Crowdsourcing Exploration

Yiangos Papanastasiou
Haas School of Business, University of California Berkeley · yiangos@haas.berkeley.edu

Kostas Bimpikis
Graduate School of Business, Stanford University · kostasb@stanford.edu

Nicos Savva
London Business School · nsavva@london.edu

Motivated by the proliferation of online platforms that collect and disseminate consumers' experiences with alternative substitutable products/services, we investigate the problem of optimal information provision when the goal is to maximize aggregate consumer surplus. We develop a decentralized multi-armed bandit framework where a forward-looking principal (the platform designer) commits upfront to a policy that dynamically discloses information regarding the history of outcomes to a series of short-lived rational agents (the consumers). We demonstrate that consumer surplus is non-monotone in the accuracy of the designer's information-provision policy. Because consumers are constantly in "exploitation" mode, policies that disclose accurate information on past outcomes suffer from inadequate "exploration." We illustrate how the designer can (partially) alleviate this inefficiency by employing a policy that strategically obfuscates the information in the platform's possession – interestingly, such a policy is beneficial despite the fact that consumers are aware of both the designer's objective and the precise way by which information is being disclosed to them. More generally, we show that the optimal information-provision policy can be obtained as the solution of a large-scale linear program. Noting that such a solution is typically intractable, we use our structural findings to design an intuitive heuristic that underscores the value of information obfuscation in decentralized learning. We further highlight that obfuscation remains beneficial even if the designer can directly incentivize consumers to explore through monetary payments.

Key words: Bayesian social learning, information provision, exploration vs. exploitation, Gittins index

1. Introduction

In the short span of just over ten years since the term was first coined, *crowdsourcing* has dramatically increased the availability of information that is relevant to a range of everyday decisions. Drawing on the experiences of members of their online communities, platforms hosting specialized content now exist that assist their users in choosing between alternative service providers (e.g., *Yelp*), products (e.g., *Epinions*), driving routes (e.g., *Waze*), physicians (e.g., *RateMDs*), holiday destinations (e.g., *TripAdvisor*), and so on.

Motivated by the proliferation of these platforms, we study an inherent inefficiency of social learning in settings characterized by decentralized information generation. In particular, the critical feature of the settings we consider is that new information is generated by individual agents

as a by-product of a self-interested choice among alternative options, and without regard for the informational externality that their experience exerts on the choices and welfare of future agents. From the perspective of the society as a whole, this translates into inefficiency which may manifest, for example, as situations where "winners keep winning," while less-explored but potentially superior options are not afforded the chance to demonstrate their worth.[1]

Since the choices of individual agents – and therefore the new information they generate – are directly related to the information they observe prior to their choice, alternative modes of information provision may result in different modes of information generation. This notion is the focus of our paper.

We consider a simple model in which a population of homogeneous agents (referred to throughout as "consumers") visit a platform sequentially, observe information pertaining to the experiences of their predecessors, and choose among alternative options ("service providers"). After receiving service from her chosen provider, each consumer reports to the platform whether the service she received was a success or a failure. Upon being selected, each provider generates a successful service outcome with a fixed probability that represents the provider's *quality* – this probability is unknown throughout, but can be learned (in the Bayesian sense) by observing the provider's history of service outcomes. At any time, the history of service outcomes is recorded by the platform, but is not necessarily observable to the consumers. Instead, there is a principal ("platform designer") who commits upfront to an *information-provision policy* which specifies the information posted on the platform at any time, given any possible recorded history. The designer's objective is to maximizing the consumers' aggregate discounted surplus over an infinite horizon. By contrast, each consumer seeks to maximize only her individual surplus through her choice of provider.

At the core of our model is the friction between the objectives of the forward-looking designer and the short-sighted consumers: the designer would like consumers to make decisions (i.e., provider choices) that benefit not only themselves (through their service experience) but also their successors (through the knowledge that their experience generates). Had consumers' actions been under the designer's full control, the designer would be faced with a classic instance of the multi-armed bandit problem (MAB; see Gittins et al. (2011)). The solution to this classic problem, which resolves the well-known "exploration-versus-exploitation" trade-off, is due to Gittins and Jones (1974), and consists of using in each period the arm of highest *Gittins index*. The challenge faced by the designer is to structure the information on which consumers base their actions, so as to

[1] Similar inefficiencies may also arise in "offline" instances of decentralized learning. For example, progress in research may be hampered by individual researchers' incentives to exploit existing knowledge with a view towards publication, rather than explore new research methods/topics; experimentation in new product development may suffer from R&D managers' preference to use proven methods that guarantee finished products; etc.

influence their decisions in a manner that serves the goal of consumer-surplus maximization. Doing so is challenging, because consumers are not naive: they are aware of both the designer's objective and the way in which information is being disclosed to them. Thus, the designer's effectiveness in managing the dynamic exploration-exploitation trade-off is directly linked to his ability to design an information-provision policy that "persuades" the self-interested consumers to take his desired actions.

We analyze first a special case of our model where there are two providers, one of which has a known quality, and use this case to highlight the qualitative nature of optimal policies. First, we evaluate the performance of policies belonging to the two extreme modes of information provision: "no-information" (NI), where the platform conceals all information in its possession at all times, and "full-information" (FI), where the platform discloses precisely all information in its possession at all times. We demonstrate that FI outperforms NI, but fails to achieve first best (i.e., the payoff achieved when the consumers' actions are under the designer's full control). The latter observation follows from existing knowledge on the MAB problem: consumers' choices under FI reduce to the "myopic" policy in the classic MAB, which is known to be suboptimal.

More importantly, we show that the designer (subject to a simple condition) can in fact achieve first best in the decentralized system, by employing a policy which is deliberately *less-than-fully informative* (i.e., a policy which lies, in a qualitative sense, between the two extremes of NI and FI). Under the optimal policy, rather than providing consumers with a precise history of service outcomes, the platform employs a coarser, "many-to-few" information structure: several histories are merged and mapped to the same configuration of information (e.g., this may take the form of a simple recommendation or a simple ranking of the alternative providers). We make precise the manner by which such policies are structured, and demonstrate how the consumers' Bayesian interpretation of the information they observe causes them to choose the designer's desired provider – interestingly, this occurs even though consumers know the designer's objective and the policy by which information is being disclosed to them.

We then turn our attention to the more involved problem of designing an information-provision policy for the designer's general problem (i.e., where the qualities of all providers are ex ante unknown). Here, we demonstrate that first best is typically infeasible, but that optimal policies maintain the feature of information obfuscation. We illustrate that the designer's problem can be formulated as a Constrained Markov Decision Process (CMDP) and show that the optimal policy can be obtained as the solution of a large-scale linear program. While such a solution is typically intractable computationally, we leverage the problem's structure to propose a heuristic solution which underscores the value of information obfuscation in decentralized learning. In particular,

we observe that our heuristic – which implements information obfuscation only suboptimally – performs close to first best, and significantly better than FI, in all our numerical experiments.[2]

Finally, we consider an extension of our model where the designer, in conjunction with his information-provision policy, can also employ monetary subsidies to directly incentivize the consumers to engage in exploration. Although the problem of optimally combining information provision with subsidies appears to be significantly more complex than its information-only counterpart, we show that the dominant class of policies is one that involves information obfuscation, consistent with the rest of our analysis. Specifically, we establish that less-than-fully informative policies allow the designer to achieve any feasible level of consumer surplus at a minimum total subsidy cost – this finding highlights the importance of information provision over and above more traditional means of resolving incentive misalignments, such as monetary transfers.

2. Related Literature

The multi-armed bandit (MAB) problem is recognized as the epitome of the *exploration-versus-exploitation* trade-off. In the classic version of the MAB problem (see Gittins et al. (2011)), a forward-looking decision maker chooses sequentially between alternative arms, each of which generates rewards according to an ex ante unknown distribution. Every time an arm is chosen, the decision maker receives a reward which, apart from its intrinsic value, is used to learn about the arm's underlying reward distribution. At any decision epoch, the decision maker may choose the arm he currently believes to be superior (exploitation), or an alternative arm with the goal of acquiring knowledge that can be used to make better-informed decisions in the future (exploration). Since its inception, the MAB problem has been extended in multiple directions to investigate exploration-versus-exploitation trade-offs that are encountered in various practical settings. For example, Caro and Gallien (2007) study dynamic assortment of seasonal goods in the presence of demand learning, while Bertsimas and Mersereau (2007) consider a marketer learning the efficacy of alternative marketing messages.[3]

In most existing applications of the MAB, a single decision maker dynamically decides on the actions to be taken while observing the outcomes of his past actions. By contrast, the problem we study in the present paper is essentially a decentralized MAB: there is a forward-looking principal

[2] The disclosure of information on the basis of coarse, less-than-fully transparent information structures appears consistent with practical observations. For example, *TripAdvisor* and *Yelp* rank providers in a manner that sometimes appears to be inconsistent with the underlying content of consumer reviews (e.g., *TripAdvisor* 2013); *Booking.com* includes in its rankings only providers that have received at least a specific number of reviews, thus withholding the initial information it receives from its users; *Netflix* and *Pandora* deliver recommendations without providing details on how these recommendations have been generated.

[3] Alizamir et al. (2013), Anand et al. (2011) and Kostami and Rajagopalan (2013) study a related trade-off between improving the quality of service and reducing waiting times in congested systems.

(the designer) who seeks to maximize the sum of discounted rewards, while actions are taken by a series of short-lived agents (the consumers). In related work, Lobel et al. (2015) consider the problem faced by a forward-looking firm selling its products through a myopic salesforce, and propose an asymptotically regret-optimal strategy that involves the firm sequentially "dropping" products deemed to be suboptimal. A similar setup to ours is used in Frazier et al. (2014) to investigate how the principal can incentivize the agents to take his desired actions by offering direct monetary payments. In their setting, the history of actions and outcomes is assumed to be common knowledge and there is, therefore, no attempt at investigating the issue of optimal information provision. In our model, the only lever that the principal uses to influence consumers' actions is his information-provision policy.

In the latter respect our work is related to, but quite distinct from, the well-developed literature on "cheap talk" (e.g., Crawford and Sobel 1982, Allon et al. 2011). In cheap-talk games, the principal privately observes the realization of an informative signal, after which he (costlessly) communicates any message he wants to the agent. In this work, there is emphasis on how the message received by the agent is interpreted, and whether any information can be credibly transmitted by the principal. By contrast, the principal in our setting commits ex ante to an information-provision policy which maps realizations of the informative signal to messages. Once this policy has been decided and implemented, the principal cannot manipulate the information he discloses (e.g., by misrepresenting the signal realization). In this case, there is no issue of how the agents will interpret the messages; rather, our focus is on how the principal should structure credible messages in a manner that internalizes the misalignment between his and the consumers' objectives.

Our paper is therefore more in the spirit of the recent stream of literature that examines how a principal can design/re-structure informative signals in ways that render agents ex ante more likely to take desirable actions. Bimpikis and Drakopoulos (2015) find that in order to overcome the adverse effects of free-riding, teams of agents working separately towards the same goal should initially not be allowed to share their progress for some pre-determined amount of time. Bimpikis et al. (2015) investigate innovation contests and demonstrate how award structures should be designed so as to implicitly enforce information-sharing mechanisms that incentivize participants to remain active in the contest. Kamenica and Gentzkow (2011) and Rayo and Segal (2010) illustrate an explicit technique for structuring informative signals – referred to as "Bayesian persuasion" – in static (i.e., one-shot) settings. In the context of decentralized learning, variants of Bayesian persuasion are employed in two recent papers. Kremer et al. (2013) focus on eliciting experimentation in an environment where outcomes are deterministic, while Che and Hörner (2014) consider a single-product setting where a designer at any time optimally "spams" a fraction of consumers to learn

about the product's quality. In both papers, once any information is received by the designer, product quality is perfectly revealed; as a result, there is initially a full-exploration period, which is then followed by full exploitation. By contrast, the main difficulty faced by the designer in our model is to effectively manage a dynamic exploration-exploitation trade-off in a stochastic environment.

The information accumulated by the platform in our model is continuously updated via consumers' reported experiences, which (through the designer's information-disclosure policy) influence the decisions of subsequent consumers. In this respect, our paper connects to the work on social learning. The basic setup involves agents (e.g., consumers) that are initially endowed with private information regarding some unobservable state of the world (e.g., product quality). When actions (e.g., purchase decisions) are taken sequentially and are commonly observable, the seminal papers by Banerjee (1992) and Bikhchandani et al. (1992) demonstrate that herds may be triggered, whereby agents rationally disregard their private information and simply mimic the action of their predecessor. This classic paradigm has since been extended in multiple directions to investigate, for example, learning in social networks (e.g., Acemoglu et al. 2011) and learning among agents with heterogeneous preferences (e.g., Lobel and Sadler 2015).

While the above papers focus on studying features of the learning process itself, another stream of literature investigates how firms can use their operational levers to steer the social-learning process to their advantage. Bose et al. (2006) and Ifrach et al. (2014) investigate dynamic pricing in the presence of social learning that occurs on the basis of actions (i.e., purchase decisions) and outcomes (i.e., product reviews), respectively. Veeraraghavan and Debo (2009) and Debo et al. (2012) consider how customers' queue-joining behavior depends on observable queue-length, and how service-rate decisions may be used to influence this behavior. Papanastasiou and Savva (2016) and Yu et al. (2013) highlight how pricing policies are affected by the interaction between product reviews and strategic consumer behavior (see also Swinney (2011)), while Papanastasiou et al. (2014) illustrate the beneficial effects of scarcity strategies when consumers learn according to an intuitive non-Bayesian rule. We contribute to this literature by investigating how the firm (platform) can influence consumer decisions and learning through its information-provision policy, a lever which may also be used in conjunction with other operational levers (e.g., pricing, inventory).

