Identification and Estimation in Search Models with Social Information*

Niccolò Lomys[†] Emanuele Tarantino[‡]

November 14, 2021

Abstract

We propose a theoretical analysis of the conditions under which estimates of search cost distributions are biased when Bayes rational agents search in the presence of social information. We extend the canonical empirical sequential and simultaneous search models by allowing a share of the agents in the population to observe the choice of one of their social connections. We find that social information changes agents' optimal search decisions. We compute the estimator of search cost distributions under various standard datasets. We find that neglecting social information typically leads to biased and inconsistent estimates of search cost distributions, with the bias sign and magnitude depending on the dataset's content. The bias magnitude is increasing in the share of agents in the population with social information. We also discuss offline estimation techniques, exogenous variations in the data, and partial identification approaches that are useful to recover correct estimates of search cost distributions.

Keywords: Search; Social Learning; Networks; Identification and Estimation of Search Costs.

JEL Classification: C5; C8; D1; D6; D8.

*We thank Daniel Ershov, Philipp Kircher, and Lorenzo Magnolfi for helpful conversations. Niccolò Lomys acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement № 714147) and from the ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

[†]Toulouse School of Economics, University of Toulouse Capitole; niccolo.lomys@tse-fr.eu.

[‡]LUISS University, EIEF, and CEPR; etarantino@luiss.it.

1 Introduction

Consumers seldom search in isolation: other consumers' choices and experiences are readily available via direct observations, communication, online social networks, popularity rankings, and product reviews. A recent theoretical literature has analyzed how observational learning (see, e.g., Kircher and Postlewaite, 2008; Hendricks, Sorensen and Wiseman, 2012; Mueller-Frank and Pai, 2016; Garcia and Shelegia, 2018; Choi, Dai and Kim, 2018; Lomys, 2020) or communication learning (Galeotti, 2010) affect individual search behavior and has shown that others' choices and experiences are an important source of information. However, structural econometric models of search typically assume that consumers make their choices in isolation, relying only on the product information they gather by inspecting firms' outlets themselves (for an extensive review of the empirical literature on search, see, e.g., Honka, Hortaçsu and Wildenbeest, 2019). One reason is that researchers may lack data on agents' connections in the social network. Another reason is that the theoretical literature has not explored the implications of social learning for the identification and estimation of empirical search models. This paper attempts to fill this gap.

Assuming that individuals search in isolation prevents researchers from asking many questions of interest. Conceptually, the questions concern the determination of the conditions under which social learning diverts consumers from searching. Empirically, the questions concern the identification and estimation of empirical search models in the presence of social learning. First, does neglecting social learning bias the estimates of search cost distributions? Second, how do the bias direction and magnitude depend on the content of the available dataset? Third, what could researchers do to determine the bias and recover correct estimates of the parameters of interest? Fourth, how do the answers to these questions depend on the specific empirical search model that is considered (e.g., sequential *vs.* simultaneous)?

Unbiased estimates of search cost distributions are crucial to quantify the benefits of policies aimed at reducing search frictions. The theoretical and empirical literature on search has documented that the presence of search costs significantly affects consumer choices, pricing behavior, and market outcomes (for extensive reviews of the theoretical literature on search, see, e.g., Baye, Morgan and Scholten, 2006; Anderson and Renault, 2018). Thus, whether people choose a product because they learned its value from searching, or because they infer its value from their connections' choices, may have important welfare consequences.

We propose a theoretical setting based on the canonical empirical sequential and simultaneous search models. We solve both models with and without social information. With social information, an agent observes the choice of one of her connections in the social network before starting her search process. We find that an agent's optimal search decisions under social information considerably differ from those without social information. We then compute the estimator of the search cost distribution with and without social information. The estimator depends on the content of the dataset. Following the empirical literature, we consider data on agents' optimal choice, search duration (that is, the number of searches), or optimal stopping. Reflecting the differences in optimal search decisions with and without social information, we show that neglecting agents' social information typically leads to biased and inconsistent estimates of search cost distributions. We relate the bias direction and magnitude to the content of the dataset and the model's primitives. We conclude by presenting potential remedies—such as offline estimation techniques, exogenous variations in the data, and partial identification approaches—that are useful to recover correct estimates of search cost distributions.

In the main text, we develop our insights in a sequential search setting (in the appendix, we consider a simultaneous search and show that all our insights remain valid). Absent social information, we consider a stylized version of the canonical sequential search model with recall by Weitzman (1979) as adopted by the empirical literature on search (Honka et al., 2019). An agent must choose between two alternatives whose utilities, high or low, are independent draws from the same distribution. Before searching, the agent knows the true utility distribution but not the realized utilities of the two alternatives. Searching for an alternative perfectly reveals its realized utility to the agent. The search process develops in two steps: after searching the first alternative, the agent decides whether to search the second alternative or not. Only the second search involves a cost (high or low). We extend this model to account for the presence of social information. Before starting her search process, with some probability, the agent observes the alternative that has been taken by one of her social connections but neither her search behavior nor her idiosyncratic search cost (search costs are i.i.d. across the two agents). This connection faces the same search problem as the agent; thus, the agent draws some inference about the realized utilities of the available alternatives from her connection's choice.

The optimal search decisions of an agent with social information differ in two respects compared to those of an isolated agent. First, an agent with social information is not indifferent about which alternative to search first, whereas an isolated agent is so. In particular, an agent with social information finds it optimal to start searching from the alternative taken by her connection. If her connection chose alternative a, with positive probability she searched alternative $a' \neq a$ as well and concluded that a delivers higher utility. Thus, in the eyes of the agent with social information, the distribution of the utility of alternative a first-order stochastically dominates that of alternative a'. Second, the expected gain from the second search for an agent with social information is lower than that of an isolated agent. This is so because an agent with social information discounts the expected gains from the second search absent social information by the probability that her connection has searched both alternatives. Searching the second alternative is valuable for the agent only if her connection has conducted only one search. Thus, the possibility to exploit her social information reduces the agent's incentives to engage in independent exploration. These two differences explain when and why neglecting social information leads to biased and inconsistent estimates of search cost distributions.

We compute the (large sample) estimator of search cost distributions. In our setting, this is equivalent to computing an estimator of the probability that an agent's search cost is high. We do so in three classes of datasets. The first contains data on agents' optimal choice, where the researcher observes the utilities of the alternatives taken by the agents in the sample. The second contains data on search duration, where the researcher observes the share of agents in the sample that conducted one search (and that of agents that conducted two searches). The third contains data on optimal stopping, where the researcher observes the share of agents who discontinue search after the first search and the utility of the first searched alternative.

Our main insight is the following: if constructed neglecting that a share of agents in the population have social information, the estimators of search cost distributions are typically biased and inconsistent. The bias direction and magnitude depend on the dataset's content and the model's primitives. When present, the bias increases as the share of agents in the population with social information grows large.

Based on the model's primitives, we distinguish between two cases. In the first case, the expected gain from the second search for an agent with social information is *larger* than the smallest search cost in the support of the search cost distribution (as is the case without social information). Then, with social information, an agent searches first for the alternative taken by her social connection; at the second search stage, she will continue searching under the same conditions in which she would do absent social information. If data on agents' optimal choices are used, neglecting social information leads to an underestimation of search costs. The reason is that an agent with social information is more likely to take a high utility alternative than an isolated agent. Instead, if data on search duration is used, neglecting social information leads to an overestimation of search costs. Since an agent searches first a high utility alternative with a larger probability than an isolated agent, she is also more likely to discontinue the process after the first search. The only case in which the search cost distribution is estimated consistently and without a bias is when data on optimal stopping are used. This is so because, at the second search stage, an agent with social information discontinues search under the same conditions (in terms of private search cost and utility of the first searched alternative) in which she would do so without social information.

In the second case, the expected gain from the second search for an agent with social information is *smaller* than the smallest search cost in the support of the search cost distribution. Then, with social information, an agent searches first for the alternative taken by her social connection; at the second search stage, in contrast to what she would do absent social information, the agent always discontinues search. If data on search duration and optimal stopping is used, neglecting social information leads to an overestimation of search costs. In both cases, the reason is that an agent with social information discontinues her search earlier/with a higher probability than an isolated agent. If data on optimal choices are considered, the search cost distribution is estimated consistently and without a bias. Since an agent with social information searches first for the alternative taken by her social connection and always discontinues search thereafter, the probability with which an agent with social information takes an alternative of a given utility is the same as that of an isolated agent.

We discuss three possible empirical strategies to recover correct estimates of search cost distributions in our context. First, if the researcher can estimate offline (consistently and without a bias) the share of agents with social information—e.g., by using detailed network data[1]—it is possible to construct an unbiased and consistent estimator of the search cost distribution. Second, we show how to exploit exogenous variations in observables (namely, in the utility distributions) to recover correct estimates of the search cost distribution. Third, we develop a partial identification approach that allows the researcher to construct useful bounds on the search cost distribution without any knowledge of the share of agents with social information. We also show how to jointly set-identify the search cost distribution and the share of agents in the population with social information.

[1] See, e.g., Kuchler, Li, Peng, Stroebel and Zhou (2021) and Bailey, Johnston, Kuchler, Stroebel and Wong (2021) for empirical analyses of peer effects using network data from Facebook.

Outline. In Section 2, we present our sequential search model. In Section 3, we characterize optimal sequential search and choice decisions with and without social information. In Section 4, we show when and why neglecting social information leads to biased and inconsistent estimates of search cost distributions under different possible datasets. In Section 5, we discuss possible approaches to recover unbiased and consistent estimates of search cost distributions in the context of our setting. In Section 6, we conclude. In Appendix A, we show that our insights hold in a simultaneous search setting as well.

2 Model

In this section, we present our sequential search model. The baseline setting is a version of the canonical sequential search model by Weitzman (1979) as developed in the empirical literature on search (see Section 2 in Honka et al., 2019), from which it borrows all the main assumptions. We then augment the baseline setting by introducing social information.

There is a countably infinite set of search problems, indexed by $n \in \mathbb{N} := \{1, 2, \ldots\}$. In search problem n, a Bayes rational agent—called agent n—must select an alternative from the set $X := \{0, 1\}$. We denote by x a typical element of X and by $\neg x$ the alternative in X other than x. Let u_n^x denote agent n's (indirect) utility from alternative x. Utilities u_n^x are i.i.d. draws—across alternatives within search problem n and across search problems—from a probability distribution over $U := \{\underline{u}, \overline{u}\}$, where $0 \leq \underline{u} < \overline{u}$. Let $\alpha := \mathbb{P}(u_n^x = \overline{u})$, and assume $\alpha \in (0, 1)$. Moreover, let $\Delta u := \overline{u} - \underline{u}$.

Agent n wishes to select the alternative with the highest realized utility. Agent n knows the true utility distribution, but not the specific pair of realized utilities (u_n^0, u_n^1). Agent n collects information about realized utilities via costly sequential search with perfect recall. Formally, agent n takes the following decisions in sequence.

1. First, agent n decides which alternative to search first; let $s_n^1 \in X$ denote agent n's decision of which alternative to search first. By searching alternative s_n^1, agent n perfectly learns its realized utility $u_n^{s_n^1}$.

2. Next, agent n decides whether to search the remaining alternative, denoted by $s_n^2 = \neg s_n^1$, and perfectly learn its realized utility $u_n^{s_n^2}$, or to discontinue search, denoted by $s_n^2 = d$.

3. Finally, agent n takes an alternative $a_n \in S_n$, where S_n is the set of alternatives agent n has searched (i.e., agent n can only take an alternative she has searched).

The first search is free. The second search involves a cost c_n which is known to agent n before starting her search process. Search costs c_n are i.i.d. across agents and are drawn from a probability distribution over $C := \{\underline{c}, \overline{c}\}$, where $0 < \underline{c} < \alpha\Delta u < \overline{c}$. Let $\beta := \mathbb{P}(c_n = \overline{c})$ and assume $\beta \in (0, 1)$.[2]

The net utility of agent n is given by the difference between the utility of the alternative she takes and the search cost she incurs. That is, at the end of the search process, agent's n net utility is

$$u_n^{a_n} - c_n(|S_n| - 1).$$

Types of Sequential Search Problem. Let θ_n be the type of sequential search problem n. Types θ_n are i.i.d. across search problems and are drawn from a probability distribution over $\Theta := \{\mathrm{A}, \mathrm{B}\}$. Let $\gamma := \mathbb{P}(\theta_n = \mathrm{A})$ and assume $\gamma \in (0, 1]$. The difference among types of sequential search problems is as follows.

