Giorgio ZANARONE*

Fiat without Authority in Firms (and Markets)

This version, January 2010

Abstract

It is often asserted that administrative control and fiat are key features of firms. Nevertheless, many inter-firm relations, from supply networks to franchising, seem equally hierarchical. To reconcile these perspectives, I develop a model where contractors are subject to the same formal obligations as employees, but, owning local assets, they care more about profits, and they are less averse to effort. I show that fiat, in the form of relational contracts where the agent obeys the principal in equilibrium, occurs both when the agent is an employee and when she is a contractor. However, the principal can give more orders to an employee (contractor) when decisions sufficiently above (close to) the level that maximizes local profits are optimal—for instance, because there are strong positive externalities from the agent to the principal, or the principal can monitor the agent cheaply. This also implies that, consistent with numerous empirical works, integration should be more frequently observed when externalities are large and monitoring costs are small.

Keywords: Fiat, Integration, Relational Contracts.

JEL codes: D23; L14; L22

* Colegio Universitario de Estudios Financieros; E-mail: gzanarone@cunef.edu. An earlier version of this paper circulated under the title “Fiat without Authority under Vertical Integration”. I am grateful to Benito Arruñada and Robert Gibbons for continuous advice and encouragements, and to Gani Aldashev, Nicholas Argyres, Marc Bellemare, Antonio Cabrales, Marco Celentani, Silke Forbes-Januszewski, Luis Garicano, Ricard Gil, Anna Grandori, Eduardo Melero, Jean-Philippe Platteau, Emmanuel Raynaud and Dean Williamson for their comments and suggestions. This study received financial support from the European Commission, through the integrated project CIT3-513420, and by the Spanish Ministry of Science and Education, through grant ECO2008-01116.
1. Introduction

In the last two decades a strong body of empirical evidence has emerged, suggesting that agency conflicts conduce to integration. For instance, franchisors own retail outlets that generate spillovers on the common brand (Yeap (2006)); motor carriers own trucks whose poor maintenance would harm the carrier’s service and reputation (Nickerson and Silverman (2003)); airline companies own regional carriers that serve routes between bad weather airports, where frequent flight rescheduling preserves the network’s reputation but causes short-term losses to the regional (Forbes and Lederman (2009)); and local governments provide in-house those services to whose quality citizens are especially sensitive (Levin and Tadelis (2009)).

The relations described in these and other studies are all plagued by a similar type of agency problem: principals (franchisors, motor carriers, major airline carriers, governments) rely on agents (outlet managers, truck drivers, managers of regional carriers, service providers) to perform onerous tasks (serving customers, driving efficiently, rescheduling flights, providing public services of high quality), which benefit the principal, but generate both benefits and costs for the agent. Importantly, agency is present independent of whether the agent is an employee or an independent contractor: for instance, restaurant chains need local managers to serve customers according to the brand’s standards.

---

1 See Lafontaine and Slade (1997, 2007) for detailed reviews of the empirical literature on vertical integration.
irrespective of whether these are franchisees or employees. Moreover, it is often the case that employees and contractors receive similar incentive contracts (Nickerson and Silverman (2003)), or incentive contracts that do not depend on the critical tasks at issue (Forbes and Lederman (2009)). This raises an obvious question: how can integration induce opportunistic agents to perform, if it does not improve their incentives?

As a possible explanation, economists have suggested long ago that, within firms, employers use fiat to resolve disputes and direct employees (Coase (1937), Simon (1951)). Particularly, early transaction cost theory has emphasized that firms can govern employees through organizational instruments, such as pervasive inspections, administrative controls and flexible contract law, which are hard to replicate in the market (Williamson (1971, 1979, 1991), Masten (1988)). More recently, Van den Steen (2009) has argued that, even when bringing transactions inside a firm does not expand the available set of formal organizational instruments, employees are less tempted to disobey orders than contractors because they do not own assets, so they have weaker incentives and are more scared to be fired. While these theories are different in focus and spirit, they all imply a monotonic relation between integration and fiat: all else equal, when the principal needs to control his agent—for instance, because the agent’s actions critically affect his profits—he should hire her as an employee, rather than as an independent contractor, because employees obey orders more promptly and without frictions.

---

2 This implies that, in these environments, one cannot assume, as in Hart and Holmstrom (2009) and in Baker et al. (2008), that integration concentrates full control over decisions within one party, eliminating the distinction between a principal who orders and an agent who implements.

3 Klein and Murphy (1997) also point out that employees lose more than contractors in case of termination.
A problem with this view is that relations between independent firms—both formal and informal—can also be strongly hierarchical (Alchian and Demsetz (1972)). For instance, courts routinely enforce the authoritative decisions of franchisors as if they were the franchisees’ employers (Hadfield (1990)). Also, car manufacturers have pervasive decision and monitoring rights on franchised dealers (Arruñada et al. (2001), Zanarone (2009)) and, even when they do not have such formal decision rights, they routinely dictate dealers’ behavior, with virtually no recourse to market-like practices such as bargaining or arbitration (Zanarone (2010)). As Van den Steen (2009) noted, this opens the question of what generates fiat between firms, and in what sense fiat within and between firms differ.

This paper acknowledges that fiat exists both in firms and in markets, and clarifies what conditions make firms or markets the optimal organizational form to achieve it. To do so, it develops a simple agency model where the principal relies on the agent to implement a non-contractible decision. Efficient decisions optimize between three components: the principal’s private benefit, which calls for high-level decisions; the agent’s private cost of effort, which calls for low-level decisions; and the value of an asset used in the production process, which calls for intermediate decisions—formally, the asset’s value is a concave function of the decision’s level. This environment is suitable to model franchise networks, supply chains, transportation networks and, in general, relationships where the values of central and local assets are imperfectly aligned, due to externalities.

