ISSN 0956-8549

Market Structure of Intermediation

By

Hongda Zhong

FINANCIAL MARKETS GROUP DISCUSSION PAPER NO. 968

July 2026

Market Structure of Intermediation*

Hongda Zhong†

December 18, 2023

Abstract

Skilled intermediaries can represent multiple investors, generating economies of scale in screening costs. However, locating competent intermediaries recreates the screening friction. I provide a framework to study the structure of intermediation chain, focusing on its length, sector size, skill distribution, and effort choice. Counterintuitively, efficient intermediation does not rely on intermediaries being more skilled. Longer chains enhance welfare and reduce the market power of the intermediary sector. Moreover, self-selection of intermediaries into two heterogeneous layers can lead to greater efficiency compared to structures with homogeneous qualities across layers. These findings have implications for delegated asset management industry.

Key Words: Intermediation Chain, Screening, Asset Management

JEL Codes: D40, L10, G20

*I thank Samuel Antill, Ulf Axelson, Jonathan Berk, Philip Bond, Mike Burkart, Georgy Chabakauri, Amil Dasgupta, Daniel Ferreira, Alessandro Gavazza, Vincent Glode, Zhiguo He, Kinda Hachem, Peter Kondor, Mina Lee, Dong Lou, Martin Oehmke, Christian Opp, Lasse Pedersen, Dimitri Vayanos, Sergio Vicente, Ansgar Walther, Andrew Winton, Shengxing Zhang, Jeffrey Zwiebel, and conference/seminar participants at the London School of Economics, Swiss Finance Institute/EPFL, Católica Lisbon School of Business & Economics, UC-Irvine, University of Minnesota, University of Colorado — Boulder, Copenhagen Business School, University of Rochester, University of Texas — Dallas, University of Houston, HKUST, Stanford University, Finance Theory Group (summer conference in Budapest and members' meeting in Michigan), OxFIT Conference, Banking Theory Brown Bag Seminar, FIRS conference, WFA conference, CICF conference, and European Summer Symposium in Financial Markets. Yang Guo and Jianing Yuan provided excellent research assistance. I gratefully acknowledge the support of the Economic and Social Research Council (ESRC New Investigator Grant: ES/T003758/1). All errors are my own.

†University of Texas at Dallas and CEPR, hongda.zhong@utdallas.edu

1 Introduction

Intermediation chains are vital in facilitating transactions across various markets. These chains consist of upstream intermediaries who screen and select downstream intermediaries, creating a layered structure. For example, direct investment is time consuming and challenging, and investors therefore often rely on financial intermediaries to access the capital market. They select financial advisors (layer 1), who recommend pension funds or fund-of-funds (layer 2), which then choose hedge funds or private equity (PE) funds (layer 3), ultimately selecting underlying assets.

Despite their prevalence, intermediation chains pose intriguing questions and lack comprehensive theoretical guidance in the literature. For example, there are almost as many equity funds as the number of public companies in the U.S..[1] If investors struggle to pick a portfolio of underlying stocks, how do they navigate the selection of funds? Additionally, why do numerous equity index mutual funds, as much as 42% (Jiang, Vayanos, and Zheng, 2020), track the S&P 500 index, offering essentially identical services? Finally, why do we witness the existence of lengthy intermediation chains, and what trade-offs do they entail?[2]

This paper constructs a novel framework based on matrix algebra to examine some key factors of intermediation chains, including chain length, sector size, skill distribution, and effort choices within each layer. To illustrate, consider a simplified example:

Introductory Example

Imagine an economy with $N_0 = 1$ mass of atomistic investors. Each investor needs a portfolio. For simplicity, the quality of the portfolio is binary. A fraction n_P of the them are good, providing a payoff of 1 when invested, while others are bad, offering no payoff. A fraction n_0 of the investors can skillfully screen the financial market to find good portfolios, while unskilled investors can only randomly select. Investor welfare W_0 is given by

$$W_0 = n_0 + (1 - n_0)\, n_P.$$

[1] By the end of 2017, the number of U.S. listed companies was 4,336 compared to 4,234 U.S. equity investment funds (including exchange-traded funds and mutual funds). See Detrixhe (2018).

[2] An illustrative example of a long intermediation chain in the financial market can be observed in the shadow banking credit intermediation chain, as highlighted by Adrian et al. (2012) and He and Li (2022).

The maximum potential welfare is 1, achieved when all investors invest in good portfolio. The source of welfare loss $(1 - W_0 = (1 - n_0)(1 - n_P))$ arises from the $1 - n_0$ unskilled investors' inability to find good portfolios.

Next, a single layer of intermediaries is introduced ($I = 1$). Each investor selects an intermediary (financial advisor) who, in turn, selects a portfolio from the market. Similar to the investors, a fraction n_S of the intermediaries are skilled and capable of finding good portfolios, while the rest are unskilled and choose portfolios at random. Skilled investors can leverage their ability to identify a skilled intermediary, while unskilled investor randomly select an intermediary. (The ability of the skilled investors to screen intermediaries is not crucial in a longer chain $I > 1$.) Welfare in this case is calculated as follows:

$$W_1 = n_0 + (1 - n_0)\left[n_S + (1 - n_S)\, n_P\right] = n_0 + (1 - n_0)\left[1 - (1 - n_S)(1 - n_P)\right].$$

Feature 1 [Efficient Intermediation Does Not Require Superior Skill]*: For any quality of the intermediary sector $n_S \in (0, 1)$, intermediation improves welfare: $W_1 > W_0$ for any $n_0, n_P \in (0, 1)$.*

This feature may seem counter-intuitive. In particular, even if the intermediary sector is less skilled than the investors on average ($n_S < n_0$), and has lower quality than the asset market ($n_S < n_P$), it remains more efficient to have an intermediary sector.

The reasoning behind this seemingly paradoxical feature is straightforward. Skilled investors can obtain good a portfolio independently or through a skilled intermediary, leaving their payoff unaffected. Unskilled investors, however, receive a higher payoff due to the probability n_S that they can encounter skilled intermediaries capable of delivering good portfolios. Even if they encounter unskilled intermediaries (probability $1 - n_S$), these intermediaries provide random portfolios as the unskilled investors would have obtained independently. Consequently, the n_S fraction of skilled intermediaries increases the chances for unskilled investors to find good portfolios.

Now, suppose there are I layers of homogeneous intermediaries. A fraction n_S is skilled in each layer. Investors select intermediaries in layer 1; intermediaries in layer i choose those in the subsequent layer; and intermediaries in the final layer select the portfolios.

The skilled intermediaries effectively create an absorbing state: They consistently locate skilled intermediaries in the following layers until the final layer delivers good portfolios. Similarly to the single-layer case ($I = 1$), skilled investors can receive good portfolios through a sequence of skilled intermediaries. Unskilled investors eventually receive bad portfolios with a probability of $(1 - n_S)^I (1 - n_P)$, that is they repeatedly find unskilled intermediaries throughout the I layers and ultimately receive a bad portfolio. The welfare in the economy is therefore

$$W_I = n_0 + (1 - n_0) \left[1 - (1 - n_S)^I (1 - n_P)\right].$$

Note that $I = 0$ and $I = 1$ nest the previously calculated W_0 and W_1 as special cases.

Feature 2 [Longer Chains Improve Efficiency]*: For any $n_S, n_0, n_P \in (0, 1)$, W_I is increasing in I and approaches the first best*

$$\lim_{I \to \infty} W_I = 1.$$

Welfare increases as the chain lengthens because unskilled investors have a higher probability of obtaining good portfolios. Each additional layer provides them with another opportunity to move towards the absorbing state created by the skilled intermediaries. Importantly, this feature holds even when investors lack the ability to screen intermediaries, for example, when all investors are unskilled ($n_0 = 0$) or, equivalently, when skilled investors cannot redirect their skills to screen intermediaries.

Feature 3 [Diminishing Marginal Benefit of a Layer]*: As the intermediary chain lengthens, the marginal welfare contribution of one additional layer diminishes*

$$\lim_{I \to \infty} W_I - W_{I-1} \to 0.$$

This feature immediately follows from the previous convergence property (even when the limit is less than 100%). The diminishing effect becomes particularly relevant when market power of the intermediary sector is introduced later in the paper. It implies that a longer chain lowers the market power for the entire intermediary sector.

Additional Ingredients and Roadmap of the Paper

Building on these features, the baseline model in Section 2 considers three additional factors. First, the success probability of skilled intermediaries may not reach 100% and may require effort input. Second, effort incentives may depend on the compensation received by each intermediary sector. Third, intermediaries may exert market power when setting compensation in equilibrium. Finally, Section 4 studies intermediaries self-selection into different layers based on their skill levels and the entry cost of each layer.

Several additional insights emerge from the baseline model. First, effort increases in the downstream of the intermediation chain (Subsection 3.1). The convergence property in Feature 2 renders investors' as well as upstream intermediaries' screening outcome inconsequential, as they can rely on the screening effort of subsequent layers. In contrast, intermediaries in the final layers are highly motivated to exert effort as their selection directly affects the delivery of good portfolios. Second, longer chains diminish the market power of the entire intermediary sector, potentially eroding the effort incentive along the chain. This creates a tradeoff between achieving probability convergence and providing an effort incentive (Subsection 3.2). Based on this tradeoff, one can derive the optimal chain length I^* for a given investor size N_0. Third, as the investor sector (N_0) grows, an asymptotically efficient structure is established (Subsection 3.3). This structure features a chain length that scales with the number of investors in a power-law manner ($I = N_0^\alpha$, where $\alpha > 0$ possibly small) and a logarithmic order of effort-making layers ($I_E = \beta \log N_0$) at the tail of the chain. Finally, when skilled intermediaries can only imperfectly screen, a simple condition is derived (Subsection 3.4) to determine whether intermediation outperforms direct market access for investors.

In cases where intermediaries can self-select into layers based on their skill types, a two-layer intermediation chain ($I = 2$) can enhance both investor payoffs and total welfare to an efficient level. The corresponding chain structures display heterogeneous skill distributions across the two layers. One high-quality layer is exclusively composed of skilled intermediaries, while the remaining intermediaries (both skilled and unskilled) enter the other layer. Under such structures, investors enjoy a high probability of finding good portfolios without the need for individual screening efforts. Other structures with mixed-quality intermediaries in both layers are less efficient, and the bimodal-efficiency feature remains robust across different effort specifications and screening technologies (Subsection 4.4).

It is worth noting that the effort patterns differ based on whether the first or second layer is of high quality. If the first layer is high quality (Subsection

4.2), intermediaries in both layers need to intensively screen. Conversely, if the second layer is of high quality (Subsection 4.3), only those in the second layer engage in rigorous screening in the asset markets, while no additional effort is exerted along the chain. Different structures can emerge as equilibrium outcomes depending on entry costs. An example of the first structure is the bank (layer 1) and loan originator (layer 2) chain, while some relevant examples featuring a high-quality second layer include the fund-of-funds (layer 1) and underlying PE or hedge fund (layer 2) chain.

As an extension and robustness analysis, I explore the scenario of multiple asset markets, as outlined in Subsection 5.1. In this case, intermediation introduces what I call "two-sided economies of scale," and this effect is best illustrated within a single-layer setting (Diamond 1984). On the one hand, skilled intermediaries exert substantial effort to screen a wide range of markets on behalf of investors, and these effort costs are shared by a large number of investors. On the other hand, skilled investors exert significant effort in identifying skilled intermediaries, and their effort costs are mitigated by the broad array of markets that skilled intermediaries can cover for them. Other extensions to the analysis encompass a more generalized compensation package (Subsection 5.2) and scenarios with more than two skill types (Subsection 5.3).

Literature Review

The literature on intermediation encompasses various perspectives and explanations for the existence of middlemen in different markets. For instance, middlemen can play a crucial role in matching buyers and sellers effectively (Rubinstein and Wolinsky 1987). In the presence of informational frictions, middlemen may have stronger incentives to become experts and provide quality certification due to their extensive transaction volume (Biglaiser 1993, Biglaiser and Friedman 1994, Lizzeri 1999). In financial markets, financial intermediaries, such as banks, naturally emerge as delegated monitors on behalf of small depositors (Diamond 1984). Market makers serve as intermediaries who match buyers and sellers in the securities market and absorb excess supply or demand (Glosten and Milgrom 1985). Intermediation can also arise when an informed party wishes to sell information through a third party, such as mutual funds (Admati and Pfleiderer 1990), or when agents collaborate to produce information, such as index providers or asset managers (Ramakrishnan and Thakor 1984, Gârleanu and Pedersen 2018). While each paper presents interesting explanations for intermediation, many of these models simplify the market structure by featuring only a single layer of intermediaries. Instead of providing another rationale for intermediation, my focus is on the market structure of intermediation chains, offering predictions on the chain length, sector size, skill distribution, and effort choices. Even in the single-layer context, many of existing stories rely on the intermediaries being more skilled than their clients, whereas my model shows that superior skill is not a necessity for efficient intermediation.

Several recent papers have also explored intermediation chains. Glode and Opp (2016) and Glode, Opp, and Zhang (2019) demonstrate that intermediaries mitigate informational frictions within each trading layer, increasing the likelihood of efficient transactions. When the chain is sufficiently long, the information asymmetry within each layer diminishes, allowing for fully efficient trades. He and Li (2022) illustrate how a credit chain formed by intermediaries facilitates firms with long-term projects in accessing cheaper funding from overlapping-generation households in the form of short-term debt. While intermediation reduces the inefficiency associated with trading debt claims among households, it introduces rollover risks for intermediaries. The optimal length of the credit chain is endogenously determined, striking a balance between the trade-offs. From a different perspective, Dasgupta and Maug (2021) argue that excessive layers of inefficient delegation can arise when decision makers have reputation concerns and prefer to pass decisions to a third party, preserving their reputation in case of failure. In addition to offering a distinct mechanism from these studies, my model provides a novel insight that even short chains can achieve full efficiency. Among many other differences, I demonstrate that efficiency is typically achieved through the final layers of the chain, and motivating effort in the initial layers proves to be both difficult and unnecessary.[3]

While most of the aforementioned papers take the participants in intermediary sectors as fixed, I draw inspiration from Kaniel and Orlov (2020) who study how agents with privately observed skill levels decide to join and leave the (single-layer) intermediary sector. In my work, I also examine the endoge-

[3]This feature is in sharp contrast to models without effort choice, such as Glode and Opp (2016), Glode, Opp, and Zhang (2019), and He and Li (2022), where each layer resolves a fraction of the inefficiencies gradually.

nous participation of agents in the intermediary sectors, but I abstract from the contracting problem between the agent and intermediary sectors. Instead, I focus on the self-selection of agents into different layers of the intermediation chain.