- If $\theta_n = \mathrm{A}$, agent n is isolated. Her sequential search problem is exactly as described above.
- If $\theta_n = \mathrm{B}$, agent n has social information. There is a (fictitious) Bayes rational agent n_0 who faces a sequential search problem of type A (i.e., $\mathbb{P}(\theta_{n_0} = \mathrm{A}) = 1$), and: (i) has the same realized utilities as agent n (i.e., $(u^0_{n_0}, u^1_{n_0}) = (u^0_n, u^1_n)$); (ii) has an idiosyncratic search cost c_{n_0} drawn independently of c_n, but from the same distribution. Before engaging in sequential search (as described above), agent n observes the alternative a_{n_0} taken by agent n_0. Agent n, however, neither observes agent n_0's search cost nor observes agent n_0's search decisions.

3 Optimal Decisions

In this section, we characterize the optimal sequential search and choice decisions for each of the two types of search problem.

[2]These parametric assumptions make a sequential search problem of type A non-trivial (see below). Absent these assumptions, in a sequential search problem of type A, an agent would always search both alternatives, or would always search only one alternative, irrespective of her search cost.

3.1 Sequential Search Problem of Type A

First Search Stage. Since the utilities of the two alternatives are i.i.d., agent n is indifferent about which alternative to search first; thus, she decides which alternative to search first uniformly at random. That is,

$$s_n^1 = \frac{1}{2} \circ 0 + \frac{1}{2} \circ 1,$$

where we denote by $\sum_x \xi(x) \circ x$ the mixture that assigns probability $\xi(x)$ to alternative x. Assuming that agents break ties uniformly at random whenever indifferent (at a search or choice stage) captures the idea that agents do not prefer an alternative over the other because of its label and that labels do not convey any information about agents' behavior.

Since the utilities of the two alternatives are i.i.d., we have

$$u_n^{s_n^1} = \begin{cases} \overline{u} & \text{with probability } \alpha \\ \underline{u} & \text{with probability } 1-\alpha \end{cases}.$$

Second Search Stage. Agent n searches the second alternative if the expected gain from doing so is larger than or equal to her search cost, and discontinues her search otherwise. Given the utility of the first alternative searched, $u_n^{s_n^1}$, agent n's expected gain from the second search is

$$V_{\mathrm{A}}\left(u_n^{s_n^1}\right) := \mathbb{E}\Big[\max\left\{u - u_n^{s_n^1}, 0\right\}\Big] = \begin{cases} 0 & \text{if } u_n^{s_n^1} = \overline{u} \\ \alpha\Delta u & \text{if } u_n^{s_n^1} = \underline{u} \end{cases}. \tag{1}$$

Thus, since $0 < \underline{c} < \alpha\Delta u < \overline{c}$, we have the following:

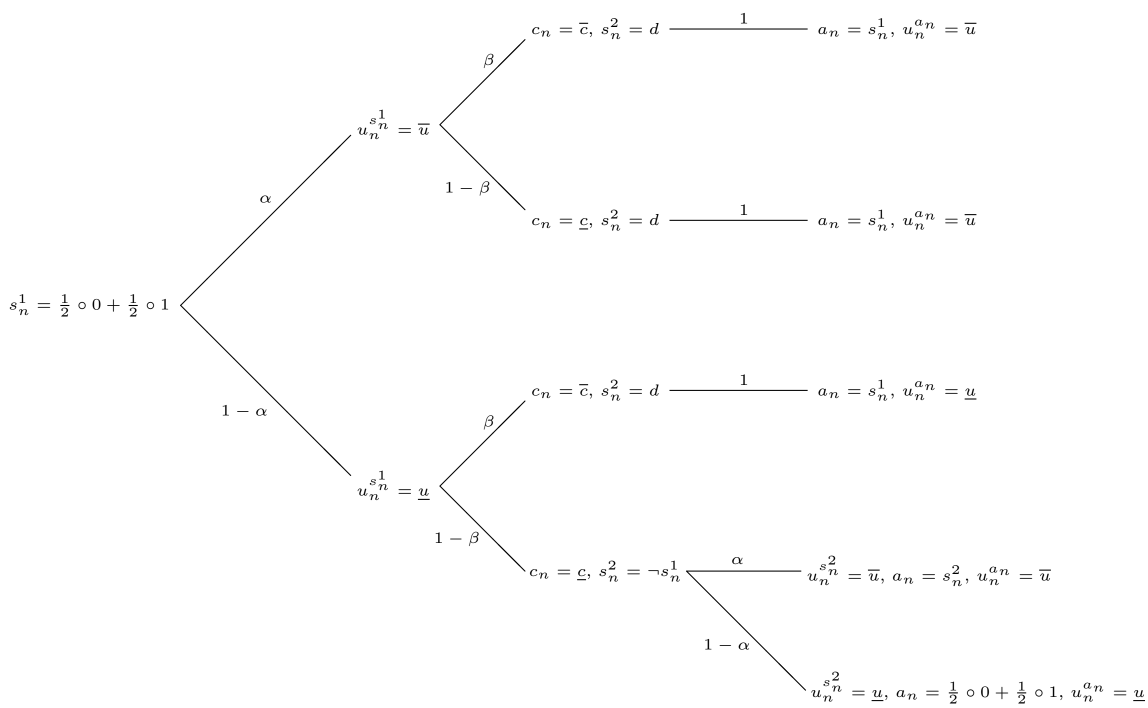
$$s_n^2 = \begin{cases} d & \text{if } u_n^{s_n^1} = \overline{u} \\ d & \text{if } u_n^{s_n^1} = \underline{u} \text{ and } c_n = \overline{c} \\ \neg s_n^1 & \text{if } u_n^{s_n^1} = \underline{u} \text{ and } c_n = \underline{c} \end{cases}.$$

Choice Stage. If agent n searched only one alternative, she takes that alternative; if agent n searched both alternatives, she takes the alternative with the highest utility, randomizing uniformly if the two alternatives have the same utility. That is,

$$a_n = \begin{cases} s_n^1 & \text{if } s_n^2 = d \\ s_n^2 & \text{if } s_n^2 = \neg s_n^1 \text{ and } u_n^{s_n^2} = \overline{u} \\ \frac{1}{2} \circ 0 + \frac{1}{2} \circ 1 & \text{if } s_n^2 = \neg s_n^1 \text{ and } u_n^{s_n^2} = \underline{u} \end{cases}.$$

Decision Tree. Summing up, agent n's decisions in a sequential search problem of type $\theta_n = \text{A}$ are as in the decision tree in Figure 1.

Figure 1: Decision Tree for a Sequential Search Problem of Type $\theta_n = \text{A}$.



3.2 Sequential Search Problem of Type B

First Search Stage. In contrast to a sequential search problem of type A, agent n is not indifferent about which alternative to search first. This is so because agent n's belief about the utilities of the two alternatives is affected by her social information—i.e., by the choice of agent n_0, which is endogenously generated by n_0's optimal decisions in a sequential search problem of type A. In particular, from agent n's viewpoint, there are two possibilities, each of which occurs with positive probability (see the decision tree in Figure 1):

1. Agent n_0 did not search alternative $\neg a_{n_0}$. In this case, agent n_0's choice is uninformative about the utility of alternative $\neg a_{n_0}$.

2. Agent n_0 searched alternative $\neg a_{n_0}$. In this case, since agent n_0 took alternative a_{n_0}, it must be that $u_n^{a_{n_0}} \geq u_n^{\neg a_{n_0}}$ and, with positive probability, $u_n^{a_{n_0}} > u_n^{\neg a_{n_0}}$.

Therefore, agent n's belief about the utility of alternative a_{n_0} strictly first-order stochastically dominates her belief about the utility of alternative $\neg a_{n_0}$. Thus, according to Weitzman (1979)'s optimal search rule, agent n searches alternative a_{n_0} at the first search stage:

$$s_n^1 = a_{n_0}.$$

This is the first difference between a sequential search problem of type A and a sequential search problem of type B.

Then, we have

$$u_n^{s_n^1} = u_{n_0}^{a_{n_0}} = \begin{cases} \overline{u} & \text{with probability } \alpha + \alpha(1-\alpha)(1-\beta) \\ \underline{u} & \text{with probability } \alpha\beta(1-\alpha) + (1-\alpha)^2 \end{cases}, \tag{2}$$

where the probabilities can be calculated from the decision tree in Figure 1.

Second Search Stage. Agent n searches the second alternative if the expected gain from the second search is larger than or equal to her search cost, and discontinues her search otherwise. To compute the expected gain from the second search, agent n must infer the probability that agent n_0 did not search alternative $\neg s_n^1$. Denote by $P(u_n^{s_n^1})$ the probability that alternative $\neg s_n^1$ was not searched by agent n_0 given that an alternative with utility $u_n^{s_n^1}$ was taken. With remaining probability, agent n_0 searched alternative $\neg s_n^1$, but nevertheless chose alternative s_n^1, in which case alternative s_n^1 is non-inferior by revealed preference. Thus, agent n's expected gain from the second search is

$$V_{\mathrm{B}}\left(u_n^{s_n^1}\right) := P\left(u_n^{s_n^1}\right)\mathbb{E}\left[\max\{u - u_n^{s_n^1}, 0\}\right] = P\left(u_n^{s_n^1}\right)V_{\mathrm{A}}\left(u_n^{s_n^1}\right).$$

From Bayes rule and the decision tree in Figure 1, it follows that

$$P(\underline{u}) := \mathbb{P}\left(s_{n_0}^2 = d \mid u_{n_0}^{a_{n_0}} = \underline{u}\right) = \frac{\beta}{\beta + (1-\alpha)(1-\beta)},$$

and so

$$V_{\mathrm{B}}\left(u_n^{s_n^1}\right) = \begin{cases} 0 & \text{if } u_n^{s_n^1} = \overline{u} \\ \frac{\alpha\beta\Delta u}{\beta+(1-\alpha)(1-\beta)} & \text{if } u_n^{s_n^1} = \underline{u} \end{cases}. \tag{3}$$

Conditional on the first searched alternative having utility $\overline{u}$, the expected gain from the second search for an agent with social information is the same as that for an isolated agent (compare equation (1) to equation (3) for $u_n^{s_n^1} = \overline{u}$). In contrast, if $u_n^{s_n^1} = \underline{u}$ the expected gain from the second search for an agent with social information is lower than that for an isolated agent (compare equation (1) to equation (3) for $u_n^{s_n^1} = \underline{u}$). This is the second difference between a sequential search problem of type A and a sequential search problem of type B.

To sum up, we find the following:

- If $\underline{c} < V_{\mathrm{B}}(\underline{u})$,

$$s_n^2 = \begin{cases} d & \text{if } u_n^{s_n^1} = \overline{u} \\ d & \text{if } u_n^{s_n^1} = \underline{u} \text{ and } c_n = \overline{c} \\ \neg s_n^1 & \text{if } u_n^{s_n^1} = \underline{u} \text{ and } c_n = \underline{c} \end{cases};$$

- If $\underline{c} > V_{\mathrm{B}}(\underline{u})$,

$$s_n^2 = d.$$

For simplicity, we ignore the case in which $\underline{c} = V_{\mathrm{B}}(\underline{u})$, as it is non-generic in the parameter space.

Choice Stage. Agent n's optimal decision is as in a sequential search problem of type $\theta_n = \mathrm{A}$.