The model shows that fiat arises in equilibrium both under integration, where the principal owns the asset and the agent is an employee, and under separation, where the

---

4 This is not always the case, as the law may award different rights to franchisees and employees. For instance, Brickley et al. (1991) find that franchisors tend to own more outlets in American states where the law protects franchisees from termination.
agent owns the asset and is a private contractor. Specifically, fiat takes the form of relational contracts where the agent routinely implements the principal’s decisions, and the principal pays her an implicit bonus covering her opportunity cost of performance.

However, the size of the bonus necessary to make the agent obey—and, consequently, the extent to which the principal can credibly order her what to do—depends in an intriguing way on whether the agent is an employee or a contractor. The argument is simple: a disobedient contractor has incentives to choose higher-level decisions than a disobedient employee, because he does not only care about saving effort, but also about raising the value of the asset she owns. Hence, a contractor is less (more) tempted to disobey—and, therefore, requires a smaller (greater) bonus to obey—decisions close to (far from) the one that maximizes the asset’s value. This implies that, when lower-level decisions are optimal or are the only ones feasible—for instance, because decisions generate few private benefits for the principal, or monitoring the agent is costly—the principal can more successfully give orders to an employee than to a contractor, and vice versa.

This result has two important implications. First, the relation between integration and the principal’s power of fiat is non-monotonic: when the environment favors low-level decisions, the principal can exert greater control over an agent by hiring her as a contractor than as an employee. Second, and consistent with the evidence cited at the beginning of the paper, changes in the environment that move forward the region of decisions enforceable by fiat—such an increase in the externalities from the agent’s actions on the principal, or a decrease in the principal’s cost of monitoring the agent, favor integration over separation.

The work most closely related to this paper is Baker et al. (2002), who first modeled how integration and separation affect the feasibility of relational contracts. This paper is
complementary to theirs, as it studies a different environment and generates different predictions. Regarding the environment, Baker *et al.* (2002) apply relational contracts to (a revised version of) the holdup problem studied by Klein *et al.* (1978), Williamson (1979), and Grossman and Hart (1986). They focus on a setting where the agent’s actions are hidden to the principal, and study how integration and separation affect the parties’ temptation to renegotiate the implicitly agreed bonus ex post, thus modifying the agent’s ex ante incentives. Conversely, this paper applies relational contracts to the complementary problem of fiat (a principal’s ability to tell the agent what to do) studied by Simon (1951), Williamson (1971, 1991) and Van den Steen (2009), and shows that it may be more easily solved within or between firms, depending on the environment. Regarding the predictions, the analysis in Baker *et al.* (2002) implies that integration should be observed when the agent’s task is not too important and, therefore, the promised bonus is small. The reason is that integration minimizes the agent’s bargaining power and, therefore, her temptation to renegotiate the bonus upwards. Conversely, in this paper, integration should be observed when the agent’s task is important. The reason is that integration makes the agent indifferent towards the value of the asset she uses and, therefore, reduces the bonus necessary to make her implement decisions far from those that maximize the asset’s value.

The rest of this paper is organized as follows. Section 2 introduces the model’s definitions and assumptions. Section 3 discusses the choice between integration and separation in a spot environment. Section 4 studies the same choice in a relational

---

5 Halonen (2002) also models a repeated holdup problem, and shows that joint ownership may be optimal in the repeated game, even though it is suboptimal in the static one.
environment, and presents the main result of the paper. Section 5 discusses additional results and implications for the theory of the firm. Section 6 concludes.

2. The environment

Consider a principal $P$, who relies on agent $A$ to implement a decision $d$. Implementing the decision generates a private (that is, non-contractible) profit $B(d)$ for the principal, and a private cost $C(d)$ for the agent. I assume that, when $P$ wants $A$ to implement a decision different from the one $A$ would spontaneously take—in a sense that will be clarified in section 3—he must communicate the desired decision to $A$. This may be due to the fact that $P$ privately observes the $B(d)$ function, or that he has private communication skills acquired in an un-modeled stage. As we shall see in section 4, this permits to interpret equilibria where the agent implements decisions different from the ones that are privately optimal for her as hierarchical relations based on fiat.

To perform her task, the agent uses an asset, whose residual value $A(d) = V(d) - O(d)$ is non-contractible, and is affected by $d$. The asset can be either owned by the agent (separation) or by the principal (integration), in which case $A$ is $P$’s employee. I assume ownership of the asset conveys the right to appropriate its residual value (Holmstrom and Milgrom (1991, 1994), Holmstrom (1999), Gibbons (2005)), and that no contractible measures of $A$’s performance are available. Hence, $A$’s formal
incentives to perform are completely determined by the choice between integration and separation.\footnote{In franchising, there is empirical evidence that firms tend to appropriate most of the profits generated by the assets they own (Krueger (1991), Maness (1996)).}

I assume that, while P can observe whether A has implemented the decision \(d\), courts cannot, so \(d\) cannot be made the object of a formal contract. This implies that, unlike in Masten (1988) and Williamson (1991), P is equally (un)able to enforce formal contracts based on A’s performance under integration and separation. Finally, I assume that \(B(d)\) and \(V(d)\) are increasing in \(d\) and concave, that \(O(d)\) and \(C(d)\) are increasing in \(d\) and convex, that \(B(0) = A(0) = C(0)\), and that \(O'(0) = C'(0) = 0\).

As an example of the environment described above, consider chain restaurants. P could be MacDonald’s, and A the manager of a representative restaurant. The decision \(d\) could represent how strict MacDonald’s design and cleanliness standards are, \(B(d)\) would be the value of McDonald’s brand, \(V(d)\) would be the restaurant’s long-term revenues from serving customers according to the standards, \(O(d)\) would be the restaurant’s forgone profit from following the standards, instead of offering a customized service, or a low quality service, and \(C(d)\) would be the local manager’s stress for managing a restaurant by the standards. Finally, integration and separation would represent the cases where the manager is a MacDonald’s employee and franchisee, respectively.
3. Spot governance

In this section, I assume P and A meet only once, with no opportunities to trade in the future. If $d$ were contractible, P and A would agree on an upfront payment A would receive for implementing the efficient decision $d^{FB} > 0$, given by the first order condition

$$B_d(d) + A_d(d) = C_d(d)$$

(1)

However, since $d$ is non-contractible, P and A cannot, in general, achieve the first best in a spot environment. To achieve the second best, they must choose, ex ante, between assigning ownership of the asset to P (integration) or A (separation).

