Information Acquisition, Referral, and Organization

Simona Grassi

Department of Strategy
HEC
Université de Lausanne

simona.grassi@unil.ch

Ching-to Albert Ma

Department of Economics
College of Arts and Sciences
Boston University

ma@bu.edu

October 2015

Abstract

Each of two experts may provide service for a client. In one state, one expert has a lower service cost than the other expert; in another state, the opposite is true. Each expert may also exert effort to acquire information about a client's service cost. Effort and acquired signal are private information. In a market, an expert may refer the client to the other for a fee. In equilibrium, only one expert exerts effort and refers successfully, yet effort and referral are inefficient. If experts form an organization, they can transfer costs among themselves. Within such an organization, an expert who refers bears the service cost incurred by the referred expert. Referral efficiency can be restored at the expense of cost-reduction incentives. An organization has a lower expected cost if and only if referral efficiency is more important than cost incentives.

Keywords: information acquisition, referral, organization, integration, cost-reduction incentive

JEL: D00, D02, D80, D83

Acknowledgement: For their comments, we thank Francesca Barigozzi, Jacopo Bizzotto, Jean-Philippe Bonardi, Giacomo Calzolari, Mark Dusheiko, George Georgiadis, Katharina Huesmann, Izabella Jelovac, Andrew Jones, Liisa Laine, Henry Mak, Debby Minehart, Marco Ottaviani, Raphael Parchet, Peter Zweifel, and seminar participants at the Antitrust Division of the US Department of Justice, Boston University, the European Economic Association congress, the Final Report Seminar for Harkness fellows 2012-2013 in New York, the Annual International Industrial Organization Conference in Boston, Indiana University–Purdue University Indianapolis, the MIT/Harvard/Boston University Health Economics Seminar, the Nordic Health Economists' Study Group Workshop in Oslo, the Norwegian School of Economics in Bergen, the Ph.D. Seminar in Health Economics and Policy in Grindelwald (Switzerland), the University of Lausanne, the University of Lugano, and the University of Oslo. Coeditor David Martimort and two referees gave us very valuable suggestions. Simona Grassi is indebted to the Commonwealth Fund, the Careum Foundation, and to her mentor Joe Newhouse while being a Harkness/Careum Fellow at the Harvard Medical School.

1 Introduction

We study an economic system consisting of experts who provide services to clients. An expert may invest in effort to find out a client's state-contingent service costs, as well as to reduce overall costs. We consider efficiencies in a referral market and within an expert organization that can assign cost responsibilities. For each institution, we study an expert's information-acquisition and cost-reduction incentives, and experts' incentives to refer clients to each other.

Information acquisition and task assignment are topical in policy forums. In the U.S. healthcare reform, cost-control measures are being phased in after the Affordable Care Act took effect in 2014. The Center for Medicare and Medicaid Services, the federal agency that administers the insurance programs for the elderly and the indigent, has been encouraging providers (such as general practitioners, specialists, and hospitals) to form so-called Accountable Care Organizations (ACOs).[1] Such organizations are supposed to reduce cost through better care coordination achieved by referrals among physicians (see Song, Sequist, and Barnett (2014)). Other professionals, such as accountants and lawyers, refer clients to each other, whether they operate in the market or within an organization. How do experts' performances compare in the market and within an organization? We provide a framework for these comparisons.

In our model, each of a set of clients would like to obtain service from one of two experts. A client's case can be easy or complicated. An easy case is always less expensive to service than a complicated one. However, the two experts have different cost comparative advantages: Expert 1 has a lower service cost than Expert 2 if the case is easy; conversely, Expert 2 incurs a lower cost than Expert 1 if the case is complicated. The complexity of a client's case is unknown. An expert may exert some effort to obtain information about the case. The effort generates an informative signal, and, as a convention, a higher signal indicates a higher likelihood of a complicated case, so a higher (expected) cost. The service from an expert gives a fixed benefit to a client, and each client

[1] For a description of ACOs, see: https://www.cms.gov/medicare/medicare-fee-for-service-payment/sharedsavingsprogram/downloads/aco-narrativemeasures-specs.pdf

pays a fixed tariff for the service.[2]

We first study how experts operate in a referral market. After, say, Expert 1 has exerted an effort and observed a signal, he may make a referral offer to Expert 2: the client and the service tariff are transferred from Expert 1 to Expert 2 if Expert 2 pays a referral price. The problems facing these experts are: i) effort is hidden action, unknown to anyone except the expert who exerts it, and ii) the signal generated by effort is hidden information, unknown to anyone except the expert who has exerted the effort.

Despite asymmetric-information problems, there is an equilibrium in which Expert 1 exerts effort, and successfully refers clients to Expert 2 if and only if their signal is above a threshold (a higher signal indicating a higher expected cost). An expert's incentive is to avoid complicated and costly clients, so in equilibrium Expert 2 only gets lemons from Expert 1. However, Expert 2 has a cost advantage in complicated cases. Expert 1 credibly exploits this cost advantage when setting the referral price, so the referred lemons will be accepted.

Expert 2's acceptance decision is based on comparing the referral price with the average cost given that signals are above the threshold. For efficiency, Expert 2 should have compared the referral price with the actual expected cost, but this is Expert 1's private information. This discrepancy is common in adverse-selection models. As a result, Expert 1's referral and information-acquisition decisions do not internalize all cost savings due to cost comparative advantage, and are never first best.

Expert 2's equilibrium strategy, however, is completely different. He will neither exert effort nor make any referral. The cost comparative advantage for Expert 1 is for the client with an easy case, but there is no equilibrium in which Expert 2 refers a client to Expert 1. A simple case is more profitable than a complicated case. If in equilibrium Expert 2 was successful at referring a client at a signal, he would also refer the client if the signal had become higher (indicating a higher cost).

[2] A stylized example is this. A consumer needs to file a tax return. Simple returns are less time-consuming than complex returns. However, a tax preparer is more cost effective than an accountant for a simple return, and vice versa.

In other words, Expert 2 would always refer lemons, never peaches. Expert 2's referral, therefore, would not let Expert 1 exploit his cost comparative advantage. Without any success in referral in equilibrium, Expert 2 does not exert effort.

We then study expert organizations. The referral market equilibrium is inefficient because an expert is unconcerned about the cost consequence to be borne by the expert who accepts the referral. Our premise is that an organization differs from the market because it can make cost information available *ex post*. In an organization, when an expert refers a client, the referring expert can be held responsible for the cost incurred by the referred expert. We call this the *cost-transfer protocol*. An expert now can fully internalize benefits of cost comparative advantage. If Expert 1's signal indicates that Expert 2 has a lower expected cost, he simply refers the client to Expert 2 and, under the cost-transfer protocol, reaps the cost savings. (Song, Sequist, and Barnett (2014) identify an ACO exactly as an organization in which "physicians...share the consequences of each other's referral decisions".)

We also examine a drawback of the cost-transfer protocol. We enrich our model by allowing each expert to choose a cost-reduction effort when serving a client. This adds another hidden action. When experts operate in the market, each is responsible for his cost, so cost effort must be efficient. Cost reduction is orthogonal to information acquisition and referral in the market. This is no longer true for an organization that uses the cost-transfer protocol.

Our point is that the cost-transfer protocol introduces a new tradeoff. When experts can reduce their costs, the magnitude of cost saving determines whether a market performs better than an organization. If cost saving by effort is small, cost comparative advantage dominates cost effort, so an expert organization performs better than the market. If cost saving by effort is large, the opposite is true. Ours is also a theory about whether referrals should be among experts within a firm under the cost-transfer protocol, or among independent experts in the market.

We consider various extensions of the basic model. First, we discuss constraints on experts' capacities, and variable returns. We qualify how various results should be properly interpreted.

Second, we endogenize clients' tariffs by a Bertrand game. Finally, we let cost comparative advantage potentially be big so that a client may be a lemon to one expert, but a peach to another. There, we show that referrals by both experts may arise in equilibrium.

Our paper is related to the literatures on credence goods, referrals, and organizations. In contrast to models of credence goods (see the Dulleck and Kerschbamer (2006) survey), we simplify experts' price and treatment decisions. Here, an expert sets one price and has no control over costs. Furthermore, many models of credence goods are on interactions between experts and clients. Instead, we study the interactions between experts via referral and information acquisition.

Garicano and Santos (2004) study referrals between two experts who have different productivities and costs in generating revenue from a project by exerting efforts. An expert can choose between implementing a project himself, or referring it to the other expert. Referral of a project is subject to asymmetric information because a project's potential can be either high or low, which is privately known by an expert. Equilibrium referrals via fixed price or revenue-sharing contracts are often inefficient. In our model, private information is in the form of a continuous signal, rather than a binary signal. The kinds of inefficiency in our model are also different. First, experts' efforts to acquire information are inefficient. Second, an expert's equilibrium referrals do not internalize social cost savings. Third, when experts form an organization, we allow the transfer of costs, which leads to shirking.

Referrals incentives have been studied in models of consumers searching for experts' advice; see Arbatskaya and Konishi (2012), Bolton, Freixas, and Shapiro (2007), Inderst and Ottaviani (2009), and Park (2005). Experts face a tradeoff between honestly advising clients to build a good reputation, and reaping a quick profit at the client's expense. We do not model search or reputation here, but we show that even without threats from consumers, referrals may occur.

Referrals are studied in the health literature. In the health sector, insurers set up incentive mechanisms for referrals between providers, say, between general practitioners and specialists (see Shumsky and Pinker (2003) and Mariñoso and Jelovac (2003)). We do not follow a contract-design approach but our result suggests that physician organizations may lead to efficienct referrals. Also, market referrals with financial transfers are uncommon in the medical sector. However, our analysis of how an organization provides incentive to refer clients is relevant to the organizational approach currently advocated in the health domain. We will revisit this after we have presented results.

We contribute to organizational economics. Our hypothesis that costs become transferable when experts merge is similar to the reallocation of ownership rights within a firm. Schmidt (1996) argues that the allocation of ownership rights has an important impact on the allocation of information about the firm. Garicano (2000), Garicano and Santos (2004) and Fuchs and Garicano (2010) argue that organizations can better match clients to experts, and this is supported by evidence of obstetric practices in Epstein, Ketcham, and Nicholson (2010). We have argued that when cost comparative advantage is internalized, matches will be efficient, so we explain why better matches happen.

However, we point out the degradation of work incentives when costs are transferred in an organization. This possibility has also been raised by Frandsen and Rebitzer (2015), who show that free-riding problems in ACOs may erode savings from better care coordination. Cebul *et al.* (2008) and Rebitzer and Votruba (2011) provide evidence on the adverse effects of coordination failures in the health care delivery system in the U.S. The free-riding and work-incentive deficiency should be weighed against better referrals, which, according to Able (2013), is the mechanism by which ACOs reduce aggregate medical expenditures and improve Medicare patient health.

The paper is organized as follows. In Section 2, we set up the model and derive the first best. Section 3 studies a market in which experts can refer clients to each other at a price. In Section 4, we present organizations and compare them to the market and the first best. We also provide specific perspectives on the relevance of our theory to legal and medical professionals. In Section 5 we consider a number of robustness issues. Section 6 concludes. An Appendix contains proofs of results.

2 The model

2.1 Clients and experts

Each of a set of clients needs the service from one of two experts. These clients may be consumers who seek services from professionals such as lawyers, doctors, or engineers. Alternatively, a company may have projects that require inputs from outside contractors, and these projects correspond to the clients while the contractors are the experts. We let there be a continuum of clients, with the total mass normalized at 1. Each client is characterized by a state or a type. Each client's state or type is independently and identically distributed on the binary support $\{\omega_1, \omega_2\}$ with a probability 1/2 on each state. We discuss the equal prior assumption in Section 5.

There are two experts, namely Expert 1 and Expert 2. Each expert can provide a service to any number of clients. This amounts to an assumption that experts have enough capacities. We further assume that the cost of service (including effort disutility, see below) is linear in the number of clients served. We do not aim to construct a theory on organizations and incentives based on returns to scale or fixed costs, so nonbinding capacity and constant returns are natural assumptions. We assume that each expert gives the same benefit to a client.

Experts differ by their service costs that are dependent on a client's states. The following table defines each expert's cost contingent on a client's type:

	state ω_1	state ω_2
Expert 1's cost	c_L	c_H
Expert 2's cost	$c_L + \Delta$	$c_H - \Delta$

where $0 < c_L < c_L + \Delta < c_H - \Delta < c_H$ (so $2\Delta < c_H - c_L$). If a client's state is ω_1, Expert 1's service cost, c_L, is lower than Expert 2's, $c_L + \Delta$, but if a client's state is ω_2, Expert 2's service cost is lower. (In Section 5 we consider an alternative cost configuration: $0 < c_L < c_H - \Delta < c_L + \Delta < c_H$.)

The cost saving Δ is assumed to be symmetric between the experts for convenience. *Ex ante* each expert has the same expected cost of providing services to clients. State ω_1 can be thought of as a "good" or "easy" state: the service cost is lower, either c_L for Expert 1 or $c_L + \Delta$ for Expert 2. State ω_2 corresponds to a "bad" or "complicated" state with service cost either $c_H - \Delta$ or c_H.