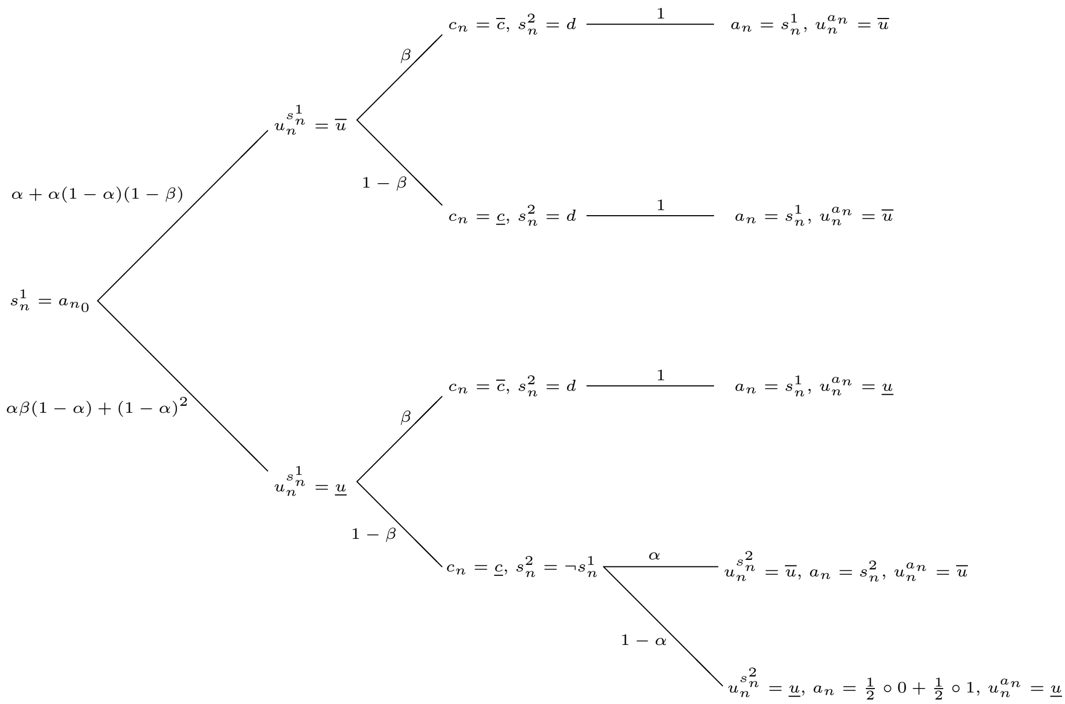
Decision Tree. Summing up, agent n's decisions in a sequential search problem of type $\theta_n = \mathrm{B}$ are as in the decision trees in Figure 2 (if $\underline{c} < V_{\mathrm{B}}(\underline{u})$) and Figure 3 (if $\underline{c} > V_{\mathrm{B}}(\underline{u})$).

Figures 2 and 3 illustrate how the two differences between types of sequential search problem shape an agent's optimal search behavior. First, the two figures show that the probability with which an agent with social information searches first an alternative with utility $u \in U$ is the same as the probability with which an isolated agent takes an alternative with that same utility (see equation (2)). Second, as Figure 3 shows, if $\underline{c} > V_{\mathrm{B}}(\underline{u})$, an agent with social information discontinues search after searching a first alternative with utility $\underline{u}$, whereas an isolated agent would have continued search (see Figure 1). These differences are key to understand when and why a bias arises when search cost distributions are estimated neglecting social information (see the next section).

4 Identification and Estimation

Consider a researcher who knows or can estimate (consistently and without a bias) the utility distribution and the support of the search cost distribution. Given this knowledge, the researcher wants to identify and estimate the search cost

Figure 2: Decision Tree for a Sequential Search Problem of Type $\theta_n = B$ if $\underline{c} < V_{\mathrm{B}}(\underline{u})$.

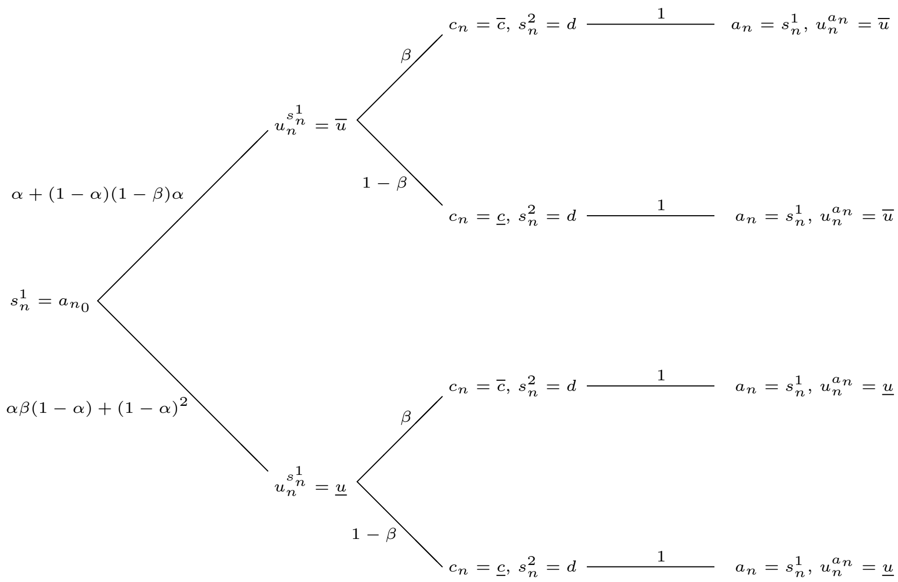


distribution. In our parametric setting, this amounts to identifying and estimating $\beta := \mathbb{P}(c_n = \overline{c})$. Data available to the researcher can come in many forms. We consider the following standard possibilities (in all cases, we consider i.i.d. observations and we denote by N the sample size).

Data on Choice. The researcher observes the share of agents in the sample who take an alternative with utility $\underline{u}$ and the share of agents in the sample who take an alternative with utility $\overline{u}$. Equivalently, the researcher observes the utility of the alternative taken by each agent in the sample, but has no data on individual search behavior.

Data on Search Duration. The researcher observes the share of agents in the sample who conducted only one search and the share of agents in the sample who conducted two searches, without observing the utilities of the searched alternatives. Equivalently, for each agent in the sample, the researcher observes the number of searches conducted, without observing the utilities of the searched alternatives.

Figure 3: Decision Tree for a Sequential Search Problem of Type $\theta_n = B$ if $\underline{c} > V_{\mathrm{B}}(\underline{u})$.



Data on Optimal Stopping. The researcher observes the share of agents in the sample who discontinue search when the utility of the first searched alternative is $\underline{u}$. Equivalently, for each agent in the sample, the researcher observes the utility of the first searched alternative and whether the agent discontinues her search process after the first search.

For each case, we show how the researcher can use the available data to construct an unbiased and consistent estimator of the the parameter β when all agents in the population are isolated (i.e., when $\gamma = 1$). Estimating search cost distributions under the assumption that all agents are isolated is what is routinely done in the empirical literature on search. Next, we show when and why the constructed estimators become biased and inconsistent if a positive share of agents in the population have social in formation (i.e., when $\gamma < 1$); depending on the data available, neglecting social information may lead to under- or over-estimation of search cost distributions.

4.1 Data on Choice

Formally, the researcher observes $\underline{u}_N^a$, where

$$\underline{u}_N^a := \frac{\sum_{n=1}^N \mathbb{1}_{\{u_n^{a_n}=\underline{u}\}}}{N}. \tag{4}$$

By the weak law of large numbers,

$$\underline{u}_N^a \xrightarrow{p} \mathbb{E}[\underline{u}_N^a] = \mathbb{P}(u_n^{a_n} = \underline{u}). \tag{5}$$

Moreover, note that

$$\begin{aligned} \mathbb{P}(u_n^{a_n} = \underline{u}) &= \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})\mathbb{P}(\theta_n = \mathrm{A}) \\ &\quad + \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B})\mathbb{P}(\theta_n = \mathrm{B}) \\ &= \gamma\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) + (1-\gamma)\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}). \end{aligned} \tag{6}$$

From the decision trees in Figures 1–3, we calculate that

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) = \alpha\beta(1-\alpha) + (1-\alpha)^2 \tag{7}$$

and

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) = \begin{cases} \left[\alpha\beta(1-\alpha) + (1-\alpha)^2\right][1-\alpha(1-\beta)] & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ \alpha\beta(1-\alpha) + (1-\alpha)^2 & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u}) \end{cases}. \tag{8}$$

Therefore, from equations (6)–(8), we have

$$\mathbb{P}(u_n^{a_n} = \underline{u}) = \begin{cases} \left[\alpha\beta(1-\alpha) + (1-\alpha)^2\right][1-(1-\gamma)\alpha(1-\beta)] & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ \alpha\beta(1-\alpha) + (1-\alpha)^2 & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u}) \end{cases}. \tag{9}$$

All Agents Are Isolated. Assume that $\gamma = 1$. From equations (5)–(7), we have that the solutions for $\hat{\beta}_N^1$ to equation

$$\alpha\hat{\beta}_N^1(1-\alpha) + (1-\alpha)^2 = \underline{u}_N^a,$$

given by

$$\hat{\beta}_N^1 = \frac{\underline{u}_N^a}{\alpha(1-\alpha)} - \frac{1-\alpha}{\alpha}. \tag{10}$$

is an unbiased and consistent estimator of β.

The Role of Social Information. Next, assume that $\gamma < 1$. Moreover, suppose the researcher neglects social information and estimates the parameter β using the estimator $\hat{\beta}_N^1$ in equation (10).

If $\underline{c} < V_{\mathrm{B}}(\underline{u})$,

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) > \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) \tag{11}$$

(compare equation (7) to equation (8) for $\underline{c} < V_{\mathrm{B}}(\underline{u})$). To understand equation (11), suppose that $\theta_n = \mathrm{B}$. Then, agent n searches first alternative a_{n_0}: $s_n^1 = a_{n_0}$. However, if $\underline{c} < V_{\mathrm{B}}(\underline{u})$, then agent n searches alternative $\neg s_n^1$ (i.e., $s_n^2 = \neg s_n^1$) if $u_n^{s_n^1} = \underline{u}$ and $c_n = \underline{c}$, which occurs with positive probability. Moreover, with positive probability, $u_n^{\neg s_n^1} = \overline{u}$ (because, with positive probability, the two alternatives have different utility and agent n_0 discontinues search even if the utility of the first searched alternative is $\underline{u}$), in which case agent n takes alternative $a_n = s_n^2$ (so that $u_n^{a_n} = \overline{u}$). Therefore, $\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) < \mathbb{P}(u_n^{a_{n_0}} = \underline{u}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})$ (where the last equality holds because $\theta_{n_0} = \mathrm{A}$), and equation (11) follows. As a result, $\mathbb{P}(u_n^{a_n} = \underline{u})$ depends on γ and, in particular

$$\mathbb{P}(u_n^{a_n} = \underline{u}) < \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})$$

(compare equation (7) to equation (9) for $\underline{c} < V_{\mathrm{B}}(\underline{u})$). Therefore, the estimator $\hat{\beta}_N^1$ in equation (10) becomes biased downward and inconsistent. More formally,

$$\hat{\beta}_N^1 \xrightarrow{p} \mathbb{E}\left[\hat{\beta}_N^1\right] = \beta - (1-\gamma)(1-\beta)[1-\alpha(1-\beta)]$$

and

$$\operatorname{Bias}\left(\hat{\beta}_N^1, \beta\right) := \mathbb{E}\left[\hat{\beta}_N^1\right] - \beta = -(1-\gamma)(1-\beta)[1-\alpha(1-\beta)] < 0. \tag{12}$$

That is, search costs are underestimated: the true search cost distribution first order stochastically dominates the estimated search cost distribution.

In contrast, if $\underline{c} > V_{\mathrm{B}}(\underline{u})$, then

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) \tag{13}$$

(compare equation (7) to equation (8) for $\underline{c} > V_{\mathrm{B}}(\underline{u})$). To understand equation (13), suppose that $\theta_n = \mathrm{B}$. Then, agent n searches first alternative a_{n_0}: $s_n^1 = a_{n_0}$. Moreover, if $\underline{c} > V_{\mathrm{B}}(\underline{u})$, then agent n always discontinues search: $s_n^2 = d$. Therefore, $a_n = s_n^1 = a_{n_0}$, and so $\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) = \mathbb{P}(u_n^{a_{n_0}} = \underline{u}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})$, from which equation (13) follows. As a result, $\mathbb{P}(u_n^{a_n} = \underline{u})$ does not depend on γ (see equation (9) for $\underline{c} > V_{\mathrm{B}}(\underline{u})$) and so the estimator $\hat{\beta}_N^1$ in equation (10) remains unbiased and consistent.

