RAND Journal of Economics
Vol. 51, No. 2, Summer 2020
pp. 323–345

Incentive-compatible advertising on nonretail platforms

Kfir Eliaz*,**

and

Ran Spiegler*,***

Nonretail platforms enable users to engage in noncommercial activities, while generating user information that helps ad targeting. We present a model in which the platform chooses a personalized ad-display rule and an advertising fee (which depends on the targeted user group). The policy that maximizes the platform's advertising revenues creates an incentive for advertisers to strategize targeting. We provide a condition for incentive-compatibility of the first-best policy, and highlight the forces that make it harder to satisfy. We apply our result to examples of platforms. Our analysis of social networks turns out to be related to the "community-detection" problem.

1. Introduction

■ Recent years have seen a proliferation of online institutions that can be described as "nonretail platforms." Users of these platforms access them on a regular basis, in order to engage in activities such as reading texts, listening to music, exchanging messages, cultivating social links, etc. In particular, when they access the platform, it is *not* for the purpose of buying from advertisers. If a user buys from an advertiser as a result of being exposed to an ad posted on the platform, the transaction takes place off it; it will have no effect on his activity on the platform, and it is quite likely that the platform does not even monitor whether the transaction has taken place. However, the transaction may temporarily depress the user's demand for similar products, thus diminishing the effectiveness of advertising them.

Of course, nonretail platforms are at least as old as the village message board. What is special about the modern online version is that users' activity on the platform leaves a massive

* Tel-Aviv University; kfire@tauex.tau.ac.il, rani@post.tau.ac.il.
** University of Utah.
*** University College London.

This is a substantial revision of a paper formerly circulated (and promoted on YouTube by a movie trailer) under the title "Incentive-Compatible Advertising on a Social Network." Financial support from ISF grant no. 1153/13 is gratefully acknowledged. We thank Yair Antler, Francis Bloch, Olivier Compte, Andrea Galeotti, Nikita Roketskyi, Omer Tamuz, Neil Thakral, numerous seminar audiences, and the referees of this Journal, for helpful comments. We are especially grateful to the editor, David Myatt, for fruitful suggestions.

trail of information that may be correlated with their consumption tastes in various areas. As a result, the platform can help advertisers achieve better targeting, which in turn helps the platform increase its advertising revenues. Here are a few examples of what we have in mind.

Online radio stations like Pandora collect information about users' musical tastes (in this respect, they differ from traditional radio), and can use that to target ads for unrelated products. For instance, whether a user likes Country Music may be correlated with his politics and lifestyle preferences. But of course, he does not access Pandora for the purpose of being informed about political candidates or buying vegan food.

Email services may use the content of personal emails to target users. If a user's emails start featuring numerous references to babies, he may experience increased exposure to diaper ads on his email account, although buying diapers is obviously not the user's primary objective when checking his email.

Messaging platforms such as Whatsapp or Snapchat may be unable or unwilling to use the content that users generate, for technical or legal reasons. However, the structure of the social network among users may provide information about their types. For instance, if users exhibit *homophily*—that is, they associate with like-minded individuals—then a large cluster in the network indicates that its members are likely to have similar tastes.

Content sharing platforms such as Reddit are message boards that publish user-generated content, and may monitor the content that users produce or consume.

Many platforms exhibit combinations of these features. For instance, social media platforms like Instagram or Twitter can use the network structure of their users as well as the content that they generate. Although not all of these real-life examples of nonretail platforms currently use this form of targeted advertising, the potential to do so is inherent in them.[1]

In this article, we study novel incentive issues that arise in advertising on nonretail platforms. The source of the potential incentive problem is that advertisers have private information regarding the consumer-preference types they would like to target. The platform relies on their targeting requests to allocate display ads to individual users, utilizing its own private information about users. When advertising fees vary with the targeting request, it becomes a *strategic* decision that involves trading off the likelihood of a transaction against the fee. Indeed, real-life ad-tech intermediaries help advertisers cope with such trade-offs by searching for the target audience that gives the "best bang for the buck." This may involve diverting the client's ad to a less-than-ideal audience to save costs.[2]

Users' ad-generated (offline) purchases can affect their willingness to make subsequent purchases—for example, because they are temporarily satiated. However, this change in their consumption-driven behavior has no visible effect on their platform activity, which is not commerce-oriented to begin with. We will see that as a result of this feature, the platform may want to *diversify* the type of ads it shows to an individual user. Even if a Country Music fan is relatively unlikely to be interested in vegan food, exposing him to such ads every once in a while may increase the long-run expected number of transactions generated by such a user. The article's basic insight is that this diversification motive creates an incentive for advertisers to misrepresent their ideal targeting. Our aim is to understand the conditions in which this incentive problem prevents the platform from attaining its first-best.

In our model, there is a group of consumers with constant access to some nonretail platform. Each consumer comes in one of two (private) preference types. A type can describe whether the consumer is interested in "healthy food," whether he likes "highbrow" movies, whether he enjoys outdoor recreational activities, etc. The platform obtains a noisy aggregate signal about

[1] Search engines are an example of hybrid platforms with both retail and nonretail features. Consumers use search engines to specifically look for a product or service to buy, but they also use them to find information unrelated to any transaction. We do not address such platforms in this article.

[2] For example, AdEspresso.com is a company that offers to help small businesses launch advertising campaigns on social media. On their website they wrote, "The audience you choose will directly affect how much you're paying…if your perfect audience is just more expensive, that's just the way it goes."

the profile of consumers' types. It then enables advertisers to post personalized display ads. Each advertiser is characterized by the quality of its match with each consumer type—defined as the probability of transaction conditional on the consumer's exposure to the firm's ad. This is the advertiser's private information (in the main version of our model, advertisers receive no additional information about consumers). *Ex ante*, each advertiser communicates to the platform the type of consumers it wishes to target. Thus, ads are classified into "types" according to the targeting request that accompanies them. Advertiser-platform communication of this kind exists in reality. For instance, Pandora offers ad targeting based (among other things) on users' listening habits.

We assume that consumers' exposure to ads is governed by a personalized, stationary display rule that the platform designs. Specifically, the platform tailors a mixture of ad types to each consumer—based on its posterior belief regarding his type (derived from its signal)—such that the ad he is exposed to at any period is drawn independently according to this mixture. Ads are like "billboards" and transactions occur offline, unmonitored by the platform. As soon as the consumer transacts with an advertiser, he switches to a "satiation" mental state in which he is inattentive to ads, and he switches back to the attentive state of mind with some constant per-period probability that captures the propensity for repeat purchases. Thus, thanks to the simplifying assumption of stationary display rules, we can depict the consumer's experience at the platform as a personal two-state Markov process, where certain transition probabilities are determined by the platform's personalized display rule.

The platform's objective is to maximize total advertisers' surplus—defined as their long-run number of transactions per period, and calculated according to consumers' personal Markov processes—and to extract it by means of advertising fees. Because the platform is uncertain about consumers' types and their mental state at any given period, its optimal display rule may be *interior*—that is, it may expose individual consumers to *both* ad types. As mentioned above, this turns out to generate a motive for advertisers to *strategize* their targeting request.

Our first observation is that optimal display rules approximately minimize the amount of time that it takes a nonsatiated consumer to transact. As a result, the optimal probability that an advertiser is displayed to a particular consumer is approximately proportional to the *square root* of the platform's posterior probability that the firm's product fits the consumer's type. In contrast, the advertising fee that fully extracts an advertiser's surplus is proportional to the *prior probability* that consumers like its product. This discrepancy ends up discriminating against products with mass appeal, and it may give advertisers an incentive to target the minority consumer group (if the reduced exposure is more than compensated for by the reduced fee).

Under what conditions on the environment's primitives can the platform design an incentive-compatible (IC) policy (consisting of a display rule and an advertising-fee schedule) that maximizes and fully extracts advertisers' surplus? Our interest in this question is twofold. First, it serves as a useful theoretical benchmark for the platform's design problem. Second, and perhaps more interestingly, it can be interpreted in the spirit of "welfare theorems" in the competitive-equilibrium literature. We can regard the nonretail platform as a market institution for allocating consumers' limited attention (the scarce resource in this environment) to advertisers. The full-surplus-extraction requirement is a zero-profit condition that captures competitive behavior among advertisers. Our question then becomes: *Can an efficient allocation of platform users' attention to advertisers be supported by a competitive market*? We do *not* study "second-best" policies when the first-best is not implementable: this is a challenging problem that requires different analytical techniques and belongs to a different article.

Our basic result is a necessary and sufficient condition for the implementability of the platform's objective (assuming that exogenous parameters are such that the optimal display rule is interior—otherwise, our condition is merely sufficient). The condition is an inequality that incorporates two quantities: (i) on the LHS, a measure of the platform's *uncertainty* about consumers' types (or how *uninformative* its signal is); and (ii) on the RHS, a simple expression that involves the characteristics of the consumer population. We illustrate this result in a simple example of a content platform where the kind of content a user consumes is a noisy signal of his type.

The chief merit of this characterization is that it isolates the platform-specific details and summarizes them by the LHS's measure of uncertainty. The inequality's RHS summarizes the consumers' features that are independent of the platform. Therefore, examining various types of platforms is reduced to studying their induced uncertainty measure. Another virtue of the inequality is that it makes comparative statics quite transparent. The inequality is *easier to satisfy* when the signal becomes *more informative*, when consumers are *less attentive* to ads, when the gap between high- and low-quality match probabilities is *smaller*, and when repeat purchases are *more* frequent.

However, it is important to emphasize that in the main applications we examine in this article, the two sides of the inequality are *interdependent*—that is, the uncertainty inherent in the signal varies with the consumer-type distribution. The reason is that the platform brings together multiple consumers, such that the platform's signal about an individual consumer depends on the types of other platform users he interacts with. Indeed, in Section 4 we apply our characterization result to such "social interaction platforms." We begin with a simple "warm-up" example of an email platform, which gets a noisy signal about a pair of users' types through the content of their email exchange.

Our main application examines a social network, where the only information that is available to the platform is the network structure. This structure has informational value because the probability of a link between two users is a function of their types. We show that in this context, the first-best is not implementable if the consumer-type distribution is either too asymmetric or too uniform. We then ask whether a larger network makes it easier for the platform to implement its objective. Following the Network Science literature on *community detection*, we assume that users' propensity to form links decreases with network size, such that the expected degree of an individual node grows only *logarithmically* in n. Applying a recent result on the community-detection problem (Abbe and Sandon, 2015), we obtain a sufficient condition for the implementability of the platform's objective for large n in terms of parameters of the network-generating process. Thus, our analysis uncovers a connection between the community-detection problem in Network Science and the economic question of incentivizing targeted advertising on social networks.

