Chapter 1
Inducing Exploration in Service Platforms

Kostas Bimpikis and Yiangos Papanastasiou

Abstract Crowd-sourced content in the form of online product reviews or recommendations is an integral feature of most Internet-based service platforms and marketplaces, including Yelp, TripAdvisor, Netflix, and Amazon. Customers may find such information useful when deciding between potential alternatives; at the same time, the process of generating such content is mainly driven by the customers' decisions themselves. In other words, the service platform or marketplace "explores" the set of available options through its customers' decisions, while they "exploit" the information they obtain from the platform about past experiences to determine whether and what to purchase. Unlike the extensive work on the trade-off between exploration and exploitation in the context of multi-armed bandits, the canonical framework we discuss in this chapter involves a principal that explores a set of options through the actions of self-interested agents. In this framework, the incentives of the principal and the agents towards exploration are misaligned, but the former can potentially incentivize the actions of the latter by appropriately designing a payment scheme or an information provision policy.

1.1 Introduction

An important function of most Internet-based platforms that act as intermediaries between customers and service providers is the provision of information regarding the quality of the potential alternatives faced by the consumers. As the service platform landscape continues to evolve, the dominant form of generating such information is through *crowdsourcing*: after transacting with a service provider, a customer

Kostas Bimpikis
Graduate School of Business, Stanford University, e-mail: kostasb@stanford.edu

Yiangos Papanastasiou
Haas School of Business, University of California, Berkeley, e-mail: yiangos@haas.berkeley.edu

may provide feedback on the provider's performance; this feedback is recorded by the platform and may become available to subsequent customers and assist them with their decision-making.

While soliciting feedback from customers is both straightforward and cost-effective, the crowdsourcing process through which information about the quality of the providers is generated is inherently inefficient from a system perspective, since it relies on the customers' self-interested choices. For an illustration of this inefficiency, consider the following example: a customer arrives at the platform and is presented with a choice between two providers, A and B. Provider A has eight "good" reviews and two "bad" reviews; Provider B has one of each. Given the available information (we assume that the customer is risk-neutral), provider A appears to be the better option; thus, the customer chooses A, and subsequently provides feedback on her choice. In fact, as long as provider A maintains a higher number of "good" reviews than "bad," he will always be preferred to provider B. However, this may not be the optimal outcome from a system perspective, which here refers to the outcome that maximizes the expected utility of the entire population of customers, because the customers' self-interested choices do not generate sufficient information on provider B to determine that he is, in fact, the inferior option.

The above example describes a phenomenon known in the experimentation literature as "under-exploration," as the self-interested individuals tend to take actions that "over-exploit" the information available to them. This chapter takes the perspective of a principal (e.g., the platform designer) who is interested in the efficient generation of information in such a system, where efficiency entails balancing exploration against exploitation with the goal of maximizing a long-run objective. Because the principal cannot dictate to the agents which action to take, she must find ways of incentivizing them to take system-optimal actions. Although we discuss a number of ways of achieving this, our main focus is on the active use of *information disclosure*, and in particular on the design of informational mechanisms that incentivize exploration in decentralized learning settings.

1.2 Related Literature

Studying the tradeoff between exploration and exploitation has a long research tradition in the context of the *multi-armed bandit* problem. In its classic version, a forward-looking decision maker makes a choice sequentially among a set of 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. When deciding which arm to play, the decision maker faces the tradeoff between the arm that she currently believes to be superior (exploitation) given the information she has at her disposal, or an alternative arm with the goal of acquiring knowledge that can be used to make better-informed decisions in the future (exploration). Since its inception, the multi-armed bandit framework has found

numerous applications in various real-world settings (e.g., Caro and Gallien, 2007, study dynamic assortment of seasonal goods in the presence of demand learning, while Bertsimas and Mersereau, 2007, consider learning in the context of developing marketing strategies).

In most existing applications of the multi-armed bandit framework, a single decision maker dynamically decides on the actions to be taken while observing the outcomes of her past actions. As such, the decision maker fully internalizes the benefits of exploration when taking actions that may not be optimal as far as maximizing her present payoff is concerned. In contrast, this chapter focuses on settings that can be essentially cast as decentralized multi-armed bandits problems: there is a forward-looking principal (the designer) who seeks to maximize a long-term objective, while actions are taken by a series of (short-lived) agents. In particular, we discuss recent work along this direction that is mostly motivated by the growing popularity of online recommendation platforms. In a nice contribution, Kremer et al. (2014) focus on eliciting experimentation in an environment where outcomes are deterministic, while Papanastasiou et al. (2017) consider a stochastic environment, in which the designer is effectively tasked with managing a dynamic exploration-exploitation trade-off. Furthermore, Che and Horner (2017) consider a single-product setting where a designer at any time optimally "spams" a fraction of consumers to learn about the product's quality. Frazier et al. (2014) aim to investigate how the principal can incentivize the agents to take her desired actions by offering direct monetary payments, i.e., their focus is not on the role of information disclosure policies (there is no ex ante or ex post asymmetry of information between the designer and the agents). Finally, Hörner and Skrzypacz (2016) also survey recent related work that combines ideas from experimentation, learning, and strategic interactions, with a particular emphasis on understanding how information but also delegation can be employed to deal with agents' incentives.[1]

Given the emphasis on the role of information the principal shares with the agents, the work we discuss here is related to, but quite distinct from, the well-developed literature on "cheap talk" (e.g., Crawford and Sobel, 1982; Allon et al., 2011). In cheap-talk games, the principal privately observes the realization of an informative signal, after which she (costlessly) communicates any message she 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. In contrast, the principal in the settings we consider 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 she discloses (e.g., by misrepresenting the signal realization). In this case, there is no issue of how the agents will interpret the messages; rather, the focus is on how the principal should structure credible messages in a manner that internalizes the misalignment between her and the consumers' objectives.

[1] Kleinberg and Slivkins (2017) also presented recently a comprehensive tutorial related to these issues.

As such, this chapter discusses work that is 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 (2016) 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. (2017b) 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.

