Cream skimming in financial markets*

Patrick Bolton
Columbia University and NBER

Tano Santos
Columbia University and NBER

Jose A. Scheinkman
Columbia University, Princeton University and NBER

May 22, 2015

ABSTRACT

We propose a model where investors may choose to acquire costly information that identifies good assets and purchase these assets in opaque (OTC) markets. Uninformed investors access an asset pool that has been cream-skimmed by informed investors. When the quality composition of assets for sale is fixed there is too much information acquisition and the financial industry extracts excessive rents. In the presence of moral hazard in origination, the social value of information varies inversely with information acquisition. Low quality origination is associated with large rents in the financial sector. Equilibrium acquisition of information is generically inefficient.

*This article was previously circulated under the title "Is the Financial Sector too Big?" We thank seminar participants at the Bank of Spain, Cambridge University, CEMFI, EHESS, HEC-Paris, HKUST, INSEAD, London Business School, London School of Economics, MIT-Sloan, Rochester, Rutgers Business School, Studienzentrum Gerzensee, Toulouse School of Economics, University of Virginia-McIntire, Washington University-Olin Business School, Wharton, and the 6th Banco de Portugal Conference on Monetary Economics for their feedback and suggestions. We also thank Guido Lorenzoni and Marzena Rostek for detailed comments and Rishab Guha and Linan Qiu for research assistance. This paper was awarded the IV Jaime Fernández Araoz Prize.

This paper is forthcoming in *The Journal of Finance.*

What does the financial industry add to the real economy? What is the optimal organization of financial markets, and how much talent is required in the financial industry? We revisit these fundamental questions in light of recent events and criticisms of the financial industry. The core issue underlying these questions is whether the financial industry extracts excessively *high rents* from the provision of financial services and whether these rents attract too much talent.[1] Figure V Panel B from (Philippon and Reshef, 2012, p. 1569) plots the evolution of US wages (relative to average non-farm wages) for three subsegments of the finance services industry: credit, insurance and 'other finance.' 'Credit' refers to banks, S&Ls, and other similar institutions, 'insurance' to life and P&C insurers, and 'other finance' to the financial investment industry and investment banks. The most remarkable finding is that the bulk of the growth in remuneration in the financial industry took place in the 'other finance' sector.

In this paper we attempt to explain the outsize remuneration in this latter sector by modeling a financial industry that is composed of two sectors: an organized, regulated, standardized, and transparent market where most retail ('*plain vanilla*') transactions take place, and an informal, *opaque* sector, where *informed* professional investors trade and '*bespoke*' services are offered to clients. We refer to the transparent, standardized, markets as organized exchanges and call the opaque sector the over-the-counter (OTC) sector even though in practice not all OTC markets are opaque (e.g. markets for foreign exchange are quite transparent). A central idea we develop is that although OTC markets provide special valuation services to sellers of securities, their opacity also allows informed dealers to extract informational rents. What is more, OTC markets cannibalize organized exchanges by '*cream-skimming*' the juiciest deals away from retail investors.[2] What is worse, informed investors don't just get access to the best assets but they also benefit from the cream-skimming activities of other informed dealers which help them extract better terms from the asset sellers. The reason is that cream skimming worsens the quality of the pool of assets flowing into the exchange, thereby lowering the price a seller of a good asset can obtain in the exchange as an outside option. This is how informed dealers are able to extract outsize terms from the sellers of good assets. Cream-skimming and outsize remuneration are two sides of the same coin. Moreover, the outsize informational rents in OTC markets attract talent to the financial industry, thereby enhancing the cream-skimming externality and further increasing the information rents extracted by informed dealers. In this way, growth in OTC markets does not lead to the dissipation of these rents. On the contrary, dealer remuneration rises as the OTC sector grows, even when growth is not due to increased demand.

The role of the financial sector in our model is to provide liquidity by allowing asset originators to sell their assets to investors. A related role is the valuation of assets for sale when assets vary in quality. This valuation service is offered by informed dealers in OTC markets, but although originators are willing to pay dealers to help them identify valuable assets, in our model there is no social benefit from these valuation services. That is, unless more accurate valuations also serve the role of providing better origination incentives. Otherwise, costly information acquisition by professional investors only serves the purpose of *cream-skimming* the best assets, which is valuable to the trading parties but has no social value. This result is related to the general observation first made by Hirshleifer (1971) about the private value of *foreknowledge information*, which in his formulation only involves "the value of *priority* in time of superior knowledge". Or, as (Stiglitz, 1989, p. 103) later put it:

> "Assume that as a result of some new information, there will be a large revaluation of some security, say from \$10 to \$50. Assume that that information will be announced tomorrow in the newspaper.....The information has only affected who gets to get the return. It does not affect the magnitude of the return. To use the textbook homily, it affects how the pie is divided, but it does not affect the size of the pie."

Of course, by identifying the most valuable assets and offering more attractive terms for those assets, informed dealers in the OTC market may also provide incentives to originators. However, we show that even when they provide better origination incentives, equilibrium entry by informed dealers into the OTC market is generically inefficient. In particular, there is too much entry into the dealer market when the marginal cost of originating good assets is high. But the equilibrium size of dealer markets can also be too small. This is the case when the marginal cost of originating good assets is small. While there is then a large "bang for the buck" in increasing the size of the informed dealer market, equilibrium incentives to enter the dealer market are too small. Indeed, when a high fraction of good assets is originated the surplus that informed dealers can cream-skim is relatively small as the difference in value between the good and average assets is small. In contrast, when the cost of origination effort is high there is little "bang for the buck" in increasing the size of the dealer market. But then private incentives to enter the dealer market are large for then the benefits from cream-skimming are large.

In our model OTC dealers' rents *increase* as more informed dealers enter the dealer market, because collectively dealers are able to extract larger cream-skimming rents. When more dealers cream-skim good assets this worsens the terms originators can get for their assets on the organized exchange, and therefore their bargaining power on OTC markets. This central mechanism in our model explains why compensation in OTC markets continued to grow even while significant new entry into this sector took place.

Our assumption that trading in OTC markets is opaque contrasts with the standard framework of trading in financial markets first developed by Grossman and Stiglitz (1980). In that class of models, privately produced information leaks out in the process of trading, and as a result *too little* costly information may be produced by 'insiders.' Since many activities in the financial industry boil down to costly private 'information acquisition', the Grossman-Stiglitz model seems to suggest that the financial sector could be too small. In contrast, our model explains how excessive rent extraction together with excessive entry into the financial industry can be an *equilibrium outcome*.

More generally, our analysis offers a novel explanation for three broad related facts about the recent evolution of the US financial services industry, as shown by Philippon and Reshef (2012) and Philippon (2015, 2012, 2008). First, the financial services industry accounts for an increasing share of GDP even after financial services exports are excluded - an increase that accelerated starting in the mid 80s. Second, this growth has been accompanied by a substantial increase in information technology (IT) spending in the financial sector. As (Philippon, 2012, Figures 5 and 6) emphasizes, while other sectors, such as retail, have also increased the fraction of IT spending these sectors have experienced a decline in their income shares in GDP. Third, in contrast to these sectors, the segment of finance that is most closely associated with OTC transactions has experienced a rise in income shares along with IT spending. Our model explains how IT efficiency gains can lead to substantial increases in compensation in brokerage and asset management even though there is entry of new talent into this sector. As IT became cheaper, OTC activities which are information intensive became more profitable relative to exchange traded activities. The increase in OTC dealers' rents that resulted from the entry of more dealers provided a reinforcement mechanism for the growth of compensation in OTC activities and prevented the dissipation of the rents from cheaper IT that was observed in other sectors.

Although our model can in principle be applied to trading activities of any securities that may be traded both on OTC markets and organized exchanges, the most direct interpretation is in terms of private placements versus public offerings of stocks. Private placements, in line with the general growth in OTC markets, have grown substantially relative to public offerings in the past two decades. Interestingly, the growth in private placements is partly the result of deregulation, following the adoption by the SEC of rule 144A in 1990, which exempts qualified institutional investors from registration requirements. As we discuss in Bolton, Santos, and Scheinkman (2012) the relaxation of registration requirements under rule 144A substantially increased secondary market liquidity and facilitated cream-skimming through private placements.[3]

The paper is organized as follows: Section 1 outlines the model. Section 2 analyzes the pure information acquisition problem by rentier investors who can become informed dealers at a cost. We show that in this situation there is always too much entry into the informed dealer market. In section 3 we introduce moral hazard in origination and show that even when there is a social value to information acquisition by dealers equilibrium entry of rentiers into the dealer market is generically inefficient. Moreover, the social value of information acquisition and trading by dealers varies inversely with private incentives to acquire information: just when it is socially optimal to limit costly investment in information by dealers, investors' private incentives to acquire information are at their highest. In our framework high marginal costs of originating high quality assets are associated with large rents in the financial sector and thus situations in which low quality assets tend to be originated are also those when the profits in the OTC sector are the highest. Section 4 concludes. All proofs are in the Appendix.

Related Literature. In his survey of the literature on financial development and growth, Levine (2005) synthesizes existing theories of the role of the financial industry into five broad functions: 1) information production about investment opportunities and allocation of capital; 2) mobilization and pooling of household savings; 3) monitoring of investments and performance; 4) financing of trade and consumption; 5) provision of liquidity, facilitation of secondary market trading, diversification, and risk management. As he highlights, most of the models of the financial industry focus on the first three functions, and if anything, conclude that from a social efficiency standpoint the financial sector is too small: due to asymmetries of information, and incentive or contract enforceability constraints, there is underinvestment in equilibrium and *financial underdevelopment*.