3.2. Spot integration

Under spot integration, A’s preferred decision maximizes $-C(d)$, which has a corner solution at $d_i^{SP} = 0$. Since $d$ is non-contractible, no matter what decision P asks her to implement, A always disobeys and chooses $d_i^{SP}$. Hence P does not even bother dictating a decision to A, and no surplus is generated.

3.3. Spot separation

Under spot separation, A’s preferred decision $d_s^{SP}$ maximizes $A(d) - C(d)$. Hence, $d_s^{SP}$ satisfies the necessary and sufficient first order condition

$$A_d(d) = C_d(d)$$

(2)
implying that $d^{EB} > d^{SP} > 0$. Since $d$ is non-contractible, no matter what decision $P$ asks her to implement, $A$ always disobeys and chooses $d^{SP}$. Anticipating this, $P$ does not even bother dictating a decision to $A$, as under integration. However, since $d^{SP} > 0$, a positive joint surplus is generated this time, which is given by

\[ JS_{SP} = JS(d^{SP}) = B(d^{SP}) + A(d^{SP}) - C(d^{SP}). \]

Since $JS_{SP} > 0 = JS_{I}^{SP}$, it immediately follows that

**Proposition 1:** In a spot environment, separation is optimal.

This is obvious: if $A$ has incentives to do more, in equilibrium, under separation than under integration, but less than the first best under both governance structures, separation must yield greater surplus than integration and, therefore, be preferred.

4. **Relational governance**

In this section I assume $P$ interacts with $A$ repeatedly, so he can use relational contracts to make $A$ implement decisions she would disobey in an arm’s-length relationship. Without loss of generality, I focus on stationary contracts where $A$ obeys $P$’s decision in every period, and $P$ pays her an end-of-period implicit bonus.\(^7\) Since $P$ tells $A$ what to do in equilibrium and $A$ obeys, relational contracts can be interpreted as hierarchical relations based on fiat. Note that, even though $A$’s incentives to obey come from the prospect of an implicit bonus, $P$ does not need, at every period, to communicate the bonus to $A$ before $A$ performs—i.e., he does not need to frame the relationship as an exchange. Instead, $P$ may

\(^7\) See Bull (1987), MacLeod and Malcomson (1989), Levin (2003), and Baker et al. (2002, 2009), for related models of relational contracts.
simply tell A at the beginning of the relationship that, if she does what P tells her to do, she will receive a bonus at the end of every period, or convince her that this is the case by repeatedly paying the bonus. Then, A will routinely obey P’s orders in equilibrium without asking about compensation, as literally implied by the notion of fiat.

Since \( d \) is non-contractible and the bonus is based on A’s compliance with \( d \), the relational contracts described above cannot be enforced by courts. Hence, they must be self-enforcing—that is, each party’s present gains from reneging must be smaller than her quasi-rents from future trade. As we will see in the next section, when the parties are moderately patient, the set of self-enforcing relational contracts—and, consequently, the principal’s ability to tell the agent what to do—differs under integration and separation, and either governance structure may be optimal, depending on the environment.

### 4.2. Relational integration

Under integration, relational contracts work as follows: at stage 1 of any period \( t \), P makes a fixed monetary transfer \( w_I \) to A, and then tells A to implement decision \( d \), such that \( d_{3P}^S < d \leq d_{PB}^P \). At the end of stage 2, P verifies that A has implemented \( d \) as told and, in case of compliance, pays her a bonus \( \beta_I \). If P fails to offer \( w_I \) at stage 1, or A fails to accept it, P immediately sells the asset to A at price \( T \), and both parties revert to the optimal spot governance structure—separation—from the current period \( t \) and thereafter. If A fails to implement \( d \) at stage 2, or P fails to pay the bonus \( \beta_I \), P sells the asset to A at price \( T \) at the end of period \( t \), and both parties revert to spot separation from period \( t+1 \) and thereafter.
Let \( r \) be \( P \)'s and \( A \)'s common interest rate. Then, the relational contract is self-enforcing if, and only if

\[
\frac{1+r}{r} [B(d) + A(d) - w_I - \beta_I] \geq \frac{1+r}{r} B(d_S^{SP}) + T
\]

\[
\frac{1+r}{r} [w_I + \beta_I - C(d)] \geq \frac{1+r}{r} [A(d_S^{SP}) - C(d_S^{SP})] - T
\]  

\[
-\beta_I + \frac{1}{r} [B(d) + A(d) - w_I - \beta_I] \geq \frac{1}{r} B(d_S^{SP}) + T
\]

\[
-C(d) + \beta_I + \frac{1}{r} [w_I + \beta_I - C(d)] \geq \frac{1}{r} [A(d_S^{SP}) - C(d_S^{SP})] - T
\]

Conditions (3) and (4) are \( P \)'s and \( A \)'s participation constraints, whereas conditions (5) and (6) are \( P \)'s and \( A \)'s D's dynamic incentive constraints, respectively. By setting \( w_I \) so that (4) binds, (3), (5) and (6) boil down, respectively, to

\[
JS(d) \geq JS(d_S^{SP})
\]

\[
JS(d) \geq JS(d_S^{SP}) + r\beta_I + \frac{r^2}{1+r} T
\]

\[
\beta_I \geq C(d) - \frac{r}{1+r} T
\]