Finally, this paper also contributes to a recent line of work which studies operational decisions in the context of Internet-enabled business models. Among others, Marinesi and Girotra (2013) examine how customer voting systems should be designed when firms seek to acquire information to improve pricing and product-design decisions; Ye et al. (2015) investigate how an online retailer should combine sponsored-search marketing with dynamic pricing; Balseiro et al. (2014) consider the problem faced by a web publisher in deciding how to allocate advertising slots between spot markets (ad exchanges) and pre-arranged contracts (reservations). In this paper, we investigate

how the information-provision policy of an online platform can be used to influenced the decisions of its users.

3. Model Description

We consider a decentralized learning setting, where a series of agents interact with a principal who manages the disclosure of information regarding the experiences of their predecessors. For concreteness, we anchor our exposition in the example of an online platform which is operated by a designer and is used by consumers to assist with their choice of service provider. We suppose that the marketplace consists of two providers, A and B; let $S = \{A, B\}$.[4] Each provider $i \in S$ is fully characterized by a probability p_i which represents the provider's service quality. Upon using provider i, a consumer receives reward equal to one with probability p_i, and equal to zero with probability $1 - p_i$; that is, service outcomes constitute independent draws from a Bernoulli distribution with success probability p_i. Initially, p_i is known to the designer and the consumers only to the extent of a common prior belief, which is expressed in our model through a Beta random variable with shape parameters $\{s_1^i, f_1^i\}$, $s_1^i, f_1^i \in \mathbb{Z}_+$.[5,6]

At the beginning of each time period $t \in T$, $T = \{1, 2, ...\}$, a single consumer visits the platform, observes information pertaining to the experiences of past consumers, and chooses a provider. We assume that upon completion of service, and before the end of period t, the consumer reports to the platform whether her experience was positive or negative (i.e., a Bernoulli success or failure). At any time t, the knowledge accumulated by the platform is summarized by the *information state* (henceforth "state") $x_t = \{x_t^A, x_t^B\}$, where $x_t^i = \{s_t^i, f_t^i\}$ and s_t^i (f_t^i) is the accumulated number of successful (failed) service outcomes for provider i up to period t (this includes the initial successes and failures, s_1^i and f_1^i, specified in the prior belief). When the system state is x_t, the Bayesian posterior belief over the quality p_i is $Beta(s_t^i, f_t^i)$, and the expected reward of the next customer to use provider i is $r(x_t, i) = \frac{s_t^i}{s_t^i + f_t^i}$ (e.g., see DeGroot 2005, Chapter 9).

At any time, the history of service outcomes (i.e., the system state x_t) is not directly observable to the consumers. Instead, there is a platform designer who *commits* upfront to a "messaging policy" that acts as an instrument of information-provision to the consumers.[7] This policy specifies the *message* that is displayed on the platform, given any underlying system state; in §7.2, we extend

[4] The general analysis in §6 can be readily extended to the case of $|S| > 2$ providers.

[5] The probability density function of a $Beta(s, f)$ random variable is given by $g(x; s, f) = \frac{x^{s-1}(1-x)^{f-1}}{B(s,f)}$, for $x \in [0, 1]$.

[6] The platform and the consumers hold the same prior belief, so that platform actions (e.g., choice of information-provision policy) do not convey any additional information on provider quality to the consumers (e.g., Bergemann and Välimäki 1997, Bose et al. 2006, Papanastasiou and Savva 2016).

[7] Commitment is a reasonable assumption in the context of online platforms, where information provision occurs on the basis of pre-decided algorithms and the large volume of products/services hosted renders ad-hoc adjustments of the automatically-generated content prohibitively costly (see also §5.4, where this assumption is relaxed).

our analysis to the case where messages may also be accompanied by monetary payments.[8] The designer's objective in choosing his messaging policy is to maximize the expected sum of consumers' discounted rewards over an infinite horizon (i.e., consumer surplus), applying a discount factor of $\delta \in [0,1)$.[9] Consumers are modelled as homogeneous, short-lived, rational agents. In our main analysis, we assume that consumers know the period of their arrival; we relax this assumption in §7.1. Upon visiting the platform, each consumer observes a message generated by the designer's policy and chooses a service provider with the goal of maximizing her individual expected reward.

The designer's choice of messaging policy, along with the consumers' choices of service provider in response to this policy, simultaneously govern the dynamics of both the learning process and the consumers' reward stream.

4. Analysis Preliminaries

Equilibrium and Model Dynamics We begin our analysis by formalizing the strategic interaction between the designer and the consumers. There are two main features of this interaction. First, the designer's *messaging policy*, which takes the platform state as an input and generates a message to be displayed by the platform to the next incoming consumer. Second, the consumers' *choice strategy*, which takes the platform's message in any given period as an input and determines the consumer's action (choice of provider).

Let $X \subseteq \mathbb{Z}_+^4$ denote the set of possible states of the platform such that $x_t \in X$ for all $t \in T$, and define the discrete set M of feasible messages that the platform can display to an incoming consumer in period t (see footnote 8). A messaging policy $g(\cdot)$ is a (possibly stochastic) mapping from the set of states X to the set of messages M; that is, a messaging policy g associates with each state $x_t \in X$ a probability $P(g(x_t) = m)$ that message $m \in M$ is displayed on the platform. Let $\mathcal{G}$ be the set of possible messaging policies. In each period t, a single consumer enters the system, observes the platform's message and chooses a service provider from the set S. The period-t consumer's choice strategy, denoted by $c_t(\cdot)$, is a mapping from the set of messages M to the set of service providers S. Let $\mathcal{C}_t$ be the set of possible choice strategies for the period-t consumer, and define $c(\cdot) := [c_1(\cdot), c_2(\cdot), ...]$.

The designer's messaging policy g along with the consumers' choice strategy c generate a *controlled Markov chain* characterized by the stochastic state-action pairs $\{(x_t, y_t); t \in T\}$, where the

[8] The generic term "message" refers to a specific configuration of information that is observed by the consumer; examples of messages include detailed outcome histories (i.e., distributions of consumer reviews), relative rankings of providers, recommendations for a specific product, etc.

[9] More generally, our analysis is relevant for cases where the platform has a different (e.g., longer-run) objective than its users. Similar objective functions as ours are commonly employed in decentralized learning models (e.g., Frazier et al. 2014, Lobel et al. 2015).

actions y_t that accompany the states x_t are determined by the designer's policy and the consumers' strategy via $y_t = c_t(g(x_t))$. When the state of the system is x_t, the expected reward of a consumer that uses provider i is $r(x_t, i) = \frac{s_t^i}{s_t^i + f_t^i}$. Transitions between system states occur as follows. The initial state x_1 is determined by the prior belief over the two providers; when the state of the system is x_t and action y_t is chosen by the period-t consumer, the state in period $t+1$, $x_{t+1} = \{x_{t+1}^A, x_{t+1}^B\}$ is determined as follows

$$x_{t+1}^i = x_t^i \text{ for } i \neq y_t, \quad x_{t+1}^i = \begin{cases} \{s_t^i + 1, f_t^i\} \text{ w.p. } \; r(x_t, i) \\ \{s_t^i, f_t^i + 1\} \text{ w.p. } 1 - r(x_t, i) \end{cases} \quad \text{for } i = y_t.$$

The above transition probabilities reflect the learning dynamics of the system: new information regarding the quality of provider i is generated in period t only if the provider is chosen by the period-t consumer.[10]

The sequence of events in our model is described in reverse chronological order as follows. Each consumer observes the designer's messaging policy and chooses a choice strategy c_t to maximize her individual expected reward. In particular, the period-t consumer's response to message m, $c_t^*(m)$ maximizes

$$E_{x_t}\left[r(x_t, c_t) \mid g(x_t) = m\right].^{11}$$

At the beginning of the time horizon, the designer (taking into account the consumers' response to any messaging policy), commits to a policy that maximizes the expected sum of consumers' discounted rewards. In particular, the designer's messaging policy $g^*(x_t)$ maximizes

$$E\left[\sum_{t \in T} \delta^{t-1} r(x_t, y_t)\right], \text{ for } y_t = c_t^*(g(x_t)).$$

Incentive-Compatible Recommendation Policies In general, multiple equilibria exist that result in the same payoff for the designer and the consumers, and the same dynamics in the learning process, not least because the same information can be conveyed from the designer to the consumers through a multitude of interchangeable messages contained in M. We follow Allon et al. (2011) in referring to such equilibria as being "dynamics-and-outcome equivalent" (DOE). In our analysis, we will employ the result of Lemma 1 below to avoid redundancies in exposition and focus attention on the informational content of equilibria, rather than on the alternative ways in which these equilibria can be implemented. Before stating the lemma, we define a subclass of messaging policies which we refer to as "incentive-compatible recommendation policies."

[10] Note that for the case of a Bernoulli reward process the current probability of success (i.e., the Bayesian probability of the next trial being a success given the current state of the system) is equal to the immediate expected reward, $r(x_t, i)$ (e.g., Gittins et al. 2011).

[11] This expectation can be computed by the period-t consumer, since the ex ante probability that the state in period t is x_t (i.e., unconditional on the message $g(x_t)$) is known to the consumer through her knowledge of the designer's policy in previous periods and the preceding consumers' best response to this policy.

Definition 1 (ICRP: Incentive-Compatible Recommendation Policy). A recommendation policy is a messaging policy defined as

$$g(x_t) = \begin{cases} A & \text{w.p.} \quad q_{x_t} \\ B & \text{w.p. } 1 - q_{x_t}, \end{cases} \tag{1}$$

where $q_{x_t} \in [0,1]$ for all $x_t \in X$. A recommendation policy is said to be incentive-compatible if for all $x_t \in X$, $t \in T$, we have $c_t^*(g(x_t)) = g(x_t)$.

Put simply, under an ICRP the platform recommends either provider A or provider B to the period-t consumer, and the consumer finds it Bayes-rational to follow this recommendation. We may now state the following result, which is analogous to the revelation principle in the mechanism-design literature, and suggests that any feasible platform payoff can be achieved through some ICRP.

Lemma 1. *For any arbitrary messaging policy g, there exists an ICRP g' which induces a DOE equilibrium in the game between the designer and the consumers.*

All proofs are provided in Appendix B. In the proof of Lemma 1, we illustrate how an ICRP can be constructed from any messaging policy so as to induce an equivalent choice strategy from the consumers. Essentially, the process consists of replacing the original messages with recommendations of the consumer actions that these messages would induce; examples of the correspondence between messaging policies and ICRPs appear in the following sections.

First Best Before analyzing the decentralized system, let us consider how the designer would direct individual consumers to the two providers, had consumers been under his *full control*. The solution to the designer's full-control problem is due to Gittins and Jones (1974) and consists of directing consumers in each period to the provider with the highest Dynamic Allocation Index, also known as the *Gittins Index*. The Gittins index for service i when in state z^i is denoted by $G_i(z^i)$ and given by

$$G_i(z^i) = \sup_{\tau > 0} \frac{E\left[\sum_{t=0}^{\tau-1} \delta^t r(x_t^i, i) \mid x_0^i = z^i\right]}{E\left[\sum_{t=0}^{\tau-1} \delta^t \mid x_0^i = z^i\right]}, \tag{2}$$

where τ is a past-measurable stopping time (i.e., measurable with respect to the information obtained up to time τ) and $r(x_t^i, i)$ is the instantaneous expected reward of provider i in state x_t^i.

In the decentralized system, the designer's ability to direct consumers to his desired provider will be limited by the consumers' self-interested behavior. Each consumer knows (i) the prior belief summarized by the initial state, x_1; (ii) the time period, t (relaxed in §7.1); and (iii) the designer's messaging policy, g. Upon visiting the platform, the consumer observes a message m, updates her belief over the current system state, x_t, and selects the provider which maximizes her individual

expected reward. As a consequence, the designer will be able to achieve first-best only if he can design a messaging policy which induces consumers to make Gittins-optimal decisions in all periods and in all system states – a sufficient condition for at least one such messaging policy to exist is the existence of an ICRP which always recommends the provider of highest Gittins index.

Throughout the following analysis we will refer to provider choices that are desirable from the platform's perspective as being "system-optimal."

5. Simple Case: Incumbent Provider B

We analyze first a simple version of our model, where there is one provider whose quality is ex ante unknown (provider A) and one incumbent provider whose quality is known with certainty (provider B). The analysis of this section serves to build intuition and highlight the main features of optimal messaging policies, within a simplified setting which is amenable to direct analytical treatment. The designer's general problem is considered subsequently in §6.

Let the prior belief over provider A's service quality be $Beta(s_1^A, f_1^A)$ and recall that the expected reward of a consumer who chooses service A in period t is given by $r(x_t, A) = \frac{s_t^A}{s_t^A + f_t^A}$, where x_t is the system state. For provider B, let the service quality be known and equal to p_B, such that the expected immediate reward of a consumer who chooses service B at any time t is simply $r_B := r(x_t, B) = p_B$. We suppose, for simplicity, that if the designer and/or the consumers are indifferent between the two providers, provider B is preferred.

