

REFERRALS*

Luis Garicano

Tano Santos

Graduate School of Business

Graduate School of Business

University of Chicago and CEPR

Columbia University and NBER

November 6, 2003

Abstract

This paper studies the matching of opportunities with talent when costly diagnosis confers an informational advantage to the agent undertaking it. When this agent is under-qualified, adverse selection prevents efficient referrals through fixed-price contracts. Spot-market contracts that rely on income sharing can match opportunities with talent but induce a team-production problem which, if severe enough, can prevent the referral of valuable opportunities. Partnership contracts, in which agents agree in advance to the allocation of opportunities and of the revenues they generate, support referrals where the market cannot, but often at the expense of distortions on those opportunities that are not referred.

“The peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in a concentrated or integrated form.” Friedrich Hayek (1945:519)

A crucial task of economic organization is to match opportunities with talent. Such matching requires that an agent first diagnose the opportunity confronted. The agent may conclude from the diagnosis that some other agent is the right person for the job, in which case he must pass it on (or *refer* it), potentially losing the rents he could obtain if he dealt with the opportunity himself. For example, a personal injury lawyer may, upon study and diagnosis, determine that the case he is working on requires more trial experience than he has.

Which institutions result in efficient referrals when the agent in charge of the diagnosis may or may not be the right person for the job? The study of this problem, which we consider the fundamental ‘referral’ problem, is the object of this paper.

An obvious way to allocate opportunities would involve trading them for a fixed price. However, this solution requires that the quality of the opportunity be observable and verifiable. This is unlikely to be the case in these markets. For example, in the injury claims market, a lawyer may determine he should refer his client to another lawyer. This transfer must take place under partial ignorance, as the lawyer receiving the claim cannot be allowed to personally evaluate the truthfulness of the client, the merits of the case, etc., since this would create the risk that he concludes an agreement that would result in the complete loss of the case to the referring lawyer.

Clearly, all agents prefer, under these conditions, to keep the most valuable opportunities and to refer the least valuable ones. As we show, this implies that whether informational asymmetries are an obstacle to the efficient allocation depends on the direction of the referral flow. A fixed-price contract exists such that the agent prefers to transfer the worst opportunities and keep the best ones. As long as the diagnosing agent is more skilled at handling the opportunity than the agent receiving it, this is the efficient allocation. In other words, ‘top-down’ diagnosis generates no inefficiency. The opposite happens when diagnosis is ‘bottom-up.’ Here, efficiency requires that a less-skilled agent keep the worst opportunities, and pass on the better ones; but any fixed price that is high enough to motivate the transfer of the best opportunities also leads to the transfer of the least valuable ones.

In this case, agents may instead choose to sell a share of the problem’s value.¹ While

this solution deals with the adverse-selection problem, it generates a team-production problem. Since the effort that an agent puts into dealing with a particular client is unobservable, the agent who is referred an opportunity and obtains only a share of the output has an incentive to free ride by providing too little effort. As a result, the spot market fails to support the right allocation precisely when the opportunities are most valuable.

We investigate these issues through a model in which agents with heterogeneous skills can draw, diagnose, and tackle opportunities. An agent may choose either to refer the opportunity to another agent or to handle it himself. Information about the quality of the opportunity is asymmetric, as only the expert agent who diagnosed it knows its true value. Moreover, effort is unobservable. Team production is endogenous, since the decision to involve another agent in the production process lies with the agent who knows about the economic opportunity. The assignment problem exists because talent, effort and value of the opportunities are complementary, but comparative advantage holds, so that it is efficient to assign the more valuable opportunities to the more skilled agents. Efficient matching then requires unskilled agents to refer valuable opportunities, and skilled agents to refer the less valuable ones. The relevant trade-off in this context, rather than the usual one between risk and incentives, is one between effort incentives and referral incentives. Allocating the opportunity requires compensating the agent who knows about its existence in a way that gives him incentives to retain the opportunity and exert effort on it when he is actually best qualified to deal with it, while giving him incentives to refer the rest. This incentive must be traded off against the risk of moral hazard on the part of the agent receiving the referral.

After characterizing the problem, and the trade-offs involved in the market for referrals, we study how agents can improve on the spot-market allocation by establishing ex-ante referral agreements. Specifically we study ‘partnerships,’ which we define as contracts that commit agents to an allocation of opportunities and of the income from these opportunities.² Partnerships weaken agents’ incentives to hold onto opportunities for which they are not best qualified, at the cost of lowering effort incentives across all opportunities. In addition, the partnerships’ ability to limit ex-post competition for opportunities greatly enhances the scope of incentives they can provide.

Entering these contracts requires making income from clients observable and verifiable, even when the same agent who draws the opportunity deals with it, and no joint production takes place. In turn, making the revenue flows observable and verifiable requires that agents jointly bill their clients. It is this joint-billing feature that is likely to cause ex-ante referral

agreements to take the legal form of partnerships.³

No previous literature has, to our knowledge, studied referrals under asymmetric information. Luis Garicano (2000) discusses a hierarchical referral process when agents' incentives are aligned, so that asymmetric information is not a problem. Joel Demski and David E. Sappington (1987), Asher Wolinsky (1993) and Curtis Taylor (1995) deal with the role of expert advisors under asymmetric information, but all these papers study the relation between the client and the expert, rather than the incentives of the expert to allocate the problem to someone more qualified.⁴ We assume instead that clients are completely uninformed and arrive randomly at the doorstep of experts, and we focus on the relationship between the different experts. A second branch of the literature, including Joseph Farrell and Suzanne Scotchmer (1988), Patrick Legros and Steven A. Matthews (1993), Eugene Kandel and Edward Lazear (1992) and Jonathan Levin and Steven Tadelis (2002), studies partnerships as revenue sharing agreements, abstracting from the allocation problem.⁵ Finally, Ronald J. Gilson and Robert H. Mnookin (1985) propose a theory of partnerships as risk sharing arrangements. As we argue in Section V we do not find the evidence on the scope of partnerships consistent with this view.

The paper is structured as follows. Section I presents the model. Section II studies the first-best allocation, which serves as a benchmark throughout. Section III discusses referrals in spot markets. Section IV considers ex-ante agreements and shows how partnerships arise to ease the referral process. Section V discusses the empirical implications of the model. Section VI concludes.

I. A MODEL OF VERTICAL REFERRALS

We consider vertical referrals, in which two agents can deal with an opportunity, but one (she) is more skilled at it than the other (he), in the sense that she adds more value to the opportunity.⁶ Our first assumption is that an expert is required to diagnose the opportunity.

Assumption 1. Opportunities cannot self sort. The allocation requires costly diagnosis which can only be provided by an expert.

In particular, we consider an economy in which risk-neutral agents draw clients, tasks or, in general, opportunities of uncertain value $v \in \{v_0, v_1\}$, with $0 < v_0 < v_1$. The low-value opportunity (v_0) is drawn with probability $\pi \in (0, 1)$, and the high-value opportunity (v_1) with probability $1 - \pi$. In order to determine the quality of the opportunity, agents must first undertake a costly diagnosis; in other words, opportunities are indistinguishable without

diagnosis.⁷ The agent who drew the opportunity then faces a choice between one of two options. First, he may choose to deal with it himself, and apply effort e to realize output. Alternatively, after diagnosis, he may transfer the opportunity, or *refer* it, to another agent, who then applies effort to it. In either case, the agent who ultimately applies effort e to the opportunity incurs a private cost $\psi(e)$, defined for all $e \in R_+$, with $\psi(0) = 0$. We assume that this function is \mathcal{C}^3 with $\psi' > 0$, $\psi'' > 0$, $\psi''' \geq 0$, and $\psi'(0) = 0$.⁸

Assumption 2. Effort is not observable or verifiable to agents other than the one who provides it. The value of the opportunity is not observable or verifiable by agents other than the one who diagnosed it.

Informational asymmetries thus favor those who obtained private information about the opportunity through their diagnosis. The assumption that only the agent who diagnosed the opportunity is informed about its value captures the idea that an economic loss is produced when the information obtained from the draw is wasted and replaced by the strategic misrepresentation of the value of this opportunity by the referring agent. This assumption can be relaxed with no impact on the analysis by assuming that both the referring agent and the agent accepting the referral may spend resources diagnosing the problem and observing noisy signals on its value. As long as the signal of the referring agent has some informational content, its communication biases the effort choice of the high-skill agent, who combines that signal with her own to form a posterior on the value of the opportunity.

The trade in opportunities thus takes place in an environment subject to both moral hazard and adverse selection. The adverse selection is due to the unobservability of the quality of the opportunity to the agent who did not diagnose it. The moral hazard is due to the unobservability of effort.

Agents are (observably) differently skilled, so that some agents add more value to each opportunity than others. Low-skill agents have skill $\theta_l = \eta$ with $\eta \in (0, 1)$, and high-skill agents have skill θ_h normalized to 1, $\theta_h = 1$.

Assumption 3. The value of the opportunity, effort and skill are complementary.

The marginal value of effort and talent is thus higher in higher value opportunities. Furthermore, output is non stochastic. We can then write output as

$$y = \theta ev.$$

Moreover, comparative advantage holds when the reservation utilities of the agents do not reverse these complementarities.

Assumption 4. Comparative advantage holds – i.e., the value of the opportunities and of the reservation utilities of the agents are such that the *efficient matching* under symmetric information has the low-skill agent deal with the less valuable opportunities and the high-skill agent deal with the more valuable ones.

The role of the referral, then, is to match a higher value problem with a more skilled agent. Formally, agents who do not tackle an opportunity gain a reservation utility, which we denote by \bar{u}^h and \bar{u}^l for the high- and low-skill agent, respectively. Then, Assumption 4 implies that:

$$\begin{aligned}
 (1) \quad & \max_e \{\eta ev_0 - \psi(e)\} + \bar{u}^h > \max_e \{ev_0 - \psi(e)\} + \bar{u}^l, \\
 (2) \quad & \max_e \{\eta ev_1 - \psi(e)\} + \bar{u}^h < \max_e \{ev_1 - \psi(e)\} + \bar{u}^l, \\
 (3) \quad & \bar{u}^l \leq \max_e \{\eta ev_0 - \psi(e)\}.
 \end{aligned}$$

That is, v_0 should be assigned to the low-skill agent (1), v_1 to the high-skill agent (2), and the low-skill agent prefers to draw opportunities rather than enjoy his reservation utility (3), so that opportunities flow into the economy.

Assumption 4 implicitly defines a ‘*first-best referral set*,’ namely, the set of high-value opportunities that meet inequality (2), given the values of v_0 , η , \bar{u}^h , and \bar{u}^l that satisfy (1) and (3). We define $v^{fb}(v_0, \eta, \bar{u}^h, \bar{u}^l)$, or v^{fb} for short, as the frontier of this set. The first-best referral set is then $v_1 \geq v^{fb}$.⁹ In this paper we focus on the role of two institutions, spot markets and ex-ante arrangements, in supporting the efficient matching, and throughout we compare the first-best referral set with the referral set supported by these two institutions.

We follow most of the literature on static (one period) asymmetric information problems in making two assumptions throughout the paper. First, we rule out the possibility of having a third party whose only function is to break the budget (i.e. an agent who consumes but does not affect output).¹⁰ Second, the parties can commit to the contracts they sign, so that no renegotiation takes place ex post.¹¹ Moreover, we restrict ourselves to deterministic mechanisms and to organizational forms that solely combine one low-skill agent with a high-skill agent.¹²

II. FIXED PRICE CONTRACTS AND THE DIRECTION OF REFERRALS

The first-best output and the corresponding welfare level serve as a benchmark throughout. The planner's problem is:

$$\max_{y_0, y_1} \left\{ \pi \left[y_0 - \psi \left(\frac{y_0}{\eta v_0} \right) + \bar{u}^h \right] + (1 - \pi) \left[y_1 - \psi \left(\frac{y_1}{v_1} \right) + \bar{u}^l \right] \right\},$$

where we have made use of the fact that $y = \theta ev$, with $\theta_h = 1$, and $\theta_l = \eta$, to replace effort with output. The first-best level of output is then given by:

$$(4) \quad \psi' \left(\frac{y_0^{fb}}{\eta v_0} \right) = \eta v_0 \quad \text{and} \quad \psi' \left(\frac{y_1^{fb}}{v_1} \right) = v_1.$$

Can this first-best allocation be implemented under asymmetric information on the quality of the opportunities? Consider first downstream referrals – i.e., those that flow from high- to low-skill agents. In this case, a fixed-price referral market (one in which the price does not depend on the value of the opportunity) can achieve the first-best allocation. Simply put, the higher skill agent prefers to refer the less valuable opportunities; a transfer of these opportunities downstream is thus incentive compatible from her perspective. Since this is the first-best allocation, we can decentralize it even under informational asymmetries.