If (5') holds, (3') must hold as well, so constraints (5') and (6') suffice for self-enforcement. By setting \( \beta_I \) so that (6') binds, (5') further boils down to

\[
JS(d) \geq JS(d_S^{SP}) + rC(d)
\]

Thus, under relational integration, \( P \) tells \( A \) to implement decision \( d_i^{RE} \), which maximizes \( JS(d) \), subject to (7). The resulting joint surplus is given by \( JS_i^{RE} = JS(d_i^{RE}) \).
4.3. Relational separation

Under separation, relational contracts work as follows: at stage 1 of any period $t$, $P$ makes a fixed monetary transfer $w_S$ to $A$, and then tells $A$ to implement decision $d$, such that $d_S^{SP} < d \leq d_{FB}$. At the end of stage 2, $P$ verifies that $A$ has implemented $d$ as told and, in case of compliance, pays her a bonus $\beta_S$. If $P$ fails to offer $w_S$ at stage 1, or $A$ fails to accept it, both parties immediately revert to spot separation from the current period $t$ and thereafter. If $A$ fails to implement $d$ at stage 2, or $P$ fails to pay the bonus $\beta_S$, both parties revert to spot separation from period $t+1$ and thereafter. This relational contract is self-enforcing if, and only if

$$\frac{1+r}{r} [B(d) - w_S - \beta_S] \geq \frac{1+r}{r} B(d_S^{SP})$$ (8)

$$\frac{1+r}{r} [A(d) + w_S + \beta_S - C(d)] \geq \frac{1+r}{r} [A(d_S^{SP}) - C(d_S^{SP})]$$ (9)

$$-\beta_S + \frac{1}{r} [B(d) - w_S - \beta_S] \geq \frac{1}{r} B(d_S^{SP})$$ (10)

$$A(d) - C(d) + \beta_i + \frac{1}{r} [A(d) + w_S + \beta_S - C(d)] \geq A(d_S^{SP}) - C(d_S^{SP}) + \frac{1}{r} [A(d_S^{SP}) - C(d_S^{SP})]$$ (11)

By setting $w_S$ so that (9) binds, (8), (10) and (11) boil down, respectively, to

$$JS(d) \geq JS(d_S^{SP})$$ (8’)

$$JS(d) \geq JS(d_S^{SP}) + r\beta_S$$ (10’)

$$\beta_S \geq A(d_S^{SP}) - A(d) + C(d) - C(d_S^{SP})$$ (11’)

13
If (10') holds, (8') must hold as well, so constraints (10') and (11') suffice for self-enforcement. By setting $\beta_S$ so that (10') binds, (11') further boils down to

$$JS(d) \geq JS(d^{SP}_S) + r[A(d^{SP}_S) - A(d) + C(d) - C(d^{SP}_S)]$$

(12)

Thus, under relational separation, P will tell A to implement the decision $d^{RE}_S$ that maximizes the joint surplus $JS(d)$, subject to (12). The resulting joint surplus is given by $JS^{RE}_S = JS(d^{RE}_S)$.

4.4. Comparison: fiat within and between firms

While fiat may occur in equilibrium under both integration and separation, constraints (7) and (12) imply that the extent to which fiat is feasible differs across governance structures. Define the principal’s power of fiat under a given governance structure as the highest-level decision beyond $d^{SP}_S$ he can enforce under that governance structure. Then, the link between governance structure and the principal’s power of fiat is clarified by the following Proposition 2: There is a critical decision $d^* > d^{SP}_S$ such that P’s power of fiat is maximized under separation in the region where $d < d^*$, and under integration in the region where $d > d^*$.

Proof: From (7) and (12) it follows that P’s temptation to renege on the bonus necessary to have decision $d$ implemented is minimized under integration if, and only if

---

8 This interpretation of the equilibrium level of $d$ as a measure of the amount of fiat exerted by P may be clearer if one assumes that, instead of one decision, A must implement a continuum of decisions indexed by $d$. Hence, a relational equilibrium where $d$ is large implies that P orders more decisions to A.
If (13) holds (does not hold), P can enforce a higher-level decision—that is, he has greater power of fiat over A—under integration (separation). The left-hand side of (13) is the function $A(d)$, and the right-hand side is the function $f(d) = A(d) - C(d)$, evaluated at its maximum $d^*_{SP}$. Both $A(d)$ and $f(d)$ are concave and have an interior maximum, and $A(d) > f(d)$ for any $d > 0$. Hence, there must be a unique decision $d^* > d^*_{SP}$ such that $A(d) > f(d^*_{SP})$ for $d < d^*$, $A(d) = f(d^*_{SP})$ for $d = d^*$, and $A(d) < f(d^*_{SP})$ for $d > d^*$. QED.

The intuition behind Proposition 2 is simple. Under integration, A only cares about saving effort, because she is an employee with no stake in the business. Hence, in order to make A obey decision $d$, the principal must fully compensate her cost of effort $C(d)$. Conversely, under separation, A also cares about raising profits, as she is residual claimant of the asset’s value. Profit maximization pushes A to spontaneously perform up to $d_{SP}^{*}$, which implies that, with respect to integration, the portion of A’s cost of effort that P must compensate to make her obey decision $d$ decreases by $C(d^*_{SP})$. However, under separation, P must also compensate A’s profit reduction from implementing $d$ instead of her preferred decision $d^*_{SP}$, which is given by $A(d^*_{SP}) - A(d)$. When P’s decision is demanding ($d > d^*$), this profit-related increase in the incentive-compatible bonus offsets the effort-related reduction $A(d^*_{SP}) - C(d^*_{SP})$, making P’s temptation not to pay greater under separation than under integration.

This result complements a classic tenet of transaction cost economics, according to which agents’ opportunism can be internalized by bringing transactions inside the firm,
where the employer directs employees by fiat (Coase (1937), Williamson (1970, 1979, 1991), Klein et al. (1978)). Indeed, when the environment is such that high-level decisions are optimal (that is, \( d > d^* \)), the tenet holds even if integration does not change the set of formal organizational instruments and the body of contract law available to the principal. However, when lower-level decisions are optimal, the principal can better direct the agent by hiring her as a contractor, rather than as an employee.