Expert 1 has a cost advantage Δ in state ω_1, while Expert 2 has an advantage in state ω_2.

The setup for experts' costs will be enriched in Section 4. Each expert's cost will be stochastic, and each expert can take an effort to reduce the expected value of his cost distribution (so the costs defined above would be expected values). We will then use the more general setup to compare markets and organizations. Until then, we use the simpler setup above. Our definition of the first best, and our results in Section 3 are unaffected by the omission of cost-reduction efforts.

We subscribe to the credence-good framework. Clients do not get to observe their states when they seek services from experts. Neither do clients get to observe how much cost an expert eventually incurs to provide the service. The only contractible event for clients is that the service is provided. To a client, for a given tariff for service provided, the experts are identical because each of them provides the same benefit.

2.2 Information acquisition

Experts do not observe clients' cost types. Each expert can acquire information about a client's cost type by exerting a costly effort. We assume that each expert has the same information-acquisition technology and effort disutility. The information comes in the form of a signal defined on a positive support, $s \in [\underline{s}, \overline{s}]$. Let $e \in \mathbb{R}_+$ denote an expert's effort, and $\phi(e)$ denote the disutility of effort. The disutility ϕ may have fixed and variable components. To acquire information, an expert has to set up an experiment, and exercise care during the investigation. These two steps correspond to the fixed and variable components. We assume that the fixed disutility is not so high as to make information acquisition worthless. On the other hand, we will assume that it is not so low that an expert will acquire information many times. That is, $\phi(0) > 0$, but $\phi(0)$ is not too big. We assume that for any strictly positive effort, ϕ is increasing and convex. We also let $\lim_{e \to 0^+} = \phi'(e) = 0$, so that the disutility due to the variable component can be arbitrarily low.

Let $f_i(s|e)$ be the density of the signal s conditional on effort e and state ω_i, $i = 1, 2$. We assume that both f_1 and f_2 are always strictly positive, and continuous. By Bayes rule, conditional on a signal s, the posterior probability of the state being ω_i is

$$\Pr(\omega_i|s,e) = \frac{f_i(s|e)}{f_1(s|e)+f_2(s|e)}, \qquad i=1,2. \tag{1}$$

We assume that for any effort, the signal satisfies *Monotone Likelihood Ratio Property (MLRP)*:

$$\frac{f_2(s'|e)}{f_2(s|e)} \geq \frac{f_1(s'|e)}{f_1(s|e)} \qquad \text{for } s' > s, \text{ each } e.$$

As a normalization, we let the signals be completely uninformative at the lowest effort, $e = 0$, so that $f_1(s|0) = f_2(s|0)$, each $s \in [\underline{s}, \overline{s}]$, and that for $e > 0$, the inequality in the MLRP definition holds as a strict inequality for each s. Under MLRP $\dfrac{f_2(s|e)}{f_1(s|e)}$ is increasing in s, so a higher value of the signal indicates a higher likelihood that the state is ω_2:

$$\Pr(\omega_2|s,e) = \frac{1}{1+\dfrac{f_1(s|e)}{f_2(s|e)}} \qquad \text{is increasing in } s.$$

For future use, we note that the *ex ante* density of signal s, given effort e, is $\Pr(\omega_1)f_1(s|e) + \Pr(\omega_2)f_2(s|e) = 0.5[f_1(s|e)+f_2(s|e)]$.

A higher effort makes signals more informative. We use the following assumption on how the densities f_1 and f_2 relate to efforts, and call it the *Informativeness Property*:

> For $e' > e$, $f_2(s|e')$ first-order stochastically dominates $f_2(s|e)$, and $f_1(s|e)$ first-order stochastically dominates $f_1(s|e')$.

A higher effort reduces the conditional cumulative density $\displaystyle\int_{\underline{s}}^{s} f_2(x|e)\mathrm{d}x$ and raises the conditional cumulative density $\displaystyle\int_{\underline{s}}^{s} f_1(x|e)\mathrm{d}x$. First-order stochastic dominance is often used in the literature to define how effort affects information. A higher effort makes a lower signal more indicative of state ω_1, while it makes a higher signal more indicative of state ω_2. We further assume that both conditional densities are differentiable in e.

2.3 First best

An allocation is an effort to be taken by an expert, and a decision rule that assigns a client to an expert according to the generated signal. The first best is an allocation that minimizes experts' expected service cost and effort disutilities. (In Subsection 4.1, we will extend the definition of an allocation and the first best to include a cost-reduction effort.) In principle, an allocation can prescribe multiple efforts to generate multiple (informative) signals and an assignment rule based on all signals. Our assumption at the beginning of Subsection 2.2 rules this out, so effort is to be exerted only once in the first best.

Let an expert take effort e. Contingent on signal s, the expected cost of servicing this client by Experts 1 and 2 are, respectively,

$$\Pr(\omega_1|s,e)c_L + \Pr(\omega_2|s,e)c_H \tag{2}$$

$$\Pr(\omega_1|s,e)\left(c_L+\Delta\right) + \Pr(\omega_2|s,e)\left(c_H-\Delta\right). \tag{3}$$

The conditional probabilities are given by (1), so Expert 2 has a cost lower than Expert 1 if and only if $f_1(s|e) \le f_2(s|e)$. For each effort e, define $\widehat{s}^{fb}(e)$ by $f_1(\widehat{s}^{fb}|e) = f_2(\widehat{s}^{fb}|e)$. By MLRP, $s \ge \widehat{s}^{fb}$ if and only if $f_1(s|e) \le f_2(s|e)$. In this notation, the cost-minimizing allocation assigns a client to Expert 2 if and only if the client's signal s is larger than $\widehat{s}^{fb}(e)$.

Given the cost-minimizing allocation, the total expected service cost and effort disutility per client is

$$\begin{aligned}&0.5\int_{\underline{s}}^{\widehat{s}^{fb}(e)}\{\Pr(\omega_1|x,e)c_L + \Pr(\omega_2|x,e)c_H\}[f_1(x|e)+f_2(x|e)]\mathrm{d}x + \\ &0.5\int_{\widehat{s}^{fb}(e)}^{\overline{s}}\{\Pr(\omega_1|x,e)(c_L+\Delta) + \Pr(\omega_2|x,e)(c_H-\Delta)\}[f_1(x|e)+f_2(x|e)]\mathrm{d}x + \phi(e). \end{aligned} \tag{4}$$

We assume that (4) is quasi-convex.[3] The first-best effort, e^{fb}, is one that minimizes (4). The

[3]Consider $-0.5\Delta\int_{\widehat{s}}^{\overline{s}}\{f_2(x|e)-f_1(x|e)\}\mathrm{d}x + \phi(e)$, which is the expected cost at threshold $\widehat{s}$ (after some constants have been dropped). The Hessian of this expected cost is:

$$\begin{bmatrix} -0.5\Delta\int_{\widehat{s}}^{\overline{s}}\left\{\frac{\partial^2 f_2(x|e)}{\partial e^2} - \frac{\partial^2 f_1(x|e)}{\partial e^2}\right\}\mathrm{d}x + \phi''(e) & 0.5\Delta\left\{\frac{\partial f_2(\widehat{s}|e)}{\partial e} - \frac{\partial f_1(\widehat{s}|e)}{\partial e}\right\} \\ 0.5\Delta\left\{\frac{\partial f_2(\widehat{s}|e)}{\partial e} - \frac{\partial f_1(\widehat{s}|e)}{\partial e}\right\} & 0.5\Delta\left\{\frac{\partial f_2(\widehat{s}|e)}{\partial s} - \frac{\partial f_1(\widehat{s}|e)}{\partial s}\right\}\end{bmatrix}.$$

Convexity requires that the Hessian is positive definite.

first-order condition is:

$$0.5\Delta \int_{\widehat{s}^{fb}(e^{fb})}^{\bar{s}} \left\{ \frac{\partial f_2(x|e^{fb})}{\partial e} - \frac{\partial f_1(x|e^{fb})}{\partial e} \right\} \mathrm{d}x = \phi'(e^{fb}). \tag{5}$$

The first best characterization has the following interpretations. First, the base costs, c_L and c_H, set up reference points only, so their values do not appear in the first-order condition (5). Second, cost saving, from c_H to $c_H - \Delta$ may be achieved, and cost increase from c_L to $c_L + \Delta$ may be avoided. The assignment of a client to Expert 2 whenever s is above a threshold is for cost effectiveness. Third, a higher effort yields more precise signals, but leads to more disutility. The left-hand side of (5) reflects the benefit. Because both f_1 and f_2 are densities, the integral in (5) would have been zero if the lower limit was set to $\underline{s}$. Now by the Informativeness Property, this integral, with lower limit at $\widehat{s}^{fb}(e^{fb}) > \underline{s}$ must be strictly positive, and it measures how strongly higher values of s leads to cost-effective assignments of clients. The right-hand side of (5) is the marginal disutility of effort.

We assume that clients are matched randomly to experts, and pay a fixed tariff, T, to the expert who renders a service. Each client obtains the same benefit from an expert and each expert's *ex ante* cost for treating a random client is equal to the average cost. The only restriction here is that T is at least the *ex ante* average cost, $(c_L + c_H)/2$. In Subsection 5.1, we endogenize the tariff (and also the initial assignment of clients) by letting experts compete in a Bertrand fashion. Our results are unchanged with endogenously chosen tariffs.[4]

3 Referral market

We look for perfect-Bayesian equilibria of the following extensive form:

Stage 0: For each client, his cost type, either ω_1 or ω_2, is drawn independently with equal probabilities. The draw is never observed by a client or an expert. Half of all clients are matched

[4] Any given value of the tariff and initial assignment define a valid subgame of the extensive form to be presented, so our analysis for arbitrary tariff and assignment is necessary even when tariffs are determined endogenously.

with Expert 1, and the other half with Expert 2.

Stage 1: An expert decides on an effort for a matched client. Then the expert observes a realization of the signal for each exerted effort. The effort and signal are the expert's private information.

Stage 2: For each client an expert chooses between keeping the client and referring the client to the other expert at a price that he chooses.

Stage 3: If an expert has received a referral at some price, the expert decides whether to accept the referral or reject it. If the expert accepts the referral, he pays the other expert the referral price, provides service to the client, incurs the cost (as the client's state eventually realizes), and receives the tariff. If he rejects the referral, the referring expert will render service and receive the tariff.

In Stage 3, an expert may not acquire information before deciding between accepting and rejecting a referral. This may be due to an expert having no access to the client until he has accepted the referral. Alternatively, information acquisition may be time consuming, and delays may be unacceptable to clients. Finally, a model with multiple rounds of information acquisition together with offers and counteroffers, is less tractable, and outside the scope here.

An expert's payoff comes from one of three events. First, if an expert has kept his own client, he gets the tariff, and incurs the service cost and effort disutility. Second, if an expert has accepted a referral, he pays the referral price, keeps the tariff, and incurs the service cost. Third, if an expert's referral has been accepted, he gets the referral price and incurs the effort disutility. Each expert has a reservation utility that is set at 0. The referral price made by an expert can be positive or negative.

An expert's strategy is defined by i) an effort in Stage 1, ii) a referral decision and price in Stage 2 as a function of the expert's own signal, and iii) a referral-acceptance decision in Stage 3 as a function of the referral price. A perfect-Bayesian equilibrium consists of a pair of strategies that are mutual best responses, and beliefs about (unobserved) effort and signals, which are updated according to strategies and Bayes rule whenever possible.

There are many unreached information sets. For example, in an equilibrium, Expert 1 may take some effort e_1, make a referral offer at price p_1 if and only if signal s is above a certain threshold. What would Expert 2 believe about Expert 1's effort and signal if Expert 1's referral price were $p_1' \neq p_1$? Also, in an equilibrium, an expert may not make any referral at all, so all referral prices are off-equilibrium. Perfect-Bayesian equilibria do not impose belief restrictions at out-of-equilibrium information sets. Multiple equilibria can be supported by many off-path beliefs (and will be discussed later). We will impose a natural and simple belief restriction to be defined in Subsection 3.2.

In the following subsections, we construct an equilibrium with the following outcome: Expert 1 exerts a strictly positive effort, but Expert 2 does not. Expert 1 refers at a price for all signals above a threshold. Expert 2 does not refer. We will begin the construction by presenting necessary conditions, then prove existence by a standard fixed-point argument.

3.1 Experts' equilibrium referral and acceptance strategies

Consider any equilibrium in which Expert 1 has taken an effort, say $e_1 > 0$, and has observed a signal s in Stage 1. In a continuation equilibrium in Stage 2, if Expert 1 makes a referral that will be accepted, it will always be at a unique price. Indeed, If Expert 2 would accept at referral prices p_1' and p_1, with $p_1 < p_1'$, then Expert 1 would never make a referral at the lower price p_1. Hence, in equilibrium, Expert 2 must reject all offers above a threshold, p_1.

