# Crowdsourcing Exploration

Yiangos Papanastasiou
London Business School · yiangosp@london.edu

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

Nicos Savva
London Business School · nsavva@london.edu

**Preliminary Draft**

In an online review platform, consumers observe information regarding the past experiences of their peers, before choosing between alternative products or services whose quality is uncertain. In turn, their choices result in the generation of new experiences which they then report back to the platform. We investigate the problem of designing optimal information-provision policies, when the platform's 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 which dynamically discloses information regarding the history of outcomes to a series of short-lived rational agents (the consumers). We demonstrate that the platform's performance is non-monotone in the accuracy of its information-provision policy. Because consumers are constantly in "exploitation" mode, policies that disclose full and accurate information on past outcomes suffer from inadequate "exploration." We find that the designer can (partially) alleviate this inefficiency by employing a policy that strategically obfuscates the information at the platform's possession – interestingly, such policies are beneficial despite the fact that consumers are aware of both the platform's objective and the precise way by which information is being disclosed to them. The problem of explicitly designing optimal information-provision policies is found to be generally intractable; nevertheless, we establish the main structural properties of optimal policies and use them to propose an effective and computationally-tractable heuristic approach.

*Key words*: Bayesian social learning, exploration vs. exploitation, multi-armed bandit, Gittins index, incentive-compatibility

*History*: October 21, 2014

---

## 1. Introduction

The proliferation of online review platforms (i.e., websites hosting consumer-generated product evaluations) marks an important shift in consumers' approach to choosing between alternative products and services. Prior to the turn of the 21st century, consumer decisions were guided primarily by firms' promotional efforts and expert opinions. These traditional sources of information

have been, to a great extent, replaced by the real-life experiences of fellow consumers, with online review platforms serving the role of aggregating and disseminating these experiences – the typical modern-day consumer consults the reviews of movie-goers on *IMDb* before deciding which movie to watch; product scores on *Amazon.com* before choosing which product to buy; service evaluations on *Angie's list* before seeking a service-provider; hotel reviews on *Hotels.com* before booking a hotel; restaurant reviews on *Yelp* before choosing where to dine; traveller suggestions on *TripAdvisor* before selecting a vacation destination; etc.

The information hosted on such platforms is predominantly crowdsourced (i.e., generated by the platform's users). In selecting a product or service among the available alternatives, consumers make the choice they believe to be superior, given the information they observe at the time of their decision; in turn, any new information they contribute to the platform consists of an evaluation of this choice. Thus, information is generated as a byproduct of self-interested consumer behavior, and not taking into account the benefit of future visitors of the platform. From the perspective of the consumer population as a whole, this translates into inefficiencies 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. Such situations are also problematic in terms of the platform's performance (i.e., long-term success and profit), which is dependent on the quality of its content.

Since consumers' choices – and therefore the information they produce – are directly related to the information at their disposal, alternative modes of information-provision to visitors of the platform may result in different levels of platform performance. This notion is the focus of this paper.

**Problem Description** We consider a simple setting where a population of homogeneous short-lived consumers visit a review platform sequentially, choose among two alternative service-providers, and then report back to the platform whether the service they received was a success or a failure.[1] Upon being selected, each provider generates a successful service outcome with a fixed probability which is ex ante uncertain and unobservable throughout. At any time, the platform records the accumulated history of service outcomes for the two providers. We assume that a "designer" commits upfront to an *information-provision policy* which specifies the information posted on the platform at any time, given any possible recorded history. Each consumer that visits the platform uses the posted information to choose between the alternative providers. The designer's goal is to maximize the expected sum of consumers' discounted rewards over an infinite

[1] The two-provider setting is used for ease of exposition; our analysis extends readily to the case of multiple providers.

horizon (i.e., consumer surplus) by choosing the information-provision policy; each consumer seeks to maximize her instantaneous expected reward by choosing a service-provider.

The friction between the objectives of the forward-looking designer and the short-sighted consumers is the focal point of our model. Had consumers' actions been under the designer's full control, the designer would be faced with a classic instance of the multi-armed bandit (MAB) problem (two-armed, discrete-time, infinite-horizon, geometrically-discounted, Bernoulli rewards; see Gittins et al. (2011)). The solution to this classic problem, which exhibits the well-known "exploration-vs-exploitation" trade-off, is due to Gittins and Jones (1974), and consists of using in each period the arm of highest *Gittins index*. In our model, however, actions are taken by self-interested consumers who have no regard for the effects of their actions on the rewards of future consumers. The challenge faced by the designer is to structure the information on which consumers base their actions, so as to influence their decisions in a manner that benefits the platform. 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-vs-exploitation trade-off is directly linked to his ability to design an information-provision policy that renders his desired actions individually-rational for the consumers.

**Overview of Results** In the special case where the success probability of one of the two providers is known ex ante, the designer's problem is found to be tractable analytically. We use this case to highlight several interesting features of the designer's problem. We examine first the performance of policies that belong to the two extreme modes of information-provision, which we refer to as the "no-information" and the "full-information" regimes. Under no-information, the platform conceals all information at its possession at all times, while under full-information it discloses all information in its possession as precisely as possible. We establish that full-information policies result in higher platform performance than no-information policies, but fail to achieve first-best (i.e., performance under the Gittins-optimal policy). The latter observation is intuitive given existing knowledge on the MAB problem: consumers' choices under full-information collapse to the suboptimal "myopic" policy in the classic version of the problem, which chooses at any time the arm believed to generate the highest expected reward given the full history of outcomes.

More interestingly, we show that the designer may be able to improve the platform's performance with respect to full-information, by employing a policy which is deliberately *less informative* (i.e., a policy which lies, in a qualitative sense, between the two extreme modes of information provision mentioned above). Under such a policy, rather than providing consumers with a precise history of service outcomes, the platform reveals information on the basis of a coarser information structure,

in a manner we make precise in our analysis. By doing so, consumers' Bayesian interpretation of the information they observe results in actions which are desirable for the platform, but not necessarily for the individual consumers that take these actions – notably, this is true even though consumers know the precise manner by which information is being disclosed to them. We illustrate that whether the designer is able to restore first-best efficiency by optimizing the information-provision policy can be deduced by a simple inspection of the system's primitives (i.e., the designer and consumers' prior beliefs).

We then turn our attention to the more involved case where the success probability of both providers is ex ante unknown. Here, we demonstrate that less-than-fully informative policies remain optimal, although it is generally not possible for the designer to achieve first-best. We show that the designer's problem reduces to the solution of successive linear programs, the parameters of which can only be obtained via dynamic-programming computation. The latter observation renders the designer's problem intractable, but is useful in establishing the main structural properties of optimal policies. Combining these properties with existing theory on Gittins indices, we propose an intuitive and computationally-tractable heuristic approach to designing information-provision policies. Our heuristic is guaranteed to generate implementable policies, admits a tractable performance bound with respect to first-best, and is shown through numerical experiments to significantly outperform full-information policies.

**Implications** Traditionally, online review platforms have offered consumers the option to sort alternatives products/services according to the average score of all available consumer reviews (e.g., *Amazon.com*). In recent times, however, this transparent mode of information-provision appears to be losing ground to modified and less transparent ranking systems. For instance, the average-based ranking is often replaced by rankings that take into account not only the average rating but also the number of reviews, in ways that are not always obvious (e.g., *Yelp*). Other platforms only include products in their rankings that have received at least a specific number of reviews (e.g., *Booking.com*) or that have been in operation for long enough (e.g., *Zagat*). Alternatively, platforms may choose to rank services based on "popularity indices" whose inputs are not disclosed (e.g., *TripAdvisor*). Furthermore, others are moving towards simple recommendations of specific products as opposed to providing lists containing review information (e.g., NetFlix).

Our work suggests that there is a good rationale behind this shift towards information-provision structures that are less transparent. While traditional systems readily allowed the consumer to "exploit" the information at the platform's possession, these offer the platform the ability to "explore." This paper provides a novel framework for analyzing the complex problem of designing information-provision policies in the context of social learning. Our model highlights the main

challenge associated with designing optimal policies, namely, to provide information that is informative enough to be of use to the consumers, but at the same time sufficiently vague to ensure that information regarding all alternative options is generated efficiently. In the latter respect, our analysis provides valuable insights on how to *crowdsource exploration* in a manner, and to a level, that is compatible with consumers' self-interested objectives.