5. Implications for the theory of the firm

This section explores some implications of Proposition 2 for the theory of the firm, regarding the link between integration, externalities, future interactions and monitoring costs; the concepts of markets, hybrids and hierarchies; the choice between public and private provision of social services; and the empirical link between organizational form and fiat.

5.2. Externalities

Redefine the value of the upstream unit as \( B(d, b) \), where \( B_b > 0, B_{db} > 0 \) and \( B_{bd} > 0 \). The relation between governance structure and the level of externality from the agent to the principal is then given by the following

**Proposition 3**: For any interest rate, there is a critical externality \( b^*(r) \) such that separation is (weakly) optimal for \( b < b^*(r) \), and integration is (weakly) optimal for \( b > b^*(r) \).
Proof: Let $d_i^{RE}(r, b)$ and $d_S^{RE}(r, b)$ be the decisions that make constraints (7) and (12) binding, respectively. It follows from (7) that $d_i^{RE}(r, b)$ increases in $b$, so we can define $b^*(r)$ as the smallest $b$ such that $d_i^{RE}(r, b) \geq d^*$. Because of Proposition 2, this implies that $d_i^{RE}(r, b) > d_S^{RE}(r, b)$ for $b > b^*(r)$, and $d_S^{RE}(r, b) > d_i^{RE}(r, b)$ for $b < b^*(r)$. Hence, it must be that $JS_i^{RE} \geq JS_S^{RE}$ for $b > b^*(r)$, with the inequality holding strictly for $d_S^{RE}(r, b) < d_i^{FB}$, and $JS_S^{RE} \geq JS_i^{RE}$ for $b < b^*(r)$, with the inequality holding strictly for $d_i^{RE}(r, b) < d_S^{FB}$. QED.

Proposition 3 provides a novel explanation for why externalities lead to integration. As shown in Table 1, this has been observed in a variety of industries: for instance, restaurant chains prefer to own upscale restaurants, dine-in restaurants, and restaurants with in-house food production, where the quality of customer service and the restaurant’s cleanliness and comfort are more critical to the chain’s reputation (Yeap (2006)); motor carriers prefer to own trucks in less-than-truckload trucking, where pick-ups and deliveries are closely interconnected and, therefore, poor maintenance and driving effort cause delays that harm the carrier’s reputation (Nickerson and Silverman (2003)); and major airline companies prefer to own regional carriers that serve routes between bad weather airports, where more flights must be rescheduled due to delays and cancellations—typically, by postponing a regional connection flight to let the major’s flight leave on time—and failure of the regional carriers to do so harms the reputation of the whole major-cum-regional network (Forbes and Lederman (2009)).

---

*See Lafontaine and Slade (1997, 2007) for detailed reviews of the empirical literature on vertical integration.*
These works emphasize that a principal-agent relationship exists both under integration and separation, so integration does not centralize decision control within the principal, as hypothesized by Hart and Holmstrom (2009) and Baker et al. (2008). Moreover, these empirical works also emphasize that integration does not expand the principal’s contractual authority over the agent, as hypothesized by Masten (1988) and Williamson (1991). Nonetheless, as the agent’s tasks become more critical for the principal, integration is observed. The explanation provided by this paper is that, as externalities increase, higher-level decisions become optimal, and these are easier to enforce by fiat under integration, because they are too far from the decisions the agent would be tempted to choose as a private contractor.

An alternative explanation for the observed link between externalities and integration, which does not involve fiat or relational contracts, has been proposed by multi-tasking models, where muting the agent’s incentives by shifting assets away from her avoids imbalances in her allocation of effort across tasks. An advantage of the theory presented here is that it does not require certain assumptions of multi-tasking models—such as the agent’s risk-aversion (Holmstrom and Milgrom (1994), Bai and Tao (2000)), or her willingness to work up to a substantial amount without formal incentives (Holmstrom and Milgrom (1991))—in order for the integration to be optimal. These assumptions may not always apply in industries where spillovers have been observed to lead to integration. For instance, previous works have noted that managers in risky retail stores receive greater, not
smaller portions of outlet profits (Norton (1988), Martin (1988), Lafontaine (1992)), which
is contrary to what one would expect if these managers were risk-averse.

5.3. Monitoring costs

Suppose that, in order to observe A’s behavior—and, thus, to effectively decide
whether to pay her the implicit bonus or not—P must incur a cost. This may include
resources spent by P, ex ante, to set up adequate administrative controls and reporting
systems, and, ex post, to collect information on A’s performance. I assume the monitoring
cost $M(d, m)$ is increasing in $d$ and $m$, with $M(0, m) = 0$, $M_d(0, m) = 0$, $M_{dm} > 0$ and
$M_{ma} > 0$, and that $B(d)$, $A(d)$ and $C(d)$ are separable functions of $d$ and of an
unobservable random term realized after $d$ is chosen, so P cannot infer A’s behavior from
the realized outcomes. To keep matters simple, I also assume the random term has 0 mean,
and that both P and A are risk-neutral, so all the equations in section 4 are unaffected,
extcept that the (expected) joint surplus is now given by $B(d) + A(d) - C(d) - M(d, m)$.
Then, monitoring costs affect the choice between integration and separation, as follows:

**Proposition 4**: For any given $r$, there is a critical level of monitoring cost $m^*(r)$, such that
integration is (weakly) optimal for $m < m^*(r)$, and separation is (weakly) optimal for
$m > m^*(r)$.