Suppose that Expert 2 accepts a referral at price p_1. How should Expert 1 choose between keeping and referring the client? Given that Expert 1 has taken effort e_1 and observed signal s, the expected payoff (net from effort disutility) from keeping the client is

$$T - \Pr(\omega_1|s, e_1)c_L - \Pr(\omega_2|s, e_1)c_H. \tag{6}$$

Because this is decreasing in s by MLRP, we conclude that Expert 1 will refer the client with signal

s whenever $s > \widehat{s}$ where

$$T - \Pr(\omega_1|\widehat{s}, e_1)c_L - \Pr(\omega_2|\widehat{s}, e_1)c_H = p_1. \tag{7}$$

Clearly, we can repeat the same steps for Expert 2's referral decision given that Expert 1 accepts a referral if the price is below a threshold. We summarize the result in the following lemma, whose proof is already in the text above.

Lemma 1 *In an equilibrium, in Stage 3 an expert's referral is accepted if and only if the referral price is at or below a threshold, and in Stage 2, an expert makes a referral if and only if the signal exceeds a threshold.*

Lemma 1 asserts that, in any equilibrium in which effort is positive, referral decisions and acceptance decisions must be threshold policies. Transmission of the private signal from information-acquisition effort must be pooling. Furthermore, an expert will refer lemons and keep peaches.

3.2 Expert 1's equilibrium referral and effort

We now focus on Expert 1's referral under the assumption that he has taken an effort. First, we introduce a belief restriction:

Definition 1 (Passive Belief) *A perfect-Bayesian equilibrium is said to satisfy passive belief if an expert's belief about the hidden effort and signal on any off-equilibrium referral price remains the same as the belief at the equilibrium price.*

Passive belief was first introduced by McAfee and Schwartz (1994) in the context of multilateral contracts.[5] Here, it says that deviations are uncorrelated trembles. Suppose that, in an equilibrium, Expert 1 takes effort e_1, and makes a referral offer at price p_1 if and only if s is above a certain

[5] "Passive belief" is a common assumption in models of foreclosure, delegation, and integrations; see Dequiedt and Martimort (2015), de Fontenay and Gans (2005), Hart and Tirole (1990), Laffont and Martimort (2000), O'Brien and Shaffer (1992), Reisinger and Tarantino (2015), and Rey and Tirole (2007). More recently, it has also been used in the consumer-search literature; see Bar-Isaac, Caruana and Cunat (2012), Buehler and Schuett (2014), and Inderst and Ottaviani (2012).

threshold. If Expert 2 receives a referral price $p_1' \neq p_1$, passive belief specifies that Expert 2 continues to believe that Expert 1 has taken effort e_1 and has made a referral because the signal is above s. The restriction requires Expert 2 to believe that his expected cost remains at the same equilibrium level even when Expert 1 offers an off-equilibrium referral price.

Lemma 2 *Under passive belief, in an equilibrium in which Expert 1 takes effort e_1 and refers whenever $s > \widehat{s}$, Expert 1's equilibrium referral price p_1 must be :*

$$p_1 = T - \Pr(\omega_1 | s > \widehat{s}, e_1)(c_L + \Delta) - \Pr(\omega_2 | s > \widehat{s}, e_1)(c_H - \Delta), \tag{8}$$

so Expert 2's equilibrium expected utility must equal the outside option.

The proof of Lemma 2 is this. When Expert 2 receives a referral, according to passive belief, he must believe that Expert 1's signal is above $\widehat{s}$. Expert 2's expected cost of providing service to the referred client is $\Pr(\omega_1 | s > \widehat{s}, e_1)(c_L + \Delta) + \Pr(\omega_2 | s > \widehat{s}, e_1)(c_H - \Delta)$. Therefore, Expert 2 accepts the referral if and only if the price is lower than $T - \Pr(\omega_1 | s > \widehat{s}, e_1)(c_L + \Delta) - \Pr(\omega_2 | s > \widehat{s}, e_1)(c_H - \Delta)$. Given this best response by Expert 2, Expert 1 optimally chooses the highest price that will be accepted. This is the definition of p_1 in (8). Clearly, Expert 2 earns a zero expected utility when he accepts a referral. (The equilibrium referral price may be negative; Expert 1 may have to pay Expert 2 in order to cover the expected cost because the tariff is low. For example, if T is just equal to the average of c_L and c_H, then T cannot cover Expert 2's expected cost, but Expert 1 will set p_1 to be negative to cover that loss. Expert 1 will do this because his loss without a referral would be even higher.)

Passive belief does rule out many other equilibria. In these equilibria, Expert 2 earns strictly more than the outside option, but referral happens less often. To see this, choose $\varepsilon > 0$, and for the same effort e_1 and some signal threshold $\widehat{s}'$ consider a referral price p_1' satisfying

$$p_1' + \varepsilon = T - \Pr(\omega_1 | s > \widehat{s}', e_1)(c_L + \Delta) - \Pr(\omega_2 | s > \widehat{s}', e_1)(c_H - \Delta).$$

Now Expert 2's strategy is to accept a referral at price p' or lower. Expert 2 believes that the signal is at least $\widehat{s}'$. If Expert 1 offers $p_1 > p_1'$, Expert 2 would change his belief; he now believes that the signal threshold has increased from $\widehat{s}'$ (perhaps all the way to $\overline{s}$). Now that the bad state is thought to be more likely, Expert 2's expected cost has increased, so he rejects p_1. Expert 1 is then stuck with having to refer at a price that leaves some rent. Passive belief rules out such a discontinuous change: when a referral is made at a higher price, Expert 2 must continue to believe that the referral threshold is $\widehat{s}'$. (See also the discussion following Proposition 1.) From now on, we will always use passive belief.

We continue to characterize Expert 1's referral threshold $\widehat{s}$. Recall that Expert 1's payoff from keeping a client with signal s is (6). Given that Expert 2 accepts a referral at price p_1, Expert 1 refers a client with signal $s > \widehat{s}$ if and only if $p_1 = T - \Pr(\omega_1|\widehat{s}, e_1)c_L - \Pr(\omega_2|\widehat{s}, e_1)c_H$. As an intermediate step, we present a basic property about experts' expected costs conditional on signals. (The proof is in the Appendix.)

Lemma 3 *For $e_1 > 0$, the equation*

$$\Pr(\omega_1|\widehat{s}, e_1)c_L + \Pr(\omega_2|\widehat{s}, e_1)c_H \equiv \frac{c_L f_1(\widehat{s}|e_1) + c_H f_2(\widehat{s}|e_1)}{f_1(\widehat{s}|e_1) + f_2(\widehat{s}|e_1)} \tag{9}$$

$$= \frac{(c_L + \Delta)\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\,dx + (c_H - \Delta)\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\,dx}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\,dx + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\,dx} \tag{10}$$

$$\equiv \Pr(\omega_1|s > \widehat{s}, e_1)(c_L + \Delta) + \Pr(\omega_2|s > \widehat{s}, e_1)(c_H - \Delta)$$

has a unique solution $\underline{s} \leq \widehat{s} \leq \overline{s}$.

Suppose that Expert 1 has chosen effort e_1. If he observes signal s, his expected cost of providing service is (9), whereas if Expert 2 gets all the clients with signals above s, Expert 2's expected cost is (10). Lemma 3 says that there must be a signal $\widehat{s}$ for which these two expected costs are equal.[6] This result stems from Expert 2's comparative advantage in providing services to clients at state

[6] If effort has not been exerted, then $f_1 = f_2$, and the signal is uninformative. The only solution for the equation is $\underline{s}$. For any strictly positive effort e_1, the solution is strictly interior.

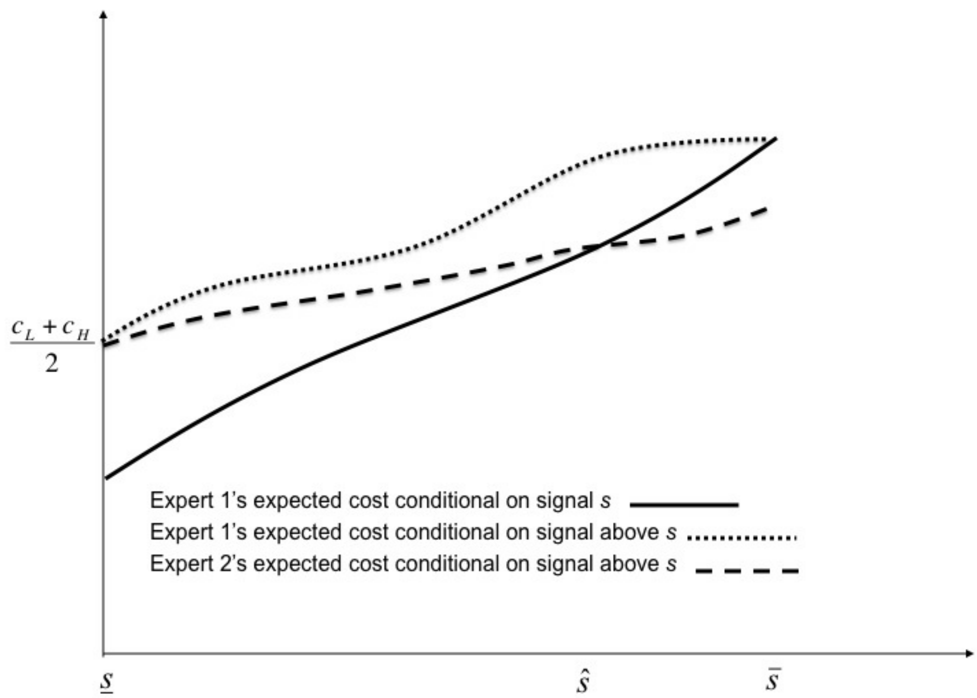

Figure 1: Expected costs and Expert 1's referral threshold $\widehat{s}$

ω_2; Expert 2's cost is Δ less than Expert 1's. Figure 1 graphs three expected costs. The solid line is Expert 1's expected cost at signal s (9). The dotted line is Expert 1's expected cost given that signals are above s:

$$\Pr(\omega_1|s > \widehat{s}, e_1)c_L + \Pr(\omega_2|s > \widehat{s}, e_1)c_H \equiv \frac{c_L \int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + c_H \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}. \tag{11}$$

This dotted line is always above Expert 1's expected cost at signal s. By MLRP, a higher s means that state ω_2 is more likely. The expected cost conditional on all signals above s must indicate a higher expected cost than at signal s. This expected cost, conditional on signals above s, of course converges to (9) at $s = \overline{s}$.

Now Expert 2's comparative advantage makes his expected cost, expression (10) and the dashed line in Figure 1, less than (11) at high values of s. (Expression (11) is identical to expression (10) at $\Delta = 0$.) But this comparative advantage diminishes as the conditional threshold s drops towards $\underline{s}$. If Expert 2 cannot exclude any possible signal Expert 1 has observed, his expected cost is simply

$(c_L + c_H)/2$. The solution $\widehat{s}$ is the intersection of the solid and dashed lines.

The significance of Lemma 3 is this. Although Expert 1's referrals pool all clients with signals higher than $\widehat{s}$, the experts nevertheless can mutually benefit from trade due to Expert 2's cost comparative advantage at state ω_2. Expert 1's referrals are all lemons, but Expert 2's has lower expected cost servicing lemons. Given effort e_1, as long as the signal is above $\widehat{s}$, the one in Lemma 3, a successful referral happens in equilibrium, as the next result shows (proof in the Appendix).

Proposition 1 *In an equilibrium in which Expert 1 exerts strictly positive effort e_1, he refers a client with a signal $s \geq \widehat{s}$ to Expert 2 at a price p_1, and Expert 2 accepts a referral if and only if Expert 1's price is at most p_1, where*

$$T - \frac{(c_L + \Delta)\displaystyle\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\,dx + (c_H - \Delta)\displaystyle\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\,dx}{\displaystyle\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\,dx + \displaystyle\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\,dx} = p_1 = T - \frac{c_L f_1(\widehat{s}|e_1) + c_H f_2(\widehat{s}|e_1)}{f_1(\widehat{s}|e_1) + f_2(\widehat{s}|e_1)}. \tag{12}$$

In (12), the first equation says that Expert 2 is indifferent between accepting all referrals of clients with signals above $\widehat{s}$ and rejecting. The second equation says that Expert 1 is indifferent between keeping client with signal $\widehat{s}$ and referring. Together they determine the continuation referral equilibrium given effort e_1. These are mutual best responses. Proposition 1 stems from classical adverse selection. Expert 1's referral is based on his private information, so client $\widehat{s}$ is his *marginal* client. Expert 2 faces the *average* client with signals above $\widehat{s}$. Adverse selection does not rule out trade because of cost comparative advantage.

Again, passive belief does rule out other continuation equilibria. For the same effort e_1, other equilibria with referral price p_1' and referral threshold $\widehat{s}'$ are possible. Let

$$T - \frac{(c_L + \Delta)\displaystyle\int_{\widehat{s}'}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta)\displaystyle\int_{\widehat{s}'}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\displaystyle\int_{\widehat{s}'}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \displaystyle\int_{\widehat{s}'}^{\overline{s}} f_2(x|e_1)\mathrm{d}x} - \varepsilon = p_1' = T - \frac{c_L f_1(\widehat{s}'|e_1) + c_H f_2(\widehat{s}'|e_1)}{f_1(\widehat{s}'|e_1) + f_2(\widehat{s}'|e_1)}.$$

Here, Expert 1's referral price is p_1', and Expert 2's equilibrium payoff is at $\varepsilon > 0$. If Expert 1 raised the price from p_1' to extract more rent, Expert 2 now would believe that the client had a very high signal, say $\overline{s}$, and would reject the higher price. This continuation equilibrium is consistent with Lemma 1, but violates passive belief. Under passive belief, any price between p_1' and $p_1' + \varepsilon$ must be accepted by Expert 2. For effort e_1, the continuation equilibrium in Proposition 1 has the most referrals.