Furthermore, the discussion here 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 (e.g., representative references along this direction include Acemoglu et al., 2011, 2014; Lobel and Sadler, 2015; Besbes and Scarsini, 2017).

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 Crapis et al. (2017) investigate dynamic pricing in the presence of social learning that occurs on the basis of actions (i.e., purchase decisions) and outcomes (i.e., product reviews), respectively. Veeraraghavan and Debo (2009) and Debo et al. (2012) consider how customers' queue-joining behavior depends on observable queue-length, and how service-rate decisions may be used to influence this behavior. Papanastasiou and Savva (2017) and Feldman et al. (2016) highlight how pricing and product-design policies are affected by the interaction between product reviews and strategic consumer behavior (see also Swinney (2011) for additional related work), while Allon and Zhang (2017) explore service-level differentiation for service organizations whose customers engage in communication through their social networks. Complementing this literature, the present chapter explores how the firm (platform) can influence consumer decisions and learning through its information-provision policy, a lever, which may also be used in conjunction with other operational levers (e.g., pricing, inventory).

Finally, the chapter is also broadly related to a recent line of work that studies operational decisions in the context of Internet-enabled business models. Among others, Marinesi et al. (2017) and Hu et al. (2013) study group-buying platforms; Balseiro et al. (2014) and Balseiro et al. (2015) consider the design and operations of ad-exchanges; Kanoria and Saban (2017) address search in two-sided platforms;

and Taylor (2016), Cachon et al. (2017), and Bimpikis et al. (2017a) explore optimal pricing and compensation policies in on-demand service platforms.[2]

1.3 Illustrative Example

The following example, which is taken from Kremer et al. (2014), provides a nice illustration of the setting and the questions we explore in this chapter.

Example 1. Agents choose sequentially between products A and B. Agent i makes her decision based on her prior on the quality of the two products and the information she obtains from the principal. In turn, the principal observes the choices and resulting payoffs of agents $1,\ldots,i-1$, and makes a recommendation to agent i, i.e., whether to purchase product A or B. The principal commits ex ante to the mechanism that generates the recommendation for agent i, i.e., the function that maps the actions and payoffs of agents $1,\ldots,i-1$ to a binary recommendation. Furthermore, agents know the mechanism set by the principal for generating recommendations and take it into account when they form their (posterior) beliefs about the quality of the two products.

Assume that the agents' common prior is that the quality of product A is uniformly distributed in $[-1,5]$ whereas the quality of product B is uniformly distributed in $[-5,5]$. Also, assume that when an agent buys a product, her (realized) payoff is equal to the quality of the product, i.e., one purchase is enough to reveal a product's (true) quality. Finally, suppose that the principal aims to explore both alternatives as soon as possible (so that she recommends the best one to future agents).

If information about past choices and outcomes were observable by the agents, the second agent would choose to take action B only if the payoff of the first agent (that would optimally take action A) was negative. Otherwise, i.e., if product A has positive payoff, the second agent (and subsequently all future agents) would choose product A and no agent would find it optimal to explore product B (which, nevertheless, could have been the optimal choice).

On the other hand, if agents do not directly observe prior choices and outcomes, the principal could induce more exploration by recommending action B to the second agent whenever the payoff associated with product A is less than one. In other words, the principal could send a binary message to the second agent: choose A if the first agent's payoff was higher than one and choose B, otherwise. Similarly, the principal can employ the following policy for the third agent: recommend choosing product B if (i) the second agent was recommended to choose product B and it turned out that B's payoff is higher than A's; or (ii) both the first and the second agent chose product A but its payoff is between 1 and 3.23. It is straightforward to show that following this policy guarantees that agents would have explored both

[2] There is also recent empirical work exploring operational issues on online marketplaces, e.g., Moon et al. (2017), Li and Netessine (2017), and Bimpikis et al. (2017c).

options by the third time period unless the payoff for product A is higher than 3.23 (one can similarly extend the policy for the fourth agent to ensure that by the fourth time period both options are explored with certainty).

In sum, agents find it optimal to follow the principal's recommendations, which, in turn, leads to more exploration and better outcomes on aggregate (assuming that the population of agents is large enough). The simple takeaway message that one can draw from this example is that by coarsening the information that the principal shares with the agents, she is able to mitigate their misalignment of interests.

1.4 Benchmark Model

Building on the discussion above, we consider a setting, where a series of agents interact with a principal who manages the disclosure of information regarding the experiences of their predecessors. For concreteness and to be in line with Sect. 1.3, we anchor our exposition in the example of an online platform which is operated by a designer and is used by customers to assist with their choice of a service provider. We assume that the marketplace features two providers, A and B; let $S = \{A, B\}$.[3] Each provider $i \in S$ is fully characterized by a probability p_i, which represents the provider's service quality. Upon using provider i, a customer receives reward equal to one with probability p_i, and equal to zero otherwise; 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 customers only to the extent of a common prior belief, which is expressed in our model through a Beta random variable with shape parameters $\{s_1^i, f_1^i\}$, with $s_1^i, f_1^i \in \mathbb{Z}_+$.[4,5]

At the beginning of each time period $t \in T$, $T = \{1, 2, \ldots\}$, a single customer visits the platform, observes information pertaining to the experiences of past customers, and chooses a provider. We assume that upon completion of service, and before the end of period t, the customer reports to the platform whether her experience was positive or negative (i.e., the realization of the Bernoulli random variable associated with her service experience). At any time t, the knowledge accumulated by the platform is summarized by the *information state* (henceforth "state") $x_t = \{x_t^A, x_t^B\}$, where $x_t^i = \{s_t^i, f_t^i\}$ and s_t^i (f_t^i) is the accumulated number of successful (failed) service outcomes for provider i up to period t (this includes the initial successes and

[3] Our analysis can be readily extended to the case of more than two providers.