**Proof**: Let $d_i^{RE}(r, m)$ and $d_s^{RE}(r, m)$ be the decisions that make constraints (7) and (12)
binding, respectively. We know from (7) that $d_i^{RE}(r, m)$ decreases in $m$, so we can define
$m^*(r)$ as the largest monitoring cost such that $d_i^{RE}(r, m) \geq d^*$. Because of Proposition 2,
this implies that $d_{r}^{RE}(r, m) > d_{s}^{RE}(r, m)$ for $m < m^*(r)$, and $d_{s}^{RE}(r, m) > d_{r}^{RE}(r, m)$ for $m > m^*(r)$. Hence, it must be that $JS_{i}^{RE} \geq JS_{s}^{RE}$ for $m < m^*(r)$, with the inequality holding strictly for $d_{s}^{RE}(r, m) < d_{FB}^{FB}$, and $JS_{s}^{RE} \geq JS_{r}^{RE}$ for $m > m^*(r)$, with the inequality holding strictly for $d_{r}^{RE}(r, m) < d_{FB}^{FB}$. QED.

As shown in Table 1, Proposition 4 is consistent with several empirical works on franchising, which show that retail outlets distant from the franchisor’s headquarters (Brickley and Dark (1987), Arruñada, Vázquez and Zanarone (2009)) or geographically dispersed (Lafontaine (1992), Lafontaine and Shaw (2005))—and, therefore, more difficult to monitor—tend to be franchised, rather than integrated. As an alternative explanation, Lafontaine and Slade (1996) have proposed that, in the presence of high monitoring costs, direct measures of the agent’s performance are noisier, so the payoffs of a risk-averse agent should be based to a greater extent on indirect measures, such as sales or profits. If compensating the agent based on profits requires transferring assets to her, this implies a move from integration towards separation. An advantage of the explanation proposed here is that it does not require agent’s risk-aversion, which, as mentioned before, seems desirable, at least in retail contracting.

5.4. Markets, hybrids and hierarchies

The model can be applied to revisit the categories of markets, hybrids and hierarchies. According to Williamson (1991), markets are good at promoting “unconsciuos cooperation”, whereby separate firms take decisions autonomously, without consulting
each other, and rely on prices as their guidance. Conversely, firms excel at promoting “purposeful cooperation” between parties who are mutually dependent, because they rely on formal organization and hierarchy to produce coordination and prevent opportunism. Hybrid organizations—long-term contracting, alliances, franchising, and the like—stand somewhere in the middle, as they aim to inject “purposeful cooperation” into relations between separate firms. The upshot is that, as “adaptation problems” between mutually dependent parties become more “consequential”, one should expect governance to evolve from markets into hybrids, and from hybrids into hierarchies.

In the model presented here, “markets” are captured by spot separation, where P does not give orders to A and A is an independent firm; “hierarchy” is captured by relational integration, where P gives orders to A and A is P’s employee; and “hybrids” are captured by relational separation, where P gives orders to A and A is an independent firm. Also, adaptation problems between mutually dependent parties can be interpreted as high levels of the parameter $b$, which measures how distant the decision that maximizes the principal’s and the agent’s joint profits is from the one a spot market contractor would spontaneously choose. Assume, now, that, in order to enter relational contracts, P and A must incur, ex ante, a positive cost $\varepsilon$—for instance, to elaborate a common language enabling P to communicate the discretionary bonus function to A. Then, we can state the following

**Proposition 5**: Spot separation is optimal when adaptation problems are non-consequential, relational separation is (weakly) optimal when adaptation problems are moderately consequential, and relational integration is (weakly) optimal when adaptation problems are highly consequential.
Proof: We know from Proposition 3 that there is $b^*(r)$ such that relational separation is (weakly) optimal for $b < b^*(r)$ and relational integration is (weakly) optimal for $b > b^*(r)$. Assume, now, that $b = 0$. Then, the first best decision is $d^{SP}_\delta$, which can be achieved under both spot and relational separation, but more efficiently so under the former, because the parties do not need to incur the cost of relational contracting $\epsilon$. Hence, we can conclude that spot separation is optimal for $b = 0$, relational separation is (weakly) optimal for $0 < b < b^*(r)$, and relational integration is (weakly) optimal for $b > b^*(r)$.

QED.

The difference between how hybrids and hierarchy are defined in Williamson (1991) and here should be familiar by now. In Williamson (1991), hybrids and hierarchy are exogenously given “discrete structural alternatives”, characterized by increasing amounts of administrative fiat and formal authority, and governed by increasingly flexible bodies of contract law, which allow for quicker and more flexible dispute resolution. Here, I focus on flexible legal systems, where the degree of authority, administrative controls and enforcement rules available to the principal does not depend on firms’ boundaries, and I show that the empirical relation between adaptation problems and governance structure predicted by Williamson (1991) continues to hold. Specifically, markets (spot separation) are optimal when adaptation problems are “non-consequential” ($b = 0$), for in that case the agent’s preferred decision coincides with the first best; hybrids (relational separation) are optimal when adaptation problems are “moderately consequential” ($0 < b < b^*(r)$), for in that case limited performance suffices, so the fact that the agent cares for profits keeps her incentives relatively aligned with the principal’s, making it easier to sustain informal
cooperation through relational contracts; and hierarchy (relational integration) is optimal when adaptation problems are “highly consequential” \((b > b^*(r))\), for in that case performance substantially above the spot market level is needed, so incentives are best aligned by making the agent an employee with no residual stake in profits.

### 5.5. Repeated interactions

Since informal relationships require the prospect of interacting repeatedly in the future, a key parameter of the model is the interest rate, which can be seen as an inverse measure of how long-term the relationship between P and A is. The link between governance structure and the interest rate is given by the following

**Proposition 6**: There is a critical interest rate \(r^*\), such that integration is (weakly) optimal for \(r < r^*\), and separation is (weakly) optimal for \(r > r^*\).

