Information Acquisition, Referral, and Organization

Simona Grassi

HEC, Institut d'Economie et
de Management de la Santé
Université de Lausanne

simona.grassi@unil.ch

Ching-to Albert Ma

Department of Economics
College of Arts and Sciences
Boston University

ma@bu.edu

February 2014

Abstract

There are two experts. Each may exert a costly effort to acquire information about clients' service costs. An expert's effort results in a signal which indicates which expert has a lower expected cost. An expert's effort is hidden action, and his signal is hidden information. In a market setting, an expert, after receiving the private signal, may refer the client to the other for a fee. When experts have comparative cost advantage in the market equilibrium, only one expert exerts costly effort and refers successfully. Equilibrium effort and referral decisions are never first best. We show that organizations can achieve the first best. These organizations include integrated firm, partnership, and mutual agreement. Each organization makes reassignment of experts' costs possible. Integration also sets up a gatekeeping protocol in which the owner-expert invests in effort to acquire information, and then refers some clients to the employee-expert. In a partnership or under a mutual agreement, each expert exerts effort and refers efficiently.

Keywords: information acquisition, referral, organization, integration, partnership

JEL: D00, D02, D80, D83

Acknowledgement: For their comments and suggestions, we thank Francesca Barigozzi, Jean-Philippe Bonardi, Giacomo Calzolari, Mark Dusheiko, Izabella Jelovac, Andrew Jones, Henry Mak, Debby Minehart, Liisa Väisänen, Peter Zweifel, and seminar participants at the Antitrust Division of the US Department of Justice, Boston University, the Final Report Seminar for Harkness fellows 2012-2013 in New York, Indiana University and Purdue University (IUPUI), the Nordic Health Economists' Study Group Workshop in Oslo, the Norwegian School of Economics (NHH) in Bergen, the Ph.D. Seminar in Health Economics and Policy in Grindelwald (Switzerland), and the University of Lausanne. Simona Grassi is indebted to the Commonwealth Fund, the Careum Foundation, and to her mentor Joe Newhouse while Harkness/Careum fellow at the Harvard Medical School.

1 Introduction

An economic system aims at performing tasks in a cost-effective way. We consider a system consisting of experts who provide services to clients. One expert may be less expensive in providing services to some clients than another expert. Furthermore, an expert may need to invest in effort to find out who is less expensive. In this paper, we study i) an expert's incentives to acquire information about clients, and ii) whether an expert has the incentive to refer clients to another expert when doing so can save costs.

Incentives on information acquisition and task assignment have occupied a topical role in many policy discussions. For example, in the U.S. health care reform, it is expected that cost-control measures will be phased in after the Affordable Care Act takes 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 and specialists, to form organizations to provide care to enrollees. (These are called Accountable Care Organizations.[1]) It is thought that coordination among providers will eliminate fragmentation and save costs.[2] Other professionals, such as accountants and lawyers, form such organizations and partnerships. Our paper is a framework to understand information acquisition and referral incentives in professional organizations.

In our model, each of a set of clients would like to obtain service from one of two experts. Each client pays a fixed tariff for the service to be provided by an expert. 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 comparative cost advantage: 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, however, is unknown. An expert may exert some

[1] See http://www.cms.gov/ACO.

[2] 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..

effort to obtain information about the case, and the effort generates a signal that indicates the likelihood of an easy case.

Here is a familiar (and stylized) example. A client has to file a tax return. Either a tax preparer or a tax lawyer can perform the filing. The filing can be easy or complicated. Although an easy case requires a lower service cost, the tax preparer has an edge. Conversely, if the case is complicated, the tax lawyer has the edge, even though complicated cases are more expensive. For example, for an easy case, the preparer's cost is \$100, while it is \$120 for the lawyer; for a complicated case, the preparer's cost is \$200, while it is \$180 for the lawyer. Each tax professional may spend some time on an initial assessment, which generates some information about the case complexity.

In this environment, we first study how experts operate in a market, in which experts operate independently but referrals are possible. After Expert 1, say, has observed a signal due to a diagnostic effort, he may make an offer to Expert 2: for a price, the client (and the associated service tariff) is transferred from Expert 1 to Expert 2 if Expert 2 pays the referral price.

The problem 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. The referral price by Expert 1 does transmit some information. Expert 1 must have received a signal that indicates his service cost is high, so that he is better off selling. Expert 2, accordingly, must believe that the likelihood for the complicated case is high, although he does not know exactly what Expert 1 has observed.

Asymmetric information does generate friction in the functioning of the referral market. However, it is common knowledge that Expert 2 has a comparative cost advantage if the case is complicated. Missing information about signals for the complicated case does not nullify this gain from trade. Indeed, we present an equilibrium in which Expert 1 successfully refers clients to Expert 2 if and only if Expert 1's signal indicates a high likelihood for a complicated case. Furthermore, Expert 1 does have an incentive to acquire information about the client.

Nevertheless, Expert 1's equilibrium referral decision and effort are never first best. Hidden

information about effort and signal does lead to distortion and inefficiencies. The basic point is that Expert 1 does not internalize all the cost savings due to comparative cost advantage. The profit motive cannot be aligned with the social cost-minimization objective.

Expert 2's equilibrium strategies in our basic model, however, are completely different. He will neither exert effort nor make any referral. The comparative cost 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. The basic fact is that a simple case is more profitable than a complicated case. If in equilibrium Expert 2 was successful at referring a client with a signal indicating a low likelihood of a complicated case, he would also refer the client if the signal indicated any higher likelihood. Expert 2's attempt to let Expert 1 exploit his comparative cost advantage is incredible. Without any success in referral in equilibrium, Expert 2 does not exert effort.

We then study how the experts can do better by forming an organization. What is it that makes the referral market equilibrium inefficient? First, an expert is unconcerned about the cost consequence to be borne by the expert accepting the referral. Second, and perhaps less obvious, is the way the clients are initially assessed. In the market, clients may be randomly matched with experts. We propose that there are two changes from market transactions when experts interact within an organization. First, an organization can make cost information available *ex post*. Second, an organization can establish assessment and referral protocols. At the same time, we maintain the assumption that an expert's effort and the acquired information are private information.

Our first solution has an expert integrating with the other. Suppose that Expert 1 buys out Expert 2; Expert 1 becomes the owner of this integrated firm and Expert 2 becomes an employee. Any cost that Expert 2 incurs in providing service to a client can now be transferred to Expert 1. This is the organizational advantage: Expert 2's cost *ex post* can now be documented and be made Expert 1's responsibility. In other words, Expert 1 now fully internalizes any saving from exploiting each expert's comparative cost advantage. Furthermore, in this integrated firm, Expert 1 will also be the gatekeeper. Expert 1 assesses all clients initially. He alone will exert the effort to acquire information. Expert 1 then gets to decide, based on the client's realized signal from effort, whether

Expert 2 or he himself will provide the service. Expert 2 will simply follow instructions, providing service to any client that Expert 1 cares to send to him. We show that integration achieves the first best.

Integrated expert organizations are not the only ones that achieve the first best. We discover that experts can form a partnership to do the same.[3] In a partnership, the experts are joint owners, and agree upon a sharing rule that splits any net proceed among themselves in all contingencies. As in an integrated organization, experts' *ex post* cost information can be used to set up the sharing rule. A partnership can also keep track of referrals so that the sharing rule can be based on whether an expert has provided service to his own client, or to a referred client. The key construction has an expert being made responsible for the cost incurred by the other expert upon a referral. That is, if Expert 1 refers a client to Expert 2, Expert 1 will have to reimburse Expert 2's cost after Expert 2 has provided service. This is a way for an expert to internalize the saving due to comparative cost advantage. Last, we show that mutual agreements that mimic the partnership contract can also achieve the first best.

Much of information economics is about information acquisition, and how hidden information impacts efficiency. The canonical model is concerned with a principal drawing an incentive contract to motivate an agent to acquire information and then truthfully reveal it. This approach has been studied extensively, and has been applied to such settings as procurement, accounting system, auction, financial service, project selection, etc. Demski and Sappington (1987) are pioneers in the study of delegated expertise and derive optimal contracts for an agent who chooses a costly effort and learns private information from it. In the context of procurement, Crémer and Khalil (1992) are among the first to study information gathering before contracts are signed; see also Crémer, Khalil, and Rochet (1998a, 1998b). It is impossible to provide a comprehensive list of papers in this literature here, but more recent contributions include Bergemann and Välimäki (2002), Dai, Lewis, and Loppomo (2006), Szalay (2009) and Iossa and Martimort (2013).