For the same reason, fixed prices also support efficient matching if opportunities flow among workers with orthogonal skills ('horizontal' referrals). Since in this case one cannot extract any value from an opportunity of a specialty other than one's own, even a fixed price of 0 supports the efficient allocation.

Consider instead a 'bottom-up' or 'upstream' referral, in which a low-skill agent draws a highly valuable opportunity and must transfer it to a high-skill agent. Given the informational asymmetry, fixed-price contracts cannot motivate the low-skill agent to transfer the best opportunities and keep the worst. The argument is standard (Akerlof, 1970). Any price that is sufficient to encourage a low-skill agent to refer a high-value opportunity leads him to refer a lower-value opportunity as well. We summarize these results in the following proposition, which we prove, together with all other results, in the Appendix.

- Proposition 1.** a. Assume that opportunities flow from higher to lower skill agents. Then there exists a fixed price, p^* , that implements the first-best allocation.
- b. Assume that opportunities flow from lower to higher skill agents. Then there does not exist a fixed price that implements the first-best allocation.

A contract other than a fixed-price contract is then necessary to support an efficient referral flow when the diagnosis is undertaken by low-skill agents. This does not mean, however, that a fixed-price contract may not be preferred. The next section discusses how different spot market arrangements, including fixed prices, mediate the upstream referral flow.

III. REFERRALS IN THE MARKET

How do spot markets deal with the informational obstacles involving the upstream referral of opportunities? There are three possible allocations. First, the spot market can implement the efficient matching prescribed by the first best. Second, all opportunities may flow upstream, a possibility we term over-referral; and third, opportunities may not flow at all and instead be retained by the low-skill agents who drew them. We term this last allocation under-referrals.

Throughout the paper, we assume that opportunities are scarce, relative to the measure of high-skill agents, and that all the surplus thus flows to the low-skill agents. Formally, low-skill agents make a take-it-or-leave-it offer to the high-skill agents after observing the opportunity.

III.A Adverse selection in the spot market

Consider the contracting problem of a low-skill agent who drew a high-value opportunity and would like to refer it ex post to a more skilled agent. An output-based contract consists of an offer of an opportunity by the low-skill agent in exchange for a contingent payment by the high-skill agent $s^l(y)$. The compensation of the high-skill agent is then $s^h(y) = y - s^l(y)$.

Clearly, the contract offered by the low-skill agent may signal information to the high-skill agent about the value of the opportunity drawn. Let μ be the equilibrium belief of the high-skill agent that the value of the opportunity is v_0 , and $U^l(v_i, s^l(y), \mu)$ be the utility of a low-skill agent who offers an opportunity v_i , in exchange for a contingent payment schedule $s^l(y)$, facing high-skill agents with beliefs μ .

We limit our attention to pure strategy equilibria, both pooling and separating. A separating equilibrium in this context is one in which the contract offered allows the high-skill agent to differentiate high (v_1) from low (v_0) value opportunities. In a pooling equilibrium, the low-skill agent's take-it-or-leave-it offer is independent of the value of the opportunity drawn.

Our setup defines a game with many Perfect Bayesian Equilibria (PBE). A PBE is a set of strategies and a set of beliefs such that the strategies are optimal, given the beliefs at each node, and the beliefs are obtained through Bayesian updating of previous beliefs using the

equilibrium strategies and observed actions. Consistent with most of the literature on signaling games, we select reasonable equilibria by requiring that they satisfy the ‘intuitive criterion’ of In-Koo Cho and David Kreps (1987).¹³ In our context, this requires that, given a candidate equilibrium in which the utility of the low-skill agent who draws v_i is given by $U(v_i)$, there do not exist output contingent payment schedules $s^l(y)$ such that $U^l(v_1, s^l(y), 0) > U(v_1)$ while $U(v_0) > U^l(v_0, s^l(y), 0)$. The reason is that if these contracts were to exist, a high-skill agent who observes them would believe that the contract is being offered by an agent who drew a valuable opportunity, which would then lead to the failure of such a candidate equilibrium.

III.A.1 Incentive constraints

The contract must solve two incentive conflicts to implement efficient matching. First, low-skill agents with a low value opportunity may find it in their interest to refer the opportunity to the high-skill agents – an adverse-selection problem. Second, high-skill agents may choose to supply too little effort, obtain a low output, and blame the referral – a moral-hazard problem. We study next each incentive problem in turn.

The contingent payment contract must prevent the low-skill agent who drew opportunity v_0 from referring it, pretending it was v_1 . Denote the prescribed effort by e_1 , and the corresponding output that results from applying this effort to a v_1 opportunity by y_1 , with $y_1 = \theta_h e_1 v_1 = e_1 v_1$, since $\theta_h = 1$. Then the output produced by the high-skill agent under the deviation of the low-skill agent is:

$$(5) \quad e_1 v_0 = y_1 \left(\frac{v_0}{v_1} \right).$$

We denote the compensation that the low-skill agent obtains under the deviation as $\hat{s}^l = s^l(y_1 v_0 / v_1)$. Then the incentive compatibility constraint of the low-skill agent is:

$$(6) \quad y_0^{fb} - \psi \left(\frac{y_0^{fb}}{\eta v_0} \right) \geq \bar{u}^l + \hat{s}^l.$$

From the moment the opportunity is referred, a new informational problem is triggered – namely, the inability of the referring party to contract on the effort supplied by the high-skill agent. Suppose that the low-skill agent does indeed refer v_1 . If the high-skill agent performs the prescribed effort, the compensation of the low-skill agent is $s_1^l = s^l(y_1)$. For the high-skill agent to conceal the value of the opportunity, she must make the output look as if the low-skill agent had referred v_0 rather than v_1 , and she had performed the prescribed effort y_1/v_1 . By expression (5), the high-skill agent conceals her under-performance if she provides effort

$(y_1 v_0)/v_1^2$, which is lower than the one prescribed. The incentive compatibility constraint of the high-skill agent is then:

$$(7) \quad s_1^h - \psi\left(\frac{y_1}{v_1}\right) \geq \hat{s}^h - \psi\left(\frac{y_1 v_0}{v_1 v_1}\right),$$

The left-hand side of this constraint is the utility of the high-skill agent under no deviation, and the right-hand side is the utility she obtains when she deviates. The compensation must be such that:

$$(8) \quad s_1^h = y_1 - s_1^l \quad \text{and} \quad \hat{s}^h = y_1 \frac{v_0}{v_1} - \hat{s}^l,$$

where s_1^h is the compensation received by the high-skill agent if the output produced is as prescribed, and \hat{s}^h is her compensation under the deviation.

The trade-off between the adverse-selection and the moral-hazard problem is now clear. Given (6), the adverse-selection problem can be dealt with by setting a sufficiently low \hat{s}^l . However, the risk of moral hazard places limits on how low \hat{s}^l can be. If no rents are left to the low-skill agent who cheats, the high-skill agent may prefer to reduce output and pretend she was cheated. Notice that any other effort deviation by the high-skill agent produces an output level outside the set

$$\mathcal{Y} = \left\{ y_1, y_1 \frac{v_0}{v_1} \right\},$$

and, hence, identifies the high-skill agent as the shirker. With unlimited liability these deviations can be prevented by sufficiently penalizing the shirker (Legros and Matthews, 1993). For this reason, and without loss of generality, we ignore output levels outside \mathcal{Y} throughout.

Also, if a referral takes place, the high-skill agent must obtain at least her outside value, that is:

$$(9) \quad s_1^h - \psi\left(\frac{y_1}{v_1}\right) \geq \bar{u}^h.$$

Constraints (6), (7), (8) and (9) can be simplified and combined into a single constraint. The individual rationality constraint of the high-skill agent (9) must be binding, since the offer is take-it-or-leave-it. This determines s_1^h . Making use of this expression, together with equation (8), we can combine expressions (6) and (7) to yield a unique constraint. The low-skill agent makes a take-it-or-leave-it offer that maximizes his compensation, $s_1^l + \bar{u}^l$. Using (8) and (9) allows us to write both the objective and the constraints in the space of outputs and thus concentrate on the allocative efficiency of the spot market transaction. The program of the low-skill agent who draws v_1 and chooses to refer it is then:

Program \mathcal{P}^m

$$(10) \quad \max_{y_1} \left\{ y_1 - \psi \left(\frac{y_1}{v_1} \right) + \bar{u}^l - \bar{u}^h \right\},$$

subject to:

$$(11) \quad y_0^{fb} - \psi \left(\frac{y_0^{fb}}{\eta v_0} \right) + \bar{u}^h \geq \Delta^o(y_1) + \bar{u}^l,$$

where

$$(12) \quad \Delta^o(y_1) = y_1 \frac{v_0}{v_1} - \psi \left(\frac{y_1 v_0}{v_1 v_1} \right).$$

$\Delta^o(y_1)$, which we call the *over-referral deviation surplus*, plays a key role in what follows. It is the surplus that results when either the low-skill agent deviates and refers v_0 or the high-skill agent deviates and pretends v_0 has been referred. Then (11) simply says that the surplus generated when the low-skill agent retains v_0 must be larger than the surplus generated under any possible deviation.

Program \mathcal{P}^m is the problem of a low-skill agent who draws opportunity v_1 and chooses to refer it. The low-skill agent, however, can always choose to retain the opportunity. Thus, he chooses to refer whenever:

$$(13) \quad s_1^l + \bar{u}^l \geq \max_y \left\{ y - \psi \left(\frac{y}{\eta v_1} \right) \right\}.$$

In characterizing the equilibrium, we proceed in two steps. First, we characterize the solution to program \mathcal{P}^m under the assumption that (13) is met. Then, in section III.B, we discuss under what conditions this is indeed the case. We proceed in this way because it is necessary to understand the distortions required to implement separation in order to understand the referral flow.

III.A.2 Distortions of output and the separating contract

Let y_0^m and y_1^m be the output associated with opportunities v_0 and v_1 , respectively, under the separating contract. The next proposition characterizes y_0^m and y_1^m as a function of $v_1 > v^{fb}$, where v^{fb} was defined in Section I.

Proposition 2. There exists a unique v^s such that

- a. if $v_1 \leq v^s$, the separating contract is such that $y_0^m = y_0^{fb}$ and $y_1^m = y_1^{fb}$ and
- b. if $v^s < v_1$, the separating contract is such that $y_0^m = y_0^{fb}$ and $y_1^m < y_1^{fb}$.

The spot-market contract can thus never implement the first best if v_1 is sufficiently high. In fact, there may not exist any v_1 such that $v_1 \leq v^s$, in which case the spot market can never mediate transactions without distortions. This occurs whenever $v^s < v^{fb}$.

To better understand the intuition behind this proposition, turn to constraint (11) and consider the problem from the perspective of a low-skill agent who drew v_0 . When v_1 is high, so is the temptation of the low-skill agent to deviate by pretending he is referring v_1 . To avoid this, his deviation compensation must be reduced. But this, in turn, increases the deviation compensation of the high-skill agent, making her deviation more enticing. Thus, if v_1 is sufficiently high, only a reduction of the output of the high-skill agent below what is optimal makes it sufficiently unattractive for the low-skill agent to refer a low quality opportunity, without leading the high-skill agent to claim that the referring agent is cheating.

The next corollary shows that these distortions are severe enough as to eliminate the effect of the complementarity between effort and value of the opportunity. That is, while in the first-best effort would increase with the value of the opportunity, when the first best cannot be implemented, the effort of the high-skill agent is decreasing in the value of the opportunity.

Corollary 3. Assume that $v^s < v_1$, that is, the first best cannot be implemented by the separating contract. Then, the effort exerted by the high-skill agent on the opportunity referred is a decreasing function of its value.

III.B Referrals in the market

Given this separating contract, what referral flow do we expect to observe in the market? As mentioned above, there are three possibilities: efficient matching, under-referrals, and over-referrals. The next Lemma shows, however, that this last type of referral flow cannot be supported in equilibrium.