## 2. Related Literature

The multi-armed bandit (MAB) problem (see Gittins et al. (2011)) is recognized as the epitome of the *exploration-versus-exploitation* trade-off. In the classic version of the MAB problem, 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-inform 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; Bertsimas and Mersereau (2007) consider a marketer learning the efficacy of alternative marketing messages over time; Besbes et al. (2014) consider the case of non-stationary reward distributions. 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 consider in the present paper is essentially a decentralized MAB: there is a forward-looking principal (the designer) who seeks to maximize the long-term sum of discounted rewards, while actions are taken by a series of short-lived agents (the consumers). A similar setup to ours is used by 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. By contrast, in our model the only lever that the principal has control over is his information-provision policy – the designer tries to elicit desirable actions from the consumers by structuring the information on which they base their decisions.[2]

[2] Our approach is consistent with our motivating examples, since real-life platforms (e.g., Yelp) typically do not have control over the prices of the products they host and do not offer monetary incentives for consumers to use specific products. Furthermore, we note that the approach we are studying is a more cost-effective way of eliciting desirable actions from the consumers, that could also potentially be used in conjunction with other incentive mechanisms such as monetary payments.

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 and Bassamboo (2011), 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 lying about the signal realization); in this case, there is no issue of how the agent will interpret the principal's message.

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 (2014) 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. (2014) 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 social 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. 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-versus-exploitation trade-off as his state of knowledge progressively changes.

The information accumulated by the platform 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 2014). Our work is more related to the recent literature where social learning occurs on the basis of observable ex post outcomes (e.g., though online product reviews). For instance, Besbes and Scarsini (2013) investigate how the sequential nature of review-generation affects the informativeness of a product's average rating.

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 and outcomes, 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 (2014) and Yu et al. (2013) highlight how pricing policies are affected by the interaction between product reviews and strategic consumer behavior, while Papanastasiou et al. (2014) illustrate the beneficial effects of scarcity strategies when consumers learn according to an intuitive non-Bayesian rule. Our work contributes to this literature by investigating how the firm can influence consumer decisions through its information-provision policy, a lever which is potentially more cost-effective and can also be used in combination with other operational levers (e.g., pricing, inventory).

## 3. Model Description

We consider an online review 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 $S=\{A,B\}$. Each provider $i\in S$ is fully characterized by a probability $p_i$ which represents the provider's service quality. Upon using service $i$, 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 prior belief, which is expressed in our model through a Beta random variable with shape parameters $\{a_1^i,b_1^i\}$, $a_1^i,b_1^i\in\mathbb{Z}_+$. Recall that the probability density function of a $Beta(a,b)$ random variable is given by

$$g(x;a,b)=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)},\ x\in[0,1].$$

The Beta priors are chosen for their tractability in modeling Bayesian learning: after observing $s^i$ successful and $f^i$ failed trials with service provider $i$, the Bayesian posterior belief over the provider's quality, $p_i$, is $Beta(a_1^i+s^i,b_1^i+f^i)$ (e.g., DeGroot 2005, Chapter 9).

At the beginning of each time period $t \in T$, $T = \{1, 2, ...\}$, a single consumer visits the platform and chooses a service 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 combined with the knowledge contained in the prior beliefs can be summarized via the *information-state* (henceforth "state") $x_t = \{x_t^A, x_t^B\}$, where $x_t^i = \{a_t^i, b_t^i\} = \{a_1^i + s_t^i, b_1^i + f_t^i\}$ and $s_t^i$ ($f_t^i$) is the accumulated number of successes (failures) for service $i$ up to period $t$. When the state of the system is $x_t$, the expected immediate reward from using service $i$, denoted by $r(x_t, i)$, is

$$r(x_t, i) = \frac{a_t^i}{a_t^i + b_t^i}.$$

In any period, the history of service-outcomes of the alternative providers, which defines the state of the system $x_t$, is recorded by the platform and is not directly observable to the consumers, who know only the period of their arrival.[3] At the beginning of the time horizon, the designer commits to an information-provision policy which acts as an instrument of information-provision to the consumers.[4] Consumers are modelled as homogeneous, short-lived, rational agents, who seek to maximize their expected instantaneous reward through their choice of service provider. Upon visiting the platform, the period-$t$ consumer receives a "message" (i.e., observes a specific configuration of information), combines this message with the prior belief $x_1$, and chooses a service provider. The platform's objective is to maximize the expected sum of consumers' discounted rewards over an infinite horizon (i.e., consumer surplus), applying a discount factor of $\delta \in [0, 1)$.[5] 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 reward stream.

# 4. Analysis

## 4.1. Equilibrium Definition and Recommendation Policies

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*,

[3] The case where consumers do not know the period of their arrival is considered in §7.1.

[4] Commitment is a reasonable assumption in our context. Online review platforms disclose information on the basis of algorithms which are typically difficult and expensive to re-design once implemented. Furthermore, the large volume of products/services hosted on such platforms renders ad-hoc adjustments of the automatically-generated content prohibitively costly. The implications of relaxing this assumption are discussed in relevant parts of our analysis.

[5] The goal of maximizing consumer surplus is the most obvious way of expressing that the platform is interested in being helpful not only to its next visitor, but also to all its future visitors. This may be the case, for instance, when the platform's revenues are largely dependent on advertisement impressions, and therefore on the platform's overall traffic volume. We note that our model insights remain relevant for alternative, but related, objective functions that include, for instance, any increasing function of consumer surplus.

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 service provider).

Let $X \subseteq \mathbb{Z}_+^4$ be 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 at any time $t$. 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 actions $y_t$ that accompany the states $x_t$ are determined by the designer's policy and the consumers' strategy via $y_t = c_t(g(x_t))$. Transitions between system states occur as follows: the initial state $x_1$ is determined by the public prior beliefs over the two service 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 $t+1$, $x_{t+1} = \{x_{t+1}^A, x_{t+1}^B\}$ follows

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

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.[6]

We seek Bayesian Nash Equilibra, where the platform's private information in period $t$ may be viewed as the designer's type, over which the period-$t$ consumer has rational beliefs. There are two equilibrium requirements: first, given the designer's messaging policy, the period-$t$ consumer's choice strategy maximizes her individual expected reward. Second, given the consumers' response to any messaging policy, the designer chooses a policy which maximizes the total expected sum of consumers' discounted rewards. These requirements are formalized in the following definition.

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

DEFINITION 1. A messaging policy $g$ and a choice strategy $c$ form a Bayesian Nash Equilibrium (BNE) if:

1. Given a messaging policy $g$, for each $m \in M$, $c_t$ solves

$$\max_{c_t \in \mathcal{C}_t} E\left[r(x_t, c_t) \mid g(x_t) = m\right], \tag{1}$$

for all $t \in T$.[7]

2. Policy $g$ solves

$$\max_{g \in \mathcal{G}} E\left[\sum_{t=1}^{\infty} \delta^{t-1} r(x_t, y_t)\right], \text{ for } y_t = c_t(g(x_t)). \tag{2}$$

In general, multiple equilibria exist which generate the same payoff for the designer and 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 $M$. Following Allon et al. (2011), such equilibria are referred to 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" (ICRP).

DEFINITION 2 (ICRP: INCENTIVE-COMPATIBLE RECOMMENDATION POLICY). A recommendation policy is a messaging policy defined as

$$g(x_t) = \begin{cases} A & \text{w.p.} \quad q_{x_t} \\ B & \text{w.p. } 1 - q_{x_t}, \end{cases} \tag{3}$$

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 service $A$ or service $B$ to the period-$t$ consumer, and the consumer finds it Bayes-rational to follow this recommendation. We may now state the following result, which allows us to focus attention on the subclass of ICRPs.

LEMMA 1. *For any arbitrary messaging policy $g$, there exists an ICRP $g'$ such that*

$$c_t(g(x_t)) = c_t(g'(x_t)),$$

*for all $x_t \in X, t \in T$, where $c_t(g(x_t))$ and $c_t(g'(x_t))$ solve (1) under policies $g$ and $g'$, respectively.*

[7] For simplicity in exposition, we assume that if a consumer is indifferent between providers, she chooses the provider that is preferred by the designer. Alternatively, we may assume that some commonly known tie-breaking rule applies.

All proofs are relegated to the Appendix. 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 induce. As a direct consequence of Lemma 1, for any messaging policy that forms part of an equilibrium described in Definition 1, there exists an ICRP which induces a DOE equilibrium. Concrete examples of the correspondence between messaging policies and ICRPs appear in the following sections.