[3]Partnerships have been studied in Holmström (1982), Legros and Matthews (1993), Levin and Tadelis (2002), among others.

Our approach differs from that of optimal contract or mechanism design. Neither clients nor experts propose any incentive contract. For clients, the only verifiable state is whether a service has been provided, and they pay a fixed fee. For experts, in the market setting, the only verifiable state is whether a client has been referred. The point here is that organizations are the mechanisms to solve problems due to hidden action and hidden information.[4] We directly identify integrated firms and partnerships for achieving the first best. More fundamentally, we discover how organizations can achieve efficiency via making cost information verifiable within, and setting gatekeeping protocols.

We use the credence-good framework. In this literature, the assumption of experts having perfect information on clients, either given exogenously or acquired through a costly effort, is common. Another common feature is that clients not only pay a price for service, but also a fee for being diagnosed by an expert; see Dullek and Kerschbamer (2009), Emons (2001), Fong (2005), Liu (2011), Pitchik and Schotter (1987), Sülzle and Wambach (2005). Dulleck and Kerschbamer (2006) have written a recent survey of this literature.

By contrast, experts in our model obtain a noisy signal about clients. Neither do they charge any diagnostic fees. The contractible events in our model are fewer than in other models in the literature. We assume that either expert provides the same benefit to a client. Prices in the referral market convey cost information between experts but this information does not affect clients' benefits (see for example Bolton, Freixas, and Shapiro (2007) in the context of financial services.)

Our model is related to the study of referrals in Garicano and Santos (2004). In their model, two experts, with different productivities, are matched with a set of projects (which they call opportunities). Each project requires effort inputs from an expert. The more productive expert should handle the more promising projects. The basic problem is hidden information about project characteristics and experts' hidden input efforts. They consider referrals and partnerships, but experts' learning about project characteristics does not involve any decision or effort. We find different kinds of distortions in the market referral game, but that many organizations can remedy

[4]It follows trivially that applying the (generalized) Revelation Principle to our model will also yield first-best allocations. Nevertheless, we have no need to relate an abstract direct mechanism to institutions.

such distortions.

We contribute to the research in organizational economics and also provide practical contributions to the design of expert organizations. Subsection 2.3 provides many examples of such organizations, but the medical profession is perhaps the most topical as the U.S. Affordable Care Act actively promotes an organizational approach. In an empirical study, Currie and MacLeod (2013) demonstrates that better diagnostic skills improve the matching between patients and procedures leading to better health outcomes. Epstein, Ketcham, and Nicholson (2010) test whether group or solo obstetric practices are better at coordinating the match between physicians and patients. They find evidence that firms overcome asymmetric information and institutional barriers in order to achieve higher levels of specialization and coordination, which benefit patients.

Our paper presents a theory about how expert organizations can do better than the market. Within an organization, costs become verifiable, and the flow of clients from one expert to the other can be controlled. These two together enable an organization to achieve the first best. Our study is related to Garicano (2000) who studies how hierarchical organizations better assign problems to agents, and to Fuchs and Garicano (2010) who show that repeated interaction between experts within a firm leads to better information about experts' qualities. Our focus is on information acquisition, and how expert organizations can make use of this information.

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 that implement the first best. In Section 5 we consider a number of robustness issues. We first endogenize consumers' tariff for service by letting experts compete against each other by setting prices. We then consider a model with a more extreme cost-parameter configuration. Under this set of cost parameters, both experts exert efforts and refer successfully in a market equilibrium, but the allocation remains inefficient. Section 6 concludes.

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 seeking services from professionals such as lawyers, doctors, or engineers. Alternatively, a company may have a set of 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. We abstract from vertical differentiation issues. Although we can consider matching clients to experts who provide different service qualities at different costs (as in, for example, Liu, Ma and Mak (2014)), for tractability, we assume away variable benefits. In any case, this may be plausible for many applications; see Subsection 2.3.

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:

	ω_1	ω_2
Expert 1	c_L	c_H
Expert 2	$c_L + \Delta$	$c_H - \Delta$

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 or $c_L + \Delta$. State ω_2 corresponds to a "bad" or "complicated" state with service cost either $c_H - \Delta$ or c_H. However, Expert 1 has a comparative cost advantage in state ω_1, while Expert 2 has an advantage in state ω_2. If a client's state is ω_1, Expert 1's cost advantage compared to Expert 2 equals Δ, the difference between $c_L + \Delta$ and c_L. Analogously, if a client's state is ω_2, Expert 2's cost advantage over Expert 1 is Δ.

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. Hence, we will let each client pay a fixed tariff to an expert when a service is provided.

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, where ϕ is a strictly increasing and strictly convex function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$. We make the common assumption that $\phi(0) = \phi'(0) = 0$, which ensures that some effort is worthwhile.

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. 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)*:

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

As a normalization, we let the signals be completely uninformative at the least effort level, $e = 0$, 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:

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

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 Illustrative examples

In the Introduction, we already discussed an example with a tax preparer and a tax lawyer. We offer two more examples. Expectant mothers would like to be taken care of by either a midwife or an obstetrician. Each expectant mother values a successful pregnancy. However, each woman's health status is uncertain. State ω_1 can be regarded as a less complicated case of pregnancy; the midwife will be able to take care of the expectant mother at a cost lower than the obstetrician. State ω_2 can be regarded as a more complicated case of pregnancy. Here, the obstetrician has a cost advantage over the midwife. The comparative cost advantage between the two providers depends on the

health status. Each medical provider can carry out various tests and spend time to find out more about expectant mothers' health statuses. This corresponds to the information acquisition process. Similar situations in the health market can be found in cost comparisons between psychiatrist and psychologist, as well as physical therapist and orthopedic surgeons, cardiologist and cardiac surgeons, etc.

Next, consider a home owner with a home-improvement project. He needs to implement his idea for renovating a room. Either a carpenter or contractor can do the work. The carpenter is Expert 1 with a cost advantage if the project is easy, while the contractor is Expert 2 with a cost advantage if the project is difficult. Each expert can do some preliminary work to assess the difficulty of the task.

Our model is theoretical and not meant to capture all details in each of the above examples. However, in each situation above, it is plausible that experts should take some steps to find out the best cost-effective method of service delivery. The next subsection derives the first-best allocation of information acquisition effort for each expert and the assignment of clients to experts.

2.4 First best

An allocation is an effort to be taken by each expert, and a decision rule that assigns a client to an expert according to the signal that has resulted from an expert effort. The first best is an allocation—efforts and decision rules—that minimizes the expected cost of providing services to the clients by the two experts and their effort disutilities.

Let each expert take an effort e. Contingent on signal s, the conditional probability that the client's state is ω_i is given by (1), so the expected cost of servicing this client by Expert 1 and 2 are respectively,

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

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

It follows that Expert 2's cost is lower if and only if $f_1(s|e) \leq f_2(s|e)$. Hence, for any effort, the

cost-minimizing allocation assigns a client to Expert 2 if and only if $f_1(s|e) \leq f_2(s|e)$. For each effort e, define $\widehat{s}^{fb}(e)$ by $f_1(\widehat{s}^{fb}|e) = f_2(\widehat{s}^{fb}|e)$. 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

$$.5\int_{\underline{s}}^{\widehat{s}^{fb}(e)}\{\Pr(\omega_1|x,e)c_L+\Pr(\omega_2|x,e)c_H\}[f_1(x|e)+f_2(x|e)]\mathrm{d}x+$$
$$.5\int_{\widehat{s}^{fb}(e)}^{\overline{s}}\{\Pr(\omega_1|x,e)(c_L+\Delta)+\Pr(\omega_2|x,e)(c_H-\Delta)\}[f_1(x|e)+f_2(x|e)]\mathrm{d}x+\phi(e). \quad (4)$$

We assume that (4) is quasi-convex. The first-best effort, e^{fb}, is one that minimizes (4). The first-order condition is:

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

The first-best characterization illustrates the trade-off. 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 higher than a threshold is for cost effectiveness. Third, a higher effort leads to more disutility, but also raises signal informativeness. The left-hand side of (5) reflects the informativeness 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) obviously is the marginal disutility of effort.