□ **Related literature.** This article belongs to a research agenda that explores novel incentive issues in modern platforms. Our earlier exercise in this vein, Eliaz and Spiegler (2015), studied an environment in which consumers submit noisy queries to a "search platform," which responds by providing consumers with a "search pool"—that is, a collection of products that they can browse via some search process. The platform's problem is to design a decentralized mechanism for efficiently allocating advertisers into search pools and extracting their surplus. Thus, unlike nonretail platforms, the search platform's *sole* function is to match users with advertisers. Eliaz and Spiegler (2015) borrowed the Bhattacharyya Coefficient (a measure of similarity between probability distributions) from the Statistics and Machine Learning literature, and demonstrated its use in representing IC constraints. The present article further demonstrates the power of this tool in a different context, and with new technical challenges that arise from the applications (e.g., the community-detection problem in social networks).

There has been a growing interest in targeted advertising in the industrial organization (IO) literature. One strand of this literature analyzes competition between advertising firms that choose advertising intensity, taking into account the cost of advertising and the probability that their advertising messages will reach the targeted consumers. Notable articles in this literature include Iyer, Soberman, and Villas-Boas (2005), Athey and Gans (2010), Bergemann and Bonatti (2011), Zubcsek and Sarvary (2011), and Johnson (2013). A second strand of this literature studies how to optimally propagate information about a new product by targeting specific individuals in a social network. Recent articles in this strand include Galeotti and Goyal (2012) and

Campbell (2015) (see Bloch, 2016 for a survey). In these articles, consumers are targeted according to their network *location*, whereas in the present article targeting is based on the platform's information regarding their *preference types*.

Bergemann and Bonatti (2015) study a different aspect of using information about consumers for advertising purposes. In their model, a single data provider offers firms information about the potential value of matches with various consumers, where the information is obtained from consumers' online activities.

2. A Model

■ We begin with a single-consumer environment (an extension to multiple consumers is given in Section 4). Let $T = \{x, y\}$ be a set of possible consumer types, and let $\pi \geq \frac{1}{2}$ be the probability that a consumer is of type x. We consider a stationary environment in which at any time period, a consumer of type $t \in T$ is in one of two mental states: (i) a "demand state" D_t in which the consumer buys a product with positive probability when he is exposed to an ad for it (we describe below the ad-display and purchase processes), and (ii) a "satiation state" S_t in which the consumer is uninterested in consumption. The consumer's transition between states obeys the following mechanical rule. He switches from his demand state to his satiation state as soon as he buys a product. When the consumer is in his satiation state, he switches back to his demand state with independent per-period probability ε. This parameter captures consumers' propensity for repeat purchases.

Products are offered by advertisers and come in two types, also labelled x and y. We say that an advertiser is of type t if it offers a type t product. Our interpretation of this typology is as follows. We envision the consumer as a vector of unobservable personality attributes. The possible vectors are partitioned into two groups, x and y, such that every product that is offered in the market is more appealing to one of the two groups. Thus, advertisers of type t offer a *variety* of products, which all share the feature that they are more appealing to a consumer of type t. The probability that the consumer buys a product conditional on being exposed to its ad while in his demand state is θ_H (θ_L) when the product's type matches (differs from) his own type, where $\theta_L < \theta_H$. The parameters θ_L, θ_H also reflect the consumer's general attention to ads: raising both by a common factor captures greater attentiveness.

The consumer has constant, uninterrupted access to a nonretail platform. Thanks to this access, the platform obtains a signal $w \in W$ regarding his type, where W is some finite set with $|W| > 1$. This signal is received *ex ante*, once and for all, before the above stochastic process that describes consumer behavior begins. Upon observing the signal realization the platform forms a posterior belief ϕ about the likelihood that the consumer is of type x. The exact specification of the signal distribution will play a role in later sections.

How does the platform match the consumer with advertisers, given its belief regarding the consumer's type? At every time period, the platform selects an ad type according to a stationary random process we will describe momentarily and displays the ad to the consumer. Each ad expires at the end of the period and a new one is displayed in the next period. We think of ads as "billboards" : transactions between consumers and advertisers take place " offline" and the platform cannot monitor them. In particular, there is no notion of "clicking" on display ads.

The display of ads is governed by a stationary rule that the platform commits to *ex ante*. (In an environment with multiple consumers, the ad would be personalized—that is, there will be a distinct ad-display process for each consumer.) Formally, given its posterior belief, q is the probability that at any time period, the platform displays an advertiser of type x (where the probability of displaying an advertiser of type y is $1 - q$). We refer to q as the platform's *display rule*.

Given that the consumer cycles between his two mental states indefinitely, the assumption of stationary display rules means that the consumer's behavior over time obeys a two-state Markov process. Specifically, the transition probabilities between the mental states of a consumer of type x given q are given by the following matrix:

$$\begin{array}{ccc} & D_x & S_x \\ D_x & 1-[\theta_H q+\theta_L(1-q)] & \theta_H q+\theta_L(1-q), \\ S_x & \varepsilon & 1-\varepsilon \end{array} \tag{1}$$

where the matrix for type y is derived by replacing q with $1-q$. It follows that given the platform's posterior belief ϕ and display rule q, the joint invariant probability that the consumer is of type x and in state D_x is

$$\rho^x \equiv \frac{\phi\varepsilon}{\theta_H q+\theta_L(1-q)+\varepsilon} \tag{2}$$

(where ρ^y is given by replacing ϕ and q with $1-\phi$ and $1-q$ in the above expression). As will be shown shortly, this probability allows us to obtain a simple expression for the long-run average number of transactions for each advertiser.

□ **Efficient ad display.** The nonretail platform creates value by facilitating consumer-advertiser matches that potentially lead to transactions. In this subsection, we characterize the display rules that maximize the expected per-period number of transactions, which is given by the following expression:

$$q[\rho^x\theta_H+\rho^y\theta_L]+(1-q)[\rho^y\theta_H+\rho^x\theta_L], \tag{3}$$

where ρ^x is given by (2).

Let $\bar{q}$ be the value of q that maximizes (3). When it is interior, first-order conditions imply

$$\frac{\rho^x}{\rho^y}=\sqrt{\frac{\phi}{1-\phi}}. \tag{4}$$

The explicit solution is

$$\bar{q}=\frac{(\theta_H+\varepsilon)\sqrt{\phi}-(\theta_L+\varepsilon)\sqrt{1-\phi}}{(\theta_H-\theta_L)(\sqrt{\phi}+\sqrt{1-\phi})}. \tag{5}$$

Then,

$$\lim_{\substack{\varepsilon\to 0\\ \theta_L\to 0}}\bar{q}=\frac{\sqrt{\phi}}{\sqrt{\phi}+\sqrt{1-\phi}}. \tag{6}$$

Henceforth, we assume that the primitives are such that the solution is interior for all signal realizations. Equivalently,

$$\frac{\theta_L+\varepsilon}{\theta_H+\varepsilon}<\sqrt{\frac{\phi}{1-\phi}}<\frac{\theta_H+\varepsilon}{\theta_L+\varepsilon}. \tag{7}$$

This condition is easier to satisfy when θ_L and ε decrease (reflecting low propensity for repeat purchases and a large effect of product match on consumers' attentiveness) and when the platform's posterior beliefs are relatively uniform (this holds when the platform's prior is not too asymmetric and its signal is not too informative).

Consumer satiation is crucial for interior display rules. Suppose that consumers experienced no satiation at all, such that their behavior would be described by a single-state process in which they demand their favorite product at every period. Then, the platform's optimal display rule would be a corner solution: if $\phi > \frac{1}{2}$, it would display type x ads with probability 1 at every period (and type y if $\phi < \frac{1}{2}$). This case is partially approached in our two-state model when $\varepsilon = 1$, such that the consumer's satiation lasts exactly one period. In this case, condition (7) holds for the smallest set of primitives $(\varepsilon, \theta_L, \theta_H)$.

Thus, our two-state Markov process is the simplest formalization of the aspect of consumer behavior (namely, satiation) that gives rise to interior optimal display probabilities. These essentially constitute an *interior allocation of consumer attention*. Rather than allocating it entirely to one of the two advertiser type, the platform prefers to randomize between them. The property that efficient allocation of consumer attention is interior turns out to create the incentive-compatibility problem that will be the subject of Section 3.

The platform's uncertainty and ex ante *surplus.* The platform's uncertainty about the consumer's type will play an important role in subsequent sections. To define a measure of the platform's uncertainty we need to introduce the following notation. Recall that the platform observes a signal realization w from a set W of possible realizations. Letting $\mu \in \Delta(T \times W)$ denote the prior joint distribution over the consumer's type and the platform's signal, the marginal over the consumer's type is $\sum_w \mu(x, w) = \pi$, and the platform's posterior belief that the consumer is of type x, conditional on receiving the signal w, is $\mu(t = x \mid w) = \phi_w$. The platform's *ex ante* (before the signal w is realized) expected uncertainty is defined by the following quantity:

$$U = \sum_{w \in W} \mu(w)\sqrt{\mu(x \mid w)\mu(y \mid w)} = \mathbb{E}\left[\sqrt{\phi_w(1 - \phi_w)}\right].$$

Note that U takes values in $[0, \frac{1}{2}]$ and that it is the expectation of a concave function of ϕ_w. Recall the property of Bayesian posteriors,

$$\pi = \sum_{w \in W} \mu(w)\phi_w,$$

and consider a change in μ that keeps π fixed and makes the signal w *more informative* in Blackwell's sense. This is equivalent to introducing a *mean-preserving spread* to the distribution over ϕ_w. Jensen's inequality then implies that U *decreases*. Therefore, U may be viewed as a measure of the platform's average *uncertainty regarding* the consumer's type.[3]

This measure allows us to obtain a simple characterization of the maximal *ex ante* expected surplus that the platform can generate. Let q_w^t denote the probability that at any time period, the platform displays an advertiser of type t conditional on the signal w, and let ρ_w^t denote the joint invariant probability that the consumer is of type t and in state D_t. Expression (3) represents the expected per-period number of transactions, according to the platform's posterior belief for a given signal realization. The *ex ante* surplus that the platform generates is obtained by taking an expectation with respect to w:

$$\sum_{w \in W} \mu(w)\left\{q_w^x\left[\rho_w^x\theta_H + \rho_w^y\theta_L\right] + q_w^y\left[\rho_w^y\theta_H + \rho_w^x\theta_L\right]\right\}. \tag{8}$$

Under the condition (7), we can plug (5) into (8) and obtain the following expression for the maximal surplus:

$$\varepsilon\left[1 - \frac{\varepsilon(1 + 2U)}{\theta_H + \theta_L + 2\varepsilon}\right]. \tag{9}$$

This expression *decreases* with U. The intuition is that a more informative signal facilitates effective targeting and therefore increases the average number of transactions per period.