Lemma 4. No equilibrium with over-referrals exists that satisfies the intuitive criterion.

An equilibrium with over-referrals could a priori be of two types: pooling or separating. A separating equilibrium with over-referrals is ruled out by comparative advantage (Assumption 4), which says that there is not enough surplus to compensate agents for the outside value foregone if the referral of v_0 actually takes place. A pooling equilibrium with over-referrals relies instead on implausible beliefs, and does not survive the intuitive criterion. Essentially, the agent who drew a low-value (v_0) opportunity never moves away from the pooling, since

he is subsidized by the fixed price paid by the high-skill agent. Thus, the low-skill agent who drew v_1 can always signal his type by offering a contract under which those with v_0 would never want to refer.

The sharing contract of Section III.A. can implement efficient matching, but, as seen in Corollary 3, it can only do so at the expense of severe effort distortions for high values of v_1 . The next proposition shows that, for high enough v_1 , these distortions make efficient matching impossible.

Proposition 5. There exists a value \bar{v} such that for all $v_1 > \bar{v}$ the market cannot support efficient matching, and the market equilibrium is characterized instead by under-referrals.

As v_1 becomes large, the distortions imposed by the contract described in Section III.A increase (as Corollary 3 shows) and, for sufficiently large values of v_1 , eventually become too large to support separation. This means that under-referrals must prevail for high v_1 . Note that, for high values of v_1 , over-referrals could Pareto improve, even ex post, on the separating allocation, but no trade exists in equilibrium.¹⁴ We close this section by illustrating, in the context of an example, the rich market referral patterns that even a simple economy (like a quadratic one) can exhibit.

Example 1. Let $\psi(e) = e^2/2$. Also assume that $v_0 = 1.4$, $\pi = 0.75$, $\bar{w}^h = 1.15$, and $\bar{w}^l = 0$.

Figure 1 reports the possible allocations in the space (η, v_1) . Efficient matching results in regions a and b , though in region b the first-best level of effort cannot be implemented. Region c is the set of economies for which under-referrals are observed.

As shown in Proposition 5, as v_1 increases, the distortions required to implement efficient matching eventually become so large that sharing contracts are too costly, thus leading to under-referrals. Also, as η increases, the comparative advantage of the high-skill agent is not enough to compensate for the distortions that efficient matching requires, and referrals may not be supported – even for relatively low values of v_1 .

FIGURE 1 ABOUT HERE

IV. EX-ANTE CONTRACTS: PARTNERSHIPS AND RETAINERS

We study here situations in which agents can enter ex-ante agreements and make ex-ante transfers to each other. We start, in Section IV.A, by studying the partnership contract,

whereby agents agree ex ante to the allocation of opportunities and income from these opportunities. In Section IV.B we study retainer contracts, and show that ex-ante transfers allow agents to sustain pooling equilibria with over-referrals, an allocation the spot market could not support. In Section IV.C we discuss the agents' referral choices given the surplus attainable under these arrangements and the under-referral allocation. We close this section by discussing, in section IV.D, the implementation of the partnership allocation.

IV.A Partnerships and efficient matching

As we document below in our description of professional service firms, joint billing in partnerships allows the parties to implement the allocation of income and opportunities agreed in advance. This in turn allows the parties, as this section shows, to implement agreements that involve the referral of some opportunities and not of others, even when they do so at the expense of effort distortions on opportunities not referred. We study next the incentive compatibility constraints associated with the partnership, characterize the resulting allocation, and compare them with the corresponding ones in the market.

IV.A.1 The partnership problem

The partnership must deal with moral hazard on both sides of the relationship and with adverse selection on the referring agent's side. The form of the moral-hazard problem of the high-skill agent is the same as that in the spot-market problem, so (7) must also hold here. The two other constraints take a slightly different form here that we discuss in what follows.

First, consider the adverse-selection problem. Assume that v_0 has been drawn. To prevent the wasteful referral of this opportunity, the partnership prescribes effort $y_0/(\eta v_0)$ and compensation s_0^l so as to solve the adverse-selection problem; that is,

$$(14) \quad s_0^l - \psi \left(\frac{y_0}{\eta v_0} \right) \geq \bar{u}^l + \hat{s}^l,$$

where, as before, \hat{s}^l is the compensation that the low-skill agent obtains under the deviation.

Equation (14) can be compared with the corresponding constraint in the spot-market transaction, equation (6), where the effort exerted by the low-skill agent was always first best, as he kept ownership of the opportunity drawn. In contrast, the partnership can use the prescribed effort and compensation to ease the adverse-selection problem, as the low-skill agent has pledged the opportunity to the partnership.

Second, when v_1 is drawn, the contract must ensure that the low-skill agent prefers to refer the opportunity rather than keep it and pretend that v_0 was drawn. Clearly, if the low-skill agent deviates and announces v_0 , the resulting output must be consistent with this

announcement, and this can only be achieved if the effort supplied is $y_0/(\eta v_1)$. Any other deviation unambiguously identifies the low-skill agent as a shirker and thus can be prevented by penalizing him sufficiently. The partnership then prevents the relevant deviation by setting:

$$(15) \quad s_1^l + \bar{u}^l \geq s_0^l - \psi \left(\frac{y_0}{\eta v_1} \right).$$

Equation (15) should be compared with (13), which determines whether the referral of v_1 takes place in the spot-market case. In the case of spot-market sharing contract, a low-skill agent who retains v_1 maximizes over effort, as the right hand side of (13) shows. In contrast, the partnership forces the low-skill agent to provide effort consistent with y_0 . The partnership then uses (15) to ease the referral process.

Finally, the budget constraint (8) has to hold. Now we also have to consider the constraint in the case where v_0 is realized,

$$(16) \quad s_0^l + s_0^h = y_0.$$

As in the previous section we can reduce constraints (14), (15), (7), plus the budget constraints (8) and (16) to a single one and simply write incentive compatibility in the space of outputs.¹⁵ The partnership problem can then be written:

Program \mathcal{P}^P

$$(17) \quad \max_{y_0, y_1} \left\{ \pi \left[y_0 - \psi \left(\frac{y_0}{\eta v_0} \right) + \bar{u}^h \right] + (1 - \pi) \left[y_1 - \psi \left(\frac{y_1}{v_1} \right) + \bar{u}^l \right] \right\}$$

subject to

$$(18) \quad y_0 - \psi \left(\frac{y_0}{\eta v_0} \right) + y_1 - \psi \left(\frac{y_1}{v_1} \right) \geq \Delta^o(y_1) + \Delta^u(y_0),$$

where

$$(19) \quad \Delta^u(y_0) = y_0 - \psi \left(\frac{y_0}{\eta v_1} \right),$$

and $\Delta^o(y_1)$ was defined in equation (12). Here, $\Delta^u(y_0)$, the *under-referral deviation surplus*, determines the surplus appropriated by a low-skill agent who deviates by not referring v_1 and who pretends v_0 was drawn instead, reducing his effort to produce the prescribed y_0 . Constraint (18) should be compared with (11), the corresponding one in the case of the market. There, the surplus generated when the low-skill agent retains v_0 had to be larger than the surplus generated under any possible deviation, which included the possibility of over-referral of v_0 or the pretense of over-referral by part of the high-skill agent. Here, ex-ante transfers make it

possible to use the equilibrium surplus in either state to compensate for the surplus foregone when the agents abstain from deviating, which now includes deviations in y_0 .

IV.A.2 Characterization of the partnership contract

Let y_0^p and y_1^p be the output associated with opportunities v_0 and v_1 respectively under the partnership equilibrium. Then we can show the following result:

Proposition 6. There exists a unique v^p such that

- a. if $v_1 < v^p$ the unique partnership separating equilibrium is such that $y_0^p < y_0^{fb} = y_0^m$ and $y_1^m < y_1^p < y_1^{fb}$, and
- b. if $v^p \leq v_1$ the unique partnership separating equilibrium is such that $y_0^p = y_0^{fb}$ and $y_1^p = y_1^{fb}$.

That is, the partnership contract can always implement the first best if v_1 is sufficiently high. Only when v_1 is low may the partnership require distortions to mediate transactions. Notice the stark contrast with the spot-market equilibrium, which cannot support efficient matching for sufficiently large v_1 .

The intuition behind Proposition 6 can be immediately grasped by evaluating (18) at the first best. First, the complementarity between effort and the value of the opportunity results in a first-best level of surplus, the left-hand side of (18), which is increasing and convex in the value of the opportunity v_1 . In contrast, the over-referral deviation surplus can grow only at a decreasing rate. The reason is that the effort under the deviation, $(y_1^{fb}v_0)/(v_1^2)$, grows at a lower rate than the first-best effort, y_1^{fb}/v_1 , and as a consequence so does the output and $\Delta^o(y_1)$. Finally, the under-referral deviation surplus is bounded from above by y_0 . Therefore, for v_1 sufficiently large there is enough surplus to compensate both the high- and low-skill agents for the rents they forego when they do not deviate.

Conversely, if v_1 is not high enough, it *may* be the case that (18) is not met when evaluated at the first best. Notice, though, that even in the presence of distortions, the partnership can sustain a higher level of output in the high-value opportunities than the market. The reason is that a judicious choice of y_0 can relax the distortions on the valuable effort of the high-skill agent.

The extent of the effort distortion depends on the frequency of referrals prescribed by the first best, as summarized by the probability of drawing a claim of high value, $1 - \pi$. As the next corollary shows, as π increases, referrals become relatively less important than the preservation

of the effort incentives on the opportunities that are more frequent, v_0 . Partnerships then trade off effort incentives against referral incentives.

Corollary 7. y_0^p (y_1^p) is an increasing (decreasing) function of π .

IV.B Retainer contracts, pooling, and over-referrals

The ability of the agents to engage in ex-ante transfers also allows for the possibility of equilibria with over-referrals, an allocation that the spot market could not support. Consider, in particular, a *retainer* contract, that is, a put option contract where the low-skill agent pays ex ante a price P to the high-skill agent for the right to sell her ex post an opportunity in exchange for a fixed payment K .¹⁶ As in Proposition 1, such a contract cannot support the separation of good and bad opportunities – a fixed payment that leads to the transfer of v_1 will also lead to the transfer of v_0 . Moreover, an agreement to transfer only bad opportunities destroys comparative advantage and is never in the interest of the agents.

To see that a retainer contract can support equilibria with over-referrals, consider, in particular, one where the transfer ex post from the high-skill agent to the low-skill agent is equal to the maximum surplus that the best opportunity (v_1) can generate, that is $K = \max_y \left\{ y - \psi \left(\frac{y}{v_1} \right) \right\} - \bar{u}^h$. Ex ante, the low-skill agent must compensate the high-skill agent for this transfer through a fixed payment

$$P = K + \bar{u}^h - \pi \max_y \left\{ y - \psi \left(\frac{y}{v_0} \right) \right\} - (1 - \pi) \max_y \left\{ y - \psi \left(\frac{y}{v_1} \right) \right\}.$$

Clearly, given this price, the high-skill agent accepts to be on a retainer, since she obtains her reservation utility \bar{u}^h . Given that the payment is independent of the low-skill agent's report, upon exercise he reports truthfully the value of the opportunity. The retainer contract thus supports over-referrals and satisfies the intuitive criterion: ex post no deviation from this equilibrium involving signaling v_1 through a sharing contract can possibly yield more than K , the net first-best surplus associated with v_1 . Whether the retainer contract arises in equilibrium depends on whether it is in the ex-ante interest of the parties. The ex-ante welfare associated with the retainer contract is:

$$(20) \quad \bar{u}^l + \pi \max_y \left\{ y - \psi \left(\frac{y}{v_0} \right) \right\} + (1 - \pi) \max_y \left\{ y - \psi \left(\frac{y}{v_1} \right) \right\}.$$

IV.C Referrals when ex-ante contracts are feasible

Consider now the decision of a set of agents who may enter an ex-ante agreement possibly involving joint billing. Agents who can choose ex ante will choose the welfare maximizing

allocation among those that can be supported. Thus an advantage of ex-ante contracts is the elimination of Pareto-dominated equilibria.