**Proof**: Let \(d^{RE}_i(r)\) and \(d^{RE}_s(r)\) be the decisions that make constraints (7) and (12) binding, respectively. We know from (7) that \(d^{RE}_i(r)\) decreases in \(r\), so we can define \(r^*\) as the largest interest rate such that \(d^{RE}_i(r) \geq d^*\). Because of Proposition 2, this implies that \(d^{RE}_i(r) > d^{RE}_s(r)\) for \(r < r^*\), and \(d^{RE}_s(r) > d^{RE}_i(r)\) for \(r > r^*\). Hence, it must be that 

\[
J_{S}^{RE} \geq J_{S}^{RE} \quad \text{for} \quad r < r^*, \quad \text{with the inequality holding strictly for} \quad d^{RE}_s(r) < d^{FB}, \quad \text{and}
\]

\[
J_{S}^{RE} \geq J_{S}^{RE} \quad \text{for} \quad r > r^*, \quad \text{with the inequality holding strictly for} \quad d^{RE}_i(r) < d^{FB}. \quad \text{QED.}
\]

Proposition 6 suggests that, when P and A face the prospect of a more durable relationship, integration is a more effective instrument to achieve fiat than separation. The
reason is that, due to the tradeoff between effort and profits faced by the agent, the principal must pay greater bonuses to employees than contractors as the decision to be implemented moves beyond the critical level $d^*$. When the principal-agent relationship is sufficiently tight, P and A move towards the region where $d > d^*$ is feasible, in which case integration allows P to ask A for greater performance than separation.

Proposition 7 stands in contrast with Garvey (1995), who shows that long-term relationships favor separation over integration. The difference is that, in Garvey (1995), the value of assets is monotonically increasing in the parties’ performance, so taking assets away from a party, by making her the other’s employee, always increases her reneging temptation. Hence, when effort is needed from both parties, the aggregate reneging temptation can be reduced via a more even split of assets, which Garvey (1995) interprets as a movement from integration towards separation. Conversely, in this paper, the asset’s value is concave in the agent’s performance, implying that, when the agent is an employee, her temptation to renege on a large level of performance is smaller than when she owns the unit (Proposition 2). Therefore, integration achieves greater performance than separation when the parties can aim high—that is, when they are patient enough for a relational contract involving a large level of effort to be feasible at all. This modeling approach is arguably more suitable than Garvey’s (1995) one in settings where both revenues and opportunity costs from the agent’s actions are difficult to contract, and are therefore appropriated by the party who owns the assets (Maness (1996)).
5.6. Testing the relation between governance structure and fiat

While fiat, administrative controls and monitoring have been regarded as distinctive features of firms at least since Coase (1937), the empirical link between firms, monitoring and fiat remains, to my knowledge, untested. Proposition 2 in this paper suggests the relation may be non-monotonic, with fiat and monitoring being greater within firms (integration) when the environment favors higher-level decisions, and between firms (separation) when the environment favors lower-level decisions.

As in any empirical study of organizational outcomes, a crucial problem is that integration and separation are endogenous decisions that depend on the environment, so regressing measures of fiat and monitoring intensity on a dummy for integration may yield biased results (Gibbons (2005)). It turns out that, for some of the environmental parameters studied here, we can sign the bias, as follows

**Proposition 7:** Holding $d^*$ constant, A is more subject to fiat and monitoring in observed firms than in observed inter-firm relations.

**Proof:** Suppose that, in environment $x$, $JS^RE(x) > JS^RE_S(x)$, so integration is strictly optimal. Then, Proposition 2 implies that $d^RE(x) > d^*(x)$. Conversely, suppose that, in environment $x'$, $JS^RE_S(x') > JS^RE_I(x')$, so separation is strictly optimal. Then, Proposition 2 implies that $d^RE_S < d^*(x')$. Hence, for any pair of environments $x$ and $x'$ such that integration is optimal in $x$, separation is optimal in $x'$ and $d^*(x)=d^*(x')$, it must be that $d^RE_I(x) > d^RE_S(x')$. Since $M(d)$ increases in $d$, this also implies that $M(d^RE_I(x)) > M(d^RE_S(x'))$. QED.
Since $d^*$ only depends on $A(d)$ and $C(d)$, shifts in parameters such as $b$, $m$ and $r$ will leave $d^*$ unchanged, all else equal. Then, Proposition 7 implies that, if one controls for parameters affecting $A(d)$ and $C(d)$ but fails to control for $b$, $m$ and $r$, regressing measures of fiat and monitoring costs on a dummy for integration will yield biased and positive coefficients.

As an example, suppose we want to compare the level of fiat and monitoring exerted by a franchisor on franchisees and on employees who operate in similar locations (that is, with plausibly similar values of $B(d)$, $A(d)$, $M(M(d))$ and $C(d)$), but who may expect to terminate their relationship with the franchisor at different times (that is, they may face different values of $r$). Measuring the prospect of future interactions is difficult, so we may be unable to properly control for $r$.\textsuperscript{10} Then, according to Proposition 7, we should find that employees are subject to greater fiat and monitoring than franchisees. This may tempt us to invoke Coase (1937) or Williamson (1991): turning a franchisee into an employee exogenously increases the franchisor’s power of fiat and his reliance on monitoring and administrative control. However, according to Proposition 2, these conclusions may be dictated by section bias: if outlets that are optimally franchised were integrated, the franchisor’s power of fiat over their managers and his investments in monitoring technology would be even smaller.

\textsuperscript{10} Consistent with that, most empirical studies of relational contracts rely on measures of past interactions as proxies for $r$ (Corts and Singh (2004), Kalnins and Mayer (2004)). An exception is Gil and Marion (2009), where highway construction schedules issued by the Californian public administration are used as exogenous proxies for the prospect of future interactions between contractors and subcontractors.
5.7. Public vs. private bureaucracies

Finally, the model could be applied to study when the State should provide a public good or service directly, through civil servants (integration), and when it should outsource it to private contractors (separation). The model would predict State supply as the optimal solution when the good has substantial social value (high $b$). Consistent with that, Levin and Tadelis (2009) find that American cities are substantially more likely to provide in-house services to whose quality citizens are especially sensitive, and “core” services, defined as those that are provided by a broad set of cities. Moving to casuistic evidence, Hart et al. (1997) document that services whose quality has high social value, but is hard to describe and so can be easily manipulated—such as foreign policy and the operation of prisons—tend to be supplied in house.