Discussion: the stationarity assumption. Stationarity in our model has exogenous and endogenous aspects. The former arises from our assumption that ads are "billboards"; the platform cannot monitor whether the consumer pays attention to ads and whether he transacts with advertisers (e.g., he may be listening to Pandora while jogging). Therefore, it cannot learn anything about consumers beyond the signal w.

[3] We thank the editor, David Myatt, for suggesting this measure.

Even so, stationary display rules (which constitute the endogenous aspect) carry a loss of generality. If we relaxed this assumption and allowed the platform's display rule to follow some Markov process with an arbitrary number of states K, the consumer's behavior over time would obey a $2K$ -state Markov process. The optimal display rule would involve switching between x and y ads, but the precise transition rule—and therefore, the long-run number of transactions—would be difficult to characterize. We believe that the qualitative insights of our stationary model would not change, as long as we assume that advertisers do not know the initial state of the Markov process—they would still treat the allocation of ad slots at any given period as a random variable, albeit one whose distribution is difficult to calculate. The stationarity assumption is thus a simplifying approximation that enables us to tractably capture the platform's motive to diversify its ad types over time. However, the quality of this approximation will vary with $\theta_L, \theta_H, \varepsilon$, and this means that our comparative statics with respect to these parameters should be taken with a grain of salt.

Suppose that ads are not billboards, such that the platform can partially monitor whether consumers notice them—for example, through *clicks* (and let us retain the rather realistic assumption that the platform does not monitor transactions). If clicks are uncorrelated with consumers' types, then exogenous stationarity continues to hold, and pricing displays is equivalent to pricing clicks. Therefore, we can think of our stationarity and pricing assumptions as reasonable approximations to situations in which clicks convey little information about consumer types.

3. Incentive Compatibility

■ So far, we have examined the platform's ad-display policy, without taking into account its interaction with advertisers. In this section we address this side of the market. Assume there is a large, equal number of advertisers of each of the two types. For notational simplicity, we normalize the total mass of advertisers of each type to one. Each advertiser can costlessly supply any amount of its product. If the consumer acquires a product from an advertiser, the advertiser earns a fixed payoff of 1. We completely abstract from product prices.[4] The platform does not receive any information about the types of individual advertisers, and advertisers receive no information about the consumer's type (we relax this assumption in the second subsection of Section 5). Conditional on displaying an ad of type t, each of the type-t advertisers is drawn uniformly. The platform generates revenues from advertising fees. Let F_t be the per-period fee that the platform charges advertisers of type t.

(Our analysis would remain unchanged if we assumed that instead of charging a per-period fee, the platform charges a *price per display*. Likewise, we could allow prices to be a function of the platform's signal w, without any effect on our analysis. This is because advertisers in our model are risk-neutral and care only about the expected number of transactions and the expected payment. It is therefore convenient analytically to assume lump-sum transfers, even if this may appear unrealistic when taken literally.)

Denote $F = (F_x, F_y)$. The pair (q, F) constitutes the platform's *policy*. The platform's objective is to find a policy (q, F) that maximizes expected profits. Its first-best would be to choose q to maximize total surplus (as given in the first subsection of Section 2) and fully extract this surplus by means of F. Specifically, let $R_x(q)$ and $R_y(q)$ denote the gross expected per-period revenue of advertisers of types x and y, respectively, given the display rule q. Then,

$$R_x(q) \equiv \sum_{w \in W} \mu(w) q_w^x \left[\rho_w^x \theta_H + \rho_w^y \theta_L\right] \tag{10}$$

$$R_y(q) \equiv \sum_{w \in W} \mu(w) q_w^y \left[\rho_w^y \theta_H + \rho_w^x \theta_L\right]$$

[4] Allowing profit margins or the number of advertisers to vary across types would be equivalent to modifying π. For models that study the effect of platform features on product prices, see Eliaz and Spiegler (2011), Candogan, Bimpikis, and Ozdaglar (2012), and Fainmesser and Galeotti (2015).

such that maximal total surplus is $R_x(\bar{q}) + R_y(\bar{q})$, where $\bar{q}$ is the efficient display rule. The optimal fee that the platform charges advertisers of type t, denoted as $\bar{F}_t$, fully extracts the maximal surplus of these advertisers—that is, $\bar{F}_t = R_t(\bar{q})$.

The pair $(\bar{q}, \bar{F})$ is thus the platform's optimal policy. In order to implement it, however, the platform needs to know advertisers' types. But now suppose that the platform is *unable* to directly verify this information. Therefore, it relies on advertisers' self-reports—which we interpret as requests to target specific preference groups. The reports are submitted *ex ante*, once and for all—that is, we do not allow for dynamic reporting, in line with our restriction to stationary environments.[5]

One could argue that the platform need not rely on advertisers' targeting requests—in principle, it could examine each advertiser's product and figure out the quality of its match with each consumer type. However, this ad-classification task is costly, and the platform can avoid the cost by decentralizing the task. Furthermore, in many cases advertisers have private information regarding the type of consumers who are attracted to their product, thanks to prior market research. For instance, certain food items (granola bars, artificially sweetened products) are not easy to classify *a priori* in terms of their appeal to "health-conscious" consumers. Likewise, the defining lines of "highbrow" movies or holiday packages that fit "outdoorsy" tourists are quite blurred. In these cases, market studies are likely to reveal information that the platform lacks. It is implausible for the platform to replicate such studies in the myriad industries it interacts with.

A policy (q, F) is IC if no single advertiser has an incentive to misreport its type, given that every other advertiser reports truthfully. In principle, the deviation changes the reported number of advertisers of each type as well as the transition probabilities that define the consumer's Markov process. However, we assume that each advertiser neglects these complications when evaluating the deviation. In other words, advertisers behave *competitively* in the sense that they take the joint invariant probabilities ρ_w^t as given. In Appendix B, we demonstrate the validity of this approximation when the number of advertisers is large.

Thus, given (q, F), an x advertiser weakly prefers to report its type truthfully if and only if

$$\sum_{w\in W} \mu(w) q_w^x \left[\theta_H \rho_w^x + \theta_L \rho_w^y\right] - F_x \geq \sum_{w\in W} \mu(w) q_w^y \left[\theta_H \rho_w^x + \theta_L \rho_w^y\right] - F_y. \tag{11}$$

Likewise, a y advertiser weakly prefers to report its type truthfully if and only if

$$\sum_{w\in W} \mu(w) q_w^y \left[\theta_H \rho_w^y + \theta_L \rho_w^x\right] - F_y \geq \sum_{w\in W} \mu(w) q_w^x \left[\theta_H \rho_w^y + \theta_L \rho_w^x\right] - F_x. \tag{12}$$

When (q, F) satisfies these two inequalities, we say it is IC. When the optimal policy $(\bar{q}, \bar{F})$ satisfies them, we say that it is *implementable*.

When F fully extracts advertisers' surplus (which is the case under the optimal policy), the LHS of (11) and (12) is 0. The inequalities thus reduce to

$$\sum_{w\in W} \mu(w) q_w^y \left[\rho_w^x - \rho_w^y\right] \leq 0 \tag{13}$$

$$\sum_{w\in W} \mu(w) q_w^x \left[\rho_w^y - \rho_w^x\right] \leq 0.$$

Plugging the solution for $\bar{q}$ from the previous subsection and performing a bit of algebra, we obtain a simple necessary and sufficient condition for implementability of the optimal policy. To express this condition in a compact form, denote

$$\lambda = \frac{\theta_H + \varepsilon}{\theta_H + \theta_L + 2\varepsilon} \in \left(\frac{1}{2}, 1\right). \tag{14}$$

[5] The platform could design a general mechanism that severely punishes all advertisers if the number of x reports differs from the number of x advertisers. This would make truthful reporting IC. However, it implausibly relies on exact knowledge of the number of x advertisers.

Proposition 1. Suppose that $\bar{q}_w$ is interior for every w. Then, $(\bar{q}, \bar{F})$ is implementable if and only if

$$U \leq (1-\lambda)\pi + \lambda(1-\pi). \tag{15}$$

Recalling the analogy to "welfare theorems" described in the Introduction, Proposition 1 can be viewed as a characterization of environments in which competitive markets can sustain an efficient allocation of consumers' scarce attention to advertisers. This result highlights two factors. As the platform becomes *less uncertain* on average about the consumer's type (where the average uncertainty is measured by U) and as the consumer-type distribution (given by π) becomes *more symmetric*, condition (15) becomes *easier* to satisfy (when $\pi = \frac{1}{2}$, the condition necessarily holds because $U \leq \frac{1}{2}$ by definition).

It should be emphasized, however, that in all applications we will examine in Section 4, these two factors are *not* independent—that is, a change in π will also affect the platform's uncertainty, an interdependence that will have nontrivial effects. Still, distinguishing between these two factors is helpful for understanding Proposition 1—just as the supply-equals-demand representation of competitive equilibrium is useful even when both supply and demand are affected by the same exogenous shock.