5.1. First Best

It will be useful to first characterize the provider choices which result when the full-control policy described in §4 is applied to the simplified setting considered here. To begin, note that since the quality of provider B is known with certainty, the provider has a constant Gittins index of $G_B := G_B(x_t) = r_B$ (Gittins et al. 2011, Chapter 7). Therefore, if the designer finds it system-optimal to use service B in some period $t = k$, then this must also be the case in all subsequent periods $t > k$. As a result, system-optimal provider choices can be described in terms of "success thresholds" for provider A.

LEMMA 2. *System-optimal provider choices are characterized as follows:*

(i) If $G_A(x_1) \leq G_B$, then any experimentation with service A is suboptimal; that is, it is system-optimal to use service B in all periods $t \in T$.

(ii) If $G_A(x_1) > G_B$, then it is system-optimal to experiment with service A at least once in period $t = 1$. In any period $t > 1$, there exists an integer $s^*(t)$ such that if $s_t^A \geq s^*(t)$ it is system-optimal to continue experimentation with service A in period t, while if $s_t^A < s^*(t)$ it

is system-optimal to choose service B in period t and forever after. The period-t threshold $s^*(t)$ is uniquely defined by

$$s^*(t) = \{\min s_t^A \ : \ s_t^A, f_t^A \in \mathbb{Z}_+^2, \ (s_t^A - s_1^A) + (f_t^A - f_1^A) = t-1, \ G_A(x_t) > G_B\}.$$

In the first case of Lemma 2, experimentation with provider A is unattractive from the onset. Noting that $G_B = r_B$, intuitively, if the incumbent's quality is sufficiently high, then there is no rationale for the designer to engage in any experimentation with the new provider. In the second case of the lemma, experimentation with the new provider is attractive for the designer to begin with, but may cease to be so as more information about the new provider's quality is acquired in the early periods of the horizon: in any period (and provided experimentation with provider A has not already been terminated), there exists a threshold on the number of accumulated successful outcomes with provider A that is required for A to remain the system-optimal choice.

5.2. The Two Extreme Modes of Information Provision

Let us now return to the decentralized model. Each consumer (knowing the platform's messaging policy) receives a message and chooses a provider to maximize her individual expected reward. In terms of the informational content of alternative policies that may be employed by the designer, there are two extreme modes of information provision; we consider each of these in turn.

At one extreme, the designer may employ a policy which is completely uninformative, in the sense that the messages disclosed to the consumers reveal nothing about the platform's accumulated knowledge. For instance, the platform may disclose the same message to consumers at any time t (or indeed no message at all), irrespective of the underlying state x_t – policies of this kind are said to belong to the "no information" (NI) regime. Under NI, consumers' choices in every period are trivially dictated by the prior belief. As a result, either all consumers choose service A (when $r(x_1, A) > r_B$), or all consumers choose service B (when $r(x_1, A) \leq r_B$), and there is no adaptation of consumer actions to the service record of each provider.[12]

At the other extreme, the designer may employ a policy which is fully informative, that is, a policy which discloses a distinct message for every system state (e.g., a detailed outcome history, $g(x_t) = x_t$) – policies of this kind are said to belong to the "full information" (FI) regime. Under FI, each consumer chooses the provider which yields the highest immediate expected reward, given precise knowledge of the system state x_t.[13] Lemma 3 summarizes the period-t consumer's choice.

[12] The unique ICRP which corresponds to the NI regime is thus

$$g(x_t) = \begin{cases} A & \text{if } r(x_1, A) > r_B, \\ B & \text{if } r(x_1, A) \leq r_B, \end{cases}$$

[13] The unique ICRP which corresponds to the FI regime is

LEMMA 3. *Consumers' choices of provider under policies belonging to the FI regime are characterized as follows:*

(i) If $r(x_1, A) \leq r_B$, then consumers choose service B in all periods $t \in T$.

(ii) If $r(x_1, A) > r_B$, then the period-1 consumer chooses service A. In any period $t > 1$, there exists an integer $\bar{s}(t)$ such that if $s_t^A \geq \bar{s}(t)$ the period-t consumer chooses service A, while if $s_t^A < \bar{s}(t)$ service B is chosen in period t and forever after. The period-t threshold $\bar{s}(t)$ is uniquely defined by

$$\bar{s}(t) = \{\min s_t^A \ : \ s_t^A, f_t^A \in \mathbb{Z}_+^2, \ (s_t^A - s_1^A) + (f_t^A - f_1^A) = t-1, \ r(x_t, A) > r_B\}.$$

Consumers' choices in Lemma 3 display a similar structure with the system-optimal choices of Lemma 2, but a closer comparison reveals two potential sources of inefficiency of the FI regime. First, if the prior belief over provider A's quality is such that $r(x_1, A) \leq r_B$, then no experimentation with service A is undertaken by the consumers under FI. This behavior is system-optimal only when it is also true that $G_A(x_1) \leq G_B$; by contrast, if $r(x_1, A) < r_B$ and $G_A(x_1) > G_B$, the designer wishes for some experimentation to occur, but experimentation is never undertaken by the consumers. The second source of inefficiency arises when $r(x_1, A) > r_B$. In this case, experimentation with service A occurs in period $t = 1$ and is also system-optimal (this follows from $G_A(x_1^A) \geq r(x_1, A)$; see Gittins et al. (2011), Chapter 7). Nevertheless, the *extent* to which experimentation occurs can be suboptimal, in particular, if there is a discrepancy between any of the period-t thresholds $\bar{s}_A(t)$ and $s_A^*(t)$. The following lemma characterizes this discrepancy.

LEMMA 4. *The thresholds $s^*(t)$ and $\bar{s}(t)$ satisfy $s^*(t) \leq \bar{s}(t)$.*

Lemma 4 suggests that the FI regime suffers from *under-exploration*: the self-interested consumers tend to abandon learning about provider A's quality prematurely, before the system-optimal amount of experimentation has occurred; this is illustrated in the following example.[14]

EXAMPLE 1. Suppose that the prior belief over service provider A's quality is $Beta(1,1)$, service B has a known quality $p_B = 0.27$ and the discount factor is $\delta = 0.9$. Suppose further that the designer adopts a messaging policy belonging to the FI regime. In this case, the first consumer chooses provider A (expected payoff $0.5 > 0.27$). In the second period, we have $\bar{s}(2) = 1$; therefore, even if the period-1 consumer's experience was negative, the second consumer still uses provider A (expected payoff of $0.\dot{3} > 0.27$). In the third period, we have $\bar{s}(3) = 2$; therefore, if both the

$$g(x_t) = \begin{cases} A & \text{if } r(x_t, A) > r_B \\ B & \text{if } r(x_t, A) \leq r_B, \end{cases}$$

[14] Equality holds in Lemma 4 for all t when the designer's discount rate is sufficiently low, since in this case the designer is effectively myopic, as are the consumers.

period-1 and the period-2 consumers' experiences were negative, the period-3 consumer abandons experimentation with provider A (expected payoff $0.25 < 0.27$) and chooses provider B, as do all consumers thereafter. By contrast, system-optimal provider choices as described in Lemma 2 dictate further experimentation with service A; in particular, we have $s^*(3) = 1 < \bar{s}(3)$.

To conclude our discussion of the two extreme modes of information-provision, we present the next result which follows directly from, and summarizes, the preceding discussion.

PROPOSITION 1. *Denote by π^{NI} and π^{FI} the platform's expected payoff under policies belonging to the NI and FI regimes, respectively. Then*

$$\pi^{NI} \leq \pi^{FI} \leq \pi^*,$$

where π^* denotes the platform's first-best expected payoff.

Put simply, FI policies outperform NI policies, but both extreme modes of information-provision fail to achieve first best (i.e., the payoff achieved when the designer has full control over the consumers' actions). Equality holds on the left-hand side of the expression when experimentation with the new provider is never undertaken by the consumers under either the FI or NI regimes (i.e., when $r(x_1, A) \leq r_B$). Equality on the right-hand side holds when experimentation is never undertaken under the FI regime, and at the same time experimentation is never system-optimal (i.e., when $r(x_1, A) \leq r_B$ and $G_A(x_1^A) \leq G_B$).

5.3. Strategic Information Provision

By moving from NI to FI, the designer enables consumers to learn from the experiences of their predecessors and adapt their choices of provider accordingly. This results in an improvement in the platform's payoff, however, the designer fails to achieve first best. The shortfall occurs because consumers do not internalize the informational externality of their actions on future users of the platform: consumers always choose the provider which maximizes their individual reward, while the designer would sometimes prefer them to choose a different provider in order to generate information that is of value to future consumers.

In this section, we address the question of whether the designer can do better than FI in the decentralized system, and if so how. We demonstrate that (i) subject to a simple condition on the initial system state, an optimal messaging policy fully restores efficiency in the decentralized system, and (ii) optimal messaging policies are characterized by deliberate and controlled obfuscation of the information in the platform's possession. Interestingly, in order to restore first best, the designer is required to intervene to *restrict* consumers' ability to learn from each other.

We begin by establishing the simple condition under which the designer can achieve first best in the decentralized system.

PROPOSITION 2. *For initial system state x_1, let g^* be an optimal messaging policy and denote by $\pi(g^*)$ the platform's expected payoff under policy g^*. Then $\pi(g^*) = \pi^*$, unless both $r(x_1, A) \leq r_B$ and $G_A(x_1) > G_B$ hold.*

Roughly speaking, first best *cannot* be achieved by the designer only when the expected quality of the unknown provider A is initially close to, but lower than, the quality of provider B. In such cases, the new provider appears to be a promising prospect from the designer's perspective, but is never given the chance to "prove his worth" by the self-interested consumers, all of which inevitably select the incumbent provider B. When this occurs, the designer's choice of messaging policy is completely irrelevant, as there is no way of ever persuading consumers to try provider A; we shall return to this observation when we consider the designer's general problem in §6.

Let us now consider *how* the designer achieves first best in Proposition 2, assuming this is permitted by the initial state x_1. In general, there exist multiple messaging policies that achieve first best, but all such policies share the common feature of being deliberately *less-than-fully informative*: under an optimal policy, messages are structured so as to withhold at least some information regarding past consumer experiences. To illustrate the manner in which this is done, we first use Lemma 1 to anchor our discussion in the subclass of messaging policies referred to as ICRPs (see Definition 1); we then present an example that allows for more general messaging policies and highlights their common features.

By Lemma 1, if first best is achievable in the decentralized system, the recommendation policy

$$g(x_t) = \begin{cases} A & \text{if } G_A(x_t) > G_B \\ B & \text{if } G_A(x_t) \leq G_B, \end{cases} \tag{3}$$

must be an ICRP. Interestingly, this implies that consumers (in all periods and in all possible system states) rationally follow recommendations for the provider of highest Gittins index, even though such recommendations are not necessarily compatible with maximization of their own individual expected reward. To understand why this is the case, let us consider the mechanics underlying policy (3).

Recall that each consumer has knowledge of (i) the initial state, x_1; (ii) the period of her arrival, t; and (iii) the designer's messaging policy, in this case (3). Upon visiting the platform, she receives a message in the form of a recommendation for A or B. Taking the period-t consumer's perspective, consider first the event that a recommendation to use provider B is received. From Lemma 4, it follows that if the designer finds it system-optimal to recommend service B in any given period, then it must be the case that provider B is also optimal for the individual receiving this recommendation; to see this, note that the designer's "tolerance" for failed service outcomes with provider A is

higher (in any period) than that of the individual consumer – thus, a B recommendation is clearly incentive-compatible (IC).

Now, consider the event that a recommendation to use provider A is received. Lemma 4 suggests that this recommendation nests two possible types of states. The first type corresponds to cases where $s_t^A \geq \bar{s}(t)$: here, service A yields a higher expected reward for the individual consumer (i.e., provider A would have been chosen by the consumer even under perfect state information). By contrast, the second type corresponds to cases where $s^*(t) \leq s_t^A < \bar{s}(t)$: here, it is provider B that yields the highest expected reward for the individual consumer. By merging these two types of states into a single message – the A recommendation – the designer is able to elicit choice A from the consumer, even if the true underlying state is of the second type: upon being recommended provider A, the consumer updates her belief over the underlying state and concludes that, in expectation, she is better off by heeding the platform's advice. In the proof of Proposition 2, we demonstrate that the latter statement holds for customers in all periods; that is, the dynamics of the system are "well-behaved," in the sense that states can always be merged into messages that allow the designer to elicit system-optimal choices from the consumers.

By employing a messaging policy which is deliberately imprecise regarding the underlying system state, the designer is able to induce system-optimal behavior in the event that the realized state of the system results in misalignment between his and the individual consumer's preferences. Returning to the more general class of messaging policies and following this logic, in any optimal policy, states of the system where $r(x_t, A) \leq r_B$ and $G_A(x_t^A) > G_B$ hold simultaneously (i.e., states in which the designer and the consumers' preferences are misaligned) must correspond to the same message as some other state/states x_t' for which $r(x_t', A) > r_B$ and $G_A(x_t^{A'}) > G_B$ (i.e., states in which the designer and the consumers' preferences are aligned). As a consequence, optimal policies are characterized by a "many-to-few" structure, and some loss of accuracy in information-provision to the consumers is necessary.