[4] The probability density function of a $Beta(s, f)$ random variable is given by

$$g(x; s, f) = \frac{x^{s-1}(1-x)^{f-1}}{B(s, f)}, \text{ for } x \in [0, 1].$$

[5] The platform and the customers hold the same prior belief, so that platform actions (e.g., the choice of an information-provision policy) do not convey any additional information on provider quality to the customers (e.g., Bergemann and Välimäki, 1997; Bose et al., 2006; Papanastasiou and Savva, 2017).

failures, s_1^i and f_1^i, specified in the prior belief). When the system state is x_t, the Bayesian posterior belief over the quality p_i is $Beta(s_t^i, f_t^i)$, and the expected utility for the next customer if she uses i is $r(x_t, i) = s_t^i/(s_t^i + f_t^i)$.

In general, the history of service outcomes (i.e., the system state x_t) is not directly observable to the customers. Instead, there is a platform designer who *commits* upfront to a "messaging policy" that acts as an instrument of information-provision to the customers.[6] This policy specifies the *message* that is displayed on the platform, given any underlying system state. In addition, the platform may accompany messages with monetary payments to customers as a further incentive to induce them to take certain actions (in fact, Frazier et al., 2014, exclusively explores the case where all generated information is observable to customers and the platform has the discretion to incentivize their actions through monetary transfers in the form of "coupons").[7] The designer's objective in choosing her messaging policy is to maximize the expected sum of customers' discounted rewards over an infinite horizon (i.e., customer surplus), applying a discount factor of $\delta \in [0, 1)$.[8] Customers are modeled as homogeneous, short-lived, rational agents. In our main analysis, we assume that customers know the period of their arrival (however, the qualitative insights we obtain are robust to relaxing this assumption). Upon visiting the platform, each customer observes a message generated by the designer's policy and chooses a service provider with the goal of maximizing her individual expected reward.

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

1.5 Inducing Exploration

The section explores whether the designer can incentivize customers to take actions that contribute to her long-run objective of generating information about the available service providers using mainly the platform's messaging policy. At the end of the section, we also report on work that has studied the use of monetary transfers in a similar setting.

Equilibrium and Model Dynamics We begin our analysis by formalizing the strategic interaction between the designer and the customers. There are two main features of this interaction. First, the designer's *messaging policy*, which takes the

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

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

[8] More generally, our analysis is relevant for cases where the platform has a different (e.g., longer-run) objective than its users.

platform state as an input and generates a message to be displayed by the platform to the next incoming customer. Second, the customers' *choice strategy*, which takes the platform's message in any given period as an input and determines the customer's action (choice of provider).

Let $X \subseteq \mathbb{Z}_+^4$ denote the set of possible states of the platform such that $x_t \in X$ for all $t \in T$, and define the discrete set M of feasible messages that the platform can display to an incoming customer in period t (see footnote 7).

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 $\mathscr{G}$ be the set of possible messaging policies.

In each period t, a single customer enters the system, observes the platform's message and chooses a service provider from the set S. The period-t customer'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 $\mathscr{C}_t$ be the set of possible choice strategies for the period-t customer, and define $c(\cdot) := [c_1(\cdot), c_2(\cdot), ...]$.

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

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

The above transition probabilities reflect the learning dynamics of the system: new information regarding the quality of provider i is generated in period t only if the provider is chosen by the period-t customer.[9]

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

$$E_{x_t}[r(x_t, c_t) \mid g(x_t) = m].^{10}$$

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

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

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

Incentive-Compatible Recommendation Policies In general, multiple equilibria exist that result in the same payoff for the designer and the customers, and the same dynamics in the learning process, not least because the same information can be conveyed from the designer to the customers through a multitude of interchangeable messages contained in M. We follow Allon et al. (2011) in referring to such equilibria as being "dynamics-and-outcome equivalent". In our analysis, we will employ the result of Lemma 1 below to simplify the 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."

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

$$g(x_t) = \begin{cases} A & \text{w.p. } q_{x_t} \\ B & \text{w.p. } 1-q_{x_t}, \end{cases} \tag{1.1}$$

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

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

Lemma 1. *For any arbitrary messaging policy g, there exists an ICRP g' which induces a dynamics-and-outcome equivalent equilibrium in the game between the designer and the customers.*

A proof for Lemma 1 can be found in Papanastasiou et al. (2017).

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

First Best As a primer to our main analysis, we consider how the designer would direct individual customers to the two providers, had the customers' actions been under her *full control*. The solution to the designer's full-control problem is due to Gittins and Jones (1974) and consists of directing customers in each period to the provider with the highest *Gittins Index*. The Gittins index for service i when in state z^i is denoted by $G_i(z^i)$ and given by:

$$G_i(z^i) = \sup_{\tau>0} \frac{E\left[\sum_{t=0}^{\tau-1} \delta^t r(x_t^i, i) \mid x_0^i = z^i\right]}{E\left[\sum_{t=0}^{\tau-1} \delta^t \mid x_0^i = z^i\right]}, \tag{1.2}$$

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

In the decentralized system, the designer's ability to direct customers to her desired provider will be limited by the customers' self-interested behavior. Each customer knows (i) the prior belief summarized by the initial state, x_1; (ii) the time period, t; and (iii) the designer's messaging policy, g. Upon visiting the platform, the customer observes a message m, updates her belief over the current system state, x_t, and selects the provider which maximizes her individual expected reward. As a consequence, the designer will be able to achieve first-best only if she can design a messaging policy which induces customers to make Gittins-optimal decisions in all periods and in all system states—a sufficient condition for at least one such messaging policy to exist is the existence of an ICRP which always recommends the provider of highest Gittins index.

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

1.5.1 Strategic Information Disclosure

Typically, the provider will not be able to achieve the first best given that that Gittins-based recommendations (system optimal provider choice) are not incentive compatible in general. This section provides a characterization of the designer's optimal policy in the presence of incentive constraints resulting from the customers' decision making.