Agents may choose among three possible allocations. First, the under-referral allocation, in which all opportunities are dealt with by the low-skill agent. In this case, welfare is:

$$(21) \quad \bar{u}^h + \pi \max_y \left\{ y - \psi \left(\frac{y}{\eta v_0} \right) \right\} + (1 - \pi) \max_y \left\{ y - \psi \left(\frac{y}{\eta v_1} \right) \right\}.$$

Second, agents may choose to enter into a retainer contract and obtain the surplus (20). Third, they may enter into a partnership agreement that implements efficient matching and achieves the surplus level given by the objective function in \mathcal{P}^P evaluated at the partnership equilibrium allocation (y_0^p, y_1^p) .¹⁷ Then the next result follows immediately from Proposition 6.b, and is given without proof.

Corollary 8. For all $v_1 > v^p$ the partnership dominates both the over-referral and the under-referral allocations, resulting in efficient matching.

Notice that the result in Corollary 8 stands in stark contrast to Proposition 5. There, we showed that the market can never support efficient matching if v_1 is sufficiently high; here we show that the contrary is true in the partnership. What is the source of the dramatic difference between these two institutions in their ability to implement efficient matching?

First, the partnership can reduce the effort incentives of the low-skill agent on the opportunities he keeps, reducing in turn his incentives to keep the valuable ones. This was not possible in the spot market as only the low-skill agent had access to those opportunities. For this reason, the distortions in the high-skill agent's effort needed to ensure efficient matching are lower than those needed in the spot market. This enlarges the referral set. To put it differently, partnerships improve on the market allocation because they trade off better effort incentives where these are most valuable for worse effort incentives where these are least valuable, namely on those opportunities dealt with by low-skill agents.

Second, the partnership contract limits ex-post competition for opportunities by specifying ex ante with whom these may be shared. This further raises the ability of the partnership to achieve transfers without imposing the distortions required in the spot market. Recall that in the spot market the utility obtained by the high-skill agent when she works on a good opportunity is limited by competition for these opportunities.¹⁸ In turn, this equilibrium utility must be higher than her deviation utility to avoid moral hazard on her part (see equation (7)). But this effectively places a ceiling on the deviation compensation of the high-skill agent

which in turn places a floor under the punishment that may be stipulated for a low-skill agent who deviates by over-referring v_0 . This ceiling is increasing in the value of the opportunity v_1 , so eventually, for sufficiently high value opportunities, effort must be distorted. In the partnership, on the other hand, competition places no ex-post constraint on the compensation that can be obtained by the high-skill agent.

The next example illustrates the ability of the partnership to sustain the first best allocation when v_1 is sufficiently high.

Example 1 (cont.) In regions e and f of Figure 2, v_1 is sufficiently close to v_0 , so that Proposition 6.a holds, and the ex-ante contract cannot implement the first best. In region e the distortions are severe enough as to preclude referrals completely and hence under-referrals obtain, whereas in region f the partnership arises in equilibrium, but it cannot support the first-best allocation of effort. As Proposition 6.b proved, for $v_1 \geq v^p$ (as given by the line separating regions f and g) the partnership implements the first best. This concludes the example.

FIGURE 2 ABOUT HERE

In the previous example, retainers never dominated either under-referrals or the partnership allocation. Is this always the case? As the next example shows, retainers can dominate either under-referrals or the partnership allocation whenever the probability of drawing the high value opportunity, $1 - \pi$, is sufficiently high. In this case, it pays to occasionally miss-match the high-skill agent with v_0 in order to preserve her first best level of effort in v_1 .¹⁹

Example 2. Let $v_0 = 1.4$, $v_1 = 2.05$, $\eta = .9$, $\bar{u}^h = .5$, and $\bar{u}^l = .2$. For these parameters the partnership cannot implement the first best. For $\pi \in (0, .48)$, v_1 occurs frequently enough to make retainers preferred to either the under-referrals or the partnership allocations. For $\pi \in [.48, .93]$ partnerships dominate. Finally, for $\pi \in [.93, 1)$, v_0 occurs frequently and the (ex-ante) costs of occasionally miss-allocating v_1 to the low-skill agent are sufficiently low to make the under-referral allocation preferred.

IV.D Implementation

We close this section with some comments on the particular contractual arrangement implementing the partnership allocation, $[y_0^p, y_1^p]$. We have shown above that the trade-off between referral incentives and effort incentives is present even when ex-ante transfers are

possible and agents have unlimited liability. Clearly though, agents are often financially constrained and have limited liability. These institutions may affect the set of feasible allocations and determine the ultimate form of the partnership contract.

Specifically, in order to implement $[y_0^p, y_1^p]$ the partnership contract has to satisfy

$$(22) \quad \frac{\left[y_0^p - \psi \left(\frac{y_0^p}{\eta v_0} \right) \right] - \Delta^u (y_0^p)}{y_1^p - y_1^p \frac{v_0}{v_1}} \geq \left(\frac{s_1^h - \hat{s}^h}{y_1^p - y_1^p \frac{v_0}{v_1}} \right) - 1 \geq \frac{\Delta^o (y_1^p) - \left[y_1^p - \psi \left(\frac{y_1^p}{v_1} \right) \right]}{y_1^p - y_1^p \frac{v_0}{v_1}}.$$

Once s_1^h and \hat{s}^h are chosen to meet (22), s_1^l and \hat{s}^l follow immediately from (8). Finally, given these choices, s_0^l is set to satisfy (14) and (15), which is always possible. Incentive compatibility is then only a restriction on the “slope” of the compensation of the high-skill agent in \mathcal{Y} . The left hand side of (22) is less than zero whereas the right hand side is greater than -1 . Thus, in the set \mathcal{Y} , the compensation of the high skill agent grows with output but less than proportionally. With unlimited liability, the level of the high-skill agent’s compensation is undetermined and we can always set $s_1^h = y_1^p$ and \hat{s}^h so that (22) is met. Under limited liability, on the other hand, sharing of the output produced with v_1 is a necessary feature of the resulting partnership contract, as observed in practice and documented below. Indeed, if we set $s_1^h = y_1^p$, it follows from (14) and (15) that $\hat{s}^l < 0$, independently of the value of s_0^l . If instead we impose that $\hat{s}^l \geq 0$ then it has to be that $s_1^l > 0$ and thus $s_1^h < y_1^p$.

Finally, cognitive limitations may force the partnership to restrict itself to contracts that are linear in output. Clearly this is not without efficiency losses. The trade-off between referral incentives and effort incentives though is still present. In particular, it can be immediately shown that if linearity is imposed, incentive compatibility strictly requires sharing of the output produced by the high skill agent and thus that the first best can never be implemented.

V. EVIDENCE

The simple model we have presented can help us to understand referral contracts in markets and partnerships. Moreover, the model results, particularly the need to balance incentives for effort provision and for referrals, can illuminate a broad set of practices in professional service organizations. Finally, the theory developed here has implications for the scope of professional partnerships. We proceed now to explore these three sets of implications.

V.A Contractual incentives for referrals in markets and partnerships

We document the existence and characteristics of the two main arrangements studied in this paper: ex-post, spot-market, referral arrangements through income sharing, and ex-ante,

where agents agree in advance to the allocation of opportunities and the revenue they generate.

An interesting example of a referral market is the injury claim referral market in the state of New York, which has been studied by Stephen J. Spurr (1988 and 1990). In this market, a lawyer who knows of a client with an injury claim may refer him to another lawyer. Both lawyers enter into a referral agreement that is sanctioned by the New York Bar Association.²⁰ As in our model, the referring and receiving lawyers divide up the income obtained from the claim. The contract only specifies the output shares corresponding to each lawyer, and does not bind any of the parties to devote a minimum time or effort to pursuing the claims. Spurr documents the existence of referral contracts involving substantial sharing of the recovery of the claim between the referring lawyer and the one who ends up doing the work.²¹ Regrettably, this data set does not allow us to test whether under-referrals characterize (as our model predicts) spot-market transactions.²²

Concerning ex-ante ‘partnership-like’ arrangements designed to encourage referrals, accounting, law and consulting firms, as well as some commercial banks rely on output sharing to decentralize the allocation of opportunities.²³ Some of them explicitly rely on objective distribution systems that reward referrals and business origination, imposing some income sharing even when it is the partner who drew the opportunity the one who does the work required.²⁴ The most common system, known as Hale & Dorr,²⁵ determines each partner’s share of the firm’s income according to *work done* (60%), *business origination* (30%), and *profit credit* (10%) (James D. Cotterman, 1995). Other systems rely directly on business origination, and award a share of the income to the partner who can claim to have brought the client to the firm (Howard Mudrick, 1990).

V.B Effort incentives and referral incentives in professional service firms

It is clear from our conversations with professional service firm members that partnerships are aware that the trade-off between incentives for referrals and for effort provision is key to the success of the partnership. Too low shares to originating partners may lead to under-referrals (or in the parlance of law firms, “hoarding of cases”); too high shares may lead to over-referrals and too little effort by receiving partners. We present here two notorious examples of the consequences of failing to deal with this conflict.

First, consider the acrimonious divorce between the consulting arm of Andersen Worldwide (then Andersen Consulting, now called Accenture) and its auditing arm (then Arthur Andersen, now plain Andersen).²⁶ The organization, initially only an auditor, entered consulting in 1954 to take advantage of the consulting opportunities that appeared while conducting

its auditing work. The resulting conflict was that, like in our analysis, while an auditing partner may not be best qualified to deal with a particular consulting opportunity, he may be able to extract some rents from dealing with it. The initial contractual solution was to divide up all the income obtained in all areas and share it equally between auditors and consultants, independent of their respective contribution to the actual work (a lock-step system). This system provided excellent incentives for the auditors to pass on the consulting opportunities to specialized consultants; however, it provided low effort incentives and resulted in low utility for the consultants relative to their outside values. In the ‘Florida accord,’ reached in 1988, the partners agreed to separate the income from the consulting arm and the income from auditing, and to transfer 15% of the income of the most profitable side to the least profitable side. This resulted in a net transfer of 15% of the consulting (AC) income to the auditing (AA) side of the firm. While this new contract improved effort incentives for consultants, it reduced sharply the referral incentives of the auditors, who could now choose between earning all the income of a consulting opportunity by keeping it, or just 15% of it by referring it to the consultants. As a result, auditors started to under-refer opportunities and keep them within the auditing arm (AA). The consultants (AC) considered this ‘internal competition’ unacceptable, and proceeded to request, in 1998, outside arbitration to accomplish the division of Andersen Worldwide in two separate partnerships.²⁷

A similar difficulty in providing sufficient incentives for referrals while maintaining effort incentives led to the fall of Watson, Leavenworth, Kelton & Taggart, a premier Park Avenue law firm considered one of the giants of patent and trademark law (see Weingarten (1981)). It disappeared when “the client-share system encouraged Watson, Leavenworth, Kelton & Taggart lawyers to guard their clients’ affiliation against intrusion by others, creating an atmosphere of competition among partners. Some partners suspected others of hoarding cases ...” The law firm eventually died when one of its star rainmakers decided to leave the law firm and “would likely take substantial accounts with him, including the Nestle Co. Inc., one of Watson, Leavenworth, Kelton & Taggart’s biggest clients.”

V.C Scope of the partnership

Our analysis points out that agents should create ex-ante referral arrangements or partnerships with those agents to whom they are likely to need to refer opportunities but with whom their skills overlap partially, so that a substantial threat of appropriability of the opportunity and the resulting misallocation exists. On the other hand, agents need not create a partnership when the opportunities belong to entirely different skill spaces (such as family

law and corporate law) or entirely different geographical areas. In fact, referrals between firms with sharply different geographical or product scope should not involve sharing contracts.

Indeed, referral contracts do not involve referral shares in instances when there is no threat of appropriation by the referring agent because he is not at all qualified or able to perform the referred opportunity. For example, law and other professional service firms form referral networks that allow one member of the network to refer to another clients whose problems fall outside their area of expertise or jurisdiction. Unlike in the arrangements discussed above, “fees are normally not shared among the firms, and law firms within networks do not charge for referrals.”²⁸ Once a client is referred, each firm separately charges for its services.

Also consistently with our discussion, law and consulting firms specialize in related and overlapping areas of the law, where the threat of appropriation is highest, and not in others. For example, “Wall Street” law firms like Cravath or Cleary offer their clients legal advice on securities law, mergers and acquisitions, commercial bank laws, etc., but not in maritime law. If a client happens to have a problem in such a clearly differentiated domain, they are referred to another firm that specializes in that field.