We next study Expert 1's effort incentive. If e_1 is an equilibrium effort, given that Expert 2 will accept a referral at price p_1, Expert 1's referral threshold is in (7). Recalling that the *ex ante* density of s is $0.5[f_1(s|e_1) + f_2(s|e_1)]$, we write Expert 1's expected payoff per client as

$$\int_{\underline{s}}^{\widehat{s}} 0.5[(T - c_L)\Pr(\omega_1|x, e_1) + (T - c_H)\Pr(\omega_2|x, e_1)][f_1(x|e_1) + f_2(x|e_1)]\mathrm{d}x +$$
$$p_1 \int_{\widehat{s}}^{\overline{s}} 0.5[f_1(x|e_1) + f_2(x|e_1)]\mathrm{d}x - \phi(e_1).$$

From the definition of $\widehat{s}$ in (7), the first integral above is Expert 1's expected utility when he keeps the client (s below $\widehat{s}$), while the second is the expected utility when he successfully refers (s above $\widehat{s}$). Using the expressions for the conditional probabilities of the states ω_1 and ω_2, we simplify the payoff per client to

$$\left[T - \frac{c_L + c_H}{2}\right] - \phi(e_1) + 0.5\int_{\widehat{s}}^{\overline{s}} \{[p_1 - (T - c_L)]f_1(x|e_1) + [p_1 - (T - c_H)]f_2(x|e_1)]\}\,\mathrm{d}x. \tag{13}$$

The first term in (13) is the expected payoff from treating a randomly chosen client; effort has a cost, the second term, but generates an expected benefit, the difference between the referral price p_1 and what Expert 1 would have obtained if he had kept the client (the integral).

In an equilibrium in which Expert 1's effort is positive, his equilibrium effort e_1^* and the referral threshold $\widehat{s}$ maximize (13) subject to the definition of $\widehat{s}$ in (7). The first-order condition characterizes Expert 1's equilibrium effort:[7]

$$0.5\int_{\widehat{s}}^{\overline{s}} \left\{[p_1 - (T - c_L)]\frac{\partial f_1(x|e_1)}{\partial e_1} + [p_1 - (T - c_H)]\frac{\partial f_2(x|e_1)}{\partial e_1}\right\}\mathrm{d}x = \phi'(e_1). \tag{14}$$

A higher effort raises the density f_2 more than the density f_1 at high signals s by the Informativeness Property, so $\int_{\widehat{s}}^{\overline{s}} \frac{\partial f_1(x|e_1)}{\partial e_1}\mathrm{d}x < \int_{\widehat{s}}^{\overline{s}} \frac{\partial f_2(x|e_1)}{\partial e_1}\mathrm{d}x$. Also because $c_H > c_L$, for any p_1 between $T - c_H$ and $T - c_L$, the term inside the curly brackets in (14) must be strictly positive.

[7] In fact, the constraint (7) is redundant because the unconstrained maximization of (13) with respect to e_1 and $\widehat{s}$ yields that constraint anyway.

3.3 Expert 2's equilibrium effort

We now turn to Expert 2's equilibrium effort and referrals. Indeed, one might have thought that some "symmetry" might apply so Expert 2 could exploit cost comparative advantage. The answer is negative, as stated in the next Proposition (proof in the Appendix).

Proposition 2 *In any equilibrium Expert 2 does not exert any effort or make any referral.*

Expert 1 has cost comparative advantage in the good state ω_1, so if there was any referral to exploit that advantage, Expert 2 would have to refer clients with low signals. Lemma 1, however, says that an expert would like to keep only clients with low signals; an expert will never refer peaches, only lemons. If Expert 1 believed that Expert 2 was referring clients with low signals, Expert 2 would cheat and refer clients with high signals. However, for clients with signals above a threshold Expert 1's expected costs will never be lower than Expert 2's. There is no possibility of mutually beneficial trade. Given that in equilibrium Expert 2 does not refer, there is no incentive for him to acquire information.

Because Expert 2 does not make any equilibrium referral, *all* referral price offers to Expert 1 are off-equilibrium, so passive belief has no bite. For later use, we note that the highest posterior belief on the bad state ω_2 can be written as $\Pr(\omega_2|\overline{s}, \widetilde{e})$ where $\widetilde{e} = \text{argmax}_e\, f_2(\overline{s}|e)/f_1(\overline{s}|e)$. We say that an expert has the *most pessimistic belief* if he believes that the other expert has taken effort $\widetilde{e}$ and has observed the signal $\overline{s}$.

3.4 Equilibrium information acquisition and referral

Now we put together our earlier results and state the following (proof in the Appendix).

Proposition 3 *There is an equilibrium characterized by the triple $[e_1^*, \widehat{s}^*, p_1^*]$ such that i) $(e_1^*, \widehat{s}^*)$ maximize (13) given p_1^*, and ii) p_1^* is given by (12) at $\widehat{s}$ and e_1^*. The equilibrium strategies and beliefs are:*

1) Expert 1 chooses effort e_1^* and refers a client with any signal $s > \widehat{s}^*$ at price p_1^* and keeps other *clients.*

2) Expert 2 chooses zero effort, does not refer, and accepts a referral if and only if the referral price is at most p_1^*.

3) If Expert 2 receives a referral at a price different from p_1^**, he continues to believe that Expert 1 has chosen effort* e_1^* *and referred a client with signal* $s > \widehat{s}^*$.

4) If Expert 1 receives a referral offer from Expert 2 at any price, Expert 1 has the most pessimistic belief (he believes that Expert 2's effort is $\widetilde{e}$ *and his signal is* $\overline{s}$*, where* $\widetilde{e} = \text{argmax}_e\, f_2(\overline{s}|e)/f_1(\overline{s}|e)$*).*

In Proposition 3, the first three points in the strategy and belief description follow directly from the previous two subsections. The fourth point is about Expert 1's response against off-equilibrium referral prices. We specify that Expert 1 has the most pessimistic belief. This is necessary to deter Expert 2 from deviating to a positive effort, and referring a client when the signal indicates an expected loss. If Expert 1 believes that Expert 2 has taken no effort, he will accept any offer p when $T - (c_L + c_H)/2 - p \geq 0$. Now if Expert 2 chooses effort $e_2 > 0$, and he observes an s where his expected cost is higher than the *ex ante* cost: $\Pr(\omega_1|s, e_2)(c_L + \Delta) + \Pr(\omega_2|s, e_2)(c_H - \Delta) > (c_L + c_H)/2$, Expert 2 will profit by successfully referring at $p = T - (c_L + c_H)/2$. Of course, this is inconsistent with Proposition 2, so Expert 1 believing Expert 2 having taken zero effort cannot be part of off-equilibrium belief. To support the equilibrium, we have chosen the most pessimistic off-equilibrium belief when no price is ever offered in equilibrium, and show in the proof that Expert 2 has no profitable deviation from zero effort. The final step in the proof is a standard, fixed-point argument for the existence of $[e_1^*, \widehat{s}^*, p_1^*]$.

Expert 2 does not exert any effort, which, of course, is inefficient. What about Expert 1's equilibrium effort and referrals? Using Proposition 1, and the first-order condition for Expert 1's equilibrium effort, we write down the conditions for the referral equilibrium $[e_1^*, \widehat{s}^*, p_1^*]$:

$$T - \frac{(c_L+\Delta)\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x + (c_H-\Delta)\int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x}{\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x + \int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x} = p_1^* = T - \frac{c_L f_1(\widehat{s}^*|e_1^*) + c_H f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \quad (15)$$

$$0.5 \int_{\widehat{s}^*}^{\overline{s}} \left\{ [p_1^* - (T - c_L)] \frac{\partial f_1(x|e_1^*)}{\partial e_1} + [p_1^* - (T - c_H)] \frac{\partial f_2(x|e_1^*)}{\partial e_1} \right\} \mathrm{d}x = \phi'(e_1^*). \quad (16)$$

Proposition 4 *In an equilibrium, Expert 1's effort and referral threshold cannot be first best. Furthermore, given equilibrium effort e_1^*, Expert 1's referral threshold $\widehat{s}^*$ is too high, $f_2(\widehat{s}^*|e_1^*) > f_1(\widehat{s}^*|e_1^*)$, so Expert 1 sometimes keeps a client even when his expected service cost is higher than Expert 2's.*

At the equilibrium referral threshold, Expert 1's expected cost at $\widehat{s}^*$ equals Expert 2's expected cost when signals are all above $\widehat{s}^*$. Given $\Delta > 0$ and MLRP, equality of the Expert 1's "marginal" cost and Expert 2's "average" cost requires $f_2(\widehat{s}^*|e_1^*) > f_1(\widehat{s}^*|e_1^*)$. (See also the proof of Proposition 4 in the Appendix.)

What about Expert 1's equilibrium effort? Using (15), we rewrite (16) as

$$0.5 \left[\frac{c_H - c_L}{2}\right] \int_{\widehat{s}^*}^{\overline{s}} \left\{ \frac{2 f_1(\widehat{s}^*|e_1^*) \dfrac{\partial f_2(x|e_1^*)}{\partial e_1} - 2 f_2(\widehat{s}^*|e_1^*) \dfrac{\partial f_1(x|e_1^*)}{\partial e_1}}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \right\} \mathrm{d}x = \phi'(e_1^*). \quad (17)$$

The left-hand side of this expression is the marginal benefit of effort. We already have noted that $c_H - c_L > 2\Delta$, so that compared to the first best, the cost differential $c_H - c_L$ affects the marginal benefit more strongly than the cost saving Δ. However, we have $\widehat{s}^*$ strictly higher than the value at the cost-effective threshold (where $f_2(s|e_1^*) = f_1(s|e_1^*)$), so the integral is smaller. Moreover, the weight on the partial derivative $\partial f_2/\partial e_1$ is smaller than 1, while the weight on $\partial f_1/\partial e_1$ is larger than 1. These two effects reduce the marginal benefit. In sum, the equilibrium effort may be smaller or larger than the first best.[8]

[8] In fact, the proof in the Appendix shows that even if Expert 1 referred a client to Expert 2 if and only if the signal indicated the bad state to be more likely, Expert 1's effort would still be excessive.

Finally, we remark that the characterization of the referral equilibrium in (15) and (16) does not prove uniqueness. Although examples that we have constructed so far have all produced a unique equilibrium, we have not proven it. Formally, Lemma 3 does show that for any given effort, (15) admits a unique solution for the price and referral threshold. It remains possible, however, that (16) admits multiple solutions in effort. However, our characterization applies to every equilibrium.

4 Organizations

Equilibria in the referral market are inefficient. An expert organization can perform better. As we have hypothesized in the Introduction, the key difference between an open market and an organization is that service costs *ex post* become verifiable within an organization. The reassignment of cost responsibiity is possible. We present a definition:

Definition 2 (Cost-transfer Protocol) *Referrals are said to follow the cost-transfer protocol when the referring expert bears the client's cost when service is provided by the referred expert.*

The cost-transfer protocol allows an organization to solve the adverse selection problem by transferring costs between experts. When an expert within an organization refers a client to a fellow expert, he is to be held responsible for the costs to be incurred by the referred expert. In other words, an expert fully internalizes the cost consequence of referring the client to another expert.

Using the cost-transfer protocol, many organizations, such as integration and partnership, can achieve the first best. Expert 1 buys out Expert 2, becomes the owner, performs all information acquisition, and refers clients whose signals indicate a higher likelihood of the bad state. Expert 2 becomes an employee, and any cost incurred will be the firm's responsibility. Obviously, Expert 2 buying out Expert 1 achieves the same. Alternatively, the experts can form a partnership. Here, each expert will acquire information, and refers efficiently. The partnership contract specifies that an expert making a referral fully reimburses the service expense.

The (simplistic) solution relies on an expert does nothing other than providing service at a predetermined set of costs. We now consider a richer environment in which an expert's service includes an additional input: he also supplies an effort that may reduce costs. Cost responsibility implies an incentive of cost reduction. But the cost-transfer protocol in organizations such as integration and partnership will mute this incentive. We will demonstrate a tradeoff between referral efficiency and cost efficiency, but first we extend the basic model to include cost reduction.