Consider the $\lambda \to 1$ limit, where a consumer rarely buys a product from a poor-match advertiser and where the propensity for repeat purchases is low. Condition (15) simplifies into

$$U \leq 1 - \pi. \tag{16}$$

To see the intuition behind this inequality, recall that the optimal display probability $\bar{q}^t_w$ in the $\lambda \to 1$ limit is proportional to the *square root* of $\mu(t \mid w)$. Thus, although a product with high $\mu(t \mid w)$ gets an advantage in terms of display probability, the square root factor *softens* this advantage. By comparison, the fee paid by an advertiser that submits the report t, is proportional to $\mu(t)$. This difference introduces an incentive for x advertisers to target the minority group of y consumers, as the reduced exposure to their ideal audience may be outweighed by the reduced fee. When the consumer-type distribution becomes more symmetric (such that $1 - \pi$ goes up), the gap between the fees paid by advertisers of different types shrinks, mitigating the misreporting incentive. As the signal becomes more informative and reduces the platform's uncertainty, the values of $\mu(t \mid w)$ get closer to 0 or 1, such that the "square root effect" vanishes, again mitigating the misreporting incentive. Finally, recall that the platform conditions the display probabilities on w, whereas advertisers are uninformed of w at the time they submit their reports. When the signal is highly informative, an advertiser that chooses to misreport knows it will be displayed with high (low) probability to consumers with low (high) probability of transacting with it, and this is another force that mitigates the misreporting incentive.

Let us now move away from the $\lambda \to 1$ regime. The probability ε of exiting the satiation state and the match-quality parameters θ_L and θ_H determine the weights of π and $1 - \pi$ in the RHS of (15). Because $\lambda \in (\frac{1}{2}, 1)$ and $\pi \geq \frac{1}{2}$, a decrease in λ relaxes the inequality (15) and makes the optimal policy easier to implement. More specifically, as consumers become *more attentive* to ads (in the sense that θ_L and θ_H increase by the same factor), and as the propensity for repeat purchases *declines*, the condition for implementing the optimal policy becomes *harder* to meet.

The following example illustrates Proposition 1.

□ **An example: content platforms.** In a prevalent type of nonretail platforms, users upload and consume content. A common feature of these platforms is that the content that a user consumes reflects not only his personal taste but also the *availability* of various types of content. In particular, a user may fail to consume his ideal content if it is scarce or not prominent. The following is a stylized example of advertising on a content platform in which users are purely consumers of content.

A content consumer faces a supply of m content items. The type of each item is x (y) with independent probability α $(1-\alpha)$. The consumer always consumes exactly one of the available items. The consumed item fails to match his type only when none of the available items do. Hence, the *only* case in which the consumer fails to consume his ideal type of content is when all m items are of the other type.

In this setup, the distribution over the platform's posterior belief is easy to compute. Whenever the signal realization w is such that both content types are available, the consumer's behavior fully reveals his type, such that $\phi_w = 1$ ($\phi_w = 0$) if he consumes an item of type x (y). For such realizations of w, we have $\phi_w(1-\phi_w) = 0$. When the signal realization w is such that all content items are of the same type, an event whose probability is $\alpha^m + (1-\alpha)^m$, the consumer's behavior is completely uninformative of his type, such that $\phi_w = \pi$. For such realizations of w, we have $\phi_w(1-\phi_w) = \pi(1-\pi)$. It follows that

$$U = [\alpha^m + (1-\alpha)^m] \cdot \sqrt{\pi(1-\pi)}.$$

Applying Proposition 1, the optimal policy is implementable if and only if

$$[\alpha^m + (1-\alpha)^m] \leq (1-\lambda)\sqrt{\frac{\pi}{1-\pi}} + \lambda\sqrt{\frac{1-\pi}{\pi}}.$$

The LHS decreases as m rises and α gets closer to $\frac{1}{2}$. This captures the intuition that a large supply of content that is drawn from a heterogeneous population of producers increases the chances that consumers' observed content consumption will reveal their tastes, and therefore enhances the informativeness of the platform's signal.

4. Social-Interaction Platforms

■ Our analysis thus far has focused on the case of a single, isolated consumer. Yet, nonretail platforms typically operate by bringing together multiple consumers, whose *social interaction* reveals information about their individual characteristics. Friendship patterns in a social network or the content of an email exchange are cases in point. In order to capture this feature of nonretail platforms in our applications, we now assume that there are *multiple consumers* and that the platform obtains an *aggregate* signal about their individual types. From a technical point of view, what makes this extension interesting is that the informativeness of the platform's signal intrinsically depends on the distribution of consumer types.

This extension requires a minor adaptation of our formalism. The set of consumers is $\{1, \ldots, n\}$ and $t_i \in T_i = \{x, y\}$ denotes the type of consumer i. Denote $T = T_1 \times \cdots \times T_n$. The prior distribution μ is defined over $T \times W$. Consumers' types are i.i.d., where $\pi \geq \frac{1}{2}$ continues to denote the prior probability that $t_i = x$. The conditional signal distribution $\mu(w \mid t)$ treats consumers symmetrically: for every permutation $f : T \to T$ there exists a permutation $g : W \to W$ such that $\mu(w \mid t) = \mu(g(w) \mid f(t))$. As a result, the distribution over the platform's posterior belief regarding the type of any consumer is the same for all consumers.

Thanks to the latter observation, the analysis in Sections 2 and 3 is easily adapted to this multiconsumer setting. First, we replace $\mu(t \mid w)$ with $\mu(t_i \mid w)$, which represents the probability that consumer i 's type is t_i conditional on the aggregate signal w. Accordingly, we replace the notation ϕ_w with $\phi_{i,w}$, which represents the probability that consumer i's type is x conditional on w. Second, we replace the notation q_w^t with the notation $q_{i,w}^t$, which represents the probability that an ad of type t is displayed at any period to consumer i given the aggregate signal w. Once these adjustments are made, the optimal display rule is given by (5) for all consumers; expression (9) describes the platform's first-best payoff *per consumer*; and most importantly for the purposes of this section, the implementability condition (15) is completely unchanged, once the definition of U replaces $\mu(t \mid w)$ with $\mu(t_i \mid w)$.

□ **The Bhattacharyya Coefficient.** Another minor adaptation that is useful for the examples in this section concerns the measure of the platform's uncertainty. The fact that U is defined in terms of the distribution over the platform's posterior belief (regarding the type of any individual consumer) makes it cumbersome in applications because it requires us to compute this distribution. Fortunately, we can slightly rewrite this measure such that it is defined entirely in terms of the conditional distribution $\mu(w \mid t_i)$, which is more straightforward to derive from primitives.

Define

$$S \equiv \sum_{w \in W} \sqrt{\mu(w \mid t_i = x)\mu(w \mid t_i = y)}. \tag{17}$$

The symmetry of μ with respect to consumers' labels implies that S is the same for all consumers i.

In the Statistics and Machine Learning literature, S is known as the *Bhattacharyya Coefficient* that characterizes the distributions $\mu(\cdot \mid x)$ and $\mu(\cdot \mid y)$.[6] It is a measure of *similarity* between these two conditional distributions. From a geometric point of view, this is an appropriate similarity measure because S is the direction cosine between two unit vectors in $\mathbb{R}^{|W|}$, $(\sqrt{\mu(w \mid x)})_{w \in W}$ and $(\sqrt{\mu(w \mid y)})_{w \in W}$. The value of S increases as the angle between these two vectors shrinks; $S = 1$ if the two vectors coincide; and $S = 0$ if they are orthogonal.

More importantly, S is a measure of the uncertainty faced by the platform. The stochastic matrix $(\mu(\cdot \mid t_i))_{t_i \in \{x,y\}}$ can be viewed as an information system in Blackwell's sense. Eliaz and Spiegler (2015) established that S decreases with the Blackwell informativeness of this information system. Indeed, elementary algebra establishes an immediate connection between S and our earlier measure of signal informativeness:

$$U = S\sqrt{\pi(1-\pi)}.$$

This enables us to restate the condition for the implementability of the platform's optimal policy:

$$S \leq (1-\lambda)\sqrt{\frac{\pi}{1-\pi}} + \lambda\sqrt{\frac{1-\pi}{\pi}}. \tag{18}$$

This is the condition that will serve us in the remainder of this section. The proofs of the results in this section (as well as the extensions in Section 5) use attractive features of the measure S.

Throughout the section, we restrict attention to the large λ regime—that is, the case of low propensity for repeat purchases and low consumer attention to low-match products. This simplifies expressions without changing the substance of our analysis.

□ **Targeting based on email content.** Consider two users of an email service. A user of type x has a baby whereas a user of type y does not. This may affect their preferences over a wide range of products. For example, type x is more likely to be interested in home entertainment and diapers, whereas type y is more likely to be interested in concerts or alcohol. The platform that operates the email service monitors users' *sent mail* folder. The aggregate signal w is a pair $w = (w_1, w_2)$, where $w_i = x$ indicates that the users' sent emails contain references to babies and $w_i = y$ indicates that they do not. The two users are friends who exchange messages with each other. The platform's algorithm cannot tell whether a user refers to his own baby or his friend's baby. Moreover, assume that a user of type x always sends emails that mention his baby, whereas

[6] See Basu, Shioya, and Park (2011) and Theodoris and Koutroumbas (2008). A related concept is the *Hellinger distance* between distributions, given by $H^2(x, y) = 1 - \sqrt{S(x, y)}$.

a user of type y mentions babies with positive probability $\xi \in (0, 1)$ *only* when his friend has one. The conditional signal distribution $\mu((w_1, w_2) \mid (t_1, t_2))$ is given by the following table:

$(t_1, t_2)\backslash(w_1, w_2)$	x, x	x, y	y, x	y, y
x, x	1	0	0	0
x, y	ξ	$1-\xi$	0	0
y, x	ξ	0	$1-\xi$	0
y, y	0	0	0	1

Let us now derive a condition for implementability of the first-best in this example. Consider the distribution over aggregate signals conditional on user 1's type. First, the signals (y, y) and (y, x) are impossible when $t_1 = x$ because by assumption, such a user type always sends an email with reference to babies. Second, the signal (x, y) is impossible when $t_1 = y$, because this babyless type can refer only to babies if his friend has a baby, which would then imply $w_2 = x$, a contradiction. It follows that the only signal realization that contributes a nonzero term to the Bhattacharyya Coefficient is $w = (x, x)$. Therefore,

$$S = \sqrt{\mu((x, x) \mid t_1 = x)\mu((x, x) \mid t_1 = y)} = \sqrt{(\pi + (1 - \pi)\xi) \cdot \pi\xi}.$$

Condition (18) in the $\lambda \to 1$ limit is thus

$$\sqrt{(\pi + (1 - \pi)\xi) \cdot \pi\xi} \leq \sqrt{\frac{1 - \pi}{\pi}}. \tag{19}$$

Note that although π appears on both sides of this inequality, it represents different things: The RHS involves the prior belief over a given user's type, whereas the LHS uses the prior over his *correspondent*'s type. Because the two users are *ex ante* identical, the same value of π features in the two sides of (18). Thus, a change in the type distribution (given by π) has *two* effects: On one hand, it impacts the platform's uncertainty about the correspondent's type via the Bhattacharyya Coefficient S; on the other hand, it affects the maximal uncertainty threshold for implementability (i.e., the RHS of (18)) The reason is that the platform's information about an individual user depends on his platform-mediated interaction with the other user, which in turn depends on his type. It follows that the platform's optimal policy is implementable in the $\lambda \to 1$ limit if and only if

$$\pi^3\xi(1 - \xi) + \pi^2\xi + \pi \leq 1.$$

When π is sufficiently close to $\frac{1}{2}$, the condition holds for all ξ. But for every sufficiently high π, there exists $\xi^*(\pi)$ such that the condition fails for all $\xi > \xi^*(\pi)$. Thus, a more symmetric distribution of consumer types is unambiguously better for implementability of the first-best. As we will later see, this is *not* true in general. Additionally, as babyless users become less likely to mention babies in their emails, the platform's signal gets more informative, and the condition for implementability becomes easier to satisfy.