In this paper, we study if and how the first best can be implemented under missing information. We assume that an expert's effort to acquire information is unobservable. Furthermore, we let each expert privately observe the signal from his effort. The first best then requires that each expert privately takes an effort, and then assigns the client either to himself or the other expert according to the signal he privately observes. We consider two types of institutions. First, we analyze a market

for referrals, and second, we consider organizations based on prespecified financial agreements.

We assume that clients pay a fixed, exogenous 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. Each client has no reason to choose the less expensive expert. In Subsection 5.1, we let the experts set tariffs and compete in a Bertrand fashion. Our results are unchanged with endogenously chosen tariffs. We will show that while the referral market does not implement the first-best allocation, some organizational arrangements between experts may achieve efficiency despite dull incentives from clients.

3 Referral market

In this section, we study a referral market. The two experts have no financial agreement between them before their interaction. They can only refer clients between themselves. Each client pays a tariff T to an expert who provides a service. The extensive form is as follows:

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

Stage 2: For each client that an expert has been matched with, the expert decides on an effort. Then the expert observes a realization of the signal for each client.

Stage 3: An expert chooses between keeping the client (for whom he has taken effort and on whom he has observed a signal) and referring the client to the other expert at a price that he chooses.

Stage 4: 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 the referral game, each expert has the same strategy space. First, Expert i chooses an unobservable effort, e_i, $i = 1, 2$, for each client, and then observes a private signal, s, about the client's cost state. Second, based on the effort and the realized signal, Expert i either keeps a client or offers the client to the other expert at a price, p_i. Third, when Expert j receives a referral at some price p_i, $j = 1, 2$, $j \neq i$, he either accepts the referral by paying price p_i or rejects it; in other words, an expert chooses the range of prices for accepting a referral. The tariff T follows a client, and is payable to the expert who actually renders the service.

An expert's payoff comes from one of three events. First, if an expert has kept his own client, his payoff is the tariff minus the client's cost and the effort disutility. Second, if an expert has accepted a referral, his payoff is the tariff minus what he has paid the referring expert and the client's cost. Third, if an expert's referral has been accepted, his payoff is his referral price minus the effort disutility. Each expert has a reservation utility that is set at 0.

The referral market is plagued by missing information about efforts and signals. Trivially, there is a no-trade equilibrium. Each expert exerts zero effort, never does any referral, and rejects referrals at any price above $(c_L + c_H)/2$. In fact, when $T < (c_L + c_H)/2$, each expert cannot obtain his reservation utility, and clients never receive services. At $T > (c_L + c_H)/2$, in the no-trade equilibrium, the total cost is $(c_L + c_H)/2$ when all clients obtain service. No cost saving ever occurs.

One might have thought that given both hidden action and hidden information, there cannot be any equilibria other than the no-trade equilibrium. Surprisingly, we present equilibria with some effort and some cost saving. In an equilibrium in which an expert makes no effort, however, a referral has no real consequence, since either expert has exactly the same *ex ante* expected cost of providing service, and we simplify our discussion by assuming that an expert makes no referral if in equilibrium he exerts no effort.

3.1 Experts' equilibrium referral and acceptance strategies

We now consider continuation equilibria starting at Stage 3. We begin by assuming that an expert, say, Expert 1, has taken an effort, say $e_1 > 0$, and has observed a signal s in Stage 2. Now Expert 2's strategy specifies the set of prices for which he will accept an offer. Clearly, in any equilibrium, this acceptance strategy takes the form of a threshold: Expert 2 accepts a referral if and only if the price offered by Expert 1 is less than p_1. (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 p_1.)

Suppose that Expert 2 accepts a referral at price p_1. How should Expert 1 choose between keeping the client and referring him to Expert 2? 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

$$\begin{aligned} & T - \Pr(\omega_1|s,e_1)c_L - \Pr(\omega_2|s,e_1)c_H \\ = \; & T - \frac{f_1(s|e_1)}{f_1(s|e_1)+f_2(s|e_1)}c_L - \frac{f_2(s|e_1)}{f_1(s|e_1)+f_2(s|e_1)}c_H. \end{aligned} \tag{6}$$

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

$$T - \frac{f_1(\widehat{s}|e_1)}{f_1(\widehat{s}|e_1)+f_2(\widehat{s}|e_1)}c_L - \frac{f_2(\widehat{s}|e_1)}{f_1(\widehat{s}|e_1)+f_2(\widehat{s}|e_1)}c_H = p_1. \tag{7}$$

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. Proofs of all other results are in the Appendix.

Lemma 1 *In an equilibrium in which an expert takes a strictly positive effort, in Stage 4 his referral is accepted if and only if the referral price is below a threshold, and in Stage 3, the expert makes a referral if and only if the signal exceeds a threshold.*

We complete the equilibrium description by specifying beliefs. Suppose that in equilibrium Expert 1 exerts effort $e_1 > 0$, and makes a referral at price p_1. If Expert 2 receives a referral at

price lower than p_1, he continues to believe that the effort is e_1, but if the referral is at a price higher than p_1, Expert 2 believes that Expert 1 has not exerted any effort, and optimally rejects that referral. In the next subsection, we will provide the characterization of the referral price.

Lemma 1 asserts that referral decisions and acceptance decisions must be threshold policies, but does not imply that an expert must exert a strictly positive effort. In the case of an expert making no effort in equilibrium, we assume that he does not make a referral.

3.2 Expert 1's equilibrium referral and effort

We proceed to characterize the signal threshold for making a referral, as well as the price threshold for accepting one. Suppose that in equilibrium Expert 1 takes a strictly positive effort e_1 and refers the client whenever $s > \widehat{s}$. When Expert 2 receives a referral offer, he must believe that the client's signal is at least $\widehat{s}$. Recall that (1) is the conditional probability of state i given a signal and an effort, and that the density of signal s given effort e_1 is $0.5[f_1(s|e_1) + f_2(s|e_1)]$. Using Bayes rule to update his beliefs given Expert 1's strategy (refer if and only if $s > \widehat{s}$), Expert 2's expected cost of providing service to the referred client is

$$\begin{aligned}
&\Pr(\omega_1|s > \widehat{s}, e_1)(c_L + \Delta) + \Pr(\omega_2|s > \widehat{s}, e_1)(c_H - \Delta) \\
&= \frac{(c_L + \Delta)\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta)\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x},
\end{aligned} \tag{8}$$

so he accepts a referral if and only if

$$T - \frac{(c_L + \Delta)\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta)\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x} \geq p_1. \tag{9}$$

Next consider Expert 1's payoff from keeping a client with signal s; this is expression (6). Given that Expert 2 accepts a referral at price p_1, Expert 1 refers a client with signal s if and only if

$$p_1 \geq T - \frac{f_1(s|e_1)}{f_1(s|e_1) + f_2(s|e_1)}c_L - \frac{f_2(s|e_1)}{f_1(s|e_1) + f_2(s|e_1)}c_H. \tag{10}$$

First we present some basic properties about experts' expected costs conditional on signals.

Lemma 2 *The equation*

$$\frac{(c_L+\Delta)\int_s^{\overline{s}} f_1(x|e_1)dx + (c_H-\Delta)\int_s^{\overline{s}} f_2(x|e_1)dx}{\int_s^{\overline{s}} f_1(x|e_1)dx + \int_s^{\overline{s}} f_2(x|e_1)dx} = \frac{c_L f_1(s|e_1) + c_H f_2(s|e_1)}{f_1(s|e_1)+f_2(s|e_1)} \tag{11}$$

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

Lemma 2 is based on the comparison of expected costs. Suppose that Expert 1 has chosen effort e_1. If he observes signal s, his expected cost of providing service to the client is

$$\frac{c_L f_1(s|e_1) + c_H f_2(s|e_1)}{f_1(s|e_1)+f_2(s|e_1)}. \tag{12}$$

Now if he refers a client whenever the signal is above s, then Expert 2's expected cost of providing service to the referred client is (8). Lemma 2 says that there must be a signal $\widehat{s}$ for which Expert 1's expected cost at signal $\widehat{s}$ is equal to Expert 2's expected cost conditional on Expert 1's signal higher than $\widehat{s}$.

This uniqueness result is based on Expert 2's comparative advantage in providing services to clients in state 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 in (12). The dotted line is Expert 1's expected cost given that signals are above s:

$$\frac{c_L\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x + c_H\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x}. \tag{13}$$

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 (details are in the proof). This expected cost conditional on signals above s of course converges to (12) at $s=\overline{s}$.