### 4.2. First-Best Provider Choices

It is instructive to think of the described model as a decentralized multi-armed bandit (MAB), in which the history of arm choices and outcomes is directly observable to a forward-looking designer, but the next arm to be pulled is chosen by a short-lived (and therefore myopic) consumer. The challenge faced by the designer is to regulate the information provided to each consumer, so as to achieve the highest possible platform performance, which is measured in terms of aggregate (discounted) consumer surplus.

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

$$G_i(z^i) = \sup_{\tau>0} \frac{E\left[\sum_{t=0}^{\tau-1} \delta^t r(x_t^i, i) \mid x_0^i = z^i\right]}{E\left[\sum_{t=0}^{\tau-1} \delta^t \mid x_0^i = z^i\right]}, \tag{4}$$

where $\tau$ is a past-measurable stopping time 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. An individual consumer's decision process is as follows. Upon entering the platform, the consumer has at his disposal three pieces of information: (i) his prior belief over the quality of alternative service providers, $x_1$; (ii) the time period, $t$ (relaxed in §7.1); (iii) the message he receives from the platform, $m$. Using this information, the consumer updates his belief over the current system state, $x_t$, and selects the service provider which maximizes his 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 choices in *all* periods. Using definition 2, we note that a sufficient condition for at least one such messaging

policy to exist is the existence of an ICRP which recommends in every period $t$ 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. An Incumbent Provider $B$

We analyze first a special version of our model, in which the market consists of one service provider whose quality is ex ante uncertain (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 several 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 $Beta(a_1^A, b_1^A)$ and recall that the expected immediate reward of a consumer who chooses service $A$ in period $t$ is given by $r(x_t, A) = \frac{a_t^A}{a_t^A + b_t^A}$, 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, the "safer" provider $B$ is preferred.

It will be useful to first characterize the provider choices which result when the full-control policy described in §4.2 is applied to the simplified setting considered here. To begin, note that if the quality of provider $B$ is known with certainty, then the provider has a constant Gittins index given by $G_B := G_B(x_t) = r_B$ (Gittins et al. 2011, Chapter 7). The implication of provider $B$'s constant index is that 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 periods $t > k$. As a result, system-optimal provider choices can be described in terms of "success thresholds" for the uncertain provider A.

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

- *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 $t \in T$.*
- *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*

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

In the first case of Lemma 2, experimentation with provider A is unattractive from the onset. Recalling that $G_B = r_B$, intuitively, if the incumbent provider'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 Lemma 2, experimentation with the new provider is attractive for the designer to begin with, but may cease to be so as more information about the service 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 of service provider.

### 5.1. The Two Extremes of Information Provision

Now let us return to the decentralized case. When consumers act autonomously, the designer can influence their decisions only through his messaging policy. In terms of the informational content of alternative policies, there are two extreme modes of information-provision. At one extreme, the designer may choose a policy which is completely uninformative (i.e., in any period $t$, the message received by the consumer reveals nothing about the system state); at the other extreme, the designer may choose a policy which is fully informative (i.e., in any period $t$, the message received by the consumer perfectly reveals the system state).

Consider first policies that are completely uninformative. As an example of such a policy, suppose that the designer discloses the same message to consumers at any time $t$ (or indeed no message at all), irrespective of the information that has been collected by the platform from previous consumers – policies of this kind are said to belong to the "no information" (NI) regime. Under the NI regime, consumers' choices in every period are trivially dictated by their prior belief, $x_1$. As a result, either all consumers choose service $A$ (when $r(x_1, A) > r_B$), or all consumers choose service $B$ (when $r(x_1, A) \leq r_B$). The unique ICRP which corresponds to the NI regime is therefore

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

and does not allow for any adaptation of consumers' provider-choices in response to the accumulated history of service outcomes.

Next, consider policies which are fully informative. For example, suppose that the designer adopts the messaging policy $g(x_t) = x_t$ – policies of this kind are said to belong to the "full information" (FI) regime. Under the FI regime, each consumer receives perfect state information, calculates her expected reward from using either provider, and chooses the provider which yields the highest expected reward. The unique ICRP which corresponds to the FI regime is

$$g(x_t) = \begin{cases} A & \text{if } r(x_t, A) > r_B \\ B & \text{if } r(x_t, A) \leq r_B, \end{cases} \tag{5}$$

and consumers' choices of provider in any period $t$ are summarized in Lemma 3.

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

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

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

Consumers' choices in Lemma 3 display a similar structure with the system-optimal choices of Lemma 2, but a closer comparison suggests two potential sources of inefficiency of the FI regime. First, note that 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 the FI regime. 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 potential source of inefficiency arises when $r(x_1, A) > r_B$. In this case, experimentation with service $A$ occurs under the FI regime at $t = 1$ and is also system-optimal (this follows from $G_A(x_1^A) \geq r(x_1, A)$; Gittins et al. (2011), Chapter 7); nevertheless, the *extent* to which experimentation occurs can be suboptimal, namely, if there is a discrepancy between any of the 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 $s^*(t) \leq \bar{s}(t)$.*

Equality holds 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. When the designer is not myopic, the FI regime will generally suffer from *under-experimentation*. That is, the self-interested consumers tend to abandon experimentation with the new provider $A$ before the system-optimal amount of experimentation has occurred; the following example illustrates.

EXAMPLE 1. Suppose that the prior belief over service provider $A$'s quality is $Beta(1,1)$, service $B$ has a known quality $p_2 = 0.27$ 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 $\bar{s}(2) = 0$; therefore, 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 $\bar{s}(3) = 1$; therefore, if both the 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 $s^*(3) = 0$.

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 $\pi^{NI}$ and $\pi^{FI}$ the platform's expected performance under policies belonging to the $NI$ and $FI$ regimes, respectively. Then*

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

*where $\pi^*$ denotes first-best expected platform performance.*

Put simply, FI policies outperform NI policies, but both extreme modes of information-provision fail to achieve first-best. 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 $G_A(x_1^A) \leq G_B$).

### 5.2. Strategic Information Provision

Messaging policies that provide consumers with a precise history of service outcomes maximize the expected reward of the individual period-$t$ consumer, but are not generally optimal from the platform's perspective. The main goal of this section is to demonstrate that in order to improve the platform's performance, the designer is required to *decrease* the accuracy of information-provision to the consumers (with respect to FI policies). We will be concerned with investigating the mechanics of this notion and discussing issues associated with the implementation of optimal messaging policies.

The result of Lemma 4 proves to be particularly important. When consumers have access to perfect state information, the possibility of premature abandonment of experimentation with provider $A$ arises. Thus, for the designer to achieve system-optimal provider choices, he is required to use his messaging policy to *prolong* the experimentation process – this can be achieved by adhering to a *coarser* information-provision structure.

Let us first establish the conditions under which first-best is achievable. Recall that for this to be the case, there must exist an ICRP which recommends in every period $t$ the service provider with highest Gittins index; that is, a policy which generates provider recommendations according to

$$g(x_t) = \begin{cases} A & \text{if } G_A(x_t) > G_B \\ B & \text{if } G_A(x_t) \leq G_B, \end{cases} \tag{6}$$

must be incentive-compatible (IC). Lemma 5 below suggests that whether this is the case can be verified by a simple inspection of the initial system state.

LEMMA 5. *If recommendation policy* (6) *is IC for* $t=1$*, then it is IC for all* $t \in T$*.*

If provider $B$ is recommended in the first period, and this recommendation is IC for the period-1 consumer, the result holds trivially since no information on provider $A$ is ever generated and subsequent periods are essentially repetitions of the first period (i.e., $B$ is recommended in all periods and is IC in all periods). If provider $A$ is recommended in the first period, and this recommendation is IC for the period-1 consumer, then the designer can implement a policy that recommends provider $A$ for as long as is system-optimal (even if provider $A$ does not maximize the immediate expected reward of the consumers), and all consumers will find it rational to follow the recommendations.

The mechanics underlying Lemma 5 are more interesting than the result itself. From the consumers' perspective, consider 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 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 than that of the individual consumer. Next, 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 of $s_t^A \geq \bar{s}(t)$: 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 of $s^*(t) \leq s_t^A < \bar{s}(t)$: here, it is provider $B$ that yields the highest expected reward for the individual consumer. However, by merging these two types of states into the $A$ recommendation, the designer is able to elicit a choice of provider $A$, even if the true underlying state is of the second type – upon being recommended provider $A$, the consumer rationally concludes that she is better off by heeding the platform's advice.