4.1 Cost-reduction effort

We enrich the model in Subsection 2.1 with general cost reduction. First, a client's service cost is now randomly distributed on a positive support $[\underline{c}, \overline{c}]$. Next, each expert has a second hidden action: a cost-reduction effort $r \geq 0$ (besides the information-acquisition effort). Then we define four distributions G_i^j on $[\underline{c}, \overline{c}]$, where G_i^j is the distribution of Expert j's service cost in state ω_i, $i, j = 1, 2$. The following table defines experts' expected costs across states and at effort r:

	state ω_1	state ω_2
Expert 1's expected cost	$\int_{\underline{c}}^{\overline{c}} c \mathrm{d}G_1^1(c\|r) = c_L - r$	$\int_{\underline{c}}^{\overline{c}} c \mathrm{d}G_2^1(c\|r) = c_H - r$
Expert 2's expected cost	$\int_{\underline{c}}^{\overline{c}} c \mathrm{d}G_1^2(c\|r) = c_L + \Delta - r$	$\int_{\underline{c}}^{\overline{c}} c \mathrm{d}G_2^2(c\|r) = c_H - \Delta - r$

so $G_i^j(\cdot|r)$ is Expert j's cost distribution at effort r and state i, and c_L, $c_L + \Delta$, $c_H - \Delta$, and c_H are expected costs at zero cost effort. The effort r reduces each expert's expected cost by r in each state.[9] An expert incurs a disutility $\psi(r)$ from effort r, and the function ψ is increasing and convex, with $\lim_{r \to 0} \psi'(r) = 0$, and $\lim_{r \to \Delta} \psi'(r) = +\infty$. The cost effort and its disutility have the usual interpretation: task management, work hours, attention, etc. The last assumptions of the Inada sort ensure that cost comparative advantage is always valid because expected cost reduction never exceeds Δ. The cost effort is to be taken when an expert provides service. We define the efficient cost effort by $r^* \equiv \text{argmax}_r[r - \psi(r)]$, and the first-best net cost reduction $\gamma \equiv r^* - \psi(r^*)$.

Cost-reduction and information-acquisition efforts are orthogonal in the first best and in the

[9] We can make the cost reductions different across states. This adds nothing conceptually, but burdens with more notation.

market. When each expert is fully responsible for service cost, he chooses effort r^* which results in the cost saving γ. In the first best and the market-referral model, both experts take cost effort r^*, so we simply redefine c_L, $c_L + \Delta$, $c_H - \Delta$, and c_H by lowering each by γ. The first best and equilibria now refer to these redefined values. Characterizations of equilibrium information effort and referral remain the same. More important, equilibria derived in the previous section do not depend on cost saving, γ.[10]

4.2 Tradeoff in an organization: cost comparative advantage versus cost effort

In the enriched model, we use the full-support cost distribution assumption. Contracts for the efficient cost effort by exploiting shifting cost-distribution supports are infeasible. An expert takes effort r^* if and only if he is fully responsible for costs. In principle, organizations can employ partial cost-sharing contracts (the referring expert, for example, being responsible for 50% of cost) so that efforts between 0 and r^* can be implemented. For brevity, we do not consider partial cost-sharing contracts because they would not change economic principles. In other words, we continue to adopt the cost-transfer protocol: all service costs are to be borne by the referring expert.

The cost-transfer protocol, defined above, eliminates adverse selection, but it also eliminates the cost effort. The referred expert will not take effort to realize the net cost saving γ. This is the main tradeoff. Now we adopt the accounting convention that the tariff stays with the expert who initiates the referral, so under cost-transfer protocol all referrals will be accepted with a zero price. We continue with the assumption that half of all clients are matched with one expert—although the two experts operate within an organization. Consider Expert 1, and suppose that he has taken information-acquisition effort e_1, and receives a signal s on a client. His own expected service cost is $\Pr(\omega_1|s, e_1)(c_L - \gamma) + \Pr(\omega_2|s, e_1)(c_H - \gamma)$ because he chooses cost effort r^*. Upon a referral, Expert 2 is not responsible for service cost, so he takes zero cost effort. From Expert 1's perspective, if the client is referred to Expert 2, Expert 1 pays a service cost $\Pr(\omega_1|s, e_1)(c_L + \Delta) - \Pr(\omega_2|s, e_1)(c_H - \Delta)$.

[10]The key Lemma 3 is unaffected because if each of c_L and c_H is reduced by γ, the solution to the equation there remains the same. The value of the equilibrium price will also be reduced by γ, so the equilibrium effort becomes the same.

Clearly, Expert 1 refers the client if and only if doing so saves cost or if signal s is larger than $\widetilde{s}_1$ defined by

$$f_1(\widetilde{s}_1|e_1)(c_L-\gamma)+f_2(\widetilde{s}_1|e_1)(c_H-\gamma)=f_1(\widetilde{s}_1|e_1)(c_L+\Delta)+f_2(\widetilde{s}_1|e_1)(c_H-\Delta). \tag{18}$$

Simplifying (18), we obtain the following (proof in the Appendix):

Lemma 4 *At effort e_1, Expert 1 refers a client to Expert 2 if and only if signal s is higher than $\widetilde{s}_1$, where*

$$\frac{f_1(\widetilde{s}_1|e_1)}{f_2(\widetilde{s}_1|e_1)}=\frac{\Delta-\gamma}{\Delta+\gamma}\leq 1. \tag{19}$$

The threshold $\widetilde{s}_1$ is first best at $\gamma=0$, and increases to $\overline{s}$ as γ increases to Δ.

Expert 1 will take cost effort r^* for his clients to lower his expected cost by γ, but under the cost-transfer protocol, Expert 2 will not. If a signal indicates that Expert 2's expected cost is lower, it must be because the bad state is much more likely than 1/2. Furthermore, as the net saving from cost effort γ increases, cost comparative advantage becomes less important, so referrals become less likely.

From Lemma 4, Expert 1's total expected cost from effort e_1 is

$$0.5\left\{\begin{array}{l}\displaystyle\int_{\underline{s}}^{\widetilde{s}_1}[f_1(x|e_1)(c_L-\gamma)+f_2(x|e_1)(c_H-\gamma)]\mathrm{d}x+\\ \displaystyle\int_{\widetilde{s}_1}^{\overline{s}}[[f_1(x|e_1)(c_L+\Delta)+f_2(x|e_1)(c_H-\Delta)]\mathrm{d}x\end{array}\right\}+\phi(e_1). \tag{20}$$

Expert 1's payoff is not aligned with the social return to information-acquisition effort. In the first best, Expert 2 chooses cost effort r^*, but, in an organization, Expert 2 chooses zero cost effort. We present (proof in the Appendix):

Lemma 5 *Expert 1 does not choose the first-best information-acquisition effort in the cost-transfer protocol except at $\gamma=0$. As γ increases to Δ, Expert 1's information-acquisition effort decreases to 0.*

Information-acquisition effort is beneficial only if it leads to referrals. When cost effort is ineffective ($\gamma=0$), Expert 1 chooses the first-best information effort because he internalizes cost comparative advantage. As γ increases, cost reduction becomes more important, and Expert 1's *ex ante* expected cost becomes lower, so he refers less often. In the limit when $\gamma = \Delta$, Expert 1 does not acquire information.

Lemmas 4 and 5 hold in a symmetric fashion for Expert 2. We now state these results (but omit their proofs).

Lemma 6 *At effort e_2, Expert 2 refers a client to Expert 1 if and only if signal s is lower than $\widetilde{s}_2$, where*

$$\frac{f_1(\widetilde{s}_2|e_2)}{f_2(\widetilde{s}_2|e_2)} = \frac{\Delta + \gamma}{\Delta - \gamma} \geq 1. \tag{21}$$

The threshold $\widetilde{s}_2$ is first best at $\gamma = 0$, and decreases to $\underline{s}$ as γ increases to Δ.

Analogous to (20), Expert 2's expected cost from effort e_2 is

$$0.5 \left\{ \begin{array}{c} \displaystyle\int_{\underline{s}}^{\widetilde{s}_2} [f_1(x|e_2)c_L + f_2(x|e_2)c_H]\mathrm{d}x + \\ \displaystyle\int_{\widetilde{s}_2}^{\overline{s}} [f_1(x|e_2)(c_L + \Delta - \gamma) + f_2(x|e_2)(c_H - \Delta - \gamma)]\mathrm{d}x \end{array} \right\} + \phi(e_2).$$

Lemma 7 *Expert 2 does not choose the first-best information-acquisition effort in the cost-transfer protocol except at $\gamma = 0$. As γ increases to Δ, Expert 2's information-acquisition effort decreases to* 0*.*

4.3 Comparison between organization and market

Let $\widetilde{e}_1$ and $\widetilde{e}_2$ be Expert 1's and Expert 2's information efforts in the cost-transfer protocol. Expert 1 provides service to clients with signals below $\widetilde{s}_1$, and refers those with signals above. Expert 2 provides service to clients with signals above $\widetilde{s}_2$, and refers those with signals below. Referrals are always accepted, but the expert who receives a referral will take zero cost effort. The total equilibrium expected costs per client is

$$0.5\left\{\begin{array}{c} 0.5\int_{\underline{s}}^{\widetilde{s}_1}[f_1(x|\widetilde{e}_1)(c_L-\gamma)+f_2(x|\widetilde{e}_1)(c_H-\gamma)]\mathrm{d}x+ \\ 0.5\int_{\widetilde{s}_1}^{\overline{s}}[f_1(x|\widetilde{e}_1)(c_L+\Delta)+f_2(x|\widetilde{e}_1)(c_H-\Delta)]\mathrm{d}x+\phi(\widetilde{e}_1) \end{array}\right\}$$
$$+0.5\left\{\begin{array}{c} 0.5\int_{\underline{s}}^{\widetilde{s}_2}[f_1(x|\widetilde{e}_2)c_L+f_2(x|\widetilde{e}_2)c_H]\mathrm{d}x+ \\ 0.5\int_{\widetilde{s}_2}^{\overline{s}}[f_1(x|\widetilde{e}_2)(c_L+\Delta-\gamma)+f_2(x|\widetilde{e})(c_H-\Delta-\gamma)]\mathrm{d}x+\phi(\widetilde{e}_2) \end{array}\right\}.$$

These four integrals correspond to different cases of experts retaining and referring clients. We simplify the expected cost per client in the cost-transfer protocol to

$$\left\{\frac{c_L+c_H}{2}\right\}-0.5\left\{\gamma\int_{\underline{s}}^{\widetilde{s}_1}[f_1(x|\widetilde{e}_1)+f_2(x|\widetilde{e}_1)]\mathrm{d}x+\Delta\int_{\widetilde{s}_1}^{\overline{s}}[f_2(x|\widetilde{e}_1)-f_1(x|\widetilde{e}_1)]\mathrm{d}x\right\}$$
$$-0.5\left\{\gamma\int_{\widetilde{s}_2}^{\overline{s}}[f_1(x|\widetilde{e}_2)+f_2(x|\widetilde{e}_2)]\mathrm{d}x+\Delta\int_{\widetilde{s}_2}^{\overline{s}}[f_2(x|\widetilde{e}_2)-f_1(x|\widetilde{e})]\mathrm{d}x\right\}$$
$$+0.5[\phi(\widetilde{e}_1)+\phi(\widetilde{e}_2)]\equiv EC_t(\gamma). \tag{22}$$

Next, consider a market equilibrium.[11] Recall from Subsection 3.4 that the equilibrium allocation is given by Expert 1's effort e_1^* and referral threshold $\widehat{s}^*$, and Expert 2's zero effort and lack of referral. Each expert uses the first-best cost effort for a γ net cost reduction. The total expected cost per client in the equilibrium is

$$0.5\left\{\begin{array}{c} \int_{\underline{s}}^{\widehat{s}^*}[f_1(x|e_1^*)c_L+f_2(x|e_1^*)c_H]\mathrm{d}x+ \\ \int_{\widehat{s}^*}^{\overline{s}}[f_1(x|e_1^*)(c_L+\Delta)+f_2(x|e_1^*)(c_H-\Delta)]\mathrm{d}x+\phi(e_1^*) \end{array}\right\}+0.5\left\{\frac{c_L+c_H}{2}\right\}-\gamma.$$

Here, each expert takes cost effort, so the net cost saving γ applies to each client. Expert 1 takes equilibrium effort e_1^*, and obtains a signal. The first integral is the expected cost of Expert 1's clients with signals below the equilibrium threshold $\widehat{s}^*$, and the second integral is the expected cost of Expert 1's referred clients. Expert 2 neither takes effort nor refers in equilibrium, so his expected cost is one half of the sum of c_L and c_H. We simplify the total equilibrium expected cost in the market equilibrium to

$$\left\{\frac{c_L+c_H}{2}\right\}-0.5\left\{\Delta\int_{\widehat{s}^*}^{\overline{s}}[f_2(x|e_1^*)-f_1(x|e_1^*)]\mathrm{d}x-\phi(e_1^*)\right\}-\gamma\equiv EC_m(\gamma). \tag{23}$$

[11] If there are many equilibria, pick any one for the comparison to follow.

Finally, from Subsection 2.3, we subtract γ from (4) evaluated at the first-best effort to obtain the expected cost under the first best, and we call this $EC_{fb}(\gamma)$. The following presents the tradeoff between the market and the expert organization under cost-transfer protocol (the proof in the Appendix).