Berk and Van Binsbergen (2022) investigate the presence of charlatans in high-skill professions, similar to the unskilled intermediaries in my model. In their single-layer intermediation model, they find that the existence of charlatans may improve investor welfare by increasing competition for skilled agents, thereby making prices more competitive. My work complements their findings by considering the distribution of charlatans along the intermediation chain and offering insights into their impact on the overall efficiency. Related to the applications of the model, in context of the delegated asset management, Gennaioli, Shleifer, and Vishny (2015) show that delegation to a less skilled fund manager may be possible due to the trust that investors place in the fund managers. My paper offers a rational story for this phenomenon.

Finally, the layered structure of the model lends itself to the literature on corporate hierarchy, including Calvo and Wellisz (1979), Qian (1994), Chen (2017), and Garicano (2000). This literature study the efficient distribution of labor force across different layers in order to achieve better monitoring or knowledge sharing. Interpreting my model in this context, I study a different underlying friction — to select the correct (skilled) workers for production. Novel to the literature and relevant in practice is that each layer contains agents with different types. For example, CEOs need to select the suitable high-level managers who in turn needs to select the suitable lower-level managers who in turn select skilled workers. The results also provide new insights. A multi-layer structure can improve the success probability up to an endogenous limit. Simultaneously, each layer becomes less critical, lowering the total wage bill.

2 Baseline Model

2.1 Model Setup

Consider an economy with N_0 atomistic investors. Each investor needs an investment portfolio. There is also a unit mass of intermediaries forming an I-layer chain between the investors and the asset market. The term "player" is used generically to refer to both investors and intermediaries. The special case where $I = 0$ corresponds to investors directly invest in the asset market without any intermediaries.

Investors receive portfolios in the following process. Each investor chooses an intermediary from layer 1, who in turn chooses an intermediary from layer 2, and this process continues until the players in the final layer I (or investors if $I = 0$) choose portfolios. Note that "choosing an intermediary" does not necessarily mean selecting only one, but rather choosing a "portfolio" of intermediaries with certain average quality. As will be clear, the average quality (skill level) of the selected (set of) intermediary is the only relevant variable.

For simplicity, I consider binary quality types, that is some of the players are skilled and similarly, some of the portfolios are good. I broadly refer to the fraction of good portfolios/skilled players in layer i as the quality of that layer, denoted by n_i. As a notational convention, the investor sector is indexed layer 0 and the asset market is indexed layer $I+1$, with n_0 and $n_{I+1} \equiv n_P$ denoting the fraction of skilled investors and good portfolios. For now, I rule out the self-selection of intermediaries into different layers (which will be the key consideration in Section 4) and simply assume each layer $i = 1, 2, ..., I$ has a homogeneous structure. The sector size $N_i \equiv \frac{1}{I}$ and the quality $n_i \equiv n_S$, where n_S is the total number (equivalently fraction) of skilled intermediaries.

A good portfolio generates 1 unit of payoff and a bad one generates 0. One can interpret "good portfolio" as one with superior risk-return tradeoff, and the associated payoff as abnormal returns. Due to the scope of the paper, I assume away the pricing aspect of portfolios and assume that they are non-rival, meaning that multiple investors can choose good portfolios without affecting their payoffs.[4] Skilled players in layer $i \in \{0, 1, ..., I\}$ have screening technologies: They can exert effort e_i and incur cost $c_i(e_i)$ to find skilled players or good portfolios in the next layer with an improved probability $P_i(e_i, n_{i+1})$. The remaining unskilled players do not possess this technology. They randomly choose portfolios or intermediaries in the next layer, and find a good (skilled) one with the unconditional probability n_{i+1}. To simplify exposition in the baseline model, I assume binary effort level $\{e, 0\}$ with $P_i(e, n_{i+1}) = p_e \in (n_{i+1}, 1]$ and $P_i(0, n_{i+1}) = n_{i+1}$. I normalize $c(0) = 0$ and write $c_e = c(e)$. Later in Sec-

[4] Essentially, all players in the economy are price takers, and there are sufficient supply of good portfolios. The key friction in the market is the difficulty to locate the good portfolios amid the uncertainty in quality.

tion 4, I allow effort choice e_i to be continuous. Figure 1 illustrates a typical structure in this intermediated economy.

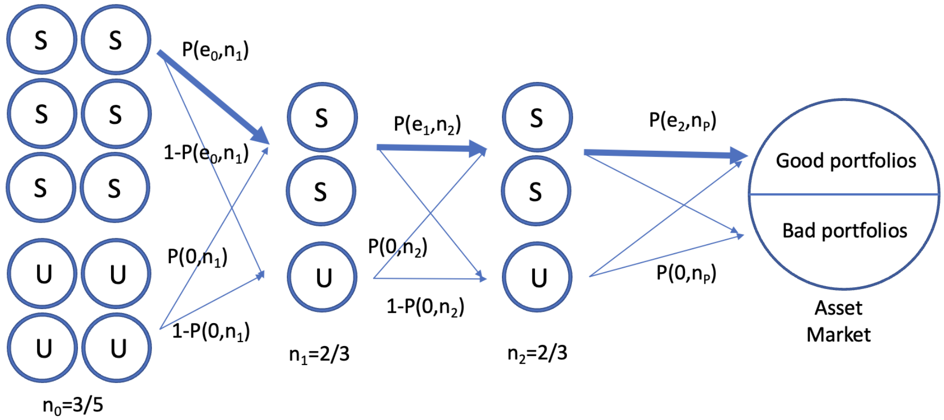

Figure 1 illustrates the structure of the economy. The abbreviations "S" and "U" represent skilled and unskilled players. In this example, $n_0 = \frac{3}{5}$ because 6 out of the 10 investors are skilled. Similarly, in both layers 1 and 2, $n_1 = n_2 = 2/3$. The probabilities associated with each arrow indicate the likelihood of the corresponding outcome. For instance, the probability that a skilled intermediary in layer 1 finds a skilled intermediary in layer 2 is given by $P(e_1, n_2)$. For readability, some possibilities are omitted in the figure, whereas thicker arrows indicate more likely outcomes due to effort input.

Intermediaries effort incentive comes from the performance pay r_i for delivering good portfolios to investors.[5] Investors therefore receive a net payoff

$$r_0 = 1 - \sum_{i=1}^{I} r_i$$

from investing in a good portfolio.[6] There are different ways to introduce market power to the multi-layered intermediary sectors, and they are essentially different ways to endogenize r_i which captures how the total surplus is being shared among investors and each intermediary sector.

Despite the lack of consensus, Shapley value, in my view, is a natural way to endow each sector some market power. More specifically, each layer

[5]I do not allow the intermediary fee to depend on individual skill types. Similarly, I do not allow skilled or unskilled intermediaries to signal through different contracts or certification system. Essentially, any mechanisms that reveal skill types eliminate the only friction in the model, thereby trivializing the problem. In practice, such a revelation may exists but is unlikely to be perfect as "pooling" arrangements are often observed in many markets. For instance, the "2-20" fee structure (2% management fee and 20% carried interest) is common in the hedge fund industry.

[6]In Section 5.2, I show that introducing a fixed fee f that is based on the total asset under management and independent of the quality of the portfolios does not materially change the analysis.

(including the investor layer) is a player in a cooperative game, and there are a total of $I + 1$ players. A coalition of $J \le I$ players has zero value, if it does not include the investor sector. A coalition of J players that includes the investor sector generates the value of a $J-1$-layer intermediation chain. With this naturally defined characteristic function, the set of $\{r_i\}$ ensures that each sector receive its Shapley value in the game. The detailed mathematical expression for Shapley value and the determination of r_i will be provided in Subsection 2.3. As a simple example, in the single layer case $I = 1$, the intermediaries and the investors will equally share the surplus brought by having the intermediary sector in the game. In a multi-layer setting, the Shapley value can also be microfounded in a bargaining game where investors and each intermediary layer negotiate the division of surplus (see Stole and Zwiebel 1996 for details).

An alternative modeling choice to endogenize r_i is to assume that the intermediary sector is competitive in that the unskilled ones break even in all layers (Section 4 implements this alternative).[7] Most of the results in Section 3 are robust to different ways to endogenize r_i. However, Proposition 2 no longer applies as its insight is based on how intermediary market power (which no longer exists with a competitive fee r_i) varies with chain length.

2.2 Payoffs and Welfare

The payoffs to investors and intermediaries quickly become complicated as the chain gets longer $I > 1$, due to the many different paths that a investor may receive a good portfolio. As part of the technical contribution of this paper, I employ matrix algebra to achieve general tractability.

Define the following screening matrix

$$\mathbb{P}_i \equiv \begin{pmatrix} P_i(e_i, n_{i+1}) & 1 - P_i(e_i, n_{i+1}) \\ P_i(0, n_{i+1}) & 1 - P_i(0, n_{i+1}) \end{pmatrix}, \quad i = 0, 1, 2, ..., I. \tag{1}$$

which captures the probability that a skilled (resp. unskilled) player in layer i finds a skilled/unskilled player in layer $i+1$ (or a good/bad portfolio if $i = I$).

[7] Another possibility is to assume that intermediaries break even on average ex-ante before they learn their skill. The outcome is qualitatively identical to the specification in Section 4.

A skilled player in layer $i \geq 0$ receives

$$
\begin{aligned}
\Pi_{S,i} = \quad & \max_{e_i} \underbrace{N_0 \quad n_0 \quad 1-n_0 \quad \Pi_{j=0}^{i-1}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{investor flow through a skilled player}} \frac{1}{N_{S,i}} r_i \times \\
& \underbrace{1 \quad 0 \,\Big)\, \Pi_{j=i}^{I}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of a skilled player finding a good product}} \qquad -c_i(e_i).
\end{aligned}
\tag{2}
$$

and similarly, an unskilled player in layer $i \geq 0$ receives

$$
\begin{aligned}
\Pi_{U,i} = \quad & \underbrace{N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{i-1}\mathbb{P}_j \begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\text{investor flow through an unskilled player}} \frac{1}{N_{U,i}} r_i \times \\
& \underbrace{0 \quad 1 \,\Big)\, \Pi_{j=i}^{I}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of an unskilled player finding a good product}} \qquad ,
\end{aligned}
\tag{3}
$$

where $N_{S,i} = n_i N_i$ and $N_{U,i} = (1-n_i) N_i$ denote the mass of skilled and unskilled players in layer i. Even though $N_{S,i} = n_S \frac{1}{I}$ and $N_{U,i} = (1-n_S)\frac{1}{I}$ for $i \geq 1$ can be immediately simplified in the baseline model, I preserve the generality of setup for later use in Section 4, where the self-selection of the intermediaries by their skill types creates more complex structures.

To understand the payoff functions, the matrix product in the first row of (2) captures the probability that a investor finds an skilled (resp. unskilled in (3)) intermediary in layer i through the initial $i-1$-layer chain. There are a total of N_0 investors, and the total investor flow is shared by the number of skilled (resp. unskilled in (3)) intermediaries in layer i. If a good portfolio is delivered to a investor, the intermediary receives r_i. The matrix product in the second row of (2) captures the probability that a skilled (resp. unskilled in (3)) player can find a good portfolio through the subsequent chain. Skilled intermediaries also pay the effort cost $c_i(e_i) \in \{0, c_e\}$ in (2).

It is useful to note that expressions (2) and (3) are general enough to admit the investor payoff as a special case. When $i = 0$, the first row simplifies to r_0 which is an investors' payoff from a good portfolio, and the second row is the probability that they receive a good portfolio through the chain.[8]

[8] The following convention applies: $\prod_{i=0}^{-1} \mathbb{P}_j$ is the identity matrix.

The welfare in the economy is defined as the sum of the payoffs of all investors and intermediaries. It captures the total payoff generated by investments and deducts the total effort and entry costs incurred by skilled players. The total welfare as a function of the chain length and number of investors can be expressed as:

$$\begin{aligned} W_I &\equiv \textstyle\sum_{i=0}^{I} \Pi_{S,i} N_{S,i} + \Pi_{U,i} \left(N_i - N_{S,i}\right) \\ &= \underbrace{N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{i=0}^{I} \mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{total payoff in the game}} - \underbrace{\sum_{i=0}^{I} c_i(e_i) N_{S,i}}_{\text{total effort costs by skilled players}}, \end{aligned} \tag{4}$$

Note that the intermediation fees (r_i) are purely a transfer among players and does not directly affect welfare since they cancel out in investor and intermediary payoffs.