In contrast to this literature, our model emphasizes the fifth function in Levine's list: secondary market trading and liquidity provision. In addition, where the finance and growth literature only distinguishes between bank-based and market-based systems (e.g. Allen and Gale (2000)), a key distinction in our model is between markets in which trading occurs on a bilateral basis at prices and conditions that are not observable by other participants, and organized exchanges with multilateral trading at prices observed by all.[4]

Our paper contributes to a small literature on the optimal allocation of talent to the financial industry. An early theory by Murphy, Shleifer, and Vishny (1991) (see also Baumol (1990)) builds on the idea of increasing returns to ability and rent seeking in a two-sector model to show that there may be inefficient equilibrium occupational outcomes, where too much talent enters one market since the marginal private returns from talent could exceed the social returns. More recently, Philippon (2010) has proposed an occupational choice model where agents can choose to become workers, financiers or originators. The latter originate assets which have a higher social than private value, and need to obtain funding from financiers. In general, as social and private returns from investment diverge it is optimal in his model to subsidize entrepreneurship. Biais, Rochet, and Woolley (2015b) propose a model of innovation with learning about risk and moral hazard, which can account for the simultaneous growth in the size of the financial industry and remuneration in the form of rents to forestall moral hazard. Neither Murphy et al. (1991), Biais et al. (2015b), or Philippon (2008) distinguish between organized exchanges and OTC markets in the financial sector, nor do they allow for excessive informational rent extraction through cream-skimming.

More recently, three independent studies have analyzed problems with some common features to ours: First, Glode, Green, and Lowery (2012) also model the idea of excessive investment in information as a way of strengthening a party's bargaining power. However, Glode et al. (2012) do not consider the occupational choice question of whether too much young talent is attracted towards the financial industry. Second, Fishman and Parker (2012) also consider a model in which informed buyers can extract information rents. Their model generates multiple Pareto ranked equilibria with different levels of information acquisition. As in our framework, they find that too much information may be produced in equilibrium. They also show that information acquisition can give rise to what they refer to as valuation runs, where sellers of assets that cannot be valued are unable to find buyers. Third, Biais, Foucault, and Moinas (2015a) analyze a model where high-frequency traders can gain a valuation advantage over other traders. They show that investment in speed by some traders (co-location, faster computers, microwave transmission, etc.) can give rise to equilibrium inefficiencies, if most of the benefits of added speed are in the form of cream-skimming from slower to faster traders.

Finally, our paper relates to the small but burgeoning literature on OTC markets, which, to a large extent, has focused on the issue of financial intermediation in the context of search models.[5] These papers have some common elements to ours, in particular the emphasis on bilateral bargaining in OTC markets, but their focus is on the liquidity of these markets and they do not address issues of cream-skimming or occupational choice.

I. The model

We consider a three periods ($t = 0, 1, 2$) competitive economy divided into two sectors: a productive and a financial sector.

A. Agents

There are two types of risk-neutral agents in this economy, each with a unit mass. The first type is *originators* of assets: These can be thought of as either company founders who sell their business, or as bankers who originate loans. Each originator has not only a storable endowment ω but also has access to a productive asset that yields either $\rho \geq 1$ or υ in period 2, with $\upsilon > \rho$. The probability that the asset yields υ is $a \in (0, 1)$. In the baseline model we take a to be exogenously given, and in a later generalization we allow originators to exert costly effort to increase a. The realization of asset returns is independent across originators. Moreover, originators only learn the quality of the asset at time $t = 2$. For simplicity, we assume that originators can only consume at time $t = 1$, so that their utility function is given by $U(c_1, c_2) = c_1$. As a result they must sell the asset they have originated at $t = 1$.

The second type of agents is *rentiers* who start out in period 0 with a storable endowment ω, which they may consume in either period 1 or 2. Their preferences are represented by the utility function: [6]

$$u(c_1, c_2) = c_1 + c_2. \tag{1}$$

Rentiers are uninformed investors. But at time $t = 0$ they can choose to become informed dealers by acquiring the required valuation skills at some positive utility cost. Formally, we assume that each rentier $d \in [0, 1]$ who chooses to become a dealer incurs a cost $\varphi(d) \geq 0$ (in units of utility), where $\varphi(\cdot)$ is a strictly convex and increasing function of d. Once the cost $\varphi(d)$ is incurred dealers are able to perfectly identify date-2 asset returns at $t = 1$ and stand ready to acquire assets from originators in that period. For simplicity, we allow each dealer to acquire at most one asset at $t = 1$. Throughout our analysis we denote by μ the endogenous measure of rentiers who become dealers. Our assumptions on the cost function $\varphi(.)$ allow us to restrict attention to a set of dealers of the form $[0, \mu]$.

We are interested in situations in which dealers are on the short side of the market, with more good projects for sale than dealers can purchase. We thus assume that $\varphi(d) = \infty$ for $d \geq \bar{d}$, with:

$$a > \bar{d}. \tag{2}$$

B. Financial Markets

A novel feature of our model is the *dual structure* of the financial system: Assets can be traded in either an *over-the-counter* (OTC) dealer market or on an organized exchange. Dealers operate in the OTC market, so that information about asset values resides in this market. On the organized exchange, on the other hand, all asset trades are uninformed trades between rentiers and originators.

In period 1 an originator has two options: she can sell her asset for the competitive equilibrium price p on the organized exchange or she can go to a dealer in the OTC market and negotiate a sale for a price p^d. As already mentioned, we assume that each dealer can acquire at most one unit of the asset, so that the total number of assets sold in the OTC market equals μ.

We begin our analysis by first taking the price p as given and derive equilibrium trade and price p^d in the OTC market given p. When a asset is identified by a dealer as worth υ then the gains from trading this asset over the counter are $(\upsilon - p)$. We take the price p^d at which this sale is negotiated between a dealer and an originator as the outcome of bilateral bargaining.[7] An originator receiving an offer from a dealer thereby learns that she has a good asset, so that bargaining is effectively under symmetric information. We therefore can take the solution to this bargaining game to be the *Asymmetric Nash Bargaining Solution* where the dealer has a given bargaining power $(1-\kappa)$ and the originator has bargaining power $\kappa \in (0,1)$ (see Nash Jr (1950); Nash (1953)).[8] The price p^d is then given by:

$$p^d = \arg\max_{s\in[p,\upsilon]} \{(s-p)^{\kappa}(\upsilon-s)^{(1-\kappa)}\} \qquad \Rightarrow \qquad p^d = \kappa\upsilon + (1-\kappa)p.$$

A justification for the Nash bargaining solution based on the sequential strategic approach to bargaining can be found in Binmore, Rubinstein, and Wolinsky (1986) (BRW). They show that the asymmetric Nash bargaining solution can be justified as an equilibrium outcome of a strategic model of bargaining with uncertain termination time. The threat point is given by the outcome in the event that the bargaining process does break down. In that event our assumptions is that the dealer simply consumes his endowment and the originator sells her asset in the exchange. In BRW, the parameter κ reflects the relative subjective probabilities that each party attributes to the breakdown of the bargaining process. In the sequential strategic justification for the asymmetric Nash bargaining solution developed by BRW, the number of dealers relative to originators does not affect κ. Still, we explore the implications of letting κ vary with the measure of dealers in the next subsection. We show that letting κ vary with the size of the dealer sector does not change our main results. Accordingly, we maintain the assumption of a constant κ throughout most of our analysis in the interest of simplicity.

We also assume that the measure of uninformed rentiers, which recall is normalized to 1, is large enough to be able to absorb all asset sales in period 1, so that the equilibrium price of assets sold on the exchange is equal to their expected payoff.[9]

Since $p^d < \upsilon$, and since dealers are on the short side of the market, they will acquire μ good assets. The remaining good assets will necessarily flow to the competitive exchange, so that the expected value of an asset, and therefore the competitive price p on the exchange, equals:

$$\left(\frac{a-\mu}{1-\mu}\right)\upsilon + \left(\frac{1-a}{1-\mu}\right)\rho,$$

which can be rewritten as:

$$p = \rho + \left(\frac{a-\mu}{1-\mu}\right)(\upsilon - \rho). \tag{3}$$

The price on the exchange equals the value of a bad asset plus the incremental value of a good asset times the expected fraction of good assets sold on the exchange. Note that an increase in the measure of dealers μ means that more good assets are *cream-skimmed* off the exchange. This lowers the fraction of good assets on the exchange and consequently the price p. Moreover, as a result of the reduction in p the price p^d is also reduced. Thus, remarkably, as the measure of dealers μ increases the fraction of the surplus that dealers extract in each OTC transaction, $\upsilon - p^d = (1-\kappa)(\upsilon - p)$, also increases.

It is a dominant strategy in period 1 for originators to attempt to first approach a dealer. They understand that with probability a the underlying value of their asset is high, in which case they are able to negotiate a sale with a dealer at price $p^d > p$ with probability $m \in [0,1]$. If they are not able to sell their asset for price p^d to a dealer, originators can turn to the organized market in which they can sell their asset for p. We assume that the probability m is simply given by the ratio of the total mass of dealers μ to the total mass of high quality assets up for sale by originators a so that

$$m = \frac{\mu}{a}. \tag{4}$$

Notice that (2) guarantees that $m \in [0,1)$.

II. Equilibrium Size of the Dealer Market

A. Payoffs of Dealers and Originators

Let $U(\mu)$ be the expected payoff of an originator when the measure of dealers is μ. Similarly let $V(d \mid \mu)$ be the expected payoff of dealer $d \leq \mu$ when the measure of dealers is μ. The originator's expected payoff at time 0 is:

$$U(\mu) = \omega + am(\mu)p^d(\mu) + (1 - am(\mu))\, p(\mu) \tag{5}$$

$$\text{where:}\quad p^d(\mu) = \kappa\upsilon + (1-\kappa)\, p(\mu) \quad \text{with} \quad p(\mu) = \rho + \left(\frac{a-\mu}{1-\mu}\right)(\upsilon-\rho), \tag{6}$$

$$\text{and:}\quad m(\mu) = \frac{\mu}{a}. \tag{7}$$

In expression (5) if the originator draws an asset yielding υ, which occurs with probability a, and gets matched to a dealer, which happens with probability $m(\mu)$, he is able to sell the asset for $p^d(\mu)$, the price for high quality assets in the dealers' market. If either of these outcomes does not occur, an event with probability $(1 - am(\mu))$, the originator must sell his asset in the uninformed exchange for a price $p(\mu)$.