Proposition 5 *The expected cost per client is lower under cost-transfer protocol than in a market equilibrium if and only if the net cost saving γ is below a threshold $\widehat{\gamma}$, $0 < \widehat{\gamma} < \Delta$.*

When net cost saving γ vanishes, the cost-transfer protocol achieves the first best. At $\gamma = 0$, we have $EC_t(\gamma) = EC_{fb}(\gamma)$. The market equilibrium never achieves the first best. However, the market equilibrium always achieves γ cost saving because each expert bears his own costs. As γ increases from 0, expected costs in the first best, market, and cost-transfer protocol fall. Both $EC_{fb}(\gamma)$ and $EC_m(\gamma)$ fall at a unit rate as γ increases. What about the expected cost $EC_t(\gamma)$? As γ increases, referrals become less often under cost-transfer protocol, and information effort becomes less important (see the last four lemmas). As a result, $EC_t(\gamma)$ falls at a rate less than 1 as γ increases. Beyond a critical value, expected cost of the market equilibrium becomes lower than cost-transfer protocol. The critical value $\widehat{\gamma}$ is obtained by the solution of $EC_m(\widehat{\gamma}) = EC_t(\widehat{\gamma})$.

We illustrate the three expected costs $EC_m(\gamma)$ and $EC_t(\gamma)$ and $EC_{fb}(\gamma)$ in the following Figure 2. The first-best cost $EC_{fb}(\gamma)$ is a parallel downward shift of the cost in the market $EC_m(\gamma)$. The cost $EC_t(\gamma)$ in the expert organization is at the first-best level at $\gamma = 0$, but decreases less steeply than the other two.

The basic economics principle is this. In the market, experts work hard to reduce their own costs, but an expert acquires private information, so referrals are subject to adverse selection. In an organization, an expert works hard only if he is responsible for the cost, but cost-transfer protocol avoids asymmetric information. According to Proposition 5, transfer-cost protocol in an organization performs better than the market if and only if cost comparative advantage is more important than cost saving.

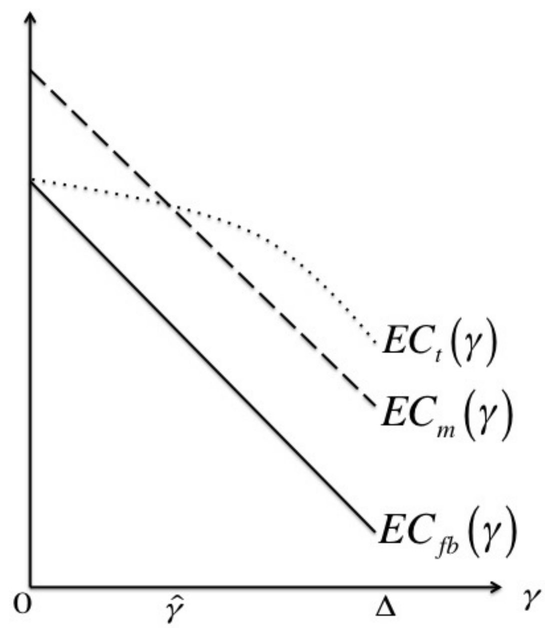

Figure 2: Expected cost and critical cost reduction $\widehat{\gamma}$

4.4 Perspectives on legal and medical organizations

Our theory offers a new perspective—tradeoff between adverse selection in the market and shirking within an organization. Proposition 5 predicts that experts form professional organizations when the cost comparative advantage from referrals is important, but that experts operate as solo-practitioners when work effort is more important. Using U.S. census data, Garicano and Hubbard (2009) show that lawyers form partnerships when they consult for corporations in markets such as banking, environment, and real estate developments. Our theory provides the foundation for the claim in Garicano and Hubbard (2009). The complexity in commercial dealing likely calls for disparate knowledge, so cross-field referrals are critical. Lawyers in domestic, insurance, and criminal ligitations more likely work as independent practitioners. Noncommercial cases are more idiosyncratic, so a lawyer's own effort is more critical.

Another illustration of our theory is in the malpratice-liability and personal litigations. According to Parikh (2006/2007), "top-end" lawyers in medical malpractice and product liability work in large practices, but "low-end" lawyers in automobile and "slip-and-fall" accidents work in solo practices. Our theory provides the rationale for the difference in the work organization of these lawyers. Top-end lawyers deal with more complex cases, so coordination between experts is important. By contrast, low-end lawyers may not have to rely on referrals that often.

Referral fees and fee-splitting are common among legal professionals, so our model applies straightforwardly. Nevertheless, our theory can also provide a normative view on the health care sector. In most countries, medical doctors are prohibited from obtaining financial benefits when they refer patients. The restriction is likely a safeguard against conflict of interests.[12] In our model, referral is a financial transaction, so it is inconsistent with the current practice. However, within Accountable Care Organizations, which are promoted by the U.S. healthcare reform, cost consequences of referrals are internalized. Proposition 5 implies that when referrals with prices are disallowed, ACOs may achieve more cost savings when cost comparative advantage is important. However, as Frandsen and Rebitzer (2015) point out, free-riding problems in ACOs can be severe, so cost comparative advantage must be balanced against muted work incentives in ACOs.

5 Robustness

We now discuss a number of robustness issues with the basic model of market referral. First, we assume only two states, ω_1 and ω_2. This can be regarded as a normalization given that we consider only two experts. If there are many (even a continuum of) states, then we proceed by first defining the subset of states for which Expert 1 is less expensive than Expert 2, and then call that subset ω_1. Second, we assume that the two states are equally likely. If they are not, the posterior probabilities in (1) will be modified by prior probabilities attached to the conditional densities f_1 and f_2. However, MLRP is unaffected, and it remains valid that Expert 2's cost of providing service to a client is lower than Expert 1 if and only if the client's signal is higher than a threshold. Our

[12] However, Pauly (1979) argues that referral with prices can improve patient welfare when markets are imperfectly competitive. Biglaiser and Ma (2007) show that a physician's self-referral at a price may lead to higher quality, thus benefitting some consumers.

computation is made easier by states being equally likely, but this assumption does not lead to any conceptual difficulties.

We have ignored capacities and variable returns. Here, there is another source of cost comparative advantage. The initial matching process may favor, say, Expert 1, who now has too many clients. Decreasing returns may lead him to refer some clients to Expert 2 even before he undertakes any effort (so has received no signal). It is a complication that may interfere with the construction of Expert 2's equilibrium belief about the referred clients' states. An analysis will have to start with the initial match between clients and experts. However, we feel that this is beyond the scope of our current research.

Capacity and variable returns may also change the comparison between market and organization. Clearly, an organization is better able to enjoy economies of scale, manage capacities, or both. The market is likely better modeled by a random initial match, but an organization can channel clients to its experts efficiently. The details in Proposition 5 may have to be altered but the basic principle of tradeoff between adverse selection and cost-reduction incentive remains valid.

In the rest of this section, we discuss two issues in details. First, we endogenize the tariff T rather than take it as given. And second, we study the equilibrium of the referral game when the cost advantage Δ is larger than the average cost, relaxing our assumption $\Delta < (c_L + c_H)/2$.

5.1 Equilibrium tariff

We modify the referral market game in Section 3 to include a Bertrand-competition game. That is, in Stage 0, at the time when the clients' types are drawn, each expert announces a tariff. Consumers observe these tariffs, and choose an expert for service. A consumer promises to pay the required tariff to the chosen expert or to a referred expert, if any, when service is provided.[13]

Recall that each expert can serve a client at an expected cost $(c_L + c_H)/2$. If an expert neither

[13] An expert cannot revise the tariff after he has exerted effort and obtained the signal. Both effort and signal are unobserved to the consumer. The issue of commitment is beyond the scope here. Furthermore, an expert cannot dump a client after the signal has been observed.

puts in effort nor refers a client, his tariff cannot be lower than $(c_L+c_H)/2$. Indeed, we now construct an equilibrium in which both experts set tariffs at $(c_L+c_H)/2$. Given this pair of (identical) tariffs, the continuation equilibrium is the market equilibrium in Section 3. Expert 2 neither exerts effort nor refers. When Expert 2 accepts a referral from Expert 1, his expected payoff is 0; see (12) in Proposition 1. Given Expert 1's tariff $(c_L + c_H)/2$, and the continuation equilibrium, it is optimal for Expert 2 to offer $(c_L + c_H)/2$.

Given that Expert 2 sets the tariff at $(c_L + c_H)/2$, Expert 1 will have no clients if he sets a higher tariff. Now suppose Expert 1 undercuts Expert 2 slightly, offering to provide service at a tariff just below $(c_L + c_H)/2$. All clients will solicit services from Expert 1. Expert 1 then follows the continuation equilibrium in Subsection 3.2 for each client. Therefore, in equilibrium Expert 1 will set the same tariff $(c_L + c_H)/2$, but all clients must first subscribe to Expert 1. After Expert 1 has observed a client's signal, he refers the client to Expert 2 if and only if the signal is higher than $\widehat{s}$, the equilibrium threshold in (12).

Our construction is similar to a standard Bertrand game with firms having different (and constant) marginal production costs: in equilibrium the more efficient firm sets a price equal to the marginal cost of the less efficient firm. Expert 1 is more efficient because he invests in information acquisition in the continuation equilibrium. Here, the "more efficient" Expert 1 sets the same tariff as the "less efficient" Expert 2, but takes all the surplus from trade. In equilibrium all clients initially seek services from Expert 1, who later refers some to Expert 2.[14]

5.2 Experts with large cost comparative advantage

We have assumed that the cost comparative advantage parameter Δ is smaller than $(c_H - c_L)/2$, so for both experts, the service cost in state ω_1 is lower than in state ω_2. This is our interpretation

[14] Clients do know about experts' cost comparative advantage. (In equilibrium, they pick Expert 1 even when both offer the same tariff.) If they do not, their strategies must not depend on experts' identities, so clients pick experts randomly when tariffs are identical. An equilibrium may then fail to exist when tariffs can be chosen from a continuous set. The more efficient Expert 1 always undercuts slightly. The usual way to restore existence is to discretize possible tariff offers. If the difference between possible tariffs is sufficiently small (like one cent), Expert 1 will just undercut to capture all clients when Expert 2 offers $T = (c_L+c_H)/2$.

for the state ω_1 being good and state ω_2 being bad. However, the value of Δ can be larger than $(c_H - c_L)/2$. In this case, we have $c_H - \Delta < c_L + \Delta$. For Expert 2, if the client's state is ω_1, the service cost becomes higher than if the state is ω_2. Now, to Expert 2 ω_1 looks like a bad state, while ω_2 looks like a good state (but the opposite is true for Expert 1). This cost specification actually allows equilibrium referrals from each expert to the other.

The derivation of Expert 1's equilibrium strategy, and Expert 2's beliefs remain unchanged, and Proposition 4 in Subsection 3.4 continues to hold. We only wish to note that Expert 2's expected cost of providing service is decreasing in Expert 1's referral threshold, so the expression in (10) is decreasing in $\widehat{s}$; in Figure 1, the dashed line is downward sloping.

For Expert 2, suppose now that he has taken effort e_2. We construct an equilibrium strategy for Expert 2's referral and effort. Again, in an equilibrium, Expert 1 accepts a referral if the price is below a threshold, say p_2. Given the tariff, if Expert 2 uses effort e_2 and receives signal s, he refers if and only if

$$p_2 \geq T - \frac{(c_L + \Delta)\, f_1(s|e_2)}{f_1(s|e_2) + f_2(s|e_2)} - \frac{(c_H - \Delta)\, f_2(s|e_2)}{f_1(s|e_2) + f_2(s|e_2)}. \tag{24}$$

By MLRP, and $c_L + \Delta > c_H - \Delta$, the expected cost in (24) is decreasing in s, so the right-hand side of (24) is increasing in s. By passive belief, in equilibrium, the signal threshold for Expert 2's referral at p_2 is $\widehat{s}_2$ such that (24) holds as an equality at $s = \widehat{s}_2$. In equilibrium, Expert 2 refers a client to Expert 1 if and only if $s < \widehat{s}_2$. This is the key difference. A higher value of the signal s indicates a higher likelihood of state ω_2. Expert 2's expected cost is decreasing in the signal, so he refers a client if and only if the signal is lower than a threshold. This is favorable news to Expert 1.

