RECOMMENDER SYSTEMS AS MECHANISMS FOR SOCIAL LEARNING*

YEON-KOO CHE AND JOHANNES HÖRNER

This article studies how a recommender system may incentivize users to learn about a product collaboratively. To improve the incentives for early exploration, the optimal design trades off fully transparent disclosure by selectively overrecommending the product (or "spamming") to a fraction of users. Under the optimal scheme, the designer spams very little on a product immediately after its release but gradually increases its frequency; she stops it altogether when she becomes sufficiently pessimistic about the product. The recommender's product research and intrinsic/naive users "seed" incentives for user exploration and determine the speed and trajectory of social learning. Potential applications for various Internet recommendation platforms and implications for review/ratings inflation are discussed. *JEL Codes:* D82, D83, M52.

I. INTRODUCTION

Most of our choices rely on the recommendations of others. Whether selecting movies, picking stocks, choosing hotels, or shopping online, shared experiences can help us make better decisions. Internet platforms are increasingly organizing user recommendations for various products. Amazon (books) and Netflix (movies) are two well-known recommenders, but there is a recommender for almost any "experience" good: Pandora for music, Google News for news headlines, Yelp for restaurants, TripAdvisor for hotels, RateMD for doctors, and RateMyProfessors for professors, to name just a few. Search engines such as Google, Bing, and Yahoo crowdsource users' search experiences and "recommend" relevant websites to other users. Social media such as Facebook

*This article was previously circulated under the title "Optimal Design for Social Learning." We thank numerous audiences and many colleagues for their comments. We thank our research assistants, Han Huynh, Konstantin Shamruk, and Xingye Wu, as well our discussants, Alex Gershkov and Frédéric Koessler. Yeon-Koo Che was supported by Global Research Network program through the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF Project number 2016S1A2A2912564). Johannes Hörner thanks the National Science Foundation (grant SES-1530608 and SES-1530528) for financial support. Corresponding author: Yale University, 30 Hillhouse Ave., New Haven, CT 06520, USA, and Toulouse School of Economics, 21 Allee de Brienne, 31015 Toulouse, France, Email: joh.horner@gmail.com.

The Quarterly Journal of Economics (2018), 871–925. doi:10.1093/qje/qjx044.
Advance Access publication on December 20, 2017.

and LinkedIn do the same for another quintessential experience good—friends.

These platforms play a dual role in social learning—discovering new information ("exploration") and disseminating it to users ("exploitation"). How the latter role can be performed effectively through methods such as collaborative filtering has received ample attention and remains a primary challenge for recommender systems.[1] The former role has received comparatively less attention. However, it is also important. Many product titles (e.g., of songs, movies, or books) are fairly niche and ex ante unappealing,[2] so that few people will find them worthwhile to explore on their own even at a zero price.[3] Nonetheless, exploring these products can be socially valuable, because some of them are ultimately worthy of consumption and their "discovery" will benefit subsequent users. However, the lack of sufficient initial discovery, known as the "cold start" problem, often leads to the demise of worthy products and startups. The challenge lies in the fact that users, on whom the recommender relies for discovery, do not internalize the benefit accruing to future users.

The current article studies how a recommender may design its policy to overcome that challenge. Specifically, we consider a model in which a designer (e.g., a platform) decides whether to recommend a product (e.g., a movie, a song, or breaking news) to users who arrive continuously after the product's release. The designer's recommendation is based on the information she collects from internal research or user feedback, both of which take

1. Recommenders employ a variety of algorithms to predict users' preferences based on their consumption histories, their demographic profiles and their search and click behaviors. The Netflix Prize of 2006–2010 illustrates the challenge associated with finding an efficient algorithm to make accurate predictions (see https://en.wikipedia.org/wiki/Netflix_Prize). See Schafer, Konstan, and Riedl (1999) and Bergemann and Ozmen (2006) for stylized descriptions of collaborative filtering.

2. Obscure titles become increasingly significant due to the proliferation of self-production. For instance, self-publishing, once considered vanity publishing, has expanded dramatically in recent years with the availability of print on demand and e-books. Bowker Market Research estimates that more than 300,000 self-published titles were issued in 2011 (*New York Times*, "The Best Book Review Money Can Buy," August 25, 2012). While still in its infancy, 3D printing and similar technologies anticipate a future that will feature an even greater increase in self-manufactured products.

3. For subscribers of the platform, the marginal price of streaming titles is essentially zero. However, users face the nonzero opportunity cost of forgoing other valuable activities, including streaming other better-known titles.

the form of breakthrough news: when the product is of high quality, the designer receives a signal confirming it ("good news") at a Poisson rate proportional to the number of users having consumed that product. We then identify an optimal recommendation policy, assuming that the designer maximizes user welfare and has full commitment power. We later justify these features in the context of Internet recommender systems.

It is intuitive and shown to be optimal that a product must be recommended to all subsequent users if the designer receives good news. The key question is whether and to what extent she (the designer) should recommend the product even when no news has arrived. This latter type of recommendation, called "spam,"[4] is clearly undesirable from an exploitation standpoint, but it can be desirable from an exploration standpoint. Indeed, ignoring user incentives, a classical prescription from the bandit literature calls for a blast of spam, or full user exploration, even against the interest of the user as long as the designer's belief about the product is above a certain threshold. However, such a policy will not work, for users will ignore the recommendation and refuse to explore if their prior belief is unfavorable. For the policy to be incentive compatible, the users' beliefs must be sufficiently favorable toward the product when called on to explore that product. The designer can create such beliefs by sending spam at a suitably chosen rate.

The exact form of spam and its optimal magnitude depends on the specific context. We explore three different realistic contexts. The first is when the designer can privately send a personalized recommendation of a product to each agent. In this case, the optimal policy selects a fraction of randomly selected agents to receive spam. In the second setting of interest, the designer's recommendations become publicly observable to all agents arriving thereafter. In this case, spam takes the form of a once-and-for-all recommendation campaign (or product ratings), which lasts for a certain time. The third is when the designer privately recommends horizontally differentiated products to agents with heterogeneous preferences. In this setting, the optimal policy determines the breadth of agent types receiving spam on either product.

For each of these settings, the optimal recommender policy involves hump-shaped dynamics. In particular, the optimal recommendation must "start small." Immediately after the release

4. Throughout, the term "spam" means an unwarranted recommendation, more precisely a recommendation of a product that has yet to be found worthy of said recommendation.

of a product, few, if any, will have explored the product, so recommending this newly released product is likely to be met with skepticism. Therefore, the recommender can spam very little in the early stages, and learning occurs at a slow pace. Accordingly, the recommender initially selects a small fraction of agents for personalized recommendations (in the private recommendation context), a low probability of triggering the once-and-for-all recommendation campaign (in the public recommendation context), and a narrow bandwidth of agents for product matching (in the heterogeneous preferences context). Over time, however, the recommendation becomes credible, so the designer selects a higher fraction of agents, a higher probability, or an increased breadth of agents for spam, depending on the contexts. Consequently, the pace of learning accelerates. In the first two contexts, the absence of news eventually makes the recommender sufficiently pessimistic about the product's quality. At that point, the designer abandons spam altogether.

The main insights and findings are shown to be robust to a number of extensions: vertical heterogeneity in user preferences, user uncertainty about the product release time, the presence of behavioral types that follow the designer's recommendations without any skepticism, the designer's investment in learning, and a more general signal structure.

Although our analysis is primarily normative, it has potential applications in several aspects of Internet platforms. Search engines determine the display order of web pages based on algorithms such as PageRank, which rely heavily on users' past search activities. Hence, they are susceptible to the entrenchment problem: the pages that users found relevant in the past are ranked highly and are thus displayed more prominently, attracting more visits and reinforcing their prominent rank, whereas newly created sites are neglected regardless of their relevance. One suggested remedy is the random shuffling of the display order to elevate the visibility of underexposed and newly created pages (Pandey et al. 2005). This is indeed a form of spam, as suggested by our optimal policy. Although our analysis is consistent with this remedy, it also highlights the incentive constraint: the frequency of random shuffling must be kept at a low level so that the searchers enlisted to "explore" these untested sites will find them ex ante credible.

A similar concern about user incentives arises when newly launched social media platforms try to recruit users via

unsolicited "user-initiated" invitations. Some social media sites are known to have "blasted" invitations to a mass of unsuspecting individuals, often unbeknownst to the inviters through some dubious form of consent.[5] Our theory cautions against such aggressive spam campaigns, for they will undermine the credibility of the recommender. For invitees to perceive that their acquaintances have initiated unsolicited invitations, the frequency of invitations must be kept at a credible level.[6]

A similar implication can be drawn for reviews/ratings inflation, which is common in many online purchase sites.[7] Ratings are often inflated by sellers—as opposed to the platforms—who have every reason to promote their products, even against the interests of consumers.[8] However, platforms have instruments at their disposal to control the degree of ratings inflation, such as filters that detect false reviews, requiring users to verify their purchase before posting reviews and allowing them to vote for "helpful" reviews.[9] Our analysis suggests that some degree of inflation is desirable from the perspective of user exploration, but keeping inflation under control is in the best interest of the platform/recommender to maintain its credibility.

Finally, our article highlights the role of internal research conducted by the recommender. An example of internal research is Pandora's music genome project, which famously hires musicologists to classify songs according to some 450 attributes. While such research is costly, it can provide significant benefits. As we show later, internal research serves as a substitute for costly user

5. Indeed, users may turn against such social media sites. A class action suit filed under *Perkins v. LinkedIn* alleged that LinkedIn's "Add Connections" feature allowed the platform to scrape users' email address books and send out multiple messages reminding recipients to join these users' personal networks. LinkedIn settled the suit for $13 million. See "LinkedIn will pay $13M for sending those awful emails," *Fortune*, October 5, 2015.

6. Note, however, that Section VII.C suggests that such a policy may be optimal for platforms facing a large fraction of naive invitees.

7. Jindal and Liu (2008) find that 60% of the reviews on Amazon have a rating of 5.0, and approximately 45% products and 59% of members have an average rating of 5.

8. Luca and Zervas (2016) suggest that as much as 16% of Yelp reviews are suspected to be fraudulent.

9. Mayzlin, Dover, and Chevalier (2014) find that Expedia's requirement that a reviewer verify her stay to review a hotel resulted in fewer false reviews at Expedia compared with TripAdvisor, which has no such requirement.

exploration and enhances the recommender's credibility and helps speed/scale up users' exploration.

The rest of the article is organized as follows. Section II uses a simple example to illustrate the main idea of the paper. Section III introduces a model. Sections IV, V, and VI characterize the optimal policy in three different contexts, serving as the main analysis. Section VII extends the results in a variety of ways. Section VIII describes related literature. Section IX concludes.

II. Illustrative Example

We begin with a simple example that highlights the main idea of incentivized exploration. Suppose a product, say a movie, is released at time $t = 0$, and a unit mass of agents arrive at each time $t = 1, 2$. The quality of the movie is either "good" ($\omega = 1$), in which case the movie yields a surplus of 1 to each agent, or "bad" ($\omega = 0$), in which case it yields a surplus of 0. The quality of the movie is unknown at the time of its release, with prior $p^0 := \Pr[\omega = 1] \in [0, 1]$. Watching the movie costs each agent $c \in (p^0, 1)$; thus, without further information, the agents would never watch the movie.

At time $t = 0$, the designer receives a signal $\sigma \in \{g, n\}$ (from its marketing research, for example) about the quality of the movie with probabilities:

$$\Pr[\sigma = g \mid \omega] = \begin{cases} \rho_0 & \text{if } \omega = 1; \\ 0 & \text{if } \omega = 0, \end{cases}$$

and $\Pr[\sigma = n \mid \omega] = 1 - \Pr[\sigma = g \mid \omega]$. In other words, the designer receives good news only when the movie is good; but she also may receive no news (even) when the movie is good.[10]

Suppose the designer has received no news at $t = 0$ but a fraction α of agents watch the movie at $t = 1$. Then, the designer

10 Thus, it follows that the designer's posterior at time $t = 1$ on $\omega = 1$ is 1 with a probability of $\rho_0 p^0$ (in the event that she receives good news) and

$$p_1 = \frac{(1 - \rho_0)p^0}{(1 - \rho_0)p^0 + 1 - p^0},$$

with a probability of $1 - \rho_0 p^0$ (in the event that she receives no news).

again receives conclusively good news with probability:

$$\Pr[\sigma = g \mid \omega] = \begin{cases} \alpha & \text{if } \omega = 1; \\ 0 & \text{if } \omega = 0. \end{cases}$$

The feature that the signal becomes more informative with a higher fraction α of agents experimenting at $t = 1$ captures the learning benefit that they confer to the $t = 2$ agents.

The designer commits to a recommendation policy that maximizes social welfare. Specifically, she recommends the movie to a fraction of agents in each period based on her information at that point in time.[11] The designer discounts the welfare in period $t = 2$ by a factor $\delta \in (0, 1)$.

The designer's optimal policy is then as follows. First, the designer is truthful at time $t = 2$, as lying can only reduce welfare and can never improve the incentive for experimentation at $t = 1$. Consider now time $t = 1$. If good news has arrived, the designer would recommend the movie to all agents. Suppose no news has been received but the designer nevertheless recommends—or "spams"—to a fraction α of the agents. The agents receiving the recommendation cannot determine whether the recommendation is genuine or spam; instead, they would form a posterior:

$$P_1(\alpha) := \frac{\rho_0 p^0 + \alpha p^0 (1 - \rho_0)}{\rho_0 p^0 + (1 - \rho_0 p^0)\alpha}.$$

If the designer spams to all agents (*i.e.*, $\alpha = 1$), then they will find the recommendation completely uninformative, and hence $P_1(1) = p^0$. Since $p^0 < c$, they would never watch the movie. By contrast, if the designer spams rarely (*i.e.*, $\alpha \simeq 0$), then $P_1(\alpha) \simeq 1$, *i.e.*, they will be almost certain that the recommendation is genuine. Naturally, the agents receiving a recommendation will

11 The designer would not gain from a stochastic recommendation policy. To see this, compare two choices: i) the designer recommends the movie to a fraction α of agents, and ii) the designer recommends it to all agents with probability α. For agents in $t = 1$, the two options are the same in terms of welfare and thus in terms of incentives. For agents in $t = 2$, the good news is learned with probability $p^0(\rho_0 + (1 - \rho_0)\alpha)$ in either way. This equivalence means that public recommendation entails no loss. This equivalence holds only because no experimentation is prescribed in $t = 2$, and breaks down in our general model where experimentation is prescribed over a duration of time, with the optimal spammed fraction featuring temporal correlation.

definitely watch the movie in this case. Because the recommendation is more credible the less the designer spams, $P_i(\alpha)$ is decreasing in α. In particular, there is a maximal fraction $\hat{\alpha} =: \frac{(1-c)\rho_0 p^0}{c(1-\rho_0 p^0)-p^0(1-\rho_0)}$ of agents who can be induced to experiment.

Social welfare,

$$W(\alpha) := p^0(\rho_0 + (1-\rho_0)\alpha)(1-c)(1+\delta) - \alpha(1-p^0)c,$$

consists of the benefit from the good movie being recommended (the first term) and the loss borne by the $t = 1$ agents from a bad movie being recommended (the second term). In particular, it also includes the benefit experimentation by the $t = 1$ agents confers to the $t = 2$ agents (captured by the term $p^0(1-\rho_0)\alpha(1-c)\delta$).