A dealer's equilibrium expected payoff is:

$$V(d \mid \mu) = \omega - \varphi(d) + (1-\kappa)(\upsilon - p(\mu)), \tag{8}$$

and a rentier's equilibrium payoff is simply ω the value of the rentier's endowment.

The next proposition follows directly from the properties of p as a function of μ (see (3)).

Proposition 1: *The utility of dealer d is an increasing and convex function of the measure of dealers μ.*

The key observation is that dealers in the OTC market *cream skim* the good assets and thereby impose a *negative externality* on the organized market. Cream skimming thus improves terms for dealers in the OTC market – $(1-\kappa)(\upsilon - p(\mu))$– and worsens them for originators looking to sell assets. Indeed, the additional rents that accrue to dealers can only come at the expense of the originator's rents given that the price $p(\mu)$ on the exchange reflects the fair expected value of the assets.[10]

One implication of Proposition 1 is that originators' expected payoff is a decreasing and concave function of μ. This is a more subtle observation than may appear at first. Indeed, notice that an increase in the number of dealers has two effects on an originator's payoffs (5). On the one hand the probability of being matched with an informed dealer goes up as μ rises, which benefits those originators with good assets. But an increase in the number of dealers also lowers prices on the exchange due to the greater *cream skimming*, which in turn leads dealers to bid less for assets in OTC markets. On net, originators are hurt by the reduction in asset prices that follows from the increase in μ.

Another important observation for what comes later is that the difference in payoff between a dealer $d \in [0, \mu]$ and the least efficient dealer μ, $V(d \mid \mu) - V(\mu \mid \mu) = \varphi(\mu) - \varphi(d)$, is also increasing in the measure of dealers μ. In particular, this difference in payoffs between the marginal and infra-marginal dealers is independent of the bargaining power $\kappa \in (0, 1)$.

B. Equilibrium and Optimum Measure of Dealers

Under the assumption embedded in equation (2), all dealers will be able to purchase a good asset in any equilibrium for a profit, given that no more than $\bar{d}$ rentiers will ever choose to become dealers. Under free entry into the dealer market, the equilibrium measure of dealers $\widehat{\mu}$ must be such that the marginal dealer $\widehat{d} = \widehat{\mu}$ breaks even:

$$(1 - \kappa)(\upsilon - p(\widehat{\mu})) - \varphi(\widehat{\mu}) \leq 0 \ (= 0 \text{ if } \widehat{\mu} > 0), \tag{9}$$

where the notation emphasizes the dependence of the equilibrium price on the exchange on the number of dealers (note that there may be multiple solutions to (9)).

Given that $\varphi(d)$ is increasing in d and $p(\mu)$ is decreasing in μ, it follows immediately that when

$$(1 - \kappa)(\upsilon - p(0)) - \varphi(0) = (1 - \kappa)(1 - a)(\upsilon - \rho) - \varphi(0) > 0 \tag{10}$$

every equilibrium is such that $\widehat{\mu} > 0$. In words, when the costs of becoming dealers are small at least for some rentiers, then these rentiers will enter the dealer market. Their cream skimming activities, in turn, will make it more attractive for other rentiers with higher costs to enter this market.

What is the socially optimal measure of dealers μ^S? For any given measure of dealers μ the total expected surplus $W(\mu)$ generated in the economy is simply the sum of total payoffs for each type of agent, rentiers, dealers and originators:

$$W(\mu) = 2\omega + \mu\upsilon + (1 - \mu)p(\mu) - \int_0^{\mu} \varphi(x)dx.$$

Or, substituting for $p(\mu)$:

$$W(\mu) = 2\omega - \int_0^{\mu} \varphi(x)dx + \rho + a(\upsilon - \rho) \tag{11}$$

It immediately follows that $W(\mu)$ is decreasing in μ, so that the socially optimal measure of dealers is $\mu^S = 0$. We are thus able to conclude that:

Proposition 2: *If condition (10) holds all equilibria involve inefficiently large dealer markets:* $\widehat{\mu} > 0$.

The economic logic behind Proposition 2 is straightforward: from an ex ante perspective all that informed dealers do is extract a transfer of surplus away from originators. This activity creates no social value. But, in order to be able to cream-skim dealers must incur socially wasteful costs to acquire the valuation skills that allow them to discriminate good assets from bad ones. Thus, it would be efficient in this simple model to simply shut down the dealer market.[11],[12]

C. *Competition among Dealers and Dealer Rents*

How sensitive are these results to our simplifying assumption that the bargaining power κ is invariant to changes in the measure of dealers μ? How are these results affected if the originators' bargaining power somehow increases as the measure μ of dealers rises relative to the volume of good assets for sale a? We explore these questions in this subsection by letting the originator's share of the surplus from trade with a dealer, $\kappa(\mu)$, increase with μ.[13]

Consider first Proposition 2, which states that when condition (10) holds all equilibria involve excessively large dealer markets. Suppose now that $\kappa(\mu)$ is a continuously differentiable function of μ with derivative $\kappa_\mu > 0$ for all $\mu > 0$. Then, if condition (10) is slightly amended as follows:

$$(1 - \kappa(0))(1 - a)(\upsilon - \rho) - \varphi(0) > 0, \tag{12}$$

it is straightforward to verify that when condition (12) holds all equilibria involve inefficiently large dealer markets. Basically, only in the extreme case where any positive entry by informed dealers $\mu > 0$ immediately results in a $\kappa(\mu) = 1$ can excessive entry be avoided.

Consider next Proposition 1, which states that the utility of dealer d is an increasing and convex function of the measure of dealers μ. In this case, dealer d's equilibrium expected payoff is given by:

$$V\left(d \mid \mu\right) = \omega - \varphi\left(d\right) + (1 - \kappa(\mu))(\upsilon - p\left(\mu\right)), \tag{13}$$

and the change in the dealer's expected payoff with μ is:

$$V_{\mu}\left(d \mid \mu\right) = -(1 - \kappa(\mu))p_{\mu}\left(\mu\right) - \kappa_{\mu}(\mu)(\upsilon - p\left(\mu\right)). \tag{14}$$

Given that

$$p_{\mu} = -\frac{1-a}{(1-\mu)^2}(\upsilon - \rho) < 0,$$

the first term in (14) is positive, but since $\kappa_{\mu}(\mu) > 0$ the second term in (14) is negative, so that the sign of $\partial V\left(d \mid \mu\right) / \partial \mu$ is now ambiguous. However, this analysis does not consider what causes the equilibrium measure of dealers μ to increase. For the equilibrium measure of dealers to increase it must be the case that a change occurs such that the marginal dealer μ is made strictly better off being a dealer than a rentier. The cause of this change in μ would determine the sign of the change in the profitability of infra-marginal dealers.

Consider for example the effect of a decrease in the value of low quality assets, ρ, on equilibrium entry into the dealer market $\mu(\rho)$ and on an individual dealer d's equilibrium expected payoff $V(d, \rho)$. A lower ρ can be interpreted as the extension of credit to the lowest quality borrowers until then rationed out of the market, as was the case in the years immediately preceding the subprime crisis. For any ρ the marginal dealer $\mu(\rho)$ is given by the dealer who is just indifferent between becoming a dealer or not for that ρ. More formally the marginal dealer $\mu(\rho)$ is the solution to the following equation:

$$F(\mu, \rho) = \varphi\left(\mu\right) - (1 - \kappa(\mu))(\upsilon - p\left(\rho\right)) = 0.$$

Or, substituting for $p(\rho)$ using (3):

$$F(\mu, \rho) = \varphi\left(\mu\right) - (1 - \kappa(\mu))(\upsilon - \rho)\left(1 - \frac{a-\mu}{1-\mu}\right) = 0. \tag{15}$$

In general equation (15) admits multiple solutions some of which are stable and others that are not. Focusing on stable equilibria, that is, equilibria for which $F_\mu > 0$, and applying the implicit function theorem we observe that:

$$\frac{d\mu}{d\rho} = -\frac{F_\mu}{F_\rho} < 0,$$

since $F_\rho > 0$. In words: A *decrease* in the value of low quality assets ρ brings about an increase in the equilibrium size of the dealer market μ. This increase in the equilibrium number of dealers in a stable equilibrium is a result of our cream-skimming mechanism. The logic is that a decrease in the value of the low quality assets, holding fixed the number of active dealers μ, decreases the quality of assets in the exchange and therefore the price in the exchange. As a result of the lower price in the exchange (holding μ fixed) all dealers also pay less for the assets they buy in the OTC market so that all dealers' expected payoffs increase. Therefore, if equilibria are stable, equilibrium can only be restored with more entry. Since shifts in ρ result in equilibria along the supply curve of dealers, the expected payoff of all the infra-marginal dealers increases when ρ decreases. This can be more formally seen as follows: Let $V(d,\rho)$ denote the expected payoff of dealer $d < \mu(\rho)$, then

$$\begin{aligned} V(d,\rho) &= V(d,\rho) - V(\mu(\rho),\rho) + V(\mu(\rho),\rho) \\ &= \varphi(\mu) - \varphi(d) + \omega. \end{aligned}$$

So that:

$$V_\rho(d,\rho) = \varphi_\mu \frac{d\mu}{d\rho} < 0.$$

In sum, a reduction in the value of the low quality asset increases entry into the dealer market and hence raises infra-marginal dealers' payoff even if the equilibrium bargaining power of dealers $(1-\kappa(\mu))$ decreases when the size of the dealer market expands. This reasoning applies more generally. Even though the bargaining power of each dealer may decrease with entry, any increase in the number of dealers that is caused by a change in a parameter that does not affect the cost function φ or the initial endowment of all agents ω must necessarily be accompanied by an increase in the profitability of the infra-marginal dealers.