Thus, by employing a messaging policy which is less informative, 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. In effect, the designer's optimal policy allows for the use of some consumers as "guinea-pigs" for learning purposes; even though this possibility is acknowledged by the consumers, they still find it rational to take the designer's preferred action. Lemma 5 ensures that the interplay between the designer's recommendations and the state-dynamics of the system are favorable in that system-optimal provider choices remain IC in all periods, provided such a recommendation is IC for the period-1 consumer. Proposition 2 follows readily.

PROPOSITION 2. *For initial system state $x_1$, let $g^*$ be an optimal messaging policy. Denote by $\pi(g^*)$ the platform's expected performance under policy $g^*$. The following statements hold:*
- *If $r(x_1, A) > r_B$, then $\pi(g^*) = \pi^*$.*
- *If $r(x_1, A) \leq r_B$ and $G_A(x_1) \leq G_B$, then $\pi(g^*) = \pi^*$.*
- *If $r(x_1, A) \leq r_B$ and $G_A(x_1) > G_B$, then $\pi(g^*) < \pi^*$.*

Roughly speaking, first-best *cannot* be achieved by the designer when the expected quality of 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 prefer to select the safe provider $B$.[8] Furthermore, note that since the quality of provider $B$ here is deterministic, when the initial system state dictates that first-best performance cannot be achieved by the designer, this also implies that the designer's choice of messaging policy is completely irrelevant; we shall return to this observation when we consider the designer's general problem in §6.

When first-best *is* achievable, there will generally exist multiple optimal messaging policies, but all optimal policies have one feature in common: any state of the system $x_t$ for which $r(x_t, A) \leq r_B$ 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, some loss of accuracy in information-provision to the consumers is inevitable; the trade-off between the accuracy of information-provision to consumers and the platform's performance 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 $t = 4$. If experimentation with service $A$ has occurred in all previous periods (note that this can be deduced by the consumer upon observation of the platform's message), there are four possible states in period $t = 4$ (see Table 1). In three of the four possible 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.

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 examples

[8] Given that the difference in consumers' expected reward from using either provider may be low in cases where $r(x_1, A) < r_B$ and $G_A(x_1) \geq G_B$ hold simultaneously, it may be worthwhile for the provider (or indeed the designer), to offer subsidies for the initial experimentation, for example, in the form of a temporary price-reduction; such pricing decisions are beyond the scope of our analysis and are not pursued further in this paper.

**Table 1**

| $x_4^A=(a_4^A,b_4^A)$ | $P(x_4^A)$ | 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 ($G_A=0.30$) |

of optimal (pure) messaging policies (i.e., optimal deterministic 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.[9] The messages $m_1, m_2, m_3 \in M$ are arbitrary, provided the mapping from states to messages is common knowledge.

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

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

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

and concludes that she should use provider $A$, as desired by the designer. Finally, it is worth noting that any optimal messaging policy consists of a "garble" of a FI policy and is, therefore, less informative in the Blackwell sense (Marschak and Miyasawa 1968).

## Comments

1. **On the necessity of a priori commitment to a messaging policy**

   How important is the designer's a priori commitment to a messaging policy? Could the designer achieve first-best without this commitment? Under full information, consumers under-experiment with the uncertain service provider. Because of this feature of the setting considered here, it is straightforward to show that the equilibrium induced by an optimal messaging

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

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) consumers preceding her 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 the result of Lemma 5).

2. **On consumers knowing the exact period of their arrival**

   It is reasonable to assume that consumers know whether they have arrived early or late during the horizon. For convenience in analysis, our model takes this assumption to the extreme such that consumers know exactly the time period of their arrival. Alternatively, suppose that consumers have some arbitrary belief over the period of their arrival. If first-best is to be retained in this case, the policy which recommends the service of highest Gittins index in every period must remain IC for the consumers; the result of Proposition 6 in §7.1 proves that this is indeed the case. Thus, the conditions for first-best provided in Proposition 2 continue to hold under any arbitrary consumer beliefs.

While the above two assumptions can be relaxed in the analysis of this section without any loss in platform performance, this will generally not be the case in the version of the designer's problem we consider next.

## 6. Two Uncertain Providers

In the general problem setting, both service providers have an ex ante uncertain quality. The designer's problem in this case is significantly more complex. Importantly, the result of Lemma 5 no longer holds, and it will generally be the case that first-best cannot be achieved by the designer. But there is also some good news. Unlike in the simplified setting of §5, even if the designer cannot achieve first-best, his messaging policy significantly influences the platform's performance. We note that a direct characterization of "second-best" policies appears to be difficult, in particular because the rationale of recommending providers on the basis of their Gittins indices breaks down under second-best scenarios.

We approach the designer's problem in two steps. In the first step, we are concerned with identifying the structural properties of optimal messaging policies and their equivalent ICRPs. Without solving the designer's problem explicitly, we show that the optimality of less-than-fully informative policies carries through to the general setting and identify further properties associated

with optimal policies. Noting that an exact solution to the designer's problem is computationally intractable, in the second step we leverage our analysis to describe an intuitive heuristic approach to designing recommendation policies whose performance is evaluated through numerical experiments.

### 6.1. Structural Properties of Optimal ICRPs

Throughout the analysis of this section we will 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 provider $A$. Recall that in the simplified setting of §5, conditions on the initial system state $x_1$ were sufficient to guarantee the existence of a messaging policy which achieves first-best performance (see Proposition 2). As Proposition 3 and Example 2 below suggest, the condition for first-best in the general setting is less straightforward.

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 incentive-compatibility of recommendations for the ex ante *less* preferable (for the consumer) provider is a necessary and sufficient condition for achieving first-best platform performance. When the quality of provider $B$ is known with certainty (and with the help of Lemma 5), this reduces to simple conditions on the initial system state $x_1$ (Proposition 2). When both providers are of ex ante uncertain quality, a simple inspection of the initial state will not suffice: even if the initial system conditions are "favorable" for the designer (i.e., $r(x_1, A) > r(x_1, B)$ and $G_A(x_1) > G_B(x_1)$), the state dynamics of the system may be inherently inefficient. To demonstrate, we present the following example.

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). As Proposition 3 suggests, first-best cannot be achieved in this system because no ICRP exists which recommends provider $B$ with positive probability in period $t = 5$. To see why, note that the consumer's expected reward is maximized by choosing provider $A$ in period $t = 5$, irrespective of provider $A$'s outcome history.

Thus, Proposition 3 allows us to test, in forwards-induction fashion, whether efficiency is achievable by checking incentive-compatibility of recommendations for provider $B$. Doing so will generally rule out the possibility of first-best; the special case in which the prior beliefs over the two providers are identical is a rare (and fragile) exception, which we state 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 a full-control policy in this case 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 recommendation 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.

For the general case where the condition of Proposition 3 does not hold, what is the best possible outcome for the designer, and how can this outcome be achieved? Before moving forward to answer this question, it is instructive to take a step back. When first-best *is* achievable, the designer's objective in structuring his messaging policy is clear: the period-$t$ consumer, conditional on the information she receives upon her arrival, must rationally choose the provider of highest Gittins-index. By contrast, when first-best *is not* achievable, the designer's problem is more complex: for instance, if the designer can entice the period-$t$ consumer to "experiment" (i.e., to choose a provider which does not maximize her expected reward), it may be optimal to do so even if this is inconsistent with Gittins-optimality. To see why, note that Gittins-based reasoning depends on the designer being able to implement the Gittins-policy in all future periods (this holds under a first-best scenario) – since the ability to implement his desired actions in the future is not guaranteed under second-best scenarios, it may be in the designer's interest to explore when he can, even if this is not consistent with Gittins-optimality.