Thus, our set-up generates a theory of the size and boundaries of the partnership that supports transactions both inside the firm and in the market. The empirical implications about the types of practices that firms should include are very different from the implications of risk-sharing theories, such as Gilson and Mnookin (1985). These authors argue that law partnerships are a risk-pooling arrangement whose aim is to encourage ex-ante investment by lawyers in specialized areas which have uncertain future demand. Such risk-based theories counterfactually predict that firms should aim to diversify the types of skills they include. Our theory, instead, suggests that partnerships should aim to cover all areas in which referrals are likely to be necessary and where skills are related to such an extent that the threat of misappropriation is real.

VI. CONCLUSIONS

The referral problem is pervasive, not only in the professional services, from where most of our examples above proceed, but also in many other fields where diagnosis is costly. For example, a literary agent may know that some other agent would be better at promoting a writer currently in her roster. Or a university professor may know that an advisor other than himself would be a better fit for the star student he is currently advising. We have studied how spot markets and ex-ante arrangements deal with this referral problem. We have shown

that, if an under-qualified agent is in control of the information flow, the spot market can implement the efficient match between skills and opportunities only through income sharing, at the expense of distortions in the provision of effort of agents with higher skills. As the potential quality of the opportunity increases these distortions become larger, to the point where they are so large that the spot-market disappears and referrals do not take place.

Partnerships, whereby agents agree in advance to a particular allocation of income and opportunities, facilitate the provision of incentives and the exchange of referrals. In particular, they allow agents to break two key constraints of spot-market transactions. First, by punishing an agent who does not actually share an opportunity, they weaken his incentives to hold onto opportunities for which he is not best qualified. Second, by limiting his trading partners *ex ante*, they limit *ex-post* competition and thus expand the range of punishments available for an agent who wrongly refers an opportunity he should have kept. As a result of these advantages, partnerships can always implement efficient matching when opportunities are valuable enough. The drawback to such arrangements is lower effort across all opportunities, since punishing individuals who do not share opportunities weakens their effort incentives. We thus suggest that the boundaries of partnerships reflect trade-offs between facilitating the exchange of referrals and effort incentives. In particular, partnership contracts should appear between agents who are somewhat specialized, so that they have comparative advantage in different types of opportunities, but share the same skills to some extent. It is among these agents that the threat of appropriation of opportunities is most important.

There are a number of possible extensions of our analysis. First, we have assumed that the referring agent can either appropriate the full value of the opportunity he draws, or, alternatively, that he can commit to share the income from the opportunity. Some relevant cases may fall in-between, where keeping some opportunities in the partnership contract requires distorting incentives so much that a break-up is preferred.²⁹ Second, to focus on the important informational asymmetries between agents, our model has ignored the role that the client may play in the referral process. Integrating the client in the analysis is an important issue for future work.

We have shown that distortions naturally arise in a referral context even without risk aversion or limited liability. We believe, however, that it is this trade-off between effort incentives and referral incentives, rather than the ‘tenuous trade-off’ between risk and incentives (Prendergast (1999)) that plays the key role in the contractual design of professional service firms.

A. APPENDIX

Proof of Proposition 1. a. We construct a fixed price equilibrium that supports the referral downstream of v_0 . Incentive compatibility requires that:

$$(A.1) \quad \max_e \{ev_1 - \psi(e)\} \geq \bar{u}^h + p \geq \max_e \{ev_0 - \psi(e)\}$$

$$(A.2) \quad \max_e \{\eta ev_0 - \psi(e)\} - p \geq \bar{u}^l.$$

Since there are more low-skill agents competing for low value opportunities, they are kept at their reservation values, so that from (A.2), $p^* = \max_e \{\eta ev_0 - \psi(e)\} - \bar{u}^l$, and condition (A.2) is trivially met. As for (A.1) first notice that,

$$\bar{u}^h + p^* = \max_e \{\eta ev_0 - \psi(e)\} + \bar{u}^h - \bar{u}^l > \max_e \{ev_0 - \psi(e)\}$$

and where the last inequality follows from (1). Next notice that,

$$\bar{u}^h + p^* = \max_e \{\eta ev_0 - \psi(e)\} + (\bar{u}^h - \bar{u}^l) < \max_e \{\eta ev_1 - \psi(e)\} + (\bar{u}^h - \bar{u}^l) < \max_e \{ev_1 - \psi(e)\},$$

where, once again, the last inequality follows from (2). Hence condition (A.1) is met.

b. Assume the price p_{as} supports the referral of v_1 and the non referral of v_0 . Then it must be that

$$\max_e \{\eta ev_0 - \psi(e)\} > \bar{u}^l + p_{as} > \max_e \{\eta ev_1 - \psi(e)\}.$$

But $g_l(v) = \max_e \{\eta ev - \psi(e)\}$ is increasing in v , a contradiction. This concludes the proof of Proposition 1.

Proofs for Section III

Let $\varepsilon(v_1) = \left(\partial e_1^{fb} / \partial v_1\right) \left(v_1 / e_1^{fb}\right)$ the elasticity of the first best level of effort of the high-skill agent with respect to v_1 . Then we can prove the next

Lemma A0. (a) The first best level of effort, e_1^{fb} is a strictly increasing and concave function of v_1 . (b) $\lim_{v_1 \rightarrow \infty} e_1^{fb} = \infty$. (c) $0 < \varepsilon(v_1) \leq 1$.

Proof: (a) It follows from an immediate application of the implicit function theorem and $\psi'' > 0$ and $\psi''' \geq 0$.

(b) Assume to the contrary that $e_1^{fb} \rightarrow \bar{e} < \infty$, then because e_1^{fb} is a strictly increasing and concave function of v_1 it follows that it asymptotes to \bar{e} from below, and, in the limit it has to be then that

$$\lim_{v_1 \rightarrow \infty} \frac{\partial e_1^{fb}}{\partial v_1} = 0.$$

But,

$$\frac{\partial e_1^{fb}}{\partial v_1} = \frac{1}{\psi''(e_1^{fb})} \geq \frac{1}{\psi''(\bar{e})} > 0,$$

for $\psi''' \geq 0$ and ψ is defined for all e in R_+ so $\psi''(\bar{e})$ is well defined, and the contradiction follows.

(c) Expanding the first order condition with respect to e_1^{fb} around 0, we obtain,

$$1 = \frac{1}{v_1} \psi'(e_1^{fb}) = \frac{1}{v_1} \psi''(c) e_1^{fb},$$

for $c \in (0, e_1^{fb})$ and hence

$$1 \leq \frac{1}{v_1} \psi''(e_1^{fb}) e_1^{fb} \Rightarrow \frac{1}{\psi''(e_1^{fb})} \frac{v_1}{e_1^{fb}} = \varepsilon(v_1) \leq 1,$$

as $\psi''' \geq 0$. That $\varepsilon(v_1) > 0$ follows immediately from (a) above. This completes the proof of Lemma A0.

Define with some abuse of notation

$$(A.3) \quad S(y) = y - \psi\left(\frac{y}{v_1}\right) \quad \text{and} \quad \Delta^o(e; v_1) = ev_0 - \psi\left(e \frac{v_0}{v_1}\right),$$

and

$$(A.4) \quad \Omega^u(y_0) = \psi\left(\frac{y_0}{\eta v_0}\right) - \psi\left(\frac{y_0}{\eta v_1}\right),$$

that is the *net* surplus that accrues to the low-skill agent when he under-refers the high value opportunity. The incentive compatibility constraint of the partnership, expression (18), can then be written as

$$(A.5) \quad S(y_1) \geq \Delta^o(y_1) + \Omega^u(y_0).$$

We are interested in the characterization of $S(y_1^{fb})$, $\Delta^o(y_1^{fb})$ and $\Omega^u(y_0^{fb})$ as functions of v_1 and of $\Delta^o(e; v_1)$ as a function of e .

Lemma A1. (a) $S(y_1^{fb})$ is a strictly increasing and strictly convex function of v_1 . (b) $\Omega^u(y_0^{fb})$ is a strictly increasing and strictly concave function of v_1 . (c) $\Delta^o(y_1^{fb})$ is a strictly increasing and concave function of v_1 , with $\Delta^o(y_1^{fb}) \rightarrow \infty$ as $v_1 \rightarrow \infty$. (d) $\Delta^o(e; v_1)$ is a strictly increasing and strictly concave function of e for all $e < e_1^{fb}$.

Proof: (a) and (b). These are immediate from $\psi'' > 0$.

(c) First, basic computations show that

$$\frac{\partial \Delta^o(y_1^{fb})}{\partial v_1} = e_1^{fb} \frac{v_0}{v_1} \left(\varepsilon(v_1) + (1 - \varepsilon(v_1)) \frac{1}{v_1} \psi' \left(e_1^{fb} \frac{v_0}{v_1} \right) \right) > 0,$$

by Lemma A0. Next, tedious, but straightforward calculations show that,

$$\begin{aligned} \frac{\partial^2 \Delta^o(y_1^{fb})}{\partial v_1^2} = & - \frac{\left(e_1^{fb} v_0 \right)^2}{v_1^3} \psi'' \left(e_1^{fb} \frac{v_0}{v_1} \right) (\varepsilon(v_1) - 1)^2 - v_0 \frac{\psi''' \left(e_1^{fb} \right)}{\psi'' \left(e_1^{fb} \right)} \left(1 - \frac{1}{v_1} \psi' \left(e_1^{fb} \frac{v_0}{v_1} \right) \right) \\ & - 2\psi' \left(e_1^{fb} \frac{v_0}{v_1} \right) \frac{e_1^{fb} v_0}{v_1^3} (1 - \varepsilon(v_1)) \leq 0, \end{aligned}$$

by the first order condition of e_1^{fb} and Lemma A0. Finally,

$$\Delta^o \left(y_1^{fb} \right) > S \left(y_1^{fb} \right) \frac{v_0}{v_1},$$

as $\psi'' > 0$. The result now follows from L'Hopital's rule and Lemma A0 (b).

(d) It follows from the first order condition for e_1^{fb} and the fact that $\psi'' > 0$, which completes the proof.

Proof of Proposition 2. With some abuse of notation define $S(e; v_i) = ev_i - \psi(e)$ for $i = \{0, 1\}$ and S_e by the derivative of this expression with respect to e . Then, substituting (9) with equality (given market clearing) in (10), the optimization is:

$$\max_{e \in \mathbb{R}^+} \left\{ (ev_1 - \psi(e)) : \Delta^o(e; v_1) \leq e_0^{fb} \eta v_0 - \psi \left(e_0^{fb} \right) \right\}$$

This problem requires maximizing a strictly concave function in a convex set subject to an increasing concave constraint. The Lagrangian is

$$L = S(e; v_1) + \lambda \left[e_0^{fb} \eta v_0 - \psi \left(e_0^{fb} \right) - \Delta^o(e; v_1) \right]$$

And the sufficient and necessary Kuhn-Tucker conditions are:

$$(A.6) \quad L_e = S_e - \lambda \Delta_e^o = 0$$

$$(A.7) \quad L_\lambda = \left[e_0^{fb} \eta v_0 - \psi \left(e_0^{fb} \right) - \Delta^o(e, v_1, v_0) \right] \geq 0$$

$$(A.8) \quad \lambda \geq 0$$

$$(A.9) \quad \lambda \left[e_0^{fb} \eta v_0 - \psi \left(e_0^{fb} \right) - \Delta^o(e; v_1) \right] = 0$$

To characterize the solution of this program, note that for some values of v_1 the constraint is met with the first best effort. In particular, define v^s as the value of v_1 that solves,

$$e_0^{fb} \eta v_0 - \psi(e_0^{fb}) = \Delta^o(y_1^{fb}) - (\bar{u}^h - \bar{u}^l).$$

By Lemma A1 (c), the solution exists and it is unique. We have thus two cases:

Case (i) $v_1 \leq v^s$: For these values of v_1 , first best effort e_1^{fb} can be implemented, with $\lambda = 0$. To see this, note that first order condition (A.6) is met since $S_e(e_1^{fb}; v_1) = 0$. Condition (A.7) is also met with strict inequality for all $v_1 < v^s$ by the definition of v^s . Trivially, (A.8) and (A.9) are met. Thus for $v_1 \leq v^s$ the first best can be implemented.