The optimal policy is to spam up to $\hat{\alpha}$, if W is increasing in α, *i.e.*, if the social value of experimentation at date 1 justifies the cost:

$$p^0 \geqslant \hat{p}^0 := \frac{c}{(1-\rho_0)(1+\delta)(1-c)+c}. \tag{1}$$

Note that the right-hand side is strictly less than c when $\rho_0 < \frac{\delta}{1+\delta}$. In that case, if $p^0 \in (\hat{p}^0, c)$, the designer will spam some of the agents at $t = 1$ to consume against their myopic interest.

III. Model

Our model generalizes the example in terms of its timing and information structure. A product is released at time $t = 0$, and, for each time $t \in [0, \infty)$, a unit mass of agents arrives and decides whether to consume the product. The agents are assumed to be myopic and (in the baseline model) ex ante homogeneous. Consuming the good costs each agent $c \in (0, 1)$, which can be the opportunity cost of time spent or the price charged. The product is either "good," in which case each agent derives the (expected) surplus of 1, or "bad," in which case the agent derives the (expected) surplus of 0. The quality of a product is a priori uncertain but may be revealed over time.[12] At time $t = 0$, the probability of

12. The agents' preferences may involve an idiosyncratic component that is realized ex post after consuming the product; the quality then captures only their common preference component. The presence of an idiosyncratic preference component does not affect the analysis because each agent must decide based on the expected surplus that he will derive from his consumption of the product.

the product being good, or simply "the prior," is p^0. We consider all values of the prior, although the most interesting case will be $p^0 \in (0, c)$, which makes nonconsumption myopically optimal.

Agents do not observe previous agents' decisions or their experiences. Instead, the *designer* mediates social learning by collecting information from past agents or her own research and disclosing all or part of that information to the arriving agents.

III.A. The Designer's Signal

The designer receives information about the product in the form of breakthrough news. Suppose a flow of size $\alpha \geqslant 0$ consumes the product over some time interval $[t, t + \mathrm{d}t)$. Then, the designer learns during this time interval that the product is "good" with probability $\lambda(\rho + \alpha)\mathrm{d}t$ if the product is good ($\omega = 1$) and with 0 probability if the product is not good ($\omega = 0$), where $\lambda > 0$ measures the rate at which user consumption produces breakthrough news and $\rho > 0$ is the rate at which the designer obtains the information regardless of the agents' behaviors.[13] In a reduced form, the signal structure describes the extent to which consumers who explore a product contribute to a recommender's learning about that product.[14] The background learning, parameterized by ρ, can arise from the designer's own product research (e.g., Pandora's music genome project). It may also arise from a flow of fans who do not mind exploring the product; that is, they face a zero cost of exploration. The designer begins with the same prior p^0 as the agents, and the agents do not have access to free learning.

III.B. The Designer's Recommendation Policy

Based on the information received, the designer provides feedback to the agents. Since agents' decisions are binary, without loss of generality, the designer simply decides whether to recommend the product. The designer commits to the following policy: at time t, she recommends the product to a fraction $\gamma_t \in [0, 1]$ of (randomly selected) agents if she learns that the product is good, and

13. Section VII.E extends our model to allow for (conclusively) bad news and (conclusively) good news. Our qualitative results continue to hold in this more general environment.

14. Avery, Resnick, and Zeckhauser (1999) and Miller, Resnick, and Zeckhauser (2004) take a structural approach to elicit honest reviews via monetary incentives.

she recommends the product to or *spams* a fraction $\alpha_t \in [0, 1]$ if no news has arrived by t. The recommendation is private in the sense that each agent observes only the recommendation made to him; that is, he does not observe recommendations made to the others in the past or present. (We consider public recommendations in Section V.) We assume that the designer maximizes the intertemporal net surplus of the agents, discounted at rate $r > 0$, over the (measurable) functions (α, γ), where $\alpha := \{\alpha_t\}_{t \geqslant 0}$ and $\gamma := \{\gamma_t\}_{t \geqslant 0}$.

III.C. The Designer's Beliefs

The designer's information at time $t \geqslant 0$ is succinctly summarized by the *designer's belief* about $\omega = 1$, which is 1 if good news has arrived by that time or some $p_t \in [0, 1]$ if no news has arrived by that time. The "no news" posterior, or simply *posterior* p_t, must evolve according to Bayes's rule. Specifically, suppose for time interval $[t, t + \mathrm{d}t)$, (total) exploration occurs at rate $\mu_t = \rho + \alpha_t$, where ρ is background learning and α_t is the flow of agents exploring at time t. If no news has arrived by $t + \mathrm{d}t$, then the designer's updated posterior at time $t + \mathrm{d}t$ must be

$$p_t + \mathrm{d}p_t = \frac{p_t(1 - \lambda(\rho + \alpha_t)\,\mathrm{d}t)}{p_t(1 - \lambda(\rho + \alpha_t)\,\mathrm{d}t) + 1 - p_t}.$$

Rearranging and simplifying, the posterior must follow the law of motion:

$$\dot{p}_t = -\lambda(\rho + \alpha_t)p_t(1 - p_t), \tag{2}$$

with the initial value at $t = 0$ given by the prior p^0. Notably, the posterior decreases as time passes, as "no news" leads the designer to become pessimistic about the product's quality.

III.D. Agents' Beliefs and Incentives

In our model, agents do not directly observe the designer's information or her beliefs. However, they can form a rational belief about the designer's beliefs. They know that the designer's beliefs are either 1 or p_t, depending on whether good news has been received by time t. Let g_t denote the probability that the designer has received good news by time t. This probability g_t is pinned down by the martingale property, that is, that the designer's posterior

must, on average, equal the prior:

$$g_t \cdot 1 + (1 - g_t)p_t = p^0. \tag{3}$$

Notably, g_t rises as p_t falls; that is, the agents find it increasingly probable that news has arrived as time progresses.

In addition, for the policy (α, γ) to be implementable, the agents must have an incentive to follow the recommendation.[15] Because the exact circumstances surrounding the recommendation (whether the agents receive the recommendation because of good news or despite no news) are kept hidden from the agents, their incentives for following the recommendation depend on their posterior regarding the designer's information:

$$q_t(p_t) := \frac{g_t\gamma_t + (1 - g_t)\alpha_t p_t}{g_t\gamma_t + (1 - g_t)\alpha_t}.$$

The denominator accounts for the probability that an agent will be recommended to consume the product, which occurs if either the designer receives good news (the first term) or the designer receives no news but selects the agent for spam (the second term); the numerator accounts for the probability that the agent receives a recommendation when the product is good. An agent will have an incentive to consume the product if and only if the posterior that the product is good is no less than the cost:

$$q_t(p_t) \geqslant c. \tag{4}$$

III.E. The Designer's Objective and Benchmarks

The designer chooses a (measurable) policy (α, γ) to maximize social welfare, namely,

$$\mathcal{W}(\alpha, \gamma) := \int_{t \geqslant 0} e^{-rt} g_t \gamma_t (1 - c)\, \mathrm{d}t + \int_{t \geqslant 0} e^{-rt}(1 - g_t)\alpha_t (p_t - c)\, \mathrm{d}t,$$

where (p_t, g_t) must follow the required laws of motion: equations (2) and (3).[16] Welfare consists of the discounted value of

15. There is also an incentive constraint for the agents not to consume the product when the designer does not recommend it. Because this constraint will not be binding throughout—because the designer typically desires more exploration than the agents—we ignore it.

16. We allow the designer to randomize over (α, γ), although our proof of Proposition 1 in Appendix A shows that such a policy is never optimal.

consumption—$1 - c$ in the event of good news and $p_t - c$ in the event of no news—for those the designer recommends to consume the product.

To facilitate the characterization of the optimal policy, it is useful to consider the following benchmarks:

i. No social learning: the agents receive no information from the designer; hence, they decide solely based on the prior p^0. When $p^0 < c$, no agent consumes.
ii. Full transparency: the designer truthfully discloses her information—or her beliefs—to the agents. Formally, full disclosure is implemented through the policy of $\gamma_t \equiv 1$ and $\alpha_t = 1_{\{p_t \geq c\}}$, which fulfills the exploitation goal of the designer, maximizing the short-term welfare of the agents.
iii. First-best policy: the designer optimizes her policy (α, γ) to maximize $\mathcal{W}$ subject to equations (2) and (3). By ignoring the incentive constraint (4), the first-best captures the classic trade-off between exploitation and exploration, as studied in the bandit literature (see Rothschild 1974; Gittins, Glazebrook, and Weber 2011). Comparing first-best and full transparency thus highlights the designer's exploration goal.
iv. Second-best policy: in this regime, the focus of our study, the designer optimizes her policy (α, γ) to maximize $\mathcal{W}$ subject to equations (2), (3), and (4). Comparing second-best and first-best policies highlights the role of incentives.

III.F. Applicability of the Model

The salient features of our model conform to Internet platforms that recommend products such as movies, songs, and news headlines. First, the assumption that the recommender is benevolent is sensible for platforms that derive revenue from subscription fees (e.g., Netflix and Pandora) or advertising (e.g., Hulu), as maximizing subscriptions leads them to maximize the gross welfare of users.[17]

17. An Internet platform earning ad revenue from user streaming may be biased toward excessive recommendations. Even such a platform recognizes that recommending bad content will result in users leaving the platform, and it will try to refrain from excessive recommendations.

Second, the assumption of recommender commitment power is plausible if the recommender can resist the temptation of over-recommending a product (to a level that would result in users ignoring its recommendations). A recommender can achieve commitment power by building a good reputation. If a recommender handles multiple titles, a simple way to build reputation is to limit the number of titles he or she recommends;[18] users may then "punish" deviation by ignoring future recommendations. Another way to build reputation is by hardwiring a recommendation technology. For example, Pandora's music genome project puts a severe bottleneck on the number of tunes that can be recommended.[19]

Third, our model does not consider monetary incentives for exploration. Indeed, monetary incentives are rarely used to compensate for the online streaming of movies, music, and news items or for user feedback on these items.[20] Monetary incentives are unreliable if the quality of exploration is difficult to verify. For instance, paying users to stream a movie or a song or to post a review does not necessarily elicit genuine exploration. Even worse, monetary incentives may lead to a biased reviewer pool and undermine accurate learning.

Finally, a central feature of our model is "gradual" user feedback, which makes social learning nontrivial. This feature may result from noise in the reviews due to unobserved heterogeneity

18. If the recommender handles many products that are, say, identically distributed with varying release times, the optimal policy will involve recommending a constant fraction of the products each time. Netflix, for instance, used to recommend 10 movies to a user, and it currently presents a "row" of recommended movies for each genre.

19. An industry observer comments that "the decoding process typically takes about 20 minutes per song (longer for dense rap lyrics, five minutes for death metal) and Westergren points out 'Ironically, I found over the years that the fact that we couldn't go fast was a big advantage. . . . The problem that needs solving for music is not giving people access to 2.5 million songs. The trick is choosing wisely'" (Linda Tischler, "Algorhythm and Blues," *Fast Company*, December 1, 2005; http://www.fastcompany.com/54817/algorhythm-and-blues).

20. Attempts made in this regard have been limited in scope. For instance, the Amazon Vine Program rewards selected reviewers with free products, and LaFourchette.com grants discounts for (verified) diners who write reviews and make reservations via their site. See Avery, Resnick, and Zeckhauser (1999) and Miller, Resnick, and Zeckhauser (2004), who study the design of monetary incentives that encourage users to share product evaluations.

in preferences or from infrequent user feedback, which is particularly the case with headline-curation and song-selection sites.[21]

IV. Optimal Recommendation Policy

We now characterize the first-best and second-best policies. We first observe that in both cases, the designer should always disclose the good news immediately; that is, $\gamma_t \equiv 1$. This follows from the fact that raising the value of γ_t can only increase the value of objective $\mathcal{W}$ and relax equation (4) without affecting any other constraints. We thus fix $\gamma_t \equiv 1$ throughout and focus on the designer's optimal spam policy α.

Next, by using equation (3) and $\gamma_t = 1$, the incentive constraint (4) simplifies to:

$$\alpha_t \leqslant \hat{\alpha}(p_t) := \min\left\{1, \frac{(1-c)(p^0 - p_t)}{(1-p^0)(c-p_t)}\right\} \tag{5}$$

if $p_t < c$ and $\hat{\alpha}(p_t) := 1$ if $p_t \geqslant c$. In words, $\hat{\alpha}(p_t)$ is the maximum spam that the designer can send, subject to the posterior $q_t(p_t)$ of the recommended agents being no less than the cost c. We thus interpret $\hat{\alpha}(p_t)$ as the designer's spamming capacity.

The capacity depends on the prior p^0. If $p^0 \geqslant c$, then the agents have myopic incentives to explore, even at the prior. From then on, the designer can keep the agents from updating their beliefs by simply spamming all agents, inducing full exploration at $\hat{\alpha}(p_t) = 1$ for all p_t.[22] Therefore, equation (4) is never binding in this case.

By contrast, if $p^0 < c$, the constraint is binding. In this case, it is optimal to set $\alpha_t = \hat{\alpha}(p_t)$, which keeps the posterior q_t of the recommended agents equal to c. More important, $\hat{\alpha}(p_t) < 1$ in this case, so not all agents are spammed. Intuitively, if the designer spams all agents (i.e., $\alpha = 1$), they will find the recommendation completely uninformative; therefore, their posterior equals p^0. Since $p^0 < c$, they will never consume the product. By

21. Due to breakthrough news, the mix of news items changes rapidly, making it difficult for users to send feedback and for the platform to adjust its selection based on their feedback in real time. Likewise, a significant number of Pandora users use the service while driving or working, which limits their ability to send feedback ("thumbs up" or "thumbs down").

22. Of course, this is possible because agents are not told whether the recommendation is the result of news or simply spam. Formally, the martingale property implies that $q_t(p_t) = p^0$ if $\alpha = 1$.

contrast, if the designer rarely spams (i.e., $\alpha \simeq 0$), then the posterior of the recommended agents will be close to 1; that is, they will be almost certain that the recommendation is genuine. Naturally, there is an interior level of spam that will satisfy incentive compatibility. The spamming capacity $\hat{\alpha}(p_t)$ is initially zero and increases gradually over time. Immediately after the product's release, the designer has nearly no ability to spam because good news never arrives instantaneously and the agents' prior is unfavorable. Over time, however, $\hat{\alpha}(p_t)$ increases because even when no news is received and p_t falls as a result, the arrival of good news becomes increasingly probable. The designer can thus build her credibility and expand her capacity to spam as time progresses.

In essence, spamming "pools" recommendations across two very different circumstances: when good news has arrived, on the one hand, and when no news has arrived, on the other. Although the agents in the latter case will never knowingly follow the recommendation, pooling the two circumstances for recommendations enables the designer to incentivize the agents to explore—as long as the recommendation in the latter circumstance is kept sufficiently infrequent/improbable. Because the agents do not internalize the social benefits of exploration, spamming becomes a useful tool for the designer's second-best policy. We next characterize the optimal recommendation policy.

PROPOSITION 1.