4.2 Data on Search Duration

Formally, the researcher observes d_N, where

$$d_N := \frac{\sum_{n=1}^N \mathbb{1}_{\{s_n^2 = d\}}}{N}.$$

By the weak law of large numbers,

$$d_N \xrightarrow{p} \mathbb{E}[d_N] = \mathbb{P}\Big(s_n^2 = d\Big). \tag{14}$$

Moreover, note that

$$\begin{aligned} \mathbb{P}\Big(s_n^2 = d\Big) &= \mathbb{P}\Big(s_n^2 = d \mid \theta_n = \mathrm{A}\Big)\mathbb{P}(\theta_n = \mathrm{A}) + \mathbb{P}\Big(s_n^2 = d \mid \theta_n = \mathrm{B}\Big)\mathbb{P}(\theta_n = \mathrm{B}) \\ &= \mathbb{P}\Big(s_n^2 = d \mid \theta_n = \mathrm{A}\Big)\gamma + \mathbb{P}\Big(s_n^2 = d \mid \theta_n = \mathrm{B}\Big)(1-\gamma). \end{aligned} \tag{15}$$

From the decision trees in Figures 1–3, we calculate that

$$\mathbb{P}\Big(s_n^2 = d \mid \theta_n = \mathrm{A}\Big) = \alpha + (1-\alpha)\beta \tag{16}$$

and

$$\mathbb{P}\Big(s_n^2 = d \mid \theta_n = \mathrm{B}\Big) = \begin{cases} \alpha + (1-\alpha)\beta + \alpha(1-\alpha)(1-\beta)^2 & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ 1 & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u}) \end{cases}. \tag{17}$$

Therefore, from equations (15)–(17), we have

$$\mathbb{P}\Big(s_n^2 = d\Big) = \begin{cases} \alpha + (1-\alpha)\beta + (1-\gamma)\alpha(1-\alpha)(1-\beta)^2 & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ \gamma[\alpha + (1-\alpha)\beta] + (1-\gamma) & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u}) \end{cases}. \tag{18}$$

All Agents Are Isolated. Suppose that $\gamma = 1$. From equations (14)–(16), we have that the solutions for $\hat{\beta}_N^2$ to equation

$$\alpha + (1-\alpha)\hat{\beta}_N^2 = d_N,$$

given by

$$\hat{\beta}_N^2 = \frac{d_N}{1-\alpha} - \frac{\alpha}{1-\alpha}. \tag{19}$$

is an unbiased and consistent estimator of β.

The Role of Social Information. Next, assume that $\gamma < 1$. Moreover, suppose the researcher neglects social information and estimates the parameter β using the estimator $\hat{\beta}_N^2$ defined by (19).

Independently of whether $\underline{c} < V_{\mathrm{B}}(\underline{u})$ or $\underline{c} > V_{\mathrm{B}}(\underline{u})$,

$$\mathbb{P}\left(s_n^2 = d \mid \theta_n = \mathrm{A}\right) < \mathbb{P}\left(s_n^2 = d \mid \theta_n = \mathrm{B}\right). \tag{20}$$

(compare equation (16) to equation (17)). To understand equation (20), suppose that $\theta_n = \mathrm{B}$. Then, agent n searches first alternative a_{n_0}, $s_n^1 = a_{n_0}$, and so $\mathbb{P}\left(s_n^2 = d \mid \theta_n = \mathrm{B}\right) = \mathbb{P}\left(u_{n_0}^{a_{n_0}} = \overline{u}\right)$. Since the utility of the two alternatives is different with positive probability and agent n_0 searches both alternatives with positive probability, $\mathbb{P}\left(u_{n_0}^{a_{n_0}} = \overline{u}\right) > \mathbb{P}\left(u_{n_0}^{s_{n_0}^1} = \overline{u}\right) = \mathbb{P}\left(u_n^{s_n^1} = \overline{u} \mid \theta_n = \mathrm{A}\right)$ (where the last equality holds because $\theta_{n_0} = \mathrm{A}$). Therefore, $\mathbb{P}\left(u_n^{s_n^1} = \overline{u} \mid \theta_n = \mathrm{A}\right) < \mathbb{P}\left(u_n^{s_n^1} = \overline{u} \mid \theta_n = \mathrm{B}\right)$. Since agent n always stops searching if $u_n^{s_n^1} = \overline{u}$, equation (20) follows. As a result, $\mathbb{P}(u_n^{a_n} = \underline{u})$ depends on γ and, in particular,

$$\mathbb{P}\left(s_n^2 = d\right) > \mathbb{P}\left(s_n^2 = d \mid \theta_n = \mathrm{A}\right)$$

(compare equation (16) to equation (18)). Therefore, the estimator $\hat{\beta}_N^2$ in equation (19) becomes biased upward and inconsistent. More formally,

$$\hat{\beta}_N^2 \xrightarrow{p} \mathbb{E}\left[\hat{\beta}_N^2\right] = \begin{cases} \beta + (1-\gamma)\alpha(1-\beta)^2 & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ \gamma\beta + (1-\gamma) & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u}) \end{cases}.$$

and

$$\mathrm{Bias}\left(\hat{\beta}_N^2, \beta\right) := \mathbb{E}\left[\hat{\beta}_N^2\right] - \beta = \begin{cases} (1-\gamma)\alpha(1-\beta)^2 > 0 & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ (1-\gamma)(1-\beta) > 0 & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u}) \end{cases}. \tag{21}$$

That is, search costs are overestimated: the estimated search cost distribution first order stochastically dominates the true search cost distribution. Clearly, the bias is larger when $\underline{c} > V_{\mathrm{B}}(\underline{u})$ because in this case, if $\theta_n = \mathrm{B}$, agent n always discontinues search at the second search stage (and not only when $c_n = \overline{c}$).

4.3 Data on Optimal Stopping

Formally, the researcher observes $d_N^{\underline{u}}$, where

$$d_N^{\underline{u}} := \frac{\sum_{n=1}^N \mathbb{1}_{\{s_n^2 = d\}} \mathbb{1}_{\left\{u_n^{s_n^1} = \underline{u}\right\}}}{\sum_{n=1}^N \mathbb{1}_{\left\{u_n^{s_n^1} = \underline{u}\right\}}}.$$

By the weak law of large numbers,

$$d_N^{\underline{u}} \xrightarrow{p} \mathbb{E}\left[d_N^{\underline{u}}\right] = \mathbb{P}\left(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}\right). \tag{22}$$

Moreover, note that

$$\begin{aligned}\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}\Big) &= \mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{A}\Big)\mathbb{P}\Big(\theta_n = \mathrm{A} \mid u_n^{s_n^1} = \underline{u}\Big) \\ &+ \mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{B}\Big)\mathbb{P}\Big(\theta_n = \mathrm{B} \mid u_n^{s_n^1} = \underline{u}\Big)\end{aligned} \tag{23}$$

From the decision trees in Figures 1–3, we calculate that

$$\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{A}\Big) = \beta, \tag{24}$$

$$\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{B}\Big) = \begin{cases}\beta & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ 1 & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u})\end{cases}, \tag{25}$$

and

$$\begin{aligned}\mathbb{P}\Big(\theta_n = \mathrm{A} \mid u_n^{s_n^1} = \underline{u}\Big) &= \frac{\gamma}{1 - \alpha(1-\beta)(1-\gamma)}, \\ \mathbb{P}\Big(\theta_n = \mathrm{B} \mid u_n^{s_n^1} = \underline{u}\Big) &= \frac{(1-\gamma)[1-\alpha(1-\beta)]}{1 - \alpha(1-\beta)(1-\gamma)}.\end{aligned} \tag{26}$$

Therefore, from equations (23)–(26), we have

$$\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}\Big) = \begin{cases}\beta & \text{if } \underline{c} < V_{\mathrm{B}}(\underline{u}) \\ \beta + \frac{(1-\beta)(1-\gamma)[1-\alpha(1-\beta)]}{1-\alpha(1-\beta)(1-\gamma)} & \text{if } \underline{c} > V_{\mathrm{B}}(\underline{u})\end{cases}. \tag{27}$$

All Agents Are Isolated. Assume that $\gamma = 1$. From equations (22)–(24), we have that

$$\hat{\beta}_N^3 = d_N^{\underline{u}} \tag{28}$$

is an unbiased and consistent estimator of β.

The Role of Social Information. Next, assume that $\gamma < 1$. Moreover, suppose the researcher neglects social information and estimates the parameter β using the estimator $\hat{\beta}_N^3$ in equation (28).

If $\underline{c} < V_{\mathrm{B}}(\underline{u})$, then

$$\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{A}\Big) = \mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{B}\Big) \tag{29}$$

(compare equation (24) to equation (25) for $\underline{c} < V_{\mathrm{B}}(\underline{u})$). The equality in equation (29) immediately follows by observing that, conditional on the first searched alternative having utility $\underline{u}$ (i.e., $u_n^{s_n^1} = \underline{u}$), agent n discontinues search if and only if $c_n = \bar{c}$ independently of whether $\theta_n = \mathrm{A}$ or $\theta_n = \mathrm{B}$. Thus, $\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}\Big)$ does not depend on γ (see equation (27) for $\underline{c} < V_{\mathrm{B}}(\underline{u})$) and so the estimator $\hat{\beta}_N^3$ in equation (28) remains unbiased and consistent.

In contrast, if $\underline{c} > V_{\mathrm{B}}(\underline{u})$, then

$$\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{A}\Big) < \mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{B}\Big) \tag{30}$$

(compare equation (24) to equation (25) for $\underline{c} > V_{\mathrm{B}}(\underline{u})$). The inequality in equation (30) immediately follows by observing that, conditional on $u_n^{s_n^1} = \underline{u}$, agent n discontinues search if and only if $c_n = \overline{c}$ when $\theta_n = \mathrm{A}$ and always discontinues search if $\theta_n = \mathrm{B}$. Thus, $\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}\Big)$ depends on γ and, in particular,

$$\mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}\Big) > \mathbb{P}\Big(s_n^2 = d \mid u_n^{s_n^1} = \underline{u}, \theta_n = \mathrm{A}\Big)$$

(compare equation (24) to equation (27) for $\underline{c} > V_{\mathrm{B}}(\underline{u})$). Therefore, the estimator $\hat{\beta}_N^3$ in equation (28) becomes biased upward and inconsistent. More formally,

$$\hat{\beta}_N^3 \xrightarrow{p} \mathbb{E}\Big[\hat{\beta}_N^3\Big] = \beta + \frac{(1-\beta)(1-\gamma)[1-\alpha(1-\beta)]}{1-\alpha(1-\beta)(1-\gamma)}$$

and

$$\mathrm{Bias}\Big(\hat{\beta}_N^3, \beta\Big) := \mathbb{E}\Big[\hat{\beta}_N^3\Big] - \beta = \frac{(1-\beta)(1-\gamma)[1-\alpha(1-\beta)]}{1-\alpha(1-\beta)(1-\gamma)} > 0. \tag{31}$$

That is, search costs are overestimated: the estimated search cost distribution first order stochastically dominates the true search cost distribution.

5 Potential Remedies

Table 1 sums up the results presented in Section 4. In particular, for each of the possible datasets we consider, the table illustrates whether and when neglecting social information leads to biased and inconsistent estimates of search cost distributions and, when so, it also illustrates the bias direction. When present, equations (12), (21), and (31) show that the bias increases as the share of agents in the population with social information grows large (i.e., as γ decreases).

Table 1: Summary of Identification and Estimation in a Sequential Search Model.

	Estimation Bias		
	Data on Choice	Data on Search Duration	Data on Optimal Stopping
$\underline{c} < V_{\mathrm{B}}(\underline{u})$	< 0	> 0	NO
$\underline{c} > V_{\mathrm{B}}(\underline{u})$	NO	> 0	> 0

Given her knowledge about the environment, the researcher may be willing to make assumptions about the model's primitives that are useful to recover correct estimates of search cost distributions. For instance, many (theoretical) search models (see, e.g., Varian, 1980; Stahl, 1989; Ellison and Wolitzky, 2012) assume that a share of agents in the population have zero (or negligible) search cost; thus, such agents know the utilities of all alternatives. In the context of our model, such assumption would correspond to assume that $\underline{c} < V_{\mathrm{B}}(\underline{u})$ and estimating $\beta \coloneqq \mathbb{P}(c_n = \overline{c})$ would amount to estimating the share of agents in the population for which search is truly costly. If the researcher is willing to assume that $\underline{c} < V_{\mathrm{B}}(\underline{u})$, then our analysis shows that an unbiased and consistent estimate of β can be obtained if data on agents' optimal stopping decisions are available (see Table 1).

