay to have others, including themselves, deactivating their TikTok and Instagram accounts.

3 Platform-optimal algorithm

In this section, we obtain and characterize the algorithm the platform implements in equilibrium. First, we show that users find it optimal to report their private signals truthfully. The platform, in turn, designs a feed for each user that is excessively driven by similarity. Intended to maximize engagement, such a feed worsens user learning as the population grows until it asymptotically vanishes.

The equilibrium concept is Bayesian Nash Equilibrium (BNE). The platform chooses an algorithm $\mathscr{F}$, while each user chooses a message m_i to maximize

$$\mathbb{E}_i[U_i|\theta_i,\mathscr{F}] = \lambda\left(v(e_i) - \beta(\theta_i - m_i)^2 - (1-\beta)\mathbb{E}_i\left[\sum_{j\in\mathscr{F}_i^{e_i}}\frac{(m_i - m_j(\theta_j))^2}{e_i}|\theta_i,\mathscr{F}\right]\right) - (1-\lambda)\mathbb{E}_i[(a_i-\theta)^2|\theta_i,\mathscr{F}],$$

and an action a_i (after learning the messages in her feed) to maximize

$$-(1-\lambda)\mathbb{E}_i\left[(a_i-\theta)^2|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right].$$

In this framework, for any algorithm the platform picks, users disclose their private signals in their messages.

Proposition 3.1. *Given any algorithm $\mathscr{F}$, every user plays truthtelling in equilibrium, i.e., $m_i^* = \theta_i$ for all $i \in \mathcal{U}$.*

Proof. See Appendix A. □

Because of improper priors, the platform cannot affect first-order moments through the feed it designs, and hence, user i believes that, in expected terms, every other user will play θ_i. Deviating from truthtelling is then not profitable.

Having shown that users play truthtelling in equilibrium, we derive the platform-optimal algorithm, denoted by $\mathscr{P}$. First, we show that maximizing profits, or total expected engagement $\sum_{i=1}^n \mathbb{E}_p[e_i]$, is equivalent to maximizing each user's expected engagement $\mathbb{E}_p[e_i]$.

Lemma 3.2. *It is equivalent for the platform to maximize total user engagement and maximizing each user's individual engagement separately.*

Proof. We know from Proposition 3.1 that, given user i, engagement e_i and a feed $\mathscr{F}_i^{e_i}$, user i plays $m_i = \theta_i$ in equilibrium: messages are not affected by the feed. Hence, there are no interdependencies across feeds: the order in which the platform ranks user j in i's feed does not affect anyone else. Finally, user i's expected engagement is a function of her expected within-the-platform utility, affected by the truthful message θ_i and her feed $\mathscr{F}_i^{e_i}$. Hence, maximizing the sum of all users' expected engagement is equivalent to maximizing each of them individually. □

Now, we intuitively explain how the platform designs the platform-optimal algorithm $\mathscr{P}$, and what the optimal action a_i^* taken by user i is after reading her feed $\mathscr{P}_i^{e_i}$. The formal details are left to the Appendix A as part of the proof of Proposition 3 below. From the point of view of the platform, and because of truthful reporting, user i's within-the-platform utility simplifies to $\mathbb{E}_p[u_i(k-1,\theta_i,\theta_{-i},\mathscr{F},\theta_i)] = \mathbb{E}_p[(v(k)-(1-\beta)\frac{1}{k}\sum_{j\in\mathscr{F}_i^{k-1}}(\theta_i-\theta_j)^2]$ when she has read $k-1$ messages. The probability of staying after reading the k-th message is then $\mathbb{E}_p[g(u_i(k,\theta_i,\theta_{-i},\mathscr{F},\theta_i))]$. To maximize this probability, the platform chooses a user j to be included next in the feed among those who have not been chosen yet, i.e., $j \in \mathcal{U}\backslash\mathscr{F}_i^k$. As g is increasing in u_i, maximizing g is equivalent to maximizing u_i. Moreover, note that conformity is the only term in which the platform can affect user i's within-the-platform utility at this stage. Hence, j is chosen according to

$$j = \underset{j\in\mathcal{U}\backslash\mathscr{F}_i^k}{\operatorname{argmax}}\{-\mathbb{E}_p[(\theta_i-\theta_j)^2]\},$$

and j is the user whose message has not yet been shown and minimizes the loss coming from conformity. The platform-optimal algorithm $\mathscr{P}$ is precisely the one which, when applied to user i, ranks other users in reverse order regarding their loss in conformity with her. In other words, for any $k \leq n$, the feed $\mathscr{P}_i^k$ shows the messages of the k users who conform the most with her. This happens, crucially, from the perspective of the platform, which is unaware of the particular realizations of the users' private signals. In short, the algorithm $\mathscr{P}$ applied to user i induces a feed given by:

$$\begin{aligned}
\mathscr{P}_i^1 &= \underset{j\in N}{\operatorname{argmax}}\{-\mathbb{E}_p[(\theta_i-\theta_j)^2]\},\\
\mathscr{P}_i^2 &= \mathscr{P}_i^1 \cup \underset{j\in N\backslash\mathscr{P}_i^1}{\operatorname{argmax}}\{-\mathbb{E}_p[(\theta_i-\theta_j)^2]\},\\
&\vdots \qquad\qquad \vdots\\
\mathscr{P}_i^k &= \mathscr{P}_i^{k-1} \cup \underset{j\in N\backslash\mathscr{P}_i^{k-1}}{\operatorname{argmax}}\{-\mathbb{E}_p[(\theta_i-\theta_j)^2]\}.
\end{aligned} \tag{3}$$

For an explicit example of how the platform designs $\mathscr{P}$ leveraging $\mathbf{\Sigma}$, please refer to Appendix B.

Proposition 3.3. *In equilibrium, the platform chooses the algorithm $\mathscr{P}$ as specified in Equation (3). In other words, the algorithm that maximizes user engagement is the one that, for each user i, designs a feed in which others appear in reverse order regarding the expected loss in conformity with user i they induce.*

Proof. The formal derivation of $\mathscr{P}$ can be found in Appendix A. □

The information friction between the platform and the users is crucial. On the one hand, the platform chooses the feed so as to maximize the loss in conformity, which effectively means maximizing $-\mathbb{E}_p[(\theta_i-\theta_j)^2]$ through the choice of j. But $-\mathbb{E}_p[(\theta_i-\theta_j)^2] =$

$-\sigma_{ii} - \sigma_{jj} + 2\sigma_{ij}$, so

$$j = \underset{j \in \mathcal{U}/\mathscr{F}_i^k}{\operatorname{argmax}} \{-\sigma_{jj} + 2\sigma_{ij}\}.$$

On the other hand, from the user's perspective, expected conformity is $-\mathbb{E}_i[(\theta_i - \theta_j)^2|\theta_i] = -\sigma_{jj} + \frac{\sigma_{ij}^2}{\sigma_{ii}}$. User i's knowledge of θ_i notably changes the expression compared to that of the platform, and we observe that, given σ_{jj}, user i would prefer to be matched with some $j \in \mathcal{U}$ either very similar or very opposite to her. As the platform is less informed, it only selects very similar users to user i and, on top of that, fixes the weight of similarity to 2, when the user would prefer it to depend on $\frac{1}{\sigma_{ii}}$. All this drives the user to an *excessive* similarity bubble, while she would prefer to observe a more diverse feed.

Proposition 3.4. *The platform excessively weights similarity between users when designing its optimal algorithm.*

Proof. As indicated above, the platform selects the next user j in the feed $\mathscr{F}_i^k$ according to $j = \underset{j \in \mathcal{U}/\mathscr{F}_i^k}{\operatorname{argmax}} \{-\sigma_{jj} + 2\sigma_{ij}\}$, while user i would prefer j to be selected according to

$$j = \underset{j \in \mathcal{U}/\mathscr{F}_i^k}{\operatorname{argmax}} \left\{-\sigma_{jj} + \frac{\sigma_{ij}^2}{\sigma_{ii}}\right\}.$$

□

When variances are homogeneous, meaning each user has an equally precise posterior of θ, the feed becomes a reverse ranking based on similarities. Then, users are displayed according to how correlated they are to the one reading the feed.

Corollary 3.5. *If variances are homogeneous, i.e., $\sigma_{ii} = \sigma_{jj} = \sigma^2$ for all $i, j \in \mathcal{U}$, the platform-optimal algorithm $\mathscr{P}$ ranks uniquely in terms of similarity.*

Proof. If variances are homogeneous, $-\mathbb{E}_p[(\theta_i - \theta_j)^2] = -2\sigma^2 + 2\sigma_{ij}$, and users are ranked following a weakly decreasing order regarding their covariance to user i (if there were ties, they would be broken randomly). Hence, $\mathscr{P}_i^1$ is the first user in the ranking, $\mathscr{P}_i^2$ is the second, and so on. □

Note that the user's expected conformity is $-\mathbb{E}_i[(\theta_i - \theta_j)^2|\theta_i] = -\sigma^2 + \frac{\sigma_{ij}^2}{\sigma^2}$ in this particular case, and the main interpretation of the difference between what the platform maximizes and what the user would like to be maximized remains the same. Crucially, the way messages are displayed in the feed influences users actions. The next result provides a formal expression for user i's optimal action a_i^*.

Proposition 3.6. *User i's optimal action after reading e_i messages, for any algorithm $\mathscr{F}$ is*

$$a_i^* = \frac{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}^t}{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}^{-1}\mathbb{1}^t},$$

where $\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}$ *is the restriction of* $\boldsymbol{\Sigma}$ *to the users in* $\mathscr{F}_i^{e_i}$ *and* $\boldsymbol{\theta}^t_{\mathscr{F}_i^{e_i}}$ *is the vector of private signals of the users in* $\mathscr{F}_i^{e_i}$.

Proof. User i's optimal action maximizes $\mathbb{E}_i[(a_i - \theta)^2]$ given the observed messages $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$. Hence, the optimal action is

$$a_i^* = \mathbb{E}_i[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}] = \frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}^t_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}$$

by Lemma A.1. □

To understand the impact of the platform-optimal algorithm on social welfare, we must examine its effect on *learning*, which refers to how information gathered on the platform improves decision-making. Before doing so, we make an additional assumption for tractability purposes. We assume that users' variances are homogeneous, i.e., that $\sigma_{ii} = \sigma_{jj}$ for all users i, $j \in \mathcal{U}$. Thus, we are in the case described by Corollary 3.5. Note that the disparity between what users prefer to observe in their feed and what the platform provides is maintained. To clearly indicate that we are now working under homogeneous variances, we will denote the platform-optimal algorithm as $\mathscr{C}$, referring to the "closest" algorithm, as now the platform simply matches users with those who are most similar, or closest, to them.