Thus, a direct characterization of second-best information-provision policies appears to be difficult. We seek first to extract the main features of optimal policies. To do so, we take the approach of assuming that the designer, given some initial system state $x_1$, has identified and implemented an optimal messaging policy.[10] We then ask which equilibrium conditions the corresponding ICRP must satisfy, with the goal of linking these conditions back to features of optimal messaging policies. We begin by defining an optimal ICRP $g^*$ as

$$g^*(x_t) = \begin{cases} A & \text{w.p.} \quad q^*_{x_t} \\ B & \text{w.p. } 1 - q^*_{x_t}, \end{cases} \tag{7}$$

[10] Existence of an optimal messaging policy follows trivially from the fact that the designer can always opt for full-information.

where $q^*_{x_t} \in [0,1]$ is the probability that the designer recommends provider $A$ when the state is $x_t$. Note that we allow also for randomized recommendations: conceivably, randomization may be beneficial from the perspective of satisfying consumers' IC constraints, even though the generated recommendation under such a policy will not always be system-optimal. Next, define $v(x_t)$ as the designer's expected value-to-go under policy $g^*$ when the state of the system is $x_t$. Finally, let $w(x_t, y_t) := r(x_t, y_t) + E[v(x_{t+1}) \mid x_t, y_t]$ and define for any period $t$ the "incentive-compatible sets"

$$IC^t_i = \{x_t : w(x_t, i) \geq w(x_t, i'),\ r(x_t, i) \geq r(x_t, i')\},$$

and the "non-incentive-compatible sets"

$$NC^t_i = \{x_t : w(x_t, i) > w(x_t, i'),\ r(x_t, i) < r(x_t, i')\},$$

where $i \neq i'$ and $i, i' \in S$. The sets $IC^t_i$ ($NC^t_i$) contain those states of the system which are possible at time $t$ and in which the designer and consumers would (would not) take the same action under perfect state information. With the above notation in hand, we state the following result.

PROPOSITION 4. *The optimal ICRP exhibits the following features:*

1. *In any period $t$,*

$$q^*_{x_t} = \begin{cases} 1 & \text{if } x_t \in IC^t_A \\ 0 & \text{if } x_t \in IC^t_B. \end{cases}$$

2. *In any period $t$, we have $0 < q^*_{x_t} < 1$ in at most one state belonging to one of the sets $NC^t_i$. For all remaining states $x_t \in NC^t_i$, we have $q^*_{x_t} = 0$ or $q^*_{x_t} = 1$.*

In any period $t$, the optimal ICRP $g^*$ maps any possible system state $x_t$ to a recommendation that is rationally followed by the period-$t$ consumer. In the proof of Proposition 4, we show that the period-$t$ recommendations, in equilibrium, solve the following linear program:

$$\begin{aligned} \max_{0 \leq q_k \leq 1} \quad & \sum_{k \in X} p_k q_k [w(k, A) - w(k, B)] \\ \text{s.t.} \quad & \sum_{k \in X} p_k (1 - q_k)[r(k, B) - r(k, A)] \geq 0, \end{aligned} \tag{8}$$

where $p_k$ denotes the probability that the system state in period $t$ is $k$. The objective function expresses the designer's goal to maximize the expected sum of the period-$t$ consumer's reward and his value-to-go. The constraint captures incentive-compatibility of a $B$ recommendation for the period-$t$ consumer; the corresponding constraint for incentive-compatibility of an $A$ recommendation is shown to be redundant, via an argument similar to that used in Proposition 3.

The optimal ICRP (i.e., the solution to (8)) exhibits the structural features described in Proposition 4; these are explained as follows. States that belong to the sets $IC^t_i$ are those in which the

designer and the period-$t$ consumer would take the same action under perfect state information. In these states, the optimal ICRP always recommends to the consumer the provider which maximizes her expected reward. The designer's gain by doing so is twofold: first, the action taken by the consumer is system-optimal (i.e., increases the objective function in (8)); second, recommendations of provider $i$ in states $x_t \in IC_i^t$ generate slack in the consumers' IC constraint (i.e., increases the left-hand side of the constraint in (8)). This slack is useful for the designer, because it allows him to recommend provider $i$ to the period-$t$ consumer in cases where the realized state of the system dictates that provider $i$ is system-optimal, but does not maximize the period-$t$ consumer's expected reward. On an intuitive level, upon entering the platform and receiving a recommendation to use service $i$, the period-$t$ consumer knows that this recommendation is more likely, but not guaranteed, to maximize her expected reward.

States that belong to the sets $NC_i^t$ are those in which the designer and the period-$t$ consumer would *not* take the same action under perfect state information. This part of the designer's problem can be described as a resource allocation subproblem. Recommendations in states $x_t \in IC_i^t$ generate IC slack (the "resource") which is to be allocated to recommendations in states $x_t \in NC_i^t$ (the "activities"). Recommending the system-optimal provider in such states incurs an IC penalty (the "cost" of each activity), while recommending the consumer-optimal provider generates *additional* IC slack (but results in a system-suboptimal consumer action). The solution to the resource-allocation subproblem consists of the designer selecting the optimal way of generating and using IC slack to induce system-optimal behavior in states $x_t \in NC_i^t$. Since the linear program has a single constraint, randomized recommendations are generated by the optimal ICRP in at most one state in any period $t$ – intuitively, the consumers' IC constraint for provider $B$ is binding whenever the designer would like to recommend $B$ to a larger extent than is afforded by consumers' incentives: the designer "gets away" with as many system-optimal recommendations as possible by pushing the period-$t$ consumer's incentive-compatibility to the limit.

It is worth noting that the designer's ability to direct consumers to provider $B$ is directly related to the slack generated in states where the provider is indeed the consumer's preferred option. A by-product of this feature is that, when the prior expectation over the two providers' quality is significantly different, there will exist initially a period of time in which the designer recommends, and consumers use, only provider $A$. To illustrate, in Example 2 above, notice that any ICRP recommends provider $A$ with probability one in periods $t \in [1, 8]$. Because $IC_B^t = \{\emptyset\}$ for $t \in [1, 8]$, provider $B$ will never be recommended in these periods, even if this is the system-optimal choice. More generally, any ICRP will initially recommend only service $A$ for at least $\lambda$ periods, where

$$\lambda = \left\{ \sup t : \frac{a_1^A}{a_1^A + b_1^A + t} > \frac{a_1^B}{a_1^B + b_1^B} \right\}$$

### 6.2. A Gittins-Based Heuristic Approach

When the designer has full-control over consumers' choices of provider, the system-optimal action in state $x_t$ is determined by comparing the Gittins indices of the alternative providers. By contrast, in the decentralized system the provider of highest index is not necessarily system-optimal, in particular because consumers' future IC constraints are effectively endogenous to the designer's chosen recommendation policy: what the designer recommends today affects what he is able to recommend in the future in a non-obvious manner. The optimal ICRP given an initial state $x_1$ is the solution to an infinite-horizon *Constrained* Markov Decision Process (see Altman 1999), in which the consumers' IC constraints are viewed as constraints on the designer's feasible actions when the system is in state $x_t$; the difficulty of the problem is further compounded by the fact that the period-$t$ constraints faced by the designer are endogenous to his policy in periods $[1, t-1]$.

In theory, the proof-technique of Proposition 4 can be used in a DP-based algorithm to extract the optimal ICRP. Unfortunately, this approach suffers from the "curse of dimensionality" and is computationally intractable even for the case of two service providers. (To illustrate, the number of possible states after just $t = 20$ periods is of the order $10^{12}$.) Furthermore, the problem of optimal information-provision in a decentralized MAB setting appears to be new to the literature, and as such, there are no existing heuristic approaches to designing information-provision policies. In this section, we use the preceding analysis to propose a Gittins-based heuristic approach. This approach (i) avoids DP computation, (ii) is guaranteed to generate IC policies, (iii) admits a tractable performance bound with respect to first-best, and (iv) is found to perform significantly better than FI policies.

To begin, we point out that the difficulty in solving the designer's problem exactly arises from the fact that the functions $w(x_t, y_t) = r(x_t, y_t) + E[v(x_{t+1}) \mid x_t, y_t]$, which are used as inputs to the designer's period-$t$ policy design LP (see proof of Proposition 4), are complicated by concerns regarding incentive-compatibility of future recommendations (in particular, these concerns manifest in the functions $v(x_{t+1})$). We take the simple approach of replacing the functions $w(x_t, y_t)$ with the Gittins indices $G_{y_t}(x_t)$. Gittins index values are described by Gittins and Wang (1992) as the sum of the immediate expected reward and a "learning component" which represents the potential future benefits of learning from the observed outcome. This description fits the functions $w(x_t, y_t)$, but with the difference that the latter are sensitive to the IC constraints of future consumers. We may say that the approach of replacing $w(x_t, y_t)$ by $G_{y_t}(x_t)$ is therefore one of (temporarily) disregarding the IC constraints of *future* consumers.