III. Moral Hazard in Origination

In this section we broaden and deepen our analysis by considering how information acquisition in dealer markets can improve originators' incentives to select higher quality assets. To the extent that an originator with a good asset is able to get a better price in the OTC market than on the organized exchange, $p^d > p$, this should encourage her to put more effort to originate a good asset. This positive effect must, of course, be weighed against the value destroying effects of rent extraction through cream skimming to determine the overall efficiency of OTC dealer markets. By introducing moral hazard in origination we not only take into account an important real effect of information acquisition in financial markets, but we are also thereby able to uncover the fundamental mechanism that underlies inefficient entry into dealer markets. As our analysis below makes clear, the key source of inefficiency is that information acquisition by dealers is motivated by the rent $\upsilon - p^d$ they are extracting, which rises rapidly with new entry into the dealer market when origination of good assets is costly and slowly when origination of good assets involves low effort costs. As a result, equilibrium entry into the dealer market is high when information acquisition by dealers has a low social value, and it is low when it has a high social value. This is the main result emerging from our analysis below.

A. Assumptions and Definition of Equilibrium

The probability of originating a good asset $a \in [\underline{a}, 1]$ can now be increased by originators at effort cost $\psi(a - \underline{a}, \theta)$. The parameter θ shifts the marginal cost of effort and for definiteness we assume $\psi_{a\theta} > 0$. Given that all originators are ex-ante identical and face the same incentives, we restrict attention to symmetric equilibria where all originators choose the same a. We begin by imposing mild assumptions on the function $\psi(a - \underline{a}, \theta)$ and $\varphi(d)$ that guarantee uniqueness of an interior *symmetric equilibrium* $(\widehat{a}, \widehat{\mu})$ such that $\widehat{a} > \widehat{\mu}$.

Assumption A1: (i) $\psi_a(0, \theta) = 0$ for each θ ; (ii) $\psi_a \geq 0$ and if $a > \underline{a}$, $\lim_{\theta \to \infty} \psi_a(a - \underline{a}, \theta) = \infty$ (iii) $\psi_{aa} > 0$; (iv) $\psi_{aaa} > 0$; and (v) $\psi_{a\theta} > 0$.

Assumption A2: (i) $\varphi_d > 0$; (ii) there exists $\overline{d} < 1$ such that $\lim_{d \longrightarrow \overline{d}} \varphi(d) = +\infty$, (iii) $\overline{d} \leq \underline{a}$ and (iv) $\varphi(0) < (1 - \kappa)(\upsilon - \rho)$.

Assumption A3: $(1 - d)\varphi(d)$ is non-decreasing in $[0, \overline{d}]$.

The assumption that $\overline{d} \leq \underline{a}$ is a sufficient condition that guarantees that $\widehat{a} > \widehat{\mu}$. It is possible to entirely dispense with this condition at the cost of a somewhat more involved analysis.[14] The assumption that $\psi_{aaa} > 0$ amounts to assuming a convex marginal origination cost; this is a standard assumption in much of the contract theory literature. Finally, Assumption A3, which plays a role in establishing uniqueness of equilibrium, simply requires that dealers' costs of becoming informed rise sufficiently fast with d.

Definition of Equilibrium: An equilibrium is given by: (i) prices $\widehat{p}$ and $\widehat{p}^d$ in period 1 at which the organized and OTC markets clear; (ii) occupational choices by rentiers in period 0, which map into an equilibrium measure of dealers $\widehat{\mu}$ and rentiers $(1-\widehat{\mu})$ (with all rentiers $d \in (0, \widehat{\mu})$ strictly preferring to become dealers); and (iii) incentive compatible effort choices $\widehat{a}$ by originators, which in turn map into an equilibrium matching probability $\widehat{m} \equiv m(\widehat{a}, \widehat{\mu})$.

It should be obvious that $\widehat{a} = 1$ can never be an equilibrium, since if $\widehat{a} = 1$ any originator that defects to $a = \underline{a}$ would be better off. Similarly $\widehat{a} = \underline{a}$ and $\widehat{\mu} = 0$ can never be an equilibrium, since under Assumption A2(iv) it is a best response for the most efficient dealers to enter the dealer market when $\widehat{a} = \underline{a}$. Moreover, under Assumption A1(i) it is a best response for originators to choose $\widehat{a} > \underline{a}$ when there is a strictly positive measure of dealers. In what follows we define $(\widehat{a}, \widehat{\mu})$ to be an interior equilibrium if $\widehat{a} > \underline{a}$ (so that $\underline{a} < \widehat{a} < 1$) and $\widehat{\mu} > 0$. The choice of a by an originator is, of course, not observable to rentiers and dealers. However, all investors can form rational expectations on originators' best effort choice. Also, recall that originators do not know for sure in period 1 what quality of asset they have been able to originate, and they can only discover that they have a good asset when they get an offer p^d from a dealer.

B. Existence and Uniqueness of Equilibrium

The equilibrium $(\widehat{a}, \widehat{\mu})$ in this extended model is given by the solution to two first-order conditions, one for optimal origination effort a and the other for optimal entry into the dealer market μ. Consider a candidate equilibrium $(\widehat{a}, \widehat{\mu})$: The utility of an originator that chooses action a is:

$$U(a|\widehat{a}, \widehat{\mu}) = \omega - \psi(a - \underline{a}, \theta) + a\widehat{m}\widehat{p}^d + (1 - a\widehat{m})\,\widehat{p}. \tag{16}$$

The first-order condition for origination effort a is thus:

$$\psi_a \left(a - \underline{a}, \theta\right) = \widehat{m} \left(\widehat{p}^d - \widehat{p}\right). \tag{17}$$

The originator chooses the effort level $a > \underline{a}$ that equalizes the marginal cost of effort with the marginal benefit of effort, which is given by the probability of being matched to an informed dealer when the originator produces a good asset, $\widehat{m}$, times the price improvement from selling the good asset in the OTC market $\left(\widehat{p}^d - \widehat{p}\right)$. As for entry into the dealer market, since φ is smooth and strictly increasing, we must have:

$$\varphi\left(\widehat{\mu}\right) = (1 - \kappa)\left(\upsilon - \widehat{p}\right). \tag{18}$$

That is, the expected rents associated with cream skimming received by the marginal dealer just compensate for the costs of acquiring information. All dealers with $d < \widehat{\mu}$ are infra-marginal and earn positive rents. All potential dealers with $d > \widehat{\mu}$ choose to stay uninformed. In light of these considerations we are able to establish the following proposition.

Proposition 3: *Under Assumptions A1-A3 there exists a unique interior equilibrium* $(\widehat{a}(\theta), \widehat{\mu}(\theta))$.

Moreover, we can also establish a key comparative statics result with respect to changes in the origination effort cost parameter, θ. When θ is high the marginal cost of originating good assets is high. In that case the fraction of good assets a that is originated is not very responsive to financial market incentives.

Proposition 4: *Assume A1-A3 hold. Then* $\widehat{a}_\theta < 0$, $\widehat{\mu}_\theta > 0$ *and consequently* $\widehat{p}_\theta < 0$.

That is, while it is to be expected that the equilibrium origination of good assets decreases as origination effort costs rise with the parameter θ, it is striking that the equilibrium measure of dealers is actually *increasing* with θ. The more costly it is to originate good assets the more dealers enter the OTC market in equilibrium. The reason is that when the cost of originating good assets increases there are fewer good assets that are brought to the uninformed exchange. This has the effect of lowering the price of assets on the exchange $\widehat{p}$. The lower price on the exchange, in turn, increases the incentives to acquire information, and therefore, the measure of informed dealers $\widehat{\mu}$

that enter in equilibrium, as seen in condition (18).

Proposition 4 highlights the central mechanism at work in this model. If $\widehat{\mu}_\theta > 0$ this means that the equilibrium incentives to acquire information for dealers are rising with the costs of originating good assets. In other words, with a higher θ the marginal cost of originating good assets is higher (since by assumption $\psi_{a\theta} > 0$) so that the social value of dealers' information is lower. However, dealers' equilibrium incentives to acquire information are then higher! That is the basic tension arising from cream skimming: the returns from cream skimming are rising when a higher fraction of good assets are distributed in the dealer market, so that the fraction of good assets sold on the exchange, $\widehat{a} - \widehat{\mu}$, is smaller. When θ is high, the best-response of originators $\widehat{a}(\theta)$ is relatively inelastic to changes in $\widehat{\mu}(\theta)$, but the fraction of good assets sold on the exchange $(\widehat{a}(\theta) - \widehat{\mu}(\theta))$ is highly elastic to changes in $\widehat{\mu}(\theta)$. This is why cream-skimming activities are rewarded the most when their social value is the lowest.

Recall that the net income of the marginal dealer must be zero and hence the net income of an inframaginal dealer d is given by $\varphi(\widehat{\mu}(\theta)) - \varphi(d)$. It follows that:

Corollary 5: *If $d < \widehat{\mu}(\theta)$ the income of dealer d is an increasing function of θ. That is, dealer d is better off the higher is the marginal cost of originating quality assets.*

Taken together the results in Proposition 4 and Corollary 5 show that situations where the quality of asset origination is low are also those where the rents accruing to the dealer sector are large.