2.3 Equilibrium Definition

The exogenous parameters of the baseline model include the number of investors N_0, the chain length I (which can be endogenized in Subsection 3.2), the fraction of skilled investors, intermediaries, and good portfolios n_0, n_S, and n_P, respectively. The equilibrium conditions of the baseline model are as follows:

1. Optimal effort choices: The effort levels $\{e_i^* | i \geq 0\}$ are individually optimal for skilled investors or intermediaries, as given by equation (2).[9]

2. Intermediary fees $\{r_i^* | i \geq 1\}$ are such that each (intermediary or investor) sector receives its Shapley value in the chain:

$$\underbrace{r_i^* N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix} - c\left(e_i^*\right) n_i N_i}_{\text{total payoff to layer } i \text{ intermediaries}} = \frac{1}{I+1} \sum_{s=0}^{I-1} \frac{C_{I-1}^{s}}{C_I^{s+1}} \quad \tilde{W}_{s+1} - \tilde{W}_s \quad . \tag{5}$$

Condition (5) needs some explanation. The left-hand side is the total payoff to layer $i = 0, 1, ..., I$.[10] The total mass of investors receiving good portfolio is

[9] For simplicity, I focus on symmetric equilibria in which all skilled intermediaries in a given layer choose the same effort level. In equilibrium, all skilled intermediaries within a layer face the same optimization problem.

[10] It can be easily verified that that the expression in the right-hand side equals $\Pi_{U,i} N_{U,i} + \Pi_{S,i} N_{S,i}$.

$N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. The portfolio must flow through one of the intermediaries in each layer. Players in layer i receive r_i^* for each good portfolio delivered to investors, and the skilled ones (mass $n_i N_i$) incurs effort cost if $e_i^* \neq 0$. The right-hand side is Shapley value associated with layer i. By definition, the Shapley value of layer i in the chain is its average marginal contribution to all possible coalitions formed by the other players. A coalition consisting only of intermediaries does not generate welfare. Therefore, a coalition of size $s+1$ sectors only generates welfare if one of the sectors is investors. Denote by $\tilde{W}_s$ the total welfare generated by an s-layer intermediation chain with N_0 investors and $N_j = \frac{1}{I}$ $(j = 1, 2, ..., s)$ intermediaries in each layer. There are C_I^{s+1} ways to form a coalition with $s+1$ sectors, among which C_{I-1}^{s} includes the investor sector with an s-layer intermediation chain. For the s-layer chain, the marginal value of having one additional layer is $\tilde{W}_{s+1} - \tilde{W}_s$, and hence the expression (5).

Technical Details: There may be multiple equilibria. For instance, if $p_e - n_P < c_e$, that is, investors lack effort incentives when directly screening the asset market, a trivial equilibrium emerges where no intermediaries make effort, and all intermediary fees $r_i \equiv 0$ for $i \geq 1$. Another, more efficient equilibrium exists where skilled intermediaries in some final I_E layers exert effort, as detailed in Proposition 2 later. Instead of selecting a specific equilibrium, I characterize the range of all possible equilibrium outcomes. Given the equilibrium multiplicity, various coalition values $\hat{W}_s$ may arise depending on the selected equilibrium. I adopt the convention that a coalition with $J \in [0, I]$ layers of intermediaries mirrors the effort pattern of the full chain, where the final I_E layers exert effort. This convention on effort pattern in a coalition can also be supported as an equilibrium outcome in a J-layer chain when N_0 is large and each layer inherits the same fee r_i^* as in the full chain.

3 Long Chain: Screening Efficiency and Market Power Reduction

The added ingredients provide further insights beyond the introductory example. Firstly, subsection 3.1 underscores the difficulty of incentivizing effort in the upstream layers of the chain. Subsequently, in Subsection 3.2, I demonstrate that the overall market power of the intermediary sector diminishes with the lengthening of the chain. Leveraging this observation, I show optimal chain length I^* that maximizes total welfare or investor payoff is generally finite. Additionally, as the number of investors N_0 grows, I establish an efficient chain length characterized by a power-law relationship $I^* \sim \log N_0$ in Subsection 3.3. Finally, Subsection 3.4 furnishes an explicit criterion to determine whether intermediation outperforms direct market access for investors.

3.1 Effort Pattern along the Chain

I start by characterizing the first set of equilibrium variables — effort choices in each layer e_i^*. If skilled intermediaries in layer i do not make effort, they are equivalent to unskilled intermediaries, which immediately implies that no one in previous layers need to exert effort. Therefore, only skilled intermediaries in some final I_E layers make effort, i.e. when $i > I - I_E$ for some cutoff $I_E \geq 0$. Mathematically, the screen matrices are given by

$$\begin{aligned} \mathbb{P}_j &= \begin{pmatrix} n_S & 1-n_S \\ n_S & 1-n_S \end{pmatrix} \text{ if } j \leq I - I_E \\ \mathbb{P}_j &= \begin{pmatrix} p_e & 1-p_e \\ n_S & 1-n_S \end{pmatrix} \text{ if } j \geq I - I_E + 1 \end{aligned} \tag{6}$$

To find the cutoff layer, consider the effort decision for the skilled intermediaries in the first effort-making layer, $i = I - I_E + 1$. Since no one in the previous layers makes effort, each intermediary in this layer randomly receives $\frac{N_0}{N_I}$ investors.[11] The effort decision by layer $I - I_E + 1$ only affects $\mathbb{P}_{I-I_E+1}$. Hence, condition (2) implies that skilled players in layer i exerts effort if

$$\frac{N_0}{N_I} r^*_{I-I_E+1} \left\{ \begin{pmatrix} 1 & 0 \end{pmatrix} \left[\begin{pmatrix} p_e & 1-p_e \\ n_S & 1-n_S \end{pmatrix} - \begin{pmatrix} n_S & 1-n_S \\ n_S & 1-n_S \end{pmatrix} \right] \Pi_{j=I-I_E+2}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \geq c_e.$$

Simple algebraic manipulation yields the following key expression that char-

[11] Indeed, it should be easy to check that $\begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \mathbb{P}_j = \mathbb{P}_j$. Hence, the first row in (2) simplifies to $N_0 (1-n_S) \frac{1}{N_{U,i}} r_i = \frac{N_0}{N_I} r_i$.

acterizes the effort pattern along the chain[12]

$$\frac{N_0}{N_I} r^*_{I-I_E+1} \left(p_e - n_S\right)^{I_E} \geq c_e. \tag{7}$$

Condition (7) reveals a robust economic insight on its own: It is more difficult to incentivize effort in the upstream layers of the chain. Mathematically, (7) is less likely to hold for a larger I_E. Hence, upstream layers (with a smaller layer index $I - I_E + 1$) are less likely to start making effort. Intuitively, the success of an upstream intermediary's screening effort depends on whether the selected downstream intermediaries can find a good portfolio. Each subsequent layer amplifies this "indirectness," which is mathematically captured by the term $\left(p_e - n_S\right)^{I_E}$ in (7). In contrast, skilled intermediaries in the final layer rely on their own effort to find a good portfolio without contamination from other intermediaries. Hence, downstream intermediaries have stronger effort incentives. This insight is robust to a general probability function $P_i(e_i, n_{i+1})$ with continuous effort choice.[13] Rearranging terms, I have the following result:

Proposition 1 *There exists a cutoff layer*

$$I_E \equiv \min\left\{I, \left\lfloor \log_{p_e - n_S} \frac{c_e N_I}{r^*_{I-I_E+1} N_0} \right.\right. , \tag{8}$$

where $\lfloor \cdot \rfloor$ denotes the floor function, such that skilled intermediaries in layer i make effort if and only if $i > I - I_E$. Additionally, the probability that an investor finds a good portfolio is

$$P^{(I_E)} = p^\dagger \left[1 - \left(p_e - n_S\right)^{I_E + 1}\right], \tag{9}$$

where

$$p^\dagger \equiv \frac{n_S}{1 - p_e + n_S}. \tag{10}$$

[12]The key to the derivation is to observe that $\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 - n_S & -(1-n_S) \\ 0 & 0 \end{pmatrix} = (1 - n_S)\begin{pmatrix} 1 & -1 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \end{pmatrix} \mathbb{P}_j = (1 - n_S)\begin{pmatrix} 1 & -1 \end{pmatrix}$.

[13]When the probability function $P_i(e_i, n_{i+1})$ is continuous rather than binary, condition (7) becomes

$$\frac{N_0}{N_I} r^*_{I-I_E+1} \Pi^I_{i=I-I_E+1} \left(P_i(e_i, n_{i+1}) - P_i(0, n_{i+1})\right) \frac{\partial P_{I-I_E+1}(e_{I-I_E+1}, n_{I-I_E+2})}{\partial e_{I-I_E+1}} = c'_{I-I_E+1}(e_{I-I_E+1}).$$

Each subsequent layer in the chain reduces the marginal benefit of effort by $P_i(e_i, n_{i+1}) - P_i(0, n_{i+1})$.

As an important step in characterizing the long-chain result, Proposition 1 determines the effort pattern and the success probability. To better understand (9), first note that $\begin{pmatrix} p^\dagger & 1-p^\dagger \end{pmatrix}$ is the left invariant distribution under the screening matrix $\mathbb{P}$ in (6).[14] The convergence of the investor's success probability towards the invariant probability $p^\dagger$ occurs at an exponential rate when an additional effort-making layer is introduced. As a numerical example for the converging speed, let's assume that a skilled intermediary can find another skilled intermediary with certainty through effort ($p_e = 1$), and a fraction $n_S = 0.5$ intermediaries are skilled. In this case, the limiting probability $p^\dagger$ is equal to 1, and after $I_E = 4$ rounds of intermediation, the success probability reaches approximately $1 - 0.5^4 \approx 94\%$.

The observation that each layer of the intermediation chain gradually solves inefficiencies is indeed a common feature in the literature (Glode and Opp 2016, Glode, Opp, and Zhang 2019, and He and Li 2022). By incorporating effort decisions, my model highlights a novel theoretical observation that the resolution of inefficiencies is primarily achieved by the final (I_E) layers of the chain.

Another related observation that will be useful in Subsection 3.2 is that, for a given compensation structure r_i^* and sector size N_i, the number of effort-making layers I_E is independent of the chain length I because it does not appear in (7). Intuitively, the effort incentive depends only on the number of subsequent layers until the asset market and is not influenced by the preceding layers.

I conclude this subsection with an observation from the financial market. Financial advisors arguably represent the first intermediary layer next to investors. My model predicts that financial advisors have little incentive to screen. Empirically, Egan, Matvos, and Seru (2019) have shown that as many as 7% of financial advisors have a misconduct record, with the number exceeding 15% in some of the largest firms. If one views financial advisory firms as another intermediary layer in the chain, the model predicts that skilled intermediaries are more likely to identify other skilled ones in the next layer, creating an association between good firms and good financial advisors. These authors indeed find supportive evidence from the opposite perspective: some firms specialize in misconduct and cater to unskilled investors. On the other end of the intermediation chain, private equity funds or hedge funds directly

[14] It can be easily verified that $\begin{pmatrix} p^\dagger & 1-p^\dagger \end{pmatrix} \mathbb{P} = \begin{pmatrix} p^\dagger & 1-p^\dagger \end{pmatrix}$ and $n_S < p^\dagger < p_1$.

invest in asset markets. Anecdotally, they typically attract top talents in the financial market, and the turnover of unskilled fund managers is usually rapid. As a result, these fund managers also exert high effort to identify abnormal returns for assets under their management.

3.2 Intermediary Fee r_i^* and Optimal Chain Length

In this subsection, I complete the equilibrium characterization by calculating the second set of equilibrium variables — intermediary fee r_i^*. A key qualitative result is that a sufficiently long chain can eliminate the market power of the intermediary sector. Consider the equilibrium fee r_i^* determined by the Shapley value in (5). When the coalition size is at least $s \geq I_E + 1$ (one investor sector and I_E intermediary sectors), there is no marginal welfare creation ($W(s+1) = W(s)$), as it is always those intermediaries in the final I_E layers exerting effort. Therefore, an intermediary layer only contributes to the total welfare if the existing coalition forms a chain fewer than I_E layers ($s \leq I_E$). After some manipulations, condition (5) becomes:

$$r_i^* P^{(I_E)} = \frac{1}{I+1} \sum_{s=0}^{I_E - 1} \frac{s+1}{I} \quad \big(P^{(s+1)} - P^{(s)} \quad - c_e n_S \frac{N_i}{N_0} \quad + \mathbf{1}_{i \geq I - I_E + 1} c_e n_S \frac{N_i}{N_0}. \tag{11}$$

Note that the first term (the average welfare contribution) in (11) is independent of the layer index i (recall $N_i = \frac{1}{I}$ is a constant). Hence the equilibrium fee only varies based on whether layer i needs to be compensated for their effort input in equilibrium. The initial $I - I_E$ layers do not make effort, and hence their fees ($i \leq I - I_E$) are given by

$$r_i^* = r_{ne}^* \equiv \frac{1}{(I+1)\, P^{(I_E)}} \sum_{s=0}^{I_E - 1} \frac{s+1}{I} \quad \left(P^{(s+1)} - P^{(s)} \right) - c_e n_S \frac{1}{I N_0} \quad . \tag{12}$$

Skilled intermediaries in the final I_E layers make effort and the corresponding fees ($i \geq I - I_E + 1$) are therefore

$$r_i^* = r_e^* \equiv \frac{1}{(I+1)\, P^{(I_E)}} \sum_{s=0}^{I_E - 1} \frac{s+1}{I} \quad \big(P^{(s+1)} - P^{(s)} \quad - c_e n_S \frac{1}{I N_0} \quad + c_e n_S \frac{1}{P^{(I_E)} I N_0}. \tag{13}$$

Ignoring the effort compensation term that is vanishing small $c_e n_S \frac{1}{N_0}$ as $N_0 \to \infty$ (this approximation is for illustration only and not required for the proof), the total compensation to the intermediary sector is approximately given by

$$r_i^* P^{(I_E)} I \approx \sum_{s=0}^{I_E-1} \frac{s+1}{I+1} \left(P^{(s+1)} - P^{(s)}\right) = \sum_{s=0}^{I_E-1} \frac{s+1}{I+1} n_S \left(p_1 - n_S\right)^{s+1}. \qquad (14)$$

Essentially, the total market power of the intermediary sector (measured by $r_i^* I$ or more precisely $r_{ne}^* (I - I_E) + r_e^* I_E)$ can be seen as a weighted average of the marginal increase in the success probability due to each one of the I_E effort-making layers. Recall from the discussion at the end of Subsection (3.1) that I_E does not directly depend on I. Therefore, as the chain lengthens, the market power of the intermediary sector diminishes ($\frac{I_E}{I} \to 0$). Intuitively, when there are many layers, each layer becomes less pivotal and contributes less to the total welfare (Feature 3 of the introductory example). However, the investor sector is indispensable to create welfare and hence the result.