(i) The first-best policy prescribes exploration

$$\alpha^{FB}(p_t) = \begin{cases} 1 & \text{if } p_t \geqslant p^*; \\ 0 & \text{if } p_t < p^*, \end{cases}$$

where

$$p^* := c\left(1 - \frac{rv}{\rho + r(v + \frac{1}{\lambda})}\right),$$

and $v := \frac{1-c}{r}$ denotes the continuation payoff on the arrival of good news.

(ii) The second-best policy prescribes exploration at

$$\alpha^{SB}(p_t) = \begin{cases} \hat{\alpha}(p_t) & \text{if } p_t \geqslant p^*; \\ 0 & \text{if } p_t < p^*. \end{cases}$$

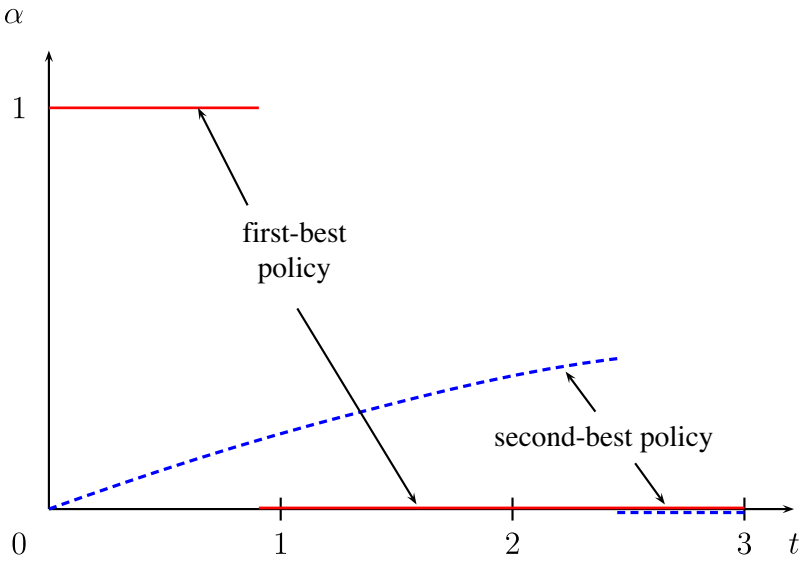

FIGURE I

Path of α for $(c, \rho, p^0, r, \lambda) = (\frac{2}{3}, \frac{1}{4}, \frac{1}{2}, \frac{1}{10}, \frac{4}{5})$

(iii) If $p^0 \geqslant c$, then the second-best policy implements the first-best policy, and if $p^0 < c$, then the second-best policy results in slower exploration/learning than the first-best policy. Whenever $p^0 > p^*$, the second-best policy induces more exploration/learning than both no social learning and full transparency.

The first-best and second-best policies have a cutoff structure. They induce maximal feasible exploration, which equals 1 under the first-best policy and the spamming capacity $\hat{\alpha}$ under the second-best policy—as long as the designer's posterior remains above the threshold level p^*. Otherwise, no exploration is chosen. The optimal policies induce interesting learning trajectories, which are depicted in Figure I for the case of $p^0 < c$.

The optimality of the cutoff policy and the associated cutoff can be explained by the main trade-off the designer faces for any given belief p:

$$\underbrace{\lambda p v \left(\frac{1}{\frac{\lambda \rho}{r} + 1} \right)}_{\text{value of exploration}} - \underbrace{c - p}_{\substack{\text{cost of} \\ \text{exploration}}} . \tag{6}$$

To understand the trade-off, suppose that the designer induces an additional unit of exploration at p, which entails flow costs for the exploring agents (the second term) but yields benefits (the first term). The benefits are explained as follows: with probability p, the product is good, and exploration will reveal this information at rate λ, which will enable the future generation of agents to collect the benefit of $v = \frac{1-c}{r}$. This benefit is discounted by the rate $\frac{1}{\frac{\lambda p}{r}+1}$ at which the good news will be learned through background learning, even with no exploration. Note that the benefits and the costs are the same under the first-best and second-best policies.[23] Hence, the optimal cutoff p^* (which equates them) is the same.

If $p^0 \geqslant c$, the designer can implement the first-best policy by simply spamming all agents as long as $p_t \geqslant p^*$. The agents comply with the recommendation because their belief is "frozen" at $p^0 \geqslant c$ under that policy. Admittedly, informational externalities are not particularly severe in this case because early agents will have an incentive to consume on their own. Note, however, that full transparency does not implement the first-best policy in this case, as agents will stop exploring once p_t reaches c. In other words, spamming is crucial to achieve the first-best, even in this case.

In the more interesting case with $p^0 < c$, the second-best policy cannot implement the first-best policy. In this case, the spamming constraint for the designer is binding. As seen in Figure I, spamming capacity is initially zero and increases gradually. Consequently, exploration starts very slowly and builds up gradually over time until the posterior reaches the threshold p^*, at which point the designer abandons exploration. Throughout, the exploration rate remains strictly below 1. In other words, learning is always slower under the second-best policy than under the first-best policy, even though the total exploration is the same (due to the common threshold). Since the threshold belief is the same under both regimes, the agents are encouraged to experiment longer under the second-best regime than under the first-best regime, as Figure I shows. In either case, as long as $p^0 > p^*$, the second-best policy implements higher exploration/learning than either no social learning or full transparency, outperforming each of these benchmarks.

23. In particular, the benefit of forgoing exploration, that is, relying solely on background learning, is the same under both regimes. This feature does not generalize to some extensions, as noted in Sections VII.A and VII.E.

Comparative statics reveal further implications. The values of (p^0, ρ) parameterize the severity of the cold-start problem facing the designer. The lower these values, the more severe the cold-start problem. One can see how these parameters affect optimal exploration policies and the speed of social learning.

COROLLARY 1.

(i) As p^0 increases, the optimal threshold remains unchanged in both the first-best and the second-best policies. The learning speed remains the same in the first-best policy but increases in the second-best policy.

(ii) As ρ increases, the optimal threshold p^* increases, and the total exploration decreases under both the first-best and the second-best policies. The speed of exploration remains the same in the first-best policy but increases in the second-best policy, provided that $p^0 < c$.[24]

Unlike under the first-best policy, the severity of the cold-start problem affects the rate of exploration under the second-best policy. Specifically, the more severe the cold-start problem, in the sense of (p^0, ρ) being smaller, the more difficult it is for the designer to credibly spam the agents, thereby reducing the rate of exploration the designer can induce.

In our model, background learning seeds the exploration; for example, if $\rho = 0$, the designer has no credibility, and no exploration ever takes place. This observation has certain policy implications. For example, Internet recommenders such as Pandora make costly investments to raise ρ, which can help the social learning in two ways. First, as shown by Corollary 1 (ii), such investments act as a substitute for agents' exploration.[25] This substitution helps lower the exploration costs of agents and speeds up learning in the second-best regime, particularly in the early stages when incentivizing user exploration is costly. Second, the designer investments have an additional benefit in the second-best regime: background learning makes spamming credible, which allows the designer to induce a higher level of user exploration at each t. Importantly, this effect is cumulative, or dynamically multiplying,

24. Recall that we are assuming that $\rho > 0$. If $\rho = 0$, then no exploration can be induced when $p^0 < c$.

25. Indeed, an increase in ρ raises the opportunity costs of exploration, calling for its termination at a higher threshold under both the first-best and the second-best policies.

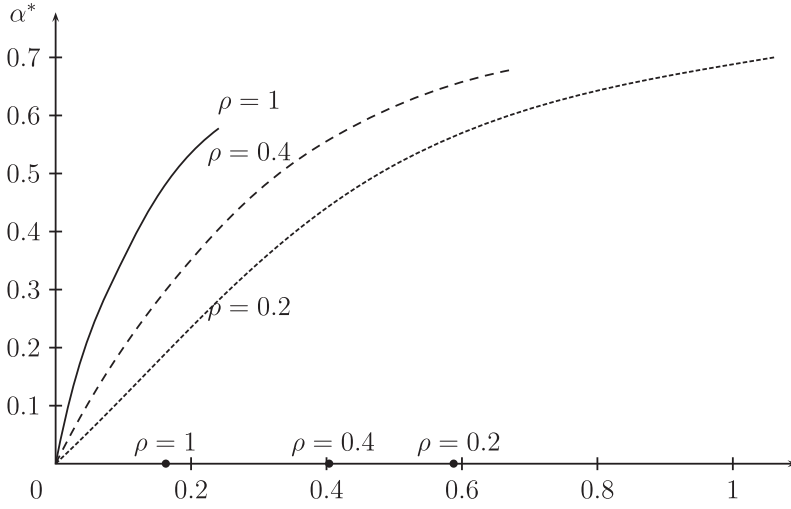

FIGURE II

(Second-Best) Spamming as a Function of ρ (Here, $(k, \ell_0, r, \lambda) = (\frac{2}{5}, \frac{1}{3}, \frac{1}{2}, 1)$)

The dots on the x-axis indicate stopping times under the first-best policy.

as increased exploration makes subsequent spamming more credible, which in turn enables further exploration. Figure II shows that this indirect effect can accelerate social learning significantly: as ρ rises, the time required to reach the threshold belief is reduced much more dramatically under the second-best policy than under the first-best policy. We will see in Section VII.D how this effect causes the designer to front-load background learning when she chooses it endogenously (at a cost).

V. PUBLIC RECOMMENDATIONS

Thus far, we have assumed that recommendations are private and personalized, meaning that agents can be kept in the dark about the recommendations that other users have received. Such private/personalized recommendations are an important part of the Internet recommender system; Netflix and Pandora make personalized recommendations based on their users' past viewing and listening histories, respectively. Likewise, search engines personalize the ranking of search items based on users' past search behaviors. However, some platforms make their recommendations public and thus commonly observable to all users. The case in

point are product ratings. Ratings provided by Amazon, Yelp, Michelin, and Parker on books, restaurants, and wines are publicly observable. In this section, we study the case in which the designer's recommendation at each time becomes publicly observable to all agents who arrive thereafter.[26]

Public recommendations are clearly not as effective as private recommendations in terms of incentivizing user exploration. Indeed, the optimal private recommendation identified earlier is not incentive compatible when made public. If only some fraction of users are recommended to explore, the action reveals the designer's private information, and users will immediately recognize that the recommendation is merely spam and thus ignore the recommendation. Hence, if the designer wishes to trigger user exploration, she must adopt a different approach. We show that although spam becomes less effective as an incentive when recommendations are public, it is still part of the optimal policy.

To focus on a nontrivial case, we assume $p^* < p^0 < c$, where p^* is the threshold belief under the first-best and second-best private recommendation policies (defined in the previous section).[27] As we show shortly, given this assumption, the designer can still induce agents to explore through a public recommendation policy, but the policy must be random.

To begin, observe first that if the designer receives news at any point in time, she will thereafter recommend the product to all agents. Plainly, the sharing of good news can only increase agents' welfare and relax their incentives, just as before.

26. In practice, users can access past ratings directly or indirectly through search engines. For instance, Amazon makes all reviews visible to users; Yelp explicitly allows users to see monthly ratings trends for each restaurant, which often span many years. Whether users can observe past and current recommendations is an important consideration for our analysis. Indeed, if only current recommendations are observable, the designer can implement optimal private recommendations, as described in Section IV, via public recommendations. For instance, to spam one out of seven users, the designer can divide each interval $[t, t + \mathrm{d}t)$ into seven equal-sized subintervals, pick one at random, and spam only those users arriving in that subinterval. This policy is clearly incentive compatible (an agent is unable to discern whether he is being targeted at random or good news has arrived) and achieves virtually the same payoff as the optimal private recommendation with an arbitrarily "fine" partitioning of time intervals.

27. If either $p^0 \geqslant c$ or $p^0 \leqslant p^*$, the first-best policy is achievable via the public recommendation policy. In the former case, the designer can spam fully until her belief reaches the threshold p^*; then, the agents do not update their beliefs, and they are therefore happy to follow the designer's recommendation. In the latter case, the first-best policy prescribes no exploration, which is trivial to implement.

Next, to see why the recommendation policy must be random, suppose that the designer commits to spamming—that is, to recommend the product to users despite having received no news—at some deterministic time t for the first time. Since the recommendation is public, all agents observe it. Because the probability of the good news arriving at time t, conditional on not having done so before, is negligible, the agents will put the entire probability weight on the recommendation being merely spam and ignore it. Hence, deterministic spam will not work. Consider the random policy described by $F(t)$, the probability that the designer starts spam by time t. Here, we heuristically derive $F(t)$, taking several features of the optimal policy as a given. Appendix B establishes these features carefully.

First, if the designer's belief falls below p^* at any point in time, assuming that no news has been received by then, the designer will stop exploration (or cease spamming). This follows from the optimal trade-off between exploitation and exploration identified earlier under the optimal (private) recommendation policy. Let t^* be the time at which the designer's posterior reaches the threshold belief p^*, provided that no agents have experimented and news has never been received.[28] Clearly, if the designer does not trigger spam by time t^*, she will not trigger spam after that time, which implies that the distribution F is supported at $[0, t^*]$. Second, once the optimal policy sends spam to all agents at some random time $t < t^*$, continuing to spam thereafter does not change the agents' beliefs; the agents have no grounds to update their beliefs. Hence, once they have incentives to explore, all subsequent agents will have the same incentive. Consequently, the optimal policy will continue recommending the product until the designer's belief falls to p^*.

Given these features, the distribution F must be chosen to incentivize users to explore when they are recommended to do so for the first time. To see how, we first obtain the agents' belief

$$q_t = \frac{p^0 e^{-\lambda\rho t}(\lambda\rho + h(t))}{p^0 e^{-\lambda\rho t}(\lambda\rho + h(t)) + (1 - p^0)h(t)},$$

on being recommended to explore for the first time, where $h(t) := \frac{f(t)}{1-F(t)}$ is the hazard rate of starting spam. This formula is

28. More precisely, $t^* = -\frac{1}{\lambda\rho} \ln\left(\frac{p^*}{1-p^*} \frac{1-p^0}{p^0}\right)$, according to equation (2) with $\alpha_t = 0$.

explained by Bayes's rule. The denominator accounts for the probability that the recommendation is made for the first time at t either because the designer receives news at time t (which occurs with probability $\lambda\rho p^0 e^{-\lambda\rho t}$) or because the random policy F triggers spam for the first time at t without having received any news (which occurs with probability $(p^0 e^{-\lambda\rho t} + 1 - p^0)h(t)$). The numerator accounts for the probability that the recommendation is made for the first time at t and that the product is good. For the agents to have incentives to explore, the posterior q_t must be no less than c, a condition that yields an upper bound on the hazard rate:

$$h(t) \leqslant \frac{\lambda\rho p^0(1-c)}{(1-p^0)(c-(1-c)e^{\lambda\rho t}) - p^0(1-c)}.$$

Among other things, this implies that the distribution F must be atomless. As is intuitive and formally shown in Appendix B, the incentive constraint is binding for the optimal policy (i.e., $q_t = c$), which gives rise to a differential equation for F, alongside the boundary condition $F(0) = 0$.[29] Its unique solution is

$$F(t) = \frac{p^0(1-c)(1-e^{-\lambda\rho t})}{(1-p^0)c - p^0(1-c)e^{-\lambda\rho t}}, \tag{7}$$

for all $t < t^*$. Since the designer never spams after t^* (when $p = p^*$ is reached), $F(t) = F(t^*)$ for $t > t^*$.