Now, let us analyze how personalization algorithms affects information gathering. The scenario we study is precisely that of a large platform size and high engagement values, reflecting the substantial growth in social media usage in recent years, both in terms of the number of users and the time spent on platforms.[14] Learning is defined as the increase in expected action utility resulting from reading messages. When a user picks the optimal action, its expected value is the posterior variance of θ conditional on the messages in the feed:

$$\mathbb{E}\left[(a_i - \theta)^2|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right] = \mathbb{E}\left[(\mathbb{E}_i[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}] - \theta)^2|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right] = \mathrm{Var}\left[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right].$$

Applying Lemma A.1, we explicitely obtain the posterior variance:

$$\mathrm{Var}\left[\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right] = \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}.$$

This expression allows us to calculate the improvement in decision-making after reading a feed for any algorithm and any society characterized by $\mathcal{U}$ and $\boldsymbol{\Sigma}$. Note that, in this model, the posterior variance is weakly lower than σ_{ii} for each user i, meaning users cannot be worse off in terms of learning after reading their feeds. We let the platform grow large now, expanding a given user base $\mathcal{U}$ by assuming that the covariances between

14 See, for example, the number of social media users from 2011 to 2028 (forecasted) https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

new users and existing users are drawn from a continuous distribution with a cumulative distribution function supported in $[-\sigma^2, \sigma^2]$ and centered at 0. The resulting covariance matrix (the expanded $\mathbf{\Sigma}$) is symmetric and positive definite.

The manner in which $\mathscr{C}$ selects the feed becomes very apparent when platform size grows large. With a vast pool of users, conformity is maximized by choosing someone almost identical to the user. This creates a feed of close copies, resulting in an echo chamber where learning diminishes. Note that this is not what the user desires: she would prefer matches with very similar or very different individuals, which would additionally increase learning. However, the next result formally shows that, asymptotically, the platform-optimal algorithm induces no learning. This is independent of the specific engagement level of the user.

Proposition 3.7. *Under the closest algorithm $\mathscr{C}$, and, for any engagement level e_i, user i's learning becomes negligible as $n \to \infty$:*

$$\lim_{n\to\infty} \mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}_i^{e_i}}] = \sigma^2.$$

Proof. See Appendix A. □

This finding is in stark contrast to classic learning models where the wisdom of the crowd enhances learning as the population grows. Here, the platform's strategic role in feed selection undermines learning, making it vanish. In conclusion, the optimal algorithm for the platform not only creates excessive echo chambers but also harms long-term learning in large populations. These issues are significant in public debate, raising concerns about the impact of social media platforms on social welfare. The approval of the DSA and DMA in the European Union addresses these concerns. In particular, the DSA forces platforms to include the non-strategic reverse-chronological algorithm that was used before personalization algorithms as an option for users. The next section is devoted to an analysis of alternative algorithms, including the already mentioned reverse-chronological algorithm, the user-optimal algorithm, and the breaking-echo-chambers algorithm.

4 The reverse-chronological algorithm and other alternatives

The reverse-chronological algorithm, which will be denoted by $\mathscr{R}$, displays friends' posts in the (reverse) order they were written. Before the implementation of personalization algorithms, every social media platform relied on this simple method of presenting the feed, which is not strategic at all. In this model, we understand the reverse-chronological algorithm as a *random* algorithm in which a post will be at the top of the feed with probability $\frac{1}{n-1}$. Consequently, this algorithm does not create echo chambers, and, as we show below, when the platform size is large, it outperforms the platform-optimal algorithm

in terms of learning, but not in terms of conformity. The effect on overall utility depends on how users weight sincerity, conformity and learning. For small populations, however, it is not even the case that individuals learn better under the reverse-chronological algorithm, so we are far from stating unambiguously that the reverse-chronological algorithm is a feasible substitute for the platform-optimal algorithm, which motivates our search for a better alternative.

Given that the closest algorithm, $\mathscr{C}$, shows a feed of likeminded users when platform size grows large, learning vanishes (Proposition 3.7). The random nature of $\mathscr{R}$ yields better learning asymptotically:

$$\lim_{n\to\infty} \text{Var}[\theta|\boldsymbol{\theta}_{\mathscr{R}_i^{e_i}}] = \frac{\sigma^2}{e_i}.$$

This is, of course, not surprising. However, note that the higher the engagement (the e_i), the better for learning, but that $\mathscr{R}$ yields lower engagement than $\mathscr{C}$ because it is worse for conformity. This trade-off arises when we compare the expected utility under both algorithms.

Proposition 4.1. *Given $\boldsymbol{\Sigma}$, the closest algorithm outperforms the reverse-chronological algorithm in large populations if and only if*

$$\lambda > \max_{i\in\mathcal{U}} \left\{ \frac{1}{1+\left(\frac{\mathbb{E}_i[e_i^{\mathscr{R}}]\alpha(\mathbb{E}_i[e_i^{\mathscr{C}}-e_i^{\mathscr{R}}])+(1-\beta)\sigma^2\sum_{j\in\mathscr{R}_i^{\mathbb{E}_i\left[e_i^{\mathscr{R}}\right]}}(1-\rho_{ij}^2)}{\sigma^2(\mathbb{E}_i[e_i^{\mathscr{R}}]-1)}\right)} \right\}.$$

For a general $\boldsymbol{\Sigma}$ and assuming the expected correlation between every pair of users i and j is zero, i.e., $\mathbb{E}[\rho_{ij}] = 0$ for all i, j, the condition is given by:

$$\lambda > \frac{1}{1+\left(\frac{\mathbb{E}_i[e_i^{\mathscr{R}}](\alpha(\mathbb{E}_i[e_i^{\mathscr{C}}-e_i^{\mathscr{R}}])+(1-\beta)\sigma^2(1-\text{Var}[\rho_{ij}]))}{\sigma^2(\mathbb{E}_i[e_i^{\mathscr{R}}]-1)}\right)}.$$

Moreover, the closest algorithm is always worse than the reverse-chronological algorithm in terms of learning.

Proof. The second part of the proposition was already shown above. Regarding the first result, we just compare expected utilities for the user when n grows large. For each user

i, they are given, respectively, by:

$$\lim_{n\to\infty} \mathbb{E}_i\left[U_i(\mathscr{C})\right] = \lambda\,\alpha\,\mathbb{E}_i[e_i^{\mathscr{C}}] - (1-\lambda)\sigma^2;$$

$$\lim_{n\to\infty} \mathbb{E}_i\left[U_i(\mathscr{R})\right] = \lambda\left(\alpha\mathbb{E}_i[e_i^{\mathscr{R}}] - (1-\beta)\frac{\sigma^2}{\mathbb{E}_i[e_i^{\mathscr{R}}]}\sum_{j\in\mathscr{R}_i^{\mathbb{E}_i\left[e_i^{\mathscr{R}}\right]}}(1-\rho_{ij}^2)\right) - (1-\lambda)\frac{\sigma^2}{\mathbb{E}_i[e_i^{\mathscr{R}}]}.$$

□

Crucially, the closest algorithm maximizes user engagement, so in expectation user i reads more posts and derives higher intrinsic utility from doing so: $\alpha\,\mathbb{E}_i[e_i^{\mathscr{C}} - e_i^{\mathscr{R}}] > 0$. Hence, both terms in the numerator of

$$\frac{\mathbb{E}_i[e_i^{\mathscr{R}}]\alpha\,\mathbb{E}_i[e_i^{\mathscr{C}} - e_i^{\mathscr{R}}] + (1-\beta)\sigma^2\sum_{j\in\mathscr{R}_i^{\mathbb{E}_i\left[e_i^{\mathscr{R}}\right]}}(1-\rho_{ij}^2)}{\sigma^2(\mathbb{E}_i[e_i^{\mathscr{R}}]-1)}$$

are positive. It is likely, then, that for many specifications of the model, the closest algorithm dominates the reverse-chronological algorithms.

Less intuitive is the case of small platform size, in which the comparison between both algorithms is more complicated. Of course, within-the-platform utility is always better under the closest algorithm, but we cannot state unambiguously which of the two algorithms is better for learning. Two effects must be taken into account when it comes to the latter. First, when feeds are of small length, even if engagement is the same under $\mathscr{C}$ and $\mathscr{R}$ (which, in general, will not be the case), we cannot state in general that learning is worse under $\mathscr{C}$. The next example illustrates this.

Consider a tiny network composed of four individuals ($n = 4$), and assume, for this exercise, that the same feed length is the same for both algorithms, $k = 3$ (this will not happen in general, as the closest algorithm will provide a longer feed, but we want to show that even in this case the reverse-chronological algorithm does not guarantee better learning). Assume that the distribution of signals, conditional on θ, is as follows:

$$\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}\right); \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 & 0.7 & 0.5\\ 0.8 & 1 & 0.3 & 0.6\\ 0.7 & 0.3 & 1 & 0.4\\ 0.5 & 0.6 & 0.4 & 1\end{pmatrix}.$$

The closest algorithm induces, for user 1, a feed given by $\mathscr{C}_1^k = \{1,2,3\}$. Assume that a particular realization of the reverse-chronological algorithm induces the feed $\mathscr{R}_1^k = \{1,3,4\}$. Posterior variances are $\text{Var}[\theta|\{\theta_1,\theta_2,\theta_3\}] = 0.58$ for the closest algorithm and $\text{Var}[\theta|\{\theta_1,\theta_3,\theta_4\}] = 0.68$ for the reverse-chronological algorithm. Surprisingly, $\mathscr{C}$ yields better learning. The covariance to user 1 is not the unique driving force; the correlations

between other users in the feed also play a role. In fact, the tension between different forces provokes that learning is not monotonic under the closest algorithm, as we can observe in Figure 1. However, when platform size grows, the similarities to user 1 become the dominant factor, and as a consequence, the closest algorithm performs worse than the reverse-chronological algorithm. This can also be observed in Figure 1, where we plot realizations of learning under $\mathscr{R}$ and $\mathscr{C}$ as n increases for a population growing from $n = 30$ to $n = 5000$, constant engagement $k = 30$ and parameters $\lambda = 0.5$ and $\beta = 0.2$.