Now, I introduce the main result of this section that characterizes the outcome as the chain length I varies.

Proposition 2 *Given the size of the investor sector N_0, the chain length I that maximizes the number of effort-making levels I_E is generally finite. Furthermore, when the chain becomes excessively long, no intermediary makes effort*

$$\lim_{I \to \infty} I_E^*(I) = 0.$$

In the case of a relatively short chain, all intermediary layers have effort incentives ($I_E = I$) as their efforts significantly impact the probability of delivering a good portfolio (condition (7) is more likely to hold with a small I_E). Consequently, due to their substantial welfare contribution, the corresponding fee r_i^* is relatively large. Here, the lengthening of the chain improves the success probability (condition (9)) and simultaneously reduces the market power of intermediaries ($r_i^* I$ decreases).

As the chain becomes longer, the marginal contribution of each layer diminishes, leading to a decrease in the equilibrium fee r_i^*. Consequently, the number of effort-making layers I_E eventually decreases, with the extreme case being no layers making effort ($I_E = 0$).

A direct implication of this proposition is that the optimal chain length is generally interior. However, I refrain from presenting a formal result because the exact optimal I^* depends on the objective function. Depending on the objective—whether it is maximizing investor payoff, total welfare, or the probability of finding good portfolios—the optimal length varies. Nonetheless, the fundamental tradeoff remains consistent: a longer chain reduces the market power of intermediaries but simultaneously diminishes their effort incentives.

3.3 Large Investor Sector: An Asymptotic Efficient Result

Up to this point, I have characterized the equilibrium outcome for any given number of investors N_0. Now, I demonstrate that if the investor sector is sufficiently large, as is the case in many markets, a long chain construction can achieve asymptotic efficiency.

Proposition 3 *Suppose the number of investors N_0 is sufficiently large. Let the chain length $I = N_0^\alpha$ for any constant $\alpha \in (0,1)$. Then as $N_0 \to \infty$, there exist a sequence of equilibria where the number of effort-marking layers I_E satisfies:*

$$\lim_{N_0\to\infty} \frac{I_E}{\ln N_0} = -\frac{1-\frac{1}{\alpha}}{\ln (p_e - n_S)} > 0 \tag{15}$$

In addition, both the skill and unskilled investor payoff approach $p^\dagger$:

$$\lim_{N_0\to\infty} \Pi_{S,0} = \lim_{N_0\to\infty} \Pi_{U,0} = p^\dagger.$$

Finally, the welfare in the economy almost entirely accrues to investors:

$$\lim_{N_0\to\infty} \frac{W_I}{N_0} = p^\dagger.$$

The result is by construction. The chain length is constructed as a power function of the investor size N_0. Given such a construction, I show that the number of effort-making layers I_E grows logarithmically with N_0. This pattern is primarily driven by the fact that effort incentive decreases exponentially as a layer is farther away from the asset market (refer to condition (7)). The presence of many effort-making layers ($I_E \to \infty$) drives the success probability towards the invariant probability $p^\dagger$ (see Proposition 1). The intermediaries' effort costs are shared among the large number of investors, and hence the welfare conclusion.

The reason why welfare almost entirely accrues to the investors is similar to the intuition behind Proposition 2: The fraction of effort-making layers is vanishingly small in the entire chain $\frac{I_E}{I} \to 0$. The presence of many non-effort-making layers eliminates the market power of the entire intermediary sector. As a result, the total fee paid to the intermediary sector (i.e. $r^*_{ne}(I - I_E) + r^*_e I_E$) becomes vanishingly small. Investors therefore receive almost all of the surplus in the economy.

The introductory example can be considered as a special case where effort ensures a good outcome with certainty ($p_e = 100\%$). In this case, the limiting probability is also $p^\dagger = 100\%$. Consequently, Proposition 3 demonstrates that a long chain can create first-best efficiency even when each layer contains both skilled and unskilled intermediaries and the intermediary sector possesses market power.

3.4 Is Intermediation Beneficial?

Now I can provide a simple condition for when intermediation is efficient. If investors directly screen in the asset market ($I = 0$), their payoff depends on whether the skilled investors find it worthwhile to exert effort. When

$$p_e - n_s < c_e, \tag{16}$$

no investors make effort and they each receive the payoff of

$$W_0 = n_S.$$

Since a long intermediation chain can deliver a payoff of $W_\infty = p^\dagger$ in the limit (Proposition 3), it is obvious that intermediation improves efficiency.

The comparison is more complicated when skilled investors have effort incentives on their own (i.e., when condition (16) fails). In this case, the average payoff to an investor without intermediation is

$$W_0 = n_0 (p_e - c_e) + (1 - n_0) n_S.$$

Compared with the intermediated payoff $p^\dagger$, the following tradeoff emerges. There are two benefits. First, intermediation creates economies of scale on effort costs c_e, because it eliminates the need for each investor to exert effort. Second, intermediation improves the payoff to the unskilled investors from n_S to $p^\dagger$. There is, however, also a downside of intermediation: It reduces the probability of skilled investors receiving good portfolios from p_e to $p^\dagger$. The next proposition offers a succinct condition on the efficiency of intermediation.

Proposition 4 *Assume the investor sector N_0 is sufficiently large, a long intermediation chain can improve welfare and investor payoff if*

$$\begin{aligned} c_e & \geq \tfrac{1}{n_0}\left[n_0 p_e + (1-n_0)\, n_S - \tfrac{n_S}{1-p_e+n_S}\right] \\ & = (p_e - n_S)\left(1 - \tfrac{n_S}{n_0(1-p_e+n_S)}\right) \end{aligned} \tag{17}$$

It is useful to note that condition (17) nests (16) as a subset. Intuitively, intermediation is beneficial if the economies of scale on effort costs c_e is significant, including the case when it is prohibitively costly for investors to exert effort on their own (16). In addition, when the fraction of the skilled investors n_0 is sufficiently small, intermediation is beneficial.

Finally, it is important to note efficient intermediation does not require the intermediary sector being more skilled $n_S > n_0$. For instance, even when there is no cost-saving benefit $c_e = 0$ and more investors are skilled $n_0 \in (n_S, 1)$, intermediation may still be beneficial if p_e is sufficiently large. This result is also robust to the possibility that unskilled intermediaries may deliver a strictly worse outcome than random selection.[15] In the introductory example, $p_e = 100\%$ — a case that guarantees (17).

4 Intermediary Self-Selection and Short Chain Efficiency

In the baseline model, I have assumed that each intermediary layer has homogeneous quality n_S. This assumption can be motivated by the possibility of learning on the job — skill is only realized after the chain has formed. An alternative and equally plausible setting is that intermediaries can choose where to operate in the chain based on their still types. In this section, I extend the model to study this possibility. The self-selection of intermediaries generates implications on endogenous sector sizes (N_i) and skill distribution (n_i) along the chain.

[15] One only needs to replace n_S in condition (17) by the corresponding probability that can be achieved by an unskilled intermediary.

The key finding is that a short two-layer chain with heterogeneous qualities can achieve asymptotic efficiency. The structure is largely determined by the relative entry costs associated with each layer. Welfare as well as investor payoff are non-monotonic with respect to the structure. Efficiency is achieved by having either the first layer (Subsection 4.2) or the second layer (Subsection 4.3) consist entirely of skilled intermediaries, while the other layer contains both skill types. In contrast, when both layers contain mixed-quality intermediaries, efficiency is reduced (Subsection 4.4). The results shed light on lender-originator chain, and the investment chain in the delegated asset management industry.

4.1 Introducing Intermediary Self-Selection

In this section, I modify the model to incorporate the self-selection of intermediaries. The basic chain structure and the process of portfolio selection are the same as in the baseline model. Unlike before, intermediaries know their skill types and can pay entry cost C_i in order to operate in layer i. I simplify away the intermediary market power and assume a constant fee $r_i \equiv r$ that is determined by the unskilled intermediaries breakeven condition. The payoff functions of the skilled and unskilled players in layer i only require a small modification to the baseline model 2 and 3 in order to incorporate the entry costs C_i.

$$\Pi_{S,i} = \max_{e_i} \underbrace{N_C \quad n_0 \quad 1-n_0 \quad \Pi_{j=0}^{i-1}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{investor flow through a skilled intermediary}} \frac{1}{N_{S,i}} r \underbrace{1 \quad 0 \Big) \Pi_{j=i}^{I}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of a skilled intermediary finding a good product}} - c(e_i) - C_i. \tag{18}$$

and

$$\Pi_{U,i} = \underbrace{N_C \Big(n_0 \quad 1-n_0 \Big) \Pi_{j=0}^{i-1}\mathbb{P}_j \begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\text{investor flow through an unskilled intermediary}} \frac{1}{N_{U,i}} r \underbrace{0 \quad 1 \Big) \Pi_{j=i}^{I}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of an unskilled intermediary finding a good product}} - C_i \tag{19}$$

Welfare in the economy is defined as

$$W = \underbrace{N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{i=0}^{I} \mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{total payoff in the game}} - \underbrace{\sum_{i=0}^{I} \left(c(e_i) N_{S,i} + C_i N_i\right)}_{\text{total effort and entry costs}}, \tag{20}$$

which is the total payoff generated by investments net of screening and entry costs incurred by (skilled) investors and intermediaries. For notational simplicity, $C_0 = 0$ — investors do not have entry costs.

I also enrich the model to allow for continuous effort choices $e \in \mathbf{E}$ among some feasible set $\mathbf{E}$. The qualitative results only depend on the following natural assumptions on the screening technology $P(e,n)$ and $c(e)$.

Assumption 1: $P(0,n) = n$, $P(\infty,n) \to 1$ for $n > 0$, $P(e,0) = 0$, and $P(e,1) = 1$.

Perhaps only one assumption ($P(\infty,n) \to 1$) requires explanation: as long as good intermediaries or portfolio exists, sufficiently large effort input results in finding them almost certainly. This assumption allows the possible outcome to be first best for exposition. Subsection 4.4 relaxes this assumption.

Assumption 2: $\frac{\partial P(e,n)}{\partial e} \geq 0$, $\frac{\partial^2 P(e,n)}{\partial e^2} \leq 0$, $\frac{\partial P(\infty,n)}{\partial e} \to 0$, and $\frac{\partial P(e,n)}{\partial n} \geq 0$.

Naturally, effort improves the probability of finding a good intermediary or portfolio, but its marginal effect diminishes. Higher quality n naturally also helps the screening outcome. A simple example is

$$P(e,n) = \frac{(1+e)\,n}{(1+e)\,n + 1 - n} = \frac{(1+e)\,n}{1+en}. \tag{21}$$

Assumption 3: $c(0) = 0$, $c(\infty) = \infty$, $c'(0) = 0$, $c'(e) > 0$, and $c''(e) > 0$.

The (marginal) cost of effort is initially zero and increases in a convex fashion to infinity. A simple example of this cost function is

$$c(e) = \frac{1}{2} e^2. \tag{22}$$

An equilibrium now contains three sets of variables: 1. the number of skilled and unskilled intermediaries in each layer $\left\{N_{S,i}^*, N_{U,i}^* | 1 \leq i \leq I\right\}$, 2. the effort choice of investors and intermediaries in each layer $\{e_i^* | 0 \leq i \leq I\}$, and 3. the equilibrium fee r^*. These $3I + 2$ equilibrium variables are determined by the following equilibrium conditions.

1. [$I + 1$ conditions] All effort choices $\{e_i^* | 0 \leq i \leq I\}$ are optimal, deter-

mined by skilled players (18).

2. [I conditions] Unskilled intermediaries break even whenever they operate in a layer $i = 1, 2, ..., I$:

$$\Pi_{U,i} = 0. \tag{23}$$

Alternatively, if a layer does not include unskilled intermediaries, then $N^*_{U,i} = 0$ is the corresponding equilibrium condition and (23) holds with inequality $\leq$.

3. [$I - 1$ independent conditions] Skilled intermediaries are indifferent between the different layers that they operate in

$$\Pi_{S,i} = \Pi^*_G, \tag{24}$$

for some constant $\Pi^*_G \geq 0$ independent of i. Alternatively, if a layer does not include skilled intermediaries, then $N^*_{S,i} = 0$ is the corresponding equilibrium condition and (24) holds with inequality $\leq$.

4. [2 conditions] Intermediary market clearing by skill types

$$\sum_{i=1}^{I} N^*_{S,i} = n_S \text{ and } \sum_{i=1}^{I} N^*_{U,i} = 1 - n_S.$$

4.2 High Quality First Layer

The first of the two possibilities to achieve asymptotic efficiency with a short two-layer chain ($I = 2$) is by having a significantly higher entry cost for the first layer ($C_1 - C_2 \to \infty$). The structure features a small but a high-quality first layer ($n_1 = 1$, $N^*_1 \to 0$). Formally,

Proposition 5 *Suppose $I = 2$. As $N_0 \to \infty$, there exists a sequence of entry costs $C_1(N_0)$ and $C_2(N_0)$ such that the corresponding sequence of equilibria have the following properties:*

1. Only a small number of skilled intermediaries join layer 1: $N^*_{S,1} = N^*_1 \to 0$. The remaining skilled as well as all unskilled intermediaries join layer 2: $N^*_{U,2} = 1 - n_S$ and $N^*_{S,2} \to n_S$.

2. Investors do not exert effort $e^*_0 = 0$. Skilled intermediaries in both layers exert high effort $e^*_i \to \infty$ ($i = 1, 2$).

