DISCUSSION PAPER SERIES

DP17619

MARKET STRUCTURE OF INTERMEDIATION

Hongda Zhong

INDUSTRIAL ORGANIZATION AND BANKING AND CORPORATE FINANCE

ISSN 0265-8003

MARKET STRUCTURE OF INTERMEDIATION

Hongda Zhong

Discussion Paper DP17619
Published 30 October 2022
Submitted 18 October 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Industrial Organization
- Banking and Corporate Finance

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Hongda Zhong

MARKET STRUCTURE OF INTERMEDIATION

Abstract

Finding good products often requires costly screening effort. A skilled intermediary can represent many consumers, generating economies of scale in effort costs. However, consumers still need to find skilled intermediaries, re-creating the original screening friction. I provide a model of intermediary market structure, focusing on the length of the intermediation chain, the sector size, skill distribution, and effort choice in each layer. Even when the intermediary sector is less skilled than consumers and has lower quality than the product market, efficient intermediation can occur. Furthermore, when intermediaries privately know their skill types, a short (two-layer) chain allows different types to self-select into different layers, approximating the first-best outcome: consumers receive good products with negligible screening costs. Finally, when skill types are unknown, thereby rendering intermediaries' self-selection ineffective, a long chain can restore first-best efficiency. These results shed light on the structure of social media influencers and the asset management industry.

JEL Classification: D40, G00, L11

Keywords: Screening

Hongda Zhong - zhonghongda@gmail.com
London School of Economics and CEPR

Market Structure of Intermediation*

Hongda Zhong†

September 29, 2022

Abstract

Finding good products often requires costly screening effort. A skilled intermediary can represent many consumers, generating economies of scale in effort costs. However, consumers still need to find skilled intermediaries, re-creating the original screening friction. I provide a model of intermediary market structure, focusing on the length of the intermediation chain, the sector size, skill distribution, and effort choice in each layer. Even when the intermediary sector is less skilled than consumers and has lower quality than the product market, efficient intermediation can occur. Furthermore, when intermediaries privately know their skill types, a short (two-layer) chain allows different types to self-select into different layers, approximating the first-best outcome: consumers receive good products with negligible screening costs. Finally, when skill types are unknown, thereby rendering intermediaries' self-selection ineffective, a long chain can restore first-best efficiency. These results shed light on the structure of social media influencers and the asset management industry.

Key Words: Screening, social media influencers, funds, index, intermediation chain

*I thank Samuel Antill, Ulf Axelson, Philip Bond, Mike Burkart, Georgy Chabakauri, Amil Dasgupta, Daniel Ferreira, Alessandro Gavazza, Vincent Glode, Zhiguo He, Peter Kondor, Dong Lou, Martin Oehmke, Christian Opp, Lasse Pedersen, Dimitri Vayanos, Sergio Vicente, Ansgar Walther, Andrew Winton, Shengxing Zhang, 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, the Finance Theory Group (Budapest) Summer Conference, the Oxford Financial Intermediation Theory Conference, and Banking Theory Junior Brown Bag Seminar. 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.

†London School of Economics and CEPR, h.zhong2@lse.ac.uk

1 Introduction

Many transactions are facilitated by intermediation chains, in which upstream intermediaries screen and select downstream intermediaries, who ultimately select products. For instance, rather than making self-directed investments, an investor can choose a pension fund or a fund-of-funds manager, who in turn selects the underlying hedge funds or private equity funds, which ultimately select portfolios of underlying assets. As another example, instead of visiting shopping malls, consumers nowadays can choose from social media platforms (YouTube, TikTok, or Instagram) which in turn select and recommend a customized set of influencers to consumers. The influencers eventually select and recommend products to their followers. Despite the prevalence of intermediation in various economic activities, its market structure is far from obvious and sometimes puzzling. For example, 42% of equity index mutual funds track the S&P 500 index (Jiang, Vayanos, and Zheng 2020). Why is there not just one fund, given that these S&P 500 tracking funds essentially provide homogeneous services? Why would consumers shop online through essentially unregulated influencers rather than visit shopping malls that typically have some quality control measures? Why are some intermediation chains short while others are long?[1]