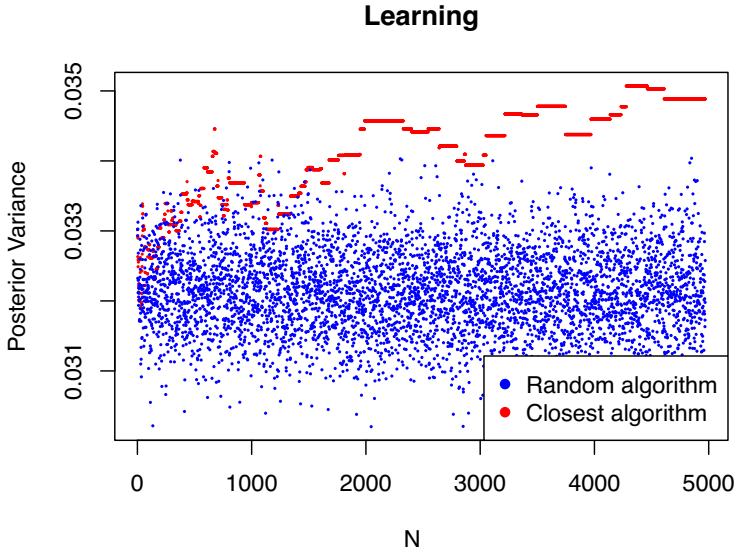

Figure 1: User's posterior variance as population grows; engagement is fixed to $k = 30$.

The second effect regarding learning when the user base is small relies on the fact that $\mathscr{C}$ induces higher engagement. Reading more posts is weakly better in learning terms, so even if we might intuitively think that learning is worse under $\mathscr{C}$ because like-minded individuals are brought to the feed, this might be counterbalanced by the greater number of messages to learn from. The example above illustrate this case, too. Assume, as the simplest scenario, that $\mathscr{C}$ induces engagement $e_1^{\mathscr{C}} = 2$ for user 1 and hence shows the posts of users 2 and 3, while $\mathscr{R}$ induces engagement $e_1^{\mathscr{R}} = 1$ and shows the message of user 4. The posterior variances are, respectively, $\text{Var}[\theta|\{\theta_1, \theta_2, \theta_3\}] = 0.58$ and $\text{Var}[\theta|\{\theta_1, \theta_4\}] = 0.75$. Note that the extra message user 1 reads under $\mathscr{C}$ is key, as otherwise $\text{Var}[\theta|\{\theta_1, \theta_2\}] = 0.9$.

The reverse-chronological algorithm might be an alternative to enhance learning, but it does not seem realistic that bringing it back to social media platforms would work: most users are better off under the closest algorithm, even if platform size is large. While the goal of the DSA is to target the harms coming from personalization algorithms (mainly, as we show in Section 3, the excessive similarity in the feeds), it does not seem realistic to think of users going back to the reverse-chronological algorithm by themselves.[15] And

this can be understandable: personalization algorithms had the objective of maximizing engagament and they certainly succeded (Guess et al., 2023). The platform-optimal algorithm is a sophisticated tool designed to please the user.

Another alternative worth exploring is to offer small modifications of the platform-optimal algorithm. Recently, platforms like X or Facebook have been adding features that "give context" or "promote fact-checked content". While keeping their personalized feeds, they sometimes incorporate a sponsored message, trying to improve users' information. Next, we study how such a modified platform-optimal algorithm might work in this model. We create the breaking echo-chambers algorithm, $\mathscr{B}$, by simply adding a user with opposite views to the closest feed. Formally, for every $i \in \mathcal{U}$, $\mathscr{B}_i(k) = \mathscr{C}_i(k-1)$, and $\mathscr{B}_i(1)$ is precisely the user with the highest negative correlation to user i.

The next result shows that, when platform size grows large, $\mathscr{B}$ allows the user to correctly learn the state of the world, maximizing learning at no cost in conformity. Remember that user i's expected conformity is $\sigma^2 - \frac{\sigma_{ij}^2}{\sigma^2}$. Hence, she is indifferent between a covariance of $\sigma_{ij} = -\sigma^2$ or $\sigma_{ij} = \sigma^2$, so, asymptotically, the breaking echo chambers algorithm incurs no penalty when maximizing learning. Because conformity and learning are simultaneously maximized, this algorithm converges to a utilitarian optimal algorithm. Note that, in contrast, the finite case is ambiguous: if there is a *very* opposite user in the pool, neither conformity nor engagement will be that harmed and learning will be significantly improved. However, it might be the case that no such user is available, conformity and engagement are punished, and even though there is an improvement in learning, the platform-optimal algorithm provides higher utility.

In fact, when platform size is large, the breaking echo chambers algorithm has an effect on users similar to that of the plataform aggregating information and displaying it publicly. When platform size is not that large, the ability of the breaking echo chambers algorithm to achieve perfect learning should be at least questioned. Summarizing: when platform size is large, the breaking echo chambers algorithm works as a utilitarian optimal one, but it is not the case when platform size is finite. Hence, it is still a must to analyze how to achieve the implementation of the utilitarian optimal algorithm in general.

Proposition 4.2. *When platform size grows large, the breaking echo chambers algorithm outperforms the closest algorithm and converges to a utilitarian optimal algorithm.*

Proof. See Appendix A. □

Platforms have already implemented algorithm modifications that promote content intended to improve user information: for example, in 2021, Twitter (now X) launched

15 This becomes even more complicated when platforms, like Instagram currently, make the button for accessing a reverse-chronological feed difficult to find and provide the personalized feed by default each time a user logs in.

"Birdwatch", which became widespread in 2023, a feature where contributors could give context under a post. As in our model we do not allow the platform to know the messages of the users, we cannot build a similar feature in which the platform could aggregate them to obtain (and potentially share) an estimator for θ, but this is what the breaking echo chambers algorithm does in practice when platform size grows large. The purpose is the same: each user would read (and learn) the state of the world and at the same time derive some instantaneous utility from interacting with friends.

In any case, implementing such an algorithm has some obvious drawbacks: it requires some regulatory enforcement (the platform has no incentives to implement it by itself), and its long-term viability in the real world remains questionable. Although opposite content might be enforced, maybe through sponsored public service announcements with regular frequency or by directly incorporating dissimilar views into the feed, any user may simply choose to disregard artificially added content and, perhaps naively, opt not to engage with it.

So far, we have explored the current institutional alternative to the platform-optimal algorithm, the reverse-chronological algorithm, and also an artificial improvement to the platform-optimal algorithm, the breaking echo-chambers algorithm. However, none of these alternatives are fully satisfactory, as either their performance or their viability is questionable. There is, however, one alternative we have not yet explored: the utilitarian optimal algorithm. This algorithm is characterized by maximizing social welfare, and, by Lemma 3.2, this is equivalent to maximizing the expected utility of each user. Nonetheless, we cannot provide a closed-form expression for this algorithm. We do, however, offer an example that can be found in Appendix B. We will denote the user-optimal algorithm as $\mathscr{U}$, and the next section is dedicated to exploring how, under the imposition of horizontal interoperability in a competitive market, platforms are compelled to implement it.

5 The need for horizontal interoperability

So far, we worked in our baseline model where a monopolist platform caters to a pool of n users who are already on board. In this scenario, the platform is not concerned about user capture but focuses solely on maximizing the time users spend on the platform—their engagement. This mirrors the current landscape of social media platforms. Large platforms like Instagram, TikTok, or X operate as monopolists within their specific niches: for instance, if someone wants to join a community for sharing pictures with friends, she would likely choose Instagram. While other sites may exist, the critical factor is that her friends are on Instagram. Network effects—a key feature of social media platforms wherein the platform's value increases as more users join and engage—protect these large incumbents. Consequently, platforms have strong incentives to grow their user base to offer greater network benefits than their rivals. This creates a high barrier to entry for

new competitors, who must offer a vastly superior service to overcome the network effects and attract users.[16]

Network effects are particularly significant when it comes to algorithms, as they heavily rely on platform size.[17] The larger the network, the more possibilities for optimizing feeds and, eventually, the higher the expected utility for users. This is evident for the user-optimal algorithm: since platform and user incentives are aligned, a larger pool from which the platform can curate a feed translates to higher expected utility. However, the situation is more nuanced for the closest algorithm, as two opposing forces come into play when the platform size increases. On one hand, within-the-platform utility increases due to better matching possibilities. On the other hand, learning might decrease (we know that learning does not behave monotonically for small size increases, but that it asymptotically vanishes). Intuitively, the strategic role of the platform means the first force should dominate: the feed is chosen to maximize within-the-platform utility, with the effects on learning being a secondary consequence.

The next results provides a necessary and sufficient condition on the parameter λ for the closest algorithm to feature network effects. Before presenting it, let us introduce slight changes in notation. Let us refer to $U_i^n(.)$ to user i's utility when platform size is n, and similarities are captured by $\mathbf{\Sigma}$. Then, $U_i^{n+1}(.)$ refers to user i's utility function when platform size has grown to $n+1$, and similarities are captured by the extension of $\mathbf{\Sigma}$ as described in Section 4. Moreover, we denote e_i user i's engagement when platform size is n, and $\tilde{e}_i$ user i's engagement when platform size is $n+1$. Formally, we say that an algorithm features network effects if and only if the expected utility the user derives from joining the platform increases with platform size, i.e., $\mathbb{E}_i[U_i^{n+1}(\tilde{e}_i, \theta_i, \theta_{-i}, \mathscr{F}_i)] \geq \mathbb{E}_i[U_i^n(e_i, \theta_i, \theta_{-i}, \mathscr{F}_i)]$ for all n. From then on, we will work under the case of $\mathscr{C}$ featuring network effects, as it is the standard in this literature.

Proposition 5.1. *Denoting by $\mathscr{C}(n)$ the closest algorithm applied to platform size n, and by $\Delta \operatorname{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}}] = \mathbb{E}_i[\operatorname{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}(n+1)}] - \operatorname{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}(n)}]]$ the expected difference in learning when the platform size increases from n users to $n+1$, we have that the closest algorithm features network effects if and only if*

$$\lambda \geq \frac{1}{1 + \frac{1}{\Delta \operatorname{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}}]} \left(\alpha \, \mathbb{E}_i[\tilde{e}_i - e_i] + (1-\beta)\mathbb{E}_i[\nu(n, \tilde{e}_i, e_i)]\right)},$$

where $\nu(n, \tilde{e}_i, e_i) = \frac{2\tilde{e}_i(\tilde{e}_i - e_i)(3+2(e_i+\tilde{e}_i))+6(e_i-\tilde{e}_i)\tilde{e}_i n+3n^2}{3e_i(2+n)^2} > 0.$

Proof. See Appendix A. □

16 This was the case, for example, when Facebook was launched and then replaced MySpace as the leading social networking site in 2009.