3. The payoffs of all investors approach the first-best outcome: $\Pi_{S,0} = \Pi_{U,0} \to 1$.

4. The entry cost for the first layer dominates that for the second layer: $C_1 - C_2 \to \infty$.

In this equilibrium, the small amount of skilled intermediaries in layer 1 each receives a large amount of investors and therefore has strong effort incentive to screen the lower-quality intermediaries in layer 2. Skilled intermediaries in layer 2 also exert high effort to identify good portfolios, after receiving significant investor delegation from layer 1 intermediaries. Investors, however, do not need to exert any effort ($e_0^* = 0$) because all intermediaries in layer 1 are good, but nonetheless can receive good portfolios almost certainly ($P(e_1^*, n_2^*)P(e_2^*, n_P) \to 1$). The unskilled intermediaries in layer 2 do not wish to deviate and join layer 1 because they do not have the skill to deliver a skilled intermediary from layer 2 and therefore cannot break even on the high entry cost.

It is worth noting that the entry costs $C_1(N_0)$ and $C_2(N_0)$ need to be designed appropriately so that the efficient outcome can be attained at a diminishing cost to investors, that is the intermediation fee $r^* \to 0$ and simultaneously provides sufficient effort incentives in both layers ($\frac{r^* N_0}{N_{S,i}} \to \infty$). The main technical challenge arises from the strategic complementarity of effort across layers. If skilled intermediaries in layer 2 exert higher effort, finding them becomes more valuable, and skilled intermediaries in layer 1 exert higher effort. Conversely, higher effort by skilled intermediaries in layer 1 leads to more investor delegation to skilled intermediaries in layer 2, motivating them to exert higher effort as well. This self-fulfilling nature poses a challenge to the existence of an equilibrium with the desired properties. The proof in the appendix provides detailed construction of the entry costs to address this challenge.[16]

As an application of Proposition 5, the model sheds light on the market structure of banks (layer 1) and loan originators (layer 2). It is clear that forming a bank entails a much higher cost than operating as a loan originator ($C_1 - C_2 \to \infty$). As a result, the first layer – the banking sector – has relatively few players ($N_{S,1}^* \to 0$) relative to the second-layer loan originators and the number of depositors N_0. Banks are often considered very skilled in screening

[16]To establish the existence of the equilibrium, I employ a constructive approach. First, I construct a sequence of equilibrium variables $N_{S,1}^*$ and e_2^* as $N_0 \to \infty$. Using these variables, I solve for the corresponding values of r^* and e_1^* by satisfying the incentive compatibility conditions of skilled intermediaries. Next, I demonstrate that by appropriately designing $N_{S,1}^*$ and e_2^*, the desired equilibrium properties $r^* \to 0$ and $P(e_1^*, n_2^*) \to 1$ are satisfied. This ensures that the equilibrium price approaches zero, and the probability of finding a good portfolio converges to one. Finally, I establish that the entry costs C_1 and C_2 can be calculated from the equilibrium variables and verify $C_1 - C_2 \to \infty$.

borrowers ($n_1^* = 1$). Consequently, investors (depositors) do not need to screen banks' lending skills ($e_0^* = 0$). On the other hand, the second layer, consisting of originators, is larger ($N_2^* \approx 1$) and exhibits mixed quality ($n_2^* \approx n_S < 1$).

4.3 High Quality Second Layer

The second possibility to achieve asymptotic efficiency with a two-layer chain ($I = 2$) is by having a significantly higher entry cost for the second layer $C_2 - C_1 \to \infty$. The structure features a small but a high-quality second layer ($n_2 = 1$, $N_2^* \to 0$). Formally,

Proposition 6 *Suppose $I = 2$. For any sequences of $C_1 \to 0$ and $C_2 \to \infty$, the corresponding sequence of equilibria have the following properties:*

1. Only a small number of skilled intermediaries join layer 2: $N_{S,2}^* = N_2^* \to 0$. The remaining skilled as well as all unskilled intermediaries join layer 1: $N_{U,1}^* = 1 - n_S$, $N_{S,1}^* \to n_S$.

2. No investors or intermediaries in layer 1 exert effort: $e_0^* = e_1^* = 0$. Skilled intermediaries in layer 2 exert high effort $e_2^* \to \infty$.

3. The payoffs of all investors approach the first-best outcome for any N_0: $\Pi_{S,0} = \Pi_{U,0} \to 1$.

The high quality of layer 2 implies that investors as well as layer 1 intermediaries do not need to make effort ($e_0^* = e_1^* = 0$). The large amount of investor delegation received by skilled intermediaries in layer 2 ($\frac{N_0}{N_{S,2}^*}$) provides strong effort incentive and as a result, $P(e_2^*, n_P) \to 1$. The unskilled intermediaries do not deviate and join layer 2 because their inability to effectively screen the asset market it impossible for them to break even on the large entry cost C_2.

While the market structure in Proposition 6 might appear as the flip side of that in Proposition 5, there are crucial qualitative differences. Efficiency in Proposition 5 necessitates a large number of investors $N_0 \to \infty$, but the structure in Proposition 6 does not. When layer 1 has high quality, all skilled intermediaries in the chain must exert high effort, demanding a large number of investors to share the costs for achieving asymptotic efficiency. Conversely, when layer 2 has high quality, only the arbitrarily small number of intermediaries in this layer incur substantial screening and entry costs ($N_{S,2}^* \left(c(e_2^*) + C_2\right) \to 0$). Consequently, the total cost remains negligible even with a finite number of investors.

This structure aligns well with several real-world examples. In the asset management industry, it is arguably more costly to develop and maintain a reputable index like the S&P 500 (layer 2) compared to registering as a passive mutual fund manager (layer 1). Consequently, only a few widely recognized indices are followed by a significant portion of passive index mutual funds. The index providers exert effort to screen and construct portfolios, while the passive mutual funds typically track the indices through replication. Investors benefit from this structure as they can invest in the chosen index without the need for extensive evaluation of different mutual funds or the asset market.

To the extent that it is more challenging (costly) to create financial products (layer 2) than to become a financial advisor (layer 1), the model predicts there will be a large number of financial advisors recommending similar products to investors. This is consistent with the evidence documented by Egan, Matvos, and Seru (2019).

When the entry costs are similar $C_1 \approx C_2$, the structure characterized in Proposition 6 still holds, although the size of the second layer (N_2) may no longer be vanishingly small. Consequently, efficiency requires the investor sector to be sufficiently large $N_0 \to \infty$. This is numerically demonstrated in Figure 2.

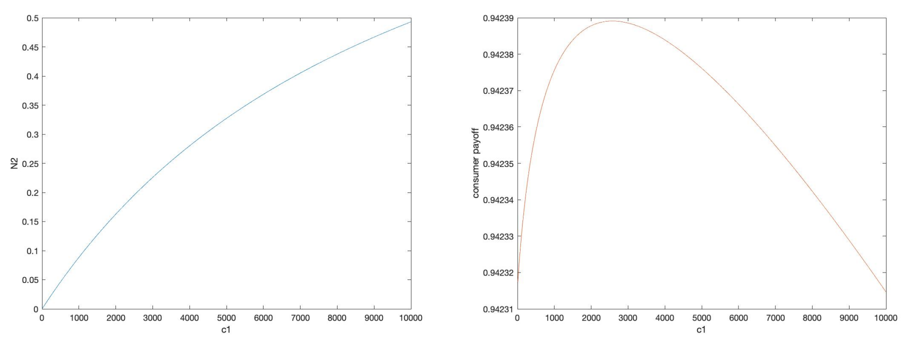

Figure 2 illustrates the relationship between the size of the second layer N_2 (left panel) and investor payoff (right panel) as a function of the first-layer entry cost C_1, which ranges from 10 to 10^4. The numerical example assumes the following parameter values: $N_0 = 10^9$, $n_0 = 0.6$, $n_P = 0.4$, $n_S = 0.5$, and $C_2 = 10^4$.

The graph on the left panel shows that as the first-layer entry cost C_1 increases, the size of the second layer N_2 also increases. In particular, when C_1 approaches the level of C_2, the size of the second layer expands to include all skilled intermediaries, $N_2 \to n_S$. The graph on the right panel displays the investor payoff as a percentage of the first-best outcome. It demonstrates that investor payoff stays approximately 94% of the first-best level in this numerical example.

A relevant example of this case is the relationship between fund-of-funds or pension funds (layer 1) and underlying private equity (PE) and hedge funds (layer 2). The entry costs to become a fund manager or a fund-of-funds manager are likely to be similar ($C_1 \approx C_2$). As a result, both intermediary sectors can have significant representation in the economy, with the more skilled managers residing in the second layer. This observation aligns with anecdotal evidence that PE and hedge funds typically attract top talent in the financial industry.

To conclude this subsection, it is worth discussing the potential deviation of investors to bypass layer 1 and directly engage with layer 2 intermediaries. By doing so, investors could avoid paying the layer 1 fees r^*. However, such a deviation is generally suboptimal in the case of Proposition 5, where investors crucially rely on the screening efforts of the high-quality layer 1 intermediaries to identify the skilled intermediaries in layer 2. Hence, this structure is robust against deviation.

On the other hand, in the case of Proposition 6, where all layer 2 intermediaries are skilled, investors do have an incentive to deviate and interact directly with them. The equilibrium structure therefore requires commitments to the chain. In practice, such commitments are often achieved through restricted access or some un-modeled services provided by layer 1 intermediaries. For example, in the context of index-tracking mutual funds and index providers, terminal investors usually lack the capacity to directly invest in hundreds of stocks to replicate an index. Similarly, in the financial advisor or fund-of-funds example, layer 1 intermediaries (financial advisors and fund-of-funds managers) may provide exclusive access to underlying funds that investors either do not have or are unaware of. It is important to note that due to the deviation incentives, this structure is less robust. Disruptions can occur, as seen in the rise of index exchange-traded funds (ETFs), which provide investors with a cost-effective means of accessing indices and significantly impact the traditional business model of index-tracking mutual funds.

4.4 Mixed Quality Layers, Compact Effort Space, and Imperfect Screening

Propositions 5 and 6 characterize the two most efficient outcomes that can be attained by a short two-layer chain. The shared feature is that the two layers have extremely heterogeneous quality: one layer with only skilled intermediaries and the other with everyone else. A natural question is what lies in between the two extreme cases. When both layers contain mixed skill types ($n_i \in (0,1)$), there is a positive probability (bounded below by $\Pi_{i=1}^{3}(1-n_i)$) that unskilled investors keep finding unskilled intermediaries and bad portfolios. In addition, all skilled players have some effort incentives, thereby reducing the economies of scale on effort costs. Consequently, the welfare is bounded away from first best. Formally,

Corollary 1 *When the entry costs C_1 and C_2 are such that both layers contain mixed skill types ($n_i \in (0,1)$), investor payoffs and welfare are bounded away from the first-best level for any intermediary structure $\{N_{S,i}, N_{U,i} | i = 1,2\}$.*

Another natural question is what if the effort set $\mathbf{E}$ is compact, and the screening outcome of the skilled players is less than perfect even at the maximum effort level. Mathematically,

$$P(\max_{e \in \mathbf{E}} e, n_{i+1}) \equiv p_e < 1.$$

In this case, the main results in this section (Propositions 5 and 6) have some minor changes. First, the feature that effort and entry costs are vanishingly small due to the economies of scale remains robust. In fact, if the effort level and costs are bounded rather than infinite, the cost per investor vanishes even quicker. Second, the success probabilities are now different as characterized by the following corollary.

Corollary 2 *The maximum payoff per investor that can be attained by having a high quality first layer ($n_1 = 1$) is*

$$p_e^2 + (1-p_e)\, n_P. \tag{25}$$

The maximum payoff per investor that can be attained by having a high quality second layer ($n_2 = 1$) is

$$p_e.$$

The expression in (25) is the probability that a skilled intermediary in layer 1 can find a good portfolio. With p_e probability it finds a skilled intermediary in layer 2, who in turn finds a good portfolio with probability p_e and with

$1-p_e$ probability it finds a unskilled intermediary in layer 2, who in turn finds a good portfolio with probability n_P.

Comparing the two cases, it is easy to see that the latter (high quality second layer $n_2 = 1$) dominates because all investors delegate the screening to a small set of skilled intermediaries, which is also the most efficient outcome that can be attained in this economy. However, the former structure (high quality first layer $n_1 = 1$) is locally efficient relative to small perturbations in the distribution of skill types. The logic is easy to see. Suppose some small amount of unskilled intermediaries enter the first layer, that is $n_1 = 1 - \epsilon$, for some small $\epsilon > 0$ and $n_2 \approx n_S$. A skilled investor receives a payoff that is strictly bounded above by (25) due to the effort cost required to find a skilled intermediary in layer 1. An unskilled investor receives a payoff that is strictly below (25) due to the positive probability ϵ that he may find a unskilled layer 1 intermediary. Hence, the bimodal feature characterized by Corollary 1 remains robust.