To shed light on these questions, I construct a parsimonious model of intermediation chain, focusing on its length, the sector size, skill distribution, and effort choices in each layer of the chain. Specifically, I consider an economy where consumers need to consume a set of products and can rely on an I-layer intermediation chain to improve their chance of receiving good products. Each consumer chooses an intermediary from the first layer, who in turn chooses an intermediary from the second layer, and so on. Those in the final layer eventually choose a product from each product market. Some products are good, generating one unit of utility, whereas others are bad and do not generate any utility. Likewise, some intermediaries and consumers are skilled, and the rest are unskilled. Skilled ones can exert costly effort to increase the probability of finding good products or skilled intermediaries in the next layer. Unskilled ones can only find a random product or intermediary. Intermediaries receive performance-based fees from consumers whenever a good product flows to consumers through them. These fees also provide effort incentives and are determined in equilib-

[1]A good example of a long intermediation chain in the financial market is the shadow banking credit intermediation chain, see Adrian et al (2012) and He and Li (2022).

rium by the worst intermediaries' breakeven condition. Intermediaries with different skill types are in fixed supply. Upon observing a noisy signal of their types, they can decide where to operate in the chain by paying the entry costs associated with that layer.[2] Such a self-selection by different intermediaries generates the endogenous market structure (size, skill distribution, and effort choice along the intermediation chain), which is the key focus of this paper. In Subsection 5.2, I further introduce an arbitrarily small dissipative cost associated with each layer and endogenize the optimal chain length. Despite the exponentially growing number of outcomes as the chain becomes longer, the matrix algebra keeps the model in a tractable form.

When consumers directly screen product markets without intermediaries ($I = 0$), two inefficiencies arise. First, all (skilled) consumers need to individually exert costly efforts to screen product markets. Second, the probability of finding a good product is bounded away from 1. Having intermediaries improves both aspects due to consumer delegation. Even though the intermediation fee from each consumer is negligible, the collective amount can be significant such that skilled intermediaries have large effort incentives and almost certainly deliver good products to many consumers. However, intermediation also comes at a cost: consumers now need to screen the intermediary market and find a skilled one. Whether intermediation is efficient is not obvious.

Three main insights emerge from the model.

First, efficient intermediation does not rely on the intermediary sector having superior skill than consumers or better quality than the product markets. This is because intermediation creates two-sided economies of scale from both receiving many consumers and screening many product markets. The intuition is best illustrated with a single layer of intermediation ($I = 1$, Section 3).[3] On the one hand, because a skilled intermediary can screen many products on behalf of consumers, skilled consumers exert high effort to almost certainly find a skilled intermediary. This effect is robust to the possibility that finding a skilled intermediary may be more difficult than finding a good product. On the other hand, due to lots of consumer delegation, skilled intermediaries exert high effort to almost certainly find good products, compensating for the possibility that the intermediary sector may have worse skills than consumers on

[2]This entry cost can be interpreted as the necessary investment in order for an intermediary to operate in that particular layer. Alternatively, it could be interpreted as the compliance costs to meet the necessary regulatory requirements set by the regulators.

[3]Having multiple layers $I > 1$ strengthens this result in that it no longer requires large numbers of consumers or markets (Sections 4 and 5).

average. Consequently, the skilled consumers can achieve first-best payoff asymptotically: they almost certainly receive good products from each market with arbitrarily small screening/compensation costs per product. The unskilled consumers' payoff is also improved. However, because they lack the ability to find a skilled intermediary through effort input, their probability of finding a good product is bounded away from 100%. While the market structure of single-layer intermediation is largely exogenous and the mechanism is related to Diamond (1984), the analysis is a useful benchmark and leads to more complex and efficient chain structures that are novel to the literature.

The second insight stemming from the model is that a short two-layer intermediation chain ($I = 2$, Section 4) can fully resolve the inefficiencies when intermediaries can self-select into layers based on their skill types (i.e., intermediaries receive precise signals about their types). Structures with a heterogeneous skill distribution across the two layers can deliver asymptotic first-best. More specifically, there is one high-quality layer consisting only of skilled intermediaries, and the rest (all unskilled and the remaining skilled) enter the other layer. The benefit of having the high-quality layer is that consumers no longer need to exert screening effort since they always find a skilled intermediary in this high-quality layer. If the first layer is high quality (Subsection 4.1), they screen the second layer intensively and almost certainly find skilled ones, who in turn screen the product market intensively, delivering good products. If instead the second layer is high quality (Subsection 4.2), they screen the product markets, and no one else makes any effort in the chain. In both cases, consumers' payoff approximates first best. The high-quality layer is typically associated with high entry cost that deters entry of unskilled intermediaries and has smaller size.