Examining F reveals various features of the optimal policy. First, as with private recommendation, the exploration under the second-best policy is single-peaked, though in a probabilistic sense. The expected exploration starts "small" (i.e., $F(t) \approx 0$ for $t \approx 0$) but accelerates over time as the designer's credibility increases (i.e., $F(t)$ is strictly increasing as t increases), and it stops altogether when p^* is reached.

While spam is part of the optimal public recommendation, its randomness makes it less effective at converting a given probability of good news into incentives for exploration, leading to a reduced level of exploration. This reduced effectiveness can be

29. As mentioned earlier, q_t remains frozen at c from then on (until exploration stops).

seen as follows:

$$F(t) = \frac{(1-c)p^0 - (1-c)p^0 e^{-\lambda\rho t}}{(1-p^0)c - (1-c)p^0 e^{-\lambda\rho t}} < \frac{(1-c)p^0 - (1-c)p_t}{(1-p^0)c - (1-c)p_t}$$
$$< \frac{(1-c)(p^0 - p_t)}{(1-p^0)(c-p_t)} = \hat{\alpha}_t,$$

where both inequalities use $p^0 < c$, and the first follows from $p_t = p^0 e^{-\int_0^t \lambda(\rho+\alpha_t)\,dt} < p^0 e^{-\lambda\rho t}$. Consequently, the rate of exploration is, on average, slower under public recommendations than under private recommendations:

PROPOSITION 2. Under the optimal public recommendation policy, the designer recommends the product at time t if good news is received by that time. If good news is not received and a recommendation is not made by time $t \leqslant t^*$, the designer triggers spam according to $F(t)$ in equation (7), and the spam lasts until her belief reaches p^* in the event that no good news arrives by that time. The induced exploration under optimal public recommendations is, on average, slower—and the level of welfare attained is strictly lower—than that under optimal private recommendations.

A direct computation shows that $F(t)$ is increasing in p^0 and ρ, leading to the comparative statics similar to Corollary 1:

COROLLARY 2. As p^0 or ρ increases, the rate of user exploration increases under optimal public recommendations.

As before, these comparative statics suggest the potential role of product research by the designer.

VI. MATCHING PRODUCTS TO CONSUMERS

Categorizing products has become an important tool that online recommenders use to inform users about their characteristics and identify target consumers. In the past, movies and songs were classified by only a handful of genres; now recommenders categorize them into numerous subgenres that match consumers' fine-grained tastes.[30] In this section, we show how a designer

30. Netflix has 76,897 micro-genres to classify the movies and TV shows available in their library (see "How Netflix Reverse Engineered Hollywood,"

can match a product to the right consumer type through user exploration.

To this end, we modify our model to allow for horizontal preference differentiation. As before, a product is released at $t = 0$, and a unit mass of agents arrives at every instant $t \geqslant 0$. However, the agents now consist of two different preference types, type a and type b, with masses m_a and m_b, respectively. We assume that $m_b > m_a = 1 - m_b > 0$.[31]

The agent types are known to the designer from, say, their past consumption histories. However, the product's fit with each type is initially unknown, and the designer's objective is to discover the fit so that she can recommend it to the right type of agent. Specifically, the product is of type $\omega \in \{a, b\}$, which constitutes the unknown state of the world. A type-ω agent enjoys the payoff of 1 from a type-ω product but 0 from a type-ω' product, where $\omega \neq \omega' \in \{a, b\}$.[32] The common prior belief is $p^0 = \Pr[\omega = b] \in [0, 1]$. At any point, given belief $p = \Pr[\omega = b]$, a type-b agent's expected utility is p, while a type-a agent's expected utility is $1 - p$. We call them the product's *expected fits* for the two types. The opportunity cost for both types is $c > 0$. Therefore, each agent is willing to consume the product if and only if its expected fit is higher than the cost.

Through consumption, an agent learns whether the product matches his taste; if so, he reports satisfaction at rate $\lambda = 1$. As before, the designer receives feedback in the form of conclusive news, with arrival rates that depend on the agents' exploration behaviors. Specifically, if fractions (α_a, α_b) of type-a and type-b agents explore, then the designer learns that the product is of type $\omega = a, b$ at the Poisson rate $\alpha_\omega m_\omega$. (For the sake of simplicity,

The Atlantic, January 2014). For example, a drama may now be classified as a "critically acclaimed irreverent drama" or a "cerebral fight-the-system drama," and a sports movie may be classified as an "emotional independent sports movie" or a "critically acclaimed emotional underdog movie." Likewise, Pandora classifies a song based on 450 attributes ("music genes"), leading to an astronomical number of subcategories.

31. If $m_a = m_b$, the optimal policy remains the same (as described in Proposition 3), except that beliefs do not drift when all agents experiment.

32. The current model can be seen as a simple variation of the baseline model. If both types value the product more highly, say, in state $\omega = b$ than in state $\omega = a$, then preferences are vertical, as in the baseline model. The key difference is that the preferences of the two types are horizontally differentiated in the current model.

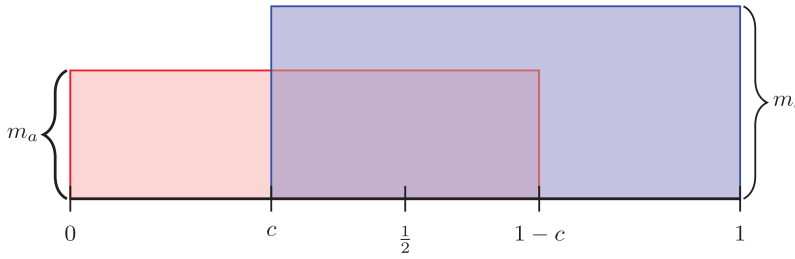

FIGURE III

Rates of Exploration by Two Types of Agents under Full Transparency

we assume that there is no background learning.) Hence, if the product type is not learned, the belief p drifts according to

$$\dot{p} = p(1-p)(\alpha_a m_a - \alpha_b m_b).$$

Note that the designer's belief can drift up or down depending on how many agents of each type are exploring. In particular, if both types explore fully ($\alpha_a = \alpha_b = 1$) but no feedback occurs, the designer's belief that the product is of type b decreases at a rate proportional to $m_b - m_a$.

Under full transparency, agents will behave optimally given the correct belief: a type-b agent (type-a agent) will consume if and only if $p \geqslant c$ ($1 - p > c \Leftrightarrow p < 1 - c$), as depicted in Figure III.

We next consider the first-best and second-best policies, under the assumption that $c < \frac{1}{2}$. This latter assumption means that the product is so popular that both types of agents are willing to consume it even when uncertainty is high (denoted the "overlapped" region in Figure III, which includes $p = \frac{1}{2}$).[33]

As before, if the designer receives news, sharing that news is trivially optimal. Hence, a policy is described by a pair (α_a, α_b) of spamming rates for the two types—the probabilities with which alternative types are recommended to consume in the event of no news—as a function of p.

33. If $c > \frac{1}{2}$, no learning occurs if the prior is in the range of $[1-c, c]$, as neither type is willing to consume. For more information on the case of an unpopular product, see the proof of Lemma D in Section C.1 of the Online Appendix, which treats both $c < \frac{1}{2}$ and $c > \frac{1}{2}$.

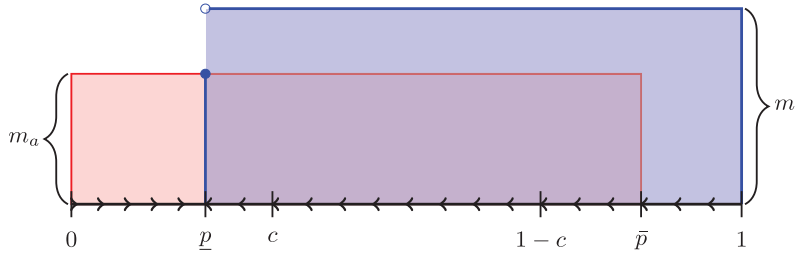

FIGURE IV

Rates of Exploration by Two Types of Agents under the First-Best Policy

LEMMA 1. The first-best policy is characterized by two thresholds, $\underline{p}$ and $\bar{p}$, with $0 < \underline{p} < c < 1-c < \bar{p} < 1$, such that

$$(\alpha_a^{FB}, \alpha_b^{FB}) = \begin{cases} (1, 0), & \text{for } p < \underline{p}, \\ (1, \frac{m_a}{m_b}), & \text{for } p = \underline{p}, \\ (1, 1), & \text{for } p \in (\underline{p}, \bar{p}], \\ (0, 1), & \text{for } p \in (\bar{p}, 1]. \end{cases}$$

The logic of the first-best policy follows the standard exploration–exploitation trade-off.[34] The policy calls for each type to explore the product as long as its expected fit exceeds a threshold: $\underline{p}$ for type b and $1-\bar{p}$ for type a. Due to the informational externalities, these thresholds are lower than the opportunity costs. In other words, the policy prescribes exploration for a type even when the product's expected fit does not justify the opportunity cost. As seen in Figure IV (compared with Figure III), the first-best policy results in wider exploration than full transparency.

Some features of the policy are worth explaining. The designer's belief drifts to $\underline{p}$ from either side (as depicted by the arrows in Figure IV), unless conclusive news arrives. The reason is that $m_b > m_a$, which results in a downward drift of belief when both types of agents explore. The behavior at $p = \underline{p}$ is also of interest. In this case, all type-a agents consume, but only a mass m_a of

34. The current model resembles that of Klein and Rady (2011), who study two players strategically experimenting with risky arms that are negatively correlated. Lemma 1 is similar to their planner's problem. The difference is that we allow for asymmetry in the size of the two agent types in our model. Of course, the main analyses are quite distinct: we focus on an agency problem, whereas they focus on a two-player game.

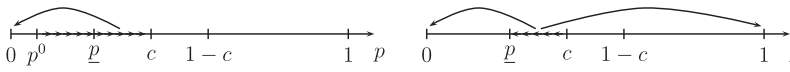

FIGURE V

Evolution of the Designer's Belief, First When Only Type-a Agents Explore (Left Panel) and Then When All Agents Do (Right Panel)

the type-b agents do, so that the belief remains constant without any updating. However, learning does not stop. Eventually, the product type will be revealed with probability 1.

Not surprisingly, the first-best policy may not be incentive compatible. The second-best policy illustrates how incentive considerations affect the optimal policy.

PROPOSITION 3. The second-best policy is described as follows:

(i) If $p^0 < c$, then $\alpha_t^{SB} = (\alpha_a^{SB}, \alpha_b^{SB}) = (1, 0)$ until p_t (which drifts up) reaches c; thereafter, the first-best policy is followed.

(ii) If $p^0 > 1 - c$, then $\alpha_t^{SB} = (0, 1)$ until p_t (which drifts down) reaches $1 - c$; thereafter, the first-best policy is followed.

(iii) If $p^0 \in [c, 1 - c]$, then the first-best policy is followed.

If $p^0 \in [c, 1 - c]$, the first-best policy is incentive compatible. Since both types of agents initially have incentives to explore, being told to explore is (weakly) good news (meaning that the designer has not learned that the state is unfavorable). By contrast, if $p^0 \notin [c, 1 - c]$, the first-best policy may not be incentive compatible.

Suppose, for instance, that $p^0 < c$. In this case, type-b agents will refuse to explore. Therefore, only type-a agents can be induced to explore. We explain the second-best policy in this case with the aid of two graphs: Figure V, which tracks evolution of the designer's belief, and Figure VI, which tracks the evolution of agents' beliefs, both assuming no breakthrough news. Since only type-a agents explore in the initial phase (times $t \leqslant 1$), the designer's belief will drift up as long as no news obtains, as seen in Figure V (left panel). During this phase, type-a agents are recommended the product regardless of whether the designer learns the state is a, so the induced belief remains constant for both types.[35]

Next, suppose the designer's belief reaches $p_t = c$. Thereafter, the first-best policy becomes incentive compatible. The reason is

35. Recall that the designer can never learn that the state is b if only type-a agents explore.

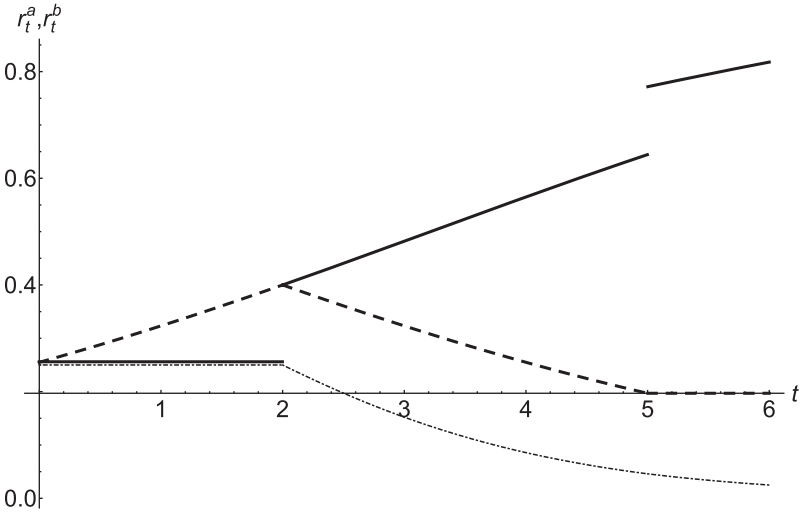

FIGURE VI

Evolution of Agents' Beliefs When the Designer Receives No News
($c = \frac{2}{5}, m_b = \frac{2}{3}, m_a = \frac{1}{3}$)

In dash-dot (red online), the type-a agent's belief; in solid (blue online), the type-b agent's belief; in dash (green online), the designer's belief.

that although the belief drifts down thereafter, as depicted in Figure V (right panel), the designer can induce both types to become optimistic about the product's fitness for them by simply recommending the product to them. A type-a agent becomes more optimistic (the belief drifts down), since she knows that type-b agents might be exploring, and were the true state known to be b, she would be told not to consume. Hence, being told to explore is good news. Meanwhile, a type-b agent's optimism jumps up to $q(p_t) = c$ at time $t = 2$ because being told to explore is proof that the designer has not learned that the state is a. Thereafter, a type-b agent becomes more optimistic (her belief drifts up) because being told to explore means that the designer has not learned that the state is a. At some point ($t = 5$), the designer's belief reaches $\underline{p}$. Because only a fraction of type-b agents get spammed, being told to explore is another piece of good news (suggesting that perhaps the designer has learned that the state is b). Type-b agents' belief jumps up and drifts up further from then on. To foster such optimism for both types of agents, the designer simply needs to keep the recommended agents uninformed about whether

the recommendation is genuine or spam and about which recommendation is made to the other agents.[36]

The optimal policy shares some common features with that in our baseline model. First, the second-best policy induces wider user exploration than would be possible under full transparency. In particular, once p_t drifts below c, type b-agents will never explore under full transparency, but they will continue to explore under the second-best policy. Second, compared with the first-best policy, the scope of early exploration is "narrower"; the exploration begins with the most willing agents with a high expected fit—type-a agents in the case of $p^0 < c$—and then gradually spreads to the agents who are initially less inclined to explore, which is a manifestation of "starting small" in the current context.

VII. Extensions

We now extend the baseline model analyzed in Section IV to incorporate several additional features. The detailed analysis is provided in Section D of the Online Appendix; here, we illustrate the main ideas and results.