Now Expert 2's comparative advantage makes his expected cost (8), the dashed line in Figure 1, less than (13) at high values of s. 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 $(c_L+c_H)/2$ (again, details are in the proof).

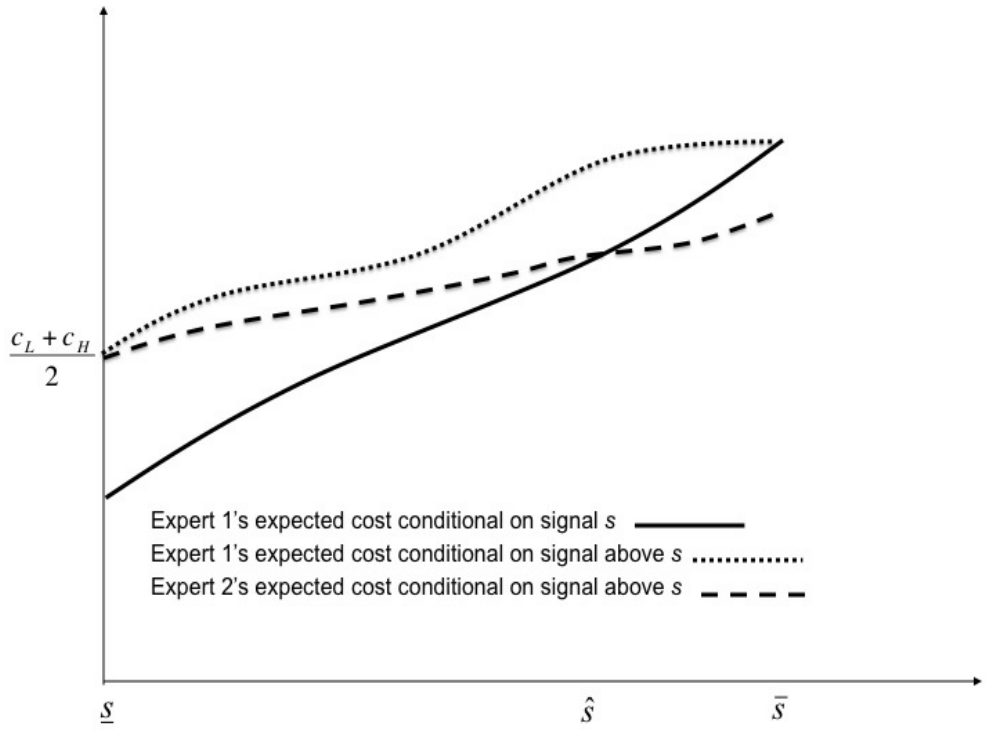

Figure 1: Expected costs and Expert 1's referral threshold $\widehat{s}$

The significance of Lemma 2 is this. Although Expert 2 can only infer that Expert 1's signal is above a certain threshold—the referral only revealing some information about Expert 1's signal—the experts nevertheless can mutually benefit from trade due to Expert 2's comparative cost advantage at state ω_2. Given effort e_1, as long as the signal is above $\widehat{s}$, the one in Lemma 2, a successful referral happens in equilibrium, as the next result shows.

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*

$$T - \frac{(c_L + \Delta)\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)dx + (c_H - \Delta)\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)dx}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)dx + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)dx} = p_1 = T - \frac{c_L f_1(\widehat{s}|e_1) + c_H f_2(\widehat{s}|e_1)}{f_1(\widehat{s}|e_1) + f_2(\widehat{s}|e_1)}. \quad (14)$$

We simply use the value of $\widehat{s}$ in Lemma 2 to set up the price p_1 at which Expert 1 can successfully refer clients with signals above $\widehat{s}$. In this equilibrium, Expert 2 is just indifferent between accepting

the referral and rejecting it. Expert 1's "marginal" client is one who has a signal $\widehat{s}$, but for each $s > \widehat{s}$, Expert 1 earns a surplus more than his payoff from rendering service himself.

We next study Expert 1's effort incentive. From Proposition 1, a continuation equilibrium is given by Expert 1's referral threshold $\widehat{s}$ and Expert 2's price acceptance threshold p_1. Hence, suppose that Expert 2 accepts a referral if and only if the referral price is at most p_1.

Recalling that the *ex ante* density of s is $0.5[f_1(s|e_1)+f_2(s|e_1)]$ and that each expert is matched initially with half of all clients, we write Expert 1's expected payoff per client from effort e_1 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

$$0.5\left[\int_{\underline{s}}^{\widehat{s}}\{[T-c_L]f_1(x|e_1)+[T-c_H]f_2(x|e_1)\}\,\mathrm{d}x + p_1\int_{\widehat{s}}^{\overline{s}}[f_1(x|e_1)+f_2(x|e_1)]\mathrm{d}x\right]-\phi(e_1), \quad (15)$$

which can be rewritten as

$$\left[T-\frac{c_L+c_H}{2}\right]-\phi(e_1)+0.5\int_{\widehat{s}}^{\overline{s}}\{[p_1-(T-c_L)]f_1(x|e_1)+[p_1-(T-c_H)]f_2(x|e_1)]\}\,\mathrm{d}x. \quad (16)$$

The first term in (16) 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 (16) subject to the definition of $\widehat{s}$ (7).[5] The first-order condition characterizes Expert 1's equilibrium effort:

$$0.5\int_{\widehat{s}}^{\overline{s}}\left\{[p_1-(T-c_L)]\frac{\partial f_1(x|e_1)}{\partial e_1}+[p_1-(T-c_H)]\frac{\partial f_2(x|e_1)}{\partial e_1}\right\}\mathrm{d}x=\phi'(e_1). \quad (17)$$

[5]In fact, the constraint (7) is redundant because the unconstrained maximization of (16) with respect to e_1 and $\widehat{s}$ yields that constraint anyway.

A higher effort raises the density f_2 more than the density f_1 at high signals s by the Informativeness Property. Expert 1 therefore more likely avoids the higher cost c_H due to equilibrium referral. Given any p_1 that is at least $T - (c_L + c_H)/2$, the first-order condition (17) admits a strictly positive e_1 as a solution.

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. Could the arguments in the above subsection for Expert 1's referrals help us construct a continuation equilibrium in which Expert 2 could refer successfully?

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

Proposition 2 says that Expert 2 lacks any effort incentive. This is in stark contrast to Lemma 2 and Proposition 1 which establish Expert 1's referral incentives. Lemma 1 says that an expert would like to keep only clients with low signals. An expert has no concern for cost comparative advantage. His only interest is to trade those clients who are likely in state ω_2 in exchange for a price. As long as an expert can successfully refer, he will refer those clients with high signals.

The comparative cost advantage of Expert 2 over Expert 1 at state ω_2 does allow mutually beneficial referrals from Expert 1 to Expert 2, as Lemma 2 and Proposition 1 demonstrate. The same cost advantage implies that Expert 1's expected costs will never be less than Expert 2's for those clients with signals above a threshold. There is no room for mutually beneficial trade because both experts prefer clients with lower expected costs. In the first best, Expert 2 should refer to Expert 1 those clients with signals below a threshold. But this is impossible in an equilibrium.

3.4 Referral Equilibrium

An equilibrium is characterized by the triple $[e_1^*, \widehat{s}^*, p_1^*]$ such that i) $(e_1^*, \widehat{s}^*)$ maximize (16) given p_1^*, and ii) p_1^* is given by (14) at $\widehat{s}$ and e_1^*. The equilibrium strategies are as follow.

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

2. Expert 2 chooses zero effort, does not refer, and accepts a referral if and only if the referral price is no higher than p_1^*.

3. If Expert 2 receives a referral at a price higher than p_1^*, he believes that Expert 1 has not taken any effort and refers a client irrespective of the signal, and hence rejects the offer.

4. If Expert 2 receives a referral at a price lower than p_1^*, he continues to believe that Expert 1 has followed the equilibrium strategy, and hence accepts the offer.