17 Many other services, such as privacy protection tools, accessibility or design do not depend on platform size.

Network effects are, therefore, platform-specific and proprietary. A platform with a small user base will provide low expected utility to its users, even if implementing the user-optimal algorithm $\mathscr{U}$. Consequently, users gravitate towards the large incumbent, causing the market to tip in its favor. The incumbent platform has no incentives to deviate from the platform-optimal algorithm, effectively trapping users. This is where the need for horizontal interoperability in social media platforms becomes apparent. Horizontal interoperability would enable a user from platform A to see posts from friends on platform B and vice versa. In other words, the algorithm implemented by platform A could match users from A with those from B, while also accessing their previous posts and interactions.[18] Some industries, such as the cell and email industries, have already become interoperable: for example, a Yahoo user can send an email to a Gmail user seamlessly.

Although horizontal interoperability is a measure potentially applicable in many markets that feature network effects, it is particularly beneficial here for two main reasons. First, the implementation of interoperability removes entry barriers created by network effects, shifting them from the platform level to the market level and distributing them among all market players. This levels the playing field, increasing competition and contestability (Crémer et al., 2000; Kades and Scott Morton, 2020). This argument is applicable to almost any market with network effects but may not hold in markets with few non-interoperable features. For example, if messaging apps like WhatsApp or Telegram were mandated to become interoperable, users not concerned about privacy would have little reason to switch from WhatsApp. Even if switching costs are low, the lack of significant non-interoperable features in messaging apps means users would likely remain with the monopolist, WhatsApp. In social media platforms, however, algorithms are a key non-interoperable feature: while platform A implements $\mathscr{C}$, platform B could implement $\mathscr{U}$. Thus, as network effects are shared, platforms must compete at the algorithm level. This is the second and crucial reason for implementing horizontal interoperability. The primary way a platform can differentiate itself is through its personalized feed algorithm. Without platform-specific network effects, users can freely choose the feed that offers the best expected utility.

In a simplified setting where interoperability eliminates the incumbent's advantage from network effects, platforms are compelled to implement the user-optimal algorithm $\mathscr{U}$. Otherwise, users will migrate to a competitor implementing it. This argument is key: horizontal interoperability would naturally induce platforms to adopt the utilitarian algorithm. The following part of this section discusses the potential benefits and weaknesses

18 Although we consider a complete network in this paper, the results apply to general networks and are more relevant in this section. With interoperability, users can maintain their neighborhood regardless of which platform each friend is a member of. This is similar to the mobile phone industry, where the focus is on whether a friend has a mobile phone, not the company providing the service.

of horizontal interoperability in social media platforms, its implementation challenges, and its current status in European legislation, particularly regarding the DMA.

Apart from the benefits already outlined, horizontal interoperability makes network effects a public good and then induces competition in all dimensions of non-interoperable features. Following our example, if two platforms were to implement the user-optimal algorithm, they would be equally attractive to a potential user. However, they could still compete in other dimensions such as service quality, user interface quality, or privacy and security. Hence, interoperability induces innovation in the non-interoperable features. By eliminating entry barriers generated by network effects, and given that entry cost is relatively low in social media, market contestability is also enhanced. Quite intuitively, large platforms will oppose interoperability: it disadvantages platforms with significant network effects, as consumer adoption decisions are no longer influenced by size. Conversely, smaller platforms would fear losing if they competed *for the market* and thus prefer interoperability to be able to compete *in the market* (Belleflamme and Peitz, 2020).[19]

The main weakness of horizontal interoperability in social media platforms is the challenge its implementation constitutes, both in practical terms and regarding the consequences for privacy. While it does not seem too complicated to develop an standard of the basic features (see Kades and Scott Morton (2020) for an overview on standarization), platforms would need to share private data. This includes not only the messages that their users post (which in most cases are public), but also individual-level data regarding their interactions, as this allows for the calculation of similarities. Opening such data flows to third parties will raise privacy and security concerns.[20] Moreover, interoperability poses a challenge in services that promise end-to-end encryption. Cryptographers widely agree that maintaining encryption between different apps may prove challenging, if not impossible.

Aiming at "preventing gatekeepers from imposing unfair conditions on business and end users and at ensuring the openess of important digital services",[21] the European Comission has introduced interoperability as a regulatory measure in the European Union through the Digital Markets Act (DMA), passed in July 2022. Under this act, "gatekeeper" platforms and services are mandated to provide interoperability for chats with users on other services.[22] However, despite horizontal interoperability gaining traction

19 Still, becoming interoperable is always a decision for the small platform to make. Regulators just require large platforms to make it possible.

20 For the interested reader, we refer to Bourreau and Krämer (2022) for a detailed discussion of privacy and security risks of interoperability in digital markets.

21 This quote is extracted from *Questions and Answers: Digital Markets Act: Ensuring fair and open digital markets*, available at https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_20_2349/QANDA_20_2349_EN.pdf.

as a regulatory measure in the EU, its actual implementation in social media platforms remains distant, as currently only messaging services are addressed.[23]

6 Conclusion

We have developed a model of communication and learning through a personalized feed. An engagement-maximizing platform excessively weights conformity when designing feeds, aligning with existing evidence on echo chambers and filter bubbles (Pariser, 2011). The platform overemphasizes users' desire for conformity, resulting in severely impaired learning. Pariser argues that individualized personalization through algorithmic filtering could lead to intellectual isolation and social fragmentation as the product of being surrounded only by like-minded individuals. Our paper theoretically demonstrates that this is the price to pay, as Guess et al. (2023) show empirically, when platforms are free to manage information exchanges to maximize profits and, hence, engagement. Institutional efforts to improve this situation have relied on the reverse-chronological algorithm, but our analysis suggests that it may not be sufficient. Users enjoy receiving recommended content, and while a random selection might enhance learning, the associated disutility may outweigh the benefits. Additionally, the likelihood of users disconnecting prematurely increases, meaning that even if diverse content enhances learning, users do not consume enough of this content.

The breaking echo chambers algorithm is a promising alternative when the platform size is large, which is the case for most social media platforms today. However, its practical implementation may be challenging. We propose horizontal interoperability as a solution, arguing that it is not just a *silver bullet* but a highly advantageous measure for the market we are analyzing. Algorithms are a non-interoperable feature of social media platforms, and the primary way platforms differentiate themselves from competitors is by finding the best algorithm for users. Competition would naturally lead to the implementation of healthier algorithms that fulfill users' desire for conformity while significantly enhancing learning.

Further research avenues emerge from this work. Beyond addressing the technical difficulties in the proper priors version of this model, we aim to explore how the algorithms we study affect *polarization*, defined as the sum of the squares of the differences between each user's beliefs about θ and the average belief. Additionally, we plan to further analyze horizontal interoperability. Interoperability might offer broader benefits than those

22 Gatekeeper platforms, defined as those entities exerting substantial market influence and possessing or expected to possess a firmly established and enduring market position, are designated by the European Commission. They are Alphabet, Amazon, Apple, ByteDance, Meta and Microsoft.

23 For the interested reader, we refer to Bourreau and Krämer (2023) for an overview on horizontal and vertical interoperability in the DMA.

discussed in this paper. For example, Farronato et al. (2024) show that when users have heterogeneous preferences, a single platform might not be as effective as multiple platforms: network effects and platform differentiation offset each other when the market tips. In principle, interoperability might resolve this issue: network effects would occur at the market level, maximizing them, while platform differentiation would still exist. Analyzing the effects of interoperability in a dynamic setting of competing platforms where heterogeneous users can multi-home is a natural extension of this work. Specifically, we aim to address two key questions: firstly, whether the necessary standards for interoperability could restrain innovation, and secondly, whether super-large platforms can maintain their dominance over time due to factors beyond algorithm competition.

References

Abreu, Luis and Doh-Shin Jeon, "Homophily in social media and news polarization," working paper, available at *https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3468416*, 2019.

Acemoglu, Daron and Asuman Ozdaglar, "Opinion dynamics and learning in social networks," *Dynamic Games and Applications*, 2011, *1* (1), 3–49.

_ , _ , and James Siderius, "A model of online misinformation," NBER Working Paper no 28884, available at *https://siderius.lids.mit.edu/wp-content/uploads/sites/36/2022/09/fake-news-July-20-2022.pdf*, 2023.

Agarwal, Saharsh, Uttara M. Ananthakrishnan, and Catherine E. Tucker, "Deplatforming and the control of misinformation: Evidence from Parler," working paper, available at *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4232871*, 2022.

Allcott, Hunt and Matthew Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 2017, *31* (2), 211–236.

_ , _ , and Lena Song, "Digital addiction," *American Economic Review*, 2022, *112* (7), 2424–2463.

Belleflamme, Paul and Martin Peitz, "The competitive impacts of exclusivity and price transparency in markets with digital platforms," *Concurrences*, 2020, (1), 2–12.

Benhabib, Jess and Alberto Bisin, "Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions," *Games and Economic Behavior*, 2005, *52* (2), 460–492.

Benzell, Seth and Avinash Collis, "Regulating digital platform monopolies: The case of Facebook," working paper, available at *https://www.aeaweb.org/conference/2023/program/paper/SNNd74ni*, 2022.

Bernheim, B. Douglas, "A theory of conformity," *Journal of Political Economy*, 1994, *102* (5), 841–877.

Biddle, Sam, "Facebook uses artificial intelligence to predict your future actions for advertisers, says confidential document," *The Intercept*, 2018, *13* (04), 2018.

Biglaiser, Gary, Jacques Crémer, and André Veiga, "Should I stay or should I go? Migrating away from an incumbent platform," *RAND Journal of Economics*, 2022, *53* (3), 453–483.

Bonatti, Alessandro and Gonzalo Cisternas, "Consumer scores and price discrimination," *Review of Economic Studies*, 2020, *87* (2), 750–791.

Bourreau, Marc and Jan Krämer, "Interoperability in digital markets," available at https://sitic.org/wordpress/wp-content/uploads/Interoperability-in-Digital-Markets.pdf, 2022.

_ **and** _ , "Horizontal and vertical interoperability in the DMA," available at https://cerre.eu/wp-content/uploads/2023/12/ISSUE-PAPER-CERRE-DEC23DMA-Horizontal-and-Vertical-Interoperability-Obligations.pdf, 2023.

Bursztyn, Leonardo, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth, "When product markets become collective traps: the case of social media," working paper, available at https://ssrn.com/abstract=4596071, 2023.

Chamley, Christophe, *Rational herds: Economic models of social learning*, Cambridge University Press, 2004.

Chitra, Uthsav and Christopher Musco, "Analyzing the impact of filter bubbles on social network polarization," in "in" 2020, pp. 115–123.

Cialdini, Robert B. and Noah J. Goldstein, "Social influence: Compliance and conformity," *Annual Review of Psychology*, 2004, *55* (1), 591–621.

Crémer, Jacques, Patrick Rey, and Jean Tirole, "Connectivity in the commercial Internet," *Journal of Industrial Economics*, 2000, *48* (4), 433–472.