5 Robustness and Extensions

5.1 Multiple Asset Markets

The model can be easily generalized to incorporate M different segmented markets, where investors need M portfolios, one from each market. In this case, if (skilled) investors directly screen M different markets, they need to exert effort costs M times. Similarity, I assume that in the case of an intermediation chain, those skilled ones in the final layer need to incur effort costs $c_I(e_I)$ in each one of the M markets. As such, the payoff functions only need some minor adjustments for the effort cost in the final layer ($i = I$) and the fact there are now $N_0 M$ total transactions. Extending the payoff functions (18) and (19) to M portfolios, skilled players receive

$$\Pi_{S,i} = \max_{e_i} \underbrace{N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{i-1} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{investor flow through a skilled player}} \frac{1}{N_{S,i}} M r_i \underbrace{\left. \begin{matrix} 1 & 0 \end{matrix} \right) \Pi_{j=i}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of a skilled player finding a good product}} - c_i(e_i) \left[1 + \mathbf{1}_{i=I} (M-1)\right] - C_i, \tag{26}$$

and unskilled players receive

$$\Pi_{U,i} = \underbrace{N_0 \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{i-1}\mathbb{P}_j \begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\text{investor flow through an unskilled player}} \tfrac{1}{N_i - N_{S,i}} M r_i \underbrace{\begin{matrix} 0 & 1 \end{matrix}\Big) \Pi_{j=i}^{I}\mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of an unskilled player finding a good product}} - C_i. \tag{27}$$

All insights from the previous sections remain valid. In fact, the proofs for Propositions 5 and 6 in the appendix are for the generic case with M markets. However, the case with multiple markets indeed introduces a new channel: two-sided economies of scale on the effort costs, making it easier for intermediation to be efficient. This new channel can be best illustrated in the single layer case $I = 1$.

Proposition 7 *Suppose $M, N_0 \to \infty$ and the entry cost $C_1 = aM\left(N_0\right)^b$ for some constants $a > 0$, $b \in (0, 1)$, then for any $n_S > 0$, skilled investors' payoff per portfolio approaches the first best ($\frac{\Pi_{S,0}}{M} \to 1$).*

The unskilled investors' payoff per portfolio approaches

$$\frac{\Pi_{U,0}}{M} \to P(0, n_S) + \left[1 - P(0, n_S)\right] P(0, n_P), \tag{28}$$

which is higher than $P(0, n_P)$ — their payoff without intermediation ($I = 0$).

Intuitively, intermediation generates two-sided economies of scale from both receiving large investor delegations N_0 and screening many asset markets M. First, the large investor delegation motivates intermediaries' effort, even when the fee r^* individually is small ($r^* \to 0$ but $r^* N_0 \to \infty$). Intermediaries' effort costs are shared by the many investors that they represent.

Second, since skilled intermediaries can screen many markets M on behalf of investors, the skilled investors in turn has high effort incentive to find a skilled intermediary. Investors' one-off screening costs in the intermediary market are shared by the large number of good portfolios a skilled intermediary can deliver.

The payoff to the unskilled investors also improves similar to the introductory example. The skilled intermediaries offer these investors higher probability of finding good portfolios. The entry cost C_1 is designed so that the equilibrium fee r^* is simultaneously small individually and sufficiently big collectively ($r^* \sim \frac{C_1}{MN_0} = (N_0)^{b-1} \to 0$ and $r^* N_0 \sim N_0^b \to \infty$).

The two-sided economies of scale in this model bear resemblance to Diamond's (1984) theory of financial intermediation, where banks serve as delegated monitors on behalf of many small depositors. Similar to the two-sided economies of scale in my model, banks in Diamond (1984) receive deposits from many depositors and monitor many underlying firms to generate efficiency.

5.2 Fixed Fee and Performance Fee

In this paper, the focus is on the market structure of the intermediary sector. To simplify the analysis and isolate the effects of endogenous price and effort choice, the compensation of intermediaries has been reduced to a simple performance pay, where they receive a payment of r when a good portfolio is delivered to a investor. However, one natural extension is to introduce a fixed pay component based on the flow of investor capital through an intermediary, irrespective of the ultimate performance.

Formally, the intermediary payoff now consists of two components. Let f represent the fee that an intermediary can collect whenever a portfolio (regardless of its quality) is delivered to an investor through that intermediary. Additionally, as before, if the portfolio is good, the intermediary charges a performance-based fee r, resulting in a total compensation of $f + r$.

For simplicity, I consider the case where intermediaries do not know their skills $\theta = \frac{1}{2}$. Similar to 2 but with the additional fixed compensation term f, the break-even condition of an intermediary in layer i can be expressed as

$$\frac{MN_0}{N_i} \underbrace{\left[f + r \begin{pmatrix} n_0 & 1 - n_0 \end{pmatrix} \Pi_{j=0}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right]}_{\text{total compensation}} - n_S c(e_i) \left[1 + \mathbf{1}_{i=I} (M-1) \right] - C_i = 0. \tag{29}$$

The main finding of this analysis is that the introduction of a fixed fee f as a parameter does not lead to new equilibrium outcomes. The equilibrium results remain unchanged.

Proposition 8 *Any equilibrium outcome $\{N_i^*, e_i^*, r^*\}$ associated with entry costs C_i and a fixed fee $f > 0$ can be implemented by a set of modified entry costs $\hat{C}_i = C_i - f\frac{MN_0}{N_i}$ without a fixed fee $\hat{f} = 0$.*

This proposition is straightforward to verify since the intermediaries' break-even condition (29) remains unchanged by the modification of entry costs and the elimination of the fixed fee f. Moreover, the effort incentives of all participants are unaffected by the fixed fee and entry cost modifications. Intuitively, the fixed pay f can be viewed as a subsidy to the entry cost C_i, and therefore, its removal does not impact the equilibrium outcome.

It is worth noting that with additional model features, it is possible to endogenize the fixed fee f. For instance, in a model where intermediaries can incur a cost K to become skilled, an additional equilibrium condition

$$\Pi_{S,i} - \Pi_{U,i} = K,$$

would arise, allowing for the determination of the equilibrium fixed fee f^*. However, as the focus of this paper is not on the determination of the compensation structure, this question is left for future research.

5.3 More than Two Skill Types

In this extension, the model relaxes the assumption of binary skill types and considers T discrete types of intermediaries with different screening technologies. Each intermediary type $t = 1, 2, ..., T$ in layer i can input effort $e_{i,t}$ and, based on the prior distribution $\delta_{i+1} \in \Delta^T$ can find an intermediary in layer $i + 1$ with a distribution $P_t(e_{i,t}, \delta_{i+1})$

$$P_t(e_{i,t}, \delta_{i+1}) \in \Delta^T \equiv \{x \in \mathbb{R}^T | x_i \in [0, 1], \sum x_i = 1\}, \tag{30}$$

Clearly, this setup nests the main model as a special case. The following T-dimensional screening matrix replaces the 2-dimensional matrix (1) in the main model.

$$\mathbb{P}_i \equiv \begin{pmatrix} P_1(e_{i,1}, \delta_{i+1}) \\ ... \\ P_t(e_{i,t}, \delta_{i+1}) \\ ... \\ P_T(e_{i,T}, \delta_{i+1}) \end{pmatrix}, \quad i = 0, 1, 2, ..., I. \tag{31}$$

I intuitively argue, without providing the technical details, that the insights from Sections 5.1 to 3 remain robust in this extended setup.

For the single-layer specification $I = 1$ as in Proposition 7, when both N_0

and $M \to \infty$, all types of intermediaries have a strong incentive to exert effort due to delegation as long as their effort inputs result in first-order stochastic dominance in P_t. As a result, investor payoff improves due to better screening outcomes at a negligible fee in equilibrium.

The intuition for the two-layer specification $I = 2$ similarly carries over. It is possible to design a pair of entry costs C_1 and C_2 such that only the most skilled type (denoted by type 1 without the loss of generality) enters either the first or the second layer. In the former case, those type-1 intermediaries in the first layer screen intensively and almost certainly find another type-1 intermediary in the second layer who in turn screens intensively in the asset market. As in Proposition 5, investors here do not need to screen. In the latter case, similar to Proposition 6, investors as well as all intermediaries in the first layer exert no effort, whereas the type-1 intermediaries in the second layer screens hard in the asset market. In both cases, investor payoff can approach the first best outcome.

Finally, when skill types are unknown and all intermediary layers share the same prior skill distribution denoted by δ, if intermediaries of a same type also use the same effort levels across layers, i.e., $e_{i,t} \equiv e_t$, then the screening matrix $\mathbb{P}_i$ is again layer-independent, similar to that in Proposition 1. As the chain becomes sufficiently long $I \to \infty$, the limiting matrix $\mathbb{P}^I$ depends on the converging properties of $\mathbb{P}$. Suppose a sufficiently high effort e_1 results in a type-1 intermediary almost certainly finding another type-1 intermediary or the best portfolio ($P_1(e_1, \delta) \to \begin{pmatrix} 1 & 0 & ... & 0 \end{pmatrix}$ as $e_1 \to \infty$) and other types have non-vanishing probabilities of finding a type-1 intermediary ($P_{t,1}(e_t, \delta) > \epsilon$ for any e_t and some $\epsilon > 0$), then the limiting probability of finding the best portfolio approaches 1. The structure in Proposition 3 remains valid as an implementation of the efficient outcome.

6 Conclusion

I introduce a model of intermediary market structure, focusing on key aspects such as skill distribution, sector size, effort choices, and the number of layers in the intermediation chain. The model demonstrates that superior quality or skill is not necessarily required for efficient intermediation. The findings also highlight the importance of self-selection and the length of the intermediation chain in achieving efficient outcomes.

Other features of the asset management industry can also be considered. For example, there may be different types of skills: Some agents are more skilled in selecting the superior portfolio, others may be more skilled in monitoring and improving the performance of the assets under management. How do different types of skills distribute along the chain?

The framework developed in this paper can serve as a valuable tool to model various other economic activities. For instance, when applied to corporate hierarchical structures, the model mirrors the hierarchical layers of managers, each responsible for monitoring and selecting their subordinates. This hierarchical cascade, akin to the intermediation chain, allows each layer to correct mistakes in lower layers but simultaneously introduces new errors into the system. In this context, some interesting questions could be the promotion pattern, career trajectory of workers, and job design in a corporation.

The model can also be applied to analyze platform economy, where platforms are skilled first layer intermediaries that screen the influencers – the second layer intermediaries who ultimately screen product markets. It is interesting to consider additional features such as rebates, quality choices by manufacturers, pricing, and consumer reviews on the influencers. The collection and strategic disclosure of information by the platforms may also yield new insights.

I look forward to future research can build upon this model to delve into these specific applications and uncover further insights and policy implications for various intermediary sectors.

References

[1] Admati, Anat R., and Paul Pfleiderer. "Direct and indirect sale of information." Econometrica: Journal of the Econometric Society (1990): 901-928.

[2] Adrian, Tobias, Adam Ashcraft, Hayley Boesky, and Zoltan Pozsar, "Shadow Banking," Federal Reserve Bank of New York Staff Reports, 2012, 2 (458), 603–618.

[3] Berk, Jonathan B., and Jules H. Van Binsbergen. "Regulation of charlatans in high-skill professions." The Journal of Finance 77.2 (2022): 1219-1258.

[4] Berk, Jonathan B., and Richard C. Green. "Mutual fund flows and performance in rational markets." Journal of political economy 112.6 (2004): 1269-1295.

[5] Biglaiser, Gary. "Middlemen as experts." The RAND journal of Economics (1993): 212-223.

[6] Biglaiser, Gary, and James W. Friedman. "Middlemen as guarantors of quality." International journal of industrial organization 12.4 (1994): 509-531.

[7] Calvo, Guillermo A., and Stanislaw Wellisz. "Hierarchy, ability, and income distribution." Journal of political Economy 87.5, Part 1 (1979): 991-1010.

[8] Chen, Cheng. "Management quality and firm hierarchy in industry equilibrium." American Economic Journal: Microeconomics 9.4 (2017): 203-244.

[9] Detrixhe, John. "There are now almost as many equity funds as there are stocks for them to invest in." Quartz, May 9th 2018

[10] Dasgupta, Amil, and Ernst G. Maug. "Delegation Chains." Available at SSRN 3952744 (2021).

[11] Diamond, Douglas W. "Financial intermediation and delegated monitoring." The review of economic studies 51.3 (1984): 393-414.

[12] Egan, Mark, Gregor Matvos, and Amit Seru. "The market for financial adviser misconduct." Journal of Political Economy 127.1 (2019): 233-295.

[13] Garicano, Luis. "Hierarchies and the Organization of Knowledge in Production." Journal of political economy 108.5 (2000): 874-904.

[14] Gârleanu, Nicolae, and Lasse Heje Pedersen. "Efficiently inefficient markets for assets and asset management." The Journal of Finance 73.4 (2018): 1663-1712.

[15] Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. "Money doctors." The Journal of Finance 70.1 (2015): 91-114.

[16] Glode, Vincent, and Christian Opp. "Asymmetric information and intermediation chains." American Economic Review 106.9 (2016): 2699-2721.

[17] Glode, Vincent, Christian C. Opp, and Xingtan Zhang. "On the efficiency of long intermediation chains." Journal of Financial Intermediation 38 (2019): 11-18.

[18] Glosten, Lawrence R., and Paul R. Milgrom. "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders." Journal of financial economics 14.1 (1985): 71-100.

[19] He, Zhiguo, and Jian Li. Intermediation via Credit Chains. No. w29632. National Bureau of Economic Research, 2022.

[20] Jiang, Hao, Dimitri Vayanos, and Lu Zheng. Tracking biased weights: Asset pricing implications of value-weighted indexing. No. w28253. National Bureau of Economic Research, 2020.

[21] Kaniel, Ron, and Dmitry Orlov. "Intermediated asymmetric information, compensation, and career prospects." Compensation, and Career Prospects (February 4, 2020) (2020).

[22] Lizzeri, Alessandro. "Information revelation and certification intermediaries." The RAND Journal of Economics (1999): 214-231.

[23] Qian, Yingyi. "Incentives and loss of control in an optimal hierarchy." The review of economic studies 61.3 (1994): 527-544.

[24] Ramakrishnan, Ram TS, and Anjan V. Thakor. "Information reliability and a theory of financial intermediation." The Review of Economic Studies 51.3 (1984): 415-432.

[25] Rubinstein, Ariel, and Asher Wolinsky. "Middlemen." The Quarterly Journal of Economics 102.3 (1987): 581-593.

[26] Stole, Lars A., and Jeffrey Zwiebel. "Intra-firm bargaining under non-binding contracts." The Review of Economic Studies 63.3 (1996): 375-410.