In the Appendix, we provide a fixed-point argument for the existence of equilibria. We proceed to characterize the equilibrium effort. Expert 2 does not exert any effort, which is of course 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 $[e_1^*, \widehat{s}^*, p_1^*]$:

$$T - \frac{(c_L + \Delta)\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x + (c_H - \Delta)\int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x}{\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x + \int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x} = p_1^* = T - \frac{c_L f_1(\widehat{s}^*|e_1^*) + c_H f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \quad (18)$$

$$0.5\int_{\widehat{s}^*}^{\overline{s}} \left\{ [p_1^* - (T - c_L)]\frac{\partial f_1(x|e_1^*)}{\partial e_1} + [p_1^* - (T - c_H)]\frac{\partial f_2(x|e_1^*)}{\partial e_1} \right\} \mathrm{d}x = \phi'(e_1^*). \quad (19)$$

Proposition 3 *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 retains a client even when his expected service cost is higher than Expert 2's.*

Proposition 3 asserts that Expert 1's equilibrium actions will not be first best. In fact, as shown in the proof, even if Expert 1 referred a client to Expert 2 according to the first-best cost-effectiveness rule (referring to Expert 2 if and only if Expert 2's expected cost is lower conditional on the observed signal, hence using a threshold $\widehat{s}^*$ defined by $f_2(\widehat{s}^*|e_1^*) = f_1(\widehat{s}^*|e_1^*)$), Expert 1 would invest in too much effort. The first best is driven by the cost-saving parameter Δ. Expert 1's profit,

however, depends on the client's types, and hence, the cost differential $c_H - c_L$. By assumption we have $c_H - c_L > 2\Delta$, so the profit motive would induce a stronger equilibrium effort incentive under the first-best cost-effectiveness rule.

Furthermore, conditional on equilibrium effort e_1^*, Expert 1 may keep a client even when Expert 2's service cost is lower: $f_2(\widehat{s}^*|e_1^*) > f_1(\widehat{s}^*|e_1^*)$. If the equilibrium referral threshold adhered to cost effectiveness, then $f_2(\widehat{s}^*|e_1^*) = f_1(\widehat{s}^*|e_1^*)$. At $\widehat{s}$, Expert 1's service cost would be

$$\frac{c_L f_1(\widehat{s}^*|e_1^*) + c_H f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \equiv \frac{c_L + c_H}{2}.$$

However, Expert 2 would anticipate that referral would be coming from any signal above $\widehat{s}$, so Expert 2's expected cost for providing service to a referred client was (see Proposition 1)

$$\frac{(c_L + \Delta)\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta)\int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}. \tag{20}$$

According to Proposition 1, this would have to be the same as $(c_L + c_H)/2$. However, according to MLRP, $\dfrac{\int_{\widehat{s}^*}^{\overline{s}} f_2(x|e_1^*)\mathrm{d}x}{\int_{\widehat{s}^*}^{\overline{s}} f_1(x|e_1^*)\mathrm{d}x} > \dfrac{f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*)}$, so (20) must be strictly bigger than

$$\frac{(c_L + \Delta) f_1(\widehat{s}^*|e_1^*) + (c_H - \Delta) f_2(\widehat{s}^*|e_1^*)}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} = \frac{c_L + c_H}{2},$$

which implies that the cost-effective referral threshold is infeasible. In fact, the proof demonstrates that $f_2(\widehat{s}^*|e_1^*) > f_1(\widehat{s}^*|e_1^*)$.

What about Expert 1's equilibrium effort? Using (18), we rewrite (19) as

$$0.5\left[\frac{c_H - c_L}{2}\right]\int_{\widehat{s}^*}^{\overline{s}} \left\{ \frac{2f_1(\widehat{s}^*|e_1^*)\dfrac{\partial f_2(x|e_1^*)}{\partial e_1} - 2f_2(\widehat{s}^*|e_1^*)\dfrac{\partial f_1(x|e_1^*)}{\partial e_1}}{f_1(\widehat{s}^*|e_1^*) + f_2(\widehat{s}^*|e_1^*)} \right\} \mathrm{d}x = \phi'(e_1^*). \tag{21}$$

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.

4 Organizations

Equilibria in the referral market are inefficient. We now present expert organizations for the implementation of the first best. We posit two possible differences between an expert organization and an open market. First, we let organizations make cost *ex post* verifiable within. Second, we let organizations decide on how clients initially are allocated to experts before any effort is taken. This implies a referral protocol. Otherwise, organizations still face problems of hidden action due to unobserved expert efforts, and hidden information due to experts' private observation of signals.

4.1 Integration

Under integration one expert buys out the other. To begin, we let Expert 1 buy out Expert 2. Compared to the market setting, there are two differences. First, in the integrated firm, Expert 1 receives all the earnings from providing services to clients, and bears all the costs. Integration transfers Expert 2's service revenues and costs to Expert 1. For the formation of the integrated firm, Expert 1 simply pays Expert 2 a flat fee in exchange for the transfer. Expert 1 does not offer any incentive contract to Expert 2; the flat fee is the only compensation Expert 2 receives. Therefore, Expert 2 does not exert any effort.

Second, Expert 1, now the owner, can determine how the clients are to be screened. The integrated organization adopts a gatekeeping protocol. Expert 1, the owner himself, receives all the clients, and for each client he exerts an effort, and observes a signal. Expert 1 then decides whether to provide service himself or to refer to Expert 2, who, as an employee, is obligated to provide service. Because Expert 2's service cost has been transferred to the integrated organization, his sole role is to respond to Expert 1's referrals by providing services to clients.

Only Expert 1 makes decisions in the integrated firm. We now explain how Expert 1 chooses effort and how he decides which clients to keep and which to refer. The key is that Expert 1 now completely internalizes the benefits and costs of effort, as well as the comparative cost savings from referrals.

Given an effort, say e_1, Expert 1 decides whether he himself provides service or lets Expert 2 do so based on the signal s he observes. As we have explained in Subsection 2.4, Expert 2's expected service cost is lower if and only if $f_1(s|e_1) \leq f_2(s|e_1)$ (see (3) and (2)). Hence, by MLRP, Expert 1 assigns a client to Expert 2 if and only if the client's signal s is larger than $\widehat{s}(e_1)$ where $f_1(\widehat{s}|e_1) = f_2(\widehat{s}|e_1)$.

Expert 1 then chooses an effort to maximize expected profits:

$$T - \phi(e_1) - 0.5\int_{\underline{s}}^{\widehat{s}(e_1)} \{\Pr(\omega_1|x,e_1)c_L + \Pr(\omega_2|x,e_1)c_H\}[f_1(x|e_1) + f_2(x|e_1)]\mathrm{d}x + \tag{22}$$
$$0.5\int_{\widehat{s}(e_1)}^{\overline{s}} \{\Pr(\omega_1|x,e_1)(c_L + \Delta) + \Pr(\omega_2|x,e_1)(c_H - \Delta)\}[f_1(x|e_1) + f_2(x|e_1)]\mathrm{d}x.$$

Clearly, Expert 1's effort choice to maximize (22) is the same as an effort choice to minimize total cost (4) for the first best in Subsection 2.4. Expert 1 therefore chooses an effort satisfying (5).

We have started with Expert 2 buying out Expert 1. In fact, Expert 2 buying out Expert 1 achieves the same. Integration now transfers Expert 1's service revenues and costs to Expert 2, who pays Expert 1 a flat fee. Expert 2 now adopts a gatekeeping protocol dictating that he himself must screen all clients. Then Expert 2 decides which clients to serve and which clients to refer to Expert 1 for service, based on the signal Expert 2 observes.

The key is that the owner-expert internalizes the cost saving under integration. Now we have assumed that the effort costs are identical. In practice, one expert may have a comparative advantage in information acquisition, so it will be more efficient if the expert who has the effort-cost advantage to integrate, so he will be responsible for information acquisition.

Because the integrated organization achieves the first best, it must generate a higher surplus than the two experts' total profit in the referral market equilibrium. Hence, there is a fee for an

expert to agree to an integration proposal.

4.2 Partnership

A partnership is an agreement between the two experts on how to share any revenues and costs among themselves.[6] We retain the assumption that each expert privately chooses effort, and privately observes signals about clients. As with the integrated firm, the basic distinction between a market environment and a partnership is that experts' service costs can be verified *ex post*. Unlike integration, the partnership cannot dictate whether an expert refers a client to the other. We let each expert be matched randomly with a half of all clients.