Demange, Gabrielle, "Simple visibility design in network games," working paper, available at https://hal.science/hal-03925344v1/document, 2023.

DeMarzo, Peter M, Dimitri Vayanos, and Jeffrey Zwiebel, "Persuasion bias, social influence, and unidimensional opinions," *Quarterly Journal of Economics*, 2003, *118* (3), 909–968.

Ellman, Matthew and Fabrizio Germano, "What do the papers sell? A model of advertising and media bias," *Economic Journal*, 2009, *119* (537), 680–704.

Evens, Tom, Karen Donders, and Adelaida Afilipoaie, "Platform policies in the European Union: Competition and public interest in media markets," *Journal of Digital Media & Policy*, 2020, *11* (3), 283–300.

Farronato, Chiara, Jessica Fong, and Andrey Fradkin, "Dog eat dog: Balancing network effects and differentiation in a digital platform merger," *Management Science*, 2024, *70* (1), 464–483.

Franck, Jens-Uwe and Martin Peitz, "Market power of digital platforms," *Oxford Review of Economic Policy*, 2023, *39* (1), 34–46.

Galeotti, Andrea, Benjamin Golub, Sanjeev Goyal, and Rithvik Rao, "Discord and harmony in networks," *available at https://arxiv.org/pdf/2102.13309*, 2021.

Golub, Benjamin and Matthew O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, 2010, *2* (1), 112–149.

Grayot, James D., "Dual process theories in behavioral economics and neuroeconomics: A critical review," *Review of Philosophy and Psychology*, 2020, *11* (1), 105–136.

Greene, Steven, "Social identity theory and party identification," *Social science quarterly*, 2004, *85* (1), 136–153.

Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow et al., "How do social media feed algorithms affect attitudes and behavior in an election campaign?," *Science*, 2023, *381* (6656), 398–404.

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya, "Curtailing false news, amplifying truth," working paper, available at https://shs.hal.science/halshs-04315924/document, 2023.

Habib, Hussam, Maaz Bin Musa, Fareed Zaffar, and Rishab Nithyanand, "To act or react: Investigating proactive strategies for online community moderation," *available at https://arxiv.org/pdf/1906.11932*, 2019.

Hartigan, John A., *Bayes Theory*, Springer Science & Business Media, 1983.

Horwitz, Jeff et al., "The Facebook files," *The Wall Street Journal,* available online at: https://www.wsj.com/articles/the-facebook-files-11631713039, 2021.

Hu, Lin, Anqi Li, and Xu Tan, "A rational inattention theory of echo chamber," working paper, available at https://arxiv.org/pdf/2104.10657.pdf, 2021.

Hwang, Elina H. and Stephanie Lee, "A nudge to credible information as a countermeasure to misinformation: Evidence from twitter," working paper, available at https://ssrn.com/abstract=3928343, 2021.

Jackson, Matthew O., Suraj Malladi, and David McAdams, "Learning through the grapevine and the impact of the breadth and depth of social networks," *Proceedings of the National Academy of Sciences*, 2022, *119* (34).

Jadbabaie, Ali, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, 2012, *76* (1), 210–225.

Kades, Michael and Fiona Scott Morton, "Interoperability as a competition remedy for digital networks," *Washington Center for Equitable Growth Working Paper Series*, 2020.

Kosinski, Michal, David Stillwell, and Thore Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, 2013, *110* (15), 5802–5805.

Kranton, Rachel and David McAdams, "Social connectedness and information markets," working paper, available at https://sites.duke.edu/rachelkranton/files/2022/12/Social_Connectedness___Information_Markets-Dec-17_2022-final.pdf, 2022.

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), 831–870.

Loomba, Sahil, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nature Human Behaviour*, 2021, *5* (3), 337–348.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Ali Jadbabaie, "A theory of non-Bayesian social learning," *Econometrica*, 2018, *86* (2), 445–490.

Morton, Fiona Scott, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A. Douglas Melamed, and Jamie Morgenstern, "Market structure and antitrust subcommittee report," George J. Stigler Center for the Study of the Economy and the State, available at https://research.chicagobooth.edu/-/media/research/stigler/pdfs/market-structure-report.pdf, 2019.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Rand, "Shared partisanship dramatically increases social tie formation in a Twitter field experiment," *Proceedings of the National Academy of Sciences*, 2021, *118* (7), e2022761118.

Mostagir, Mohamed and James Siderius, "Learning in a post-truth world," *Management Science*, 2022, *68* (4), 2860–2868.

_ **and** _ , "When do misinformation policies (not) work?," available online at: https://siderius.lids.mit.edu/how-do-we-stop-misinformation/, 2023.

_ **and** _ , "When should platforms break echo chambers?," available online at: https://siderius.lids.mit.edu/wp-content/uploads/sites/36/2023/01/quarantine-v6.pdf, 2023.

Mudambi, Maya and Siva Viswanathan, "Prominence reduction versus banning: An empirical investigation of content moderation strategies in online platforms," working paper, available at https://core.ac.uk/download/pdf/542548963.pdf, 2022.

Mueller-Frank, Manuel and Claudia Neri, "A general analysis of boundedly rational learning in social networks," *Theoretical Economics*, 2021, *16* (1), 317–357.

_ , Mallesh M. Pai, Carlo Reggini, Alejandro Saporiti, and Luis Simantujak, "Strategic management of social information," working paper, available at *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4434685*, 2022.

Pariser, Eli, *The filter bubble: How the new personalized web is changing what we read and how we think*, Penguin, 2011.

Pennycook, Gordon and David G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, 2019, *188*, 39–50.

_ and David G Rand, "The psychology of fake news," *Trends in Cognitive Sciences*, 2021, *25* (5), 388–402.

Popiel, Pawel, "Addressing platform power: The politics of competition policy," *Journal of Digital Media & Policy*, 2020, *11* (3), 341–360.

Reuter, Jonathan and Eric Zitzewitz, "Do ads influence editors? Advertising and bias in the financial media," *Quarterly Journal of Economics*, 2006, *121* (1), 197–227.

Sagioglou, Christina and Tobias Greitemeyer, "Facebook's emotional consequences: Why Facebook causes a decrease in mood and why people still use it," *Computers in Human Behavior*, 2014, *35*, 359–363.

Silverman, Craig, "This analysis shows how viral fake election news stories outperformed real news on Facebook," *BuzzFeed,* available at https://newsmediauk.org/wp-content/uploads/2022/10/Buzzfeed-_This_Analysis_Shows_How_Viral_Fake_Election_News_Stories_Outperformed_Real_News_On_Facebook_-_BuzzFeed_News.pdf, November 16, 2016.

Solon, Olivia, "Facebook's failure: Did fake news and polarized politics get Trump elected?," *The Guardian,* available at *https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories*, November 10, 2016.

Sunstein, Cass R., "A prison of our own design: Divided democracy in the age of social media," *Democratic Audit UK*, 2017.

Törnberg, Petter, "Echo chambers and viral misinformation: Modeling fake news as complex contagion," *PLoS one*, 2018, *13* (9).

A Omitted proofs

Proof of Proposition 3.1

Proof. User i chooses message $m_i \in \mathbb{R}$ to maximize her expected utility, knowing her private signal θ_i and the algorithm $\mathscr{F}$. I.e., user i picks m_i to maximize:

$$\mathbb{E}_i[U_i|\theta_i,\mathscr{F}] = \lambda\left(v(e_i) - \beta(\theta_i - m_i)^2 - (1-\beta)\mathbb{E}_i\left[\sum_{j\in\mathscr{F}_i^{e_i}}\frac{(m_i - m_j(\theta_j))^2}{e_i}|\theta_i,\mathscr{F}\right]\right)$$
$$-(1-\lambda)\mathbb{E}_i[(a_i-\theta)^2|\theta_i,\mathscr{F}].$$

This is equivalent to maximizing

$$-\beta(\theta_i - m_i)^2 - (1-\beta)\left(m_i^2 + \sum_{j\in\mathscr{F}_i^{e_i}}\mathbb{E}_i\left[\frac{m_j(\theta_j)^2}{e_i}|\theta_i,\mathscr{F}\right] - 2m_i\sum_{j\in\mathscr{F}_i^{e_i}}\mathbb{E}_i\left[\frac{m_j(\theta_j)}{e_i}|\theta_i,\mathscr{F}\right]\right).$$

The first order condition with respect to m_i yields

$$m_i = \beta\theta_i + (1-\beta)\frac{1}{e_i}\sum_{j\in\mathscr{F}_i^{e_i}}\mathbb{E}_i\left[m_j(\theta_j)|\theta_i,\mathscr{F}\right]. \tag{4}$$

As this holds for all $j \in \mathcal{U}$, we substitute in this expression $m_j(\theta_j) = \beta\theta_j + (1-\beta)\frac{1}{e_j}\sum_{l\in\mathscr{F}_j^{e_j}}\mathbb{E}_j\left[m_l(\theta_l)|\theta_j,\mathscr{F}\right]$ for all $j \in \mathscr{F}_i^{e_i}$, and then we repeat the procedure for all $l \in \mathscr{F}_j^{e_j}$ and so on. Users' knowledge of $\mathscr{F}$ and Σ is crucial at this point, allowing us to commute the sum and the expectation operators. We can iterate on this procedure as many times as desired:[24]

$$\begin{aligned}
m_i &= \beta\theta_i + (1-\beta)\frac{1}{e_i}\sum_{j\in\mathscr{F}_i^{e_i}}\mathbb{E}_i\left[m_j(\theta_j)|\theta_i,\mathscr{F}\right]\\
&= \beta\theta_i + (1-\beta)\frac{1}{e_i}\sum_{j\in\mathscr{F}_i^{e_i}}\mathbb{E}_i\left[\beta\theta_j + (1-\beta)\frac{1}{e_j}\sum_{l\in\mathscr{F}_l^{e_j}}\mathbb{E}_j[m_l(\theta_l)|\theta_j,\mathscr{F}]\right]\\
&= \beta\theta_i + (1-\beta)\beta\theta_i + \frac{(1-\beta)^2}{e_ie_j}\sum_{j\in\mathscr{F}_i^{e_i}}\left(\sum_{l\in\mathscr{F}_l^{e_j}}\mathbb{E}_i[\mathbb{E}_j[m_l(\theta_l)|\theta_j,\mathscr{F}]|\theta_i,\mathscr{F}]\right) = ...\\
&= \beta\theta_i\sum_{r=0}^{m}(1-\beta)^r + \frac{(1-\beta)^m}{e_ie_j...e_m}\sum_{j\in\mathscr{F}_i^{e_i}}\left(...\left(\sum_{p\in\mathscr{F}_m^{e_m}}\mathbb{E}_i[...[\mathbb{E}_m[m_p(\theta_p)|\theta_m,\mathscr{F}]...]|\theta_i,\mathscr{F}]\right)...\right).
\end{aligned} \tag{5}$$

This expression holds for all $m \in \mathbb{N}$, so we can take limits when $m \to \infty$. On the one hand, $\lim_{m\to\infty}\sum_{r=0}^{m}(1-\beta)^r = \frac{1}{\beta}$, and, hence, the first term in Equation (5) is simply θ_i.