□ **Social networks.** We now turn to our main application in this article, where the platform operates a social network—that is, it enables consumers to form social links with each other. Whether a pair of consumers is linked depends stochastically on their types. The network structure does not evolve over time. We focus entirely on the informational content of the network structure itself, and ignore other aspects of the users' activity on the network that may generate valuable information for advertisers. This will enable us to establish a theoretical connection with an interesting question in the Network Science literature.

Formally, a social network is a random nondirected graph in which consumers are nodes. The set W can thus be redefined as the set of all possible networks. From now on, we will refer to elements in N as consumers or nodes interchangeably. We assume that μ obeys a random graph process known as the *stochastic block model* (SBM). An SBM is characterized by a triplet (n, σ, P), where n is the number of nodes, σ is a probability vector over k types and P is a $k \times k$

symmetric matrix, where the entry P_{ij} gives the independent probability that a node of type i forms a link with a node of type j. In the case of $k = 2$ that fits our model, the type distribution σ is represented by π, and the connectivity matrix P is characterized by three parameters: p_x, the probability that two x types connect; p_y, the probability that two y types connect; and p_{xy}, the probability that different types connect. The components σ and P generate a joint distribution μ over consumer-type profiles and social networks that treats consumers symmetrically, as assumed at the beginning of this section.

The following are two natural specifications of the $k = 2$ SBM. Under *homophily*, agents with similar characteristics are more likely to connect. The connectivity matrix in this case can be captured by two parameters: $p_x = p_y = \alpha$ and $p_{xy} = \beta < \alpha$.[7] An alternative specification is *extroversion/introversion*: some agents have a greater propensity to form social links than others. This case, too, can be represented with two parameters $\alpha > \beta$, such that $p_x = \alpha^2$, $p_y = \beta^2$, and $p_{xy} = \alpha\beta$.

An example: a three-node network with perfect homophily. Let $n = 3$ and assume that nodes i and j are linked in w if and only if $t_i = t_j$. The network is pinned down by the profile of consumer types. In particular, the only networks that are realized with positive probability are the fully connected graph and the graphs in which exactly two nodes are connected. We can use this observation to calculate $\mu(w \mid t_1)$. For example, the probability that the network is fully connected conditional on $t_1 = x$ is π^2, whereas the probability of this network conditional on $t_1 = y$ is $(1 - \pi)^2$; the probability of the network in which only 1 and 2 are linked conditional on $t_1 = x$ is $\pi(1 - \pi)$; and so forth.

Let w_{ijl} denote the fully connected network, and let w_{ij} denote the network in which only nodes i and j are linked. Then,

$$S = \sqrt{\mu(w_{ijl} \mid x)\mu(w_{ijl} \mid y)} + \sqrt{\mu(w_{jl} \mid x)\mu(w_{jl} \mid y)} + 2\sqrt{\mu(w_{ij} \mid x)\mu(w_{ij} \mid y)} = 4\pi(1 - \pi).$$

Therefore, the first-best is implementable in the λ → 1 limit if and only if

$$4\pi(1 - \pi) \leq \sqrt{\frac{1 - \pi}{\pi}}. \tag{20}$$

As in the email example of the second subsection of Section 4, the parameter π plays a double role in this inequality, thanks to the assumption that all consumers are *ex ante* identical. On the RHS (the upper bound on expected uncertainty that is required for implementing the first-best), π represents the platform's prior belief regarding the type of user i. On the LHS, π expresses the platform's prior belief about the other users. Hence, a change in π affects *both* the implementability threshold on the RHS and the Bhattacharyya Coefficient on the LHS.

Inequality (20) simplifies into

$$16\pi^3(1 - \pi) \leq 1.$$

This condition is satisfied as long as $\pi \gtrapprox 0.92$. That is, implementability of the first-best requires a highly *asymmetric* type distribution. Contrast this with our findings in the second subsection of Section 4, where implementability depended on a relatively *symmetric*-type distribution. However, as the next result shows, this conclusion relies on perfect homophily, which is not a generic property: slight perturbation of the connectivity matrix in the example would lead to nonimplementability for sufficiently large π.

Proposition 2. Fix $n \geq 2$ and a generic P. There exist $\pi^*, \pi^{**} \in (\frac{1}{2}, 1)$ with the following property. For every $\pi \in (\pi^*, 1) \cup (\frac{1}{2}, \pi^{**})$ there exists $\lambda^*(\pi)$ such that for every $\lambda > \lambda^*(\pi)$, the optimal policy is not implementable.[8]

[7] Bramoulle et al. (2012) develop an alternative model of homophily, in which the number of nodes is random as well.

[8] There is no necessary relation between π^* and π^{**}.

Thus, a moderately asymmetric consumer-type distribution is necessary for implementing the optimal policy under the SBM (in the high λ regime). To get an intuition for this result, recall our discussion of condition (15), which highlighted two interdependent factors that facilitate implementing the first-best: relative symmetry of the consumer-type distribution and lower uncertainty (about consumer types) for the platform.

The case of $\pi \approx 1$ is simple. For generic P and fixed n, there is an upper limit to the network's informational content, which implies a positive lower bound on the Bhattacharyya Coefficient that is *independent* of π. Therefore, a very asymmetric type distribution overweighs whatever positive effect it may have on the uncertainty factor.

The case of $\pi \approx \frac{1}{2}$ involves different reasoning. Here the network is very uninformative about the nodes' types. For example, in the homophily case with high α and low β, the network is very likely to consist of two fully connected components, yet these will tend to be similar in size and it will be difficult to identify the type of consumers that belong to each component. Thus, both S and $(1 - \pi)/\pi$ will be close to 1 in the $\pi \to \frac{1}{2}$ regime, and it is not clear *a priori* which effect is stronger. However, it turns out that when π is close to $\frac{1}{2}$, the uncertainty factor overweighs the type-distribution symmetry factor.

In the three-node example, we saw that implementing the first-best is impossible for most values of π, even though the SBM was maximally informative given the network size. This raises the question of whether increasing n would help implementing the first-best. The following result gives a positive answer.

Proposition 3. Fix a generic (π, P). There exists n^* such that the optimal policy is implementable for all SBMs (n, π, P) with $n > n^*$.

Thus, for a large enough network, incentive compatibility does not constrain implementing the optimal policy. The proof involves a simple "law of large numbers" argument. For illustration, consider the extreme case of perfect homophily, where $\alpha = 1$ and $\beta = 0$. Any realized network consists of two fully connected components. When n is large, the probability that the larger component consists of x consumers is close to 1. As $n \to \infty$, the network becomes arbitrarily informative, such that S becomes arbitrarily close to 0, and the condition for implementability of the optimal policy is satisfied.

To get a quantitative sense of Proposition 3, consider the following table, which provides values of n^* for various specifications of the homophily case:

π	α	β	λ	n^*
0.6	0.1	0.05	0	1124
0.6	0.1	0.02	0	356
0.75	0.1	0.05	0	485
0.75	0.1	0.02	0	151
0.6	0.01	0.005	0	12, 060
0.6	0.01	0.002	0	3762
0.999	0.1	0.05	0	748
0.75	0.1	0.05	0.833	231
0.75	0.1	0.02	0.833	72

This table illustrates the forces that affect implementability of the optimal policy via their effect on the Bhattacharyya Coefficient of a signal that indicates whether there is a link between two given nodes (see (A3) in Appendix A).

Up to now we assumed that the likelihood of forming links does not change as we increase the network size, such that the expected degree of a node was linear in n. However, in the context of social networks, it makes sense to assume that the average number of links that a node forms grows at a slower rate than the network size. As a result, the network will become sparser as it grows larger. In this case, it is not clear whether a larger network will be more informative than a smaller one, and therefore it is not clear whether the optimal policy will be easier to implement.

To address this question, we turn to a literature within Network Science known as *community detection* (see Mossel, Neeman, and Sly, 2012; Abbe and Sandon, 2015, and the references therein). The objective in this literature is to identify with high probability the types of nodes in a given network, under the assumption that the network was generated by a known SBM. The literature looks for conditions on the SBM parameters that are necessary and sufficient for identifying node types and for implementing the identification with computationally efficient algorithms. These conditions capture the extent to which the network is informative about node types. Because this is also a crucial consideration in our model, the community-detection literature allows us to obtain simple sufficient conditions for implementability of the optimal policy if the network-formation process obeys an SBM.

Following the practice in the community-detection literature, assume that the expected degree of a node grows *logarithmically* with n. Specifically, we assume that the connectivity matrix Pdepends on n, such that

$$p_x = a^2 \frac{\ln(n)}{n} \qquad p_{xy} = b^2 \frac{\ln(n)}{n} \qquad p_y = c^2 \frac{\ln(n)}{n},$$

where a, b, c are arbitrary constants. To derive a *sufficient*condition for implementability of the optimal policy, we borrow existing necessary and sufficient conditions for (asymptotic) *exact recovery*of two asymmetric "communities." By exact recovery, we mean that for a given large network, there exists an algorithm that can identify the type of each node with a probability arbitrarily close to 1. If exact recovery is feasible, then the network is almost perfectly informative. This implies that S is close to 0 and therefore the condition for implementability of the optimal policy holds.