The key feature of partnership is that it can keep track of referrals. The partnership can distinguish whether services are provided by the originally matched expert or by the referred expert. Otherwise, we follow the standard requirement that a partnership determines a sharing rule that must always be budget balanced. That is, the sum of the two experts' monetary payoffs must be always equal to the net monetary resources resulting from services. Since cost information becomes verifiable within the partnership, a sharing rule can be based on realized costs. We introduce new notation to describe the partnership sharing rule.

First, recall that Expert 1's service is either low, at c_L, or high, at $c_{H;}$; Expert 2's service cost is either low, at $c_L + \Delta$, or high at $c_H - \Delta$. Suppose that Expert i, $i = 1, 2$, provides services to his own clients, each of whom having either low or high costs, *ex post*. Let the masses of these clients with realized low and high costs be q_{iL} and q_{iH} , respectively.[7] Next, suppose that Expert i provides services to referred clients, each of whom having either low or high costs, *ex post*. Let the masses of these referred clients with realized low and high costs be q_{iL}^R and q_{iH}^R , respectively. For example, among those clients serviced by Expert 1, q_{1H} is the mass of Expert 1's original clients whose costs have turned out to be high c_H, while q_{2L}^R is the mass of clients referred by Expert 1

[6] This is the usual definition; see Legros and Matthews (1993), and Levin and Tadelis (2002).

[7] Given that a client's state is either ω_1 or ω_2, the masses of clients with low and high costs *ex post* are related. Nevertheless, the distinction between q_{iL} and q_{iH} is made for a conceptual reason: the partnership sharing rule must be budget balanced no matter how costs turn out.

to Expert 2, and whose service costs at Expert 2 have turned out to be low $c_L + \Delta$.

The *ex post* costs from Expert 1's services are $q_{1L}c_L + q_{1H}c_H + q^R_{1L}c_L + q^R_{1H}c_H$. Likewise, the *ex post* costs from Expert 2's services are $q_{2L}(c_L + \Delta) + q_{2H}(c_H - \Delta) + q^R_{2L}(c_L + \Delta) + q^R_{2H}(c_H - \Delta)$. In each case, an expert's service cost can be separated into two parts: those services he provides to his original clients, and those he provides to clients referred to him. The partnership's total cost is just the sum of costs incurred by the experts. The partnership's total revenue is T times the total mass of clients.

Let S_1 and S_2 be the sharing rule, a function of the cost profile $(q_{1L}, q_{1H}, q_{2L}, q_{2H}, q^R_{1L}, q^R_{1H}, q^R_{2L}, q^R_{2H})$, such that $S_1 + S_2$ equals the total net revenue for any cost profile. The accounting rule is that each expert's revenue and realized cost go into the partnership pool. The sharing rule then distributes the proceeds to the two experts, as a function of the cost profile.

We will show that the following sharing rule implements the first best:

$$\begin{aligned} S_1 &= 0.5T - [q_{1L}c_L + q_{1H}c_H] - [q^R_{2L}(c_L + \Delta) + q^R_{2H}(c_H - \Delta)] \\ S_2 &= 0.5T - [q_{2L}(c_L + \Delta) + q_{2H}(c_H - \Delta)] - [q^R_{1L}c_L + q^R_{1H}c_H]. \end{aligned} \tag{23}$$

Here, each expert gets a half of the total revenue, $0.5T$. For costs, Expert 1 bears the costs of services he provides to his own clients, as well as the costs of services of clients that he refers to Expert 2. The sharing rule is symmetric with respect to the experts. Obviously, under this sharing rule the budget is balanced; the sum of S_1 and S_2 in (23) equals the net revenue, for any cost profile.

The partnership game is a natural adaptation of the extensive form in Subsection 3.1, so we do not have to write it down. There are just a few differences. First, there is no referral price, although an expert is free to accept or reject a referral. Second, in Stage 1, the partnership decides on a sharing rule. Finally, at the end, after services have been rendered, the sharing rule will be implemented, and the accounts settled.

We consider the sharing rule in (23), and derive the perfect-Bayesian equilibria of the partnership game. First, notice that under (23), an expert is not responsible for the cost of any client that is

referred to him. Hence, an expert's belief about a referred client's type is irrelevant. We specify that in an equilibrium, an expert accepts any referral.

Second, under the sharing rule (23), for any client that is matched with an expert, say Expert 1, that expert is responsible for the service cost. If Expert 1 provides service himself, then either q_{1L} or q_{1H} will increase by one unit, depending on whether the client's state turns out to be ω_1 or ω_2. If Expert 1 refers the client, then either q^R_{2L} or q^R_{2H} will increase correspondingly.

Suppose now Expert 1 has chosen an effort, say e_1, and observed a signal, s. His posterior belief that the client's state is ω_i given by (1), and for convenience it is rewritten here:

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

If Expert 1 provides service, then his expected cost is $c_L \Pr(\omega_1|s,e_1) + c_H \Pr(\omega_2|s,e_1)$. If Expert 1 refers this client to Expert 2, the sharing rule (23) specifies that he will still be responsible for Expert 2's cost, which is $(c_L+\Delta)\Pr(\omega_1|s,e_1) + (c_H-\Delta)\Pr(\omega_2|s,e_1)$. Clearly, Expert 1 refers a client if and only if $\Pr(\omega_2|s,e_1) \geq 0.5$, which is the first-best rule.

Third, an expert has to decide on an effort, given that he will use the first-best referral rule. However, given that the expert fully internalizes a client's expected cost, his objective is the same as the choice of effort to minimize the service cost and effort disutility (4). An expert's equilibrium effort must be first best.

The first best under a partnership may also be implemented by a *mutual agreement*. The experts can construct a bilateral contract between themselves. Again, *ex post* cost information can be used. The mutual agreement stipulates that an expert can refer any number of clients to the other expert but must bear the service costs incurred by the recipient expert. The bilateral agreement mimics the equilibrium under partnership. Each expert fully internalizes the service cost of a client who is initially matched with him, so has an incentive to invest in effort to find out if he himself or the other expert is more cost effective.

5 Robustness

We now discuss a number of robustness issues. 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 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 comparative cost 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 (and hence 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 integration and partnership. Under integration, the owner-expert must carry out information acquisition to implement the first best because he is gatekeeper In partnership, each expert can exert effort,and both are gatekeepers. Any capacity constraint and decreasing returns must favor partnership over integration. On the other hand, increasing returns favor integration.

In the rest of this section, we discuss two other robustness issues in detail. 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, a violation of our assumption $\Delta < (c_L + c_H)/2$.

5.1 Equilibrium tariff

We have assumed that an expert receives a fixed tariff T for service rendered to each client. We can straightforwardly endogenize tariffs. Let the market consist of Expert 1 and Expert 2. We append a Bertrand-competition stage before the referral game in Section 3. That is, we add Stage 0 in which each expert announces a tariff. Consumers observe these tariffs, and choose an expert for service. A consumer pays the required tariff when he chooses an expert.[8]

Recall that each expert can serve a client at an expected cost $(c_L + c_H)/2$. If an expert neither 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 (14) 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 (14).

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. Here, the "more efficient" Expert 1 sets the same tariff as

[8] 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 "less efficient" Expert 2, but takes all the surplus from trade.[9]

Finally, we consider how experts compete with organizations. Let the market consist of Expert 1, Expert 2, and an expert organization such as one in Section 4. Under either integration or partnership, the organization achieves the first best, so its service cost and effort disutility per client is

$$\int_{\underline{s}}^{\widehat{s}^{fb}} \{\Pr(\omega_1|x,e^{fb})c_L + \Pr(\omega_2|x,e^{fb})c_H\}[f_1(x|e^{fb}) + f_2(x|e^{fb})]\mathrm{d}x + \tag{24}$$
$$\int_{\widehat{s}^{fb}}^{\overline{s}} \{\Pr(\omega_1|x,e^{fb})(c_L+\Delta) + \Pr(\omega_2|x,e^{fb})(c_H-\Delta)\}[f_1(x|e^{fb}) + f_2(x|e^{fb})]\mathrm{d}x + \phi(e^{fb}),$$

where $\widehat{s}^{fb}$ satisfies $f_1(\widehat{s}^{fb}|e^{fb}) = f_2(\widehat{s}^{fb}|e^{fb})$. Because this is first best, the expected service costs and disutility in (24) are smaller than any other expert can achieve in the referral equilibrium in Subsection 3.4. By the same Bertrand-competition argument, the organization will win all clients at a tariff lower than what each individual expert can offer**.** If there are many expert organizations, then Bertrand competition will force the equilibrium tariff to the level in (24).