24 Abusing notation, we iterate m times and also refer to the m-th user as m.

On the other hand, the second term vanishes as $m \to \infty$. Hence, $m_i^* = \theta_i$ for all $i \in \mathcal{U}$ and we have truthtelling for any algorithm $\mathscr{F}$ and any engagement levels $\{e_i\}_{i\in\mathcal{U}}$. □

Proof of Proposition 3.3

Proof. The probability that, under algorithm $\mathscr{F}$, user i stays for one more period after staying for k is given by $g(u_i(k, \mathscr{F}))$. To maximize such probability, the platform chooses $\mathscr{F}_i(k) = \arg\max_{j\in\mathcal{U}\setminus\mathscr{F}_i^{k-1}}\{\mathbb{E}\left(g(u_i(k, \mathscr{F}))\right)\}$. As g is strictly increasing on u_i and the expectation preserves the order, this is equivalent to maximizing user i's expected inside-the-platform utility.

Given truthful reporting and noting that $v(.)$ is independent of the algorithm $\mathscr{F}$, we can write the platform's objective as finding the user $j \in \mathcal{U}\setminus\mathscr{F}_i^{k-1}$ that maximizes $-\mathbb{E}_p\left(\sum_{l\in\mathscr{F}_i^{k-1}}(\theta_i - \theta_l)^2 + (\theta_i - \theta_j)^2\right)$ or simply $-\mathbb{E}_p\,(\theta_i - \theta_j)^2)$. Thus, maximizing the probability of user i staying for one more period is equivalent to minimizing the conformity cost of such period.

Let us prove next that the algorithm that maximizes expected engagement is the same that maximizes within-the-platform expected utility. Given algorithm $\mathscr{F}$, the probability of staying at least until period e_i is $\Pi_{j=1}^{e_i} g(u_i(j, \mathscr{F}))$, and the probability of staying precisely until period e_i is then

$$\Pi_{j=1}^{e_i} g(u_i(j, \mathscr{F}))\left(1 - \Pi_{j=e_i+1}^{n} g(u_i(j, \mathscr{F}))\right).$$

Now, let us take two feeds, namely $\mathscr{F}_i$ and $\mathscr{F}_i'$, such that they are identical except from two users that are interchanged, i.e., there are users t and t' such that

$$\mathscr{F}_i(t) = \mathscr{F}_i'(t') \text{ and } \mathscr{F}_i(t') = \mathscr{F}_i'(t).$$

Moreover, they satisfy $-\mathbb{E}_p\left((\theta_i - \theta_t)^2\right) > -\mathbb{E}_p\left((\theta_i - \theta_t')^2\right)$. All this means that in the feed $\mathscr{F}_i$, the user who penalizes conformity the least is shown before. Without loss of generality we can assume that $t = 1$ and $t' = 2$, and then our goal is to show that such feed $\mathscr{F}_i$ yields higher expected engagement. Notice, first, that $g(u_i(1, \mathscr{F})) > g(u_i(1, \mathscr{F}'))$ because conformity is higher. Expected engagement under $\mathscr{F}$ reads as

$$g(1, \mathscr{F})\left[\left(1 - \Pi_{j=2}^{n} g(u_i(j, \mathscr{F}))\right) + g(u_i(2, \mathscr{F}))\left(2(1 - \Pi_{j=3}^{n} g(u_i(j, \mathscr{F})) + 3...)\right)\right].$$

Given that $g(u_i(j, \mathscr{F})) = g(u_i(j, \mathscr{F}'))$ for $j > 2$, we define c, which has the same value for both algorithms, as $c := \left(2(1 - \Pi_{j=3}^{n} g(U_i(j, \mathscr{F}_i)) + 3...)\right)$ to ease notation. Then, given that $v(e_i) = \alpha\, e_i$, $g(u_i(1, \mathscr{F})) = g(u_i(2, \mathcal{F}'))$ and $g(u_i(1, \mathscr{F}')) = g(u_i(2, \mathscr{F}))$, we have that

$$\begin{aligned}\mathbb{E}_p\,(e_i \mid \mathscr{F}) &= g(1, \mathscr{F})\left[\left(1 - \Pi_{j=2}^{n} g(u_i(j, \mathscr{F}))\right) + g(u_i(2, \mathscr{F}))C\right] \\ &\geq g(1, \mathscr{F}')\left[\left(1 - \Pi_{j=2}^{n} g(u_i(j, \mathscr{F}'))\right) + g(u_i(2, \mathscr{F}'))C\right] = \mathbb{E}_p\,(e_i \mid \mathscr{F}').\end{aligned}$$

This shows that any feed that is not reverse-ordered following the expected loss in conformity $\mathbb{E}_p((\theta_i - \theta_j)^2)$ is always dominated. □

Lemma A.1. *The posterior distribution of θ conditional on $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$ is given by*

$$\theta|\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} \sim \mathcal{N}\left(\frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}, \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}\right),$$

where $\mathbb{1}$ is an n-vector of ones, $\boldsymbol{\Sigma}_{\mathcal{S}_i}$ is the restriction of $\boldsymbol{\Sigma}$ to the users in $\mathscr{F}_i^{e_i}$, and $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}$ is the vector of private signals of the users in $\mathscr{F}_i^{e_i}$.

Proof. Let us assume, for simplicity, that the signals user i observes in her personalized feed $\mathscr{F}_i^{e_i}$ are $\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} = \{\theta_1, ..., \theta_k\}$. We know that $(\theta_1 \dots \theta_k) \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})$ because of the properties of the multinormal distribution. Now, the posterior distribution of θ conditional on $\theta_{\mathscr{F}_i^{e_i}}$ is proportional to the likelihood function:

$$\begin{aligned} g(\theta|\theta_{\mathscr{F}_i^{e_i}}) &\propto \left(2\pi \det(\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})\right)^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}})^t \boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathscr{F}_i^{e_i}})\right] \\ &= \left(2\pi \det(\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})\right)^{-1/2} \exp\left[-\frac{1}{2}\left(\theta^2 \mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t - 2\theta \mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} + \boldsymbol{\theta}^t_{\mathscr{F}_i^{e_i}}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}\right)\right]. \end{aligned}$$

Multiplying by the constant $\sqrt{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}\sqrt{\det(\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}})}$, we obtain:

$$\begin{aligned} g(\theta|\theta_{\mathscr{F}_i^{e_i}}) &= \sqrt{\frac{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}{2\pi}} \exp\left[-\frac{1}{2}\left(\theta^2 \mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t - 2\theta \mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}} + \frac{(\boldsymbol{\theta}^t_{\mathscr{F}_i^{e_i}}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1})^2}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}\right)\right] \\ &= \sqrt{\frac{\mathbb{1}\boldsymbol{\Sigma}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\theta - \frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}}{\sqrt{\frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}}}\right)^2\right]. \end{aligned}$$

This is the distribution function of a normal random variable with mean $\frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}$ and variance $\frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}$. Thus,

$$\theta|\theta_{\mathscr{F}_i^{e_i}} \sim \mathcal{N}\left(\frac{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\boldsymbol{\theta}_{\mathscr{F}_i^{e_i}}}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}, \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{F}_i^{e_i}}\mathbb{1}^t}\right)$$

as we wanted to show. □

Proof of Proposition 3.7

Proof. By assumption, $g(.) \in (0, 1)$, so that even if there is no penalty in conformity and within-the-platform utility is just given by $v(e_i)$, user i's engagement is a finite number. Let us call such number k. Given the generating process for new users, for every $\varepsilon > 0$,

there is some $\bar{n} \in \mathbb{N}$ such that if $n > \bar{n}$, there are user i's neighbors $j_1, ..., j_k$ such that $\rho_{i,j_r} > 1 - \varepsilon$ for all $r \in \{1, ..., k\}$. On the other hand, applying the Cauchy-Schwarz inequality to the correlations between the pairs formed by user i and two other users, say j_r and j_l, we get

$$\rho_{j_r,j_l} \geq \rho_{j_r,i}\rho_{j_l,i} - \sqrt{(1-\rho^2_{j_r,i})(1-\rho^2_{j_l,i})}.$$

Using the ε-bounds derived above, we obtain:

$$\rho_{j_r,j_l} \geq (1-\varepsilon)^2 - 2\varepsilon = 1 - 4\varepsilon + \varepsilon^2 \quad \forall\, j_r, j_l.$$

Now, assume engagement is e_i. As $e_i \leq k$, users from $\mathscr{C}_i^{e_i} \subset \mathcal{U}$ are taken from the set of k users specified above. Let us now define $\delta = 4\varepsilon - \varepsilon^2$. For every $\delta > 0$, there is some $\tilde{n}$ such that if $n > \tilde{n}$, the feed induced by the closest algorithm $\mathscr{C}_i^{e_i}$ verifies that if $j_r, j_l \in \mathscr{C}_i^{e_i}$,[25] $\rho_{j_r,j_l} > 1 - \delta$ (it is enough to choose ε accordingly). Hence, we have that for the matrix $\mathbf{A}$ defined as

$$\mathbf{A} := \sigma^2 \begin{pmatrix} 1 & 1-\delta & \dots & 1-\delta \\ 1-\delta & 1 & \dots & 1-\delta \\ \vdots & \vdots & \ddots & \vdots \\ 1-\delta & \dots & 1-\delta & 1 \end{pmatrix},$$

$\mathbf{A} \leq \mathbf{\Sigma}_{\mathscr{C}_i^{e_i}}$, where $\leq$ refers to element-wise ordering and $\mathbf{\Sigma}_{\mathscr{C}_i^{e_i}}$ is the covariance matrix for the users in $\mathscr{C}_i^{e_i}$. Now, we need an auxiliary result:

Lemma A.2. *In this particular case,* $\boldsymbol{A} \leq \mathbf{\Sigma}_{\mathscr{C}_i^{e_i}}$ *implies* $\mathbf{\Sigma}^{-1}_{\mathscr{C}_i^{e_i}} \leq \boldsymbol{A}^{-1}$.