From the perspective of computation, the proposed approach allows the designer's recommendation policy to be constructed in a forwards-induction manner, using sequential solutions of the LP described in the proof of Proposition 4, thus alleviating the need for DP computation (see

also §6.3). From the perspective of feasibility of the extracted policy, note that in the period-$t$ policy-design LP, the IC constraints of the period-$t$ consumer are kept intact, thus guaranteeing that the policy generated by the heuristic approach will be incentive-compatible for the period-$t$ consumer, and consequentially for consumers in all periods $t \in T$.

Before presenting the results of our numerical study, it is also worth noting that the proposed approach admits a tractable upper bound on the difference between expected performance achieved under the heuristic policy and first-best performance. Define for any period $t$ the "approximate incentive-compatible sets"

$$AIC_i^t = \{x_t : G_i(x_t) \geq G_{i'}(x_t),\ r(x_t, i) \geq r(x_t, i')\},$$

and the "approximate non-incentive-compatible sets"

$$ANC_i^t = \{x_t : G_i(x_t) > G_{i'}(x_t),\ r(x_t, i) < r(x_t, i')\},$$

where $i \neq i'$ and $i, i' \in S$. These sets mirror the sets $IC_i^t$ and $NC_i^t$ of the previous section, but take into account the approximation of the functions $w(x_t, y_t)$ introduced by the heuristic. By design, the heuristic policy $\hat{g}$ will exhibit the same structural features as the optimal policy described in Proposition 4. We may therefore also define the following sets

$$\hat{U}_i^t = \{x_t : x_t \in ANC_i^t, \hat{q}_{x_t} = \mathbf{1}_{\{i=A\}}\},$$

for $i \in S$, which contain those states where it is system-optimal for the designer to recommend provider $i$ (even though this is not optimal for the period-$t$ consumer) and the designer does so, under policy $\hat{g}$, with probability one ($\hat{q}_{x_t}$ represents the probability with which the designer recommends service $A$ under policy $\hat{g}$). Finally, let

$$U^t = \cup_{i \in S} \left( ANC_i^t \setminus \hat{U}_i^t \right)$$

be the set of states at time $t$ in which the designer is forced, with at least some probability, to recommend a socially-suboptimal choice. We may then state the following result which utilizes Glazebrook (1982).

PROPOSITION 5. *Under policy $\hat{g}$, let $p_{x_t}$ denote the probability that the system state is $x_t$ when the period is $t$. The difference between $\pi^*$ and $\pi(\hat{g})$ is bounded from above by*

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

The above bound has an intuitive form, in that it accumulates a penalty whenever the heuristic policy fails to recommend the provider of highest Gittins index; furthermore, the penalty incurred in each such state is characterized simply by the Gittins-suboptimality of the recommended provider.

### 6.3. Numerical Study

In this section we present numerical results relating to the computation and performance of the Gittins-based heuristic proposed in the previous section. Our numerical experiments consist of two steps. In the first step, given an initial system state $x_1$ we generate and store offline the ICRP corresponding to the Gittins-based heuristic, that is, the mapping from system states to incentive-compatible recommendations. In the second step, we use Monte Carlo simulation to evaluate the performance of the generated ICRP and compare it against the performance of first-best (full-control) policies and policies belonging to the FI and NI regimes. We discuss each step in turn.

**ICRP Extraction** The inputs to the subroutine used to extract the Gittins-based ICRP are (i) the initial system state $x_1$, (ii) the designer's discount factor $\delta$, 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 the discount factor $\delta$. For each period $t$, we solve the approximate version of the designer's LP (i.e., the LP defined in (8), but with the modification of replacing $w(x_t, y_t)$ with Gittins indices $G_{y_t}(x_t)$), and 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 $p_{x_{t+1}}$. Extracting IC recommendations becomes computationally cumbersome after approximately $80-100$ periods depending on $x_1$ (owing to the large number of possible system states). We observe that using strategic recommendations beyond the point when computation becomes prohibitively slow is at most marginally beneficial in terms of system performance; for example, see Figure 1. Therefore, to increase computational efficiency, in our experiments we extract and use Gittins-based recommendations only for an initial time window (50 periods), after which an FI policy is employed, which requires no upfront computation and is guaranteed to be incentive-compatible.

**Performance Evaluation** Having extracted the heuristic ICRP, we compare its performance against that of full-control policies, FI and NI policies. We conduct Monte Carlo simulations and employ the *Bayesian approach* (e.g., see Caro and Gallien (2007)). In Table 2, we report for different initial states $x_1$ the average platform performance achieved under three policies: (i) $\pi^*$ achieved under a full-control Gittins-optimal policy; (ii) $\pi(\hat{g})$ achieved when a Gittins-based heuristic policy is employed in periods $t \in [1, 50]$, after which the designer discloses perfect state information; (iii) $\pi^{FI}$ achieved under a FI (full-information) policy which discloses perfect state information in all periods; (iv) $\pi^{NI}$ achieved under a NI (no-information) policy. We report also the upper bound on the difference $\pi^* - \pi(\hat{g})$ suggested by Proposition 5, which appears under the label $b$.

The upper part of the table contains results for initial states which are unfavorable for the designer, in the sense that there is misalignment between the provider of highest ex ante expected

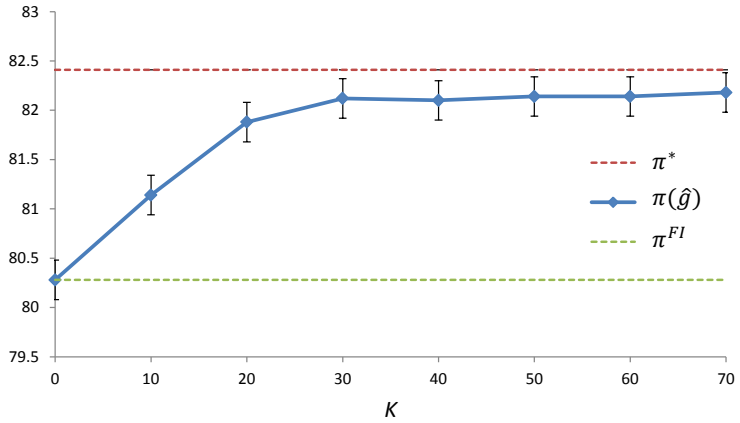


**Figure 1** **Expected platform performance when Gittins-based recommendations are used in periods $t \in [1, K]$, after which an FI policy is employed ($\pi(\hat{g})$); error bars mark 95% confidence intervals. Values $\pi^*$ and $\pi^{FI}$ represent expected performance under Gittins-optimal (i.e., full-control) and FI policies respectively. Parameter values: $x_1 = \{(11, 3), (2, 1)\}$, $\delta = 0.99$.**

reward and the provider of highest ex ante Gittins index; the lower part of the table contains results for favourable initial states. Note that across all initial states considered in Table 2, $\pi^*$ is not achievable in a decentralized system (for the upper part of the table, this follows from the unfavorable initial state; for the lower part, this can be verified via arguments such as the one used in Example 2). We observe that across all instances considered: (i) the Gittins-based heuristic performs close to first-best, especially when the initial system state is "favorable," and (ii) the performance of the Gittins-based recommendation system is statistically better than that of FI policies, particularly when the initial system state is "unfavorable."

**Table 2** **Performance of alternative policies at different initial states $x_1$. Numbers in parentheses denote standard errors. $\pi(\hat{q})$ are based on the use of the heuristic ICRP for the first 50 periods of the horizon, and use of FI policy thereafter. Parameter values: $\delta = 0.99$.**

| $x_1 = \{(a_1^A, b_1^A), (a_1^B, b_1^B)\}$ | $\pi^*$ | $\pi(\hat{g})$ | $\pi^{FI}$ | $\pi^{NI}$ | $b$ |
|---|---|---|---|---|---|
| $\{(6,3),(1,1)\}$ | 71.80 (0.07) | 71.65 (0.07) | 69.02 (0.06) | 66.67 (0.00) | 0.69 (0.01) |
| $\{(11,3),(2,1)\}$ | 82.41 (0.03) | 82.15 (0.03) | 80.28 (0.03) | 78.57 (0.00) | 0.49 (0.01) |
| $\{(22,6),(2,1)\}$ | 82.10 (0.07) | 81.64 (0.06) | 79.17 (0.05) | 78.57 (0.00) | 1.15 (0.01) |
| $\{(11,3),(6,3)\}$ | 80.04 (0.04) | 79.99 (0.04) | 79.65 (0.04) | 66.67 (0.00) | 0.28 (0.01) |
| $\{(1,1),(3,6)\}$ | 55.47 (0.07) | 55.34 (0.07) | 54.59 (0.07) | 50.00 (0.00) | 0.35 (0.01) |
| $\{(1,1),(6,12)\}$ | 55.19 (0.1) | 55.11 (0.1) | 54.05 (0.10) | 50.00 (0.00) | 0.26 (0.00) |
| $\{(2,1),(5,3)\}$ | 74.91 (0.07) | 74.86 (0.07) | 74.27 (0.07) | 66.67 (0.00) | 0.32 (0.00) |
| $\{(4,2),(5,3)\}$ | 72.81 (0.06) | 72.81 (0.06) | 72.18 (0.06) | 50.00 (0.00) | 0.36 (0.00) |

# 7. Extensions

## 7.1. Imperfect Knowledge of Consumers' Arrival Times

In our analysis, we have assumed that consumers know the exact time-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, the following result suggests that our approach is in fact a robust one.