Proposition 4. In the $n \to \infty$ limit, the optimal policy is implementable whenever

$$\pi(a-b)^2 + (1-\pi)(c-b)^2 \geq 2. \tag{21}$$

Note that in the homophily case we have $a = c$, whereas the extroversion/introversion case satisfies $b = \sqrt{ac}$. Thus, Proposition 4 implies the following.

Corollary 1. In the $n \to \infty$ limit, the optimal policy is implementable in the homophily case whenever

$$(a-b)^2 \geq 2,$$

whereas in the extroversion/introversion case, the optimal policy is implementable whenever

$$(\pi a + (1-\pi)c)(\sqrt{a} - \sqrt{c})^2 \geq 2.$$

Thus, when connectivity increases logarithmically with network size, a sufficient condition for implementability of the optimal policy for a large network is that the homophily or extroversion/introversion effects are strong enough.

5. Extensions

■ In this section we examine two extensions of our model.

□ **More than two types.** Our first extension concerns the number of product/preference types. Throughout this article we assumed there are only two types. Suppose that there are $K > 2$ types, denoted as $x_1, \ldots, x_K$. Suppose that the high-quality match probability θ_H applies whenever advertisers' and consumers' types coincide, and that the low-quality match probability θ_L applies in any other case.

Consider the case in which the optimal display rule is interior—that is, $\bar{q}_w^{x_k} > 0$ for every $w \in W$ and $k = 1, \ldots, K$. Then, it is straightforward to show that a necessary and sufficient condition for implementability of the optimal policy is that for every $k, j = 1, \ldots .K$,

$$S(k, j) \leq \lambda \sqrt{\frac{\mu(x_k)}{\mu(x_j)}} + (1 - \lambda)\sqrt{\frac{\mu(x_j)}{\mu(x_k)}},$$

where $\mu(x_k)$ is the *ex ante* probability that a consumer is of type x_k, and $S(k, j)$ is the Bhattacharyya Coefficient of $(\mu(w \mid x_k))_{w \in W}$ and $(\mu(w \mid x_j))_{w \in W}$.

This exercise also demonstrates the usefulness of the Bhattacharyya Coefficient. In the two-type case we could use the measure U, which is defined in terms of the distribution over the platform's posterior belief because that belief was reduced to a scalar. This is no longer the case when $K > 2$. In contrast, the Bhattacharyya Coefficient continues to serve as a meaningful measure of the extent to which the platform's signal separates a given pair of types.

□ **Partially informed advertisers.** Up to now we assumed that advertisers are entirely uninformed of the realization of w. Relaxing this assumption raises a natural question: Can the platform benefit from releasing information to the advertisers? Our first result in this subsection is a negative answer to this question. This finding then raises an immediate follow-up question: When advertisers can partially retrieve the platform's information, how much can they learn without destroying the platform's ability to implement the optimal policy?

To address the first question, return to the *single-consumer* case of Sections 2 and 3, and suppose that before an advertiser submits its report to the platform, it receives a signal s regarding the realization of w. The signal is independent of t conditional on w. Let r be the joint distribution over the platform's signal w and the advertiser's signal s. We allow the advertisers' signals to be correlated conditional on w. The platform does not observe the advertisers' signals.

We extend the incentive-compatibility requirement such that it needs to hold for every realization of s. In principle, because an advertiser's type now consists of both its product type and its information, one would like the pair (q, F) to condition on both. In other words, theoretically advertisers need to report both components of their type. However, because the optimal display rule is only a function of advertisers' product types, it is easy to show that the platform's ability to implement the optimal policy is unaffected if it also requires advertisers to report their signal. Therefore, we will continue to assume that advertisers report only their product type, and this report is the only input that feeds (q, F). Then, the original IC constraints (13) are exactly the same, except that the term $\mu(w)$ is replaced with $r(w \mid s)$. We require advertisers' IR constraint to bind *ex ante*—that is, on average across their signal realizations.

It follows that the necessary and sufficient condition for implementability of the optimal policy can be written as follows. For every realization of s and every $t, t' \in \{x, y\}$,

$$\sum_{w \in W} r(w \mid s) q(t \mid w) \left[\rho_w^t - \rho_w^{t'}\right] \leq 0. \tag{22}$$

By Blackwell's ranking of information systems, r' is less informative than r if there is a system of conditional probabilities $(p(s \mid s'))_{s,s'}$, such that for every w, s,

$$r'(s \mid w) = \sum_{s'} p(s \mid s') r(s' \mid w).$$

The following result establishes that the platform benefits from withholding information from advertisers.

Proposition 5. (i) If the optimal policy is implementable under r, then it is implementable under any r' that is less informative than r.

(ii) If advertisers are fully informed of the platform's signal (i.e., $r(w \mid w) = 1$ for every w), the optimal policy is not implementable when λ is sufficiently large.

The reason why releasing information about w to advertisers cannot help the platform is standard—it means that IC constraints that previously held only on average are now required to hold for all signals. Part (ii) of the result establishes that this monotonicity result is not vacuous: giving advertisers full information about the platform's signal will prevent it from implementing its optimal policy when λ is large.

Suppose that the platform cannot prevent advertisers from learning *part* of its own signal; how much information can it afford to give away? In the remainder of this section, we analyze this question in the context of the *social network* application of the third subsection of Section 4. In particular, consider an SBM and assume that each advertiser gets information by sampling a random subset of no more than d nodes (out of the total of n nodes in the network), and learning the subgraph of w over these d nodes. Recall that w is realized according to a given SBM. Hence, the Bhattacharyya Coefficient can be defined for any subgraph of w consisting of k nodes, $k = 1, \dots, n$ (where the connectivity matrix is fixed). Denote this coefficient by $S(k)$.

Proposition 6. Suppose each advertiser is informed of the subgraph induced by w over a random subset of at most d nodes. If

$$S(n-d) \leq \left[\lambda\sqrt{\frac{\pi}{1-\pi}} + (1-\lambda)\sqrt{\frac{1-\pi}{\pi}}\right] - \left[\frac{d}{n-d}\cdot\frac{\sqrt{2}-1}{2\sqrt{\pi(1-\pi)}}\right] \tag{23}$$

then the optimal policy is implementable.

Note that the term in the first bracket on the RHS of (23) is precisely the RHS of (18), the necessary and sufficient condition for implementing the optimal policy, whereas the term in the second bracket is some positive constant that increases in d. The term on the LHS measures the uncertainty in the subgraph that advertisers do *not* observe.

When π and the connectivity matrix are fixed, inequality (23) is stated entirely in terms of d and n. We can therefore express $S(n-d)$ as a function of d, and use the upper bounds on $S(k)$ that we derived in the third subsection of Section 4 to get a closed-form upper bound on d, such that the optimal policy is implementable for any value of d below that bound. Finally, the comparative statics with respect to d are consistent with our previous results. When d increases, the RHS of (23) clearly goes down, whereas $S(n-d)$ goes up because a smaller network is a less informative signal. Thus, a larger d makes it more difficult to satisfy the sufficient condition.

Appendix A: Proofs

Proposition 1. From (7), it follows that (5) characterizes the optimal display policy. Plugging this expression for q^t_w into the $IC(x, y)$ constraint (13) yields the following inequality:

$$\sum_{w\in W} \mu(w)\cdot\sqrt{\mu(x \mid w)\mu(y \mid w)} \leq \sum_{w\in W} \mu(w)\cdot[\lambda\mu(y \mid w) + (1-\lambda)\mu(x \mid w)]. \tag{A1}$$

Note that $\mu(w)\mu(t \mid w) = \mu(t, w)$ and $\sum_{w\in W}\mu(x, w) = \pi$. The above inequality can thus be rewritten as (18). If we carry out a similar exercise for $IC(y, x)$, we obtain the inequality

$$U \leq \lambda\pi + (1-\lambda)(1-\pi).$$

By assumption, $\lambda, \pi \geq \frac{1}{2}$. Therefore, the only inequality that matters is (18).

Proposition 2. The proof will use the following property of the Bhattacharyya Coefficient.

Remark 1. Suppose that for any consumer i, we can represent w as a pair, $w = (g_1, g_2)$, such that $\mu(g_1, g_2 \mid t_i) \equiv \mu(g_1 \mid t_i)\mu(g_2 \mid t_i)$. For every $k = 1, 2$, define

$$S_k = \sum_{g_k} \sqrt{\mu(g_k \mid t_i = x)\mu(g_k \mid t_i = y)}.$$

Then, $S = S_1 \cdot S_2$.

Remark 1 says that the Bhattacharyya Coefficient induced by a collection of signals that are independent conditional on the consumer's type is the product of the signals' coefficients. The property follows immediately from the coefficient's definition, and therefore the proof is omitted.

Our method of proof is to obtain two different lower bounds on S, and use these bounds to derive π^* and π^{**}.

(i) Fix a node i. Suppose that the platform were informed of the realized network w, as well as of t_j for all $j \neq i$. This would clearly be a (weakly) more informative signal of t_i than learning w only. Moreover, conditional on learning $(t_j)_{j \neq i}$, the link status between any $j, h \neq i$ has no informational content regarding t_i (this follows from the assumption that the SBM is known and from Remark 1). Therefore, in order to calculate a lower bound on S, we can consider a signal that consists of $(t_j)_{j \neq i}$ and the link status between i and every other j.

Let us calculate the Bhattacharyya Coefficient of the signal that consists of learning t_j and whether nodes i and j are linked:

$$\begin{aligned}
&\sqrt{\pi p_x \cdot \pi p_{xy}} + \sqrt{\pi(1 - p_x) \cdot \pi(1 - p_{xy})} \\
&+ \sqrt{(1 - \pi)p_{xy} \cdot (1 - \pi)p_y} + \sqrt{(1 - \pi)(1 - p_{xy}) \cdot (1 - \pi)(1 - p_y)} \\
&= \pi\left(\sqrt{p_x p_{xy}} + \sqrt{(1 - p_x)(1 - p_{xy})}\right) + (1 - \pi)\left(\sqrt{p_y p_{xy}} + \sqrt{(1 - p_y)(1 - p_{xy})}\right).
\end{aligned}$$

Because signals that correspond to different nodes $j \neq i$ are independent conditional on t_i, Remark 1 implies that the Bhattacharyya Coefficient of the signal that consists of $(t_j)_{j \neq i}$ and the link status between i and every other j is

$$\left[\pi\left(\sqrt{p_x p_{xy}} + \sqrt{(1 - p_x)(1 - p_{xy})}\right) + (1 - \pi)\left(\sqrt{p_y p_{xy}} + \sqrt{(1 - p_y)(1 - p_{xy})}\right)\right]^{n-1}.$$

Recall that by construction, this expression is weakly below S. Without loss of generality, let

$$\sqrt{p_x p_{xy}} + \sqrt{(1 - p_x)(1 - p_{xy})} \leq \sqrt{p_y p_{xy}} + \sqrt{(1 - p_y)(1 - p_{xy})}.$$

Then, S is weakly above

$$\delta \equiv \left(\sqrt{p_x p_{xy}} + \sqrt{(1 - p_x)(1 - p_{xy})}\right)^{n-1}.$$

For generic P (in particular, when all matrix entries get values in (0,1)), this term is strictly positive.