VII.A. Vertically Heterogeneous Preferences

The preceding section considers agents whose preferences are horizontally differentiated. Here, we consider agents whose preferences are vertically differentiated. Suppose that the agents have two possible opportunity costs: $1 > c_H > c_L > p^0$. (As in the baseline model, we assume background learning at rate $\rho > 0$.) A low-cost agent is more willing to explore the product than a high-cost agent, so the model captures the vertical heterogeneity of preferences. As in the preceding section, we assume that the designer observes the type of the agent from, say, his past consumption history.[37] For instance, the frequency of downloading or streaming movies may indicate a user's (opportunity) cost of exploration.

36. The divergence of the beliefs held by the two types of agents is sustained only through private recommendations. Hence, the optimal policy cannot be implemented through a public recommendation.

37. If the designer cannot infer the agents' costs, then her ability to induce agents to explore is severely limited. Che and Hörner (2015) show that if the agents have private information over costs drawn uniformly from [0, 1], then the second-best policy reduces to full transparency, meaning that the designer will never spam.

We illustrate how the designer tailors her recommendation policy to each type in this case.

To begin, one can extend the incentive constraint (3) to yield the spamming capacity for each type:

$$\hat{\alpha}_i(p_t) := \frac{(1-c_i)(p^0-p_t)}{(1-p^0)(c_i-p_t)},$$

for $i = L, H$. In other words, each type $i = H, L$ can be spammed with a probability of at most $\hat{\alpha}_i(p_t)$, given designer's belief p_t. Note that $\hat{\alpha}_L(p_t) > \hat{\alpha}_H(p_t)$, so a low-cost type can be spammed more than a high-cost type. The optimal policies are again characterized by cutoffs:

PROPOSITION 4. Both the first-best and the second-best policies are characterized by a pair of thresholds $0 \leqslant p_L \leqslant p_H \leqslant p^0$, such that each type $i = L, H$ is asked to explore with maximal probability (which is one under the first-best policy and $\hat{\alpha}_t(p_t)$ under the second-best policy) if $p_t \geqslant p_i$, and zero exploration otherwise. The belief threshold for the low type is the same under the two regimes, but the threshold for the high type is higher under first-best policy than under the second-best policy.

The overall structure of the optimal policy is similar to that of the baseline model: the policy prescribes maximal exploration for each type until her belief reaches a threshold (which is below its opportunity cost), and the maximal exploration under the second-best policy "starts small" and accelerates over time. Consequently, given a sufficiently high prior belief, both types are initially induced to explore. The high type's threshold is reached first, and from then on only the low type explores. Next, the low type's threshold is reached, at which point all exploration stops.

The trade-off facing the designer with regard to the low type's marginal exploration is conceptually the same as before, which explains why the low type's threshold is the same under both the first-best and the second-best policies. However, the trade-off with regard to the high type's marginal exploration is different. Unlike the baseline model, stopping the high type's exploration does not mean stopping all users' exploration; it means that only the low type will explore thereafter. This has several implications. First, the high type will explore less, making the threshold higher, compared with the case in which only the high type can explore

(a version of the baseline model). Second, this also explains why the high type will explore more under the second-best policy than under the first-best policy. The binding incentive constraint means that the low-cost type's exploration will be lower under the second-best policy, so the consequence of stopping the high-cost type's exploration is worse under the second-best policy than under first-best policy. Third, the high type's exploration makes the arrival of news more plausible, thus making the recommendation for the low type more credible. Hence, the designer "hangs on" to the higher-cost type longer than she does under the first-best policy.[38]

VII.B. Calendar Time Uncertainty

We have thus far assumed that agents are perfectly aware of the calendar time. As we argue, relaxing this assumption makes it easier for the designer to spam the agents. Indeed, if they are a priori sufficiently unsure about how long exploration has been occurring, the designer can achieve the first-best policy. Roughly speaking, uncertainty regarding calendar time allows the designer to further cloud the meaning of a "consume" recommendation, as she can shuffle not only histories of a given length (some when she has learned the state, others when she has not) but also histories of different lengths.

A simple way to introduce calendar time uncertainty is to assume that the agents do not know when they have arrived relative to the product's release time. In keeping with realism, we assume that the flow of agents "dries out" after a random time τ following an exponential distribution with parameter $\xi > 0$.[39]

From the designer's point of view, ξ is an "additional" discount factor to be added to r, their original discount rate. Hence, the first-best policy is the same as in the baseline model, adjusting for this

38. Che and Hörner (2015) show that this structure holds more generally, for instance, when agents' costs are continuous, as drawn from an interval.

39. An alternative modeling option would be to assume that agents hold the improper uniform prior on the arrival time. In that case, the first-best policy is trivially incentive compatible, as an agent assigns probability one to an arrival after the exploration phase is over. Not only is an improper prior conceptually unsatisfying, but it is also more realistic that a product has a finite (but uncertain) "shelf life," which is what the current assumption amounts to—namely, the product's shelf life expires at τ. Agents do not know τ or their own arrival time: conditional on $\{\tau = t\}$ (which they do not know), they assign a uniform prior over $[0, t]$ on their arrival time.

rate. In particular,

$$p^* = c\left(1 - \frac{(r+\xi)v}{\rho + (r+\xi)(v+\frac{1}{\lambda})}\right),$$

where $v := \frac{1-c}{r+\xi}$.

The following formalizes the intuition that, provided that the prior belief about calendar time is sufficiently diffuse, the designer is able to replicate the first-best policy.

PROPOSITION 5. There exists $\bar{\xi} > 0$ such that, for all $\xi < \bar{\xi}$, the first-best policy is incentive compatible.

This result suggests that it is easier to incentivize users to explore a product that has a long shelf life or a priori durable appeal than a product that does not. The intuition is as follows. An agent will have a stronger incentive to explore when it is more likely that she has arrived after the exploration phase is complete—that is, after the designer's belief will have reached p^* absent any good news—as any recommendation made in the postexploration phase must be an unambiguously good signal about the product. A longer shelf life ξ for the product means not only that both the exploration and the postexploration phases are longer but also that the agents will put a comparatively higher probability on arriving in the second phase.

VII.C. Naive Agents

In practice, some users are naive enough to follow the platform's recommendation without any skepticism. Our results are shown to be robust to the presence of such naive agents, with a new twist. Suppose that a fraction $\rho_n \in (0, 1)$ of the agents naively follow the designer's recommendation. The others are rational and strategic, as has been assumed thus far; in particular, they know about the presence of the naive agents and can rationally respond to the recommendation policy with the knowledge of their arrival time. The designer cannot tell naive agents apart from rational agents. For simplicity, we now assume no background learning. Intuitively, the naive agents are similar to fans (background learning) in our baseline model, in the sense that they can be called on to seed social learning at the start of product life. However, naive agents are different from fans in two ways. The naive agents incur positive costs $c > 0$, so their learning is not free, which affects the optimal recommendation policy. Second, their exploration can

only be triggered by the designer, and the designer, due to her inability to separate them, cannot selectively recommend a product to them.[40]

The designer's second-best policy has the same structure as before: at each time t, absent any news, she spams a fraction $\alpha_t \in [0, 1]$ of randomly selected agents to explore, regardless of their types. (She recommends to all agents on the receipt of good news.) Due to the presence of naive agents, the designer may now spam at a level that may fail the rational agents' incentive constraint. Given policy α_t, mass $\rho_n\alpha_t$ of naive agents will explore, and mass $(1 - \rho_n)\alpha_t$ of rational agents will explore if and only if $\alpha_t \leqslant \hat{\alpha}(p_t)$, where $\hat{\alpha}(p_t)$ is defined in equation (5). Since the rational agents may not follow the recommendation, unlike the baseline model, the mass of agents who explore may differ from the mass of those who receive spam. Clearly, the most the designer can induce to explore is

$$\hat{e}(p_t) := \max\{\rho_n, \hat{\alpha}(p_t)\} \geqslant \rho_n\alpha_t + (1 - \rho_n)\alpha_t \cdot 1_{\{\alpha_t \leqslant \hat{\alpha}(p_t)\}}.$$

PROPOSITION 6. In the presence of naive agents, the second-best policy induces exploration at rate

$$e^{SB}(p_t) = \begin{cases} \hat{e}(p_t) & \text{if } p_t \geqslant p^*; \\ 0 & \text{if } p_t < p^*, \end{cases}$$

where p^* is defined in Proposition 1—but with $\rho = 0$.

The presence of naive agents adds an interesting feature to the optimal policy. To explain, assume that $p^* < p^0 < c$. Recall that $\hat{\alpha}(p_t) \approx 0 < \rho_n$ for $t \approx 0$, implying that $\hat{e}(p_t) = \rho_n$ in the early stages, meaning that the optimal policy always begins with a "blast" of spam to all agents; that is, $\alpha_t^{SB} = 1$. Of course, the rational agents will ignore the spam, but the naive agents will listen and explore. Despite their naiveté, their exploration is real, so the designer's credibility and her capacity $\hat{\alpha}(p_t)$ to spam the rational agents increase over time. If $\rho_n < \hat{\alpha}(p^*)$, then $\hat{\alpha}(p_t) > \rho_n$ for all $t > \hat{t}$, where $\hat{t} \in (0, t^*)$ is such that $\rho_n = \hat{\alpha}(p_{\hat{t}})$.[41] This means that starting at $\hat{t}$, the designer switches from blasts of spam to a more controlled spam campaign at $\alpha_t = \hat{\alpha}(p_t)$, targeting rational agents (as well as

40. We assume that the naive agents are still sophisticated enough to mimic what rational agents would say when the designer asks them to reveal themselves.

41. The threshold time t^* is the same as that defined in Section IV, except that $\rho = 0$.

naive ones). If $\rho_n \geqslant \hat{\alpha}(p^*)$, however, the designer will keep on blasting spam to all agents and thus rely solely on the naive agents for exploration (until she reaches p^*).

The blasting of spam in the early phases is reminiscent of aggressive campaigns often observed when a new show (e.g., a new original series) or a new platform is launched. Although such campaigns are often ignored by sophisticated users, our analysis shows that they can be optimal in the presence of naive users.

VII.D. Costly Product Research

For platforms such as Pandora and Netflix, product research by the recommender constitutes an important source of background learning. Product research may be costly for a recommender, but as highlighted earlier, it may contribute to social learning. To gain more precise insights into the role played by the recommender's product research, we endogenize background learning. Specifically, we revisit the baseline model, except that now the designer chooses the background learning $\rho_t \geqslant 0$ at the flow cost of $c(\rho_t) := \rho_t^2$ at each time $t \geqslant 0$. While a closed-form solution is difficult to obtain, a (numerical) solution for specific examples provides interesting insights. (The precise formulation and method of analysis are detailed in Section D.4 of the Online Appendix.)

Figure VII illustrates the product research under the second-best policy and full transparency.

In this example, as in the baseline model, user exploration α_t follows a hump-shaped pattern; it starts small but accelerates until it reaches a peak, after which it completely ceases. The intuition for this pattern is the same as before. The interesting new feature is the front-loading of the designer's product research ρ_t^{SB}. As can be seen in Figure VII, ρ_t^{SB} is highest at $t = 0$ and falls gradually. Eventually the product research stops, but well after user exploration stops.[42]

42. The latter feature may be surprising because our cost function satisfies $c'(0) = 0$. In this example, designer learning eventually stops because the benefit of product research decreases exponentially as p_t approaches 0. Hence, unlike in the baseline model, learning is incomplete, despite the arbitrarily small marginal cost at low levels of background learning. Note also that ρ_t^{SB} has a kink at the time that agent exploration ceases, and it can be increasing just prior to that time, as shown in Figure VII. Because the prospect of future learning through agents' exploration winds down, the incentives to learn via ρ increase, which can more than offset the depressing effect of increased pessimism about the state.

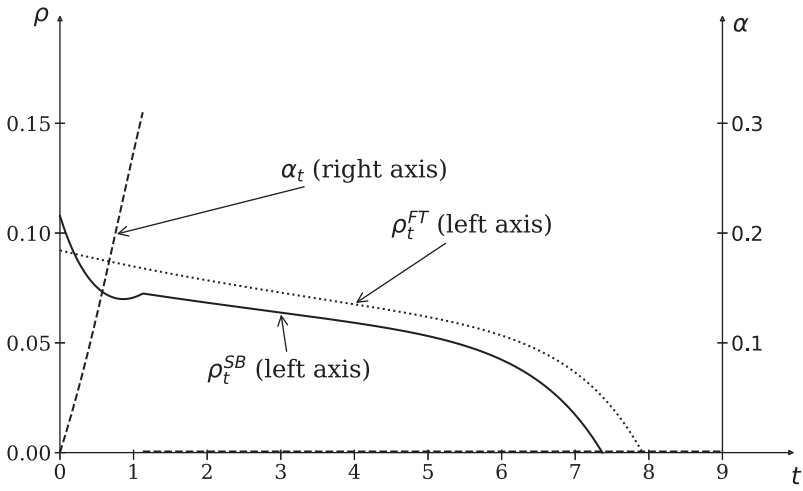

FIGURE VII

Functions ρ and α ($r = 0.01$, $\lambda = 0.01$, $c = 0.6$, $p^0 = 0.5$)

The front-loading of ρ reflects three effects. First, the marginal benefit from learning is high in the early phases when the designer is most optimistic. Second, as noted earlier, the designer's learning and the agents' exploration are "substitutes" for learning, and the value of the former is particularly high in the early phases when the latter is highly constrained. Third, background learning increases the designer's capacity to credibly spam the agents, and this effect is strongest in the early phases due to its cumulative nature mentioned earlier.

These three effects are seen more clearly via comparison with the full-transparency benchmark, where the designer optimally chooses its research (denoted in Figure VII by ρ_t^{FT}) against agents choosing $\alpha_t \equiv 0$, their optimal behavior under full transparency. The first two effects are present in the choice of ρ_t^{FT}. In fact, the substitute effect is even stronger here than in the second-best policy because agents never explore here, which explains why ρ_t^{FT} exceeds ρ_t^{SB} for a wide range of t. Very early, however, the third effect—relaxing the incentive constraint—proves quite important for the second-best policy, which is why $\rho_t^{SB} > \rho_t^{FT}$ for a very low t. In short, the front-loading of designer learning is even more pronounced in the second-best policy

compared with the full-transparency benchmark due to the incentive effect.[43]

VII.E. A More General Signal Structure

Thus far, our model has assumed a simple signal structure that features only good news, which is a reasonable assumption for many products whose priors are initially unfavorable but can be improved dramatically through social learning. However, for some other products, social learning may involve the discovery of poor quality. Our signal structure can be extended to allow for such a situation via "bad" news.[44] Specifically, news can be either good or bad, where good news reveals $\omega = 1$ and bad news reveals $\omega = 0$, and the arrival rates of the good news and bad news are $\lambda_g > 0$ and $\lambda_b > 0$, respectively, conditional on the state. More precisely, if a flow of mass α consumes the product over some time interval $[t, t + \mathrm{d}t)$, then during this time interval, the designer learns that the product is "good" with probability $\lambda_g(\rho + \alpha)\mathrm{d}t$ and "bad" with probability $\lambda_b(\rho + \alpha)\mathrm{d}t$. Note that we retain the assumption that either type of news is perfectly conclusive.