Case (ii) $v_1 \geq v^s$: Obtain effort $e = e_1^m$ by setting the IC constraint at equality:

$$(A.10) \quad \left[e_0^{fb} \eta v_0 - \psi(e_0^{fb}) - \Delta^o(e_1^m; v_1) \right] = 0$$

Equation (A.10) has a unique solution e_1^m for all $v_1 \geq v^s$ with $e_1^m < e_1^{fb}$. To see this notice that by Lemma A1 (d) $\Delta^o(e; v_1)$ is increasing and concave in e . Notice that $e_0^{fb} \eta v_0 - \psi(e_0^{fb}) > \Delta^o(0; v_1)$. Now let $e_0^h = \operatorname{argmax}_e S(e; v_0)$. Clearly, $e_0^h < e_1^{fb}$. Also $S(e_0^h; v_0) > e_0^{fb} \eta v_0 - \psi(e_0^{fb})$, since the high-skill agent is more productive. Finally, notice that $\Delta^o(e_0^h; v_1) > S(e_0^h; v_0)$, thus there is a solution to (A.10) in $[0, e_1^{fb})$. Finally, define λ^m as follows,

$$(A.11) \quad \lambda^m = \frac{v_1 - \psi'(e_1^m)}{v_0 [1 - \psi'(e_1^m \frac{v_0}{v_1}) \frac{1}{v_1}]}$$

e_1^m and λ^m make (A.6) hold. Notice that $\lambda^m > 0$, since both numerator and denominator are positive. To see this, notice that $\psi'(e_1^{fb}) = v_1$, and that $e_1^m < e_1^{fb}$. As a result, both $\psi'(e_1^m) < v_1$ and $\psi'(e_1^m \frac{v_0}{v_1}) < 1$. Then equations (A.6), (A.7), (A.8), and (A.9) hold for λ^m and e_1^m , and thus Proposition 2.b holds. This completes the proof of Proposition 2.

Proof of Corollary 3. If the first best cannot be implemented, then e_1^m is determined by:

$$y_0^{fb} - \psi\left(\frac{y_0^{fb}}{\eta v_0}\right) = e_1^m v_0 - \psi\left(e_1^m \frac{v_0}{v_1}\right) - (\bar{u}^h - \bar{u}^l),$$

and applying the implicit function theorem,

$$\frac{\partial e_1^m}{\partial v_1} = -\frac{e_1^m}{v_1^2} \left(\frac{\psi'\left(e_1^m \frac{v_0}{v_1}\right)}{1 - \frac{1}{v_1} \psi'\left(e_1^m \frac{v_0}{v_1}\right)} \right) < 0,$$

as $e_1^m < e_1^{fb}$. This completes the proof of Corollary 3.

Proof of Lemma 4. In the candidate pooling equilibrium, the utility of the low-skill agent, independently of the value of the opportunity at hand, is $U(v_i) = q + \bar{u}^l, i = 0, 1$ where q is the pooling price. Since low-skill agents make a take-it-or-leave-it offer, q is given by

$$(A.12) \quad q = \pi \max_y \left\{ y - \psi \left(\frac{y}{v_0} \right) \right\} + (1 - \pi) \max_y \left\{ y - \psi \left(\frac{y}{v_1} \right) \right\} - \bar{u}^h.$$

Recall that $U^l(v_i, s^l(y), \mu)$ is the utility of a low-skill agent who offers an opportunity v_i , in exchange for a contingent payment schedule $s^l(y)$, facing high-skill agents with beliefs μ . Then to prove that the candidate pooling equilibrium does not satisfy the intuitive criterium, it is enough to show that there exist output contingent payment schedules $s^l(y)$ such that, first, $U(v_0) > U^l(v_0, s^l(y), 0)$ while, second, $U^l(v_1, s^l(y), 0) > U(v_1)$. We proceed by constructing such a contingent payment schedule.

Recall as well that outputs outside the set $\mathcal{Y} = \left\{ y_1, y_1 \frac{v_0}{v_1} \right\}$ can be ignored. Let $s^l(y_1) = s_1^l$ and $s^l(y_1 \frac{v_0}{v_1}) = \hat{s}_0^l$. We thus have that $U^l(v_1, s^l(y), 0) = s_1^l + \bar{u}^l$ and that $U^l(v_0, s^l(y), 0) = \hat{s}_0^l + \bar{u}^l$.

First, the schedule has to be such that low-skill agents with v_0 prefer the candidate pooling equilibrium to the sharing schedule (even if they can successfully refer the problem and ‘pass themselves’ for holders of a v_1 opportunity):

$$(A.13) \quad q + \bar{u}^l \geq \hat{s}_0^l + \bar{u}^l.$$

\hat{s}_0^l must be such that the high-skill agent taking on the sharing contract prefers not to deviate, equation (7), and the participation constraint of the high-skill agent has to be met (9). Combining equations (A.13), (7), and (9) together with (8), simplifying, and using the definition of the over-referral deviation surplus $\Delta^o(y_1)$ in (12) the value of y_1 in the candidate separating equilibrium must satisfy:

$$(A.14) \quad q + \bar{u}^l \geq \Delta^o(y_1) + \bar{u}^l - \bar{u}^h,$$

where the right hand side of (A.14) is the utility that a low skill-agent with v_0 obtains when he refers his opportunity pretending it was v_1 .

Second, the sharing schedule must be such that the low-skill agents with v_1 prefer it to the utility they could get in the candidate pooling equilibrium, $s_1^l + \bar{u}^l \geq q + \bar{u}^l$, or equivalently

$$(A.15) \quad S(y_1) + \bar{u}^l - \bar{u}^h \geq q + \bar{u}^l,$$

where $S(y_1)$ was defined in (A.3).

To show that the pooling equilibrium does not satisfy the intuitive criterion, it suffices then to show that there always exists a y_1 such that (A.14) and (A.15) hold. The proof then

hinges on the properties of $S(y_1)$ and $\Delta^o(y_1)$, which we investigate next. The reader can turn to Figure 3 for the intuition of this proof.

Given the value of q , expression (A.12), and that $\pi \in (0, 1)$,

$$(A.16) \quad S(0) = \Delta^u(0) = 0 < q + \bar{u}^h \quad \text{and} \quad S\left(y_1^{fb}\right) = \Delta^o\left(y_1^{fb} \frac{v_1}{v_0}\right) > q + \bar{u}^h.$$

Given our assumption on $\psi(\cdot)$, both $S(y_1)$ and $\Delta^o(y_1)$ are strictly concave functions of y_1 . Then it follows from (A.16) that $S(y_1)$ and $\Delta^o(y_1)$ can intersect $q + \bar{u}^h$ only twice. Moreover, define $\Theta(y_1) = S(y_1) - \Delta^o(y_1)$. It follows from the fact that $\psi''' \geq 0$ that the function $\Theta(y_1)$ is strictly concave, and hence the functions $S(y_1)$ and $\Delta^o(y_1)$ can only intersect twice: At $y_1 = 0$ and at $y_1 = \tilde{y} > y_1^{fb}$ such that, by (A.16), $S_{y_1}(\tilde{y}) < 0$ and $\Delta_{y_1}^o(\tilde{y}) > 0$ so that $\Delta^o(y_1)$ is everywhere below $S(y_1)$ for $y_1 \in (0, \tilde{y})$ (see Figure 3). Finally, define

$$\Delta^o(\bar{y}_1) = q + \bar{u}^h \quad \text{where} \quad \bar{y}_1 < y_1^{fb} \quad \text{and} \quad S(\underline{y}_1) = q + \bar{u}^h \quad \text{where} \quad \underline{y}_1 < y_1^{fb} \frac{v_1}{v_0}$$

It is immediate that $\underline{y}_1 < \bar{y}_1$. Then, for all $y \in (\underline{y}_1, \bar{y}_1)$, $S(y) > q + \bar{u}^h > \Delta^o(y)$, so (A.14) and (A.15) are met, and there always exists a contingent contract that breaks the candidate pooling equilibrium. This completes the proof of Lemma 4.

[FIGURE 3 ABOUT HERE]

Proof of Proposition 5. When the first best does not hold the market level of effort, e^m , is the solution to constraint (11), which we can write as

$$C = e^m v_0 - \psi\left(e^m \frac{v_0}{v_1}\right) \quad \text{where} \quad C = y_0^{fb} - \psi\left(\frac{y_0^{fb}}{\eta v_0}\right) + \bar{u}^h - \bar{u}^l,$$

a positive constant. Notice that $\psi\left(e^m \frac{v_0}{v_1}\right) < \psi\left(e^m\right) \frac{v_0}{v_1}$ as $\psi'' > 0$ and $\frac{v_0}{v_1} < 1$. Hence,

$$C > \frac{v_0}{v_1} [e^m v_1 - \psi(e^m)].$$

Next, the difference between the utility of the low-skill agent when he keeps v_1 and the one he obtains when he refers through a sharing contract is,

$$\begin{aligned} \Delta^{u-s}(v_1) &= \max_e \{e \eta v_1 - \psi(e)\} - \left[e^m v_1 - \psi(e^m) - (\bar{u}^h - \bar{u}^l) \right] \\ &> \frac{v_1}{v_0} \left[\frac{v_0}{v_1} \max_e \{e \eta v_1 - \psi(e)\} - C \right] + (\bar{u}^h - \bar{u}^l). \end{aligned}$$

But

$$\lim_{v_1 \rightarrow \infty} \left[\frac{\max_e \{e\eta v_1 - \psi(e)\}}{v_1} \right] = \lim_{v_1 \rightarrow \infty} \tilde{e}\eta = \infty,$$

where $\tilde{e} = \operatorname{argmax} \{e\eta v_1 - \psi(e)\}$, the first equality follows from L'Hopital's rule, and the last from an argument identical to Lemma A0 (b). But hence $\lim_{v_1 \rightarrow \infty} \Delta^{u-s}(v_1) = \infty$, so under-referrals eventually dominates the spot sharing contract. This completes the proof of Proposition 5.

Proofs of Section IV

Let $W[y_1, y_0]$ be the welfare function defined in the space of output allocations resulting from opportunities v_1 and v_0 respectively. It is immediate to prove that $W[\cdot, \cdot]$ is quasiconcave.

As for the indifference curves associated with $W[\cdot, \cdot]$, a straightforward application of the implicit function theorem shows that:

$$\frac{dy_0}{dy_1} = - \left(\frac{1 - \pi}{\pi} \right) \left[\frac{1 - \frac{1}{v_1} \psi' \left(\frac{y_1}{v_1} \right)}{1 - \frac{1}{\eta v_0} \psi' \left(\frac{y_0}{\eta v_0} \right)} \right] < 0 \quad \text{for all } y_0 < y_0^{fb} \quad \text{and} \quad y_1 < y_1^{fb}$$

Notice that as:

$$(A.17) \quad y_1 \rightarrow y_1^{fb} \quad \text{then} \quad \frac{dy_0}{dy_1} \rightarrow 0 \quad \text{and} \quad y_0 \rightarrow y_0^{fb} \quad \text{then} \quad \frac{dy_0}{dy_1} \rightarrow -\infty$$

Clearly, the indifference map has a satiation point in $[y_1^{fb}, y_0^{fb}]$ and bend "backwards" whenever $y_1 > y_1^{fb}$ or $y_0 > y_0^{fb}$. Next, define the $\mathcal{S}^c = \{[y_1, y_0] \text{ such that (18) is met}\}$, as the set of incentive compatible allocations, and define \mathcal{F}^c as the frontier of \mathcal{S}^c , which defines y_0 as a function of y_1 . We show next, that \mathcal{S}^c is a convex set, for which it is enough to prove that \mathcal{F}^c defines y_0 as a concave function of y_1 .

Lemma A2 The frontier of incentive compatible contracts is a strictly concave function with a maximum at $y_1^* < y_1^{fb}$.