Similarly, given her knowledge about the environment, the researcher may be willing to assume that search costs are "high" for all agents in the population.[3] In the context of our model, such assumption would correspond to assume that $\underline{c} > V_{\mathrm{B}}(\underline{u})$. If the researcher is willing to assume that $\underline{c} > V_{\mathrm{B}}(\underline{u})$, then our analysis shows that an unbiased and consistent estimate of β can be obtained if data on agents' choices are available (see Table 1).

In most circumstances, however, it may be hard to make either of such assumptions ($\underline{c} < V_{\mathrm{B}}(\underline{u})$ *vs.* $\underline{c} > V_{\mathrm{B}}(\underline{u})$). Moreover, even if the researcher is willing to make either of such assumptions, the data needed to recover unbiased and consistent estimates of β may not be available. Therefore, alternative solutions are called for.

In the rest of this section, we discuss possible approaches to recover an unbiased and consistent estimator of β. For ease of exposition, we focus on data on agents' choices. With different datasets, analogous reasoning applies.

5.1 Estimating γ Offline

Suppose the researcher can estimate γ "offline" e.g., by using detailed network data. In this section, we study when such information can be used to correctly identify and estimated the parameter of interest.

Let $\hat{\gamma} \in (0, 1)$ denote an unbiased and consistent estimator of γ. Let $\underline{\beta}_N^1$ and $\overline{\beta}_N^1$ be defined as the solutions to the following equations:

$$\alpha \underline{\beta}_N^1 (1-\alpha) + (1-\alpha)^2 = \underline{u}_N^a \tag{32}$$

$$\left[\alpha \overline{\beta}_N^1 (1-\alpha) + (1-\alpha)^2\right]\left[1-(1-\hat{\gamma})\alpha\left(1-\overline{\beta}_N^1\right)\right] = \underline{u}_N^a, \tag{33}$$

[3]However, this assumption seems less common in the search literature.

(for $\overline{\beta}_N^1$, consider the real solution in the interval $(0, 1)$). By equations (5) and (9), either $\underline{\beta}_N^1$ or $\overline{\beta}_N^1$ is an unbiased and consistent estimator of β, but it remains to determine which one.

By equations (32) and (33), we have $\underline{\beta}_N^1 < \overline{\beta}_N^1$ (for more details, recall the analysis in Section 4.1). Moreover, let $\underline{V}_{\mathrm{B}}(\underline{u})$ (resp., $\overline{V}_{\mathrm{B}}(\overline{u})$) be the value of $V_{\mathrm{B}}(\underline{u})$ defined by (3) evaluated at $\underline{\beta}_N^1$ (resp., $\overline{\beta}_N^1$). Since either $\underline{\beta}_N^1$ or $\overline{\beta}_N^1$ is an unbiased and consistent estimator of β, either $\underline{V}_{\mathrm{B}}(\underline{u})$ or $\overline{V}_{\mathrm{B}}(\overline{u})$ is an unbiased and consistent estimator of the true $V_{\mathrm{B}}(\underline{u})$ in the data generating process. Moreover, as $V_{\mathrm{B}}(\underline{u})$ is increasing in β, we have $\underline{V}_{\mathrm{B}}(\underline{u}) < \overline{V}_{\mathrm{B}}(\underline{u})$. From these observations, together with Table 1, we conclude the following:

- If $\underline{c} < \underline{V}_{\mathrm{B}}(\underline{u})$, then $\overline{\beta}_N^1$ is an unbiased and consistent estimator of β.
- If $\underline{c} > \overline{V}_{\mathrm{B}}(\underline{u})$, then $\underline{\beta}_N^1$ is an unbiased and consistent estimator of β.
- If $\underline{V}_{\mathrm{B}}(\underline{u}) < \underline{c} < \overline{V}_{\mathrm{B}}(\underline{u})$, it is not possible to detect with the available data whether $\underline{\beta}_N^1$ or $\overline{\beta}_N^1$ is an unbiased and consistent estimator of β (but one of the two is), and another approach is required.

Example. We illustrate these ideas with a numerical example. Suppose that

$$\alpha = \frac{1}{2}, \qquad \Delta u = 5, \qquad \hat{\gamma} = \frac{1}{2}, \qquad \underline{u}_N^a = \frac{5}{16}.$$

With this numerical specification, we have

$$\underline{\beta}_N^1 = \frac{1}{4}, \qquad \overline{\beta}_N^1 = \sqrt{6} - 2, \qquad \underline{V}_{\mathrm{B}}(\underline{u}) = 1, \qquad \overline{V}_{\mathrm{B}}(\underline{u}) = 4 - \sqrt{6}.$$

If $\underline{c} < 1$, the researcher uses $\overline{\beta}_N^1 = \sqrt{6} - 2$ as an estimate of β; $\underline{c} > 4 - \sqrt{6}$, the researcher uses $\underline{\beta}_N^1 = \frac{1}{4}$ as an estimate of β; if $1 < \underline{c} < 4 - \sqrt{6}$, then it is not possible to decide whether to use $\underline{\beta}_N^1$ or $\overline{\beta}_N^1$ as an estimate of β based on the data alone.

5.2 Shifts to the Utility Distribution

In this section, we ask whether shifts in the parameters determining the (indirect) utility distribution can help identify the parameter of interest.[4]

Suppose the researcher observes a shift in the value of the high utility $\overline{u}$ to $\overline{u}' > \overline{u}$.[5] Let $\underline{u}_N^a$ (resp., $\underline{u}_N^{a\prime}$) be the sample moment defined by (4) calculated using the dataset with $\overline{u}$ (resp., $\overline{u}'$). Moreover, let $V_{\mathrm{B}}'(\underline{u})$ be the value of $V_{\mathrm{B}}(\underline{u})$ defined by

[4] An important identification question in the empirical search literature is whether changes in demand originate from shifts in utility (Honka et al., 2019).

[5] The case in which the researcher observes a shift in the value of the low utility $\underline{u}$ to $\underline{u}' < \underline{u}$ can be treated analogously, with the obvious changes.

(3) evaluated at $\overline{u}'$. Finally, let $\mathbb{P}_{\overline{u}}(u_n^{a_n} = \underline{u})$ (resp., $\mathbb{P}_{\overline{u}'}(u_n^{a_n} = \underline{u})$) be the theoretical prediction defined by (9) when the high utility is $\overline{u}$ (resp., $\overline{u}'$).

Since $V_{\mathrm{B}}(\underline{u})$ is increasing in $\overline{u}$, we have $V_{\mathrm{B}}(\underline{u}) < V_{\mathrm{B}}'(\underline{u})$. Suppose that the shift in the high utility is such that $V_{\mathrm{B}}(\underline{u}) < \underline{c} < V_{\mathrm{B}}'(\underline{u})$. That is, before the shift, an agent with social information and search cost $\underline{c}$ finds it optimal to discontinue search if the utility of the first searched alternative is $\underline{u}$; in contrast, after the shift, an agent with social information and search cost $\underline{c}$ finds it optimal to search the second alternative if the utility of the first searched alternative is $\underline{u}$. Under these conditions, $\mathbb{P}_{\overline{u}}(u_n^{a_n} = \underline{u}) > \mathbb{P}_{\overline{u}'}(u_n^{a_n} = \underline{u})$. Although $\mathbb{P}_{\overline{u}}(u_n^{a_n} = \underline{u})$ and $\mathbb{P}_{\overline{u}'}(u_n^{a_n} = \underline{u})$ are not observable in the data, we know that, by the weak law of large numbers,

$$\underline{u}_N^a \xrightarrow{p} \mathbb{P}_{\overline{u}}(u_n^{a_n} = \underline{u}) \qquad \text{and} \qquad \underline{u}_N^{a\,\prime} \xrightarrow{p} \mathbb{P}_{\overline{u}'}(u_n^{a_n} = \underline{u}),$$

and so, as $N \to \infty$, we will have $\underline{u}_N^a > \underline{u}_N^{a\,\prime}$ almost surely. As a result, in a sufficiently large sample, the difference $\underline{u}_N^a - \underline{u}_N^{a\,\prime}$ is larger than zero in a statistically significant way at any desired level. By equation (9), it follows that $\hat{\beta}_1$ defined by

$$\alpha\hat{\beta}_1(1-\alpha) + (1-\alpha)^2 = \underline{u}_N^a \tag{34}$$

(i.e., $\hat{\beta}_1$ in equation (10)) can be used as an unbiased and consistent estimators of β.

In contrast, if the shift in the high utility is such that $\underline{c} < V_{\mathrm{B}}(\underline{u}) < V_{\mathrm{B}}'(\underline{u})$ or $V_{\mathrm{B}}(\underline{u}) < V_{\mathrm{B}}'(\underline{u}) < \underline{c}$, then the model's theoretical predictions remain the same before and after the shift. As a result, $\underline{u}_N^a - \underline{u}_N^{a\,\prime} \xrightarrow{p} 0$ and so it is not possible to detect with the available data which is an unbiased and consistent estimator of β.

Example. We illustrate these ideas with a numerical example. Suppose that

$$\alpha = \frac{1}{2}, \qquad \underline{u} = 2, \qquad \overline{u} = 8, \qquad \overline{u}' = 14, \qquad \underline{u}_N^a = \frac{3}{8}, \qquad \underline{u}_N^{a\,\prime} = \frac{9}{64}.$$

Note that $\underline{u}_N^a - \underline{u}_N^{a\,\prime} = \frac{15}{64}$ is larger than zero in a statistically significant way at any desired level if N is sufficiently large. Then, by solving (34) for $\hat{\beta}_1$ with $\alpha = \frac{1}{2}$ and $\underline{u}_N^a = \frac{3}{8}$ the researcher obtains

$$\hat{\beta}_1 = \frac{\underline{u}_N^a}{\alpha(1-\alpha)} - \frac{1-\alpha}{\alpha} = \frac{\frac{3}{8}}{\frac{1}{4}} - 1 = \frac{1}{2}$$

as an estimate of β.

5.3 Partial Identification

The previous remedies may not be effective either because of data limitations or because the conditions under which they are useful to find an unbiased and consistent point estimator of β are not satisfied. If so, the researcher can rely on a partial identification approach, which we discuss in this section.

Set Identification of $\boldsymbol{\beta}$. The first set estimator, which we denote by $B^{\beta}(\underline{u}_N^a)$, consists of all $\beta \in (0,1)$ which are compatible with the observed empirical moment $\underline{u}_N^a$ as a model prediction for some $\gamma \in (0,1]$. Formally, define

$$f(\alpha,\beta,\gamma) := \Big[\alpha\beta(1-\alpha) + (1-\alpha)^2\Big][1-(1-\gamma)\alpha(1-\beta)].$$

Then, using equations (5) and (9), we have

$$B^{\beta}(\underline{u}_N^a) := \Big\{\beta \in (0,1) : f(\alpha,\beta,\gamma) = \underline{u}_N^a \text{ for some } \gamma \in (0,1]\Big\}.$$

Joint Set Identification of $(\boldsymbol{\beta},\boldsymbol{\gamma})$. The second set estimator, which we denote by $B^{\beta,\gamma}(\underline{u}_N^a)$, consists of all $(\beta,\gamma) \in (0,1) \times (0,1]$ which are compatible with the observed empirical moment $\underline{u}_N^a$ as a model prediction. Formally, using equations (5) and (9), we have

$$B^{\beta,\gamma}(\underline{u}_N^a) := \Big\{(\beta,\gamma) \in (0,1) \times (0,1] : f(\alpha,\beta,\gamma) = \underline{u}_N^a\Big\}.$$

Example. We illustrate these ideas with a numerical example. Suppose that

$$\alpha = \frac{1}{2}, \qquad \underline{u}_N^a = \frac{5}{16},$$

as in Section 4.1. In this case,

$$B^{\beta}(\underline{u}_N^a) = \left[\frac{1}{4}, \frac{1}{2}\Big(\sqrt{10}-2\Big)\right),$$

and

$$B^{\beta,\gamma}(\underline{u}_N^a) = \left\{(\beta,\gamma) \in (0,1)\times(0,1] : \beta \in \left[\frac{1}{4}, \frac{1}{2}\Big(\sqrt{10}-2\Big)\right) \text{ and } \gamma = \frac{3-2\beta(2+\beta)}{2(1-\beta^2)}\right\}.$$

6 Conclusion

Search frictions are an important determinant of consumer choices, pricing behavior, and market outcomes. Therefore, quantifying search frictions in markets is a central empirical exercise to assess the welfare consequences of policy interventions. In modern interconnected societies, individuals seldom search in isolation. Others' choices and experiences are readily available via direct observations, communication, online social networks, popularity rankings, and product reviews. In this paper, we propose a theoretical model to explore how social information affects individual search behavior and the resulting observable outcomes—such as choice, search duration, and information acquisition decisions—that are used to estimate empirical search models. In particular, we extend the canonical empirical sequential and simultaneous search models by allowing a share of the agents in the population to observe the choice of one of their social connections. Our main insight is that neglecting social information typically leads to biased and inconsistent estimates of search cost distributions. The bias direction and magnitude depend on the dataset's content and the model's primitives. When present, the bias increases as the share of agents in the population with social information grows large. We also discuss possible empirical strategies to recover correct estimates of search cost distributions.