Expert 1 receives all those clients with signals below $\widehat{s}_2$, so he accepts Expert 2's referral if and only if

$$T - \frac{c_L \displaystyle\int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + c_H \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x}{\displaystyle\int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x} \geq p_2,$$

where the various integrals average out those signals below $\widehat{s}_2$ across the two states. Given effort

e_2, an equilibrium in referrals exists if there are price p_2 and threshold $\widehat{s}_2$ such that

$$T - \frac{(c_L + \Delta)\, f_1(\widehat{s}_2|e_2) + (c_H - \Delta)\, f_2(\widehat{s}_2|e_2)}{f_1(\widehat{s}_2|e_2) + f_2(\widehat{s}_2|e_2)} = p_2 = T - \frac{c_L \displaystyle\int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + c_H \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x}{\displaystyle\int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x}. \tag{25}$$

This is the characterization of the referral equilibrium for Expert 2, as Proposition 1 is for Expert 1. Such price p_2 and threshold $\widehat{s}_2$ satisfying (25) must exist. Indeed, by MLRP, $c_L < c_H$, and $c_L + \Delta > c_H - \Delta$, the ratio in the left-hand side of (25) is decreasing in $\widehat{s}_2$, while the ratio in the right-hand side is increasing.[15]

For the continuation equilibrium with price p_2 and threshold $\widehat{s}_2$, Expert 2's per-client expected payoff from effort e_2 can be simplified to

$$\left[T - \frac{c_L + c_H}{2}\right] + 0.5 \int_{\underline{s}}^{\widehat{s}_2} \left\{ [p_2 - (T - c_L - \Delta)] f_1(x|e_2) + [p_2 - (T - c_H + \Delta)] f_2(x|e_1)] \right\} \mathrm{d}x - \phi(e_2). \tag{26}$$

This has the same interpretation of Expert 1's expected payoff in (13). Expert 2's optimal effort is one that maximizes (26), and its first-order condition is

$$0.5 \int_{\underline{s}}^{\widehat{s}_2} \left\{ [p_2 - (T - c_L - \Delta)] \frac{\partial f_1(x|e_2)}{\partial e_2} + [p_2 - (T - c_H + \Delta)] \frac{\partial f_2(x|e_2)}{\partial e_2} \right\} \mathrm{d}x = \phi'(e_2). \tag{27}$$

Expert 2's equilibrium strategy is therefore characterized by price p_2, threshold $\widehat{s}_2$, and effort e_2 satisfying (25) and (27).

Using the same steps as in the proof of Proposition 4, we can show that Expert 2's equilibrium effort is never first best. Furthermore, given the equilibrium effort e_2, Expert 2's equilibrium referral threshold $\widehat{s}_2$ satisfies $f_2(\widehat{s}_2|e_2) < f_1(\widehat{s}_2|e_2)$, so Expert 2 sometimes retains a client even when his expected service cost is higher than Expert 1's.

[15] MLRP implies that the distribution f_1 is first-order stochastically dominated by f_2.

6 Conclusion

We posit a theory about how an organization can overcome market frictions due to hidden action and hidden information. This is a novel approach in the study of credence goods. The extant literature has looked at individual experts operating in a market to serve clients. There has been a lack of focus on how organizations may change experts' incentives. Although an organization can overcome adverse selection by the cost-transfer protocol, this leads to reduced work incentives. We derive a theory of the firm based on cost of adverse selection in the market compared to cost of reduced work incentive within an organization.

We have made some simplifying assumptions. It may be interesting to study the referral game when clients' benefits, not just their costs, are uncertain. Can referral convey information about benefits? Can a client rely on an expert to tell him that a service is not worthwhile? Our experts are profit maximizers. If one considers the health market as a specific application, physicians are known to be altruistic, so the pure profit-maximization assumption is invalid. It will be interesting to study how altruistic experts will play the referral game.

In the details of our model, we have also made a number of assumptions. Multiple rounds of information efforts are assumed away. Nor are multiple rounds of referral price offers allowed. We have also made use of the constant returns to scale in services. Any of these issues may be relaxed for a more general model.

Appendix

Proof of Lemma 3: By MLRP, $\dfrac{f_2(x|e_1)}{f_1(x|e_1)}$ is increasing in x, so for any s we have

$$\frac{\int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x} \equiv \frac{\int_s^{\bar{s}} \frac{f_2(x|e_1)}{f_1(x|e_1)} \cdot f_1(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x}$$

$$> \frac{\int_s^{\bar{s}} \frac{f_2(s|e_1)}{f_1(s|e_1)} f_1(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x} = \frac{f_2(s|e_1)}{f_1(s|e_1)}. \tag{28}$$

It follows that

$$\frac{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x} < \frac{f_1(s|e_1)}{f_1(s|e_1) + f_2(s|e_1)}$$

and

$$\frac{\int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x} > \frac{f_2(s|e_1)}{f_1(s|e_1) + f_2(s|e_1)}.$$

Therefore, at any $s < \bar{s}$,

$$\frac{c_L \int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x + c_H \int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x} > \frac{c_L f_1(s|e_1) + c_H f_2(s|e_1)}{f_1(s|e_1) + f_2(s|e_1)}. \tag{29}$$

Applying L'Hospital's rule, we have

$$\lim_{s \to \bar{s}} \frac{(c_L + \Delta) \int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta) \int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\bar{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\bar{s}} f_2(x|e_1)\mathrm{d}x} = \frac{(c_L + \Delta) f_1(\bar{s}|e_1) + (c_H - \Delta) f_2(\bar{s}|e_1)}{f_1(\bar{s}|e_1) + f_2(\bar{s}|e_1)}$$

$$= \frac{c_L f_1(\bar{s}|e_1) + c_H f_2(\bar{s}|e_1) - \Delta[f_2(\bar{s}|e_1) - f_1(\bar{s}|e_1)]}{f_1(\bar{s}|e_1) + f_2(\bar{s}|e_1)} < \frac{c_L f_1(\bar{s}|e_1) + c_H f_2(\bar{s}|e_1)}{f_1(\bar{s}|e_1) + f_2(\bar{s}|e_1)}.$$

We have shown that at s sufficiently near $\bar{s}$ (10) is smaller than (9).

Now at $\underline{s}$, we have

$$\frac{(c_L + \Delta) \int_{\underline{s}}^{\bar{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta) \int_{\underline{s}}^{\bar{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\underline{s}}^{\bar{s}} f_1(x|e_1)\mathrm{d}x + \int_{\underline{s}}^{\bar{s}} f_2(x|e_1)\mathrm{d}x} = \frac{c_L \int_{\underline{s}}^{\bar{s}} f_1(x|e_1)\mathrm{d}x + c_H \int_{\underline{s}}^{\bar{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\underline{s}}^{\bar{s}} f_1(x|e_1)\mathrm{d}x + \int_{\underline{s}}^{\bar{s}} f_2(x|e_1)\mathrm{d}x}$$

$$> \frac{c_L f_1(\underline{s}|e_1) + c_H f_2(\underline{s}|e_1)}{f_1(\underline{s}|e_1) + f_2(\underline{s}|e_1)}.$$

We have shown that at s sufficiently near $\underline{s}$, (10) is larger than (9). Therefore, the equation in the lemma must have a solution $\widehat{s}$.

Finally, for uniqueness, rewrite the equation in the lemma as

$$\frac{(c_L+\Delta)+(c_H-\Delta)\dfrac{\left[\displaystyle\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x\right]/f_2(s|e_1)}{\left[\displaystyle\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x\right]/f_1 s|e_1)}\bullet\dfrac{f_2(s|e_1)}{f_1(s|e_1)}}{1+\dfrac{\left[\displaystyle\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x\right]/f_2(s|e_1)}{\left[\displaystyle\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x\right]/f_1 s|e_1)}\bullet\dfrac{f_2(s|e_1)}{f_1(s|e_1)}}=\frac{c_L+c_H\dfrac{f_2(s|e_1)}{f_1(s|e_1)}}{1+\dfrac{f_2(s|e_1)}{f_1(s|e_1)}}. \tag{30}$$

By MLRP, the inverse hazard rates satisfy $\left[\displaystyle\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x\right]/f_2(s|e_1) > \left[\displaystyle\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x\right]/f_1(s|e_1)$; see also (28) above. As s changes, the rates of change of the left-hand and right-hand sides of (30) will never be identical. As separate functions, the graphs of (9) and (10) can only cross each other once. In other words, there can only be one solution.

Proof of Proposition 1: The two equations in (12) include the equation in Lemma 3, which already establishes a solution for $\widehat{s}$. We then set the value of p_1 according to (12). From Lemma 1, equilibrium referrals are those with signals above a threshold, so we simply set Expert 1's referral threshold at $\widehat{s}$. From Lemma 1, Expert 2 accepts a referral if and only if the price is below a threshold, so we set Expert 2's acceptance threshold at p_1.

Proof of Proposition 2: Assume, to the contrary, that Expert 2 exerts a strictly positive effort e_2 in an equilibrium. By Lemma 1, Expert 2 refers a client if and only if the client's signal is above a threshold, say $\widetilde{s}$. Let this referral be made at a price p_2 which will be accepted by Expert 1 in equilibrium.

At signal $\widetilde{s}$, Expert 2's expected cost is $(c_L+\Delta)\Pr(\omega_1|\widetilde{s},e_2)+(c_H-\Delta)\Pr(\omega_2|\widetilde{s},e_2)$

$$\begin{aligned}
&= \frac{(c_L+\Delta)f_1(\widetilde{s}|e_2)+(c_H-\Delta)f_2(\widetilde{s}|e_2)}{f_1(\widetilde{s}|e_2)+f_2(\widetilde{s}|e_2)} \\
&< \frac{(c_L+\Delta)\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+(c_H-\Delta)\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x} \qquad (31)\\
&< \frac{c_L\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+c_H\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x} \qquad (32)
\end{aligned}$$

where the inequality in (31) follows from MLRP (see also (29) in the proof of Lemma 3). Now the derivative of (31) with respect to Δ is

$$\frac{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x-\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}<0,$$

where the inequality is due to $f_2(\bullet|e_2)$ first-order stochastically dominating $f_1(\bullet|e_2)$, an implication of MLRP. Hence, (31) is decreasing in Δ. By reducing the value of Δ to zero, we obtain (32), Expert 1's expected cost of providing service to a client conditional on Expert 2's signal being at least $\widetilde{s}$.

In sum, because

$$\frac{(c_L+\Delta)f_1(\widetilde{s}|e_2)+(c_H-\Delta)f_2(\widetilde{s}|e_2)}{f_1(\widetilde{s}|e_2)+f_2(\widetilde{s}|e_2)}<\frac{c_L\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+c_H\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}$$

it is impossible to find p_2 to satisfy

$$T-\frac{(c_L+\Delta)f_1(\widetilde{s}|e_2)+(c_H-\Delta)f_2(\widetilde{s}|e_2)}{f_1(\widetilde{s}|e_2)+f_2(\widetilde{s}|e_2)}\le p_2\le T-\frac{c_L\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+c_H\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x+\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x} \qquad (33)$$

a condition for an equilibrium. This is a contradiction.

Proof of Proposition 3: First, $(e_1^*,\widehat{s}^*)$ maximize Expert 1's expected utility (13) given Expert 2's acceptance threshold p_1^* so this is a best response. Second, p_1^* is given by (12) at $\widehat{s}$ and e_1^*, so acceptance threshold p_1^* is a best response. Expert 2's belief clearly satisfies passive belief.

We now show that Expert 2's best response is to choose no effort. Given the most pessimistic belief, Expert 1 will reject any referral price p where $T - \Pr(\omega_1|\overline{s}, \widetilde{e})c_L - \Pr(\omega_2|\overline{s}, \widetilde{e})c_H < p$, so the minimum price for Expert 1 to accept a referral is $p = T - \Pr(\omega_1|\overline{s}, \widetilde{e})c_L - \Pr(\omega_2|\overline{s}, \widetilde{e})c_H$.

Suppose that Expert 2 takes some effort, say $e_2 > 0$. Expert 2's expected utility from keeping the client at signal s is $T - \Pr(\omega_1|s, e_2)(c_L + \Delta) - \Pr(\omega_2|s, e_2)(c_H + \Delta)$. By definition, $\Pr(\omega_2|s, e_2) \leq \Pr(\omega_2|\overline{s}, \widetilde{e})$, so we have

$$\begin{aligned}
\Pr(\omega_1|s, e_2)(c_L + \Delta) + \Pr(\omega_2|s, e_2)(c_H - \Delta) &\leq \Pr(\omega_1|\overline{s}, \widetilde{e})(c_L + \Delta) + \Pr(\omega_2|\overline{s}, \widetilde{e})(c_H - \Delta) \\
&< \Pr(\omega_1|\overline{s}, \widetilde{e})c_L + \Pr(\omega_2|\overline{s}, \widetilde{e})c_H.
\end{aligned}$$

Therefore,

$$T - \Pr(\omega_1|s, e_2)(c_L + \Delta) - \Pr(\omega_2|s, e_2)(c_H - \Delta) > T - \Pr(\omega_1|\overline{s}, \widetilde{e})c_L - \Pr(\omega_2|\overline{s}, \widetilde{e})c_H = p.$$

Expert 2 cannot profit from deviating to an effort and referring some clients to Expert 1.

It remains to show that the triple $[e_1^*, \widehat{s}^*, p_1^*]$ exists. We use a standard fixed-point argument. Bound Expert 1's feasible efforts by a compact convex set, say a closed interval of the real numbers. Clearly we can let the referral threshold reside in the signal support, which is convex and compact. Finally, we can also let the referral price be an element of a compact convex set of real numbers. Define a map Ψ that takes an effort, a referral threshold, and a price onto themselves: $\Psi(e_1, \widehat{s}, p_1) = (e_1', \widehat{s}', p_1')$, where we define Ψ by

$$(e_1', \widehat{s}') = \underset{e_1, \widehat{s}}{\operatorname{argmax}}\, 0.5 \int_{\widehat{s}}^{\overline{s}} \{[p_1 - (T - c_L)] f_1(x|e_1) + [p_1 - (T - c_H)] f_2(x|e_1)]\} \,\mathrm{d}x - \phi(e_1) \quad (34)$$

$$p_1' = T - \frac{(c_L + \Delta)\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta)\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}. \quad (35)$$

Here, (34) is Expert 1's best response against Expert 2's referral-acceptance price p_1 (the same as the maximization of (13) with respect to effort and referral threshold), while (35) is Expert 2's referral-acceptance best response against Expert 1's effort e_1 and referral threshold $\widehat{s}$ (see also (12) in Proposition 1).