Appendix

Lemma 1 *Suppose $A_i \to \infty$ $(i = 1, 2, ...)$ is a sequence of positive coefficients and let $e_i^* = \arg\max_e A_i P(e,n) - c(e)$ denote the solution for any fixed $n \in (0,1)$. Then the average cost $\frac{c(e_i^*)}{A_i} \to 0$.*

Proof of Lemma 1: Consider the same problem redefined on the choice of probability p:

$$\max_p A_i p - \hat{c}(p), \tag{32}$$

where $\hat{c}(p) \equiv c(P^{-1}(p,n))$ and $e \equiv P^{-1}(p,n)$ – the inverse function of $P(e,n)$ with respect to e. Because $P(e,n)$ is increasing and concave in e, $P^{-1}(p,n)$ is increasing and convex in p. Hence, the compound function $\hat{c} = c \circ P^{-1}$ is increasing and convex in p. Since this optimization problem (32) is identical to the original problem, their solutions must also coincide $p_i^* = P(e_i^*, n)$. The first-order condition of (32) with respect to p is therefore given by

$$A_i = \hat{c}'(p_i^*). \tag{33}$$

Because $A_i \to \infty$, it must be $p_i^* \to 1$. Consider the following decomposition of $\hat{c}\left(p_i^*\right)$ and apply Darboux's theorem,

$$\hat{c}\left(p_i^*\right) = \sum_{j=1}^{i} \hat{c}\left(p_j^*\right) - \hat{c}\left(p_{j-1}^*\right. = \sum_{j=1}^{i} \hat{c}'\left(p_j^\dagger\right)\left(p_j^* - p_{j-1}^*\right. \leq \sum_{j=1}^{i} \hat{c}'\left(p_j^*\right. \left(p_j^* - p_{j-1}^*\right.$$

where $c(p_0^*)$ is conveniently defined as 0, $p_j^\dagger \in \left[p_{j-1}^*, p_j^*\right]$ is some intermediate value, and the last inequality utilizes the convexity of $\hat{c}$. In addition, $0 < p_j^* - p_{j-1}^* \to 0$. Consequently,

$$\lim_{i\to\infty} \frac{c(e_i^*)}{A_i} = \lim_{i\to\infty} \frac{\hat{c}\left(p_i^*\right)}{A_i} \leq \lim_{i\to\infty} \frac{\sum_{j=1}^{i} A_j \left(p_j^* - p_{j-1}^*\right)}{A_i}.$$

Now I show that the last term converges to 0. Consider any ϵ, there exits an k such that $1 - p_k^* < \frac{\epsilon}{2}$. For this k, there exits an $i > k$ such that $\frac{A_k}{A_i} < \frac{\epsilon}{2}$. It then follows

$$\frac{\sum_{j=1}^{i} A_j \left(p_j^* - p_{j-1}^*\right)}{A_i} = \frac{\sum_{j=1}^{k} + \sum_{j=k+1}^{i} A_j \left(p_j^* - p_{j-1}^*\right)}{A_i} < \frac{\epsilon \sum_{j=1}^{k} \left(p_j^* - p_{j-1}^*\right)}{2} + (p_i^* - p_k^*) < \epsilon.$$

This completes the proof for $\frac{c(e_i^*)}{A_i} \to 0$. ■

Proof of Proposition 1: Suppose $i = I - I_E + 1$ is the first layer in which the skilled intermediaries make effort. In addition to the effort condition for the ith layer (7), the $i-1$ layer should not have effort incentive and therefore

$$\frac{N_0}{N_I} r^*_{I_E - 1} \left(p_1 - n_S\right)^{I_E + 1} < c.$$

Taken together, they imply

$$I_E = \quad \log_{p_1 - n_S} \frac{cN_I}{r^*_{I_E} N_0} \quad .$$

To calculate $P^{(I_E)}$, note that the screening matrix can be diagonalized as follows

$$\mathbb{P} = \begin{pmatrix} 1 & 1-p_1 \\ 1 & -n_S \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & p_1 - n_S \end{pmatrix} \begin{pmatrix} p^\dagger & 1 - p^\dagger \\ \frac{1}{1-p_1+n_S} & -\frac{1}{1-p_1+n_S} \end{pmatrix},$$

where

$$\begin{pmatrix} p^\dagger & 1 - p^\dagger \\ \frac{1}{1-p_1+n_S} & -\frac{1}{1-p_1+n_S} \end{pmatrix} = \begin{pmatrix} 1 & 1-p_1 \\ 1 & -n_S \end{pmatrix}^{-1}.$$

Hence,

$$\begin{aligned}
\mathbb{P}^i &= \begin{pmatrix} 1 & 1-p_1 \\ 1 & -n_S \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \left(p_1 - n_S\right)^i \end{pmatrix} \begin{pmatrix} p^\dagger & 1 - p^\dagger \\ \frac{1}{1-p_1+n_S} & -\frac{1}{1-p_1+n_S} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \left(1-p_1\right)\left(p_1 - n_S\right)^i \\ 1 & -n_S\left(p_1 - n_S\right)^i \end{pmatrix} \begin{pmatrix} p^\dagger & 1 - p^\dagger \\ \frac{1}{1-p_1+n_S} & -\frac{1}{1-p_1+n_S} \end{pmatrix} \\
&= \begin{pmatrix} \frac{n_S + (1-p_1)(p_1-n_S)^i}{1-p_1+n_S} & \frac{(1-p_1)-(1-p_1)(p_1-n_S)^i}{1-p_1+n_S} \\ \frac{n_S - n_S(p_1-n_S)^i}{1-p_1+n_S} & \frac{1-p_1+n_S(p_1-n_S)^i}{1-p_1+n_S} \end{pmatrix} \\
&= \begin{pmatrix} p^\dagger + \left(1-p^\dagger\right)\left(p_1 - n_S\right)^i & 1 - p^\dagger - \left(1-p^\dagger\right)\left(p_1 - n_S\right)^i \\ p^\dagger - p^\dagger\left(p_1 - n_S\right)^i & 1 - p^\dagger + p^\dagger\left(p_1 - n_S\right)^i \end{pmatrix}.
\end{aligned}$$

When $I_E < I$, some initial layers together with the investor sector do not make effort. Hence, the success probability is

$$\begin{aligned}
P^{(I_E)} &= \begin{pmatrix} n_S & 1 - n_S \end{pmatrix} \mathbb{P}^{I_E} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} n_S & 1 - n_S \end{pmatrix} \begin{pmatrix} p^\dagger + \left(1-p^\dagger\right)\left(p_1 - n_S\right)^{I_E} \\ p^\dagger - p^\dagger\left(p_1 - n_S\right)^{I_E} \end{pmatrix} \\
&= p^\dagger - \left(p^\dagger - n_S\right)\left(p_1 - n_S\right)^{I_E} = p^\dagger \left[1 - \left(p_1 - n_S\right)^{I_E + 1}\right].
\end{aligned}$$

■

Proof of Proposition 2: First, note that $r^* \leq 1$. Hence, condition (8) implies that as I becomes sufficiently large

$$I_E \leq \log_{p_e - n_S} \frac{c_e}{IN_0} = \log_{p_e - n_S} \frac{c_e}{N_0} - \log_{p_e - n_S} I = a \ln I + b,$$

where ln is the natural logarithm, and $a \equiv -\frac{1}{\ln(p_e - n_S)} > 0$ and $b \equiv \log_{p_e - n_S} \frac{c_e}{N_0}$ are two constants. Consequently as $I \to \infty$, the ratio $\frac{I_E}{I} \to 0$.

Next, I show that for any $I_E \geq 1$, $\frac{I_E}{I} < \epsilon$ for some small $\epsilon > 0$, effort condition (7) fails to hold. To see this, note that condition (13) implies

$$r_e^* < \frac{\epsilon}{(I+1)\,P^{(I_E)}} \sum_{s=0}^{I_E-1} \quad P^{(s+1)} - P^{(s)} - c_e n_S \frac{1}{IN_0} \quad + c_e n_S \frac{1}{P^{(I_E)} IN_0} < \frac{\epsilon}{(I+1)\,P^{(I_E)}} + c_e n_S \frac{1}{P^{(I_E)} IN_0}.$$

Since this condition holds for any ϵ, the fact that $P^{(I_E)} > P^{(0)} = n_S$ implies

$$r_e^* < \frac{c_e}{IN_0},$$

when I is sufficiently large. Recall $N_I = \frac{1}{I}$, the above condition immediately implies that

$$\frac{N_0}{N_I} r_e^* \left(p_e - n_S\right)^{I_E} < \frac{N_0}{N_I} r_e^* < c_e.$$

Hence, it must be that $I_E = 0$ when I becomes sufficiently large. ■

Proof of Proposition 3: First, I show there exists an equilibrium where at least some intermediaries make effort $I_E \geq 1$. Using (7) and (13), I want to show that

$$IN_0 \left(p_e - n_S\right) \quad \frac{1}{(I+1)\,P^{(1)}} \frac{1}{I} \quad P^{(1)} - P^{(0)} - c_e n_S \frac{1}{IN_0} \quad + c_e n_S \frac{1}{P^{(1)} IN_0} \quad \geq c_e.$$

Using the fact that $I = N_0^\alpha$ for some $\alpha \in (0,1)$, the left-hand side is bounded below by

$$IN_0 \left(p_e - n_S\right) \frac{1}{(I+1)\,P^{(1)}} \frac{1}{I} \quad P^{(1)} - P^{(0)} \quad = \frac{N_0}{I+1} \left(p_e - n_S\right) \frac{P^{(1)} - P^{(0)}}{P^{(1)}} \to \infty.$$

Therefore, $I_E \geq 1$.

Second, I estimate a lower bound on r_e^*. From (13), we have

$$r_e^* \geq \frac{1}{P^{(I_E)}\,(I+1)} \sum_{s=0}^{I_E-1} \frac{s+1}{I} \quad P^{(s+1)} - P^{(s)} \Big) > \frac{1}{(I+1)^2} \quad 1 - \frac{P^{(0)}}{P^{(I_E)}} \quad .$$

Proposition 1 and expression (8) imply that

$$I_E = \log_{p_e - n_S} \frac{c_e}{r_e^* N_0^{1+\frac{1}{\alpha}}} > (-1 + \frac{1}{\alpha}) \log_{p_e - n_S} N + \text{const.}$$

Hence, the limit is bounded by

$$\lim_{N_0 \to \infty} \frac{I_E}{\ln N_0} \geq -\frac{1 - \frac{1}{\alpha}}{\ln\left(p_e - n_S\right)}.$$

Third, r_e^* also has an upper bound from (13):

$$r_e^* \leq \frac{I_E}{I^2} \quad 1 - \frac{P^{(0)}}{P^{(I_E)}} \quad + c_e n_S \frac{1}{IN_0}. \tag{34}$$

Again from expression (8)

$$I_E \leq (-1 + \frac{1}{\alpha}) \left(\log_{p_e - n_S} N_0 + \log_{p_e - n_S} I_E\right) + \text{const.}$$

Note that $I_E \to \infty$ implies that $\frac{\ln I_E}{I_E} \to 0$. Hence,

$$\lim_{N_0} \frac{I_E}{\ln N_0} \leq -\frac{1-\frac{1}{\alpha}}{\ln\left(p_e - n_S\right)}.$$

Taken together, we have established (15).

I now show that investor payoffs converge to $p^{\dagger}$. Since $I_E \sim \ln N_0 \to \infty$, condition (9) implies that the success probability is $p^{\dagger}$. Furthermore, because investors do not make effort in equilibrium, there is no effort cost and both skilled and unskilled investors receive the same payoff. Finally, from the upper bound on r_e^* in (34) and the fact that $\frac{I_E}{I} \sim \frac{\ln N_0}{N_0^{\alpha}} \to 0$, we have

$$r_e^* I \leq \frac{I_E}{I} \quad 1 - \frac{P^{(0)}}{P^{(I_E)}} \quad + c_e n_S \frac{1}{N_0} \to 0,$$

as $N_0 \to \infty$. The total intermediation fee therefore

$$\lim_{N_0 \to \infty} \left(I - I_E\right) r_{ne}^* + I_E r_e^* \to 0,$$

and $r_0 = 1$. In summary, investor payoff

$$\lim_{N_0 \to \infty} \Pi_{S,0} = \lim_{N_0 \to \infty} \Pi_{U,0} = r_0^* p^{\dagger} = p^{\dagger}.$$

Finally, since total welfare in the economy is bounded below by total investor payoff and bounded above by the total payoff (excluding effort cost) to investors, it is immediate that

$$p^{\dagger} \leq \lim_{N_0 \to \infty} \frac{W_I}{N_0} \leq \lim_{N_0 \to \infty} \frac{N_0 p^{\dagger}}{N_0}.$$

This completes the proof of the proposition. ■

Proof of Proposition 4: First, when condition (16) holds which automatically implies (17), a long intermediation chain always improves welfare and investor payoff:

$$W_\infty = p^{\dagger} > n_S = W_0.$$

Next, suppose (16) fails. From (10), we have

$$\begin{aligned} & W_\infty \geq W_0 \\ \Leftrightarrow \quad & p^{\dagger} \geq n_0\left(p_e - c_e\right) + \left(1 - n_0\right) n_S \\ \Leftrightarrow \quad & \tfrac{n_S}{1 - p_e + n_S} \geq n_0\left(p_e - c_e\right) + \left(1 - n_0\right) n_S \\ \Leftrightarrow \quad & n_0 c_e \geq n_0 p_e + \left(1 - n_0\right) n_S - \tfrac{n_S}{1 - p_e + n_S} \end{aligned}.$$

Dividing by n_0 and after some manipulation, the above condition becomes

$$c_e \geq \left(p_e - n_S\right) - \frac{n_S}{n_0} \frac{p_e - n_S}{1 - p_e + n_S} = \left(p_e - n_S\right)\left(1 - \frac{n_S}{n_0\left(1 - p_e + n_S\right)}\right. .$$