The trade-off between the accuracy of information provision to consumers and the platform's payoff is an issue of practical relevance. To illustrate that this trade-off need not be a steep one, and to fix the ideas discussed in this section, we revisit Example 1 but now assume that the designer employs an optimal messaging policy. We pick up the process in period $t = 4$ and consider the decision process of the period-4 consumer under alternative messaging policies. There are four possible states in period $t = 4$, each of which occurs with probability 0.25 (see Table 1). In three of these four states, the designer and the consumers prefer the same action; that is, under perfect state information consumers would make the system-optimal choice of provider. By contrast, in the fourth state listed in Table 1 consumers would not make the system-optimal choice under perfect information.

Table 1

$x_4^A = (s_4^A, f_4^A)$	$P(x_4^A)$	consumer prefers	designer prefers
$(4,1)$	0.25	A ($r_A = 0.8$)	A ($G_A = 0.87$)
$(3,2)$	0.25	A ($r_A = 0.6$)	A ($G_A = 0.71$)
$(2,3)$	0.25	A ($r_A = 0.4$)	A ($G_A = 0.52$)
$(1,4)$	0.25	B ($r_A = 0.2$)	A ($G_A = 0.30$)

How can the designer structure his messaging policy so as to induce the period-4 consumer to choose provider A when the realized state of the world is $x_4^A = (1,4)$? Below are three distinct examples of optimal messaging policies (i.e., mappings between possible states of the system and messages disclosed to the period-4 consumer), ordered from left to right in increasing order of accuracy of information provided to the period-4 consumer.[15] The messages $m_1, m_2, m_3 \in M$ are arbitrary, since the mapping from states to messages (i.e., the designer's policy) is common knowledge.

$$\left.\begin{matrix}(4,1)\\(3,2)\\(2,3)\\(1,4)\end{matrix}\right\} m_1 \qquad \begin{matrix}(4,1)\}\, m_1\\ \left.\begin{matrix}(3,2)\\(2,3)\\(1,4)\end{matrix}\right\} m_2\end{matrix} \qquad \begin{matrix}(4,1)\}\, m_1\\(3,2)\}\, m_2\\ \left.\begin{matrix}(2,3)\\(1,4)\end{matrix}\right\} m_3\end{matrix}$$

From left to right, the designer may choose to map all, three, or only two possible period-4 states to the same message. Note, however, than in any optimal messaging policy, state $(1,4)$ cannot correspond to a unique message. To see how such imprecisions in the designer's policy restore first-best, consider, for example, the third messaging policy. If the consumer receives messages m_1 or m_2 when visiting the platform, then she has perfect state information and rationally chooses service A, as indicated in Table 1. If she receives message m_3, she conducts the following calculation

$$\begin{aligned} E[r(x_4, A) \mid g(x_4) = m_3] &= \frac{2}{2+3} \times P(x_4 = (2,3) \mid g(x_4) = m_3) + \frac{1}{1+4} \times P(x_4 = (1,4) \mid g(x_4) = m_3) \\ &= 0.4 \times 0.5 + 0.2 \times 0.5 = 0.3 > 0.27 = r_B, \end{aligned}$$

and concludes that she should choose provider A, as desired by the designer. Since the consumer receives message m_3 both when the system state is $(2,3)$ and when it is $(1,4)$, provider A is chosen in both scenarios: the possibility of the state being $(2,3)$ entices the consumer to choose A even when the state is actually $(1,4)$. Finally, we note that any optimal messaging policy consists of a "garble" of FI and is, therefore, less informative than FI in the Blackwell sense (see Marschak and Miyasawa 1968).

[15] By comparison, note that a FI policy would generate a unique message for each state of the system.

5.4. Comments

Commitment vs. Cheap Talk How important is the designer's a priori commitment to a messaging policy? In the simplified setting of this section, it is straightforward to show that the equilibrium induced by an optimal messaging policy can also be supported in a dynamic cheap-talk game. To see why, suppose that the designer does not commit to a policy a priori, and engages in a cheap-talk game with the period-t consumer. If the consumer receives a recommendation to use service B, then it must be the case that $r(x_t, A) \leq r_B$, since the only deviation-proof policy for the designer is to recommend service B only if $G_A(x_t^A) \leq G_B$, which in turn implies $r(x_t, A) \leq r_B$. If the consumer receives a recommendation to use service A, then this means that (i) all preceding consumers have used provider A, and (ii) $G_A(x_t^A) > G_B$; the consumer's rational response in this case is to follow the designer's recommendation (this follows in a similar manner as that used to explain IC of A recommendations above). While the commitment assumption can be relaxed in the analysis of this section without any loss for the platform, the same cannot be said in the designer's general problem which we consider in §6.

Information vs. Payments Apart from strategic information provision, an alternative approach to persuade consumers to engage in exploration is to offer monetary exploration subsidies. (In practice, such subsidies may be implemented through provider-specific discounts, reward points, etc.) In the case where subsidies are used, we may consider $\pi - \gamma\kappa$ as the designer's objective (see also §7.2), where π is the expected sum of the consumers' discounted rewards, κ is the expected sum of discounted subsidies paid to the consumers, and γ is a nonnegative constant. If $\gamma = 0$, then subsidies are costless for the designer, and first best can be trivially achieved by paying each consumer the amount necessary for her to choose the Gittins-optimal provider. The more interesting case is when $\gamma > 0$ (see also Frazier et al. 2014). Here, notice that the use of subsidies to induce exploration automatically rules out first best, since any $\kappa > 0$ results in a payoff $\pi - \gamma\kappa \leq \pi^* - \gamma\kappa < \pi^*$, where π^* denotes first best. It follows that in any case where strategic information provision alone is capable of restoring first best (e.g., as in Proposition 2), this approach is superior to that of using monetary subsidies. In cases where first best cannot be achieved through information provision alone, it may be optimal for the designer to employ monetary subsidies; however, as we demonstrate in §7.2, even when the use of monetary subsidies is optimal, strategic information obfuscation can only benefit the designer.

6. General Case

We now consider the case where the quality of both providers is ex ante unknown. Throughout the remaining analysis we assume, without loss of generality, that $r(x_1, A) \geq r(x_1, B)$; that is, the ex ante (weakly) preferable provider for the consumer is A. We begin by stating a condition for achieving first best in the general problem.

PROPOSITION 3. *For initial system state x_1, let g^* be an optimal messaging policy. Then $\pi(g^*) = \pi^*$ if and only if there exists an ICRP which recommends service B whenever $G_B(x_t) > G_A(x_t)$.*

Proposition 3 suggests that IC of recommendations for the ex ante *less* preferable (for the consumer) provider in all periods is a necessary and sufficient condition for achieving first best payoff. When the quality of provider B is known with certainty, this reduces to the simple condition on the initial system state x_1 described in Proposition 2. By contrast, when both providers are of ex ante unknown quality, a simple inspection of the initial state will not suffice: even if the initial system conditions are "favorable" for the designer (i.e., $r(x_1, A) \geq r(x_1, B)$ and $G_A(x_1) \geq G_B(x_1)$), first-best may not be feasible owing to the dynamics of the learning process; the following example demonstrates.

EXAMPLE 2. Suppose that the prior belief over provider A's quality is $Beta(10, 2)$, the prior belief over provider B's quality is $Beta(2, 2)$, and $\delta = 0.99$. Thus, $r(x_1, A) = 0.83 > 0.5 = r(x_1, B)$ and $G_A(x_1) = 0.92 > 0.78 = G_B(x_1)$; that is, the initial state of the system is "favorable." Furthermore, note that provider A remains the system-optimal choice in periods $t \in [1, 4]$ with probability one (i.e., irrespective of the service outcomes in periods $t \in [1, 3]$). By contrast, in period $t = 5$, there is a strictly positive probability that the system-optimal provider is B (i.e., if all trials undertaken with provider A in periods $t \in [1, 4]$ fail). However, no ICRP exists which recommends provider B with positive probability in period $t = 5$; to see this, note that the consumer's expected reward is maximized by choosing provider A in period $t = 5$, irrespective of provider A's outcome history (i.e., $r(x_5, A) > r(x_5, B)$ for all possible x_5). Therefore, as Proposition 3 suggests, first best cannot be achieved.

Thus, Proposition 3 allows us to test, in forward-induction fashion, whether first best is achievable, by checking IC of B recommendations. For most initial states x_1, this test reveals that first best cannot be achieved (see Appendix C for an exception). Nevertheless, as we demonstrate in the remainder of our analysis, a policy that discloses strategically-obfuscated information still results in a significant payoff improvement with respect to FI.

6.1. Optimal ICRPs

Infeasibility of first best implies that Gittins-based recommendations are not IC in the general problem, so that the designer can no longer rely on the Gittins index theorem to construct an optimal ICRP. Here, we provide a characterization of the designer's optimal policy in the general problem.[16]

[16] The authors thank an anonymous referee for comments on this section.

By Lemma 1, the designer in our model seeks to find the best possible ICRP, that is, to choose optimally the probabilities q_{x_t} that define the recommendations received by the period-t consumer in each possible system state

$$g(x_t) = \begin{cases} A & \text{w.p.} \quad q_{x_t} \\ B & \text{w.p. } 1 - q_{x_t}, \end{cases}$$

while at the same time ensuring that any recommendation received by the period-t consumer is IC. The designer's general problem may be framed as the following *Constrained* Markov Decision Process (CMDP; see Altman (1999)),

$$\begin{aligned} \max_{g(x_t)} \; & E\left[\sum_{t\in T} \delta^{t-1} r(x_t, g(x_t))\right] \\ \text{s.t. } & E_{x_t}[r(x_t, A) \mid g(x_t) = A] \geq E_{x_t}[r(x_t, B) \mid g(x_t) = A], \; \forall t \in T, \\ & E_{x_t}[r(x_t, B) \mid g(x_t) = B] \geq E_{x_t}[r(x_t, A) \mid g(x_t) = B], \; \forall t \in T, \end{aligned} \tag{4}$$

where the constraints state that any recommendation that is generated by policy g in period t is found to be IC (and is therefore followed) by the period-t consumer. The presence of the IC constraints introduces both direct and indirect complications. The direct complication is that recommendations generated by the designer's policy in all states that *could* occur in period t must now be viewed jointly, since such recommendations are coupled by the need to satisfy the period-t consumer's IC constraints. The indirect complication is that the designer's choice of policy up to period t affects the beliefs of customers that visit the platform in periods $t+1$ onwards, and therefore (through the IC constraints) also affects the feasible region of recommendations in future periods.

To facilitate exposition of the result that follows, we introduce the following additional notation. Let X_t be the set of states that are reachable from the initial state x_1 (under some policy) in period t, so that the total state space is $X = \bigcup_{t\in T} X_t$. Denote by $\mathcal{P}_{kiz}$ the transition probability from state k to state z when provider i is used (note that these probabilities have been specified in §4), and let Δ_a denote the Dirac delta function concentrated at a.[17]

Proposition 4. *The optimal ICRP is given by*

$$q_k^* = \frac{\rho(k, A)}{\sum_{i\in S} \rho(k, i)},$$

[17] The result of Proposition 4 extends readily to the case of $|S| = n$ providers (in this case, an ICRP consists of n possible recommendations, and each recommendation must satisfy $n-1$ IC constraints per period), as well as to alternative platform objective functions (by replacing $r(k, i)$ with suitable reward functions).

where $\rho(k,i)$ *solve*

$$\begin{aligned}
\max_{\rho} \quad & \sum_{k\in X}\sum_{i\in S}\rho(k,i)r(k,i) \\
s.t. \quad & \sum_{k\in X_t}\rho(k,B)[r(k,B)-r(k,A)]\geq 0, && \forall t\in T, \\
& \sum_{k\in X}\sum_{i\in S}\rho(k,i)\left(\Delta_z(k)-\delta\mathcal{P}_{kiz}\right)=\Delta_{x_1}(z), && \forall z\in X, \\
& \rho(k,i)\geq 0, && \forall k\in X,\ i\in S. \qquad (5)
\end{aligned}$$

A few comments on the solution technique of Proposition 4 are warranted. To solve the designer's problem, the objective and constraints of the CMDP (4) are first expressed as sums of the immediate expected reward in each state-action pair, $r(k,i)$, multiplied by the time-discounted "occupancy" of the pair, $\rho(k,i)$. Then, the LP (5) optimizes over the admissible set of occupancy measures, which is described by the LP's constraints. In particular, in the context of our problem, any admissible occupancy measure must be consistent with (i) the consumers' incentives (this is captured by the period-specific inequality constraints, which ensure that each period-t consumer finds the recommendation she receives IC), and (ii) the system's dynamics (this is captured by the state-specific equality constraints, which ensure that the occupancy of each state is consistent with the system's state-transition probabilities).[18] Finally, once the optimal occupancy measure has been identified, the probabilities q_k^* are chosen in a manner that induces this measure.

To gain insight into the structure of optimal policies, it is instructive to consider a finite-horizon version of the problem, consisting of T_F time periods. In this case, applying Theorem 3.8 of Altman (1999) reveals that the optimal ICRP uses randomized recommendations in at most T_F states. As the horizon length T_F increases, the state space grows exponentially, but the number of states in which randomization occurs grows only linearly (for instance, the number of possible states for $T_F=20$ is of the order 10^{12}, but randomization occurs in at most 20 states). This suggests that optimal policies consist mainly of deterministic recommendations, relying extensively on the use of the state-merging structure identified in §5.3 to "persuade" consumers to experiment.