Proof. Let $\mathbf{A}$ be the covariance matrix selected by the closest algorithm, i.e., $\mathbf{A} = \mathbf{\Sigma}_{\mathscr{C}_i^{e_i}}$:

$$\mathbf{A} = \begin{pmatrix} 1 & a_{12} & a_{13} & \dots & a_{1e_i} \\ a_{12} & 1 & a_{23} & \dots & a_{2e_i} \\ a_{13} & a_{23} & 1 & \dots & a_{3e_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1e_i} & a_{2e_i} & a_{3e_i} & \dots & 1 \end{pmatrix}$$

Let $\mathbf{B}$ be the following matrix

$$\mathbf{B} = \begin{pmatrix} 1 & b & b & \dots & b \\ b & 1 & b & \dots & b \\ b & b & 1 & \dots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b & b & b & \dots & 1 \end{pmatrix}$$

with $b = 1 - \delta$ such that $\mathbf{B} \leq \mathbf{A}$ element-wise. We denote the elements of the inverse

25 Here we abuse notation slightly, as $\mathcal{U}$ should be $\mathcal{U}(n)$ and $\mathscr{C}_i^{e_i}$ should be $\mathscr{C}_i^{e_i}(., n)$.

matrices $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ as follows:

$$\mathbf{A}^{-1} = \begin{pmatrix} \bar{a}_{11} & \bar{a}_{12} & \bar{a}_{13} & \dots & \bar{a}_{1e_i} \\ \bar{a}_{12} & \bar{a}_{22} & \bar{a}_{23} & \dots & \bar{a}_{2e_i} \\ \bar{a}_{13} & \bar{a}_{23} & \bar{a}_{33} & \dots & \bar{a}_{3e_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{1e_i} & \bar{a}_{2e_i} & \bar{a}_{3e_i} & \dots & \bar{a}_{e_ie_i} \end{pmatrix},$$

and

$$\mathbf{B} = \alpha \begin{pmatrix} 1 & \bar{b} & \bar{b} & \dots & \bar{b} \\ \bar{b} & 1 & \bar{b} & \dots & \bar{b} \\ \bar{b} & \bar{b} & 1 & \dots & \bar{b} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{b} & \bar{b} & \bar{b} & \dots & 1 \end{pmatrix}.$$

Now, as $\mathbf{A}\mathbf{A}^{-1} = \mathbf{Id}$, $\bar{a}_{11} + a_{12}\bar{a}_{12} + a_{13}\bar{a}_{13} + ... + a_{1e_i}\bar{a}_{1e_i} = 1$. Moreover, $\mathbf{A} \geq \mathbf{B}$ implies that $\bar{a}_{11} + b\sum_{j=2}^{e_i} \bar{a}_{1j} \leq 1$. On the other hand, as $\mathbf{B}\mathbf{B}^{-1} = \mathbf{Id}$, $\alpha(1 + b\bar{b}(e_i - 1)) = 1$. Hence,

$$\bar{a}_{11} + b\sum_{j=2}^{e_i} \bar{a}_{1j} \leq \alpha(1 + b\bar{b}(e_i - 1)), \quad \forall\, b \in (0, 1).$$

This implies that $\bar{a}_{11} \leq \alpha$ and $\sum_{j=2}^{e_i} \bar{a}_{1j} \leq \alpha(e_i - 1)\bar{b}$. Following the same reasoning, we obtain

$$\bar{a}_{ii} \leq \alpha \quad \forall\, i \text{ and } \bar{a}_{ij} \leq \alpha\bar{b} \quad \forall\, j \neq i.$$

Then, $\mathbf{A}^{-1} \leq \mathbf{B}^{-1}$ as we wanted to show. □

Therefore,

$$\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{C}_i^{e_i}}\mathbb{1}^t \leq \mathbb{1}\mathbf{A}^{-1}\mathbb{1}^t \Rightarrow \frac{1}{\mathbb{1}\mathbf{A}^{-1}\mathbb{1}^t} \leq \frac{1}{\mathbb{1}\boldsymbol{\Sigma}^{-1}_{\mathscr{C}_i^{e_i}}\mathbb{1}^t} \Rightarrow \frac{1}{\mathbb{1}\mathbf{A}^{-1}\mathbb{1}^t} \leq \mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}_i^{e_i}}].$$

On the other hand, we have that $\mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}_i^{e_i}}] \leq \sigma^2$ by construction (note that $\mathrm{Var}[\theta|\theta_i] = \sigma^2$). Consequently, after calculating $\mathbb{1}\mathbf{A}^{-1}\mathbb{1}^t = \frac{e_i}{\sigma^2(1+(e_i-1)(1-\delta))}$, we finally get:

$$\frac{\sigma^2(1 + (e_i - 1)(1 - \delta))}{e_i} \leq \mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}_i^{e_i}}] \leq \sigma^2$$

for every $\delta \in (0, 1)$. Finally, we have that $\delta \to 0$ as $n \to \infty$. Then, taking limits in the above expression we obtain that $\mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}_i^{e_i}}] = \sigma^2$.

□

Proof of Proposition 4.2

Proof. The following proof consists of two parts. First, we will show that conformity is at least as good under $\mathscr{B}$ than under $\mathscr{C}$. Second, we will show that, asymptotically, learning

is perfect under $\mathscr{B}$.

Note that, by assumption, $g(.) \in (0,1)$, so that even if there is no penalty in conformity and within-the-platform utility is just given by $v(e_i)$, user i's engagement is a finite number. Let us call such number k. Now, following the reasoning in the proof of Proposition 3.7, for every $\varepsilon > 0$ there exists a $\bar{n}(\varepsilon) \in \mathbb{N}$ such that for every $n > \bar{n}(\varepsilon)$, there exists a set of k users such that $\rho_{ij} \geq 1 - \varepsilon$ for every $j = 1, ..., k$. Moreover, for every $\delta > 0$, there exists a $\tilde{n}(\delta) \in \mathbb{N}$ such that for every $n > \tilde{n}(\delta)$ there is a user l such that $\rho_{il} < \delta - 1$. Let us now take $n \geq \max\{\bar{n}(\varepsilon), \tilde{n}(\delta)\}$ and define, for any engagement e_i, $\mathscr{B}_i^{e_i} = \{l, 1, ..., e_i - 1\}$, where users in $\{1, ..., e_i - 1\}$ are taken from the pool of size k obtained before. Then, as $n \to \infty$, we will have that $\rho_{ij} \to 1$ for all $j \in \{1, ..., k\}$ and $\rho_{il} \to -1$. Then, expected conformity under the breaking echo chambers algorithm becomes

$$\mathbb{E}_i\left(\sum_{j \in \mathscr{B}_i^{k+1}} (\theta_i - \theta_j)^2 \mid \theta_i\right) = \sigma^2 \sum_{j=1}^{k+1} (1 - \rho_{ij}^2) + \sigma^2(1 - \rho_{il}^2),$$

which converges to zero as $n \to \infty$.

Now, let us study learning. First, we analyze what user i learns from user l's message:

$$\text{Var}[\theta|\theta_i, \theta_l] = \frac{1}{\mathbb{1}\boldsymbol{\Sigma}_{il}^{-1}\mathbb{1}^T} = \frac{\delta(2-\delta)}{4-\delta},$$

which coverges to zero as n grows large. Note that, as $l \in \mathscr{B}_i^{e_i}$,

$$\text{Var}[\theta|\theta_i, \theta_l] \geq \text{Var}[\theta|\theta_{\mathscr{B}_i^{e_i}}] \geq 0,$$

so $\lim_{n\to\infty} \text{Var}[\theta|\theta_{\mathscr{B}_i^{e_i}}] = 0$ and there is perfect learning under the breaking echo chambers algorithm. □

Proof of Proposition 5.1

Proof. Again, we assume that covariances are drawn from a uniform distribution $\mathcal{U}[-\sigma^2, \sigma^2]$. The platform matches the user with those featuring the highest covariances to her, and then, in terms of within-the-platform utility it means that we have to compute

$$\begin{aligned}\mathbb{E}_i[u_i^n(e_i, \theta_i, \theta_{-i}, \mathscr{C})] =& \lambda\left(v(e_i) - (1-\beta)\frac{1}{e_i}\left(e_i\sigma^2 - \frac{1}{\sigma^2}\sum_{j \in \mathscr{C}_i^{e_i}} \sigma_{ij}^2\right)\right) \\ =& \lambda\left(v(e_i) - (1-\beta)\left(\sigma^2 - \frac{1}{e_i}\sum_{j=n-e_i+1}^{n}\left(\frac{4j^2}{(n+1)^2} - \frac{4j}{n+1} + 1\right)\right)\right).\end{aligned}$$

Hence, overall user i's expected utility is given by

$$\mathbb{E}_i[U_i^n(e_i,\theta_i,\theta_{-i},\mathscr{C})] = \lambda\left(v(e_i) - (1-\beta)\left(\sigma^2 - \frac{1}{e_i}\sum_{j=n-e_i+1}^{n}\left(\frac{4j^2}{(n+1)^2} - \frac{4j}{n+1} + 1\right)\right)\right) - (1-\lambda)\mathbb{E}_i[\mathrm{Var}[\theta|\boldsymbol{\theta}_{\mathscr{C}}]],$$

and a simple rearrangement of the expression $\mathbb{E}_i[U_i^{n+1}(\tilde{e}_i,\theta_i,\theta_{-i},\mathscr{C})] - \mathbb{E}_i[U_i^n(e_i,\theta_i,\theta_{-i},\mathscr{C})]$ yields the desired inequality. □

B Example

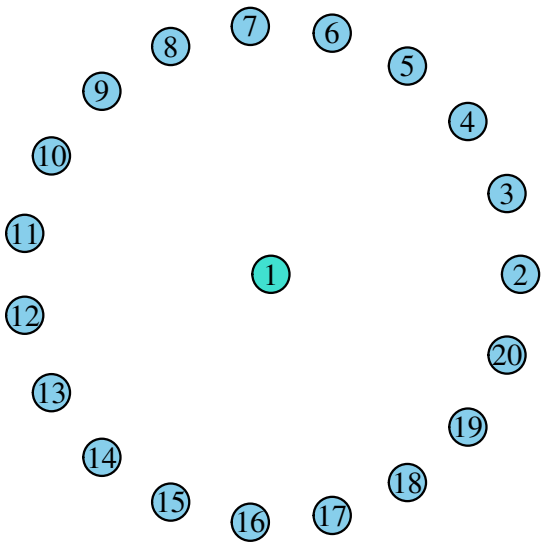

Figure 2: Platform size $n = 20$.