C. *Efficiency of Equilibrium*

We now analyze the efficiency of equilibria in the presence of moral hazard in origination. Our efficiency benchmark is the size of a dealer sector μ^S chosen by a planner that maximizes total social surplus subject to the constraint that originators will always choose an origination effort $\tilde{a}(\mu, \theta)$ that is a best response. That is, if the cost of effort parameter is θ, there are μ dealers, and every originator choses effort $\tilde{a}(\mu, \theta)$, then an individual originator has no incentive to deviate from $\tilde{a}(\mu, \theta)$. Originators' best response is then given by the following first-order condition of their optimization problem with respect to a:

$$\psi_a(a - \underline{a}, \theta) = \frac{\mu}{a}\kappa(\upsilon - \rho)\left(\frac{1-a}{1-\mu}\right). \tag{19}$$

Lemma A.2 in the appendix establishes that the best response function $\tilde{a}(\mu, \theta)$ is smooth and increasing and concave in μ. Using the best response function $\tilde{a}$ we can state the social surplus optimization problem for a planner as

$$\max_{\mu} W(\mu, \theta) = \rho + \tilde{a}(\mu, \theta)(\upsilon - \rho) - \psi\left(\tilde{a}(\mu, \theta) - \underline{a}, \theta\right) - \int_0^{\mu} \varphi(u)du. \tag{20}$$

Lemma A.3 in the Appendix shows that the social surplus function is a strictly concave function of μ, which allows us to use first-order conditions to compare the measure of dealers that maximizes welfare with the measure of dealers that obtains in equilibrium, $\widehat{\mu}$. The strict concavity of $W(\mu, \theta)$ also insures that there exists a unique $\mu^S(\theta)$ that solves the planner's problem. Our main result is:

Proposition 6: *(a) Under Assumptions A1-A3 there exists a $\widehat{\theta}$ such that for $\theta > \widehat{\theta}$, $\mu^S(\theta) < \widehat{\mu}(\theta)$.*

(b) Let

$$\Phi\left(\theta\right) = \frac{\psi_{aa}(\hat{a}(\theta) - \underline{a}, \theta)}{\psi_a(\hat{a}(\theta) - \underline{a}, \theta)}(\hat{a}(\theta) - \underline{a}).$$

If $\Phi\left(\theta\right)$ is non-decreasing in θ then there exists a $\widehat{\theta}$ such that for $0 \leq \theta \leq \widehat{\theta}$, $\mu^S\left(\theta\right) \geq \widehat{\mu}(\theta)$, with equality if $\theta = \widehat{\theta}$, and for $\theta > \widehat{\theta}$ $\mu^S\left(\theta\right) < \widehat{\mu}(\theta)$.[15]

This proposition establishes that when the cost of originating good projects is large, the equilibrium number of dealers is excessive. Recall also that Corollary 5 states that in these circumstances infra-marginal dealers would have large rents. Thus, in the model, large rents by infra-marginal dealers are a *symptom* of an excessively large dealer sector.

We illustrate this result with the following numerical example, where originators have a quadratic effort function $\psi(a) = \frac{1}{2}\theta a^2$ (with $\underline{a} = 0$) and where $\varphi(d)$, κ, ρ and υ are given the following parameter values:

$$\varphi(d) = .1 + d; \quad \kappa = .35; \quad \rho = 2; \quad \upsilon = 2.5.$$

We vary θ over the interval $[0, 0.78]$, making sure that $\widehat{a} > \widehat{\mu}$.[16] Figure 1 plots the equilibrium $(\widehat{a}(\theta), \widehat{\mu}(\theta))$ and constrained optimum $(a^S(\theta), \mu^S(\theta))$ over this interval. The example illustrates all the main effects at work in the model. In particular, as shown in Proposition 4 origination incentives are decreasing with θ and the measure of informed dealers $\widehat{\mu}$ is increasing in θ.

Panel A illustrates that both $\widehat{a}$ and a^S are decreasing in θ, but $\widehat{a}$ declines more gradually than a^S because the planner fully internalizes the costs of information acquisition for dealers, while the equilibrium only partially reflects these costs. Thus, when the marginal cost of effort θ increases beyond 0.08 the planner responds by cutting the measure of dealers μ^S, as Panel B illustrates. The reason is that dealers cost the same but provide a weaker origination incentive when θ is higher. In contrast, the equilibrium measure of dealers $\widehat{\mu}$ is always *increasing* in θ, as the cream-skimming incentives for dealers increase with θ (see Proposition 4). Therefore, equilibrium origination effort $\widehat{a}$ declines less sharply than the socially optimal effort a^S. As seen in Proposition 6-b, for the value $\widehat{\theta} = .4$ the two schedules intersect at a unique θ and in region II the measure of informed dealers is above the constrained optimum, $\widehat{\mu} > \mu^S$, whereas it is below in region I.

Figure 1 around here

In sum, the equilibrium size of dealer markets is too small when θ is small and too large when θ is large. This is a striking and at first sight counterintuitive result. When θ is small the planner gets a large origination efficiency improvement by increasing the dealer market. This leads the planner to choose a relatively high μ^S and results in a high a^S, as can be seen in Figure 1. But when the fraction of good assets that is originated a^S is high the equilibrium incentives to enter the dealer market are small because the surplus that dealers can cream-skim is then relatively small. To see this consider the extreme case where $a^S \longrightarrow 1$. The price of an asset on the exchange p then converges to υ. But this means that in the limit there can be no cream-skimming for dealers and therefore almost no entry by rentiers into the dealer market. In contrast, when θ is large the planner gets a low origination response by increasing the measure of dealers and prefers a lower μ^S and a^S as is illustrated in Figure 1. But when a^S is low the incentives to enter the dealer market are large. To see this, note that when $a^S \longrightarrow 0$, $p \longrightarrow \max\{\rho - \frac{\mu}{1-\mu}(\upsilon - \rho), 0\}$, so that the total surplus from trading in the dealer market is large (it is close to υ when p is close to zero). As a result, a lot of rentiers will be induced to enter the dealer market.

D. Comparative Statics with respect to the Cost of Information

Our analysis in the previous subsection can be straightforwardly extended to consider the effects of changes in the cost of information for dealers. We model changes in information acquisition costs for dealers by specifying a cost function $\varphi(\mu, \eta)$ with parameter $\eta > 0$ and such that $\varphi_\eta(\mu, \eta) > 0$. By making the obvious changes to Assumptions A2 and A3 above we are able to establish an analog of Proposition 3.

Assumption A2′: (i) $\varphi_\mu > 0$; (ii) There exists $\bar{d} \leq \underline{a}$ such that $\lim_{\mu \to \bar{d}} \varphi(\mu, \eta) = \infty$, for each η;(iii) $\varphi(0, \eta) < (1-\kappa)(\upsilon - \rho)$ for each η;(iv) $\lim_{\eta \to \infty} \varphi(\mu, \eta) = \infty$ for each $d > 0$; (v) $\lim_{\eta \to 0} \varphi(\mu, \eta) = 0$ for any $\mu < \bar{d}$; and (vi) $\varphi_\eta > 0$.

Assumption A3′: $(1-\mu)\varphi(\mu, \eta)$ is non-decreasing in $\mu \in [0, \overline{d}]$ for every η.

The following is an analogue to the earlier proposition establishing comparative statics with respect to θ:

Proposition 7: *Assume that A1, A2′ and A3′ hold. Then* $\widehat{a}_\eta < 0$ *and* $\widehat{\mu}_\eta < 0$.

In words, the equilibrium origination of good assets and the equilibrium measure of dealers are both decreasing functions of the parameter η measuring the cost of information acquisition for a dealer. Recall that the net surplus of the marginal dealer $\widehat{\mu}(\eta)$ must be zero in equilibrium and hence that the rent of an infra-marginal dealer $\mu < \widehat{\mu}(\eta)$ equals $I(\mu, \eta) = \varphi(\widehat{\mu}(\eta), \eta) - \varphi(\mu, \eta)$. Therefore, the change in infra-marginal rents with respect to η is given by:

$$I_\eta(\mu, \eta) = \varphi_\mu \widehat{\mu}_\eta + \varphi_\eta(\widehat{\mu}(\eta), \eta) - \varphi_\eta(\mu, \eta).$$

It follows that a decrease in the cost parameter η increases the income of an infra-marginal dealer provided that $\varphi_{\eta\mu} \leq 0$. That is, provided that the cost reduction is at least as large for a more efficient dealer.

Corollary 8: *If* $\varphi_{\eta\mu} \leq 0$ *the rent of an infra-marginal dealer increases as* η *decreases.*

Proposition 7 and Corollary 8 establish that when the cost of information for dealers decreases, the equilibrium measure of dealers, the number of OTC transactions, and the income of every dealer increase. Perhaps then this is the source of exogenous variation behind the phenomenal growth and large rents of the financial services industry over the last two decades. Indeed, one important source of information cost savings in the past three decades is the spread of IT, which has facilitated the relative valuation of assets and lowered trading costs. As we highlight in Bolton et al. (2012), during this period of rapid spread of IT we have also witnessed abnormal growth in the size of OTC derivatives, swaps, commodities, and forward markets.[17] Also, as shown in Figure V Panel B from (Philippon and Reshef, 2012, p. 1569) the abnormal growth in median compensation in the financial industry since the early 1980s is driven in large part by the compensation of broker-dealers, which constitute the main entry in their 'other finance' category. Broker-dealers, of course, are the main players in OTC markets along with units inside commercial banks and insurance companies, such as AIG's infamous Financial Products group, which have been richly rewarded during the boom years prior to the crisis. Philippon (2015) has further shown that, unlike other sectors, financial intermediation has not become cheaper as a result of investments in IT technology. To the extent that all efficiency gains from IT investments are appropriated by dealers in the form of informational rents this is not entirely surprising.

Accordingly, a simple way of explaining the stylized facts associated with the growth in this sector could be in terms of the comparative statics of this section. Of course, our analysis is cast in a static model and cannot lend itself to a full dynamic explanation. Still, our simple comparative static exercises above can explain how this IT revolution has increased the number of dealers, which in turn has led to the higher compensation of dealers. [18]

Two specific examples provide a simple illustration of the role IT technology has played in financial innovation and customization. The first is commodities forward contracts, which have been increasingly used in recent years thanks to satellite imaging technology and IT applications such as Google Earth. Due to their more accurate geographic footprint, dealers can be more sure of the production capacity of producers, what in turn allows them to cream-skim the less risky counter-parties. The second is energy derivatives such as those offered by Enron Capital and Trade Resources (ECT) a subsidiary of Enron, which set up a "gas bank"–essentially a financial intermediary between buyers and sellers of natural gas–offering both price stability and local gas-supply and demand assurance. As (Tufano, 1996, p. 139) puts it "ECT's risk managers have clear instructions to develop a hedging strategy that minimizes net gas exposures, and the company has invested millions of dollars in hardware, software, and hundreds of highly trained personnel to eliminate mismatches and ensure that fluctuations in gas prices do not jeopardize the company's existence." It is worth recalling that before its eventual collapse, Enron, and in particular ECT, received numerous awards for these innovations.