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

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

while at the same time ensuring that any recommendation received by the period-t customer is incentive compatible. The designer's general problem may be framed as

the following *Constrained* Markov Decision Process (CMDP; see Altman (1999)),

$$\begin{aligned}
&\max_{g(x_t)} E\left[\sum_{t\in T}\delta^{t-1}r(x_t,g(x_t))\right]\\
&\text{s.t. } E_{x_t}[r(x_t,A)\mid g(x_t)=A]\geq E_{x_t}[r(x_t,B)\mid g(x_t)=A],\ \forall t\in T,\\
&\qquad E_{x_t}[r(x_t,B)\mid g(x_t)=B]\geq E_{x_t}[r(x_t,A)\mid g(x_t)=B],\ \forall t\in T,
\end{aligned}\tag{1.3}$$

where the constraints state that any recommendation that is generated by policy g in period t is found to be incentive compatible (and is therefore followed) by the period-t customer.

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

To facilitate exposition of the result that follows, we introduce the following additional notation. Let X_t be the set of states that are reachable from the initial state x_1 (under some policy) in period t, so that the total state space is $X=\bigcup_{t\in T}X_t$. Denote by $\mathscr{P}_{kiz}$ the transition probability from state k to state z when provider i is used (note that these probabilities have been specified in §1.5), and let Δ_a denote the Dirac delta function concentrated at a.[11]

Proposition 1. *The optimal ICRP is given by*

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

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

$$\begin{aligned}
\max_{\rho}\ &\sum_{k\in X}\sum_{i\in S}\rho(k,i)r(k,i)\\
s.t.\ &\sum_{k\in X_t}\rho(k,B)[r(k,B)-r(k,A)]\geq 0, && \forall t\in T,\\
&\sum_{k\in X}\sum_{i\in S}\rho(k,i)\left(\Delta_z(k)-\delta\mathscr{P}_{kiz}\right)=\Delta_{x_1}(z), && \forall z\in X,\\
&\rho(k,i)\geq 0, && \forall k\in X,\ i\in S.
\end{aligned}\tag{1.4}$$

A few comments on the solution technique of Proposition 1 are warranted. To solve the designer's problem, the objective and constraints of the CMDP (1.3) are first ex-

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

pressed as sums of the immediate expected reward in each state-action pair, $r(k,i)$, multiplied by the time-discounted "occupancy" of the pair, $\rho(k,i)$. Then, the LP (1.4) optimizes over the admissible set of occupancy measures, which is described by the LP's constraints. In particular, in the context of our problem, any admissible occupancy measure must be consistent with (i) the customers' incentives (this is captured by the period-specific inequality constraints, which ensure that each period-t customer finds the recommendation she receives IC), and (ii) the system's dynamics (this is captured by the state-specific equality constraints, which ensure that the occupancy of each state is consistent with the system's state-transition probabilities).[12] Finally, once the optimal occupancy measure has been identified, the probabilities q_k^* are chosen in a manner that induces this measure.

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

1.5.2 The Value of Information Obfuscation

The "curse of dimensionality" renders the optimal solution to the designer's general problem computationally intractable. However, by combining the structural insights yielded by our analysis (i.e., state-merging, limited randomizations, sufficiency of two-message policies), it is possible to generate tractable and effective heuristic solutions. In this section, we consider one such heuristic and use it to establish that the value of information obfuscation is significant, even if this is implemented in a simple and intuitive manner (we note that the payoff under any heuristic serves as a lower bound on the payoff of the optimal policy described in Proposition 1).

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

[12] Note that the solution to LP (1.4) can also be used to retrieve the period-t customer's belief over the system state upon entry to the platform; specifically, this belief is given by $P(x_t = z) = \sum_{i\in S}\rho(z,i)/(\sum_{k\in X_t}\sum_{i\in S}\rho(k,i))$.

$$\max_{0\le q_{x_t}\le 1} \sum_{x_t\in X} p_{x_t} q_{x_t} [G_A(x_t) - G_B(x_t)]$$
$$\text{s.t.} \sum_{x_t\in X} p_{x_t}(1-q_{x_t})[r(x_t,B) - r(x_t,A)] \ge 0, \tag{1.5}$$

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

To evaluate the benefits of information obfuscation (in the sense of the Gittins-based heuristic), we conduct the numerical experiments presented in Table 1.1. The table focuses on the added "learning value" of obfuscation in comparison to that of a FI policy. Specifically, we first calculate the difference $(\pi^* - \pi^{NI})$, i.e., the difference between the platform's payoff when no social learning takes place (π^{NI}) and when social learning takes place optimally (π^*). This difference is an upper bound on the learning value that can be achieved by the designer in the decentralized system through information-provision. We then calculate the percentage of this value achieved under FI ($\Delta\pi^{FI}$) and under the Gittins-based heuristic ($\Delta\pi(\hat{g})$).

The upper half of the table pertains to initial states which are "unfavorable" for the designer, in the sense that there is an ex ante misalignment between the provider of highest expected reward and the provider of highest Gittins index; by contrast,

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

Table 1.1 Proportion of first-best learning value captured in the decentralized system by FI, defined as $\Delta\pi^{FI} = (\pi^{FI} - \pi^{NI})/(\pi^* - \pi^{NI})$, and by the Gittins-based heuristic $\hat{g}$ with $K = 50$, defined as $\Delta\pi(\hat{g}) = (\pi(\hat{g}) - \pi^{NI})/(\pi^* - \pi^{NI})$ (where π^*, π^{FI}, π^{NI} and $\pi(\hat{g})$ denote expected platform payoff under first best, FI, NI and the Gittins-based heuristic, respectively). $r(x_1, i)$ and std(x_1, i) denote, respectively, the expectation and standard deviation of the reward of provider $i \in \{A, B\}$ at the initial state x_1. Parameter values: $\delta = 0.99$.

the lower part of the table pertains to "favorable" initial states. Across all instances we consider, the heuristic performs significantly better than full information. Furthermore, we observe that the benefit is highest when the initial state is unfavorable: in such cases, under full information the customers tend to stick with the ex ante preferable provider and only rarely engage in experimentation with the alternative option. Next, notice that in each of the four subgroups of initial states, the ex ante expected reward of the two providers is maintained constant, but the variance of one of the two changes; this allows us to capture different environments in terms of the potential benefits of exploration. Here, intuitively, we observe that the benefits of information obfuscation are especially pronounced when the quality of the ex ante preferable provider is relatively certain while the quality of the alternative provider is relatively uncertain.