For any δ, we can find π^* sufficiently close to 1 such that $\sqrt{(1 - \pi^*)/\pi^*} = \delta^2 < 1$. For any $\pi > \pi^*$, let $\sqrt{(1 - \pi)/\pi} = \hat{\delta}^2$ where $\hat{\delta} < \delta$, and choose λ to be sufficiently close to 1 such that the RHS of (18) is arbitrarily close to $\hat{\delta}^2$, and therefore below δ, thus violating (18).

(ii) Let us now obtain a different lower bound on S. Once again, we use the fact that S decreases with the informativeness of the signal given by the network. For fixed n and π, this informativeness is maximal under perfect homophily—that is, when $p_x = p_y = 1$ and $p_{xy} = 0$. Assume perfect homophily, and consider an arbitrary node. Conditional on this node's type, if we learn whether it is linked to the other nodes, we do not gain any additional information from learning the links among these other nodes. The reason is that conditional on the node's type, it is linked to another node if and only if the two nodes' types are identical. Thus, knowing the node's type and its link status with all other nodes, we can entirely pin down the rest of the network. Moreover, conditional on the node's type, its link status with respect to some node is independent of its link status with respect to another node.

It follows that the signal given by the network under perfect homophily is equivalent to a collection of $n - 1$ conditionally independent signals: each signal generates a link with probability π $(1 - \pi)$ conditional on the original node's type being x (y). By Remark 1, the Bhattacharyya Coefficient for this network is thus

$$\left(\sqrt{\pi(1 - \pi)} + \sqrt{(1 - \pi)\pi}\right)^{n-1}.$$

As this expression is weakly lower than S, the following inequality is a necessary condition for the implementability of the optimal policy:

$$\left(\sqrt{4\pi(1 - \pi)}\right)^{n-1} \leq \lambda\sqrt{\frac{1 - \pi}{\pi}} + (1 - \lambda)\sqrt{\frac{\pi}{1 - \pi}}.$$

This inequality can be rewritten as follows:

$$2^{n-1}\pi^{\frac{n}{2}}(1 - \pi)^{\frac{n}{2}-1} - \lambda - (1 - \lambda)\left(\frac{\pi}{1 - \pi}\right) \leq 0. \tag{A2}$$

The inequality is binding for $\pi = \frac{1}{2}$. All we now need to show is that there exists $\pi^{**} > \frac{1}{2}$ and a function $\lambda^*(\pi)$ such that for every $\pi \in (\frac{1}{2}, \pi^{**})$ and every $\lambda > \lambda^*(\pi)$, the derivative of (A2) with respect to π is strictly positive. Straightforward calculation establishes that this is indeed the case.

Proposition 3. Fix an arbitrary node i. Suppose that we were given a signal that describes only whether there is a link between i and some given node $j \neq i$. The probability of a link conditional on $t_i = x$ is $\eta_x = \pi p_x + (1-\pi)p_{xy}$, and the probability of a link conditional on $t_i = y$ is $\eta_y = \pi p_{xy} + (1-\pi)p_y$. Therefore, the Bhattacharyya Coefficient that corresponds to this signal is

$$\sqrt{\eta_x \eta_y} + \sqrt{(1-\eta_x)(1-\eta_y)}. \tag{A3}$$

Now suppose that we are given a signal that describes whether there is a link between i and *each* of the other $n-1$ nodes. As the probability of such a link is independent across all $j \neq i$ conditional on t_i, Remark 1 implies that the Bhattacharyya Coefficient that corresponds to this signal is

$$[\sqrt{\eta_x \eta_y} + \sqrt{(1-\eta_x)(1-\eta_y)}]^{n-1}. \tag{A4}$$

Now, observe that this signal is weakly less informative than learning the entire network w. Therefore, S is weakly below the expression (A4). It follows that the following inequality is a sufficient condition for the implementability of the optimal policy:

$$[\sqrt{\eta_x \eta_y} + \sqrt{(1-\eta_x)(1-\eta_y)}]^{n-1} \leq \lambda\sqrt{\frac{1-\pi}{\pi}} + (1-\lambda)\sqrt{\frac{\pi}{1-\pi}}. \tag{A5}$$

For generic (π, P), $\eta_x \neq \eta_y$, such that $\sqrt{\eta_x \eta_y} + \sqrt{(1-\eta_x)(1-\eta_y)} < 1$. In addition, for any quadruple $(\pi, \theta_L, \theta_H, \varepsilon)$ that satisfies (7), the RHS of (A5) is bounded away from 0. Therefore, there exists n^* such that the inequality holds for every $n > n^*$.

Proposition 4. Recall that

$$S = \frac{1}{\sqrt{\pi(1-\pi)}} \sum_{w \in W} \mu(w)\sqrt{\mu(t_i = x \mid w)\mu(t_i = y \mid w)}$$

for any node i. Exact recovery means that the probability (measured according to μ) of realizations w for which $\mu(t_i = x \mid w)$ or $\mu(t_i = y \mid w)$ are arbitrarily close to 0 is arbitrarily high. Therefore, exact recovery is ensured if $S \to 0$ when $n \to \infty$.

Let $n \to \infty$. Given the preceding paragraph, we need only to derive a sufficient condition for exact recovery. By Abbe and Sandon (2015), such a network is exactly recoverable if and only if

$$\max_{r \in [0,1]} \left\{ r[\pi a^2 + (1-\pi)b^2] + (1-r)[\pi b^2 + (1-\pi)c^2] - \pi a^{2r} b^{2(1-r)} - (1-\pi) b^{2r} c^{2(1-r)} \right\}. \geq 1$$

A sufficient condition for this inequality to hold is that the maximand of the LHS is weakly greater than one for $r = \frac{1}{2}$—that is, if

$$\pi\left(\frac{a^2+b^2}{2}\right) + (1-\pi)\left(\frac{c^2+b^2}{2}\right) - \pi(ab) - (1-\pi)(cb) \geq 1,$$

which is equivalent to (21).

Proposition 5. (i) The proof is entirely rudimentary and standard. Nevertheless, we give it for completeness. By assumption, inequality (22) holds for every s. Using the definition of Blackwell informativeness, we can rewrite $r'(w \mid s)$ as

$$\begin{aligned} &= \frac{\mu(w)}{r'(s)} r'(s \mid w) = \frac{\mu(w)}{r'(s)} \sum_{s'} p(s \mid s') r(s' \mid w) \\ &= \frac{\mu(w)}{r'(s)} \sum_{s'} p(s \mid s') \frac{r(s') r(w \mid s')}{\mu(w)} = \sum_{s'} \frac{p(s \mid s') r(s')}{r'(s)} r(w \mid s'), \end{aligned}$$

where $r(s')$ is the *ex ante* probability of the signal s' under r, and $r'(s)$ is the *ex ante* probability of the signal s under r'. Now, elaborate the term

$$\frac{p(s \mid s') r(s')}{r'(s)} = \frac{\sum_w \mu(w) p(s \mid s') r(s' \mid w)}{\sum_{s''} \sum_w \mu(w) p(s \mid s'') r(s'' \mid w)}.$$

We can easily see that this term is between 0 and 1, and that

$$\sum_{s'} \frac{p(s \mid s') r(s')}{r'(s)} = 1.$$

It follows that for every s, $r'(w \mid s)$ is some convex combination of $(r(w' \mid s))_{w'}$. Therefore, given that under r, (22) holds for every s, it must hold under r' as well. (ii) Suppose that advertisers are fully informed of the realization of w. Then, the necessary and sufficient conditions for implementability of the optimal policy are that for every w,

$$\sqrt{\mu(x \mid w)\mu(y \mid w)} \leq \lambda\mu(y \mid w) + (1-\lambda)\mu(x \mid w)$$

$$\sqrt{\mu(x \mid w)\mu(y \mid w)} \leq \lambda\mu(x \mid w) + (1-\lambda)\mu(y \mid w).$$

Consider a signal realization w^* for which $\mu(x \mid w^*) \neq \frac{1}{2}$ (there must exist such a realization). The above inequalities can be written as

$$1 \leq \lambda\sqrt{\frac{\mu(y \mid w^*)}{\mu(x \mid w^*)}} + (1-\lambda)\sqrt{\frac{\mu(y \mid w^*)}{\mu(x \mid w^*)}} \tag{A6}$$

$$1 \leq \lambda\sqrt{\frac{\mu(x \mid w^*)}{\mu(y \mid w^*)}} + (1-\lambda)\sqrt{\frac{\mu(x \mid w^*)}{\mu(y \mid w^*)}}. \tag{A7}$$

Because $\mu(x \mid w^*) \neq \frac{1}{2}$, either $\mu(x \mid w^*) > \mu(y \mid w^*)$ or $\mu(x \mid w^*) > \mu(y \mid w^*)$. Assume the former, without loss of generality. As inequality (A7) is violated for $\lambda = 1$, this inequality would also be violated whenever λ is sufficiently large.

Proposition 6. Suppose an advertiser learns the subgraph of w over some subset of nodes N_1 (the size of which is n_1). We can represent w as a triple (g_1, g_2, h), where g_1 is the subgraph that the advertiser learns, g_2 is the subgraph induced by w over the remaining set of nodes $N_2 = N - N_1$ (the size of which is n_2), and h consists of all links between a node in N_1 and a node in N_2. Because w is generated by an SBM and g_1 and g_2 are defined over disjoint sets of nodes, g_1 and g_2 are independently distributed.