We conclude by highlighting a possible direction for future research. In our setting, agents are either isolated or observe the alternative taken by an isolated agent. However, the connection structure in real social networks is complex and so are communication channels and the sources of informational externalities among agents. Hence, it is hard to specify the exact content of agents' social (and non-social) information. A researcher may want to estimate structural parameters in empirical search models by making only week assumptions on agents' information. This calls for a robust empirical approach based on the theoretical literature on robust predictions in incomplete information and/or extensive form games (see, e.g., Bergemann and Morris, 2016; Doval and Ely, 2020).[6] We plan to explore these ideas in future work.

[6]A growing number of empirical papers use the Bayes correlated equilibrium notion of Bergemann and Morris (2016) to develop informational robust identification and counterfactuals in entry games (Magnolfi and Roncoroni, 2021), auctions (Syrgkanis, Tamer and Ziani, 2021), and a single agent agent discrete choice model of voting (Gualdani and Sinha, 2020).

A Simultaneous Search

A.1 Model

In this section, we present our simultaneous search model. The baseline setting is a version of the canonical simultaneous search model by Stigler (1961) as developed in the empirical literature on search (see Section 2 in Honka et al., 2019), from which it borrows all the main assumptions. We then augment the baseline setting by introducing social information.

There is a countably infinite set of search problems, indexed by $n \in \mathbb{N} := \{1, 2, \dots\}$. In search problem n, a Bayes-rational agent—called agent n—must select an alternative from the set $X := \{0, 1\}$. We denote by x a typical element of X and by $\neg x$ the alternative in X other than x. Let u_n^x denote agent n's (indirect) utility from alternative x. Utilities u_n^x are i.i.d. draws—across alternatives within search problem n and across search problems—from a probability distribution over $U := \{\underline{u}, \overline{u}\}$, where $0 \le \underline{u} < \overline{u}$. Let $\alpha := \mathbb{P}(u_n^x = \overline{u})$, and assume $\alpha \in (0, 1)$. Moreover, let $\Delta u := \overline{u} - \underline{u}$

Agent n wishes to select the alternative with the highest realized utility. Agent n knows the true utility distribution, but not the specific pair of realized utilities (u_n^0, u_n^1). Agent n collects information about realized utilities via costly simultaneous search. Formally, agent n takes the following decisions.

1. Agent n commits to search a fixed set of alternatives $S_n \in 2^X \setminus \{\emptyset\}$, where 2^X denotes the power set of the set X. By searching alternative x, agent n perfectly learns its realized utility u_n^x.
2. After actually searching the alternatives in the set S_n, agent n takes an alternative $a_n \in S_n$ (i.e., agent n can only take an alternative she has searched).

Searching one alternative is free, while searching both alternatives involves a cost c_n. Search costs c_n are i.i.d. across agents and are drawn from a probability distribution over $C := \{\underline{c}, \overline{c}\}$, where $0 \le \underline{c} < \alpha(1 - \alpha)\Delta u < \overline{c}$. Let $\beta := \mathbb{P}(c_n = \overline{c})$ and assume $\beta \in (0, 1)$.[7]

The net utility of agent n, denoted by U_n, is given by the difference between the utility of the alternative she takes and the search cost she incurs. That is, at the end of the search process, agent's n net utility is: $U_n := u_n^{a_n} - c_n(|S_n| - 1)$.

[7]These parametric assumptions make a simultaneous search problem of type A non-trivial (see below). Absent these assumptions, in a simultaneous search problem of type A, an agent would always commit to search both alternatives, or would always commit to search only one alternative, irrespective of her search cost.

Types of Simultaneous Search Problem. Let θ_n be the type of simultaneous search problem n. Types θ_n are i.i.d. across search problems and are drawn from a probability distribution over $\Theta := \{\mathrm{A}, \mathrm{B}\}$. Let $\gamma := \mathbb{P}(\theta_n = \mathrm{A})$ and assume $\gamma \in (0, 1]$. The difference among types of simultaneous search problems is as follows.

- If $\theta_n = \mathrm{A}$, agent n is isolated. Her simultaneous search problem is exactly as described above.
- If $\theta_n = \mathrm{B}$, agent n has social information. There is a (fictitious) Bayes rational agent n_0 who faces a simultaneous search problem of type A, and: (i) has the same realized utilities as agent n (i.e., $(u^0_{n_0}, u^1_{n_0}) = (u^0_n, u^1_n)$); (ii) has an idiosyncratic search cost c_{n_0} drawn independently of c_n, but from the same distribution. Before engaging in simultaneous search (as described above), agent n observes the alternative a_{n_0} taken by agent n_0. Agent n, however, neither observes agent n_0's search cost nor observes agent n_0's decisions of how many and which alternatives to search.

A.2 Optimal Decisions

In this section, we characterize the optimal simultaneous search and choice decisions for each of the two types of search problem.

A.2.1 Simultaneous Search Problem of Type A

Search Stage. Agent n's search problem consists of committing to search a fixed set of alternatives $S_n \in 2^X \setminus \{\emptyset\}$ that maximizes the expected utility to agent n from searching that set of alternatives net of search costs. That is,

$$S_n \in \operatorname*{arg\,max}_{S \in 2^X \setminus \{\emptyset\}} := [V_\mathrm{A}(S) - c_n(|S| - 1)],$$

where, for all $S \in 2^X \setminus \{\emptyset\}$,

$$V_\mathrm{A}(S) := \mathbb{E}\left[\max_{x \in S}\{u^x_n\}\right].$$

Clearly,

$$V_\mathrm{A}(S) = \begin{cases} \underline{u} + \alpha\Delta u & \text{if } S = \{0\} \text{ or } S = \{1\} \\ \overline{u} - (1-\alpha)^2\Delta u & \text{if } S = X \end{cases}. \tag{35}$$

Note that, for all $x \in X$,

$$V_\mathrm{A}(X) - c_n > V_\mathrm{A}(\{x\}) \iff c_n < \alpha(1-\alpha)\Delta u.$$

Thus, since $0 < \underline{c} < \alpha(1-\alpha)\Delta u < \overline{c}$, we have the following:

$$S_n = \begin{cases} \frac{1}{2} \circ \{0\} + \frac{1}{2} \circ \{1\} & \text{if } c_n = \overline{c} \\ X & \text{if } c_n = \underline{c} \end{cases},$$

where we denote by $\sum_x \xi(x) \circ \{x\}$ the mixture that assigns probability $\xi(x)$ to set $\{x\}$. Hereafter, we denote by s_n the alternative in S_n whenever $|S_n| = 1$.

Since the utilities of the two alternatives are i.i.d., we have:

- If $S_n = \frac{1}{2} \circ \{0\} + \frac{1}{2} \circ \{1\}$,

$$u_n^{s_n} = \begin{cases} \overline{u} & \text{with probability } \alpha \\ \underline{u} & \text{with probability } 1-\alpha \end{cases},$$

- If $S_n = X$,

$$\begin{cases} u_n^0 = u_n^1 = \overline{u} & \text{with probability } \alpha^2 \\ u_n^x = \overline{u} > \underline{u} = u_n^{\neg x} & \text{with probability } 2\alpha(1-\alpha) \\ u_n^0 = u_n^1 = \underline{u} & \text{with probability } (1-\alpha)^2 \end{cases}.$$

Choice Stage. If agent n searched only one alternative, she takes that alternative; if agent n searched both alternatives, she takes the alternative with the highest utility, randomizing uniformly if the two alternatives have the same utility. That is,

$$a_n = \begin{cases} s_n & \text{if } S_n = \frac{1}{2} \circ \{0\} + \frac{1}{2} \circ \{1\} \\ x & \text{if } S_n = X \text{ and } u_n^x > u_n^{\neg x} \\ \frac{1}{2} \circ 0 + \frac{1}{2} \circ 1 & \text{if } S_n = X \text{ and } u_n^x = u_n^{\neg x} \end{cases}.$$

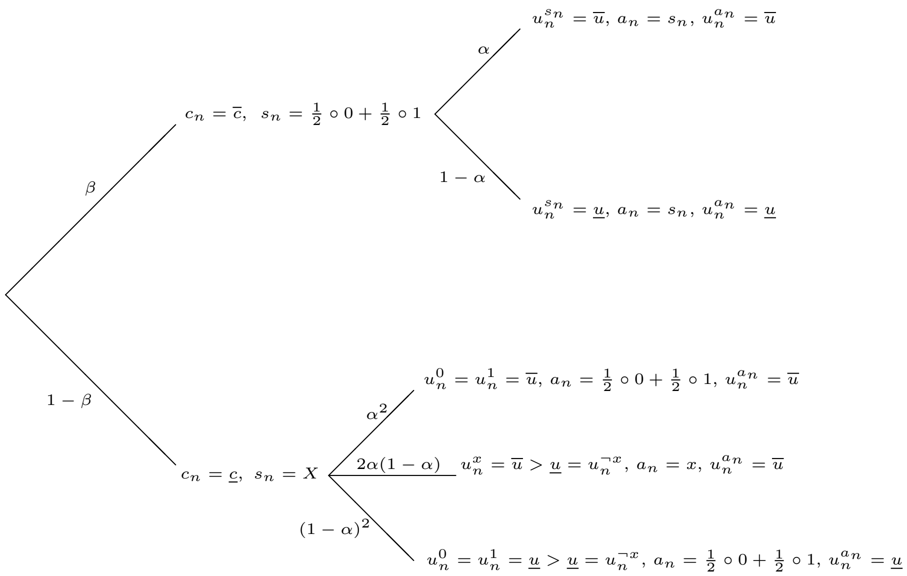
Decision Tree. Summing up, agent n's decisions in a simultaneous search problem of type $\theta_n = \text{A}$ are as in the decision tree in Figure 4.

A.2.2 Simultaneous Search Problem of Type B

Search Stage. As in a simultaneous search problem of type A, agent n's search problem consists of committing to search a fixed set of alternatives $S_n \in 2^X \setminus \{\emptyset\}$ that maximizes the expected utility to agent n from searching that set of alternatives net of search costs. That is,

$$S_n \in \underset{S \in 2^X \setminus \{\emptyset\}}{\arg\max} := [V_{\text{B}}(S) - c_n(|S| - 1)],$$

Figure 4: Decision Tree for a Simultaneous Search Problem of Type $\theta_n = \mathrm{A}$.



where, for all $S \in 2^X \setminus \{\emptyset\}$,

$$V_{\mathrm{B}}(S) \coloneqq \mathbb{E}\left[\max_{x \in S} \{u_n^x\} \mid a_{n_0} \right]. \tag{36}$$

In contrast to a simultaneous search problem of type A, if agent n decides to commit to search only one alternative, she is not indifferent about which alternative to search. This is so because agent n's belief about the utilities of the two alternatives is affected by her social information—i.e., by the choice of agent n_0, which is endogenously generated by n_0's optimal decisions in a simultaneous search problem of type A (note that the expectation in (36) is conditional on a_{n_0}). In particular, from agent n's viewpoint, there are two possibilities, each of which occurs with positive probability (see the decision tree in Figure 1):

1. Agent n_0 searched only one alternative alternative. In this case, agent n_0's choice is uninformative about the utility of alternative $\neg a_{n_0}$.

2. Agent n_0 searched both alternative. In this case, since agent n_0 took alternative a_{n_0}, it must be that $u_n^{a_{n_0}} \geq u_n^{\neg a_{n_0}}$ and, with positive probability $u_n^{a_{n_0}} > u_n^{\neg a_{n_0}}$.