5.2 Experts with large comparative cost 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 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 is unchanged, and Proposition 3 in Subsection 3.2 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 (8) is decreasing in $\widehat{s}$; in Figure 1, the dashed line is downward sloping.

[9] Expert 1 is more efficient because he invests in information acquisition.

For Expert 2, suppose now that he has taken effort e_2. Further, assume that if Expert 2 refers a client at a price no higher than p_2, Expert 1 will accept it. Let Expert 2 observe a signal s**.** If he keeps the client, he receives the tariff T, while if he refers at price p_2, he will get that price without having to provide the service. Hence, Expert 2 refers if and only if

$$p_2 \geq T - \frac{(c_L + \Delta)\, f_1(s|e_2)}{f_1(s|e_2) + f_2(s|e_2)} - \frac{(c_H - \Delta)\, f_2(s|e_2)}{f_1(s|e_2) + f_2(s|e_2)}. \tag{25}$$

By MLRP, and $c_L + \Delta > c_H - \Delta$, the expected cost in (25) is decreasing in s, so the right-hand side of (25) is increasing in s. For given e_2 and p_2, define $\widehat{s}_2$ such that at $s = \widehat{s}_2$, (25) holds as an equality:

$$p_2 = T - \frac{(c_L + \Delta)\, f_1(\widehat{s}_2|e_2)}{f_1(\widehat{s}_2|e_2) + f_2(s|e_2)} - \frac{(c_H - \Delta)\, f_2(\widehat{s}_2|e_2)}{f_1(\widehat{s}_2|e_2) + f_2(\widehat{s}_2|e_2)}.$$

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 a piece of favorable news to Expert 1.

Given Expert 2's threshold $\widehat{s}_2$, Expert 1's expected cost if he accepts the referral is

$$\Pr(\omega_1|s < \widehat{s}_2, e_2)c_L + \Pr(\omega_2|s < \widehat{s}_2, e_2)c_H \equiv \frac{c_L \int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + c_H \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x}{\int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x},$$

so he accepts Expert 2's referral if and only if

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

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

$$T - \frac{(c_L + \Delta)\, f_1(\widehat{s}_2|e_2) + (c_H - \Delta)\, f_2(\widehat{s}_2|e_2)}{f_1(\widehat{s}_2|e_2) + f_2(\widehat{s}_2|e_2)} = p_2 = T - \frac{c_L \int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + c_H \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x}{\int_{\underline{s}}^{\widehat{s}_2} f_1(x|e_2)\mathrm{d}x + \int_{\underline{s}}^{\widehat{s}_2} f_2(x|e_2)\mathrm{d}x}. \tag{26}$$

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 (26) must exist. Indeed, the solution $\widehat{s}_2$ of the following equation

$$\frac{(c_L+\Delta)\,f_1(\widehat{s}_2|e_2)+(c_H-\Delta)\,f_2(\widehat{s}_2|e_2)}{f_1(\widehat{s}_2|e_2)+f_2(\widehat{s}_2|e_2)}=\frac{c_L\int_{\underline{s}}^{\widehat{s}_2}f_1(x|e_2)\mathrm{d}x+c_H\int_{\underline{s}}^{\widehat{s}_2}f_2(x|e_2)\mathrm{d}x}{\int_{\underline{s}}^{\widehat{s}_2}f_1(x|e_2)\mathrm{d}x+\int_{\underline{s}}^{\widehat{s}_2}f_2(x|e_2)\mathrm{d}x} \tag{27}$$

must exist. Indeed, by MLRP, $c_L < c_H$, and $c_L+\Delta > c_H-\Delta$, the left-hand side of (27) is decreasing in $\widehat{s}_2$, while the right-hand side is increasing.[10]

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

$$\begin{aligned}&p_2\int_{\underline{s}}^{\widehat{s}_2}0.5[f_1(x|e_2)+f_2(x|e_2)]\mathrm{d}x\\&\int_{\widehat{s}_2}^{\overline{s}}0.5[(T-c_L-\Delta)\Pr(\omega_1|x,e_2)+(T-c_H+\Delta)\Pr(\omega_2|x,e_2)][f_1(x|e_2)+f_2(x|e_2)]\mathrm{d}x-\phi(e_2).\end{aligned}$$

which simplifies to

$$\left[T-\frac{c_L+c_H}{2}\right]+0.5\int_{\underline{s}}^{\widehat{s}_2}\left\{[p_2-(T-c_L-\Delta)]f_1(x|e_2)+[p_2-(T-c_H+\Delta)]f_2(x|e_1)]\right\}\mathrm{d}x-\phi(e_2). \tag{28}$$

This has the same interpretation of Expert 1's expected payoff in (16). Expert 2's optimal effort is one that maximizes (28), and its first-order condition is

$$0.5\int_{\underline{s}}^{\widehat{s}_2}\left\{[p_2-(T-c_L-\Delta)]\frac{\partial f_1(x|e_2)}{\partial e_2}+[p_2-(T-c_H+\Delta)]\frac{\partial f_2(x|e_2)}{\partial e_2}\right\}\mathrm{d}x=\phi'(e_2). \tag{29}$$

Expert 2's equilibrium strategy is therefore characterized by price p_2, threshold $\widehat{s}_2$, and effort e_2 satisfying (26) and (29).

We use (26) to eliminate p_2 in (29), and obtain

$$0.5\left[\frac{c_H-c_L}{2}\right]\int_{\underline{s}}^{\widehat{s}_2}\left\{\frac{2f_1(\widehat{s}_2|e_2)\dfrac{\partial f_2(x|e_2)}{\partial e_2}-2f_2(\widehat{s}_2|e_2)\dfrac{\partial f_1(x|e_2)}{\partial e_2}}{f_1(\widehat{s}_2|e_2)+f_2(\widehat{s}_2|e_2)}\right\}\mathrm{d}x=\phi'(e_2),$$

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

which has the same form as the characterization of Expert 1's effort in (21). We conclude 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. The proof of these two results follows, in a symmetric fashion, the proof of Proposition 3, and is omitted.

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. We do not expect that an organization can directly eliminate hidden action and hidden information. However, an organization can keep track of members' activities, and institute accounting information. We show that using *ex post* cost information, and establishing referral protocols work well in two organizational forms. Both integration and partnership can solve hidden action and hidden information problems.

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 market game.

Appendix

Proof of Lemma 2: By MLRP, $\dfrac{f_2(x|e_1)}{f_1(x|e_1)}$ is increasing in x, so for any s we have

$$\begin{aligned}\frac{\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x} &\equiv \frac{\int_s^{\overline{s}} \frac{f_2(x|e_1)}{f_1(x|e_1)} \cdot f_1(x|e_1)\mathrm{d}x}{\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x} \\ &> \frac{\int_s^{\overline{s}} \frac{f_2(s|e_1)}{f_1(s|e_1)} f_1(x|e_1)\mathrm{d}x}{\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x} = \frac{f_2(s|e_1)}{f_1(s|e_1)}. \end{aligned} \tag{30}$$

It follows that

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

and

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

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

$$\frac{c_L\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x + c_H\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x} > \frac{c_L f_1(s|e_1) + c_H f_2(s|e_1)}{f_1(s|e_1) + f_2(s|e_1)}. \tag{31}$$

Applying L'Hospital's rule, we have

$$\begin{aligned}\lim_{s\to\overline{s}} \frac{(c_L+\Delta)\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H-\Delta)\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x} &= \frac{(c_L+\Delta)f_1(\overline{s}|e_1) + (c_H-\Delta)f_2(\overline{s}|e_1)}{f_1(\overline{s}|e_1) + f_2(\overline{s}|e_1)} \\ = \frac{c_L f_1(\overline{s}|e_1) + c_H f_2(\overline{s}|e_1) - \Delta[f_2(\overline{s}|e_1) - f_1(\overline{s}|e_1)]}{f_1(\overline{s}|e_1) + f_2(\overline{s}|e_1)} &< \frac{c_L f_1(\overline{s}|e_1) + c_H f_2(\overline{s}|e_1)}{f_1(\overline{s}|e_1) + f_2(\overline{s}|e_1)}.\end{aligned}$$

We have shown that at s sufficiently near $\overline{s}$ the left-hand side of (11) is smaller than the right-hand side.