IV. Conclusion

Our theory helps explain the simultaneous growth in the size of the financial services industry and the compensation of dealers in the most opaque parts of the financial sector. OTC markets emerge even in the presence of well functioning exchanges. The reason is that both originators and informed dealers have an incentive to meet outside the exchange: Originators with good assets may get better offers from informed dealers than are available on the exchange, and dealers can use their information to cream-skim good assets. Our model thus offers a novel theory of endogenous segmentation of financial markets, where "smart-money" investors deal primarily in opaque OTC markets to protect their information, and uninformed investors trade on organized exchanges.

In Bolton, Santos, and Scheinkman (2014) we consider an extension of our model where we allow for smart-money trading both on the organized exchange and on OTC markets. When financiers have a choice of becoming either informed traders on the exchange or informed dealers in an OTC market, we show that in equilibrium OTC markets are always too large relative to the organized exchange. The reason is that substitution of informed trading on a transparent exchange for trading on opaque OTC markets results in worse terms of trade for originators with good assets. Therefore, to maintain the same origination incentives of good assets a larger informed financial sector is required. Given that information rents are bigger in opaque OTC markets, financiers' private incentives are to switch trading to OTC markets even if this tends to undermine originators' ex-ante incentives to originate good assets, which explains why OTC markets are too large in equilibrium.

Informed dealers profit from the opaqueness of OTC transactions and this is one reason why broker-dealers have generally resisted the transfer of trading of the most standardized OTC contracts onto organized platforms, as required by the Dodd-Frank Act of 2010[19]. This is also why the largest Wall Street firms are so intent on avoiding disclosure of prices and fees in the new exchanges set up in response to the Dodd-Frank Act.[20] Interestingly, in a heterogeneous world, firms with a high probability of generating good assets also benefit (*ex-ante*) from the option of trading in opaque markets. It is, thus, not surprising that some firms have also been keen to keep OTC markets in their present form.[21] All in all, we therefore expect that a first line of defense by the financial industry to the new regulations required under the Dodd-Frank Act is likely to be to over-customize derivatives contracts and to offer fewer standardized, plain-vanilla, contracts (which will be required to trade on organized exchanges); the second line of defense will be to set up clearinghouses that maintain opacity and do not require disclosure of quotes; and a third line will be to ensure that the operation of clearinghouses remains under the control of the main dealers.

REFERENCES

Afonso, Gara, 2011, Liquidity and congestion, *Journal of Financial Intermediation* 20, 324–360.

Allen, Franklin, and Douglas Gale, 1998, Optimal financial crises, *The Journal of Finance* 53, 1245–1284.

Allen, Franklin, and Douglas Gale, 2000, *Comparing financial systems* (MIT press).

Baumol, William J., 1990, Entrepreneurship: Productive, unproductive, and destructive, *Journal of Political Economy* 98, pp. 893–921.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2015a, Equilibrium fast trading, Journal of Financial Economics (forthcoming).

Biais, Bruno, Jean-Charles Rochet, and Paul Woolley, 2015b, Dynamics of innovation and risk, *Review of Financial Studies* 28, 1353–1380.

Binmore, Ken, Ariel Rubinstein, and Asher Wolinsky, 1986, The nash bargaining solution in economic modelling, *The RAND Journal of Economics* 17, 176–188.

Board, Simon, and Moritz Meyer-ter-Vehn, 2014, Relational contracts in competitive labour markets, *The Review of Economic Studies* 82, 490–534.

Bolton, Patrick, Tano Santos, and Jose A Scheinkman, 2012, Shadow finance, in A Blinder, A Loh, and R Solow, eds., *Rethinking the Financial Crisis* (Russell Sage Foundation).

Bolton, Patrick, Tano Santos, and Jose A Scheinkman, 2014, Cream skimming in financial markets, Columbia University. Available on SSRN.

Bolton, Patrick, Tano Santos, and Jose A Scheinkman, 2015, Savings glut and financial fragility, manuscript, Columbia University.

Diamond, Douglas W, 1997, Liquidity, banks, and markets, *Journal of Political Economy* 105, 928–956.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen, 2005, Over-the-counter markets, *Econometrica* 73, 1815–1847.

Fecht, Falko, 2004, On the stability of different financial systems, *Journal of the European Economic Association* 2, 969–1014.

Fishman, Michael J, and Jonathan A Parker, 2012, Valuation, adverse selection, and market collapses, NBER Working Paper No. 18358.

Glode, Vincent, Richard C Green, and Richard Lowery, 2012, Financial expertise as an arms race, *The Journal of Finance* 67, 1723–1759.

Goldin, Claudia, and Lawrence F Katz, 2008, Transitions: Career and family life cycles of the educational elite, *The American Economic Review Papers & Proceedings* 98, 363–369.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *The American Economic Review* 70, pp. 393–408.

Harper, Christine, Matthew Leising, and Shannon Harrington, 2009, Wall Street Stealth Lobby Defends $35 billion Derivatives Haul, *Bloomberg* August 30th.

Hirshleifer, Jack, 1971, The private and social value of information and the reward to inventive activity, *The American Economic Review* 61, pp. 561–574.

Jacklin, Charles J., 1987, Demand deposits, trading restrictions, and risk sharing, in Edward C Prescott, and Neil Wallace, eds., *Contractual arrangements for intertemporal trade* (U of Minnesota Press).

Lagos, Ricardo, and Guillaume Rocheteau, 2009, Liquidity in asset markets with search frictions, *Econometrica* 77, 403–426.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill, 2011, Crises and liquidity in over-the-counter markets, *Journal of Economic Theory* 146, 2169–2205.

Leising, Matthew, 2009, ISDA hires Rosen to fight Obama OTC derivatives plan, *Bloomberg* July 10th.

Levine, Ross, 2005, Finance and growth: Theory and evidence, in Philippe Aghion, and Steven N. Durlauf, eds., *Handbook of Economic Growth*, volume 1, Part A, 865 – 934 (Elsevier).

Lewis, Michael, 2015, *Flash Boys: A Wall Street Revolt*, first edition (W. W. Norton & Company).

Morgenson, Gretchen, 2010, Fair Game: It's Not Over Until It's in the Rules, *The New York Times* August 28th.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny, 1991, The allocation of talent: Implications for growth, *The Quarterly Journal of Economics* 106, 503–530.

Nash, John, 1953, Two-person cooperative games, *Econometrica: Journal of the Econometric Society* 128–140.

Nash Jr, John F, 1950, The bargaining problem, *Econometrica: Journal of the Econometric Society* 155–162.

Philippon, Thomas, 2008, The evolution of the us financial industry from 1860 to 2007: Theory and evidence, NYU Stern Working Paper.

Philippon, Thomas, 2010, Financiers versus engineers: Should the financial sector be taxed or subsidized?, *American Economic Journal: Macroeconomics* 2, 158–82.

Philippon, Thomas, 2012, Finance vs. Walmart: Why are Financial services so Expensive?, in A Blinder, A Loh, and R Solow, eds., *Rethinking the Financial Crisis* (Russell Sage Foundation).

Philippon, Thomas, 2015, Has the us finance industry become less efficient? on the theory and measurement of financial intermediation, *American Economic Review* 105, 1408–38.

Philippon, Thomas, and Ariell Reshef, 2012, Wages and Human Capital in the U.S. Finance Industry: 1909—2006*, *The Quarterly Journal of Economics* 127, 1551–1609.

Rothschild, Michael, and Joseph Stiglitz, 1976, Equilibrium in competitive insurance markets: An essay on the economics of imperfect information, *The Quarterly Journal of Economics* 90, 629–649.

Santos, Tano, and Jose A. Scheinkman, 2001, Competition among exchanges, *The Quarterly Journal of Economics* 116, 1027–1061.

Scannell, Kara, 2009, Big Companies Go to Washington to Fight Regulations on Fancy Derivatives, *The Wall Street Journal* July 10th.

Stiglitz, Joseph E., 1989, Using tax policy to curb speculative short-term trading, *Journal of Financial Services Research* 3, 101–115.

Story, Louise, 2010, A Secretive Banking Elite Rules Trading in Derivatives, *The New York Times* December 11th.

Tett, Gillian, 2010, Calls for radical rethink of derivatives body, *Financial Times* August 26th.

Tufano, Peter, 1996, Who manages risk? an empirical examination of risk management practices in the gold mining industry, *The Journal of Finance* 51, 1097–1137.

Vayanos, Dimitri, and Tan Wang, 2007, Search and endogenous concentration of liquidity in asset markets, *Journal of Economic Theory* 136, 66 – 104.

Vayanos, Dimitri, and Pierre-Olivier Weill, 2008, A search-based theory of the on-the-run phenomenon, *The Journal of Finance* 63, 1361–1398.

A. Appendix

A.1. Proof of Propositions 3 and 4

Write $U(a|\tilde{a},\mu)$ for an entrepreneur's expected payoff in period 0 when the measure of dealers is μ and when all other entrepreneurs are choosing origination effort $\tilde{a}$. If $\widehat{p}\,(\tilde{a},\mu)$ ($p^d\,(\tilde{a},\mu)$) is the competitive (*resp.* OTC) price that would prevail if there are μ dealers and the average action is $\tilde{a}$ and if $\tilde{a} \geq \mu$ then

$$\begin{aligned} U(a|\tilde{a},\mu) &= \omega_e - \psi(a-\underline{a},\theta) + am\,(\tilde{a},\mu)\,p^d\,(\tilde{a},\mu) + (1-am\,(\tilde{a},\mu))\,\widehat{p}\,(\tilde{a},\mu) \\ &= \omega_e - \psi(a-\underline{a},\theta) + \frac{a\mu}{\tilde{a}}(\kappa\upsilon + (1-\kappa)\,\widehat{p}\,(\tilde{a},\mu)) + \left(1-\frac{a\mu}{\tilde{a}}\right)\widehat{p}(\tilde{a},\mu), \end{aligned} \tag{A.1}$$

Since the average action satisfies $\tilde{a} \geq \mu$, then

$$\widehat{p}(\tilde{a},\mu) = \rho + \frac{\tilde{a}-\mu}{1-\mu}(\upsilon-\rho) \tag{A.2}$$

Thus,

$$U(a|\tilde{a},\mu) = \omega_e - \psi(a-\underline{a},\theta) + \frac{a\mu}{\tilde{a}}\kappa(\upsilon-\rho)\left(\frac{1-\tilde{a}}{1-\mu}\right) + \rho + \frac{\tilde{a}-\mu}{1-\mu}(\upsilon-\rho) \tag{A.3}$$

Differentiating $U(a|\tilde{a},\mu)$ with respect to a and setting $a=\tilde{a}$ in the FOC for a

$$\psi_a(\tilde{a}-\underline{a},\theta) = \frac{\mu}{\tilde{a}}\kappa(\upsilon-\rho)\left(\frac{1-\tilde{a}}{1-\mu}\right) \tag{A.4}$$

which must characterize any interior symmetric equilibrium origination-effort response $\tilde{a}$ to a given measure of dealers $\mu \leq \tilde{a}$. Concavity guarantees that given μ, an $\tilde{a}$ that solves (A.4) is an equilibrium action.