Proof: A straightforward application of the implicit function theorem shows that,

$$(A.18) \quad \left[\frac{dy_0}{dy_1} \right]_{\mathcal{F}^c} = \frac{1 - \frac{1}{v_1} \psi' \left(\frac{y_1}{v_1} \right) - \frac{v_0}{v_1} \left[1 - \frac{1}{v_1} \psi' \left(\frac{y_1 v_0}{v_1 v_1} \right) \right]}{\frac{\partial \Omega^u(y_0)}{\partial y_0}},$$

where $\Omega^u(y_0)$ was defined in (A.4). Rearranging the above expression and applying the implicit function theorem again:

$$1 = - \frac{\left[\frac{d^2 y_0}{dy_1^2} \right]_{\mathcal{F}^c} \frac{\partial \Delta^u(y_0)}{\partial y_0} + \frac{1}{v_1^2} \left[\psi'' \left(\frac{y_1}{v_1} \right) - \left(\frac{v_0}{v_1} \right)^2 \psi'' \left(\frac{y_1 v_0}{v_1 v_1} \right) \right]}{\left(\left[\frac{dy_0}{dy_1} \right]_{\mathcal{F}^c} \right)^2 \frac{\partial^2 \Omega^u(y_0)}{\partial y_0^2}} \Rightarrow \left[\frac{d^2 y_0}{dy_1^2} \right]_{\mathcal{F}^c} < 0,$$

as $\psi'''(\cdot) \geq 0$, and it can be immediately proved that $\frac{\partial \Omega^u(y_0)}{\partial y_0} > 0$ and $\frac{\partial^2 \Omega^u(y_0)}{\partial y_0^2} > 0$. The last part of the Lemma follows from (A.18). This completes the proof of Lemma A2.

Proof of Proposition 6. The only interesting situation is when $[y_1^{fb}, y_0^{fb}]$ does not belong to \mathcal{S}^c . By Lemma A2 the frontier is a strictly concave function (the set of incentive compatible allocations \mathcal{S}^c is convex.) Furthermore \mathcal{S}^c is a compact set. Maximization of the strictly quasiconcave welfare function on the feasible set yields a unique maximum. Clearly, the planner places the allocation in the downward sloping side of the frontier. The slope of the frontier of incentive compatible allocations, as given by (A.18), evaluated at either y_1^{fb} or y_0^{fb} is $-\infty < \left[\frac{dy_0}{dy_1} \right]_{\mathcal{F}^c} < 0$. But then by (A.17), it pays to move the allocation towards a strict interior. This shows that when the first best cannot be implemented *both* levels of output are distorted. We characterize next the region where the first best obtains.

The first best obtains whenever (A.5) is met when evaluated at y_1^{fb} and y_0^{fb} . To show where this is indeed the case, start by noticing that for $v_1 = v_0$, $S(y_1^{fb}) = \Delta^o(y_1^{fb})$ and $\Omega^u(y_0^{fb}) = 0$. By Lemma A1 (a), $S(y_1^{fb})$ is strictly convex whereas, by Lemma A1 (b) and (c), $\Delta^o(y_1^{fb}) + \Omega^u(y_0^{fb})$ is a strictly concave function of v_1 so the equation $S(y_1^{fb}) = \Delta^o(y_1^{fb}) + \Omega^u(y_0^{fb})$ has at most another solution. Let v^p be such a solution and assume first that $v^p \geq v^{fb}$, then for all $v^{fb} < v_1 < v^p$ the first best cannot be implemented as for those values of v_1 , $S(y_1^{fb}) < \Delta^o(y_1^{fb}) + \Omega^u(y_0^{fb})$. If $v^p < v^{fb}$ then the first best can be implemented for all those opportunities that satisfy Assumption 4. If on the other hand the unique solution is $v^p = v_0$ then necessarily, $S(y_1^{fb}) > \Delta^o(y_1^{fb}) + \Omega^u(y_0^{fb})$ for all $v_1 > v^{fb}$ and the first best can always be implemented. This completes the proof of Proposition 6.

Proof of Corollary 7. Clearly if the organization can implement the first best or if it cannot support any communication, then both y_1^p and y_0^p are independent of π . If, on the other hand, the organization can only support communication with distortions then (18) is binding. The derivative of y_0^p with respect to y_1^p , as given by equation (A.18), is negative in the efficient side of the frontier of incentive compatible contracts. Taking the derivative of the objective function (17) with respect to y_1^p , to obtain the first order condition, and applying the implicit function theorem to find $\frac{dy_1^p}{d\pi}$ yields the result. This completes the proof of Corollary 7.

REFERENCES

- Akerlof, George** “The Market for Lemons: Quality Uncertainty and the Market Mechanism.” *Quarterly Journal of Economics*, August 1970, 84(3), pp. 488-500.
- Bagwell, Laurie Simon and Bernheim, B. Douglas** “Veblen Effects in a Theory of Conspicuous Consumption.” *American Economic Review*, June 1996, 86(3), pp. 349-373.
- Bagwell, Kyle and Riordan, Michael H.** “High and Declining Prices Signal Product Quality.” *The American Economic Review*, March 1991, 81(1), pp. 224-239.
- Cho, In-Koo and Kreps, David** “Signaling Games and Stable Equilibria.” *Quarterly Journal of Economics*, May 1987, 102(2), pp. 179-221.
- Choi, Jay Pil** “Brand Extension as Informational Leverage.” *The Review of Economic Studies*, October, 1998, 65(4), pp. 655-669.
- Cotterman, James D. ed.** *Compensation Plans for Law Firms*. Newton Square, PA: Altman, Weil, Pensa Inc, 1995.
- Demski, Joel S. and Sappington, David E. M.** “Delegated Expertise.” *Journal of Accounting Research*, Spring 1987, 25(1), pp. 68-89.
- Farrell, Joseph and Scotchmer, Suzanne.** “Partnerships.” *Quarterly Journal of Economics*, May 1988, 103(2), pp. 279-298.
- Garicano, Luis.** “Hierarchies and the Organization of Knowledge in Production.” *Journal of Political Economy*, October 2000, 108(4), pp. 874-904.
- Gilson, Ronald J. and Mnookin, Robert H.** “Sharing Among the Human Capitalists: An Economic Inquiry into the Corporate Law Firm and how Partners Split Profits.” *Stanford Law Review*, 1985, 37(313), pp. 313-392.
- Hayek, Friedrich A.** “The Use of Knowledge in Society.” *American Economic Review*, September 1945, 35(4), pp. 519-30.
- Holmstrom, Bengt.** “Moral Hazard in Teams.” *Bell Journal of Economics*, Autumn 1982, 13(2), pp. 324-340.
- Jullien, Bruno.** “Participation Constraints in Adverse Selection Models.” *Journal of Economic Theory*, July 2000, 93(1), pp. 1-47.
- Kehrer, Kenneth.** “Trust vs. Brokerage: When both departments can handle the job, who gets the business?” *ABA Banking Journal*, October 1998, pp 95-98.
- Kandel, Eugene and Lazear, Edward P.** “Peer Pressure and Partnerships.” *Journal of Political Economy*, August 1992, 100(4), pp. 801-817.
- Laffont, Jean-Jacques and Tirole, Jean.** *A Theory of Incentives in Procurement and Regulation*, Cambridge:MIT Press, 1993.
- Landry, Scot and Nanda, Ashish.** “Family Feud (A): Andersen versus Andersen,” *Harvard Business School Case No. 9-800-064*, 1999.
- Legros, Patrick and Matthews, Steven A. ,** “Efficient and Nearly-Efficient Partnerships,” *The Review of Economic Studies*, July 1993, 60(3), pp. 599-611.
- Levin, Jonathan and Tadelis, Steve.,** “A Theory of Partnerships.” Working Paper, Stanford University, 2002.

- Lewis, Tracy and David E. Sappington**, “Countervailing Incentives in Agency Problems,” *Journal of Economic Theory*, December 1989, 49(2), pp. 249-312.
- Maggi, Giovanni and Andres Rodriguez-Clare**, “On Countervailing Incentives,” *Journal of Economic Theory*, June 1995, 66(1), pp. 238-63.
- Marcus, Ruth**, “Covington Challenge: To Stay on Top,” *The National Law Journal*, May 4, 1981.
- Mudrick, Howard**, “Partner compensation.” in *The CPA Journal Online*, at <http://www.nysscpa.org/cpajournal/old/08656260.htm>, 1997.
- Prendergast, Canice**, “The Provision of Incentives in Firms,” *Journal of Economic Literature*, March 1999, 37(1), pp.7-63.
- Rajan, Raghuram and Zingales, Luigi**. “Power in the Theory of the Firm,” *Quarterly Journal of Economics*, May 1998, 113(2), pp. 387-432.
- Schultz, Christian**. “Polarization and Inefficient Policies.” *The Review of Economic Studies*, April 1996, 63(2), pp. 331-343.
- Spurr, Stephen J.** “Referral Practices Among Lawyers: A Theoretical and Empirical Analysis.” *Law and Social Inquiry*, 1988.
- “The Impact of Advertising and other Factors on Referral Practices, with Special Reference to Lawyers.” *Rand Journal of Economics*, Summer 1990, 21(2), pp. 235-246.
- Taylor, Curtis**. “The Economics of Breakdowns, Checkups and Cures” *Journal of Political Economy* February 1995, 103(1), pp. 53-74.
- Trotter, Michael, H.** Profit and the Practice of Law, Athens and London: The University of Georgia Press, 1997.
- Weingarten, Gene**. “Breaking Up,” *The National Law Journal*, June 1, 1981.
- Wilber, James**. “Partner Compensation Systems - How Firms Distribute Owner Profits,” Altman-Weil, August (2000), downloadable at <http://www.altmanweil.com/about/articles>.
- Wolinsky, Asher**. “Competition in a Market for Informed Experts’ Services.” *Rand Journal of Economics*, Autumn 1993, 24(3), pp.380-398.

* We are grateful to Canice Prendergast for many discussions and insightful suggestions. We also benefited greatly from conversations with Ulf Axelson, Pierre-Andre Chiappori, Wouter Dessen, Doug Diamond, Robert Gertner, Tom Hubbard, Glenn MacDonald, Kevin Murphy, Michael Raith, Rafael Repullo, Matthew Rhodes-Kropf, Antoinette Schoar, Pablo Spiller, Lars Stole, Tim Van Zandt, and David C. Webb. We also thank participants in seminars at the University of Chicago, Stanford Institute of Theoretical Economics (SITE), UC Berkeley, INSEAD, LSE, Columbia University, MIT, UCLA, Duke, Harvard Business School, Rochester, CEMFI, and the Society of Economic Design in Istanbul. We are grateful to the GSB for financial support. Finally, we thank the editors and three anonymous referees.

Footnotes

¹This is the case, for example, in the New York personal injury claims market (Spurr 1988, 1990). We discuss this market in Section V.

²The economics literature (e.g. Farrell and Scotchmer, 1988, or Legros and Matthews, 1993) has used the word partnership to refer to ex-ante income sharing arrangements between agents engaged in team production. Our usage, which allows agents to allocate both opportunities and income, is more closely aligned with the legal term.

³A priori, it is possible that two agents agree to bill their clients together through the same accountant, for example. In our view, this would be, de facto, a partnership.

⁴Demski and Sappington study the problem of inducing an expert to acquire knowledge in situations when the easiest thing for him is to make a blind recommendation rather than go through the trouble of actually figuring out what went wrong. Wolinski studies the development of reputation by these experts in the presence of consumer search. Taylor analyzes the market solutions to the problem posed by the expert's incentive to always recommend treatment to an uninformed consumer.

⁵Farrell and Scotchmer (1988) study partnerships as coalition formation games in which agents divide output equally, and obtain implications for the size and composition of such part-

nerships. Legros and Matthews (1993) study incentives in deterministic partnerships (teams sharing output) and show that, under general conditions, partners choose the efficient actions. Kandel and Lazear (1992) study peer pressure in partnerships. Levin and Tadellis (2003) study the role of revenue sharing contracts in partnerships, and show that the ‘excessive’ concern for quality of a revenue sharing group of agents ensures that, when observing output quality is costly, production choices are more efficient than those made by a profit maximizing firm.

⁶When no vertical elements are present, so that each agent can handle entirely different problems, the incentive problem is not present, as section 3 shows.

⁷We do not explicitly include in the model the diagnosis cost. Instead, its existence is implicit in the fact that only the referring agent knows the value of the opportunity.

⁸ $\psi'(0) = 0$ and $\psi'' > 0$ are required to ensure that the first-best is interior for all v_0 and v_1 in R_+ . $\psi''' \geq 0$ guarantees that the ex-ante contract we introduce below is a concave program, an assumption similar to the one made by Laffont and Tirole (1993) in a different but related problem (see proposition 1.1, page 59). Most commonly used cost functions, including the CES and all polynomial functions with the right derivative signs satisfy all of our assumptions.

⁹It is straightforward to show that for a given choice of v_0 , η , \bar{u}^h , and \bar{u}^l , v^{fb} is unique.