If news arrives, the designer's posterior jumps to 1 or 0. Otherwise, it follows

$$\dot{p}_t = -p_t(1 - p_t)\delta(\rho + \alpha_t), \quad p_0 = p^0, \tag{8}$$

where $\delta := \lambda_g - \lambda_b$ is the relative arrival rate of good news, and α_t is the exploration rate of the agents. Intuitively, the designer becomes pessimistic from absence of news if good news arrives faster ($\delta > 0$) and becomes optimistic if bad news arrives faster ($\delta < 0$). The former case is similar to the baseline model, so we focus on the latter case. The formal result, the proof of which

43. To avoid clutter, we do not depict the first-best policy in Figure VII, but its structure is quite intuitive. First, user exploration under the first-best policy is the same as before: a full exploration until p falls to a particular threshold. Second, the first-best product research ρ_t^{FB} declines in t, as is the case under full transparency, due to the designer's declining belief. More important, ρ_t^{FB} is below ρ_t^{SB} everywhere. The reason is twofold: (i) more user exploration occurs under the first-best policy, which lowers optimal product research through the substitution effect, and (ii) the incentive-promoting effect of product research is absent under the first-best policy.

44. See Keller and Rady (2015) for the standard bad news model of strategic exploration.

is available in Section D.5 of the Online Appendix (which also includes the general good news case), is as follows:

PROPOSITION 7. Consider the bad news environment ($\delta < 0$). The first-best policy (absent any news) prescribes no exploration until the posterior p rises to p_b^{FB} and then full exploration at a rate of $\alpha^{FB}(p) = 1$ thereafter, for $p > p_b^{FB}$, where

$$p_b^{FB} := c\left(1 - \frac{rv}{\rho + r(v + \frac{1}{\lambda_b})}\right).$$

The second-best policy implements the first-best policy if $p_0 \geqslant c$ or if $p_0 \leqslant \hat{p}_0$ for some $\hat{p}_0 < p_b^{FB}$. If $p_0 \in (\hat{p}_0, c)$, then the second-best policy prescribes no exploration until the posterior p rises to p_b^* and then exploration at the maximum incentive-compatible level thereafter for any $p > p_b^*$,[45] where $p_b^* > p_b^{FB}$. In other words, the second-best policy triggers exploration at a later date and at a lower rate than the first-best policy.

Although the structure of the optimal recommendation policy is similar to that in the baseline model, the intertemporal trajectory of exploration is quite different. Figure VIII depicts an example with $\delta < 0$ and a sufficiently low prior belief. Initially, the designer finds the prior to be too low to trigger a recommendation, and she never spams as a result. However, as time progresses without receiving any news (good or bad), her belief improves gradually, and as her posterior reaches the optimal threshold, she begins spamming at the maximal capacity allowed by incentive compatibility. One difference here is that the optimal second-best threshold differs from that of the first-best threshold. The designer has a higher threshold, meaning that she waits longer to trigger exploration under the second-best policy than she would under the first-best policy. This is due to the difference in the trade-offs at the margin between the two regimes. Although the benefit of not triggering exploration is the same in the two regimes, the benefit

45. The maximal incentive-compatible level is $\hat{\alpha}(p_t) := \min\left\{1, \frac{\left(\frac{p_t(1-p^0)}{(1-p_t)p^0}\right)^{-\frac{\lambda g}{\delta}} - 1}{(\frac{1-p_t}{p_t})(\frac{c}{1-c}) - 1}\right\}$.

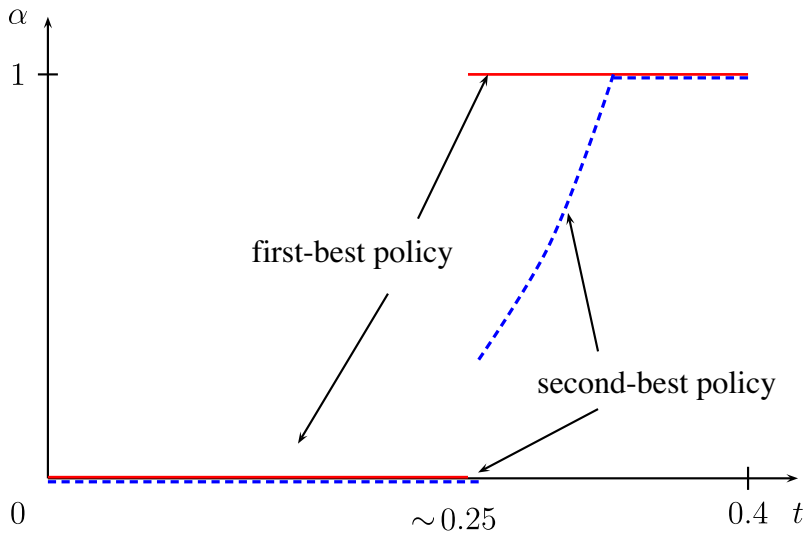

FIGURE VIII

Path of α for $\delta < 0$ and $(c, \rho, p^0, r, \lambda_g, \lambda_b) = (\frac{1}{2}, 1, \frac{2}{7}, \frac{1}{10}, 1, 2)$

of triggering exploration is lower in the second-best regime due to the constrained exploration that follows in that regime.

VIII. RELATED LITERATURE

Our article relates to several strands of the literature. First, our model can be viewed as introducing an optimal design into the standard model of social learning. In standard models (for instance, Bikhchandani, Hirshleifer, and Welch 1992; Banerjee 1993; Smith and Sørensen 2000), a series of agents take actions myopically, ignoring their effects on the learning and welfare of agents in the future. Smith, Sørensen, and Tian (2016) study altruistic agents who distort their actions to improve observational learning for posterity.[46] In an observational learning model such as that of Smith, Sørensen, and Tian (2016), agents are endowed with private signals, and the main issue is whether their actions communicate the private signals to subsequent agents. By

46. In section 4.B of their article, they show how transfers can implement the optimal policy that they derive in the case of altruistic agents.

contrast, in our model, agents do not have private information ex ante and must be incentivized to acquire it.

Whether they want to communicate such information (by providing feedback or taking an action that signals it) is an important issue, which we do not address. Instead, we simply posit a stochastic feedback (Poisson) technology. Frick and Ishii (2014) examine how social learning affects the adoption of innovations of uncertain quality and explain the shape of commonly observed adoption curves. In these papers, the information structure—what agents know about the past—is fixed exogenously. Our focus is precisely the optimal design of the information flow to the agent. Such dynamic control of information is present in Gershkov and Szentes (2009), but that paper considers a very different environment, as direct payoff externalities (voting) exist.

Much more closely related to the present article is a recent article by Kremer, Mansour, and Perry (2014). They study the optimal mechanism that induces agents to explore two products of unknown qualities. As in this article, the designer can incentivize agents to explore by manipulating their beliefs, and her ability to do so increases over time. While these themes are similar, there are differences. In their model, the uncertainty regarding the unknown state is rich (the quality of the product is drawn from some interval), but user feedback is instantaneous (trying the product once reveals its quality). In the current article, the state is binary, but the user feedback is gradual. This distinction matters for welfare and exploration dynamics. Here, the incentive problem entails a real-time delay and a nonvanishing welfare loss; in their setup, the loss disappears in the limit, as either the time interval shrinks or its horizon increases. The exploration dynamics also differ: our optimal policy induces a "hump"-shaped exploration that depends on the designer's belief, whereas their exploration dynamics—namely, how long it takes for a once-and-for-all exploration to occur—maps to the realized value of the dominant product observed in the first period. In addition, we explore extensions that have no counterpart in their model, including public recommendations and product categorization. We ultimately view the two papers as complementary.

Our model builds on the Poisson bandit process for the recommender's signal, introduced in a strategic setting by Keller, Rady, and Cripps (2005) and applied by several authors in principal-agent setups (see, for instance, Klein and Rady; Hörner and Samuelson 2013; Halac, Kartik, and Liu 2016). As in these

articles, the Poisson bandit structure provides a tractable tool for studying dynamic incentives. The main distinguishing feature of the current model is that the disclosure policy of the principal (recommender) and the resulting control of agents' beliefs serve as the main tool to control the agents' behavior.

Our article also contributes to the literature on Bayesian persuasion that studies how a principal can credibly manipulate agents' beliefs to influence their behaviors. Aumann, Maschler, and Stearns (1995) analyze this question in repeated games with incomplete information, whereas Ostrovsky and Schwarz (2010), Rayo and Segal (2010), and Kamenica and Gentzkow (2011) study the problem in a variety of organizational settings. The current article pursues a similar question in a dynamic setting. In this regard, the current article joins a burgeoning literature that studies Bayesian persuasion in dynamic settings (see Renault, Solan, and Vieille 2014; Ely, Frankel, and Kamenica 2017; Halac, Kartik, and Liu 2015; Ely 2017). The focus on social learning distinguishes our article from these other papers.[47]

Finally, the present article is related to the empirical literature on user-generated reviews (Jindal and Liu 2008; Mayzlin, Dover, and Chevalier 2014; Luca and Zervas 2016).[48] These publications suggest ways of empirically identifying manipulations in the reviews made by the users of Internet platforms such as Amazon, Yelp, and TripAdvisor. Our article contributes a normative perspective on the extent to which the manipulation should be controlled.

IX. Conclusion

Early exploration is crucial for users to discover and adopt potentially valuable products on a large scale. This article has shown how a recommendation policy can be designed to promote such early exploration. The current study offers several takeaways.

47. Papanastasiou, Bimpikis, and Savva (2017) show that the insights of the current article extend to the two-product context, although without fully characterizing the optimal mechanism. Mansour, Slivkins, and Syrgkanis (2015) develop an incentive-compatible disclosure algorithm that is near optimal regardless of the prior in a multi-armed bandit setting, while Mansour et al. (2016) allow for interactions among the agents. Avery, Resnick, and Zeckhauser (1999) and Miller, Resnick, and Zeckhauser (2004) study monetary incentives to prompt the sharing of product information.

48. Dai et al. (2014) offer a structural approach to aggregate consumer ratings and apply it to restaurant reviews from Yelp.

First, a key aspect of a user's incentives to explore is his beliefs about a product, which the designer can control by "pooling" a genuine positive signal regarding the product with spam—a recommendation without any such signal. Spamming can turn users' beliefs favorably toward the product and thus incentivize exploration by early users. Consequently, spamming is part of an optimal recommendation policy.

Second, spamming is effective only when it is properly underpinned by genuine learning. Excessive spam campaigns can backfire and harm the recommender's credibility. We have shown how a recommender can build her credibility by "starting small" in terms of the amount (in the case of private recommendations), the probability (in the case of public recommendations) and the breadth (in the case of heterogeneous tastes) of spam, depending on the context. We have also highlighted the role of the recommender's independent product research, such as that performed by Netflix and Pandora. Recommender-initiated research can not only act as a substitute for costly learning by users but also substantially increase the credibility with which the recommender can persuade agents to explore. These benefits are particularly important in the early phases of the product cycle when user exploration is weakest, causing the designer to front-load her investment.

As noted earlier, this article yields implications for several aspects of online platforms. Aside from online platforms, a potentially promising avenue of application is the adaptive clinical trial (ACT) of medical drugs and procedures. Unlike the traditional design, which fixes the characteristics of the trial over its entire duration, the ACT modifies the course of the trial based on the accumulating results of the trial, typically by adjusting the doses of a medicine, dropping patients from an unsuccessful treatment arm and adding patients to a successful arm (see Berry 2011; Chow and Cheng 2008). ACTs improve efficiency by reducing the number of participants assigned to an inferior treatment arm and/or the duration of their assignment to such an arm.[49] An important aspect of the ACT design is the incentives for the patients and doctors to participate in and stay on the trial. To this end, managing their beliefs, which can be affected when the

49. The degree of adjustment is limited to a level that does not compromise the randomized control needed for statistical power. Some benefits of ACTs are demonstrated in Trippa et al. (2012).

prescribed treatment changes over the course of the trial, is crucial. Note that the suppression of information, especially with regard to alternative treatment arms, is within the ethical boundary of the clinical trial and is a key instrument for preserving patient participation and the integrity of the experiment.[50] The insight from this study can provide some useful guidance for future research on this aspect of ACT design.

While this article provides some answers on how user exploration can be improved via recommendations, it raises another intriguing question: how does recommendation-induced user exploration influence the learning of user preferences? For instance, in the ACT context, the endogenous assignment of patients to alternative treatment arms may compromise the purity of a randomized trial and make the treatment effect difficult to identify. A similar concern arises with the dynamic adjustment of explorations conducted by online platforms, as they may make it harder to assess the effect of user exploration. A precise understanding of the tradeoff between improved user exploration and the observation of user preferences requires a careful embedding of current insight within the richer framework Internet platforms employ to understand user preferences. We leave this question for future research.

APPENDIX A: PROOF OF PROPOSITION 1

Proof. It is convenient to work with the odds ratio, $\ell := \frac{p}{1-p}$, and with the cost ratio, $k := \frac{c}{1-c}$. Using ℓ_t and substituting for g using equation (3), we can write the second-best program as follows:

$$[SB] \qquad \sup_{\alpha} \int_{t \geqslant 0} e^{-rt} \left(\ell^0 - \ell_t - \alpha_t (k - \ell_t) \right) \mathrm{d}t,$$

subject to

$$\dot{\ell}_t = -\lambda(\rho + \alpha_t)\ell_t, \ \forall t, \ \text{and} \ \ell_0 = \ell^0, \tag{9}$$

$$0 \leqslant \alpha_t \leqslant \bar{\alpha}(\ell_t), \ \forall t, \tag{10}$$

50. For instance, keeping the type of arm to which a patient is assigned—whether a control arm (e.g., placebo) or a new treatment—hidden from the patient and their doctor is an accepted practice.

where $\ell^0 := \frac{p^0}{1-p^0}$ and $\bar{\alpha}(\ell_t) := \hat{\alpha}(\frac{\ell_t}{1+\ell_t})$. Obviously, the first-best program, labeled [*FB*], is the same as [*SB*], except that the upper bound for $\bar{\alpha}(\ell_t)$ is replaced by 1.

To analyze this trade-off precisely, we reformulate the designer's problem to conform with the standard optimal control framework. First, we switch the roles of the variables so that we treat ℓ as a "time" variable and $t(\ell) := \inf\{t \mid \ell_t \leqslant \ell\}$ as the state variable, which is interpreted as the time required for a posterior ℓ to be reached. Up to constant (additive and multiplicative) terms, the designer's problem is written as follows:

For problem $i = SB, FB$,

$$\sup_{\alpha(\ell)} \int_0^{\ell^0} e^{-rt(\ell)} \left(1 - \frac{k}{\ell} - \frac{\rho\left(1-\frac{k}{\ell}\right)+1}{\rho+\alpha(\ell)}\right) \mathrm{d}\ell$$

s.t. $t(\ell^0) = 0$,

$$t'(\ell) = -\frac{1}{\lambda(\rho+\alpha(\ell))\ell},$$

$$\alpha(\ell) \in \mathcal{A}^i(\ell),$$

where $\mathcal{A}^{SB}(\ell) := [0, \bar{\alpha}(\ell)]$, and $\mathcal{A}^{FB} := [0, 1]$.

This transformation enables us to focus on the optimal recommendation policy as a function of the posterior ℓ. Given the transformation, the admissible set no longer depends on the state variable (since ℓ is no longer a state variable), thus conforming to the standard specification of the optimal control problem.