1.5.3 Minimizing Regret

In the setting we have considered so far, the designer's objective was to maximize the expected discounted sum of the customers' rewards over an infinite horizon. A related objective that has been studied in the literature is that of minimizing the designer's long-run *regret*. Typically, focusing on regret as the designer's objective simplifies the analysis (at least to some extent) and, thus, allows for a different set of results that mainly provide reasonable guarantees of performance for relatively simple strategies.

The discussion in this section follows Kremer et al. (2014) and Mansour et al. (2015). In the context of minimizing regret, the performance of a suggested policy is compared to the case that the designer knows the stochastic process generating the rewards for its customers, i.e., knows p_i for $i \in \{A, B\}$, and, thus, would always direct them towards the better of the two service providers.

Kremer et al. (2014) consider a horizon of T time periods (thus, T incoming customers) and propose the following policy for generating messages (recommendations) to them:

1. Customers are partitioned into $\lceil T/m \rceil$ blocks of m customers each. Customers belonging to the same block receive the same recommendation (and, thus, end up using the same service provider at the induced equilibrium). Customers in the first block are recommended to visit the provider that is ex ante more likely to generate higher expected rewards (based on the common prior), say provider A.
2. The designer observes the realizations of the rewards for the first m agents and computes the average empirical mean reward, $\hat{\mu}_A$, for provider A, i.e., the provider that is ex ante more likely to be a better choice for the customers.
3. Keeping $\hat{\mu}_A$ fixed throughout the horizon, the designer recommends provider B to the i-th block of customers, if $\hat{\mu}_A \in (\theta_{i-1}, \theta_i]$, where $\{\theta_i\}_{i=1}^{\lceil T/m \rceil}$ is a set of thresholds that the designer determines so that the recommendations she makes to customers are incentive compatible (essentially θ_i is such that customers would be

indifferent between following the designer's recommendation and choosing the provider that is ex ante more likely to generate higher rewards, i.e., provider A, if $\hat{\mu}_A$ was exactly equal to the threshold.[13] The first time customers are recommended to use provider B (and, as a result, end up using B), the designer computes $\hat{\mu}_B$ based on their realized rewards.

4. After the designer recommends provider B for the first time and computes $\hat{\mu}_B$, the messaging policy takes the following form:
   - If $\hat{\mu}_A \leq \theta_{i-1}$, the designer recommends the provider that corresponds to the highest empirical mean, i.e., she recommends provider A if $\hat{\mu}_A \geq \hat{\mu}_B$, and provider B otherwise.
   - $\hat{\mu}_A > \theta_i$, the designer recommends provider A.

Kremer et al. (2014) provide the following theorem for the performance of the messaging policy described above:

Theorem 1. *Setting the size of each block to $T^{2/3}\ln T$, i.e., $m = T^{2/3}\ln T$, guarantees that the average regret per customer, i.e., the expected difference between taking the best possible action and following the designer's recommendation, is bounded above by:*

$$C\frac{\ln T}{T^{1/3}},$$

where C is a constant that depends only on the priors.

In other words, the theorem above implies that as the horizon gets longer (equivalently, the population of customers getting recommendations from the platform increases), the average regret per customer becomes negligible. Thus, this simple policy that appropriately partitions customers into different blocks, achieves a reasonably good asymptotic performance compared to always choosing the best available service provider.

Mansour et al. (2015) extend Kremer et al. (2014) by considering a setting where in each of T time periods n new customers interact with the platform and simultaneously take an action (out of $k \geq 2$ potential alternatives). The payoff of each customer is determined by an underlying state (that captures the quality of each of the alternatives) and the actions of the rest of the customers in her cohort. As in Kremer et al. (2014), Mansour et al. (2015) present results on (simple) policies that are near-optimal (in a regret minimization sense) compared to the best-in-hindsight policy.[14]

[13] This is a natural generalization of the computation in the example of Sect. 1.3.

[14] Che and Horner (2017) also consider the problem of optimally designing recommendation policies in a setting where information about the quality of two potential alternatives arrives continuously over time—their setting uses the exponential bandit framework of Keller et al. (2005) as a building block.

1.5.4 Incentivizing Customers using Payments

Although we mainly focus on the role of the designer's information disclosure policy to induce system-optimal actions, it is reasonable to consider a setting where the designer may use monetary transfers as a way to promote exploration among the platform's customers. In particular, a very interesting direction for future work would be to extend the modeling framework in Sect. 1.4 to a setting where the designer can combine her messaging policy with monetary transfers with the objective of maximizing the discounted sum of the customers' expected rewards minus the corresponding transfers.

Relatedly, Frazier et al. (2014) explore the use of monetary transfers as a way to incentivize exploration when the information generated by the customers' past interactions with the service providers is available to both the designer but also to future customers, i.e., there is no ex post information asymmetry between the designer and the customers. For example, two extreme policies that one could consider (and Frazier et al. (2014) discuss briefly) would be the following:

1. The designer never compensates customers for taking an action. Then, each customers chooses the action that maximizes her one time period payoff based on the information generated by past actions. In other words, customers never "explore" and always "exploit" (based on the history of actions and payoffs that they observe). Such myopic behavior is typically suboptimal given that the rate of exploration is inefficient from the designer's perspective.
2. On the other extreme, the designer compensates customers sufficiently to induce customers to take the system-optimal action at every time period, i.e., the designer offers a payment to the customer taking action at time t, which is equal to the difference between the expected reward corresponding to the service provider with the highest Gittins index at t and the service provider that would be myopically optimal to choose given the available information. Obviously, this policy generates the system-optimal rate of exploration, but it may lead to large cumulative payments from the designer.