6.2. The Value of Information Obfuscation

The "curse of dimensionality" renders the optimal solution to the designer's general problem computationally intractable. However, by combining the structural insights yielded by our analysis (i.e., state-merging, limited randomizations, sufficiency of two-message policies), it is possible to generate tractable and effective heuristic solutions. In this section, we consider one such heuristic

[18] Note that the solution to LP (5) can also be used to retrieve the period-t consumer's belief over the system state upon entry to the platform; specifically, this belief is given by $P(x_t=z)=\frac{\sum_{i\in S}\rho(z,i)}{\sum_{k\in X_t}\sum_{i\in S}\rho(k,i)}$.

and use it to establish that the value of information obfuscation is significant, even if this is implemented in a simple and intuitive manner (we note that the payoff under any heuristic serves as a lower bound on the payoff of the optimal policy described in Proposition 4).

Consider the following Gittins-based heuristic, which combines our preceding analysis with the centralized solution to the designer's problem to deliver IC recommendations. Let p_{x_t} denote the probability that the state in period t is x_t. The heuristic is initialized by choosing the starting state x_1 and proceeds by repeating two steps. First, it solves the period-t LP

$$\begin{aligned} \max_{0\leq q_{x_t}\leq 1} \ & \sum_{x_t\in X} p_{x_t} q_{x_t} [G_A(x_t) - G_B(x_t)] \\ \text{s.t.} \ & \sum_{x_t\in X} p_{x_t}(1-q_{x_t})[r(x_t,B) - r(x_t,A)] \geq 0 \end{aligned} \tag{6}$$

and stores the solution q_{x_t} (this is the designer's recommendation policy for period t); second, the period-t solution is used along with the probabilities p_{x_t} to calculate the probabilities $p_{x_{t+1}}$. The two steps are repeated until a pre-specified period $t = K$ is reached, after which a full-information policy is employed (or, equivalently, an ICRP which always recommends the provider of highest expected reward). Essentially, in each of the first K periods of the horizon, the heuristic employs state-merging to deliver recommendations that maximize the expected Gittins index, subject to the recommendations being IC. A more detailed discussion of the heuristic and its properties is provided in Appendix A, along with a theoretical bound on its payoff with respect to first best (see Proposition 7).

To evaluate the benefits of information obfuscation (in the sense of the Gittins-based heuristic), we conduct the numerical experiments presented in Table 2. The table focuses on the added "learning value" of obfuscation in comparison to that of a FI policy. Specifically, we first calculate the difference $(\pi^* - \pi^{NI})$, i.e., the difference between the platform's payoff when no social learning takes place (π^{NI}) and when social learning takes place optimally (π^*). This difference is an upper bound on the learning value that can be achieved by the designer in the decentralized system through information-provision. We then calculate the percentage of this value achieved under FI $(\Delta\pi^{FI})$ and under the Gittins-based heuristic $(\Delta\pi(\hat{g}))$.

The upper half of the table pertains to initial states which are "unfavorable" for the designer, in the sense that there is an ex ante misalignment between the provider of highest expected reward and the provider of highest Gittins index; by contrast, the lower part of the table pertains to "favourable" initial states. Across all instances we consider, the heuristic performs significantly better than full information. Furthermore, we observe that the benefit is highest when the initial state is unfavorable: in such cases, under full information the consumers tend to stick with the ex ante preferable provider and only rarely engage in experimentation with the alternative option.

Next, notice that in each of the four subgroups of initial states, the ex ante expected reward of the two providers is maintained constant, but the variance of one of the two changes; this allows us to capture different environments in terms of the potential benefits of exploration. Here, intuitively, we observe that the benefits of information obfuscation are especially pronounced when the quality of the ex ante preferable provider is relatively certain while the quality of the alternative provider is relatively uncertain.

$x_1 = \{(a_1^A, b_1^A), (a_1^B, b_1^B)\}$	$r(x_1, A)$	$\text{std}(x_1, A)$	$r(x_1, B)$	$\text{std}(x_1, B)$	$\Delta\pi^{FI}$	$\Delta\pi(\hat{g})$
$\{(6,3),(1,1)\}$	0.67	0.15	0.5	0.29	47.2%	96.3%
$\{(12,6),(1,1)\}$	0.67	0.11	0.5	0.29	18.6%	85.0%
$\{(18,9),(1,1)\}$	0.67	0.09	0.5	0.29	6.0%	83.7%
$\{(15,6),(2,1)\}$	0.71	0.10	0.67	0.24	58.1%	97.8%
$\{(15,6),(4,2)\}$	0.71	0.10	0.67	0.18	66.0%	90.7%
$\{(15,6),(6,3)\}$	0.71	0.10	0.67	0.15	71.7%	93.0%
$\{(1,1),(3,6)\}$	0.5	0.29	0.33	0.15	87.6%	100%
$\{(1,1),(6,12)\}$	0.5	0.29	0.33	0.11	81.0%	95.9%
$\{(1,1),(9,18)\}$	0.5	0.29	0.33	0.09	80.0%	100%
$\{(1,1),(3,6)\}$	0.5	0.29	0.33	0.15	85.4%	94.6%
$\{(3,3),(3,6)\}$	0.5	0.19	0.33	0.15	85.9%	94.6%
$\{(6,6),(3,6)\}$	0.5	0.14	0.33	0.15	51.1%	96.2%

Table 2 Proportion of first-best learning value captured in the decentralized system by FI, defined as $\Delta\pi^{FI} = \frac{\pi^{FI} - \pi^{NI}}{\pi^* - \pi^{NI}}$, and by the Gittins-based heuristic $\hat{g}$ with $K = 50$, defined as $\Delta\pi(\hat{g}) = \frac{\pi(\hat{g}) - \pi^{NI}}{\pi^* - \pi^{NI}}$ (where π^*, π^{FI}, π^{NI} and $\pi(\hat{g})$ denote expected platform payoff under first best, FI, NI and the Gittins-based heuristic, respectively). $r(x_1, i)$ and $std(x_1, i)$ denote, respectively, the expectation and standard deviation of the reward of provider $i \in \{A, B\}$ at the initial state x_1. Parameter values: $\delta = 0.99$. Results were obtained using simulation and the Bayesian approach described in Caro and Gallien (2007); raw data provided in Appendix D.

7. Extensions

7.1. Imperfect Knowledge of Consumers' Arrival Times

In our main analysis, we have assumed that consumers know the exact period of their arrival, which implies that they know how many of their peers preceded them in seeking service. While this assumption may appear to be restrictive, here we show that our approach is in fact a robust one. In particular, the result that follows allows for consumers to hold arbitrary beliefs over their arrival times.[19]

PROPOSITION 5. *Let g (g^*) denote an ICRP (the optimal ICRP) when consumers have perfect knowledge of their arrival times. Then:*

[19] In order to avoid pathological cases, we impose the restriction that each consumer's actual arrival time is consistent with her belief; that is, it cannot be that the consumer arrives in a period where her belief assigns zero probability.

(i) g remains an ICRP under any arbitrary belief held by consumers over their arrival times.

(ii) If v^* is an optimal ICRP under a specific set of consumer beliefs, we have $\pi(v^*) \geq \pi(g^*)$.

In the proof of Proposition 5, we demonstrate that if a recommendation policy is IC for the case where consumers have precise knowledge of their arrival times, then the same recommendation policy is IC when consumers hold arbitrary (and possibly heterogeneous) beliefs. This result is particularly appealing, because it suggests that the designer can achieve an expected payoff equal to $\pi(g^*)$ irrespective of what consumers' beliefs may be and irrespective of whether these beliefs are observable to the designer. When consumers do not have precise knowledge of their arrival time, this allows the designer to merge states not only within each period, but also across periods – effectively, the constraints faced by the designer in delivering IC recommendations become less stringent. The extent to which this may help the designer achieve higher payoff will depend on the specific nature of consumers' beliefs. In any case, a robust and effective approach for the designer is to design his policy assuming that consumers are perfectly informed about their arrival times.

7.2. Combining Information with Monetary Subsidies

In this section, we extend our analysis to allow for cases where the platform designer, in order to elicit exploration from the self-interested consumers, can complement his information-provision policy with monetary subsidies. Consistent with extant literature (e.g., Frazier et al. 2014), we assume that the designer's objective is to maximize $\pi - \gamma\kappa$, where π is the expected sum of the consumers' discounted rewards, κ is the expected sum of discounted exploration subsidies paid to the consumers, and γ is a nonnegative constant. The extended model can be viewed as a lagrangian relaxation of a problem where the designer has a fixed budget allocated to exploration subsidies. Moreover, the pure-information model of §6 can be retrieved by setting γ sufficiently large (i.e., so that payments are prohibitively costly), while when $\gamma = 0$ (i.e., so that payments are costless) the designer can implement first best simply by paying each consumer the amount necessary for her to choose the Gittins-optimal provider (i.e., the difference between the expected reward of her preferred choice and that of the Gittins-optimal choice).

Under a full-information policy, the designer must evaluate in each system state whether the long-run benefit of incentivizing the consumer to conduct exploration (through a monetary payment) is higher than the instantaneous cost of doing so. The full-information problem has been studied by Frazier et al. (2014), who provide a characterization of the gains that are achievable at a fixed budget. Here, we consider the combination of information provision with subsidies, with a focus on establishing whether information obfuscation remains beneficial when the provider can directly incentivize exploration through subsidies.

To answer this question, we consider "messaging-with-subsidies" policies where, apart from specifying a mapping from states to messages, the designer can choose for each message $m \in M$ used in period t an accompanying subsidy plan $\{\kappa_t^i(m)\}_{i\in S}$, where the subsidy $\kappa_t^i(m) \geq 0$ can be claimed by the consumer if she chooses to use provider i. In this case, the designer faces the same trade-off as that in the full-information case described above, but with an additional layer of complexity: the payments necessary to induce exploration in any system state now also depend on the designer's choice of messaging policy, because the compensation required by each consumer depends on what she infers (regarding the providers' qualities) from the message she receives. Thus, this problem appears to be significantly more complex due to the multitude of possible combinations of payment plans and state-to-message mappings. To facilitate exposition of the result that follows, we first provide the following definition, which expands the notion of an ICRP to the case with subsidies.

Definition 2 (ICRSP: Incentive-Compatible Recommendation-with-Subsidies Policy). A recommendation-with-subsidies policy is a messaging-with-subsidies policy defined as

$$v(x_t) = \begin{cases} A, \{\kappa_t^i(A)\}_{i\in S} & \text{w.p.} \quad q_{x_t} \\ B, \{\kappa_t^i(B)\}_{i\in S} & \text{w.p. } 1 - q_{x_t}, \end{cases} \tag{7}$$

where $q_{x_t} \in [0,1]$ for all $x_t \in X$, and $\kappa_t^i(j) = 0$ for all $i, j \in S$, $i \neq j$. A recommendation-with-subsidies policy is said to be incentive-compatible if for all $x_t \in X$, each period-t consumer follows the recommendation she receives.

Under an ICRSP, the designer recommends a single provider, but now this recommendation may also be accompanied by a positive subsidy (conversely, the subsidies corresponding to providers other than the one recommended are set to zero). The use of an ICRSP constitutes information obfuscation, as each possible period-t state maps to one of two recommendations. We may now present the following result, which establishes the dominance of ICRSPs in the model with subsidies.

Proposition 6. *For any arbitrary messaging-with-subsidies policy v, there exists an ICRSP v' which achieves a (weakly) higher expected platform payoff.*

To establish Proposition 6, we show that for any arbitrary messaging-with-subsidies policy there exists an ICRSP which replicates the same consumer actions in every system state, but at a potentially *lower* total subsidy cost. More specifically, as we demonstrate in the proof of the proposition, if there exists in the general policy v some message m which (in any period) induces selection of provider i without requiring a monetary subsidy attached to that provider, then there exists an ICRSP v' which achieves the same actions but at a strictly lower total subsidy cost (thus resulting in a strictly higher platform payoff). We present an example of this effect below.

EXAMPLE 3. Suppose that the prior belief over provider A's quality is $Beta(5,5)$ (i.e., ex ante expected quality of 0.5) and that provider B has a known quality of $r_B = 0.53$; thus, the ex ante preferable provider for the consumers is B. Consider the minimum total subsidies required to induce exploration of provider A in the first two periods under a FI policy versus under an ICRSP. In the first period, for the customer to choose provider A, both policies must offer a minimum subsidy of $\kappa_1^A = p_B - r(x_1, A) = 0.53 - 0.5 = 0.03$ (i.e., assuming $\kappa_1^B = 0$). However, in the second-period:

(i) Under FI, the state x_2 is disclosed to the period-2 consumer and a minimum subsidy $\kappa_2^A(x_2) = \max\{p_B - r(x_2, A), 0\}$ must be offered in order for the consumer to choose provider A. If the period-1 trial was successful (occurs w.p. 0.5), then $\kappa_2^A = \max\{0.53 - 0.55, 0\} = 0$. If the period-1 trial was unsuccessful (occurs w.p. 0.5), then $\kappa_2^A = \max\{0.53 - 0.45, 0\} = 0.08$. Thus, the ex ante expected period-2 subsidy is $E[\kappa_2^A(x_2)] = 0.5 \times 0.08 = 0.04$.

(ii) Under an ICRSP, the designer in period 2 recommends provider A irrespective of the period-1 outcome. For this recommendation to be incentive-compatible, it must be accompanied by a subsidy of $\kappa_2^A(A) = \max\{p_B - E[r(x_2, A)], 0\} = \max\{0.53 - (0.5 \times 0.55 + 0.5 \times 0.45), 0\} = 0.03$.

Thus, in the above example, the ICRSP achieves the same consumer actions in the first and second periods as a FI-with-subsidies policy, but at a 25% lower subsidy cost.