Next we focus on $u(\ell) := \frac{1}{\rho+\alpha(\ell)}$ as the control variable. With this change of variable, the designer's problem (both second-best and first-best) is restated, up to constant (additive and multiplicative) terms.

For $i = SB, FB$,

$$\sup_{u(\ell)} \int_0^{\ell^0} e^{-rt(\ell)} \left(1 - \frac{k}{\ell} - \left(\rho\left(1-\frac{k}{\ell}\right)+1\right)u(\ell)\right) \mathrm{d}\ell, \tag{11}$$

s.t. $t(\ell^0) = 0$,

$$t'(\ell) = -\frac{u(\ell)}{\lambda\ell},$$

$$u(\ell) \in \mathcal{U}^i(\ell),$$

where the admissible set for the control is $\mathcal{U}^{SB}(\ell) := [\frac{1}{\rho+\bar{\alpha}(\ell)}, \frac{1}{\rho}]$ for the second-best problem; and $\mathcal{U}^{FB}(\ell) := [\frac{1}{\rho+1}, \frac{1}{\rho}]$. With this transformation, the problem becomes a standard linear optimal control problem (with state t and control α). A solution exists via the Filippov-Cesari theorem (Cesari 1983).

We thus focus on the necessary condition for optimality to characterize the optimal recommendation policy. To this end, we write the Hamiltonian:

$$\mathcal{H}(t, u, \ell, \nu) = e^{-rt(\ell)}\left(1 - \frac{k}{\ell} - \left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right)u(\ell)\right) - \nu\frac{u(\ell)}{\lambda\ell}. \tag{12}$$

The necessary optimality conditions are that there exist an absolutely continuous function ν: $[0, \ell^0]$ such that, for all ℓ, either

$$\phi(\ell) := \lambda e^{-rt(\ell)}\ell\left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right) + \nu(\ell) = 0, \tag{13}$$

or else $u(\ell) = \frac{1}{\rho+\bar{\alpha}(\ell)}$ if $\phi(\ell) > 0$ and $u(\ell) = \frac{1}{\rho}$ if $\phi(\ell) < 0$.

Furthermore,

$$\begin{aligned}\nu'(\ell) &= -\frac{\partial\mathcal{H}(t, u, \ell, \nu)}{\partial t} \\ &= re^{-rt(\ell)}\left(\left(1 - \frac{k}{\ell}\right)(1 - \rho u(\ell)) - u(\ell)\right) \quad (\ell - \text{a.e.}).\end{aligned} \tag{14}$$

Finally, transversality at $\ell = 0$ implies that $\nu(0) = 0$ (since $t(\ell)$ is free).

Note that

$$\begin{aligned}\phi'(\ell) = &-rt'(\ell)\lambda e^{-rt(\ell)}\ell\left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right) \\ &+ \lambda e^{-rt(\ell)}\left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right) + \frac{\rho k\lambda e^{-rt(\ell)}}{\ell} + \nu'(\ell),\end{aligned}$$

or using the formulas for t' and ν',

$$\phi'(\ell) = \frac{e^{-rt(\ell)}}{\ell}\left(r\left(\ell - k\right) + \rho\lambda k + \lambda\left(\rho\left(\ell - k\right) + \ell\right)\right), \tag{15}$$

Therefore, ϕ cannot be identically 0 over some interval, as there is at most one value of ℓ for which $\phi'(\ell) = 0$. Every solution must be "bang-bang." Specifically,

$$\phi'(\ell) \gtreqless 0 \Leftrightarrow \ell \gtreqless \tilde{\ell} := \left(1 - \frac{\lambda(1+\rho)}{r + \lambda(1+\rho)}\right) k > 0.$$

In addition, $\phi(0) = -\lambda e^{-rt(\ell)} \rho k < 0$. Therefore, $\phi(\ell) < 0$ for all $0 < \ell < \ell^*$ for some threshold $\ell^* > 0$, and $\phi(\ell) > 0$ for $\ell > \ell^*$. The constraint $u(\ell) \in \mathcal{U}^i(\ell)$ must bind for all $\ell \in [0, \ell^*)$ (a.e.), and every optimal policy must switch from $u(\ell) = \frac{1}{\rho}$ for $\ell < \ell^*$ to $\frac{1}{\rho + \tilde{\alpha}(\ell)}$ in the second-best problem and to $\frac{1}{\rho+1}$ in the first-best problem for $\ell > \ell^*$. It remains to determine the switching point ℓ^* (and establishing uniqueness in the process).

For $\ell < \ell^*$,

$$v'(\ell) = -\frac{r}{\rho} e^{-rt(\ell)}, \qquad t'(\ell) = -\frac{1}{\rho\lambda\ell},$$

so that

$$t(\ell) = C_0 - \frac{1}{\rho\lambda} \ln \ell, \quad \text{or} \quad e^{-rt(\ell)} = C_1 \ell^{\frac{r}{\rho\lambda}},$$

for some constants $C_1, C_0 = -\frac{1}{r} \ln C_1$. Note that $C_1 > 0$; or else, $C_1 = 0$ and $t(\ell) = \infty$ for every $\ell \in (0, \ell^*)$, which is inconsistent with $t(\ell^*) < \infty$. Hence,

$$v'(\ell) = -\frac{r}{\rho} C_1 \ell^{\frac{r}{\rho\lambda}},$$

and so (using $v(0) = 0$),

$$v(\ell) = -\frac{r\lambda}{r + \rho\lambda} C_1 \ell^{\frac{r}{\rho\lambda}+1},$$

for $\ell < \ell^*$. We now substitute v into ϕ for $\ell < \ell^*$ to obtain

$$\phi(\ell) = \lambda C_1 \ell^{\frac{r}{\rho\lambda}} \ell \left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right) - \frac{r\lambda}{r + \rho\lambda} C_1 \ell^{\frac{r}{\rho\lambda}+1}.$$

We now see that the switching point is uniquely determined by $\phi(\ell) = 0$, as ϕ is continuous and C_1 cancels. Rearranging terms,

we obtain

$$\frac{k}{\ell^*} = 1 + \frac{\lambda}{r + \rho\lambda},$$

which leads to the formula for p^* in the proposition (via $\ell = \frac{p}{1-p}$ and $k = \frac{c}{1-c}$). We have identified the unique solution to the program for both first-best and second-best problems and have shown that the optimal threshold p^* applies to both problems.

The second-best policy implements the first-best policy if $p^0 \geqslant c$, since $\bar{\alpha}(\ell) = 1$ for all $\ell \leqslant \ell^0$ in this case. If $p^0 < c$, $\bar{\alpha}(\ell) < 1$ for a positive measure of $\ell \leqslant \ell^0$. Hence, the second-best policy implements a lower and thus slower exploration than does the first-best policy.

As for sufficiency, we use the Arrow sufficiency theorem (Seierstad and Sydsæter 1987, Theorem 5, p. 107). This amounts to showing that the maximized Hamiltonian $\hat{\mathcal{H}}(t, \ell, \nu(\ell)) = \max_{u \in \mathcal{U}^i(\ell)} \mathcal{H}(t, u, \ell, \nu(\ell))$ is concave in t (the state variable) for all ℓ. To this end, it suffices to show that the terms inside the large parentheses in equation (12) are negative for all $u \in \mathcal{U}^i$, $i = FB, SB$. This is indeed the case:

$$\begin{aligned} &1 - \frac{k}{\ell} - \left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right) u(\ell) \\ &\quad \leqslant 1 - \frac{k}{\ell} - \min\left\{\left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right)\frac{1}{1+\rho}, \left(\rho\left(1 - \frac{k}{\ell}\right) + 1\right)\frac{1}{\rho}\right\} \\ &\quad = -\min\left\{\frac{k}{(1+\rho)\ell}, \frac{1}{\rho}\right\} < 0, \end{aligned}$$

where the inequality follows from the linearity of the expression in $u(\ell)$ and the fact that $u(\ell) \in \mathcal{U}^i \subset [\frac{1}{\rho+1}, \frac{1}{\rho}]$ for $i = FB, SB$. The concavity of the maximized Hamiltonian in t therefore follows. We thus conclude that the candidate solution is indeed optimal. □

Appendix B: Proof of Proposition 2

Proof. Write h_t^P for the public history up to time t and h_t for the private history of the designer—which includes whether she received positive feedback by time t. Let $p(h_t)$ denote the designer's belief given her private history.

- Suppose that, given some arbitrary public history h_t^P, the agent is willing to consume at t. Then he is willing to consume if nothing more is said thereafter. In other words, the designer can receive her incentive-unconstrained first-best after such a history. Since this is an upper bound of her payoff, we might assume that she implements it.
- It follows that the only nontrivial public histories are those in which the agents are not yet willing to buy. Given h_t, the designer chooses (possibly randomly) a stopping time τ, which is the time at which she first tells the agent to consume (she then receives her first-best). Let $F(\tau)$ denote the distribution that she uses to tell them to consume at time τ, conditional on her not having had good news by time τ; let $F_t(\tau)$ denote the distribution that she uses if she had positive news precisely at time $t \leqslant \tau$. We will assume for now that the designer emits a single "no consume" recommendation at any given time; we will explain why this is without loss as we proceed.
- Note that as usual, once the designer's belief $p(h_t)$ drops below p^*, she might as well resort to "truth telling," that is, telling the agents to abstain from buying unless she has received conclusive news. This policy is credible, as the agent's belief is always weakly above the belief of the designer who has not received positive news, conditional on h_t^P. Again, it gives the designer her first-best payoff; therefore, given that this is an upper bound, it is the solution. It immediately follows that $F(t^*) > 0$, where t^* is the time required for the designer's belief to reach p^* absent positive news, given that $\mu_t = \rho$ until then. If indeed $F(t) = 1$ for some $t \leqslant t^*$, then the agent will not consume when told to do so at some time $t \leqslant \max\{t' : t' \in \text{supp}(F)\}$. (His belief will have to be no more than his prior for some time below this maximum, which will violate $c > p^0$.) Note that $F_t(t^*) = 1$ for all $t \leqslant t^*$: on reaching time t^*, the designer's belief will make truth-telling optimal, so there is no benefit from delaying good news if it has occurred. Hence, at any time $t > t^*$, conditional on a "no consume" recommendation (so far), it is common knowledge that the designer has not received good news.
- The final observation: whenever agents are told to consume, their incentive constraint must be binding (unless

it is common knowledge that exploration has stopped and the designer has learned that the state is good). If this is not the case for some time t, then the designer can increase $F(t)$ (the probability with which she recommends "consume" at that time, conditional on her not having received good news yet) and raise her payoff, while keeping the hazard rate $\frac{F(\mathrm{d}t')}{1-F(t')}$ fixed at later points in time, which will leave future incentives unchanged.

Let

$$H(\tau) := \int_0^\tau \int_0^t \lambda\rho e^{-\lambda\rho s}(1 - F(s))\,\mathrm{d}s F_s(\mathrm{d}t).$$

This (nondecreasing) function represents the probability that the agent is told to consume for the first time at some time $t \leqslant \tau$, given that the designer has learned that the state is good at some earlier date $s \leqslant t$. Note that H is constant on $\tau > t^*$ and that its support is the same as that of F. Because $H(0) = 0$, $F(0) = 0$ as well.

Let $P(t)$ denote the agent's belief, conditional on the (w.l.o.g., unique) history h_t^P, such that he is told to consume at time t for the first time. For any time t in the support of F, we have

$$P(t) = \frac{p^0\left(H(\mathrm{d}t) + e^{-\rho\lambda t}F(\mathrm{d}t)\right)}{p^0\left(H(\mathrm{d}t) + e^{-\rho\lambda t}F(\mathrm{d}t)\right) + (1 - p^0)F(\mathrm{d}t)}.$$

Indifference implies that

$$P(t) = c, \text{ or } L(t) = k,$$

where $L(t)$ is the likelihood ratio

$$L(t) = \ell^0 \frac{H(\mathrm{d}t) + e^{-\rho\lambda t}F(\mathrm{d}t)}{F(\mathrm{d}t)}.$$

Combining these facts, we have, for any t in the support of F,

$$\left(\frac{k}{\ell^0} - e^{-\rho\lambda t}\right) F(\mathrm{d}t) = H(\mathrm{d}t). \tag{16}$$

[51]

This also holds for any $t \in [0, t^*]$, as both sides are zero if t is not in the support of F. Integrating H by parts yields

$$H(\tau) = \int_0^\tau \lambda\rho e^{-\lambda\rho t}(1 - F(t))F_t(\tau)\,\mathrm{d}t.$$

Integration by parts also yields

$$\int_0^\tau \left(\frac{k}{\ell^0} - e^{-\rho\lambda t}\right) F(\mathrm{d}t) = \left(\frac{k}{\ell^0} - e^{-\rho\lambda\tau}\right) F(\tau) - \int_0^\tau \lambda\rho e^{-\lambda\rho t} F(t)\,\mathrm{d}t.$$

Hence, given that $H(0) = F(0) = 0$, we can rewrite the incentive compatibility constraint for all $t \leqslant t^*$ as:

$$\left(\frac{k}{\ell^0} - e^{-\rho\lambda\tau}\right) F(\tau) = \int_0^\tau \lambda\rho e^{-\lambda\rho t}((1 - F(t))F_t(\tau) + F(t))\,\mathrm{d}t.$$

Note that this implies, given that $F_t(\tau) \leqslant 1$ for all t, $\tau \geqslant t$, that

$$\left(\frac{k}{\ell^0} - e^{-\rho\lambda\tau}\right) F(\tau) \leqslant \int_0^\tau \lambda\rho e^{-\lambda\rho t}\mathrm{d}t = 1 - e^{-\lambda\rho\tau},$$

so that

$$F(t) \leqslant \frac{1 - e^{-\lambda\rho t}}{\frac{k}{\ell^0} - e^{-\rho\lambda t}}, \tag{17}$$

an upper bound that is achieved for all $t \leqslant t^*$ if and only if $F_t(t) = 1$ for all $t \leqslant t^*$.

51. If multiple histories of "no consume" recommendations were considered, a similar equation would hold after any history h_t^P for which "consume" is recommended for the first time at t, replacing $F(\mathrm{d}t)$ and $H(\mathrm{d}t)$ with $\tilde{F}(h_t^P)$ and $\tilde{H}(h_t^P)$, respectively; $\tilde{F}(h_t^P)$ is then the probability that such a history is observed without the designer having received good news yet, while $\tilde{H}(h_t^P)$ is the probability that such a history has been observed after the designer has received good news by then. We then define $F, H : \mathbf{R}_+ \to \mathbf{R}_+$ as (given t) the expectation $F(t)$ (resp. $H(t)$) over all public histories h_t^P, for which t is the first time at which "consume" is recommended. Taking expectations over histories h_t^P gives equation (15). The remainder of the proof is unchanged.

Before writing the designer's objective, let us work out some of the relevant continuation payoff terms. First, t^* is given by our familiar threshold, which is defined by the belief $\ell_{t^*} = k\frac{\lambda\rho+r}{\lambda(1+\rho)+r}$; given that exploration occurs at rate ρ until t^*, conditional on a "no consume" recommendation, we have $e^{-\lambda\rho t^*} = \frac{\ell_{t^*}}{\ell^0}$.