PROPOSITION 6. *Let $g^*$ be an optimal ICRP when consumers have perfect knowledge of their arrival times. Then:*

*(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 6, 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 performance equal to $\pi(g^*)$ irrespective of what consumers' beliefs may be and irrespective of whether these beliefs are observable to the designer. Furthermore, the result also guarantees that the Gittins-based heuristic approach to policy-design proposed in §6.2 remains valid. When consumers do not have precise knowledge of their arrival time, the IC constraints faced by the designer become *less* stringent, because (in the language of §6.1) the designer can use consumers' probabilistic beliefs to "carry" IC slack across periods. The extent to which this may help the designer achieve higher performance will depend on the specific nature of consumers' beliefs. In any case, a robust and effective approach for the designer is to simply design his policy under the assumption that consumers are perfectly informed about their arrival times.

# 8. Conclusion

This paper represents a first attempt to understand how information-provision policies can be structured so as to "persuade" self-interested rational agents to take non-self-interested actions, in a decentralized multi-armed bandit setting characterized by the well-known dynamic exploration-versus-exploitation trade-off. To conclude this paper, we discuss briefly some further questions that arise from our analysis that may represent fruitful avenues for future research.

We have characterized the main properties of optimal information-provision policies, focusing on the case of homogeneous consumers. In reality, users of online review platforms are likely to be heterogeneous in their preferences; that is, given the same information, consumers may have a preference for different options depending on idiosyncratic factors. The consequences of consumer heterogeneity for optimal information-provision policies are unclear: presumably, heterogeneity in

preferences may decrease the need for information-manipulation by the platform, since consumers are now more likely to generate information on the alternative options irrespective of the information they receive; on the other hand, preference heterogeneity may also render the design of incentive-compatible policies harder, especially if the preferences of individual platform visitors are unobservable to the platform. Another related and interesting angle is that of capacity constraints, which may be modelled as random shocks on the consumers' actions. For example, given the information at her disposal, a consumer may wish to dine at a specific restaurant. If, however, this restaurant is fully booked, then the consumer may be forced to dine in alternative restaurant, resulting in "forced exploration." Thus, in the context of decentralized information-generation, capacity constraints may in fact be beneficial.

Furthermore, we have assumed that the quality of alternative service-providers remains fixed over time. However, in practical settings, quality levels may change over time due to factors exogenous to the social-learning process, or even endogenous, as providers may react to the reviews they receive. This raises a number of interesting questions: What is the platform's optimal information-provision policy when providers' reward-distribution is non-stationary? How does a provider set his quality level and how does he react to the reviews he receives? How does this response depend on the platform's mode of information-provision?

# Appendix

## A. Proofs

### Supporting Results

The following lemma is used in subsequent proofs.

LEMMA 6. *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. See, for example, Bellman (1956).

### 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 suppose that the consumer chooses the designer's preferred option.

We 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 $M_t^A = \{m : \text{period-t consumer chooses } A\}$ and $M_t^B = \{m : \text{period-t consumer chooses } B\}$. Now consider the recommendation policy $g'$, defined by

$$g'(x_t) = \begin{cases} A & \text{w.p. } \sum_{m \in M_t^A} P(g(x_t) = m) \\ B & \text{w.p. } \sum_{m \in M_t^B} P(g(x_t) = m). \end{cases} \tag{9}$$

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 $G_A(x_t) \leq G_B$ 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 $G_A(x_t) \leq G_B$ holds, at which point it is system-optimal to switch to $B$ and use $B$ forever after. We have $x_t = \{a_t^A, b_t^A\}$, where $a_t^A + b_t^A = (a_1^A + s_t^A) + (b_1^A + f_t^A) = a_1^A + b_1^A + t - 1$ and $s_t^A, f_t^A \geq 0$; that is, $x_t = \{a_1^A + s_t^A, t - 1 + b_1^A - s_t^A\}$. From property (iii) of Lemma 6, 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 $r_A(x_t) \leq r_B$ 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 $r_A(x_t) \leq r_B$ holds, at which point consumers switch to $B$ and use $B$ forever after. We have $x_t = \{a_t^A, b_t^A\}$, where $a_t^A + b_t^A = (a_1^A + s_t^A) + (b_1^A + f_t^A) = a_1^A + b_1^A + t - 1$ and $s_t^A, f_t^A \geq 0$; that is, $x_t = \{a_1^A + s_t^A, t - 1 + b_1^A - s_t^A\}$. Note that $r(x_t, A) = \frac{a_t^A}{a_t^A + b_t^A}$, 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 $s_t^A \geq \bar{s}(t)$ 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 $r_A(x_t, A) \leq G_A(x_t, A)$, a contradiction. We conclude that $s^*(t) \leq \bar{s}(t)$ 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 $\tau$ 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 $\pi^*$ 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 $\delta$ (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 $\delta$. When this is the case, we have $\pi^{FI} = \pi^*$, while when there is discrepancy between $\tau^*$ and $\hat{\tau}$ we have $\pi^{FI} < \pi^*$.

**Proof of Lemma 5**

The recommendation policy considered is

$$g(x_t) = \begin{cases} A & \text{if } G_A(x_t) > G_B \\ B & \text{if } G_A(x_t) \leq G_B, \end{cases} \tag{10}$$

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 (10) 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 (10) is IC in period $t$ if both of the following hold simultaneously

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

$$E[r(x_t, A) - r_B \mid g(x_t) = B] \leq 0. \tag{12}$$

The two conditions postulate that the period-$t$ consumer is better off (in expectation) by following the recommendation she receives, be it $A$ (11) or $B$ (12). Now, notice that condition (12) is guaranteed to hold by policy (10) as follows:

$$E[r(x_t, A) - r_B \mid g(x_t) = B] = E[r(x_t, A) - r_B \mid G_A(x_t) \leq G_B] \leq 0,$$

where we have first used the structure of policy (10) and then the Gittins index property $r(x_t, A) \leq G_A(x_t)$ and the fact that $G_B = r_B$. We next claim that under case (i), if condition (12) holds, then condition (11) 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 (10) (as is true under *any* recommendation policy), we have

$$\begin{aligned} r(x_1, A) - r_B &= E[r(x_t, A) - r_B] \\ &= E[r(x_t, A) - r_B \mid g(x_t) = A] P(g(x_t) = A) + E[r(x_t, A) - r_B \mid g(x_t) = B] P(g(x_t) = B). \end{aligned} \tag{13}$$

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 (12) 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 (11) is satisfied. We conclude that if (10) is IC in period $t=1$, then it is IC in all subsequent periods.