Therefore, agent n's belief about the utility of alternative a_{n_0} strictly first-order stochastically dominates her belief about the utility of alternative $\neg u_{n_0}$, and so, $\mathbb{E}\left[u_n^{a_{n_0}} \mid a_{n_0}\right] > \mathbb{E}\left[u_n^{\neg a_{n_0}} \mid a_{n_0}\right]$. Thus, if agent n commits to search only one alternative, she will search alternative a_{n_0}. This is the first difference between a simultaneous search problem of type A and a simultaneous search problem of type B.

Note that

$$u_{n_0}^{a_{n_0}} = \begin{cases} \overline{u} & \text{with probability } \alpha + \alpha(1-\alpha)(1-\beta) \\ \underline{u} & \text{with probability } \alpha\beta(1-\alpha) + (1-\alpha)^2 \end{cases},$$

where the probabilities can be calculated from the decision tree in Figure 4. Thus, in a simultaneous search problem of type B, we have

$$V_{\mathrm{B}}(S) = \begin{cases} \beta[\underline{u} + \alpha\Delta u] + (1-\beta)\left[\overline{u} - (1-\alpha)^2\Delta u\right] & \text{if } S = \{a_{n_0}\} \\ \overline{u} - (1-\alpha)^2\Delta u & \text{if } S = X \end{cases}. \tag{37}$$

Note that the probability with which an agent with social information takes an alternative with utility $u \in U$ by committing to search only one alternative is the same as the probability with which an isolated, who may commit to search one or two alternatives, takes an alternative with that same utility. Moreover, note that $V_{\mathrm{B}}(\{a_{n_0}\}) > V_{\mathrm{A}}(\{x\})$ for all $x \in X$ (compare equation (35) to equation (37) when S is a singleton). That is, the value of committing to search only one alternative (namely, alternative a_{n_0}) for an agent with social information is larger than that for an isolated agent. This is the second difference between a simultaneous search problem of type A and a simultaneous search problem of type B.

Since

$$V_{\mathrm{B}}(X) - c_n > V_{\mathrm{B}}(\{a_{n_0}\}) \iff c_n < V_{\mathrm{B}}(X) - V_{\mathrm{B}}(\{a_{n_0}\}) = \alpha(1-\alpha)\beta\Delta u,$$

by defining $\Delta(V_{\mathrm{B}}) := V_{\mathrm{B}}(X) - V_{\mathrm{B}}(\{a_{n_0}\})$, we have the following:

- If $\underline{c} < \Delta(V_{\mathrm{B}})$,

$$S_n = \begin{cases} \{a_{n_0}\} & \text{if } c_n = \overline{c} \\ X & \text{if } c_n = \underline{c} \end{cases},$$

- If $\underline{c} > \Delta(V_{\mathrm{B}})$,

$$S_n = \{a_{n_0}\}.$$

For simplicity, we ignore the case in which $\underline{c} = \Delta(V_{\mathrm{B}})$, as it is non-generic in the parameter space.

Now we have:

- If $S_n = \{a_{n0}\}$

$$u_n^{s_n} = u_{n0}^{a_{n0}};$$

- If $S_n = X$, as in a simultaneous search problem of type A,

$$\begin{cases} u_n^0 = u_n^1 = \overline{u} & \text{with probability } \alpha^2 \\ u_n^x = \overline{u} > \underline{u} = u_n^{\neg x} & \text{with probability } 2\alpha(1-\alpha) \\ u_n^0 = u_n^1 = \underline{u} & \text{with probability } (1-\alpha)^2 \end{cases}.$$

Choice Stage. If agent n searched only one alternative, she takes that alternative; if agent n searched both alternatives, she takes the alternative with the highest utility, randomizing uniformly if the two alternatives have the same utility. That is,

$$a_n = \begin{cases} s_n & \text{if } S_n = \{a_{n0}\} \\ x & \text{if } S_n = X \text{ and } u_n^x > u_n^{\neg x} \\ \frac{1}{2} \circ 0 + \frac{1}{2} \circ 1 & \text{if } S_n = X \text{ and } u_n^x = u_n^{\neg x} \end{cases}.$$

Decision Tree. Summing up, agent n's decisions in a simultaneous search problem of type $\theta_n = \text{B}$ are as in the decision trees in Figure 5 (if $\underline{c} < \Delta(V_\text{B})$) and Figure 6 (if $\underline{c} > \Delta(V_\text{B})$).

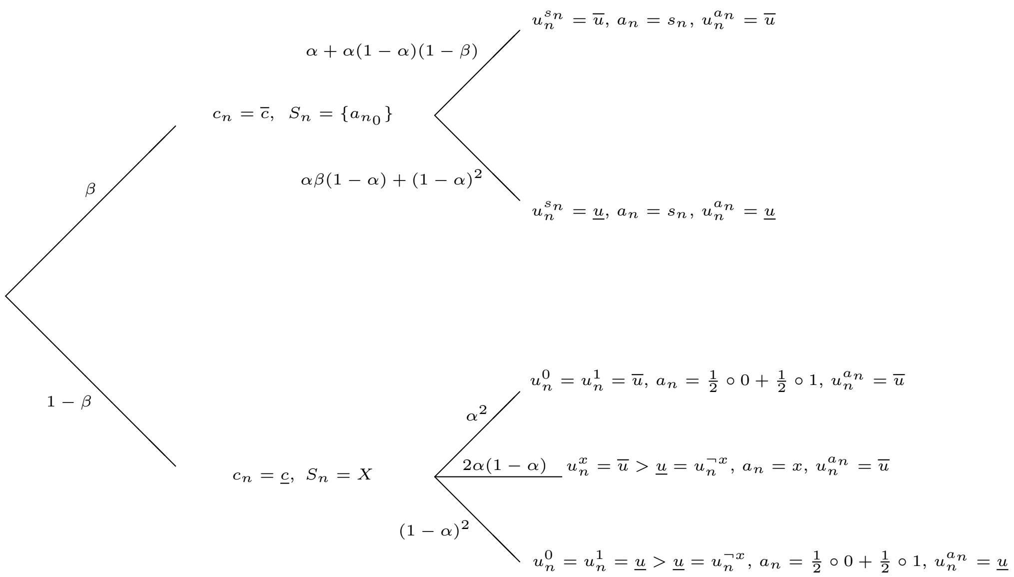
A.3 Identification and Estimation

Consider a researcher who knows (or can estimate) the utility distribution and the support of the search cost distribution. Given this knowledge, the researcher wants to identify and estimate the search cost distribution. In our parametric setting, this amounts to identifying and estimating $\beta := \mathbb{P}(c_n = \overline{c})$. Data available to the researcher can come in many forms. We consider the following standard possibilities (in all cases, we consider i.i.d. observations and we denote by N the sample size).

Data on Choice. The researcher observes the share of agents in the sample who take an alternative with utility $\underline{u}$ and the share of agents in the sample who take an alternative with utility $\overline{u}$. Equivalently, the researcher observes the utility of the alternative taken by each agent in the sample, but has no data on individual search behavior.

Data on the Number of Searches. The researcher observes the share of agents in the sample who conducted only one search and the share of agents in the sample

Figure 5: Decision Tree for a Simultaneous Search Problem of Type $\theta_n = B$ if $\underline{c} < \Delta(V_{\mathrm{B}})$.

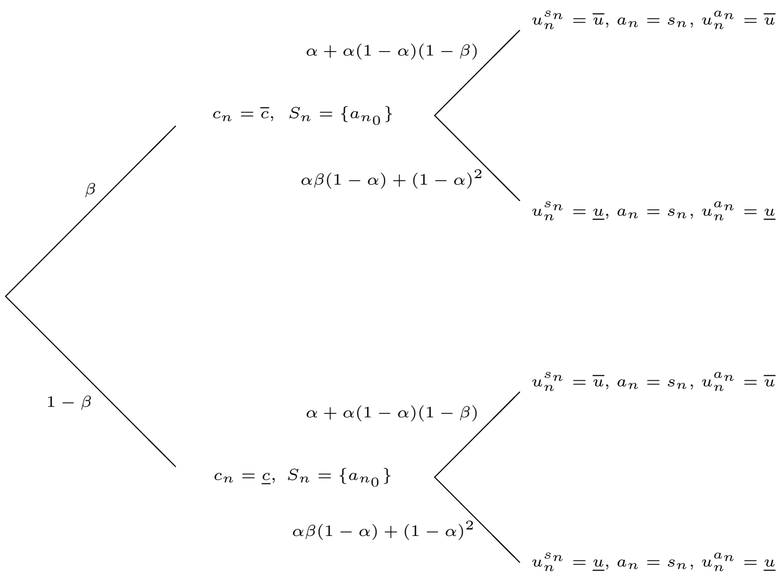


who conducted two searches, without observing the utilities of the searched alternatives. Equivalently, for each agent in the sample, the researcher observes the number of searches conducted, without observing the utilities of the searched alternatives.

We do not consider data on optimal stopping decisions as they cannot be interpreted using a simultaneous search model.

For each case, we show how the researcher can use the available dataset to construct an unbiased and consistent estimator of the the parameter β when all agents in the population are isolated (i.e., when $\gamma = 1$). Estimating search cost distributions under the assumption that all agents are isolated is what is routinely done in the empirical literature on search. Next, we show when and why the constructed estimators become biased and inconsistent when a positive share of agents in the population have social in formation (i.e., when $\gamma < 1$); depending on the available dataset, neglecting social information may lead to under- or over-estimation of search cost distributions.

Figure 6: Decision Tree for a Simultaneous Search Problem of Type $\theta_n = B$ if $\underline{c} > \Delta(V_\mathrm{B})$.



A.3.1 Data on Choice

Formally, the researcher observes $\underline{u}_N^a$, where

$$\underline{u}_N^a \coloneqq \frac{\sum_{n=1}^N \mathbb{1}_{\{u_n^{a_n} = \underline{u}\}}}{N}.$$

By the weak law of large numbers,

$$\underline{u}_N^a \xrightarrow{p} \mathbb{E}[\underline{u}_N^a] = \mathbb{P}(u_n^{a_n} = \underline{u}). \tag{38}$$

Moreover, note that

$$\begin{aligned}\mathbb{P}(u_n^{a_n} = \underline{u}) &= \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})\mathbb{P}(\theta_n = \mathrm{A}) + \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B})\mathbb{P}(\theta_n = \mathrm{B}) \\ &= \gamma\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) + (1-\gamma)\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}).\end{aligned} \tag{39}$$

From the decision trees in Figures 4–6, we calculate that

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) = \alpha\beta(1-\alpha) + (1-\alpha)^2 \tag{40}$$

and

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) = \begin{cases} \alpha\beta^2(1-\alpha) + (1-\alpha)^2 & \text{if } \underline{c} < \Delta(V_\mathrm{B}) \\ \alpha\beta(1-\alpha) + (1-\alpha)^2 & \text{if } \underline{c} > \Delta(V_\mathrm{B}) \end{cases}. \tag{41}$$

Therefore, from equations (39)–(41), we have

$$\mathbb{P}(u_n^{a_n} = \underline{u}) = \begin{cases} \alpha\beta(1-\alpha)(1-\alpha)[\gamma + (1-\gamma)\beta] + (1-\alpha)^2 & \text{if } \underline{c} < \Delta(V_\mathrm{B}) \\ \alpha\beta(1-\alpha) + (1-\alpha)^2 & \text{if } \underline{c} > \Delta(V_\mathrm{B}) \end{cases}. \tag{42}$$

All Agents Are Isolated. Assume that $\gamma = 1$. From equations (38)–(40), we have that the solutions for $\hat{\beta}_N^1$ to equation equation

$$\alpha\hat{\beta}_N^1(1-\alpha) + (1-\alpha)^2 = \underline{u}_N^a,$$

given by

$$\hat{\beta}_N^1 := \frac{\underline{u}_N^a}{\alpha(1-\alpha)} - \frac{1-\alpha}{\alpha}. \tag{43}$$

is an unbiased and consistent estimator of β.

The Role of Social Information. Next, assume that $\gamma < 1$. Moreover, suppose the researcher neglects social information and estimates the parameter β using the estimator $\hat{\beta}_N^1$ in equation (43).