¹⁰This is a standard assumption. An agent with such a function would have a strong incentive to sabotage if only minimally to channel output to himself. See e.g. Holmstrom (1982) and Legros and Matthews (1993).

¹¹The assumption of no renegotiation is reasonable in this context. First, the existing institutions, such as law firms, have an incentive to enforce contracts to facilitate future referrals. Second, agents have an incentive to develop a reputation for not renegotiating referral contracts in the hope of maintaining the credibility of future referral transactions.

¹²An alternative organizational form would have a low-skill agent fully specialized in diag-

nosis and in distribution of opportunities to either another low-skill agent or the high-skill agent. This solves the asymmetric information problems presented here, at the cost of having one of the agents not producing at all. Thus, an upper bound on the distortions we study is the price of an extra worker. Medicine comes closest to this organizational arrangement. Still, ‘gatekeepers’ (the general practitioners) do indeed diagnose and treat the simple cases.

¹³See for example Bagwell and Riordan (1991), Schultz (1996), Bagwell and Berheim (1997), and Choi (1997).

¹⁴For example, consider the case where π is close to 0 and v_1 is large. Then the cost of mismatching high-skill agents with low-value opportunities is negligible. In this case over-referrals clearly Pareto improves on under-referrals.

¹⁵As in Section 4, the low-skill agent makes a take-it-or-leave-it offer to the high-skill agent consisting of an ex-ante contract. This assumption affects only the distribution of the surplus and not the allocation, as the contract is chosen ex ante, and, absent wealth constraints, transfers always exist that lead the best ex-ante contract to be chosen.

¹⁶This contract can also be interpreted as a call option on the time of the high-skill agent to whom the value of the opportunity is truthfully communicated upon exercise. Alternatively, the contract may be interpreted as an ‘employment contract,’ where a high-skill agent “hires” a low-skill agent who receives a fixed wage and is asked to refer all diagnosed problems upstream in exchange for a fixed per-problem fee.

¹⁷Agents may always choose not to enter ex-ante arrangements and supply their opportunities or skill in the spot market. Since the partnership can always implement a spot-market contract which achieves efficient matching, the two possible outcomes of the spot-market (under-referrals and spot market contract) are subsumed in these three choices. The existence of the spot market option will however determine the reservation utility of the agents and thus affect the distribution of surplus. Since agents can freely transfer surplus, we can ignore these

distributional considerations.

¹⁸In particular, in the spot market competition for good opportunities holds the utility the high-skill agent obtains to her outside utility, $s_1^h - \psi(\frac{y_1}{v_1}) = \bar{u}^h$.

¹⁹We thank one referee for suggesting an example in this direction.

²⁰Court opinions are also an excellent source of documentation on the nature of these contracts, given the disputes they often give rise to. A recent example is *Florida Bar vs. Kevin Carson*, (91,550 Florida (1998)), in which the dispute is over an agreement where ‘the two lawyers entered into a mutually advantageous referral relationship, whereby [one of the lawyers] Mr. Vasilaros told [the other lawyer] Mr. Carson that he would pay Mr. Carson 25% of the attorney’s fee for personal injury cases that he obtained as a result of referrals made by Mr. Carson to him.’ Another high-profile dispute is documented in ‘Against O’Quinn: Ex-partner sues lawyer for \$250 million’ (*Houston Chronicle* Feb 15, 1999). Among the issues at the heart of the dispute, the plaintiff claims that the defendant ‘took cases referred by [the defendant’s] associates but did not pay him from these cases as he was supposed to.’

²¹Spurr had access, through an order of the Federal District Court of New York, to the file retainer and closing statements. These files contain the fees to be earned by each lawyer, the gross recoveries and the share of the recovery assigned to the (actual) litigation lawyer; whether the lawsuit was filed, settled, went to verdict; the verdict, and the names of the lawyers.

²²As a referee has pointed out, while the optimal contract is a step function, these contracts are simpler, linear sharing rules. A previous version of this paper, available from the authors, showed that the economics of the problem remain unaltered when cognitive limitations require the use of linear rules.

²³On accounting firm arrangements, see Mudrick (1997). On banks, where the issue is to ensure that high net-worth customers are referred by brokerage officers to trust officers, see

Kehrer (1998). Personal communications from several partners of economics consulting firms have confirmed the details of these arrangements are entirely along the lines of similar ones in law firms. Finally, on law firm compensation schemes, see Cotterman (1995), the Survey of Law Firm Economics by Altman, Weil, and Pensa (2000) and *The Commercial Lawyer*, (June/July 2000 issue).

²⁴Not all firms rely on performance measures to compensate partners. A minority (Cotterman, 1995:29) rely only on seniority to allocate rewards (the so-called lock-step method). Of the 20 top law firms in the US, only four use the pure lock-step method (*The Commercial Lawyer*, June/July 2000 issue): Cleary, Gottlieb, Steen & Hamilton, Cravath, Swaine & Moore, Davis Polk & Wardwell, and Wachtell, Lipton, Rosen & Katz. The remaining firms used performance-related methods ranging from the mild modified lock-step method of Latham & Watkins (85% lockstep and the rest is performance) to pure merit-based methods. Incentives to share business and provide effort in lock-step law firms are provided by the law firm's "culture," according to personal interviews with a Cleary partner.

²⁵Originated by Reginald H. Smith, managing partner of Hale & Dorr of Boston in the 1940s.

²⁶For the details of the case, see Landry and Nanda (1999). We thank Kathy Spier for pointing out this case and sharing her analysis with us.

²⁷As AC CEO George Shaheen put it: 'it does not make sense for AC to continue to send money, especially the amount of money we send to AA, and at the same time compete more and more in the marketplace' (Nanda, 1999:9)

²⁸"Making a Network of Lawyers" by Jonathan D. Glater in *The New York Times*, June 8, 2001. According to the same article, the consulting firm Altman Weil Inc. has identified 300 such national and international law-firm networks. Typical examples are a client who needs advice on a merger with a Brazilian company, or one who needs to register a trademark in different countries.

²⁹This case can be studied by specifying the reservation utility of agents as opportunity dependent, $\bar{u}^i = \delta\theta_i ev$, where δ is the productivity loss that results from losing the specific complementary assets of the organization. Situations in which the reservation utility of an agent is type-dependent have been investigated by Lewis and Sappington (1989), and have been characterized more generally by Maggi and Rodriguez-Clare (1995) and Jullien (2000). Firms may also choose to develop specific assets that tie employees to the firm (Rajan and Zingales, 1998) to improve referral incentives by reducing the value of defecting.

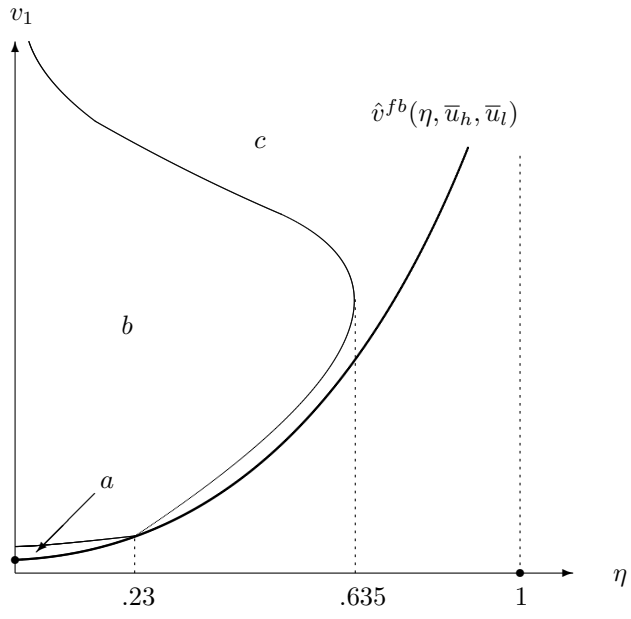


FIGURE 1

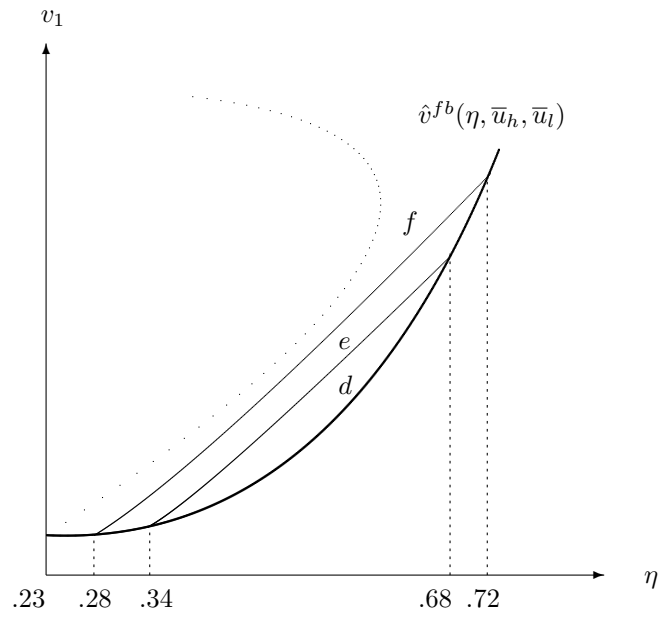


FIGURE 2

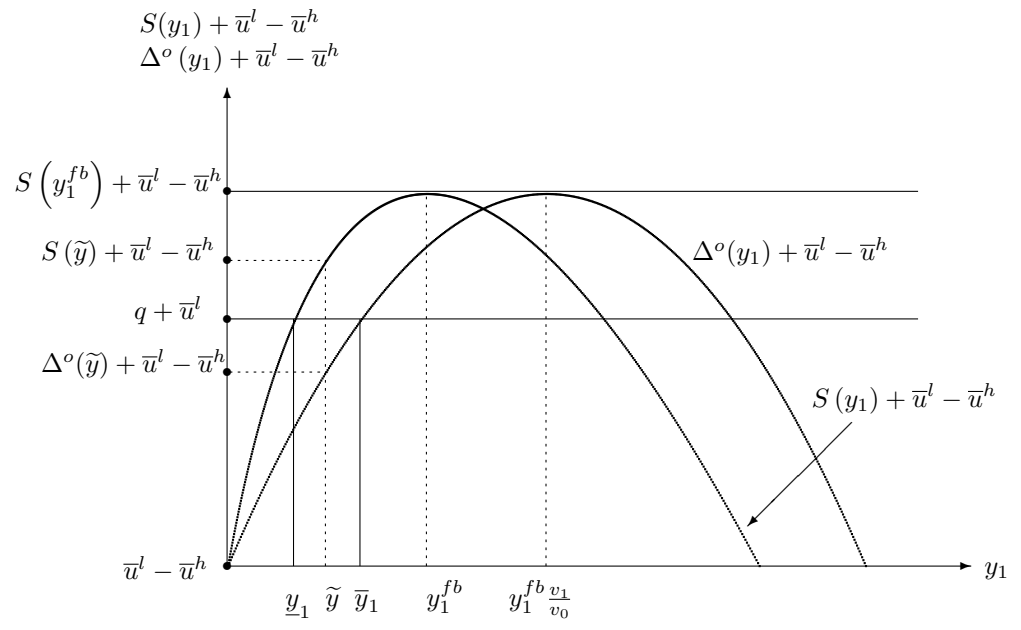


FIGURE 3

Figure 1. Region a denotes the set of economies for which the spot market implements the first best. Region b denotes the set of economies for which the spot contract supports efficient matching but with distortions. Region c is the set of economies for which the spot market cannot support any referrals.

Figure 2. Region d denotes the set of economies for which referrals do not take place. Region e denotes the set of economies where partnerships arise to support efficient matching but with distortions. Region f denotes the set of economies for which the partnership implements the first best. The dotted line denotes the frontier in the signaling case.

Figure 3. Single crossing property. $q + \bar{u}^l$ is the utility of the low-skill agents in the candidate pooling equilibrium and $S(y_1) + \bar{u}^l - \bar{u}^h$ is the utility of the low-skill agent whose draw is v_1 and signals his type with a separating contract that yields output y_1 . Finally, $\Delta^o(y_1) + \bar{u}^l - \bar{u}^h$ is the utility the low-skill agent whose draw is v_0 would obtain if he were to offer the same contract as the the low-skill agent with v_1 as a function of y_1 . Any separating contract implementing output $\tilde{y} \in [y_1, \bar{y}_1]$ breaks the pooling equilibrium.