**Proof of Proposition 2**

The first two cases listed in the proposition follow from Lemma 5, since in these cases the Gittins recommendation policy (6) is IC in period $t=1$. Next, since $r(x_1, A) \leq G_A(x_1)$ for any $x_1$, the only remaining case not covered in the proof of Lemma 5 is case (iii) of Proposition 2, namely, $r(x_1, A) \leq r_B$ 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

$$g(x_t) = \begin{cases} A & \text{w.p.} \quad q_{x_t} \\ B & \text{w.p. } 1 - q_{x_t}, \end{cases} \tag{14}$$

where $q_{x_t} \in [0,1]$ for all $x_t \in X$. Policy (14) is IC in period $t$ if both of the following hold simultaneously

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

$$E[r(x_t, A) - r(x_t, B) \mid g(x_t) = B] \leq 0. \tag{16}$$

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 $r(x_1, A) - r(x_1, B) \geq 0$ 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

$$\begin{aligned} r(x_1, A) - r(x_t, B) &= E[r(x_t, A) - r(x_t, B)] \\ &= E[r(x_t, A) - r(x_t, B) \mid g(x_t) = A] P(g(x_t) = A) \\ &\quad + E[r(x_t, A) - r(x_t, B) \mid g(x_t) = B] P(g(x_t) = B). \end{aligned} \tag{17}$$

If a $B$ recommendation is IC under policy (14), then (16) holds and the second term of (17) is non-positive. It follows that the first term of (17) is non-negative (because $r(x_1, A) - r(x_1, B) \geq 0$), so that (15) 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**

Let $X_t$ be the set of possible states at time $t$ and let $p_k = P(x_t = k \mid x_t \in X_t)$. Let $v(x_t)$ be the designer's value-to-go when the system is in state $x_t$, and let $w(x_t, i) := r(x_t, i) + E[v(x_{t+1}) \mid x_t, i]$. Let $q_{x_t} \in [0,1]$ be the probability that the designer recommends service $A$ when the state of the system is $x_t$. Then, any optimal ICRP $g^*$ must solve the following problem in all periods $t$.

$$\begin{aligned} \max_{0 \leq q_{x_t} \leq 1} \quad & E_{x_t} E_{g(x_t)} w(x_t, g(x_t)) \\ \text{s.t.} \quad & E_{x_t}[r(x_t, A) \mid g(x_t) = A] \geq E_{x_t}[r(x_t, B) \mid g(x_t) = A] && (ICA) \\ & E_{x_t}[r(x_t, B) \mid g(x_t) = B] \geq E_{x_t}[r(x_t, A) \mid g(x_t) = B] && (ICB) \end{aligned} \tag{18}$$

That is, the designer's period-$t$ policy maximizes the expected sum of current reward plus value-to-go, subject to the period-$t$ consumer's IC constraints: for the policy to be IC, the consumer must find it more preferable to follow the designer's recommendation (expected reward given by the left-hand side of the constraints) as opposed to deviating from the recommendation (expected reward given by right-hand sides).

We note first that, using an argument similar to that used to arrive at (17) in the proof of Proposition 3, it follows that the assumption $r(x_1, A) \geq r(x_1, B)$ implies that constraint (ICA) in problem (18) is redundant, since if the policy $g^*$ satisfies constraint (ICB) then it is guaranteed to satisfy (ICA). Constraint (ICA) may therefore be disregarded. Next, for the objective function, we have

$$\begin{aligned} E_{x_t} E_{g(x_t)} w(x_t, g(x_t)) &= E_{x_t} \left[ P(g(x_t) = A) w(x_t, A) + P(g(x_t) = B) w(x_t, B) \right] \\ &= E_{x_t} \left[ q_{x_t} w(x_t, A) + (1 - q_{x_t}) w(x_t, B) \right] \\ &= \sum_{k \in X_t} p_k \left[ q_k w(k, A) + (1 - q_k) w(k, B) \right] \end{aligned}$$

For the IC constraint (ICB), we have for $i \in S$,

$$\begin{aligned} E_{x_t}[r(x_t, i) \mid g(x_t) = B] &= \sum_{k \in X_t} r(k, i) \frac{P(g(x_t) = B, x_t = k)}{P(g(x_t) = B)} \\ &= \sum_{k \in X_t} r(k, i) \frac{P(g(x_t) = B \mid x_t = k) P(x_t = k)}{\sum_{z \in X_t} P(g(x_t) = B \mid x_t = z) P(x_t = z)} \\ &= \sum_{k \in X_t} r(k, i) \frac{P(g(x_t) = B \mid x_t = k) P(x_t = k)}{\sum_{z \in X_t} P(g(x_t) = B \mid x_t = z) P(x_t = z)}, \end{aligned}$$

so that

$$E_{x_t}[r(x_t, i) \mid g(x_t) = B] = M_B \sum_{k \in X_t} p_k (1 - q_k) r(k, i),$$

for $M_B = \frac{1}{\sum_{z \in X_t} p_z (1 - q_z)} \geq 1$. It follows that problem (18) can be expressed as the following linear program.

$$\begin{aligned} \max_{0 \leq q_k \leq 1} \quad & \sum_{k \in X_t} p_k q_k [w(k, A) - w(k, B)] \\ \text{s.t.} \quad & \sum_{k \in X_t} p_k (1 - q_k) [r(k, B) - r(k, A)] \geq 0 \end{aligned}$$

Note first that the feasible region of the LP implies that we have $q_k^* \in \{0, 1\}$ for all but at most one state $k$. Furthermore, there are two relevant categories of states: (1) states with $w(k, i) \geq w(k, i')$ and $r(k, i) \geq r(k, i')$; (2) states with $w(k, i) \geq w(k, i')$ and $r(k, i) < r(k, i')$. For states belonging to group (1), it is clear that any optimal solution has $q_k = 1$ for $w(k, A) \geq w(k, B)$ and $q_k = 0$ for $w(k, A) \leq w(k, B)$. To see this, note that this maximizes the contribution of state $k$ to the objective function and weakly increases (depending on whether $q_k = 0$ or $q_k = 1$) the contribution of state $k$ to the left-hand side in the consumer's IC constraint. (Note: this means that when the state of the system belongs to group (1) it is always optimal for the designer to recommend deterministically the service of highest immediate reward). States belonging to group (2), have an opposite contribution to the objective function and the constraint. Therefore, the designer chooses the $q_k$ corresponding to these states so as to increase the objective function as much as possible without violating the consumer's IC constraint.

**Proof of Proposition 5**

For a simple family of alternative bandit processes, let $\pi^*$ denote the expected sum of discounted rewards under the optimal full-control policy (i.e., the Gittins policy), and let $\pi^Z$ 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

$$\pi^* - \pi^Z \leq E_Z \left[ \sum_{t=1}^{+\infty} \delta^{t-1} \left( \max_{i \in S} G_i(x_t) - G(Z, x_t) \right) \right], \tag{19}$$

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 (19) 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 (19) incurs a penalty equal to $\mid G_A(x_t) - G_B(x_t) \mid$. Thus the expected period-$t$ penalty is $p_{x_t} q_{x_t} \mid G_A(x_t) - G_B(x_t) \mid$ for $x_t \in U^t$. Summing up across periods we have

$$\pi^* - \pi(\hat{g}) \leq \sum_{t=1}^{+\infty} \sum_{x_t \in U^t} \delta^{t-1} p_{x_t} q_{x_t} \mid G_A(x_t) - G_B(x_t) \mid \tag{20}$$

$$\leq \sum_{t=1}^{+\infty} \sum_{x_t \in U^t} \delta^{t-1} p_{x_t} \mid G_A(x_t) - G_B(x_t) \mid, \tag{21}$$

where the last simplification is not severe since in every period we have $q_{x_t} = 1$ apart for possibly one state $x_t$ (see Proposition 4).

**Proof of Proposition 6**

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

$$E[r(x_t, m) \mid g(x_t) = m, t] \geq E[r(x_t, m') \mid g(x_t) = m, t] \tag{22}$$

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 $m \in \{A, B\}$, simply 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 performance. Under any arbitrary consumer beliefs, the designer can always implement the ICRP $g^*$ and achieve $\pi(g^*)$, while he may be able to improve on this policy by implementing a policy $v^*$ which depends on the specific beliefs held by the consumers; hence, $\pi(v^*) \geq \pi(g^*)$.

## 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.

Allon, G., A. Bassamboo. 2011. Buying from the babbling retailer? the impact of availability information on customer behavior. *Management Science* **57**(4) 713–726.

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.

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.

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.*

Besbes, O., M. Scarsini. 2013. On information distortions in online ratings. *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. 2014. Disclosing information in strategic experimentation. *Working paper, Stanford University.*

Bimpikis, K., S. Ehsani, M. Mostagir. 2014. 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.

Gittins, J., Y. Wang. 1992. The learning component of dynamic allocation indices. *The Annals of Statistics* **20**(3) 1625–1636.

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.

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

Lobel, I., E. Sadler. 2014. Preferences, homophily, and social learning. *Working paper, New York University.*

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. 2014. Dynamic pricing in the presence of social learning and strategic consumers. *Working Paper, London Business School.*

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

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

Yu, M., L. Debo, R. Kapuscinski. 2013. Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Working Paper, University of Chicago*.