Frazier et al. (2014) provide a characterization of the extent to which payments from the designer to the customers can mitigate their incentive constraints and recover the optimal reward on aggregate. In particular, letting OPT denote the first best cumulative rewards, i.e., the discounted sum of the expected rewards corresponding to always choosing the service provider with the highest Gittins index, they call a point $(\alpha,\beta) \in [0,1]^2$ *achievable at discount rate* δ if there exists a payment policy for the designer, i.e., a mapping from the history of observations to payments to customers, that satisfies the following two conditions:

1. The discounted sum of the expected rewards corresponding to the policy is at least as high as $\alpha \cdot \text{OPT}$.
2. The discounted sum of the expected payments corresponding to the policy is at most as high as $\beta \cdot \text{OPT}$.

The main result in Frazier et al. (2014) is the following theorem that provides a remarkably clean characterization of the set of points that can be achieved by the designer.

Theorem 2. *Let* (α, β) *be a point in* $[0,1]^2$. *Then,* (α, β) *is achievable at discount rate* δ *if and only if the following condition holds:*

$$\sqrt{\beta} + \sqrt{1-\alpha} > \sqrt{\delta}.$$

Theorem 2 provides some insight on what the designer can (and cannot) do using payments ("coupons") to incentivize the platform's customers. A basic ingredient of the proof is a set of policies that involve mixing between the two extremes described above, i.e., the policy that involves no payments and the one where payments are large enough to induce customers to take the action with the highest Gittins index (thus, giving some idea on what type of policies may lead to good performance for the designer). Note that this is a worst-case result, i.e., it bounds the designer's performance against any distribution of rewards. Thus, the designer could potentially achieve a higher discounted sum of expected rewards in environments where the uncertainty in payoffs takes a more specific form, like the one specified in Sect. 1.4.[15] Importantly though, Theorem 2 assumes that customers can observe the entire history of actions and their corresponding rewards; thus, it does not offer any insight on how the designer might appropriately disclose information to increase the set of achievable points.

1.6 Promising Directions

So far, we have mainly used the example of an online recommendation platform to describe the main questions that motivate this chapter and illustrate a number of key findings. However, the idea that a platform (principal) can appropriately design the information flow to its users (agents) as a way to incentivize them to take actions that may not be myopically optimal is more widely applicable and, to a large extent, still unexplored. In this section, we briefly describe two other application settings that may provide interesting starting points for future work in the area.

1.6.1 Learning in Dynamic Contests

Innovation contests are gaining in popularity as a tool that firms and institutions employ to outsource their R&D and innovation efforts to the crowd. An open call

[15] However, the analysis may, in general, be quite challenging.

is placed for a project that participants compete to finish and the winners, if any, are awarded a prize. Recent successful examples include The NetFlix Prize and the Heritage Prize,[16] and a growing number of ventures like Innocentive, TopCoder, and Kaggle provide online platforms to connect innovation seekers with potential innovators.

The objective of the contest designer is to maximize the probability of reaching the innovation goal while minimizing the time it takes to complete the project. Obviously, the success of a contest depends crucially on the pool of participants and the amount of effort they decide to exert. Typically, innovation projects have the following three key features. First, progress towards the end goal takes the form of a series of breakthroughs interspersed between long intervals of seeming stagnation. Second, and quite importantly, it is not clear at the onset whether the end goal is attainable, even if it is clearly specified, or which of potentially many alternative approaches would be the best one to use. Finally, a third feature that distinguishes innovation contests from other settings involving competition among agents, is that agents can learn from one another: an agent's (partial) progress towards the goal provides useful information to the rest about the feasibility of the project and/or the best approach to follow.[17]

These three features imply that news about a participant's progress has the following interesting dual role: it makes agents more optimistic about the state of the world, as the goal is more likely to be attainable; thus, agents have a higher incentive to exert costly effort. We call this the *encouragement effect.*[18] At the same time, such information implies that one of the participants has a lead, which might negatively affect effort provision from the remaining agents as the likelihood of them beating the leader and winning the prize becomes slimmer. We refer to this as the *competition* effect. These two effects interact with one other in subtle ways over the duration of the contest, and understanding this interaction is of first-order importance for contest design.

Thus, the contest's information disclosure policy (e.g., through intermediate milestone awards) may have a large effect on the agents' participation and effort provision and, consequently, on the likelihood that the contest will be successful. In recent work, Bimpikis et al. (2017b) consider the question of *whether* and *when* should the contest designer disclose information regarding the competitors' (partial) progress with the goal of maximizing her expected payoff. Interestingly, they illustrate the benefits of non-trivial information disclosure policies, where the designer withholds information from the agents and only releases it after a certain amount

[16] The NetFlix Prize offered a million dollars to anyone who succeeded in improving the company's recommendation algorithm by a certain margin and was concluded in 2009. The Heritage Prize was a multi-year contest whose goal was to provide an algorithm that predicts patient readmissions to hospitals. A successful breakthrough was obtained in 2013.

[17] In addition to the work that we discuss here, which mainly focuses on the dynamics of learning and competition in contests, there is also an extensive body of work that explore a number of questions in a static framework, e.g., Terwiesch and Xu (2008), Ales et al. (2017), and Körpeoğlu and Cho (2017).

[18] The term "encouragement" originates from the literature on strategic experimentation (e.g., Bolton and Harris, 1999; Keller et al., 2005)

of time has elapsed. Such designs further highlight the active role that information may play in incentivizing agents to participate in the contest.

Second, they identify the role of intermediate awards as a way for the designer to implement the desired information disclosure policy—the policy that maximizes the effort provision of the agents and consequently the chances of innovation taking place. Intermediate awards are very common in innovation contests (the aforementioned NetFlix and Heritage prizes are examples of contests that have employed intermediate awards) and the discussion here provides a potential reason for their ubiquity.

A simple illustration of the main ideas in Bimpikis et al. (2017b) is the following. Consider an innovation contest that consists of well-defined milestones. For example, the goal of the Netflix prize was to achieve an improvement of 10% over the company's proprietary algorithm, with a first progress prize set at 1% improvement. In this example, reaching the milestone of 1% improvement constitutes *partial progress* towards the goal, and we assume that the agents and the designer are able to verifiably communicate this. Assume for now that the innovation is attainable with certainty given enough effort, and that agents are fully aware of that. The lack of progress towards the goal is then solely a result of the stochastic return on effort. When no information is disclosed about the agents' progress, they become progressively more pessimistic about the prospect of them winning, as they believe that someone must have made progress and that they are now lagging behind in the race towards the end goal. This might lead them to abandon the contest, thus decreasing the aggregate level of effort and consequently increasing the time to complete the innovation project.