Hence, the conclusion.■

Proof of Proposition 5: In this conjectured equilibrium, investors (regardless of their skill levels) do not need to make any effort $e^* = 0$ as all layer 1 intermediaries are skilled $N_{U,1}^* = 0$. The remaining four equilibrium variables, including skilled intermediaries' effort levels e_1^* and e_2^*, intermediary distribution $N_{S,1}^*$ (or equivalently, $n_2^* = \frac{n_S - N_{S,1}^*}{1 - N_{S,1}^*}$), and price

r^* are determined by the following conditions:

1. first-order condition of skilled layer 1 intermediaries

$$M\frac{1}{N^*_{S,1}}r^*N_0\left[P(e_2^*,n_P)-P(0,n_P)\right]\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}=c'(e_1^*), \tag{35}$$

2. first-order condition of skilled layer 2 intermediaries

$$\frac{P(e_1^*,n_2^*)}{n_S-N^*_{S,1}}r^*N_0\frac{\partial P(e_2^*,n_P)}{\partial e_2^*}=c'(e_2^*), \tag{36}$$

3. skilled intermediaries indifference condition between the two layers

$$\begin{aligned}&M\tfrac{1}{N^*_{S,1}}r^*N_0\left[P(e_1^*,n_2^*)P(e_2^*,n_P)+\left(1-P(e_1^*,n_2^*)\right)P(0,n_P)\right]-c(e_1^*)-C_1\\&=M\left[\tfrac{P(e_1^*,n_2^*)}{n_S-N^*_{S,1}}r^*N_0P(e_2^*,n_P)-c(e_2^*)\right]-C_2\geq 0,\end{aligned} \tag{37}$$

4. unskilled intermediaries break even in layer 2

$$M\frac{1-P(e_1^*,n_2^*)}{1-n_S}r^*N_0P(0,n_P)=C_2, \tag{38}$$

5. and finally, unskilled intermediaries do not wish to enter layer 1 (an inequality constraint that needs to satisfy in equilibrium)

$$M\frac{1}{N^*_{S,1}}r^*N_0\left[P(0,n_1)P(e_2^*,n_P)+\left(1-P(0,n_1)\right)P(0,n_P)\right]\leq C_1. \tag{39}$$

We consider the following conjugate problem: finding the entry costs C_1 and C_2 such that an equilibrium exists with the desired properties in the proposition. Choose a sequence of $N^*_{S,1}\to 0$ and $e_1^*\to\infty$ as $N_0\to\infty$, such that $\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}=N_0^{x_1}$ and $N^*_{S,1}=N_0^{x_2}$, where $x_2-1<x_1<x_2<0$. Such a choice is possible because $\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}$ is decreasing in e_1^* and approaches 0 as $e_1^*\to\infty$.

Taking the ratio between (35) and (36), we have

$$\frac{\frac{\partial P(e_2^*,n_P)}{\partial e_2^*}}{c'(e_2^*)}=M\frac{n_S-N^*_{S,1}}{N^*_{S,1}}\frac{P(e_2^*,n_P)-P(0,n_P)}{P(e_1^*,n_2^*)}\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}. \tag{40}$$

Since $\frac{\frac{\partial P(e_2^*,n_P)}{\partial e_2^*}}{c'(e_2^*)}$ is decreasing in e_2^* with a full support on $(0,\infty)$ and the right-hand side is increasing in e_2^*, a solution must uniquely exist. In addition, the sequence of $\left(N^*_{S,1},e_1^*\right)$ is chosen such that

$$\lim_{N_0\to\infty}\frac{\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}}{N^*_{S,1}}=\lim_{N_0\to\infty}N_0^{x_1-x_2}=0\text{ and }\lim_{N_0\to\infty}\frac{\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}}{N^*_{S,1}}N_0=\lim_{N_0\to\infty}N_0^{1+x_1-x_2}=\infty, \tag{41}$$

then (40) implies that its solution $e_2^*\to\infty$. In addition, (35) implies that $r^*\to 0$ and $r^*N_0\to\infty$. Conditions (37) and (38) jointly determine $C_2>0$ and C_1.

Next, I show that $C_1 \gg C_2$. Manipulate (37), we have

$$\begin{aligned} & M\left[\tfrac{P(e_1^*,n_2^*)}{n_S-N_{S,1}^*}r^*N_0P(e_2^*,n_P)-c(e_2^*)\right]-C_2 \\ =\ & M\tfrac{1}{N_{S,1}^*}r^*N_0\left[P(e_1^*,n_2^*)P(e_2^*,n_P)+\left(1-P(e_1^*,n_2^*)\right)P(0,n_P)\right]-c(e_1^*)-C_1 \\ \geq\ & M\tfrac{1}{N_{S,1}^*}r^*N_0\left[P(e_2^*,n_2^*)P(e_2^*,n_P)+\left(1-P(e_2^*,n_2^*)\right)P(0,n_P)\right]-c(e_2^*)-C_1, \end{aligned}$$

where the last inequality is due to the optimality of e_1^*. Let $N_0 \to \infty$, using the fact that $P(e_1^*,n_2^*), P(e_2^*,n_2^*), P(e_2^*,n_P) \to 1$, and $N_{S,1}^* \to 0$, we have

$$C_1 - C_2 \geq M\frac{1}{N_{S,1}^*}r^*N_0 - M\frac{1}{n_S-N_{S,1}^*}r^*N_0 + (M-1)\,c(e_2^*).$$

Note from (41) that $r^*N_0 \to \infty$ and $N_{S,1}^* \to 0$, so $C_1 \gg C_2$.

Furthermore, I verify the inequality condition (39). Using (37) and (38), we know that (39) is equivalent to

$$\begin{aligned} & \tfrac{1}{N_{S,1}^*}\left\{\left[P(e_1^*,n_2^*)-P(0,n_1)\right]\left[P(e_2^*,n_P)-P(0,n_P)\right]-\tfrac{c(e_1^*)N_{S,1}^*}{Mr^*N_0}\right\} \\ \geq\ & \left[\tfrac{P(e_1^*,n_2^*)}{n_S-N_{S,1}^*}P(e_2^*,n_P)-\tfrac{1-P(e_1^*,n_2^*)}{1-n_S}P(0,n_P)\right]-\tfrac{c(e_2^*)}{r^*N_0} \end{aligned} \tag{42}$$

Apply Lemma 1 to condition (35) implies that $\frac{c(e_1^*)N_{S,1}^*}{Mr^*N_0} \to 0$, as $N_{S,1}^* \to 0$. It immediately follows that the left-hand side of (42) approaches ∞ as $N_{S,1}^* \to 0$, whereas the right-hand side is bounded. Hence, the inequality condition (39) holds.

Finally, I show investor payoff $\Pi_{S,0} = \Pi_{U,0} \to M$. Since $e_0^* = 0$, $r^* \to 0$, and $P(e_1^*,n_2^*), P(e_2^*,n_P) \to 1$, we have

$$\Pi_{S,0} = \Pi_{U,0} \geq M(1-2r^*)P(e_1^*,n_2^*)P(e_2^*,n_P) \to M,$$

completing the proof. ■

Proof of Proposition 6: In this conjectured equilibrium, skilled intermediaries in the first layer do not exert effort $e_1^* = 0$ because all second layer intermediaries are skilled. Similarly investors do not make any effort $e_0^* = 0$ because layer 1 intermediaries either are unskilled or do not exert effort. Hence, the remaining equilibrium variables r^*, $N_{S,2}^*$, and e_2^* solve

$$\Pi_{S,1} = \Pi_{U,1} = M\frac{1}{1-N_{S,2}^*}r^*N_0P(e_2^*,n_P) - C_1 = 0, \tag{43}$$

$$\Pi_{S,2} = M\left[\frac{1}{N_{S,2}^*}r^*N_0P(e_2^*,n_P) - c(e_2^*)\right] - C_2 = 0, \tag{44}$$

and

$$Mr^*N_0\frac{\partial P(e_2^*,n_P)}{\partial e_2^*} = N_{S,2}^*c'(e_2^*). \tag{45}$$

Using (43) and (44) to eliminate r^*, we have

$$\frac{N_{S,2}^*}{1-N_{S,2}^*} = \frac{C_1}{Mc(e_2^*)+C_2}. \tag{46}$$

Since the left-hand side ranges from 0 to ∞ as $N_{S,2}^*$ increases from 0 to 1, the solution $N_{S,2}^*(e_2^*) < n_S$ exists for any e_2^* and $C_2 \gg C_1$. In addition, whenever $e_2^* \to \infty$ or $C_2 \gg C_1$, we have $N_{S,2}^* \to 0$. Finally, to show such an equilibrium e_2^* always exists, from (43) and

(45), we have

$$\frac{\frac{\partial P(e_2^*, n_P)}{\partial e_2^*}}{P(e_2^*, n_P)} = \frac{c'(e_2^*)}{Mc(e_2^*) + C_2}. \tag{47}$$

To see that (47) aways has solution, note that it is in fact the first-order condition associated with the following maximization problem

$$\max_e G(e) \equiv M \log P(e, n_P) - \log \left[c(e) + \frac{C_2}{M} \right. .$$

Clearly, since $G(\infty) \to -\infty$ and $G'(0) > 0$, there must exist an interior maximum, which is also a solution e_2^* to (47).

I now show $P(e_2^*, n_P) \to 1$ as $C_2 \to \infty$. Suppose otherwise, $P(e_2^*, n_P)$ is bounded away from 1, which in turn implies that $\frac{\partial P(e_2^*, n_P)}{\partial e_2^*}$ is bounded away from 0. In addition, $c(e_2^*)$ and $c'(e_2^*)$ are both bounded away from infinity. However, this contradicts with (47) because the right-hand side approaches 0 as $C_2 \to \infty$. Hence, it must be $P(e_2^*, n_P) \to 1$.

Finally, note that in (43), $N_{S,2}^* \to 0$, and $C_1 \to 0$, and it quickly follows $r^* \to 0$. Therefore, the investor payoff

$$\Pi_0 = M(1 - 2r^*) P(e_2^*, n_P) \to M,$$

completing the proof. ■

Proof of Proposition 7: It is useful to explicitly write down the intermediaries' payoff functions from (2) and (3):

$$\Pi_{S,1} = \max_{e_1} M \left. \frac{P(e_0, n_S) n_0 + P(0, n_S)(1 - n_0)}{n_S} r N_C P(e_1, n_P) - c(e_1) \right] - C_1. \tag{48}$$

$$\Pi_{U,1} = M \frac{[1 - P(e_0, n_S)] n_0 + [1 - P(0, n_S)](1 - n_0)}{1 - n_S} r N_C P(0, n_P) - C_1. \tag{49}$$

I first show that the effort inputs of both skilled investors and intermediaries approach infinity: $e_0^*, e_1^* \to \infty$. Because $\Pi_{U,1} \leq \Pi_{S,1}$, the breakeven condition (23) implies that $\Pi_{S,1} \geq 0$. In addition,

$$\frac{P(e_0, n_S) n_0 + P(0, n_S)(1 - n_0)}{n_S} \in [1, \frac{n_0 + n_S(1 - n_0)}{n_S}]$$

which is bounded by two constants. Condition (48) together with the fact that $\Pi_{S,1} \geq 0$ imply that

$$r^* \geq \frac{C_1}{MN_0 P(e_1, n_P) \frac{P(e_0, n_S) n_0 + P(0, n_S)(1 - n_0)}{n_S}} = a'(N_0)^{b-1}, \tag{50}$$

for some constant $a' > 0$. Condition (48) also implies that the intermediary's effort choice $e_1^* \to \infty$, as $N_0 \to \infty$.

Furthermore, Lemma 1 implies that the average cost of screening effort in condition (48) vanishes

$$\frac{c(e_1)}{N_0} \to 0.$$

Condition (49) together with the fact that $\Pi_{U,1} \leq 0$ imply that

$$r^* \leq \frac{C_1}{MN_0 P(0,n_P)^{\frac{[1-P(e_0,n_S)]n_0+[1-P(0,n_S)](1-n_0)}{1-n_S}}}.$$

Note that

$$\frac{[1-P(e_0,n_S)]\,n_0+[1-P(0,n_S)]\,(1-n_0)}{1-n_S} \in [1-n_0, 1],$$

is bounded between two constants. Hence, condition (50) can be strengthened to

$$r^* \sim (N_0)^{b-1} \to 0, \tag{51}$$

where the convergence to 0 is independent of M and holds whenever $N_0 \to \infty$.

Skilled investors' effort problem can be explicitly written as

$$\Pi_{S,0} = \max_{e_0} M(1-r^*)\left\{P(e_0,n_S)P(e_1,n_P) + [1-P(e_0,n_S)]\,P(0,n_P)\right\} - c(e_0). \tag{52}$$

Since $M(1-r^*)\,[P(e_1^*,n_P) - P(0,n_P)] \to \infty$ as $M \to \infty$, the first order condition of (52) immediately implies that $e_0^* \to \infty$.

Next, apply Lemma 1 to the skilled investor's optimization problem in (52), and it implies that $\frac{c(e_0^*)}{M} \to 0$.

We can now calculate investor payoffs in the limiting case as $M, N_0 \to \infty$. For skilled investors,

$$\frac{\Pi_{S,0}}{M} = (1-r^*)\left\{P(e_0^*,n_S)P(e_1^*,n_P) + [1-P(e_0^*,n_S)]\,P(0,n_P)\right\} - \frac{c(e_0^*)}{M} \to 1.$$

For unskilled investors,

$$\frac{\Pi_{U,0}}{M} = (1-r^*)\left\{P(0,n_S)P(e_1^*,n_P) + [1-P(0,n_S)]\,P(0,n_P)\right\} \to P(0,n_S) + [1-P(0,n_S)]\,P(0,n_P),$$

completing the proof. ■