8. Conclusion

This paper investigates how information provision can be used to regulate the process by which information is generated in decentralized learning contexts. We conduct our analysis within a decentralized multi-armed bandit framework that exhibits the well-known exploration-exploitation trade-off. We demonstrate how, by disclosing information that is strategically obfuscated, a principal interested in maximizing social surplus can succeed in "persuading" self-interested agents to take socially-optimal actions. We have further demonstrated that the value of information obfuscation in decentralized learning can be significant, and that this value persists even when agents' actions can be directly incentivized through monetary payments.

Similar misalignments in the objectives of the agents and the principal are inherent in many settings (e.g., see §1), however, it is important to recognize that our model makes several simplifications on dimensions which may influence information provision in specific contexts. Such dimensions include, among others, more complex principal objectives, agent heterogeneity in preferences and/or reporting propensity, behavioral biases in decision making, and external factors that promote specific agent actions. While the aforementioned simplifications present potential avenues for future work, we discuss below two further issues that are particularly intriguing.

The first is associated with variation of the quality of alternative options over time. For instance, in the review platform setting, the quality of service providers is likely to change over time. Future

work may focus on two relevant questions. First, if changes in quality are assumed to be exogenous to the learning process, then how should the platform disclose information to its users? Here, one may expect an optimal policy to include an element of "forgetting" relatively old (and therefore possibly outdated) information.[20] Second, if qualities are endogenous to the learning process (e.g., providers react to the content reported to the platform), then how does the principal's information-provision policy interact with the providers' choice of quality? In this case, the platform must consider not only its role in providing information to consumers, but also its role in affecting the providers' service quality.

The second interesting issue is that of competition. In the current paper we have assumed a "monopolistic" platform. In a setting where multiple platforms are competing for user traffic, how would the platforms structure their information-provision policies? Would platforms choose to differentiate by employing policies of different informativeness? In the short-run, if a platform elects to employ a full-information policy as opposed to a competitor's strategic-information policy, then we may expect it to attract a larger portion of the consumer population. However, the current paper suggests that the full-information platform will generate qualitatively inferior content, and may therefore suffer in the long run.

Appendix

A. The Gittins-Based Heuristic

In this section, we provide further details on the Gittins-based heuristic (6) described in §6.2.

Heuristic Design Note first that a period-by-period construction of a policy that constitutes an ICRP is permitted by the structure of the constraints in problem (4). In particular, to ensure that a policy is an ICRP, the constraints that the designer's period-t recommendations must satisfy are fully specified by the belief of the period-t customer; at the same time, the belief of the period-$t+1$ consumer follows readily from the period-t belief and the period-t policy. In the heuristic, every period-t LP respects the IC constraints of the period-t consumer (this is ensured by the single linear constraint in (6), which can be shown to be equivalent to the two period-t constraints that appear in (4); e.g., see proof of Proposition 4), so that the policy constructed is guaranteed to be an ICRP (i.e., feasible).

The heuristic operates on the basis of the state-merging property identified in §5 to maximize in each period the *expected Gittins index* of the action taken by the period-t consumer. To see how this is achieved, define for period t the sets $IC_i^t = \{x_t : G_i(x_t) \geq G_{i'}(x_t),\ r(x_t,i) \geq r(x_t,i')\}$ and $NC_i^t = \{x_t : G_i(x_t) > G_{i'}(x_t),\ r(x_t,i) < r(x_t,i')\}$, where $i \neq i'$ and $i,i' \in S$. The sets IC_i^t (NC_i^t) contain those states of the system in which the provider of highest Gittins index would (would not) be preferred by the period-t consumer under full information. The solution to each period-t LP merges states belonging to IC_i^t with states belonging to NC_i^t, with the goal of eliciting Gittins-optimal actions in states where the consumers under full information would have chosen a different action.

[20] See Besbes et al. (2014) for related work in a setting with centralized decision making.

Performance The performance of the heuristic can be evaluated by exploiting the observation that the heuristic is equivalent to a suboptimal centralized policy in the MAB problem. Specifically, let U^t be the set of states at time t in which the heuristic policy is forced, with at least some probability, *not* to recommend the provider of highest Gittins index. We may then state the following result which utilizes Glazebrook (1982).

PROPOSITION 7. *For initial system state x_1, let $\hat{g}$ denote the Gittins-based heuristic policy and let p_{x_t} denote state probabilities under policy $\hat{g}$. The following statements hold:*

1. *The difference between π^* and $\pi(\hat{g})$ is bounded by*

$$\pi^* - \pi(\hat{g}) \leq \sum_{t=1}^{+\infty} \sum_{x_t \in U^t} \delta^{t-1} p_{x_t} \mid G_A(x_t) - G_B(x_t) \mid .$$

2. *Let g^* be the optimal ICRP. If $\pi(g^*) = \pi^*$, then $\pi(\hat{g}) = \pi^*$.*

The bound accumulates a penalty (equal to the Gittins-suboptimality of the recommended provider) whenever the heuristic policy fails to recommend the provider of highest Gittins index. Since the heuristic can only perform worse than the optimal policy described in Proposition 4, this bound also serves as a lower bound on the payoff of the optimal ICRP described in Proposition 4 (we note that a limitation of this bound is that it requires numerical calculations, e.g., simulation). The second point of the proposition shows that if Gittins-based recommendations are IC everywhere, these recommendations are also chosen by the Gittins-based heuristic.

Computation The inputs to the routine used to extract the Gittins-based ICRP in our computations are (i) the initial system state x_1, (ii) the designer's discount factor δ, and (iii) a table of Gittins indices at the designer's discount factor. Computation of Gittins index tables is relatively straightforward (e.g., see Gittins et al. (2011), pp.223-224), and need only be conducted once for each value of δ. For each period t, we solve LP (6), store the solution, and then use the solution along with the current states x_t and their probabilities p_{x_t} to construct the set of possible states in period $t+1$ and calculate their probabilities $p_{x_{t+1}}$. We observe that using strategic IC recommendations beyond period 50 is only marginally beneficial in terms of system payoff but computationally cumbersome. Thus, we set the initial number of period where the heuristic actively obfuscates information to $K = 50$. After extracting the heuristic policy, we perform simulation analysis to evaluate its performance (see §6.2).

B. Proofs

Supporting Results

The following lemma is used in subsequent proofs. For proof of this lemma, see, for example, Bellman (1956).

LEMMA 5. *Let $g(a,b)$ denote the Gittins index of a Bernoulli reward process with current success probability distributed as $Beta(a,b)$, $a,b \in \mathbb{Z}^+$. The following properties hold: (i) $g(a,b) < g(a+1,b)$; (ii) $g(a,b) > g(a,b+1)$; (iii) $g(a,b) < g(a+1,b-1)$.*

Proof of Lemma 1 Given the designer's policy and the choice-strategy of the preceding consumers, the period-t consumer holds rational beliefs over the possible states of the system in period t. Upon receiving message m, the consumer's expected reward from choosing service i is given by

$$E[r(x_t,i)\mid g(x_t)=m]=\sum_{j\in X_t} r(j,i)\frac{P(g(x_t)=m,x_t=j)}{P(g(x_t)=m)}=\sum_{j\in X_t} r(j,i)\frac{P(g(x_t)=m\mid x_t=j)P(x_t=j)}{\sum_{k\in X_t}P(g(x_t)=m\mid x_t=k)P(x_t=k)}$$
$$=\sum_{j\in X_t} r(j,i)\frac{P(g(j)=m)P(x_t=j)}{\sum_{k\in X_t}P(g(k)=m)P(x_t=k)}.$$

Conditional on receiving message m, it is optimal for the consumer to use service A or service B, or the consumer is indifferent between the two providers. In the latter case, we assume that the consumer chooses the designer's preferred option. We will show, by construction, that for any arbitrary messaging policy there exists an ICRP which induces equivalent system dynamics. For some messaging policy g, define the sets $M_t^A=\{m: m\in M,\ \text{period-t consumer chooses } A\}$ and $M_t^B=\{m: m\in M, \text{period-t consumer chooses } B\}$. Now consider the recommendation policy g', defined by

$$g'(x_t)=\begin{cases} A & \text{w.p. } \sum_{m\in M_t^A}P(g(x_t)=m)\\ B & \text{w.p. } \sum_{m\in M_t^B}P(g(x_t)=m).\end{cases} \tag{8}$$

The recommendation policy g' is, by design, incentive-compatible for the period-t consumer, since we have simply replaced messages with recommendations of the service-choices that they induce. Since the above recommendation policy results in (stochastically) identical consumer choices in any period t and in any state of the system x_t, the statement of the lemma follows.

Proof of Lemma 2 Note first that if $G_A(x_t)\le G_B$ for some $t=k$, then provider B is system-optimal in period $t=k$. Furthermore, if B is used in period $t=k$ then $x_{k+1}^A=x_k^A$ so that B remains system-optimal in all periods $t>k$. The first part of the lemma follows readily. For the second part, note that A is system-optimal in period $t=1$. Furthermore, provider A remains system-optimal until the first period in which $G_A(x_t)\le G_B$ holds, at which point it is system-optimal to switch to B and use B forever after. We have $x_t=\{s_t^A,f_t^A\}$, where $s_t^A+f_t^A=s_1^A+f_1^A+t-1$; that is, $x_t=\{s_t^A,s_1^A+f_1^A+t-1-s_t^A\}$. From property (iii) of Lemma 5, we know that $G_A(x_t)$ is increasing in s_t^A; the threshold $s^*(t)$ follows from this monotonicity.

Proof of Lemma 3 Under the FI regime, consumers have perfect state information. If $r_A(x_t)\le r_B$ for some $t=k$, then provider B is chosen in period $t=k$. If B is chosen in period $t=k$ then $x_{k+1}^A=x_k^A$ so that B is chosen in all periods $t>k$. The first part of the lemma follows readily. For the second part, note that A is chosen by the consumer in period $t=1$. Furthermore, provider A is chosen by the consumers until the first period in which $r_A(x_t)\le r_B$ holds, at which point consumers switch to B and use B forever after. We have $x_t=\{s_t^A,f_t^A\}$, where $s_t^A+f_t^A=s_1^A+f_1^A+t-1$; that is, $x_t=\{s_t^A,s_1^A+f_1^A+t-1-s_t^A\}$. Next, note that $r(x_t,A)=\frac{s_t^A}{s_t^A+f_t^A}$, is increasing in s_t^A; the threshold $\bar{s}(t)$ follows from this monotonicity.

Proof of Lemma 4 By contradiction. Suppose that for some t we have $s^*(t)>\bar{s}(t)$; then, there exists some x_t with $s_t^A\ge\bar{s}(t)$ and $s_t^A<s^*(t)$. From Lemma 3, we have that consumers in state x_t prefer to use service A, which in particular implies $r_A(x_t,A)>r_B$. From Lemma 2, we have that the designer in state x_t prefers to use provider B, which in particular implies that $G_A(x_t)<G_B$. Lemmas 2 and 3 together imply $r_A(x_t,A)>r_B=G_B>G_A(x_t,A)$. However, note that from Gittins et al. (2011), pp.176-177, we know that $r_A(x_t,A)\le G_A(x_t,A)$, a contradiction. We conclude that $s^*(t)\le\bar{s}(t)$ for all $t\in T$.

Proof of Proposition 1 We establish each side of the inequality in turn. Consider first $\pi^{NI} \leq \pi^{FI}$. If $r(x_1, A) \leq r_B$, then under either policy regime consumers choose service B at all $t \in T$; therefore, in this case we have $\pi^{NI} = \pi^{FI}$. If $r(x_1, A) > r_B$ then the first consumer chooses service A under both regimes. Furthermore, under NI, consumers choose A in all $t \in T$, because all choices are made based on x_1. Under FI, consumer choices are characterized by the stopping time $\hat{\tau} = \inf\{t : r(x_t, A) \leq r_B\}$, at which time consumers switch to service B and use this service forever after (note that τ takes a finite value with positive probability provided the prior distribution $Beta(a_1^A, b_1^A)$ has positive density across its support). Thus, policies NI and FI are outcome-and-dynamics equivalent up to the stopping time $\hat{\tau}$, and we may focus on differences thereafter. Consider any realization of the stopping time $\hat{\tau}$. In period $t = \hat{\tau}$, the expected value-to-go under NI is $\frac{r(x_{\hat{\tau}}, A)}{1-\delta}$, while the expected value-to-go under FI is $\frac{r_B}{1-\delta} \geq \frac{r(x_{\hat{\tau}}, A)}{1-\delta}$. We conclude that $\pi^{NI} \leq \pi^{FI}$.

Next, note that $\pi^{FI} \leq \pi^*$ follows simply from the fact that FI is a feasible policy and π^* is first-best. We describe the conditions that specify whether FI achieves first-best or not. If $r(x_1, A) \leq r_B$, two possible cases arise: (i) $G_A(x_1) \leq G_B$, in which case consumers choose service B at all $t \in T$, and this is also system-optimal, so that $\pi^{FI} = \pi^*$; (ii) $G_A(x_1) > G_B$, in which case consumers choose service B at all $t \in T$, but it is system-optimal to use service A at least once in period $t = 1$, so that $\pi^{FI} < \pi^*$. Next, if $r(x_1, A) > r_B$ then this implies $G_A(x_1) > G_B$. Under FI, the consumer at $t = 1$ chooses service A, and this is also the system-optimal choice. Furthermore, consumer choices under FI are characterized by $\hat{\tau}$ as described above, while system-optimal choices are characterized by the stopping time $\tau^* = \inf\{t : G_A(x_t) \leq G_B\}$. Note that $G_B = r_B$ and that $G_A(x_t)$ is increasing in δ (Gittins et al. 2011, pp.32) with $\lim_{\delta \to 0} G_A(x_t) = r(x_t, A)$. Therefore, the stopping rule $G_A(x_t) \leq G_B = r_B$ collapses to the stopping rule $r(x_t, A) \leq r_B$ for sufficiently small δ. When this is the case, we have $\pi^{FI} = \pi^*$, while when there is discrepancy between τ^* and $\hat{\tau}$ we have $\pi^{FI} < \pi^*$.