In contrast, when there is uncertainty about the feasibility of the end goal, agents that have made little or no progress towards the goal become pessimistic about whether it is even possible to complete the contest. If this persists, an agent may drop out of the competition as she believes that it is not worth putting the effort for what is likely an unattainable goal, reducing aggregate experimentation in the process and decreasing the chances of reaching a feasible innovation.

This highlights the complex role that information about the agents' progress play in this environment. In the first scenario, when the competition effect is dominant (since there is no uncertainty about the attainability of the end goal), disclosing that one of the participants is ahead may deter future effort provision as it implies that the probability of winning is lower for the laggards. In the second case, when the encouragement effect dominates, an agent's progress can be perceived as good news, since it reduces the uncertainty regarding the feasibility of the end goal.[19]

Several directions may be worth pursuing following the main ideas presented in this section. Bimpikis et al. (2017b) and Halac et al. (2017) consider contests with well-defined goals that may be unattainable. Alternatively, one could consider a setting where the contest designer's goal is to obtain the best solution or implementation possible to a given project. There can exist multiple approaches that one could

[19] Bimpikis and Drakopoulos (2016) and Halac et al. (2017) also consider a strategic experimentation framework to study the interplay between a principal and the agents' incentives and how appropriately designing information disclosure mechanisms may increase welfare.

employ and observing each other's progress reveals information about their relative merits.[20] What is then the optimal way to disclose information as way to balance the tradeoff between exploring the space of alternative approaches and driving effort towards those that look most promising?

An agent's progress provides information not only about the feasibility of the innovation project or the quality of the approach she is employing but also her skill level. Thus, in a setting where there is uncertainty about how good the competition is, appropriately designing an information disclosure policy may play an important role in keeping agents engaged and willing to exert effort.[21]

1.6.2 Dealing with Misinformation

The 2016 US presidential elections and the associated "fake news" phenomenon highlighted the importance of incentivizing a new form of "exploration" in the online space. Rather than exploration with the goal of identifying the quality of a product, the term here refers to exploration with the goal of evaluating the quality of an information source; for instance, if the information source in question is a news article circulating in a social media platform, exploration refers to the process of "fact-checking" the article's content to determine its validity.

With this context in mind, Papanastasiou (2017) develops a sequential model of news propagation with endogenous fact-checking, and identifies pathological outcomes whereby fabricated news articles fail to be detected and subsequently spread throughout the society, manipulating the beliefs of the agents in the process. In this setting, there is "under-exploration" from a system perspective in the sense that individual agents do not internalize the impact of their fact-checking decisions on the actions and beliefs of their downstream peers, which in turn may result in inadequate levels of fact-checking. The study proceeds by analyzing a first-order defense against the propagation of fake news involving a social media platform that decides whether and when to intervene with the sharing of a news article by conducting its own fact-check (an approach recently adopted by Facebook).

An alternative to the platform conducting its own fact-checks is that of incentivizing the agents to do so through direct monetary payments, in a manner analogous to the setup of Frazier et al. (2014). Perhaps a more interesting avenue, however, is the design of appropriate information-disclosure mechanisms that may be able to achieve the same effect in a cost-effective way. One insight from Papanastasiou (2017) is that the fact-checking decisions of agents are influenced to a large extent by the number of times the information in question has been shared between their peers. Thus, one might expect that the concealment of such information by the plat-

[20] Girotra et al. (2010), Kornish and Ulrich (2011), Huang et al. (2014), and Jiang et al. (2016) are recent empirical studies that consider the role of learning and feedback in crowdsourcing contests and, more broadly, in the innovation process.

[21] There are also a number of notable recent papers that consider different aspects of contest design and its applications, e.g., Seel and Strack (2016), Hu and Wang (2017), and Strack (2016).

form may increase the amount of scrutiny an article undergoes, thereby reducing the propagation of fabricated information. At the same time, too much fact-checking is also an inefficient outcome: every fact-check incurs a cost to the agents which may be unnecessary. It follows that, as in Papanastasiou et al. (2017), the optimal information policy must strike a balance between allowing the agents to exploit the information generated by their peers, while also motivating them to explore (fact-check) at a system-optimal level.

In a somewhat related direction, Candogan and Drakopoulos (2017) study the tradeoff between user engagement and misinformation in the context of an online social networking platform. The content available on the platform may contain inaccuracies and false claims. The platform, which knows the quality of its content, may use a signaling device, e.g., recommend whether agents engage or not with the content, so as to induce a desired engagement behavior. A main emphasis in this line of work is the interplay between the platform's (signaling) policy and the structure of the agents' social network.

1.7 Concluding Remarks

This chapter showcases that choosing whether, when, and what information to disclose to agents may have a first-order impact on the payoff of a principal. Most of the exposition centers around the example of an online recommendation platform (e.g., Yelp or Tripadvisor) but as we highlight in Sect. 1.6 these ideas may apply to many more real-world settings. Our hope is that the discussion provided here makes clear that information disclosure policies may be effective operational levers especially in the context of online platforms that rely on their users for ensuring a high quality of service. We believe that the role of information flows in mitigating the potential misalignment of interests between a principal and an agent/set of agents is quite important and relatively unexplored, and may thus provide a fruitful avenue for future research.[22] Although the scope of the ideas presented here is quite broad, we expect that they will be particularly relevant in the design and operations of online platforms and marketplaces.

References

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

Acemoglu D, Bimpikis K, Ozdaglar A (2014) Dynamics of information exchange in endogenous social networks. Theoretical Economics 9(1):41–97

[22] There is currently significant interest in the role of information in mitigating the potential misalignment of interests between a principal and an agent/set of agents, e.g., Renault et al. (2017), Ely (2017), and Orlov et al. (2017).