On the other hand, if the average action is a and the number of dealers is $\tilde{\mu} \leq a$ then the gain of becoming informed by a dealer μ is given by:

$$\upsilon - p^d\,(a,\tilde{\mu}) - \varphi(\mu) = (1-\kappa)(\upsilon - \widehat{p}\,(a,\tilde{\mu})) - \varphi(\mu) \tag{A.5}$$

As in equation (A.2), since $a \geq \tilde{\mu}$,

$$\widehat{p}(a, \tilde{\mu}) = \rho + \frac{a - \tilde{\mu}}{1 - \tilde{\mu}}(\upsilon - \rho)$$

Since φ is strictly increasing, we obtain that given an average action $a > 0$ by originators, $\tilde{\mu}$ is an equilibrium number of dealers if and only if:

$$(1 - \tilde{\mu}(a))\varphi(\tilde{\mu}(a)) = (1 - \kappa)(\upsilon - \rho)(1 - a) \tag{A.6}$$

We are now ready to prove:

Lemma A.1: *Under Assumptions A1-A3, for each θ, there exists a unique equilibrium $(\widehat{a}(\theta), \widehat{\mu}(\theta))$ satisfying $\widehat{a}(\theta) \geq \widehat{\mu}(\theta) > 0$. In addition, $\widehat{a}_\theta < 0$ and $\widehat{\mu}_\theta > 0$.*

Proof: Consider the following system of equations:

$$\begin{aligned} \psi_a(a - \underline{a}, \theta) &= \frac{\mu}{a}\kappa(\upsilon - \rho)\left(\frac{1 - a}{1 - \mu}\right) && \text{(A.7)} \\ (1 - \mu)\varphi(\mu) &= (1 - \kappa)(\upsilon - \rho)(1 - a) && \text{(A.8)} \end{aligned}$$

As we argued above, for each θ, $(\widehat{a}(\theta), \widehat{\mu}(\theta))$ is an equilibrium if and only if it satisfies equations (A.7) and (A.8). Since, for each θ $\frac{a}{1-a}\psi_a$ is a strictly increasing function of a that maps $(\underline{a}, 1)$ onto $(0, \infty)$, there is a unique $a(\mu, \theta)$ that solves equations (A.7) with $\frac{\partial a}{\partial \mu} > 0$ and $\frac{\partial a}{\partial \theta} < 0$. Thus, using assumption A3 we obtain that $\frac{(1-\mu)\varphi(\mu)}{1-a(\mu,\theta)}$ is a strictly increasing function of μ and assumption A.2 guarantees that there is a unique $\widehat{\mu}(\theta) < \overline{d} \leq \underline{a}$ that solves

$$(1 - \mu)\varphi(\mu) = (1 - \kappa)(\upsilon - \rho)(1 - a(\mu, \theta))$$

Let $\widehat{a}(\theta) = a(\widehat{\mu}(\theta), \theta)$. The implicit function theorem applied to the system (A.7) and (A.8) guarantees that $\widehat{a}_\theta < 0$ and $\widehat{\mu}_\theta > 0$. □

A.2. Proof of efficiency results

Lemma A.2: *Under Assumption A1 the function $\tilde{a}(\mu, \theta)$ that solves equation (19) is smooth and concave in μ and such that $\tilde{a}_\mu > 0$.*

Proof: We already showed in the proof of Lemma A.1 that $\tilde{a}_{\mu}(\mu,\theta) > 0$. The implicit function theorem yields

$$\tilde{a}_{\mu}(\mu,\theta) = \frac{\kappa(\upsilon-\rho)(1-a)^2}{(1-\mu)^2(\psi_{aa}a(1-a)+\psi_a)} \tag{A.9}$$

Since $\psi_{aaa} \geq 0$,

$$\tilde{a}_{\mu\mu}(\mu,\theta) \sim -2(1-a)(1-\mu)^2(\psi_{aa}a(1-a)+\psi_a) -$$
$$(1-a)^2\left[-2(1-\mu)(\psi_{aa}a(1-a)+\psi_a) - (1-a)^2(1-\mu)^2[\psi_{aaa}\tilde{a}_{\mu}a(1-a)+(2-2a)\psi_{aa}\tilde{\mu}\right]$$
$$< 2(1-a)(1-\mu)(\psi_{aa}a+\psi_a)(\mu-a) \leq 0,$$

which completes the proof. □

Lemma A.3: *Under Assumptions A1-A3 the social surplus function $W(\mu,\theta)$ is a strictly concave function of μ.*

Proof: For any $\mu < \tilde{\mu}(\theta)$,

$$W_{\mu} = (\upsilon-\rho)\tilde{a}_{\mu}(\mu,\theta) - \psi_a(\tilde{a}a - \underline{a}(\mu,\theta),\theta)\tilde{a}_{\mu}(\mu,\theta) - \varphi(\mu)$$

and

$$W_{\mu\mu} = (\upsilon-\rho)\tilde{a}_{\mu\mu}(\mu,\theta) - \psi_{aa}(\tilde{a}_{\mu}(\mu,\theta))^2 - \psi_a(\tilde{a}-\underline{a}(\mu,\theta),\theta)\tilde{a}_{\mu\mu}(\mu,\theta) - \varphi_{\mu}(\mu)$$

Concavity follows from the concavity of $\tilde{a}$ with respect to $\mu$and $\psi_a a(\mu,\theta) < \kappa(\upsilon-\rho)$. □

Proof of Proposition 6: (a). Recall that, as seen in Proposition A.3, W is concave in μ. Then it is enough to evaluate W_{μ} at $\mu = \widehat{\mu}$, the equilibrium measure of informed dealers for a given θ and show that there exists a unique $\widehat{\theta} \in (\underline{\theta},\overline{\theta})$ such that $W_{\mu} > 0$ for $\theta \leq \hat{\theta}$ and $W_{\mu} < 0$ for $\theta > \hat{\theta}$. First using the first order condition (A.7) in (A.9) we can write

$$\tilde{a}_{\mu}(\mu,\theta) = \frac{\kappa(\upsilon-\rho)(1-a)}{(1-\mu)\left[a(1-\mu)\psi_{aa} + \frac{\mu}{a}\kappa(\upsilon-\rho)\right]} \tag{A.10}$$

The derivative of W with respect to μ, evaluated at $\mu = \widehat{\mu}$ is

$$W_{\mu}(\widehat{\mu}(\theta),\theta) = [(\upsilon-\rho)-\psi_a]\,\tilde{a}_{\mu}(\widehat{\mu}(\theta),\theta) - \varphi(\widehat{\mu}(\theta)). \tag{A.11}$$

Note that equation (A.8) and the hypothesis that for each $a > \underline{a}$, $\lim_{\theta\to\infty} \psi_a(a - \underline{a}, \theta) = \infty$ implies that as $\theta \to \infty$ the equilibrium action $\widehat{a}$ converges to $\underline{a}$. Equation (A.8) and assumption A3 guarantee that as $\theta \to \infty$, $\widehat{\mu}(\theta)$ approaches (monotonically, by Proposition 4) its maximum value. Equation (A.8) thus implies that $\psi_a(\widehat{a} - \underline{a}, \theta)$ is bounded below by a positive constant δ. Since ψ_a is convex,

$$0 < \delta \leq \psi_a(a - \underline{a}, \theta) \leq \psi_{aa}(a - \underline{a}, \theta)(a - \underline{a}).$$

Thus as $\theta \to \infty$, $\psi_{aa}(\widehat{a} - \underline{a}, \theta) \to \infty$, and equation (A.10) implies that $\tilde{a}_\mu(\widehat{\mu}, \theta) \to 0$. From (A.11), $W_u(\widehat{\mu}(\theta), \theta)$ is negative for θ large.