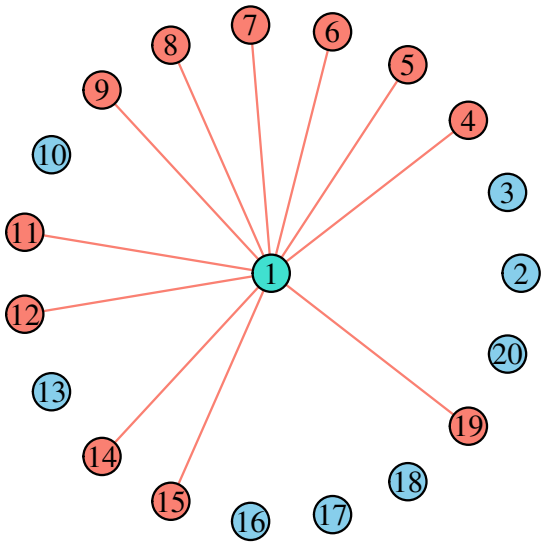

Figure 3: Platform-optimal feed.

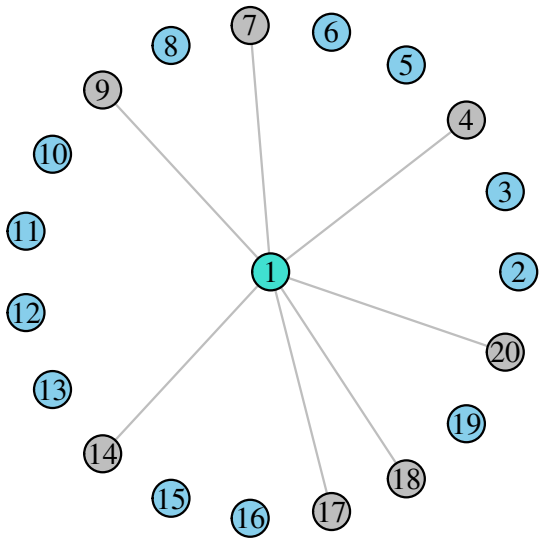

Figure 4: Reverse-chronological feed.

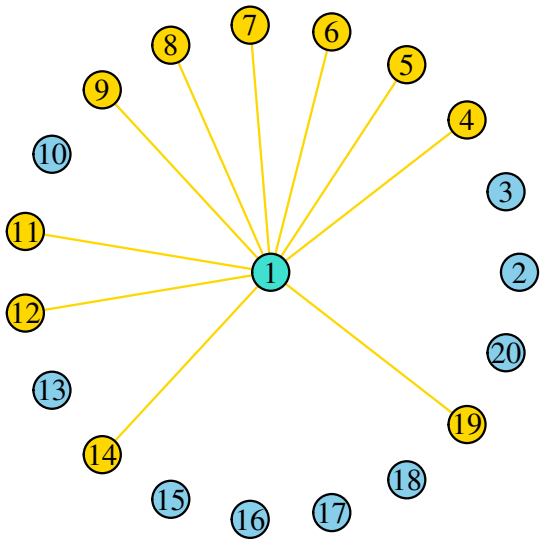

Figure 5: User-optimal feed.

Here we present the feeds user 1 would observe in a platform of size $n = 20$ (Figure 2) with similarity matrix $\mathbf{\Sigma}$ as displayed below. We fix parameters to $\alpha = 0.001$, $\lambda = 0.5$ and $\beta = 0.2$.

$$\mathbf{\Sigma} = \begin{pmatrix}
1.00 & -0.20 & -0.15 & 0.24 & 0.20 & 0.05 & 0.14 & 0.01 & 0.13 & -0.12 \\
-0.20 & 1.00 & -0.00 & -0.12 & 0.21 & 0.08 & -0.13 & -0.07 & -0.07 & 0.13 \\
-0.15 & -0.00 & 1.00 & -0.38 & -0.20 & -0.06 & -0.17 & 0.02 & -0.09 & -0.24 \\
0.24 & -0.12 & -0.38 & 1.00 & -0.23 & -0.20 & 0.04 & 0.05 & 0.03 & 0.07 \\
0.20 & 0.21 & -0.20 & -0.23 & 1.00 & -0.00 & 0.11 & -0.09 & -0.09 & 0.04 \\
0.05 & 0.08 & -0.06 & -0.20 & -0.00 & 1.00 & 0.27 & -0.17 & 0.06 & 0.06 \\
0.14 & -0.13 & -0.17 & 0.04 & 0.11 & 0.27 & 1.00 & 0.23 & 0.21 & -0.02 \\
0.01 & -0.07 & 0.02 & 0.05 & -0.09 & -0.17 & 0.23 & 1.00 & 0.10 & 0.17 \\
0.13 & -0.07 & -0.09 & 0.03 & -0.09 & 0.06 & 0.21 & 0.10 & 1.00 & 0.02 \\
-0.12 & 0.13 & -0.24 & 0.07 & 0.04 & 0.06 & -0.02 & 0.17 & 0.02 & 1.00 \\
0.21 & 0.06 & 0.04 & 0.05 & 0.01 & 0.13 & -0.02 & 0.16 & -0.02 & 0.14 \\
0.17 & 0.05 & -0.29 & 0.06 & 0.39 & -0.05 & 0.14 & -0.22 & -0.14 & -0.00 \\
-0.14 & 0.24 & 0.23 & -0.15 & -0.07 & 0.28 & 0.20 & 0.08 & -0.01 & 0.08 \\
0.14 & -0.18 & 0.20 & 0.02 & -0.11 & -0.29 & -0.34 & -0.16 & -0.04 & -0.01 \\
0.01 & 0.03 & 0.22 & 0.02 & -0.23 & -0.02 & -0.39 & -0.33 & -0.11 & 0.15 \\
-0.16 & 0.25 & -0.24 & 0.09 & 0.06 & -0.04 & -0.13 & -0.16 & 0.18 & 0.08 \\
-0.26 & 0.21 & 0.15 & -0.16 & 0.04 & -0.04 & 0.03 & -0.01 & -0.09 & 0.11 \\
-0.12 & 0.10 & -0.06 & -0.23 & 0.13 & 0.09 & -0.07 & 0.20 & -0.13 & 0.30 \\
0.35 & -0.01 & 0.15 & -0.04 & 0.06 & -0.02 & -0.22 & -0.19 & -0.01 & -0.12 \\
-0.16 & 0.10 & -0.22 & -0.16 & 0.05 & -0.02 & -0.02 & -0.09 & 0.10 & 0.06
\end{pmatrix}$$

$$\begin{pmatrix}
0.21 & 0.17 & -0.14 & 0.14 & 0.01 & -0.16 & -0.26 & -0.12 & 0.35 & -0.16 \\
0.06 & 0.05 & 0.24 & -0.18 & 0.03 & 0.25 & 0.21 & 0.10 & -0.01 & 0.10 \\
0.04 & -0.29 & 0.23 & 0.20 & 0.22 & -0.24 & 0.15 & -0.06 & 0.15 & -0.22 \\
0.05 & 0.06 & -0.15 & 0.02 & 0.02 & 0.09 & -0.16 & -0.23 & -0.04 & -0.16 \\
0.01 & 0.39 & -0.07 & -0.11 & -0.23 & 0.06 & 0.04 & 0.13 & 0.06 & 0.05 \\
0.13 & -0.05 & 0.28 & -0.29 & -0.02 & -0.04 & -0.04 & 0.09 & -0.02 & -0.02 \\
-0.02 & 0.14 & 0.20 & -0.34 & -0.39 & -0.13 & 0.03 & -0.07 & -0.22 & -0.02 \\
0.16 & -0.22 & 0.08 & -0.16 & -0.33 & -0.16 & -0.01 & 0.20 & -0.19 & -0.09 \\
-0.02 & -0.14 & -0.01 & -0.04 & -0.11 & 0.18 & -0.09 & -0.13 & -0.01 & 0.10 \\
0.14 & -0.00 & 0.08 & -0.01 & 0.15 & 0.08 & 0.11 & 0.30 & -0.12 & 0.06 \\
1.00 & -0.22 & -0.04 & 0.10 & 0.13 & 0.19 & -0.22 & 0.05 & 0.07 & -0.20 \\
-0.22 & 1.00 & -0.13 & 0.05 & -0.08 & -0.03 & 0.14 & 0.02 & -0.01 & 0.13 \\
-0.04 & -0.13 & 1.00 & -0.29 & -0.00 & -0.23 & 0.14 & 0.06 & -0.13 & -0.15 \\
0.10 & 0.05 & -0.29 & 1.00 & 0.45 & 0.14 & -0.06 & -0.03 & 0.29 & 0.11 \\
0.13 & -0.08 & -0.00 & 0.45 & 1.00 & -0.02 & -0.03 & 0.12 & 0.32 & -0.02 \\
0.19 & -0.03 & -0.23 & 0.14 & -0.02 & 1.00 & 0.03 & -0.16 & -0.07 & 0.26 \\
-0.22 & 0.14 & 0.14 & -0.06 & -0.03 & 0.03 & 1.00 & 0.00 & -0.04 & 0.21 \\
0.05 & 0.02 & 0.06 & -0.03 & 0.12 & -0.16 & 0.00 & 1.00 & 0.08 & 0.08 \\
0.07 & -0.01 & -0.13 & 0.29 & 0.32 & -0.07 & -0.04 & 0.08 & 1.00 & -0.02 \\
-0.20 & 0.13 & -0.15 & 0.11 & -0.02 & 0.26 & 0.21 & 0.08 & -0.02 & 1.00
\end{pmatrix}$$

The platform-optimal or closest algorithm, $\mathscr{P}$, ranks users according to their covariances to user i. The ranking is 19, 4, 11, 5, 12, 7, 14, 9, 6, 8, 15, 10, 18, 13, 3, 16, 20, 2, and 17. For the specific configuration of this example, the expected engagement can

be calculated to 10.8 (approximated to 11), and hence user i will learn the messages of the first 11 users in the ranking. We represent this in Figure 3, where user 1 is linked to those whose messages will be read. In turn, the user-optimal algorithm, $\mathscr{U}$, ranks users according to their overall contribution to user i's utility. The ranking is 19, 7, 4, 5, 14, 11, 9, 12, 6, 8, 15, 3, 20, 18, 10, 16, 13, 2. The expected engagement is 10.4 (approximated to 10), and hence user i observes the messages of the first 10 users in such ranking, as represented in Figure 5. Crucially, even though the order provided by each of these two algorithms is different, the set of users appearing in the realized feeds is almost the same (note that the only difference is that the feed under $\mathscr{P}$ includes user 15). Finally, the reverse-chronological algorithm $\mathscr{R}$ randomly ranks users as 14, 20, 4, 17, 18, 9, 7, 15, 6, 16, 11, 19, 12, 10, 3, 13, 2, 8, 5, yields expected engagement 6.8 (approximated to 7), and induces a feed represented in Figure 4.