Ales L, Cho SH, Körpeoğlu E (2017) Optimal award scheme in innovation tournaments. Operations Research. Forthcoming

Allon G, Zhang DJ (2017) Managing service systems in the presence of social networks. Working paper

Allon G, Bassamboo A, Gurvich I (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

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

Balseiro SR, Besbes O, Weintraub GY (2015) Repeated auctions with budgets in ad exchanges: Approximations and design. Management Science 61(4):864–884

Banerjee A (1992) A simple model of herd behavior. The Quarterly J Economics 107(3):797–817

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

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

Besbes O, Scarsini M (2017) On information distortions in online ratings. Operations Research. Forthcoming

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

Bimpikis K, Drakopoulos K (2016) Disclosing information in strategic experimentation. Working paper

Bimpikis K, Candogan O, Saban D (2017a) Spatial pricing in ride-sharing networks. Working paper

Bimpikis K, Ehsani S, Mostagir M (2017b) Designing dynamic contests. Operations Research. Forthcoming

Bimpikis K, Elmaghraby WJ, Moon K, Zhang W (2017c) Managing market thickness in online B2B markets. Working paper

Bolton P, Harris C (1999) Strategic experimentation. Econometrica 67(2):349–374

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

Cachon GP, Daniels KM, Lobel R (2017) The role of surge pricing on a service platform with self-scheduling capacity. Manufacturing & Service Operations Management 19(3):368–384

Candogan O, Drakopoulos K (2017) Optimal signaling of content accuracy: Engagement vs. misinformation. Working paper

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

Che YK, Horner J (2017) Recommender systems as incentives for social learning. Working paper

Crapis D, Ifrach B, Maglaras C, Scarsini M (2017) Monopoly pricing in the presence of social learning. Management Science 63(11):3586–3608

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

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

Ely JC (2017) Beeps. The American Economic Review 107(1):31–53

Feldman P, Papanastasiou Y, Segev E (2016) Social learning and the design of new experience goods. Management Science. Forthcoming

Frazier P, Kempe D, Kleinberg J, Kleinberg R (2014) Incentivizing exploration. In: Proceedings of the 15th ACM conference on Economics and Computation, ACM, pp 5–22

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. Management Science 56(4):591–605

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

Gittins J, Glazebrook K, Weber R (2011) Multi-armed bandit allocation indices. John Wiley & Sons
Halac M, Kartik N, Liu Q (2017) Contests for experimentation. J Political Economy 125(5):1523–1569
Hörner J, Skrzypacz A (2016) Learning, experimentation and information design. Survey prepared for the 2015 Econometric Summer Meetings in Montreal
Hu M, Wang L (2017) Simultaneous vs. sequential crowdsourcing contests. Working paper
Hu M, Shi M, Wu J (2013) Simultaneous vs. sequential group-buying mechanisms. Management Science 59(12):2805–2822
Huang Y, Vir Singh P, Srinivasan K (2014) Crowdsourcing new product ideas under consumer learning. Management Science 60(9):2138–2159
Jiang ZZ, Huang Y, Beil DR (2016) The role of feedback in dynamic crowdsourcing contests: A structural empirical analysis. Working paper
Kamenica E, Gentzkow M (2011) Bayesian persuasion. American Economic Review 101(6):2590–2615
Kanoria Y, Saban D (2017) Facilitating the search for partners on matching platforms: Restricting agent actions. Working paper
Keller G, Rady S, Cripps M (2005) Strategic experimentation with exponential bandits. Econometrica 73(1):39–68
Kleinberg RD, Slivkins A (2017) Tutorial: Incentivizing and coordinating exploration. In: Proceedings of the 18th ACM conference on Economics and Computation
Kornish LJ, Ulrich KT (2011) Opportunity spaces in innovation: Empirical analysis of large samples of ideas. Management Science 57(1):107–128
Körpeoğlu E, Cho SH (2017) Incentives in contests with heterogeneous solvers. Management Science. Forthcoming
Kremer I, Mansour Y, Perry M (2014) Implementing the "wisdom of the crowd." J Political Economy 122(5):988–1012
Li J, Netessine S (2017) Market thickness and matching (in) efficiency: Evidence from a quasi-experiment. Working paper
Lobel I, Sadler E (2015) Preferences, homophily, and social learning. Operations Research 64(3):564–584
Mansour Y, Slivkins A, Syrgkanis V, Wu ZSW (2015) Bayesian exploration: Incentivizing exploration in bayesian games. In: Proceedings of the 16th ACM conference on Economics and Computation, ACM, pp 565–582
Marinesi S, Girotra K, Netessine S (2017) The operational advantages of threshold discounting offers. Management Science. Forthcoming
Moon K, Bimpikis K, Mendelson H (2017) Randomized markdowns and online monitoring. Management Science. Forthcoming
Orlov D, Skrzypacz A, Zryumov P (2017) Persuading the principal to wait. Working paper
Papanastasiou Y (2017) Fake news propagation and detection: A sequential model. Working paper
Papanastasiou Y, Savva N (2017) Dynamic pricing in the presence of social learning and strategic consumers. Management Science 63(4):919–939
Papanastasiou Y, Bimpikis K, Savva N (2017) Crowdsourcing exploration. Management Science. Forthcoming
Rayo L, Segal I (2010) Optimal information disclosure. Journal of Political Economy 118(5):949–987
Renault J, Solan E, Vieille N (2017) Optimal dynamic information provision. Games and Economic Behavior 104:329–349
Seel C, Strack P (2016) Continuous time contests with private information. Mathematics of Operations Research 41(3):1093–1107
Strack P (2016) Risk-taking in contests: The impact of fund-manager compensation on investor welfare. Working paper
Swinney R (2011) Selling to strategic consumers when product value is uncertain: The value of matching supply and demand. Management Science 57(10):1737–1751

Taylor T (2016) On-demand service platforms. Manufacturing & Service Operations Management. Forthcoming

Terwiesch C, Xu Y (2008) Innovation contests, open innovation, and multiagent problem solving. Management Science 54(9):1529–1543

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