If $\underline{c} < \Delta(V_\mathrm{B})$,

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A}) > \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) \tag{44}$$

(compare equation (40) to equation (41) for $\underline{c} < V_\mathrm{B}(\underline{u})$). To understand equation (44), suppose that $\theta_n = \mathrm{B}$. If $\underline{c} < \Delta(V_\mathrm{B})$, we have two cases:

(i) If $c_n = \overline{c}$, agent n commits to search only alternative a_{n_0}, $S_n = \{a_{n_0}\}$, and so $\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) = \mathbb{P}(u_{n_0}^{a_{n_0}} = \underline{u}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})$ (where the last equality holds because $\theta_{n_0} = \mathrm{A}$).

(ii) If $c_n = \underline{c}$, agent n commits to search both alternatives, $S_n = X$; with positive probability, $u_n^{\neg a_{n_0}} = \overline{u}$ (because, with positive probability, the two alternatives have different utility and agent n_0 committed to search only one alternative), in which case agent n takes alternative $a_n = \neg a_{n_0}$ (so that $u_n^{a_n} = \overline{u}$); therefore, $\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{B}) < \mathbb{P}(u_n^{a_{n_0}} = \underline{u}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})$.

Since both (i) and (ii) occur with positive probability, equation (44) follows. As a result $\mathbb{P}(u_n^{a_n} = \underline{u})$ depends on γ and, in particular,

$$\mathbb{P}(u_n^{a_n} = \underline{u}) < \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \mathrm{A})$$

(compare equation (40) to equation (42) for $\underline{c} < V_{\rm B}(\underline{u})$). Therefore, the estimator $\hat{\beta}_N^1$ in equation (43) becomes biased downward and inconsistent. More formally,

$$\hat{\beta}_N^1 \xrightarrow{p} \mathbb{E}\Big[\hat{\beta}_N^1\Big] = \gamma\beta + (1-\gamma)\beta^2$$

and

$$\text{Bias}\Big(\hat{\beta}_N^1, \beta\Big) := \mathbb{E}\Big[\hat{\beta}_N^1\Big] - \beta = -(1-\gamma)\beta(1-\beta) < 0. \tag{45}$$

That is, search costs are underestimated: the true search cost distribution first order stochastically dominates the estimated search cost distribution.

In contrast, if $\underline{c} > \Delta(V_{\rm B})$, then

$$\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \text{A}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \text{B}) \tag{46}$$

(compare equation (40) to equation (41) for $\underline{c} > V_{\rm B}(\underline{u})$). To understand equation (44), suppose that $\theta_n = \text{B}$. If $\underline{c} > \Delta(V_{\rm B})$, agent n always commits to search only alternative a_{n0}, $S_n = \{a_{n0}\}$, and so $\mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \text{B}) = \mathbb{P}(u_{n0}^{a_{n0}} = \underline{u}) = \mathbb{P}(u_n^{a_n} = \underline{u} \mid \theta_n = \text{A})$ from which equation (13) follows. As a result, $\mathbb{P}(u_n^{a_n} = \underline{u})$ does not depend on γ (see equation (42) for $\underline{c} > V_{\rm B}(\underline{u})$) and so the estimator $\hat{\beta}_N^1$ remains unbiased and consistent.

A.3.2 Data on the Number of Searches

Formally, the researcher observes d_N, where

$$d_N := \frac{\sum_{n=1}^N \mathbb{1}_{\{|S_n|=1\}}}{N}.$$

By the weak law of large numbers,

$$d_N \xrightarrow{p} \mathbb{E}[d_N] = \mathbb{P}(|S_n| = 1). \tag{47}$$

Moreover, note that

$$\begin{aligned}\mathbb{P}(|S_n| = 1) &= \mathbb{P}(|S_n| = 1 \mid \theta_n = \text{A})\mathbb{P}(\theta_n = \text{A}) \\ &\quad + \mathbb{P}(|S_n| = 1 \mid \theta_n = \text{B})\mathbb{P}(\theta_n = \text{B}) \\ &= \mathbb{P}(|S_n| = 1 \mid \theta_n = \text{A})\gamma + \mathbb{P}(|S_n| = 1 \mid \theta_n = \text{B})(1-\gamma).\end{aligned} \tag{48}$$

From the decision trees in Figures 1–3, we calculate that From the decision trees in Figures 1–3, we calculate that

$$\mathbb{P}(|S_n| = 1 \mid \theta_n = \text{A}) = \beta \tag{49}$$

and

$$\mathbb{P}(|S_n| = 1 \mid \theta_n = \mathrm{B}) = \begin{cases} \beta & \text{if } \underline{c} < \Delta(V_\mathrm{B}) \\ 1 & \text{if } \underline{c} > \Delta(V_\mathrm{B}) \end{cases}. \tag{50}$$

Therefore, from equations (48)–(50), we have

$$\mathbb{P}(|S_n| = 1) = \begin{cases} \beta & \text{if } \underline{c} < \Delta(V_\mathrm{B}) \\ \gamma\beta + (1-\gamma) & \text{if } \underline{c} > \Delta(V_\mathrm{B}) \end{cases}. \tag{51}$$

All Agents Are Isolated. Suppose that $\gamma = 1$. From equations (47)–(49), we have that

$$\hat{\beta}_N^2 := d_N \tag{52}$$

is an unbiased and consistent estimator of β.

The Role of Social Information. Next, assume that $\gamma < 1$. Moreover, suppose the researcher neglects social information and estimates the parameter β using the estimator $\hat{\beta}_N^2$ in equation (52).

If $\underline{c} < \Delta(V_\mathrm{B})$, then

$$\mathbb{P}(|S_n| = 1 \mid \theta_n = \mathrm{A}) = \mathbb{P}(|S_n| = 1 \mid \theta_n = \mathrm{B}) \tag{53}$$

(compare equation (49) to equation (50) for $\underline{c} < \Delta(V_\mathrm{B})$). The equality in (53) immediately follows by observing that agent n commits to search only one alternative if and only if $c_n = \bar{c}$ independently of whether $\theta_n = \mathrm{A}$ or $\theta_n = \mathrm{B}$. Thus, $\mathbb{P}(|S_n| = 1)$ does not depend on γ (see equation (51) for $\underline{c} < \Delta(V_\mathrm{B})$) and so the estimator $\hat{\beta}_N^2$ in equation (52) remains unbiased and consistent.

In contrast, if $\underline{c} > \Delta(V_\mathrm{B})$,

$$\mathbb{P}(|S_n| = 1 \mid \theta_n = \mathrm{A}) < \mathbb{P}(|S_n| = 1 \mid \theta_n = \mathrm{A}) \tag{54}$$

(compare equation (49) to equation (50) for $\underline{c} > \Delta(V_\mathrm{B})$). To understand equation (54), note that agent n: commits to search only one alternative if and only if $c_n = \bar{c}$ if $\theta_n = \mathrm{A}$; always commits to search only one alternative if $\underline{c} > \Delta(V_\mathrm{B})$ and $\theta_n = \mathrm{B}$. As a result, $\mathbb{P}(|S_n| = 1)$ depends on γ, and in particular,

$$\mathbb{P}(|S_n| = 1) < \mathbb{P}(|S_n| = 1 \mid \theta_n = \mathrm{A})$$

(compare equation (49) to equation (51) with $\underline{c} > \Delta(V_\mathrm{B})$). Therefore, the estimator $\hat{\beta}_N^2$ in equation (52) becomes biased upward and inconsistent. More formally, we have

$$\hat{\beta}_N^2 \xrightarrow{p} \mathbb{E}\Big[\hat{\beta}_N^2\Big] = \gamma\beta + (1-\gamma)$$

and

$$\text{Bias}\Big(\hat{\beta}_N^2, \beta\Big) := \mathbb{E}\Big[\hat{\beta}_N^2\Big] - \beta = (1-\gamma)(1-\beta) > 0. \tag{55}$$

That is, search costs are overestimated: the estimated search cost distribution first order stochastically dominates the true search cost distribution.

A.4 Discussion

Table A.4 sums up the results presented in Section A.3. In particular, for each of the datasets we consider, the table illustrates whether and when neglecting social information leads to biased and inconsistent estimates of search cost distributions and, when so, it also illustrates the bias direction. When present, equations (45) and (55) show that the bias increases as the share of agents in the population with social information grows large (i.e., as γ decreases).

Table 2: Summary of Identification and Estimation in a Simultaneous Search Model.

	Estimation Bias	
	Data on Choice	Data on Number of Searches
$\underline{c} < \Delta(V_{\text{B}})$	< 0	NO
$\underline{c} > \Delta(V_{\text{B}})$	NO	> 0

The results are similar as those with a sequential search model. There is, however, one difference. With a simultaneous search setting, estimation with data on the number of searches is unbiased and consistent if $\underline{c}$ is negligible (so that $\underline{c} < \Delta(V_{\text{B}})$); in contrast, with a sequential search setting, estimation with data on the number of searches (interpreted as data on search duration in that setting) is always biased and inconsistent.

The remedies that we discuss in Section 5 for a sequential search model also apply to a simultaneous search model, with the obvious changes.

References

Anderson, Simon P. and Régis Renault (2018), "Firm Pricing with Consumer Search." *Handbook of Game Theory and Industrial Organization, Volume II.* Edward Elgar Publishing.

Bailey, Michael, Drew Johnston, Theresa Kuchler, Johannes Stroebel, and Arlene Wong (2021), "Peer Effects in Product Adoption." *American Economic Journal: Applied Economics (forthcoming).*

Baye, Michael R., John Morgan, and Patric Scholten (2006), "Information, Search, and Price Dispersion." *Handbook on Economics and Information Systems*, 1, 323–375.

Bergemann, Dirk and Stephen Morris (2016), "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games." *Theoretical Economics*, 11, 487–522.

Choi, Michael, Anovia Yifan Dai, and Kyungmin Kim (2018), "Consumer Search and Price Competition." *Econometrica*, 86, 1257–1281.

Doval, Laura and Jeffrey C. Ely (2020), "Sequential Information Design." *Econometrica*, 88, 2575–2608.

Ellison, Glenn and Alexander Wolitzky (2012), "A Search cost Model of Obfuscation." *The RAND Journal of Economics*, 43, 417–441.

Galeotti, Andrea (2010), "Talking, Searching, and Pricing." *International Economic Review*, 51, 1159–1174.

Garcia, Daniel and Sandro Shelegia (2018), "Consumer Search with Observational Learning." *The RAND Journal of Economics*, 49, 224–253.

Gualdani, Cristina and Shruti Sinha (2020), "Identification and Inference in Discrete Choice Models with Imperfect Information." *Working Paper*.

Hendricks, Kenneth, Alan Sorensen, and Thomas Wiseman (2012), "Observational Learning and Demand for Search Goods." *American Economic Journal: Microeconomics*, 4, 1–31.

Honka, Elisabeth, Ali Hortaçsu, and Matthijs Wildenbeest (2019), "Empirical Search and Consideration Sets." *Handbook of the Economics of Marketing*, 1. Elsevier, 193–257.

Kircher, Philipp and Andrew Postlewaite (2008), "Strategic Firms and Endogenous Consumer Emulation." *The Quarterly Journal of Economics*, 123, 621–661.

Kuchler, Theresa, Yan Li, Lin Peng, Johannes Stroebel, and Dexin Zhou (2021), "Social Proximity to Capital: Implications for Investors and Firms." *The Review of Financial Studies (forthcoming).*

Lomys, Niccolò (2020), “Collective Search in Networks.” *Working Paper.*

Magnolfi, Lorenzo and Camilla Roncoroni (2021), “Estimation of Discrete Games with Weak Assumptions on Information.” *Working Paper.*

Mueller-Frank, Manuel and Mallesh M. Pai (2016), “Social Learning with Costly Search.” *American Economic Journal: Microeconomics*, 8, 83–109.

Stahl, Dale O. (1989), “Oligopolistic Pricing with Sequential Consumer Search.” *The American Economic Review*, 700–712.

Stigler, George J. (1961), “The Economics of Information.” *Journal of Political Economy*, 69, 213–225.

Syrgkanis, Vasilis, Elie Tamer, and Juba Ziani (2021), “Inference on Auctions with Weak Assumptions on Information.” *Working Paper.*

Varian, Hal R. (1980), “A Model of Sales.” *The American Economic Review*, 70, 651–659.

Weitzman, Martin L. (1979), “Optimal Search for the Best Alternative.” *Econometrica*, 43, 641–654.