Proof of Proposition 2 The proof of the proposition relies on the following lemma.

LEMMA 6. *Gittins-recommendations are IC in all periods $t \in T$ if and only if a Gittins-recommendation is IC in period $t = 1$.*

Proof. The recommendation policy considered is

$$g(x_t) = \begin{cases} A & \text{if } G_A(x_t) > G_B \\ B & \text{if } G_A(x_t) \leq G_B, \end{cases} \tag{9}$$

If the above policy is IC in period $t = 1$, then this implies that either (i) $G_A(x_1) > G_B$ (designer prefers A in period 1) and $r(x_1, A) > r_B$ (consumer also prefers A in period 1), or (ii) $G_A(x_1) \leq G_B$ (designer prefers B) and $r(x_1, A) \leq r_B$ (consumer also prefers B). Under case (ii), IC of policy (9) in all periods follows trivially from the fact that each period is a repetition of the first (i.e., $x_t = x_1$ for all t).

Next, under case (i), note that policy (9) is IC in period t if both of the following hold simultaneously

$$E[r(x_t, A) - r_B \mid g(x_t) = A] \geq 0, \tag{10}$$

$$E[r(x_t, A) - r_B \mid g(x_t) = B] \leq 0. \tag{11}$$

The two conditions postulate that the period-t consumer is better off (in expectation) by following the recommendation she receives, be it A (10) or B (11). Now, notice that condition (11) is guaranteed to hold

by policy (9) since $E[r(x_t,A)-r_B \mid g(x_t)=B] = E[r(x_t,A)-r_B \mid G_A(x_t)\le G_B] \le 0$, where we have first used the structure of policy (9) and then the Gittins index property $r(x_t,A)\le G_A(x_t)$ and the fact that $G_B = r_B$. We next claim that under case (i), if condition (11) holds, then condition (10) must also hold. To see this, first note that upon entering the system and before receiving a message from the platform, the period-t consumer's expected reward from using service A is simply $E[r(x_t,A)] = r(x_1,A)$ where x_t are the possible system states in period t. Furthermore, under policy (9) (as is true under *any* recommendation policy),

$$\begin{aligned} r(x_1,A)-r_B &= E[r(x_t,A)-r_B] \\ &= E[r(x_t,A)-r_B \mid g(x_t)=A]P(g(x_t)=A) + E[r(x_t,A)-r_B \mid g(x_t)=B]P(g(x_t)=B). \end{aligned} \tag{12}$$

In the above expression, under case (i) we have $r(x_1,A)-r_B>0$ so that the left-hand side is positive. If expression (11) holds, then the second term of the right-hand-side expression is non-positive. Therefore, the first term of the right-hand side must be positive, which in particular implies that (10) is satisfied. We conclude that if (9) is IC in period $t=1$, then it is IC in all subsequent periods. □

The first two cases listed in the proposition follow from Lemma 6, since in these cases the Gittins recommendation policy (3) is IC in period $t=1$. Next, since $r(x_1,A)\le G_A(x_1)$ for any x_1, the only remaining case not covered in the proof of Lemma 6 is case (iii) of Proposition 2, namely, $r(x_1,A)\le r_B$ and $G_A(x_1)>G_B$. To prove that in this case $\pi(g^*)<\pi^*$, it suffices to point out that the first-best policy would use provider A in period $t=1$, but that provider A is never chosen in period $t=1$ by the consumers in the decentralized system, under any messaging policy.

Proof of Proposition 3 Consider an arbitrary recommendation policy

$$g(x_t) = \begin{cases} A & \text{w.p.} \quad q_{x_t} \\ B & \text{w.p. } 1-q_{x_t}, \end{cases} \tag{13}$$

where $q_{x_t}\in[0,1]$ for all $x_t\in X$. Policy (13) is IC in period t if both of the following hold simultaneously

$$E[r(x_t,A)-r(x_t,B) \mid g(x_t)=A] \ge 0, \tag{14}$$

$$E[r(x_t,A)-r(x_t,B) \mid g(x_t)=B] \le 0. \tag{15}$$

Under policy g and for initial state x_1, let X_t be the set of possible states in period t. Next, note that in any period t, before the consumer receives a recommendation, we have $E[r(x_t,i)] = r(x_1,i)$. Furthermore, recall that $r(x_1,A)-r(x_1,B)\ge 0$ by assumption. In any state $x_t\in X_t$, the consumer receives either an A or a B recommendation, the probability of which is specified by q_{x_t} and $1-q_{x_t}$ respectively. We have

$$\begin{aligned} r(x_1,A)-r(x_1,B) = E[r(x_t,A)-r(x_t,B)] =& E[r(x_t,A)-r(x_t,B) \mid g(x_t)=A]P(g(x_t)=A) \\ &+ E[r(x_t,A)-r(x_t,B) \mid g(x_t)=B]P(g(x_t)=B). \end{aligned} \tag{16}$$

If a B recommendation is IC under policy (13), then (15) holds and the second term of (16) is non-positive. It follows that the first term of (16) is non-negative (because $r(x_1,A)-r(x_1,B)\ge 0$), so that (14) holds; thus IC of a B recommendation in period t ensures IC of an A recommendation. To complete the proof, note that an ICRP achieves first-best if and only if it recommends (deterministically) in any state x_t the provider of highest Gittins index. From the above discussion it follows that this necessary and sufficient condition is equivalent to the existence of an ICRP which recommends provider B in period t in any state in which provider B has the highest Gittins index.

Proof of Proposition 4 We frame the problem as a constrained MDP and employ a Linear Programming solution approach (see Altman (1999), Chapter II). We wish to choose policy g to maximize

$$R(x_1, g) := E\left[\sum_{t\in T} \delta^{t-1} r(x_t, g(x_t))\right],$$

Define $f(x_1, g; k, y) := \sum_{t\in T} \delta^{t-1} P(x_t = k, g(x_t) = y) = \delta^{t-1} P(x_t = k, g(x_t) = y)$ (the last equality follows because state k can only be visited once), so that the objective can be expressed as

$$R(x_1, g) = \sum_{k\in X} \sum_{y\in S} f(x_1, g; k, y) r(k, y).$$

The above objective must be maximized subject to the consumers' IC constraints. Specifically, for each $t \in T$, the chosen policy must satisfy the conditions

$$E\left[r(x_t, A) - r(x_t, B) \mid g(x_t) = A\right] \geq 0,$$
$$E\left[r(x_t, B) - r(x_t, A) \mid g(x_t) = B\right] \geq 0.$$

Regarding these period-t constraints, note first that using the proof of Proposition 3 (see (16) and discussion that follows it) the first constraint is redundant, so that we need only consider the second constraint (i.e., the constraint corresponding to recommendations for the ex ante less favorable option from the consumers' perspective). Next, note that the second constraint can be written as

$$E[r(x_t, B) - r(x_t, A) \mid g(x_t) = B] = \sum_{k\in X_t} \left[r(k, B) - r(k, A)\right] \frac{P(x_t = k, g(x_t) = B)}{P(g(x_t) = B)} \geq 0,$$

which is equivalent to the constraint

$$\sum_{k\in X_t} f(x_1, g; k, B) \left[r(k, B) - r(k, A)\right] \geq 0.$$

Thus we have expressed both the objective and the constraints of the problem in terms of the "occupation measure" $f(x_1, g)$. To find the optimal recommendation policy, we first optimize over the set of occupation measures by solving the following linear program

$$\begin{aligned}
\max_{\rho} \quad & \sum_{k\in X} \sum_{i\in S} \rho(k, i) r(k, i) \\
\text{s.t.} \quad & \sum_{k\in X_t} \rho(k, B)[r(k, B) - r(k, A)] \geq 0, && \forall t \in T, \\
& \sum_{k\in X} \sum_{i\in S} \rho(k, i) \left(\Delta_z(k) - \delta \mathcal{P}_{kiz}\right) = \Delta_{x_1}(z), && \forall z \in X, \\
& \rho(k, i) \geq 0, && \forall k \in X,\ i \in S.
\end{aligned}$$

Note that the last two sets of constraints ensure that $\rho(\cdot, \cdot)$ is a probability measure over state-action pairs. We then use the solution to the above LP to derive the optimal recommendation policy (i.e., the policy which induces the optimal occupation measure); this is given by $q_k^* = \frac{\rho(k,A)}{\sum_{i\in S} \rho(k,i)}$.

Proof of Proposition 5 Suppose that the designer employs a recommendation policy g which is an ICRP when consumers have precise knowledge of the period of their arrival. We will first show that g remains an ICRP under *any* arbitrary belief held by each individual consumer regarding his arrival time (that is, we do not exclude the possibility of consumers holding heterogeneous beliefs regarding their arrival time). IC of g when consumers have precise knowledge of their arrival time implies that

$$E[r(x_t,m) \mid g(x_t)=m,t] \geq E[r(x_t,m') \mid g(x_t)=m,t] \tag{17}$$

for all $m,m' \in \{A,B\}$, $x_t \in X$ and $t \in T$. Now consider the perspective of some customer j who enters the system when the (unobservable) system state is x_t, receives a recommendation $g(x_t)=m$ and holds some arbitrary belief regarding the time period of his arrival; let this belief be described by $P(t=v)=:p_v \geq 0$ with $\sum_{v\in T} p_v = 1$. To see that consumer j finds the recommendation $g(x_t)=m$ IC for $m \in \{A,B\}$, note that

$$\begin{aligned} E[r(x_t,m) \mid g(x_t)=m] = \sum_{v\in T} p_v E[r(x_t,m) \mid g(x_t)=m, t=v] &\geq \sum_{v\in T} p_v E[r(x_t,m') \mid g(x_t)=m, t=v] \\ &= E[r(x_t,m') \mid g(x_t)=m]. \end{aligned}$$

Thus, any g which is an ICRP when consumers have precise knowledge of their arrival time remains an ICRP when consumers have arbitrary (and possibly heterogeneous) beliefs. (Note here that the designer's recommendation may result in the consumer updating his belief regarding his arrival time, in which case the above argument continues to apply under the consumer's updated arrival-time belief.) Among all possible precise-knowledge ICRPs, g^* maximizes expected platform payoff. Under any arbitrary consumer beliefs, the designer can always implement the ICRP g^* and achieve $\pi(g^*)$, while he may be able to do better by implementing a policy v^* which depends on the specific beliefs held by the consumers; hence, $\pi(v^*) \geq \pi(g^*)$.

Proof of Proposition 6 Consider an arbitrary messaging-with-subsidies policy v where each message $m \in M$ in period t is accompanied by a subsidy plan $\{\kappa_t^i(m)\}_{i\in S}$, with $\kappa_t^i(m) \geq 0$, $S=\{A,B\}$. Under policy v, define the sets $Z_t^i = \{m : m \in M,$ period-t consumer chooses provider $i\}$ for $i \in S$. In particular, this implies

$$\begin{aligned} E[r(x_t,i) \mid g(x_t)=m] + \kappa_t^i(m) &\geq E[r(x_t,j) \mid g(x_t)=m] + \kappa_t^j(m) \\ \kappa_t^i(m) - \kappa_t^j(m) &\geq E[r(x_t,j) - r(x_t,i) \mid g(x_t)=m] \end{aligned}$$

for all $x_t \in X$, $m \in Z_t^i$, $j \in S$. Notice that under policy v, the designer only incurs the subsidy cost $\kappa_t^i(m)$ when message $m \in Z_t^i$ is disclosed to the consumer (because i is the consumer's chosen provider). Now, consider an alternative policy $\hat{v}$, which uses the same state-to-message mapping as v, but accompanies each message $m \in Z_t^i$ with the subsidy plan $\{\kappa_t^i(m) = \max\{E[r(x_t,j) - r(x_t,i) \mid g(x_t)=m], 0\}, \kappa_t^j(m)=0\}$. Notice that under $\hat{v}$, messages $m \in Z_t^i$ still induce action i but at the lowest possible subsidy expenditure. Therefore, policy $\hat{v}$ uses the same messages and induces the same actions as v, but at a weakly lower subsidy cost (resulting in a weakly higher platform payoff). Note that the total expected period-t subsidy incurred under policy $\hat{v}$ from messages $m \in Z_t^i$ is

$$\bar{\kappa}_t^i = \sum_{m\in Z_t^i} P(g(x_t)=m) \max\{E[r(x_t,j) - r(x_t,i) \mid g(x_t)=m], 0\}$$

$$= \sum_{m \in Z_t^i} \max\{P(g(x_t) = m) E[r(x_t, j) - r(x_t, i) \mid g(x_t) = m], 0\}.$$

We will now construct an ICRSP which induces the same actions as $\hat{v}$ in period t, but at a weakly lower total expected period-t subsidy. Consider an ICRSP v', which takes from policy $\hat{v}$ each message $m \in Z_t^i$ along with its corresponding subsidy plan $\{\kappa_t^i(m), 0\}$ (as described above), and replaces it with the recommendation i and a subsidy plan $\{\kappa_t^i(i), 0\}$. For the recommendation i to be incentive compatible, it must be accompanied by a minimum subsidy of $\kappa_t^i(i) = \max\{E[r(x_t, j) - r(x_t, i) \mid g(x_t) = i], 0\}$. This means that under v', the total expected period-t subsidy from i recommendations is

$$\begin{aligned}
\bar{\kappa}_t^{i'} &= P(g(x_t) = i) \max\{E[r(x_t, j) - r(x_t, i) \mid g(x_t) = i], 0\} \\
&= \left[\sum_{m \in Z_t^i} P(g(x_t) = m) \right] \max \left\{ \frac{\sum_{m \in Z_t^i} P(g(x_t) = m) E[r(x_t, j) - r(x_t, i) \mid g(x_t) = m]}{\sum_{\nu \in Z_t^i} P(g(x_t) = \nu)}, 0 \right\} \\
&= \max \left\{ \sum_{m \in Z_t^i} P(g(x_t) = m) E[r(x_t, j) - r(x_t, i) \mid g(x_t) = m], 0 \right\}
\end{aligned}$$

To complete the proof, notice that $\bar{\kappa}_t^i \geq \bar{\kappa}_t^{i'}$, which implies that policy v' replicates the actions of policy $\hat{v}$, but at a weakly lower total subsidy cost (resulting in a weakly higher platform payoff). Furthermore, it follows by inspection that the inequality is strict provided the quantity $E[r(x_t, j) - r(x_t, i) \mid g(x_t) = m]$ is negative for at least some $m \in Z_t^i$ (e.g., this is the case when under the general policy v *or* under the dominating policy $\hat{v}$, there exists some $m \in Z_t^i$ for which $\kappa_t^i(m) = 0$).