From time t^* onward, if the designer has not recommended to consume, good news has not arrived. Exploration only occurs at rate ρ from that point on. This history contributes to the expected total payoff by

$$p^0(1 - F(t^*))e^{-(r+\lambda\rho)t^*}\frac{\lambda\rho}{r+\lambda\rho}\frac{1-c}{r}.$$

Indeed, this payoff is discounted by the factor e^{-rt^*}. It is positive only if the state is good, and the history is reached with probability $p^0(1 - F(t^*))e^{-\lambda\rho t^*}$: the probability that the state is good, that the designer has not received any good news, and that she has not yet spammed. Finally, conditional on that event, the continuation payoff is equal to

$$\int_0^\infty \lambda\rho e^{-rs-\lambda\rho s}\mathrm{d}s \cdot \frac{1-c}{r} = \frac{\lambda\rho}{r+\lambda\rho}\frac{1-c}{r}.$$

Next, let us consider the continuation payoff if the designer spams at time $\tau \leqslant t^*$. As previously mentioned, she will then experiment at a maximum rate until her belief drops below p^*. The stopping time $\tau + t$ that she chooses must maximize her expected continuation payoff from time τ onward, given her belief p_τ, that is,

$$W(\tau) = \max_t \left\{ p_\tau \left(1 - \frac{r}{\lambda\rho + r}e^{-(\lambda(1+\rho)+r)t}\right)\frac{1-c}{r} - (1-p_\tau)(1-e^{-rt})\frac{c}{r} \right\}.$$

The second term is the cost incurred on agents during time $[\tau, \tau + t]$ when the state is bad. The first is the sum of three terms, all conditional on the state being good: (i) $(1-e^{-rt})\frac{1-c}{r}$, the agents' flow benefit from exploration during $[\tau, \tau + t]$; (ii) $(1-e^{-\lambda(1+\rho)t})e^{-rt}\frac{1-c}{r}$, the benefit after good news has arrived by time $\tau + t$; and (iii) $e^{-(r+\lambda(1+\rho))t}\frac{\lambda\rho}{r+\lambda\rho}\frac{1-c}{r}$, the benefit from background learning after time $\tau + t$ when no good news has arrived by that time. Taking first-order conditions, this function is uniquely

maximized by

$$t(\tau) = \frac{1}{\lambda(1+\rho)} \ln\left(\frac{\ell_\tau}{k}\frac{\lambda(1+\rho)+r}{\lambda\rho+r}\right).$$

Note that we can write $W(\tau) = p_\tau W_1(\tau) - (1-p_\tau)W_0(\tau)$, where $W_1(\tau)$ ($W_0(\tau)$) is the benefit (resp., cost) from the optimal choice of t given that the state is good (resp., bad). Plugging in the optimal value of t gives

$$w_1(\tau) := \frac{rW_1(\tau)}{1-c} = 1 - \frac{r}{\lambda\rho+r}\left(\frac{\ell_\tau}{k}\frac{\lambda(1+\rho)+r}{\lambda\rho+r}\right)^{-1-\frac{r}{\lambda(1+\rho)}},$$

and

$$w_0(\tau) := \frac{rW_0(\tau)}{c} = 1 - \left(\frac{\ell_\tau}{k}\frac{\lambda(1+\rho)+r}{\lambda\rho+r}\right)^{-\frac{r}{\lambda(1+\rho)}}.$$

Note that given no good news by time t, we have $\ell_t = \ell^0 e^{-\rho t}$. It follows that

$$\begin{aligned} &k(1-w_0(t)) - \ell^0 e^{-\lambda\rho t}(1-w_1(t)) \\ &= k\left(1 - \frac{r}{\lambda(1+\rho)+r}\right)\left(\frac{k}{\ell_t}\frac{\lambda\rho+r}{\lambda(1+\rho)+r}\right)^{\frac{r}{\lambda(1+\rho)}} \\ &= Ke^{\frac{r\rho}{1+\rho}t}, \end{aligned} \tag{18}$$

with

$$K := k\frac{\lambda(1+\rho)}{\lambda(1+\rho)+r}\left(\frac{k}{\ell^0}\frac{\lambda\rho+r}{\lambda(1+\rho)+r}\right)^{\frac{r}{\lambda(1+\rho)}}.$$

For future reference, we can use the definition of ℓ_{t^*} to write

$$Ke^{\frac{r\rho}{1+\rho}t^*} = k\frac{\lambda(1+\rho)}{\lambda(1+\rho)+r}\left(\frac{k}{\ell_{t^*}}\frac{\lambda\rho+r}{\lambda(1+\rho)+r}\right)^{\frac{r}{\lambda(1+\rho)}} = k\frac{\lambda(1+\rho)}{\lambda(1+\rho)+r}. \tag{19}$$

We can finally write the objective. The designer chooses $\{F, (F_s)_{s=0}^{t^*}\}$ to maximize

$$J = p^0 \int_0^{t^*} e^{-rt} \left(\frac{1-c}{r} H(\mathrm{d}t) + e^{-\rho\lambda t} W_1(t) F(\mathrm{d}t) \right)$$
$$-(1-p^0) \int_0^{t^*} e^{-rt} W_0(t) F(\mathrm{d}t) + p^0 (1 - F(t^*)) e^{-(r+\lambda\rho)t^*} \frac{\lambda\rho}{r+\lambda\rho} \frac{1-c}{r}.$$

The first two terms are the payoffs in case a "consume" recommendation is made over the interval $[0, t^*]$ and is split according to whether the state is good or bad; the third term is the benefit accruing if no consume recommendation is made by time t^*.

Multiplying by $\frac{r}{1-c} \frac{e^{rt^*}}{1-p^0}$, the objective is rewritten as:

$$\int_0^{t^*} e^{-r(t-t^*)} \left(\ell^0 H(\mathrm{d}t) + \ell^0 e^{-\rho\lambda t} w_1(t) F(\mathrm{d}t) - k w_0(t) F(\mathrm{d}t) \right)$$
$$+ \ell^0 (1 - F(t^*)) e^{-\lambda\rho t^*} \frac{\lambda\rho}{r+\lambda\rho}.$$

We can use equation (16) (as well as $\ell^0 e^{-\lambda\rho t^*} = \ell_{t^*} = k \frac{\lambda\rho + r}{\lambda(1+\rho)+r}$) to rewrite this equation as

$$\int_0^{t^*} e^{-r(t-t^*)} \left(k(1 - w_0(t)) - \ell^0 e^{-\rho\lambda t} (1 - w_1(t)) \right) F(\mathrm{d}t)$$
$$+ (1 - F(t^*)) \frac{\lambda\rho k}{\lambda(1+\rho)+r}.$$

Using equation (18) and ignoring the constant term $\frac{\lambda\rho k}{\lambda(1+\rho)+r}$ (irrelevant for the maximization) gives

$$e^{rt^*} K \int_0^{t^*} e^{-\frac{r}{1+\rho}t} F(\mathrm{d}t) - \frac{\lambda\rho k}{\lambda(1+\rho)+r} F(t^*).$$

Integrating this objective by parts and using $F(0) = 0$ and equation (19), we obtain

$$e^{rt^*} \frac{rK}{1+\rho} \int_0^{t^*} e^{-\frac{r}{1+\rho}t} F(t)\, \mathrm{d}t + \left(k \frac{\lambda(1+\rho)}{\lambda(1+\rho)+r} - k \frac{\lambda\rho}{\lambda(1+\rho)+r} \right) F(t^*).$$

Using equation (19) once more to eliminate K, we finally obtain

$$\frac{\lambda k}{\lambda(1+\rho)+r}\left(\int_0^{t^*} re^{-\frac{r}{1+\rho}(t-t^*)}F(t)\,\mathrm{d}t + F(t^*)\right).$$

Note that this objective function is increasing point-wise in $F(t)$ for each $t \leqslant t^*$. Hence, it is optimal to set F as given by its upper bound provided by equation (17), for all $t \leqslant t^*$,

$$F(t) = \frac{\ell^0(1-e^{-\lambda\rho t})}{k-\ell^0 e^{-\rho\lambda t}},$$

and for all $t \leqslant t^*$, $F_t(t) = 1$.

To prove the last statement (on the average speed of exploration), fix any $t \leqslant t^*$. Under optimal public recommendations, spam is triggered at s according to $F(s)$ and lasts until t, unless the posterior reaches p^*. Let $T(s)$ be the time at which the latter event occurs if spam is triggered at s. Then, the expected level of exploration performed by time t under public recommendations is as follows:

$$\int_0^t (\min\{t, T(s)\} - s)\mathrm{d}F(s) \leqslant \int_0^t (t-s)\mathrm{d}F(s) = \int_0^t F(s)\,\mathrm{d}s$$
$$= \int_0^t \frac{\ell^0 - \ell^0 e^{-\lambda\rho s}}{k-\ell^0 e^{-\lambda\rho s}}\mathrm{d}s < \int_0^t \frac{\ell^0-\ell_s}{k-\ell_s}\mathrm{d}s = \int_0^t \hat{\alpha}(\ell_s)\,\mathrm{d}s,$$

where ℓ_s is the likelihood ratio at time s under the optimal private recommendation. The first equality follows from integration by parts, and the inequality holds because $\ell_s = \ell^0 e^{-\lambda\int_0^s(\tilde{\alpha}(\ell_{s'})+\rho)\,\mathrm{d}s'} < \ell^0 e^{-\lambda\rho s}$. □

COLUMBIA UNIVERSITY
YALE UNIVERSITY AND TOULOUSE SCHOOL OF ECONOMICS

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

REFERENCES

Aumann, Robert J., Michael Maschler, and Richard E. Stearns. *Repeated Games With Incomplete Information*, (Cambridge: MIT Press, 1995).

Avery, Christopher, Paul Resnick, and Richard Zeckhauser. "The Market for Evaluations," *American Economic Review*, 89 (1999), 564–584.
Banerjee, Abhijit V. "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107 (1993), 797–817.
Bergemann, Dirk, and Deran Ozmen. "Optimal Pricing with Recommender Systems," *Proceedings of the 7th ACM Conference on Electronic Commerce*, (2006), 43–51.
Berry, Donald A. "Adaptive Clinical Trials in Oncology." *National Review of Clinical Oncology*, 9 (2011), 199–207.
Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100 (1992), 992–1026.
Cesari, Lamberto. *Optimization–Theory and Applications. Problems with ordinary differential equations, Applications of Mathematics*, 17 (Berlin-Heidelberg-New York: Springer-Verlag, 1983).
Che, Yeon-Koo, and Johannes Hörner. "Optimal Design for Social Learning," mimeo (Columbia and Yale Universities, 2015).
Chow, Shein-Chung, and Mark Chang. "Adaptive Design Methods in Clinical Trials? a Review," *Orphanet Journal of Rare Diseases*, 3 (2008), 1750–1172.
Dai, Weijia et al. "Optimal Aggregation of Consumer Ratings: An Application to Yelp.com," working paper, Harvard Business School, 2014.
Ely, Jeffrey C. "Beeps," *American Economic Review*, 107 (2017), 31–53.
Ely, Jeffrey C., Alexander Frankel, and Emir Kamenica. "Suspense and Surprise," *Journal of Political Economy*, 123 (2015), 215–260.
Frick, Mira, and Yuhta Ishii. "Innovation Adoption by Forward-Looking Social Learners," working paper, Harvard, 2014.
Gershkov, Alex, and Balazs Szentes. "Optimal Voting Schemes with Costly Information Acquisition," *Journal of Economic Theory*, 144 (2009), 36–68.
Gittins, John C., Kevin D. Glazebrook, and Richard Weber. *Multi-armed Bandit Allocation Indices*, 2nd ed. (New York: Wiley and Sons, 2011).
Halac, Marina, Navin Kartik, and Qingmin Liu. "Contests for Experimentation," *Journal of Political Economy*, 125 (2017), 1523–1569.
Halac, Marina, Navin Kartik, and Qingmin Liu. "Optimal Contracts for Experimentation," *Review of Economic Studies*, 83 (2016), 1040–1091.
Hörner, Johannes, and Larry Samuelson. "Incentives for experimenting agents," *RAND Journal of Economics*, 44 (2013), 632–663.
Jindal, Nitin, and Bing Liu. "Opinion Spam and Analysis," *Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM*, (2008), 219–230.
Kamenica, Emir, and Matthew Gentzkow. "Bayesian Persuasion," *American Economic Review*, 101 (2011), 2590–2615.
Keller, Godfrey, and Sven Rady. "Breakdowns," *Theoretical Economics*, 10 (2015), 175–202.
Keller, Godfery, Sven Rady, and Martin Cripps. "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73 (2005), 39–68.
Klein, Nicolas, and S. Rady. "Negatively Correlated Bandits," *Review of Economic Studies*, 78 (2011), 693–732.
Kremer, Ilan, Yishay Mansour, and Motty Perry. "Implementing the "Wisdom of the Crowd"," *Journal of Political Economy*, 122 (2014), 988–1012.
Luca, Michael, and Georgios Zervas. "Fake It Till You Make It: Reputation, Competition and Yelp Review Fraud," *Management Science*, 62 (2016), 3412–3427.
Mansour, Yishay, Aleksandrs Slivkins, and Vasilis Syrgkanis. "Bayesian incentive-compatible bandit exploration," In *15th ACM Conf. on Electronic Commerce (EC)*, 2015.
Mansour, Yishay et al. "Bayesian Exploration: Incentivizing Exploration in Bayesian Games," mimeo, Microsoft Research, 2016.
Mayzlin, Dina, Yaniv Dover, and Judith A. Chevalier. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (2014), 2421–2455.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. "Eliciting Informative Feedback: The Peer-Prediction Method," *Management Science*, 51 (2004), 1359–1373.
Ostrovsky, Michael, and Michael Schwarz. "Information Disclosure and Unraveling in Matching Markets," *American Economic Journal: Microeconomics*, 2 (2010), 34–63.
Pandey, Sandeep et al. "Shuffling a Stacked Deck: the Case for Partially Randomized Ranking of Search Engine Results," In *Proceedings of the 31st International Conference on Very Large Data Bases*, 2005.
Papanastasiou, Yiangos, Kostas Bimpikis, and Nicos Savva. "Crowdsourcing Exploration," *Management Science*, forthcoming, 2017.
Rayo, Luis, and Ilya Segal. "Optimal Information Disclosure," *Journal of Political Economy*, 118 (2010), 949–987.
Renault, Jerome, Eilon Solan, and Nicolas Vieille. "Optimal Dynamic Information Provision," 2014, arXiv:1407.5649 [math.PR].
Rothschild, Michael. " A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9 (1974), 185–202.
Schafer, J. Ben, Joseph Konstan, and John Riedl. "Recommender Systems in e-commerce," *Proceedings of the 1st ACM conference on Electronic commerce*, (1999), 158–166.
Seierstad, Atle, and Knut Sydsæter. *Optimal Control Theory with Economic Applications* (Amsterdam: North-Holland, 1987).
Smith, Lones, and Peter Sørensen. "Pathological Outcomes of Observational Learning," *Econometrica*, 68 (2000), 371–398.
Smith, Lones, Peter Sørensen, and Jianrong Tian. "Informational Herding, Optimal Experimentation, and Contrarianism," mimeo, University of Wisconsin, 2016.
Trippa, Lorenzo et al. "Bayesian Adaptive Randomized Trial Design for Patients With Recurrent Glioblastoma," *Journal of Clinical Oncology*, 30 (2012), 3258–3263.