To explain the evidence, Levin and Tadelis (2009) argue that State employees are compensated based on hours worked, while private contractors are compensated based on quality. Insofar as hours worked are cheaper to measure, and can substitute effort in producing a given quality level, it is preferable to handle services that require high quality through civil servants, and save the (high) costs of contracting on quality directly. In a different vein, Hart et al. (1997) argue that the State can cheaply contract quality standards with its agents, independent of whether these are civil servants or private contractors. However, the State relies on these agents to propose policies on how to contain costs and enhance quality, and the agents’ initiative is non-contractible. In this setting, civil servants have less incentive than private contractors to propose policies that cut costs at the expense of quality, because they anticipate that the State, who owns the relevant facilities, would
reject “bad” proposals, so they could not use them as bargaining chips to extract salary increases.

A limitation in Levin and Tadelis (2009) is that, while they show evidence on the choice between integration and separation in public procurement, they model the (different) choice between more or less complete procurement contracts. To link their evidence to the model, they assume incomplete (complete) contracts are akin to integration (separation). However, this needs not be the case: in theory, the State could write complete, performance-based contracts with employees, or specify that private contractors must work for a certain amount of time following its instructions, as in “cost-plus” private procurement (Corts and Singh (2004)) and franchising. The assumption that time and effort are substitute inputs for quality seems also problematic, especially for tasks that require creativity and attention. This paper overcomes both limitations: on one hand, it explicitly models the choice between integration and separation. On the other hand, it only requires a non-contractible agent’s input (the action $d$, which could be interpreted as effort or attention), with no need for substitutability between effort and time.

A limitation in Hart et al. (1997) is that, by assuming perfectly contractible quality and focusing on ex ante agents’ initiative, they do not capture (possibly relevant) settings where the State dictates its preferred policies to the agents—so initiative is not a major issue—but cannot specify contractually how such policies are to be implemented. Conversely, this type of setting is nicely captured by the model presented here, which emphasizes ex post implementation, rather than ex ante initiative.
Overall, the model in this paper seems closest to the informal theory of public procurement in Williamson (1999), who argues that sovereign transactions—including foreign policy—are better handled by civil servants, because they have low-powered incentives, and, therefore, are less likely to commit “probity” hazards—that is, to sacrifice their mission in order to gain personal benefits and contain operating costs. Williamson (1999) reckons that, with low-powered incentives, civil servants may lose motivation (as in section 3 of this paper), but argues that this can be avoided through periodic promotions, and by providing “added social conditioning” (p. 329); as an example of the latter, he cites the American Foreign Service, which “places great emphasis on the intellectual and social superiority of its members…[who enjoy] being in the mainstream of foreign policy decision-making” (p. 329). The forms of social conditioning discussed by Williamson (1999) may be seen as non-monetary means to create the stream of quasi-rents $w_t + \beta_t$ modeled in section 4, which civil servants would lose if caught shirking. While reputational concerns and social conditioning will also matter to private contractors, they may be less stringent than for civil servants if the policies desired by the State highly diverge from the ones that maximize the contractor’s short-term profits, for instance, because they require high quality performance at high cost.
6. Conclusion

This paper has studied the role of fiat within and between firms. It has assumed that bringing a transaction inside the firm does not affect a principal’s ability to monitor his agent and to enforce formal contracts, but importantly affects the agent’s incentives and, that way, the principal’s ability to enforce decisions informally. The key point is that a firm’s employee cares about saving effort but does not care about the value of local assets, whereas a contractor cares about both. Hence, a contractor is less tempted than an employee to disobey moderate decisions, which are close to those that maximize the value of her assets, while an employee is less tempted to disobey higher-level decisions.

Unlike previous theories, the model predicts a non-monotonic relation between firms and fiat: a principal can more effectively enforce low-level decisions when the agent is a contractor, and high-level decisions when she is an employee, so what organizational form best serves the principal’s need to gain control and authority over the agent depends on the environment. This is consistent with the observation that many inter-firm relations, such as franchising, are governed by fiat, informal dispute resolution and close supervision of agents.

To my knowledge, there is no empirical evidence on the relation between governance structure and fiat. This paper offers a precise testable prediction: once controlling for potential differences in the available set of formal organizational instruments within and between firms—which this model abstracts from—and for the endogeneity of integration decisions, one should observe that integration reduces fiat when moderate performance is
optimal, and increases it when high-level performance is optimal. Bringing this proposition to the data is a challenge I hope to pursue in future work.

References


Table 1. Evidence on the effect of spillovers and monitoring costs on integration

<table>
<thead>
<tr>
<th>Empirical study</th>
<th>Year</th>
<th>Industry</th>
<th>Agent’s task</th>
<th>Observed variations</th>
<th>Change in model’s parameters</th>
<th>Effect on integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nickerson &amp; Silverman</td>
<td>2003</td>
<td>Trucking</td>
<td>Drive safely</td>
<td>Less-than-truckload</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Yeap</td>
<td>2006</td>
<td>Chain restaurants</td>
<td>Control service quality</td>
<td>In-house production; dine-in service; high price</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Forbes &amp; Lederman</td>
<td>2009</td>
<td>Air transportation</td>
<td>Reschedule flights</td>
<td>Bad weather airport</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Levin &amp; Tadelis</td>
<td>2009</td>
<td>Public procurement</td>
<td>Provide high quality service</td>
<td>Citizens more sensitive to quality</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Lafontaine &amp; Shaw</td>
<td>2005</td>
<td>Retailing</td>
<td>Manage outlet</td>
<td>Franchisor operates in more states</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>