Proof of Proposition 7 We prove the two points of the proposition in turn.

Proof of first point. For a simple family of alternative bandit processes, let π^* denote the expected sum of discounted rewards under the optimal full-control policy (i.e., the Gittins policy), and let π^Z denote the expected sum of discounted rewards under some alternative full-control policy Z. Glazebrook (1982) establishes that the difference between these two quantities is bounded by

$$\pi^* - \pi^Z \leq E_Z \left[\sum_{t=1}^{+\infty} \delta^{t-1} \left(\max_{i \in S} G_i(x_t) - G(Z, x_t) \right) \right], \tag{18}$$

where S is the set of bandit processes, x_t is the state of the system at time t and $G(Z, x_t)$ is the Gittins index of the bandit chosen in state x_t by policy Z (note that E_Z denotes expectation taken over realizations of x_t under policy Z).

In our decentralized model, the designer employs the Gittins-based heuristic in order to influence consumers' choices of service provider. According to this heuristic, the designer recommends the service of highest Gittins index whenever possible, taking into account the consumers' IC constraints. Note that, by design, the recommendations generated by the heuristic are guaranteed to be IC and are therefore followed by the consumer; as a result, the designer's recommendation policy may be viewed as a full-control, but nevertheless, suboptimal MAB policy. Let the sets U^t be defined as in the main text. Viewing equilibrium consumer choices as a suboptimal full-control policy, the designer uses the service of highest Gittins index with probability one in all states except those belonging to the sets U^t, $t \in T$. Thus, the contribution to the right-hand side of (18) of states $x_t \in X_t \setminus U^t$, $t \in T$, is zero. Whenever the system is in states $x_t \in U^t$ the

designer uses (recommends) the suboptimal provider with strictly positive probability, let this probability be q_{x_t}, and in this case the right-hand side of (18) incurs a penalty equal to $|\, G_A(x_t) - G_B(x_t)\,|$. Thus the expected period-t penalty is $p_{x_t} q_{x_t} \,|\, G_A(x_t) - G_B(x_t)\,|$ for $x_t \in U^t$. Summing up across periods we have

$$\pi^* - \pi(\hat{g}) \leq \sum_{t=1}^{+\infty} \sum_{x_t \in U^t} \delta^{t-1} p_{x_t} q_{x_t} \,|\, G_A(x_t) - G_B(x_t)\,| \leq \sum_{t=1}^{+\infty} \sum_{x_t \in U^t} \delta^{t-1} p_{x_t} \,|\, G_A(x_t) - G_B(x_t)\,|\,.$$

Proof of the second point. Note that if there exists a policy g^* such that $\pi(g^*) = \pi^*$ then, by Lemma 1, this implies that a recommendation policy which recommends in every period the provider of highest Gittins index is an ICRP. Next, note that the period-t objective function used in our heuristic along with the structure of the resulting period-t LP in (6) ensures that the policy extracted by the heuristic is precisely the ICRP which recommends the provider of highest Gittins index in all states (since this policy does not violate the consumers' IC constraints and maximizes the objective function of the period-t LP).

C. Identical Prior Beliefs

We present an example where first best is feasible in the decentralized system. In this example the prior belief over the quality of the two providers is identical. We state the following result as a corollary of Proposition 3 without proof.

COROLLARY 1. *Suppose $x_1^A = x_1^B$. Then there exists a messaging policy g^* such that $\pi(g^*) = \pi^*$.*

Note that when $x_1^A = x_1^B$, a full-control policy is indifferent between using service A or B at $t = 1$, and thereafter uses the service with highest Gittins index. To see how the designer can match this policy in the decentralized system, consider the following ICRP. At time $t = 1$, the designer randomizes and recommends either service with probability one half; in periods $t \geq 2$, the designer recommends the service with highest Gittins index. Incentive-compatibility for all customers under this policy is satisfied as follows: since the period-1 customer is indifferent between services, she follows the designer's recommendatioin irrespective of what this is. To the period-2 consumer, the past is perfectly symmetric, since the designer could have recommended, and observed an outcome from, any one of the two services in the first period (i.e., for any possible state $j = \{x_j^A, x_j^B\} \in X_2$ there exists an equiprobable state $k = \{x_k^A = x_j^B, x_k^B = x_j^A\}$). As a result, any recommendation that the designer makes in the second period is IC for the period-2 consumer, and the same logic applies to all consumers thereafter.

D. Figures

References

Acemoglu, D., M. A. Dahleh, I. Lobel, A. Ozdaglar. 2011. Bayesian learning in social networks. *The Review of Economic Studies* **78**(4) 1201–1236.

Alizamir, S., F. de Véricourt, P. Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157–171.

Allon, G., A. Bassamboo, I. Gurvich. 2011. "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations Research* **59**(6) 1382–1394.

Altman, E. 1999. *Constrained Markov Decision Processes*. CRC Press.

$x_1 = \{(a_1^A, b_1^A), (a_1^B, b_1^B)\}$	π^*	π^{FI}	$\pi(\hat{g})$	π^{NI}
$\{(6,3),(1,1)\}$	71.895 (0.11)	69.137 (0.10)	71.702 (0.11)	66.667 (0.00)
$\{(12,6),(1,1)\}$	71.442 (0.09)	67.558 (0.08)	70.725 (0.09)	66.667 (0.00)
$\{(18,9),(1,1)\}$	71.179 (0.09)	66.936 (0.07)	70.443 (0.08)	66.667 (0.00)
$\{(15,6),(2,1)\}$	78.152 (0.09)	75.336 (0.07)	78.006 (0.08)	71.428 (0.00)
$\{(15,6),(4,2)\}$	75.889 (0.07)	74.371 (0.07)	75.472 (0.07)	71.428 (0.00)
$\{(15,6),(6,3)\}$	74.859 (0.07)	73.890 (0.06)	74.619 (0.07)	71.428 (0.00)
$\{(1,1),(3,6)\}$	55.428 (0.10)	54.535 (0.10)	55.427 (0.10)	50.000 (0.00)
$\{(1,1),(6,12)\}$	55.042 (0.10)	54.009 (0.10)	55.042 (0.10)	50.000 (0.00)
$\{(1,1),(9,18)\}$	54.835 (0.10)	53.710 (0.10)	54.833 (0.10)	50.000 (0.00)
$\{(1,1),(3,6)\}$	55.616 (0.11)	54.795 (0.11)	55.312 (0.11)	50.000 (0.00)
$\{(3,3),(3,6)\}$	52.393 (0.08)	52.055 (0.08)	52.263 (0.08)	50.000 (0.00)
$\{(6,6),(3,6)\}$	51.227 (0.06)	50.627 (0.06)	51.180 (0.06)	50.000 (0.00)

Table 3 Simulated payoffs of the alternative policies for different initial states x_1 (numbers in parentheses denote standard errors): (i) first best (full control), π^*; (ii) full information, π^{FI}; (iii) Gittins-based heuristic with $K = 50$ (see Appendix A), $\pi(\hat{g})$; (iv) no information, π^{NI}. Parameter values: $\delta = 0.99$.

Anand, K. S., M.F. Pac, S. Veeraraghavan. 2011. Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science* **57**(1) 40–56.

Balseiro, S. R., J. Feldman, V. Mirrokni, S. Muthukrishnan. 2014. Yield optimization of display advertising with ad exchange. *Management Science* **60**(12) 2886–2907.

Banerjee, A.V. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics* **107**(3) 797–817.

Bellman, R. 1956. A problem in the sequential design of experiments. *Sankhya: The Indian Journal of Statistics* **30** 221–252.

Bergemann, D., J. Välimäki. 1997. Market diffusion with two-sided learning. *RAND Journal of Economics* **28**(4) 773–795.

Bertsimas, D., A. Mersereau. 2007. A learning approach for interactive marketing to a customer segment. *Operations Research* **55**(6) 1120–1135.

Besbes, O., Y. Gur, A. Zeevi. 2014. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Working paper, Columbia University.*

Bikhchandani, S., D. Hirshleifer, I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* **100**(5) 992–1026.

Bimpikis, K., K. Drakopoulos. 2015. Disclosing information in strategic experimentation. *Working paper, Stanford University.*

Bimpikis, K., S. Ehsani, M. Mostagir. 2015. Designing dynamic contests. *Working paper, Stanford University.*

Bose, S., G. Orosel, M. Ottaviani, L. Vesterlund. 2006. Dynamic monopoly pricing and herding. *The RAND Journal of Economics* **37**(4) 910–928.

Caro, F., J. Gallien. 2007. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science* **53**(2) 276–292.

Che, Y., J. Hörner. 2014. Optimal design for social learning. *Working paper, Columbia University.*

Crawford, V.P., J. Sobel. 1982. Strategic information transmission. *Econometrica* **50**(6) 1431–1451.

Debo, L., C. Parlour, U. Rajan. 2012. Signaling quality via queues. *Management Science* **58**(5) 876–891.

DeGroot, M. 2005. *Optimal Statistical Decisions*. John Wiley & Sons.

Frazier, P., D. Kempe, J. Kleinberg, R. Kleinberg. 2014. Incentivizing exploration. *Technical Report, Cornell University.*

Gittins, J., K. Glazebrook, R. Weber. 2011. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons.

Gittins, J., D. Jones. 1974. A dynamic allocation index for the sequential design of experiments. *Progress in Statistics* 241–266. Read at the 1972 European Meeting of Statisticians, Budapest.

Glazebrook, K. 1982. On the evaluation of suboptimal strategies for families of alternative bandit processes. *Journal of Applied Probability* **19** 716–722.

Ifrach, B., C. Maglaras, M. Scarsini. 2014. Bayesian social learning from consumer reviews. *Working paper, Columbia University.*

Kamenica, E., M. Gentzkow. 2011. Bayesian persuasion. *American Economic Review* **101**(6) 2590–2615.

Kostami, V., S. Rajagopalan. 2013. Speed-quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management* **16**(1) 104–118.

Kremer, I., Y. Mansour, M. Perry. 2013. Implementing the "wisdom of the crowd." *Journal of Political Economy, forthcoming.*

Lobel, I., A. Mani, J. Reed. 2015. Learning via external sales networks *Working Paper, New York University.*

Lobel, I., E. Sadler. 2015. Preferences, homophily, and social learning. *Operations Research, forthcoming.*

Marinesi, S., K. Girotra. 2013. Information acquisition through customer voting systems. *Working Paper, INSEAD.*

Marschak, J., K. Miyasawa. 1968. Economic comparability of information systems. *International Economic Review* **9**(2) 137–174.

Papanastasiou, Y., N. Bakshi, N. Savva. 2014. Scarcity strategies under quasi-bayesian social learning. *Working Paper, London Business School.*

Papanastasiou, Y., N. Savva. 2016. Dynamic pricing in the presence of social learning and strategic consumers. *Forthcoming, Management Science.*

Rayo, L., I. Segal. 2010. Optimal information disclosure. *Journal of Political Economy* **118**(5) 949–987.

Swinney, R. 2011. Selling to strategic consumers when product value is uncertain: The value of matching supply and demand. *Management Science* **57**(10) 1737–1751.

TripAdvisor. 2013. How rankings can be inconsistent with reviews (support forum entry). (Oct. 14th).

Veeraraghavan, S., L. Debo. 2009. Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* **11**(4) 543–562.

Ye, S., G. Aydin, S. Hu. 2015. Sponsored search marketing: Dynamic pricing and advertising for an online retailer. *Management Science, forthcoming.*

Yu, M., L. Debo, R. Kapuscinski. 2013. Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Management Science, forthcoming.*