Clearly, the Maximum Theorem applies to (34), and there is a selection of the solution $(e_1', \widehat{s}')$ which is continuous in p_1. Furthermore, p_1' in (35) is obviously continuous in e_1 and $\widehat{s}$. By Brouwer's Fixed Point Theorem, Ψ has a fixed point $(e_1^*, \widehat{s}^*, p_1^*)$.

Proof of Proposition 4: Suppose not, i.e., suppose that in an equilibrium Expert 1's effort and referral threshold are first best. Then $f_2(\widehat{s}^*|e_1^*) = f_1(\widehat{s}^*|e_1^*)$; see Subsection 2.3. From the second equation in (15) we obtain

$$\begin{aligned} p_1^* - (T - c_L) &= -\frac{(c_H - c_L) f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} = -\frac{c_H - c_L}{2} \\ p_1^* - (T - c_H) &= \frac{(c_H - c_L) f_1(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} = \frac{c_H - c_L}{2}. \end{aligned}$$

We then write (16) as

$$0.5 \left[\frac{c_H - c_L}{2}\right] \int_{\widehat{s}^*}^{\overline{s}} \left\{ \frac{\partial f_2(x|e_1^*)}{\partial e_1} - \frac{\partial f_1(x|e_1^*)}{\partial e_1} \right\} dx = \phi'(e_1^*).$$

However, by assumption $c_H - c_L > 2\Delta$. Comparing this simplified (16) with (5), we conclude that $e_1^* > e^{fb.}$, so Expert 1's effort is not first best.

Next, suppose, to the contrary, that $f_2(\widehat{s}^*|e_1^*) \le f_1(\widehat{s}^*|e_1^*)$. First, we note that

$$\frac{(c_L + \Delta) f_1(\widehat{s}^*|e_1^*) + (c_H - \Delta) f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \equiv \frac{c_L f_1(\widehat{s}^*|e_1^*) + c_H f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} + \frac{\Delta[f_1(\widehat{s}^*|e_1^*) - f_2(\widehat{s}^*|e_1^*)]}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)}.$$

Therefore, by $f_2(\widehat{s}^*|e_1^*) \le f_1(\widehat{s}^*|e_1^*)$, we have

$$\frac{c_L f_1(\widehat{s}^*|e_1^*) + c_H f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \le \frac{(c_L + \Delta) f_1(\widehat{s}^*|e_1^*) + (c_H - \Delta) f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)}.$$

Now by MLRP, we have (28):

$$\frac{\displaystyle\int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x}{\displaystyle\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x} > \frac{f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*)}.$$

It follows that

$$\begin{aligned} & \frac{(c_L + \Delta)\displaystyle\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x + (c_H - \Delta)\int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x}{\displaystyle\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x + \int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x} \\ > \quad & \frac{(c_L + \Delta) f_1(\widehat{s}^*|e_1^*) + (c_H - \Delta) f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \\ \ge \quad & \frac{c_L f_1(\widehat{s}^*|e_1^*) + c_H f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)}, \end{aligned}$$

which contradicts (15). We conclude that $f_2(\widehat{s}^*|e_1^*) > f_1(\widehat{s}^*|e_1^*)$.

Proof of Lemma 4: The expression in the lemma is obtained from solving for the f_1/f_2 ratio in (18). Clearly, at $\gamma = 0$, we have $f_1(\widetilde{s}_1|e_1) = f_2(\widetilde{s}_1|e_1)$, so $\widetilde{s}_1$ is the first-best referral threshold at effort e_1: see $f_1(\widehat{s}^{fb}|e) = f_2(\widehat{s}^{fb}|e)$ in Subsection 2.3. The right-hand side of (19) is strictly decreasing in γ, and goes to 0 as γ increases to Δ. By MLRP, we conclude that $\widetilde{s}_1$ must increase to $\overline{s}$.

Proof of Lemma 5: We drop all constants (those that involve only c_L and c_H) in (20) and then simplify it to obtain

$$-\int_{\underline{s}}^{\widetilde{s}_1} 0.5\gamma[f_1(x|e_1) + f_2(x|e_1)]\mathrm{d}x - \int_{\widetilde{s}_1}^{\overline{s}} 0.5\Delta[f_2(x|e_1) - f_1(x|e_1)]\mathrm{d}x + \phi(e_1).$$

Differentiating this with respect to e_1 and setting it to zero, we get the first-order condition:

$$0.5\left\{\int_{\underline{s}}^{\widetilde{s}_1} \gamma\left[\frac{\partial f_1(x|e_1)}{\partial e_1} + \frac{\partial f_2(x|e_1)}{\partial e_1}\right]\mathrm{d}x + \int_{\widetilde{s}_1}^{\overline{s}} \Delta\left[\frac{\partial f_2(x|e_1)}{\partial e_1} - \frac{\partial f_1(x|e_1)}{\partial e_1}\right]\mathrm{d}x\right\} = \phi'(e_1). \qquad (36)$$

Now the first-best information effort is given by (5), and we conclude that e_1 is never first best except at $\gamma = 0$.

By Lemma 4, $\widetilde{s}_1$ tends to $\overline{s}$ as γ tends to Δ. The first integral in (36) becomes arbitrarily small because the integrands are derivatives of densities, which sum to 0 over the support. Obviously, the second integral tends to 0. Hence, any e_1 satisfying (36) must tend to 0.

Proof of Proposition 5: First, at $\gamma = 0$, $EC_t(\gamma)$ in (22) is the expected cost at the first best (5). Also, at $\gamma = \Delta$, $\widetilde{s}_1 = \overline{s}$, $\widetilde{s}_2 = \underline{s}$, so $EC_t(\gamma)$ in (22) equals $\left\{\dfrac{c_L + c_H}{2}\right\} - \gamma$. Because $EC_m(\gamma)$ is the market equilibrium expected cost, it is higher than the first best. Hence, $EC_m(0) > EC_t(0)$. By inspection, we have $EC_m(\Delta) < EC_t(\Delta)$.

Next, because the market equilibrium is independent of γ, $EC_m(\gamma)$ has a derivative of -1. The expected cost in $EC_t(\gamma)$ is the result of optimal choices of information effort and referral threshold, so the envelope theorem applies. The derivative of $EC_t(\gamma)$ is the partial derivative of (22) with respect to γ:

$$\frac{\mathrm{d}EC_t(\gamma)}{\mathrm{d}\gamma} = -0.5\left\{\int_{\underline{s}}^{\widetilde{s}_1}[f_1(x|\widetilde{e}_1) + f_2(x|\widetilde{e}_1)]\mathrm{d}x + \int_{\widetilde{s}_2}^{\overline{s}}[f_1(x|\widetilde{e}_2) + f_2(x|\widetilde{e}_2)]\mathrm{d}x\right\} > -1$$

where the inequality follows from $\widetilde{s}_1 < \overline{s}$ and $\widetilde{s}_2 > \underline{s}$. Hence, as γ varies between 0 and Δ, there is only point $\widehat{\gamma}$ such that $EC_m(\widehat{\gamma}) = EC_t(\widehat{\gamma})$. The proposition follows.

References

Able, B.H. (2013). The Stark Physician Self-Referral Law and Accountable Care Organizations: Collision Course or Opportunity to Reconcile Federal Anti-Abuse and Cost-Saving Legislation, *Journal of Law & Health*, 315-347.

Arbatskaya, M., & Konishi, H. (2012). Referrals in Search Markets, *International Journal of Industrial Organization,* 30(1), 89-101.

Bar-Isaac, H., Caruana, G., & Cunat, V. (2012). Search, Design and Market Structure. *The American Economic Review*, 102 (2), 1140-1160.

Biglaiser, G., & Ma, C-t.A. (2007). Moonlighting: Public Service and Private Practice. *The RAND Journal of Economics*, 38, 1113-1133.

Bolton, P., Freixas, X., & Shapiro, J. (2007). Conflicts of Interest, Information Provision, and Competition in the Financial Services Industry, *Journal of Financial Economics*, 85(2), 297–330.

Buehler, B., & Schuett F. (2014). Certification and Minimum Quality Standards when some Consumers are Uninformed, *European Economic Review*, 70, 493–511.

Cebul, R.D., Rebitzer, J.B., Taylor, L.J., & Votruba, M.E. (2008). Organizational Fragmentation and Care Quality in the U.S. Healthcare System, *Journal of Economic Perspectives*, 22(4), 93-113.

de Fontenay, C.C., & Gans, J.S. (2005). Vertical Integration in the Presence of Upstream Competition, *The RAND Journal of Economics*, 36(3), 544-572.

Dequiedt, V., & Martimort, D. (2015). Vertical Contracting with Informational Opportunism, *American Economic Review*, 105(7), 2141-82.

Dulleck, U., & Kerschbamer, R. (2006). On Doctors, Mechanics, and Computer Specialists: The economics of Credence Goods, *Journal of Economic Literature*, 44(1), 5-42.

Epstein, A. J. Ketcham, J. D., & Nicholson, S. (2010). Specialization and Matching in Professional Services Firms, *The RAND Journal of Economics*, 41(4), 811-834.

Frandsen, B. & Rebitzer, J.B. (2015). Structuring Incentives within Organizations: the Case of Accountable Care Organizations, *Journal of Law, Economics, and Organizations*, 31, 77-103.

Fuchs, W. & Garicano, L. (2010). Matching Problems with Expertise in Firms and Markets, *Journal of the European Economic Association*, 8(2-3), 354-364.

Garber, A. M. & Skinner, J. (2008). Is American Health Care Uniquely Inefficient? *The Journal of Economic Perspectives,* 22(4), 27–50.

Garicano, L. (2000). Hierarchies and the Organization of Knowledge in Production, *Journal of Political Economy*, 108(5), 874-904.

Garicano, L. & Hubbard, T. N. (2009). Specialization, Firms, and Markets: The Division of Labor Within and Between Law Firms, *Journal of Law, Economics, & Organization.* 25(2), 339-371

Garicano, L. & Santos, T. (2004). Referrals, *American Economic Review*, 94(3), 149-173.

Hart, O., & Tirole, J. (1990). Vertical Integration and Market Foreclosure, Brookings Papers on Economic Activity (Microeconomics), 205–285.

Inderst, R., & Ottaviani, M. (2009). Misselling through agents, *The American Economic Review*, 99(3), 883-908(26).

Inderst, R., & Ottaviani, M. (2012). Competition through commissions and kickbacks, *American Economic Review,* 102(2), 780-809.

Laffont, J.-J. & Martimort, D. (2000). Mechanism design with collusion and correlation, *Econometrica*, 68, 309–342.

Mariñoso, B.G., & Jelovac, I. (2003). GPs' Payment Contracts and their Referral Practice, *Journal of Health Economics*, 22(4), 617-35.

MacAfee, P. & Schwartz, M. (1994). Opportunism in Multilateral Vertical Contracting: Nondiscrimination, Exclusivity, and Uniformity, *American Economic Review*, 84(1), 210-230.

O'Brien, D. P., & Shaffer, G. (1992). Vertical Control with Bilateral Contracts, *The RAND*

Journal of Economics, 23, 299–308.

Park, I. (2005). Cheap-Talk Referrals of Differentiated Experts in Repeated Relationships, *The RAND Journal of Economics*, 36(2), 391–411.

Parikh, S. (2006/2007). How the Spider Catches the Fly: Referral Networks in the Plaintiffs' Personal Injury Bar, *New York Law School Law Review*, 51, 243-283.

Pauly, M. V. (1979). The Ethics and Economics of Kickbacks and Fee Splitting, *Bell Journal of Economics*, 10(1), 344-352.

Rey, P., & Tirole, J. (2007), A Primer on Foreclosure, in Mark Armstrong and Robert Porter, eds., *Handbook of Industrial Organization,* vol. 3, North-Holland, 2145–2220.

Rebitzer, J. B., & Votruba, M.E. (2011). Organizational Economics and Physician Practices, NBER Working Paper No. 17535.

Reisinger, M., & Tarantino, E. (2015) Vertical Integration, Foreclosure, and Productive Efficiency, *The RAND Journal of Economics*, 46(3).

Schmidt K.M. (1996). The Costs and Benefits of Privatization: An Incomplete Contracts Approach, *The Journal of Law, Economics & Organization*, 12(1), 1-24.

Shumsky, R. A., & Pinker, E. J. (2003). Gatekeepers and Referrals in Services, *Management Science* 49(7), 839-856.

Song, Z., Sequist T.D., & Barnett, M.L. (2014), Patient Referrals: A Linchpin for Increasing the Value of Care. *JAMA*. 312(6), 597-598.