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

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

We have shown that at s sufficiently near $\underline{s}$, the left-hand side of (11) is larger than the right-hand side. Therefore, equation (11) must have a solution $\widehat{s}$.

Finally, for uniqueness, rewrite (11) as

$$\frac{(c_L+\Delta)+(c_H-\Delta)\dfrac{\left[\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x\right]/f_2(s|e_1)}{\left[\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x\right]/f_1 s|e_1)} \bullet \dfrac{f_2(s|e_1)}{f_1(s|e_1)}}{1+\dfrac{\left[\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x\right]/f_2(s|e_1)}{\left[\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x\right]/f_1 s|e_1)} \bullet \dfrac{f_2(s|e_1)}{f_1(s|e_1)}} = \frac{c_L + c_H\dfrac{f_2(s|e_1)}{f_1(s|e_1)}}{1+\dfrac{f_2(s|e_1)}{f_1(s|e_1)}}. \tag{32}$$

By MLRP, the inverse hazard rates satisfy $\left[\int_s^{\overline{s}} f_2(x|e_1)\mathrm{d}x\right]/f_2(s|e_1) > \left[\int_s^{\overline{s}} f_1(x|e_1)\mathrm{d}x\right]/f_1(s|e_1)$; see also (30) above. As s changes, the rates of change of the left-hand and right-hand sides of (32) will never be identical. As separate functions, the graphs of the left-hand and right-hand sides of (11) can only cross each other once. In other words, there can only be one solution.

Proof of Proposition 1: The two equations in (14) include equation (11). Lemma 2 already establishes a solution for $\widehat{s}$. We then set the value of p_1 according to (14). Clearly the conditions for an equilibrium

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

are satisfied.

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 $(c_L + \Delta)\Pr(\omega_1|\widetilde{s}, e_2) + (c_H - \Delta)\Pr(\omega_2|\widetilde{s}, e_2)$

$$
\begin{aligned}
&= \frac{(c_L + \Delta)f_1(\widetilde{s}|e_2) + (c_H - \Delta)f_2(\widetilde{s}|e_2)}{f_1(\widetilde{s}|e_2) + f_2(\widetilde{s}|e_2)} \\
&< \frac{(c_L + \Delta)\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x + (c_H - \Delta)\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x + \int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x} \qquad (33) \\
&< \frac{c_L\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x + c_H\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x + \int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x} \qquad (34)
\end{aligned}
$$

where the inequality in (33) follows from MLRP (see also (31) in the proof of Lemma 2). Now the derivative of (33) 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, (33) is decreasing in Δ. By reducing the value of Δ to zero, we obtain Expert 1's expected cost of providing service to a client conditional on Expert 2's signal being at least $\widetilde{s}$ in (34).

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

$$
T - \frac{(c_L + \Delta)f_1(\widetilde{s}|e_2) + (c_H - \Delta)f_2(\widetilde{s}|e_2)}{f_1(\widetilde{s}|e_2) + f_2(\widetilde{s}|e_2)} \leq p_2 \leq T - \frac{c_L\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x + c_H\int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x}{\int_{\widetilde{s}}^{\overline{s}} f_1(x|e_2)\mathrm{d}x + \int_{\widetilde{s}}^{\overline{s}} f_2(x|e_2)\mathrm{d}x} \qquad (35)
$$

a condition for an equilibrium. This is a contradiction.

Proof of Proposition 3: 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 2.4. From the second equation in (18) 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 (19) 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 (19) with (5), we conclude that $e_1^* > e^{fb.}$, so Expert 1's effort is not first best.

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

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

Now by MLRP, we have (30):

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

It follows that

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

which contradicts (18). We conclude that $f_2(\widehat{s}^*|e_1^*) > f_1(\widehat{s}^*|e_1^*)$.

Existence of Market Equilibrium

The existence of the market equilibrium follows a standard fixed-point argument. Consider bounding the feasible Expert 1's efforts by a compact convex set, say a closed interval of 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

$$(e_1', \widehat{s}') = \underset{e_1, \widehat{s}}{\operatorname{argmax}}\, 0.5 \int_{\widehat{s}}^{\overline{s}} \{[p_1 - (T - c_L)] f_1(x|e_1) + [p_1 - (T - c_H)] f_2(x|e_1)]\}\, \mathrm{d}x - \phi(e_1) \quad (36)$$

$$p_1' = T - \frac{(c_L + \Delta) \int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + (c_H - \Delta) \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}{\int_{\widehat{s}}^{\overline{s}} f_1(x|e_1)\mathrm{d}x + \int_{\widehat{s}}^{\overline{s}} f_2(x|e_1)\mathrm{d}x}. \quad (37)$$

Here, (36) is Expert 1's best response against Expert 2's referral acceptance price p_1 (the same as the maximization of (16) with respect to effort and referral threshold), while (37) is Expert 2's referral acceptance best response against Expert 1's effort e_1 and referral threshold $\widehat{s}$ (see also (14) in Proposition 1).

Clearly, the Maximum Theorem applies to (36), and there is a selection of the solution $(e_1', \widehat{s}')$ which is continuous in p_1. Furthermore, p_1' in (37) is obviously continuous in e_1 and $\widehat{s}$. By Brouwer's Fixed Point Theorem, Ψ has a fixed point $(e_1^*, \widehat{s}^*, p_1^*)$, which is the vector of mutual best responses.

References

Bergemann, D., & Välimäki, J. (2002). Information acquisition and efficient mechanism design. *Econometrica*, 70(3), 1007-1033.

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.

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.

Crémer, J., & Khalil, F. (1992). Gathering information before signing a contract. *American Economic Review*, 82(3), 566–578.

Crémer, J., Khalil, F., & Rochet, J-C. (1998a). Contracts and productive information gathering. *Games and Economic Behavior*, 25(2), 174–193.

Crémer, J., Khalil, F., & Rochet, J-C. (1998b). Strategic information gathering before a contract is offered. *Journal of Economic Theory,* 81(1), 163–200.

Currie J., & MacLeod W. B. (2013). Diagnosis and unnecessary procedure use: evidence from C-section. NBER Working Paper No. 18977.

Dai, C., Lewis, T. R., & Lopomo, G. (2006). Delegating management to experts. *The Rand Journal of Economics*, 37(3), 503-520.

Demski, J. S., & Sappington, D. E. (1987). Delegated expertise. *Journal of Accounting Research*, 25(1), 68-89.

Dulleck, U., & Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1), 5-42.

Dulleck, U., & Kerschbamer, R. (2009). Experts vs discounters: Consumer free-riding and experts withholding advice in markets for credence goods. *International Journal of Industrial Organization*, 27(1), 15-23.

Emons, W. (2001). Credence goods monopolists. *International Journal of Industrial Organization*, 19(3), 375-389.

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.

Fong, Y. F. (2005). When do experts cheat and whom do they target?. *The RAND Journal of Economics*, 36(1), 113-130.

Fuchs, W., & Garicano, L. (2010). Matching problems with expertise in firms and markets. *Journal of the European Economic Association*, 8(2-3), 354-364.

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

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5), 874-904.

Holmström, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 13(2), 324-340.

Iossa, E., & Martimort, D. (2013). Hidden action or hidden information? How information gathering shapes contract design. CEPR Discussion Papers No. 9552.

Legros, P., & Matthews, S. A. (1993). Efficient and nearly-efficient partnerships. *The Review of Economic Studies*, 60(3), 599-611.

Levin, J., & Tadelis, S. (October 2002). A Theory of Partnerships. Stanford Law and Economics Olin Working Paper 244.

Liu, T. (2011). Credence goods markets with conscientious and selfish experts, *International Economic Review*, 52(1), 227-244.

Liu, T., Ma, CA., & Mak, H. (2014). Incentives for motivated agents in a partnership, Boston University Working Paper.

Pitchik, C., & Schotter, A. (1987). Honesty in a model of strategic information transmission. *American Economic Review*, 77(5), 1032-1036.

Rebitzer, J. B., & Votruba, M.E. (2011). Organizational economics and physician practices. NBER Working Paper No. 17535.

Sülzle, K., & Wambach, A. (2005). Insurance in a market for credence goods. *Journal of Risk and Insurance*, 72(1), 159-176.

Szalay, D. (2009). Contracts with endogenous information. *Games and Economic Behavior*, 65(2), 586-625.