The necessary and sufficient condition for implementability of the optimal policy is that for every signal g_1,

$$\sum_{g_2,h} \mu(g_2, h \mid g_1) \sum_{i \in N} \sqrt{\mu(t_i = x \mid g_1, g_2, h)\mu(t_i = y \mid g_1, g_2, h)} \tag{A8}$$

$$\leq \sum_{g_2,h} \mu(g_2, h \mid g_1) \sum_{i \in N} \left[\lambda\mu(t_i = y \mid g_1, g_2, h) + (1-\lambda)\mu(t_i = x \mid g_1, g_2, h)\right]$$

and

$$\sum_{g_2,h} \mu(g_2, h \mid g_1) \sum_{i \in N} \sqrt{\mu(t_i = x \mid g_1, g_2, h)\mu(t_i = y \mid g_1, g_2, h)} \tag{A9}$$

$$\leq \sum_{g_2,h} \mu(g_2, h \mid g_1) \sum_{i \in N} \left[\lambda\mu(t_i = x \mid g_1, g_2, h) + (1-\lambda)\mu(t_i = y \mid g_1, g_2, h)\right].$$

These expressions are easily derived from the inequality (A1) given at the beginning of the proof of Proposition 1.

Because g_1 and g_2 are independent, we can write $\mu(g_2, h \mid g_1) = \mu(g_2)\mu(h \mid g_1, g_2)$. Also, observe that $\mu(t_i = x \mid g_1, g_2) = \sum_h \mu(h \mid g_1, g_2)\mu(t_i = x \mid g_1, g_2, h)$. Applying the Cauchy–Schwartz inequality, we obtain

$$\sqrt{\mu(t_i = x \mid g_1, g_2)\mu(t_i = y \mid g_1, g_2)} \geq \sum_h \mu(h \mid g_1, g_2)\sqrt{\mu(t_i = x \mid g_1, g_2, h)\mu(t_i = y \mid g_1, g_2, h)}.$$

It follows that inequalities (A8) and (A9) are implied by the following, simpler inequalities:

$$\sum_{i \in N}\left[\sum_{g_2} \mu(g_2)\sqrt{\mu(t_i = x \mid g_1, g_2)\mu(t_i = y \mid g_1, g_2)} - \lambda\mu(t_i = y \mid g_1) - (1-\lambda)\mu(t_i = x \mid g_1)\right] \leq 0$$

$$\sum_{i \in N}\left[\sum_{g_2} \mu(g_2)\sqrt{\mu(t_i = x \mid g_1, g_2)\mu(t_i = y \mid g_1, g_2)} - \lambda\mu(t_i = x \mid g_1) - (1-\lambda)\mu(t_i = y \mid g_1)\right] \leq 0.$$

Consider the top inequality (it will be easy to see that if it holds, the bottom inequality holds as well). We can break the summation over $i \in N$ into two summations over N_1 and N_2. Because g_1 and g_2 are independent, for every $i \in N_1$ we can write $\mu(t_i = x \mid g_1, g_2) = \mu(t_i = x \mid g_1)$. Similarly, for every $i \in N_2$ we can write $\mu(t_i = x \mid g_1, g_2) = \mu(t_i = x \mid g_2)$ and $\mu(t_i = x \mid g_1) = \mu(t_i = x) = \pi$. It follows that the inequality can be rewritten as

$$\sum_{i \in N_2}\left[\sum_{g_2}\left(\mu(g_2)\sqrt{\mu(t_i = x \mid g_2)\mu(t_i = y \mid g_2)} - \lambda\mu(t_i = x \mid g_1) - (1-\lambda)\mu(t_i = y \mid g_1)\right)\right]$$

$$+\sum_{i\in N_1}\left[\sum_{g_2}\left(\mu(g_2)\sqrt{\mu(t_i=x\mid g_1)\mu(t_i=y\mid g_1)}-\lambda\mu(t_i=x\mid g_1)-(1-\lambda)\mu(t_i=y\mid g_1)\right)\right]$$
$$\leq 0.$$

The top sum can be simplified into

$$n_2S(n_2)\sqrt{\pi(1-\pi)}-n_2\lambda\pi-n_2(1-\lambda)(1-\pi)$$

and the bottom sum can be grouped together as

$$\sum_{i\in N_1}\left[\sqrt{\mu(x\mid g_1)\mu(y\mid g_1)}-\lambda\mu(x\mid g_1)-(1-\lambda)\mu(y\mid g_1)\right]$$
$$\leq n_1\cdot\max_{\chi\in\{0,1\}}\max_{\varphi\in[0,1]}\left[\sqrt{\varphi(1-\varphi)}-\chi\varphi-(1-\chi)(1-\varphi)\right]$$
$$=n_1\cdot\frac{\sqrt{2}-1}{2}.$$

Plugging this term and exploiting the assumption that $\pi>\frac{1}{2}$, we can now obtain the following sufficient condition for implementability of the optimal policy:

$$n_2\left[S(n_2)\sqrt{\pi(1-\pi)}-\lambda\pi-(1-\lambda)(1-\pi)\right]+n_1\frac{\sqrt{2}-1}{2}\leq 0. \tag{A10}$$

Substituting d for n_1 and $n-d$ for n_2 yields the desired condition.

Appendix B: Finitely many advertisers

Suppose there are A advertisers of each type. When a single advertiser of type x pretends to be y, it changes its probability of display from q_w^x/A to $q_w^y/(A+1)$. Hence, the transition probability from D_x to S_x changes to

$$\theta_H\left(q_w^x+\frac{q_w^y}{A+1}\right)+\theta_L\frac{Aq_w^y}{A+1}$$

because a type x consumer will transact with probability θ_H if the displayed ad is either by one of the truthful x advertisers or by the single deviating advertiser, and with probability θ_L if the displayed ad is by one of the truthful y advertisers. Similarly, the transition probability from D_y to S_y changes to

$$\theta_H\frac{Aq_w^y}{A+1}+\theta_L\left(q_w^x+\frac{q_w^y}{A+1}\right).$$

Consequently, the invariant probability that the consumer is of type x and in state D_x is

$$\tilde{\rho}_w^x=\frac{\mu(x\mid w)\varepsilon}{\left(\frac{A}{A+1}\right)(\theta_H-\theta_L)q_w^x+\theta_H\left(\frac{1}{A+1}\right)+\theta_L\left(\frac{A}{A+1}\right)+\varepsilon}$$

and the invariant probability that he is in state D_y is

$$\tilde{\rho}_w^y=\frac{\mu(y\mid w)\varepsilon}{\left(\frac{A}{A+1}\right)(\theta_H-\theta_L)q_w^y+\theta_L+\varepsilon}.$$

In a similar manner, we can derive the invariant probabilities when a single y advertiser deviates. Note that $\tilde{\rho}\to\rho_i$ as $A\to\infty$.

It follows that an x advertiser weakly prefers to report its type if and only if

$$\sum_{w\in W}\mu(w)\frac{q_w^x}{A}\left[\theta_H\rho_w^x+\theta_L\rho_w^y\right]-F_x\geq\sum_{w\in W}\mu(w)\frac{q_w^y}{A+1}\left[\theta_H\tilde{\rho}_w^x+\theta_L\tilde{\rho}_w^y\right]-F_y.$$

This inequality is approximated by (11) when A is large, up to a normalization of the fees F_x and F_y. The IC constraint of a y advertiser is handled similarly.

References

Abbe, E. and Sandon, C. “Community Detection in General Stochastic Block Models: Fundamental Limits and Efficient Recovery Algorithms.” IEEE 56th Annual Symposium on Foundations of Computer Science, 2015, pp. 670–688.

Athey, S. and Gans, J.S. “The Impact of Targeting Technology on Advertising Markets and Media Competition.” *American Economic Review: Papers & Proceedings*, Vol. 100 (2010), pp. 608–613.

Basu, A., Shioya, H., and Park, C. *Statistical Inference: The Minimum Distance Approach*. New York: Chapman and Hall/CRC, 2011.

Bergemann, D. and Bonatti, A. "Targeting in Advertising Markets: Implications for Offline Verus Online Media." *Rand Journal of Economics*, Vol. 42 (2011), pp. 417–443.

——— and ———. "Selling Cookies." *American Economic Journal: Microeconomics*, Vol. 7 (2015), pp. 259–294.

Bloch, F. "Targeting and Pricing in Social Networks." In Y. Bramoulle, A. Galeotti and B. Rogers, eds., *The Oxford Handbook of the Economics of Networks* Oxford. New York: Oxford University Press, 2016.

Bramoulle, Y., Currarini, S., Jackson, M., Pin, P., and Rogers, B. "Homophily and Long-Run Integration in Social Networks." *Journal of Economic Theory*, Vol. 147 (2012), pp. 1754–1786.

Campbell, J. "Localized Price Promotions as a Quality Signal in a Publicly Observable Network." *Quantitative Marketing and Economics*, Vol. 13 (2015), pp. 27–57.

Candogan, O., Bimpikis, K., and Ozdaglar, A. "Optimal Pricing in Networks with Externalities." *Operations Research*, Vol. 60 (2012), pp. 883–905.

Eliaz, K. and Spiegler, R. "A Simple Model of Search Engine Pricing." *Economic Journal*, Vol. 121 (2011), pp. F329–F339.

——— and ———. "Search Design and Broad Matching." *American Economic Review*, Vol. 105 (2015), pp. 563–586.

Fainmesser, I. and Galeotti, A. "Pricing Network Effects." *Review of Economic Studies*, Vol. 83 (2015), pp. 165–198.

Galeotti, A. and Goyal, S. "Network Multipliers: The Optimality of Targeting Neighbors." *Review of Network Economics*, Vol. 11 (2012), pp. 1446–9022.

Iyer, G., Soberman, D., and Villas-Boas, J.M. "The Targeting of Advertising." *Management Science*, Vol. 24 (2005), pp. 461–476.

Johnson, J.P. "Targeted Advertising and Advertising Avoidance." *Rand Journal of Economics*, Vol. 44 (2013), pp. 128–144.

Mossel, E., Neeman, J. and Sly, A. "Stochastic Block Models and Reconstruction." mimeo, 2012.

Theodoridis, S. and Koutroumbas, K. *Pattern Recognition*. Boston, MA: Academic Press, 2008.

Zubcsek, P.P. and Sarvary, M. "Advertising to a Social Network." *Quantitative Marketing and Economics*, Vol. 9 (2012), pp. 71–107.