(b). Evaluating (A.7), (A.8) and (A.10) at $\mu = \widehat{\mu}$ substituting in (A.11) and noting that $\widehat{a}(\theta) = a(\widehat{\mu}(\theta), \theta)$ we obtain

$$\begin{aligned} W_\mu(\widehat{\mu}(\theta)) &= (\upsilon - \rho)\left[1 - \kappa\frac{\widehat{\mu}(\theta)(1 - \widehat{a}(\theta))}{\widehat{a}(\theta)(1 - \widehat{\mu}(\theta))}\right] \frac{\kappa(\upsilon - \rho)\frac{1-\widehat{a}(\theta)}{1-\widehat{\mu}(\theta)}}{\widehat{\mu}(\theta)(1 - d)\psi_{aa} + \frac{\widehat{\mu}(\theta)}{\widehat{a}(\theta)}\kappa(\upsilon - \rho)} \\ &- (1 - \kappa)(\upsilon - \rho)\frac{1 - \widehat{a}(\theta)}{1 - \widehat{\mu}(\theta)} \\ &= (\upsilon - \rho)\left(\frac{1 - \widehat{a}(\theta)}{1 - \widehat{\mu}(\theta)}\right)[\Omega(\theta) - (1 - \kappa)], \end{aligned} \tag{A.12}$$

where

$$\Omega(\theta) = (\upsilon - \rho)\left[1 - \kappa\frac{\widehat{\mu}(\theta)(1 - \widehat{a}(\theta))}{\widehat{a}(\theta)(1 - \widehat{\mu}(\theta))}\right] \frac{\kappa}{\widehat{a}(\theta)(1 - \widehat{\mu}(\theta))\psi_{aa} + \frac{\widehat{\mu}(\theta)}{\widehat{a}(\theta)}\kappa(\upsilon - \rho)}. \tag{A.13}$$

The sign of $W_\mu(\widehat{\mu}(\theta))$ is thus determined by $\Omega(\theta) \gtrless 1 - \kappa$. We now show that $\Omega_\theta < 0$. First notice that we can write the function $\Omega(\theta)$

$$\Omega(\theta) = \left(\frac{\widehat{a}}{\widehat{\mu}(\theta)}\right)\left[1 - \kappa\frac{\widehat{\mu}(\theta)(1 - \widehat{a})}{\widehat{a}(1 - \widehat{\mu}(\theta))}\right]\left(\frac{1}{1 + (1 - \widehat{a})\Lambda(\theta)}\right), \tag{A.14}$$

where we have used expression (A.7) and defined

$$\Lambda(\theta) = \frac{\psi_{aa}}{\psi_a}a. \tag{A.15}$$

Notice that the assumption in Proposition 6-b implies that $\Lambda_\theta > 0$. In addition, Assumption A1 and Proposition 4 guarantee that

$$\frac{\partial}{\partial\theta}\left(\frac{1}{1+(1-\widehat{a})\Lambda\left(\theta\right)}\right)<0 \qquad \frac{\partial}{\partial\theta}\left(\frac{\widehat{a}}{\widehat{\mu}(\theta)}\right)<0 \quad \text{and} \quad \frac{\partial}{\partial\theta}\left[1-\kappa\frac{\widehat{\mu}(\theta)(1-\widehat{a})}{\widehat{a}(1-\widehat{\mu}(\theta))}\right]<0,$$

which implies that $\Omega_\theta < 0$. □

A.3. Proof of Proposition 7

The proof of Propositions 7 parallel the proofs in section A.A.1, by considering instead of equations (A.7) and (A.8), the system:

$$\psi_a(a) = \frac{\mu}{a}\kappa(\upsilon-\rho)\left(\frac{1-a}{1-\mu}\right) \tag{A.16}$$

$$(1-\mu)\varphi(\mu,\eta) = (1-\kappa)(\upsilon-\rho)(1-a) \tag{A.17}$$

Notes

[1]Goldin and Katz (2008) document that the percentage of male Harvard graduates with positions in Finance 15 years after graduation tripled from the 1970 to the 1990 cohort, largely at the expense of occupations in law and medicine.

[2]Rothschild and Stiglitz (1976) considered a different form of cream skimming in insurance markets with adverse selection. In that setting, insurers are uninformed about risk types, but offer contracts that induce informed agents to self-select into insurance contracts. For an application of the Rothschild-Stiglitz framework to competition among organized exchanges see Santos and Scheinkman (2001).

[3]Other applications than private placements versus initial public offerings, such as derivatives markets, are also discussed in Bolton et al. (2012).

[4]The literature comparing bank-based and market-based financial systems argues that bank-based systems can offer superior forms of risk sharing, but that they are undermined by competition from securities markets (see Jacklin (1987); Diamond (1997); Fecht (2004)). This literature does not explore the issue of misallocation of talent to the financial sector, whether bank-based or market-based.

[5]See Duffie, Gârleanu, and Pedersen (2005); Vayanos and Wang (2007); Vayanos and Weill (2008); Lagos and Rocheteau (2009); Lagos, Rocheteau, and Weill (2011); Afonso (2011).

[6]We have also explored the more general specification where

$$U(c_1, c_2) = \delta_o c_1 + (1 - \delta_o)c_2, \qquad \text{and} \qquad u(c_1, c_2) = \delta_d c_1 + (1 - \delta_d)c_2,$$

for originators and dealers respectively. In the above expression $\delta_o \in \{0, 1\}$ and $\delta_d \in \{0, 1\}$ are indicator variables with prob $(\delta = 1) = \pi_o \in (0, 1)$ and and prob $(\delta_d = 1) = \pi_D \in (0, 1)$. All the results in this paper are unchanged under this more general specification.

[7] We do not assume an explicit matching protocol. One possibility is that dealers inspect all projects, then dealers are put in a line and pick only one project, then dealers bargain with the entrepreneur. For a similar protocol in a job-search context see Board and Meyer-ter-Vehn (2014)

[8]For a similar approach to modeling negotiations in OTC markets between dealers and clients

see Lagos et al. (2011).

[9]In Bolton, Santos, and Scheinkman (2015) we explore the implications of the model when the measure of uninformed rentiers is not sufficient to absorb asset sales on the exchange, so that cash-in-the-market pricing obtains as in Allen and Gale (1998).

[10]Cream skimming is the key mechanism behind our results. In Bolton et al. (2012) we show that these results remain unaffected if instead dealers trade on an exchange as in Grossman and Stiglitz (1980), with informed trades hiding behind noise trades. As long as the pool of assets available to uninformed investors, and therefore asset prices, are adversely affected by the cream skimming activities of informed investors our general results obtain.

[11]In Bolton et al. (2012) we also explore the more realistic situation where dealers come from the ranks of potential originators rather than from the pool of rentiers. There is then an even higher cost of letting potential originators undertake a career in finance: Not only is the cost of acquiring dealer skills wasted, but also the lost expected output that would have been produced in the real sector.

[12]In line with Biais et al. (2015a), our model can then be interpreted as an illustration of *front-running* through, say, high frequency trading. By this interpretation dealers are high frequency traders who, sometimes at great cost and ingenuity, can find out in advance the price at which an asset available for purchase will sell at a later date (sometimes just a few milliseconds later). The dealers then purchase the assets that they know will see a price improvement (the good assets) so as to sell them at a later date and realize an easy capital gain. Michael Lewis refers to this type of activity as "scalping" and remarks that one of the unfortunate consequences of the rise in high frequency trading is all the talent it has attracted that might have been deployed more productively elsewhere: "The new players in the financial markets, the kingpins of the future who had the capacity to reshape those markets, were a different breed: the Chinese guy who had spent the previous ten years in American universities, the French particle physicist from FERMAT lab; the Russian aerospace engineer; the Indian PhD in electrical engineering. "There were just *thousands* of these people" said Schwall. "Basically all of them with advanced degrees. I remember thinking to myself how unfortunate it was that so many engineers were joining these firms to exploit investors rather than solving public problems."" [(Lewis, 2015, p. 121) *Flash Boys*, Norton, New York]

[13] As is well known, in the classic bargaining models a la Rubinstein (1982) that build on the outside option principle of BRW, there is generally no effect on the bargaining power κ of changes in the measure μ. Another way of modeling competition could be to let originators engage in costly search for dealers and when another dealer is found to let the parties engage in bilateral bargaining. Greater competition among dealers in this more elaborate model can then be captured through changes in the probability of finding a new dealer conditional on a new search. The price on the exchange p, which is always available to the originator, acts as an additional outside option in this search and bargaining game, one that is more attractive to the originator if search is more costly or if the discount rate is higher. What obtains in this model is a general bargaining outcome $p^d = \kappa x_h + (1-\kappa)V(p,\delta)$, where $V(p,\delta)$, the reservation value of the originator, is a function of both the price on the exchange and the probability δ of finding another dealer in an additional search. Then, as long as $V_p > 0$, a rather natural assumption, all our results carry through in this more elaborate model.

[14]In Bolton et al. (2014) we analyze the model without this assumption and also consider the case where $\widehat{\mu} \geq \widehat{a}$.

[15]Note that the assumption on $\Phi(\theta)$ is met if, for instance, $\psi(a-\underline{a},\theta) = \theta(a-\underline{a})^n$.

[16] Indeed note that in this example Assumption A2(iii) is not met, as $\underline{a} = 0$. However, under our specific parametric assumptions we have $\widehat{\mu}(\theta) < \widehat{a}(\theta)$ for $\theta \in [0, .78]$ and thus $\widehat{m}(\theta) \in (0,1)$. Bolton et al. (2014) explore this example also for values of θ for which $\widehat{\mu} > \widehat{a}$.

[17]In particular, Figure 2 in Bolton et al. (2012) shows that in 1998-2011 OTC contracts for interest rate derivatives grew by a factor of 14, while exchange traded interest rate futures grew by a factor of 3. Commodity forwards and futures display a similar pattern, except that the volume in OTC commodity contracts collapsed after the crisis.

[18]Deregulation, which some commentators (*e.g.* Philippon and Reshef (2012)) suggest was responsible for the phenomenal growth of the financial services industry in the past quarter century, would have a similar effect: the decrease in costs in OTC activities would generate a larger OTC market, higher compensation for dealers, and lower ex-ante profits for entrepreneurs.

[19]The lobbying activity of banks to avoid any major changes in the organization of OTC markets has been amply documented in the press. See for example Leising (2009); Morgenson (2010); Tett (2010). Harper, Leising, and Harrington (2009) write:“ [T]he banks ... are expected to lobby to

remove any requirements that the contracts be executed on exchanges because that would cut them out of making a profit on the trades, according to lawyers working for the banks."

[20]See for example Story (2010), who reports on the efforts by the largest banks to thwart an initiative by Citadel, the Chicago hedge fund, to set up an electronic trading system that would display prices for CDSs.

[21]See Scannell (2009), who writes "Companies from Caterpillar Inc. and Boeing Co. to 3M Co. are pushing back on proposals to regulate the over-the-counter derivatives market, where companies can make *private deals* to hedge against sudden moves in commodity prices or interest rates". (Emphasis ours).

Figure 1. Equilibrium effort, $\widehat{a}$, measure of informed intermediaries $\widehat{\mu}$, together with the constrained optimum $\left(a^S, \mu^S\right)$ (dashed lines) as a function of θ.