These two-layer structures resemble several intermediation chains in practice. The social media platform and influencer chain is a good example where the first layer (platform) has high quality. The entry cost to building a social media platform is much higher than the entry cost to becoming an influencer. As a consequence, there is only a small number of platforms (e.g.,YouTube, Instagram, or TikTok) who are very skilled at selecting from a vast number of channels/influencers and recommending suitable ones to viewers/consumers. Consumers do not need to screen the handful of platforms and often receive good recommendations. Influencers typically have mixed qualities. The good ones, upon receiving lots of followers (potential consumers) from platforms, work hard to satisfy their preferences. Another example with a high quality second layer is the fund-of-funds and underlying hedge fund or PE fund chain. Fund managers, regardless of layers, have arguably comparable entry costs in that they are subject to similar regulatory requirements and need similar financial knowledge. The model implies that the second layer (hedge funds or PE funds) is populated with skilled managers. Some skilled and all unskilled managers work for fund-of-funds, but none needs to exert effort because they can rely on the underlying funds to generate returns. This prediction is supported by the anecdotal story that hedge funds and private equity funds attract top talent in the financial industry.

The third insight is that even when intermediaries cannot effectively self-select (e.g., only receive noisy or uninformative signals about their types), a long intermediation chain can restore the first-best outcome asymptotically (Subsection 5.1). Two effects drive this result. First, as the chain becomes sufficiently long, there can be sufficiently many small layers at the end of the intermediation chain. Because of the small sector size, skilled intermediaries in those final layers therefore receive a large consumer delegation. They exert high effort and almost certainly find other skilled ones in the next layer (or a good product if in the final layer), thereby creating an absorbing state. Having sufficiently many such layers ensures that consumers eventually find a good product (or equivalently, having sufficiently many opportunities to reach this absorbing state). Second, consumers and intermediaries at the front of the chain do not need to exert effort, because they can free ride on the convergence property. Hence, consumers almost certainly receive good products with negligible effort costs, restoring the first best. Several theoretical implications emerge from the convergence property. First, most of the inefficiencies are resolved by the final layers of the chain. In addition, it is exponentially more difficult to incentivize effort in initial layers. Finally, the diminishing marginal benefit of additional layers allows me to endogenize the optimal chain length with a small fixed dissipative cost associated with each layer (Subsection 5.2)

Finally, I offer some insights for a more general compensation package including the possibility of a fixed fee (Subsection 6.1) and layer-dependent compensation (Subsection 6.2). As compensation is not the key focus of this paper, I leave a more comprehensive analysis of these interesting topics to future research. The insights are also robust to introducing more than two skill types (Subsection 6.3).

Literature Review

There is a large literature on intermediation. For example, middlemen may exist because they are superior in matching buyers or sellers (Rubinstein and Wolinsky 1987). Under informational frictions, middlemen may also have a stronger incentive to become experts and provide quality certification due to the large transaction volume they handle (Biglaiser 1993, Biglaiser and Friedman 1994, Lizzeri 1999). In financial markets, financial intermediaries, such as banks, arise naturally as the delegated monitors on behalf of many small depositors (Diamond 1984). Market makers match buyers and sellers in the securities market and absorb excess supply or demand (Glosten and Milgrom 1985). Intermediation may also emerge when the informed party wishes to sell information through a third party, such as mutual funds (Admati and Pfleiderer 1990), or when agents join forces to produce information, such as index providers or asset managers (Ramakrishnan and Thakor 1984, Gârleanu and Pedersen 2018). While each paper provides an interesting story for intermediation, its market structure is often simplified away in that all these models feature only one layer of intermediaries. Instead of offering another reason for intermediation, I take some shared features in many intermediated markets (namely, intermediaries generate economies of scale and receive performance-based pay) and focus on the market structure of the intermediation chain (its length, the skill, effort, and size distribution of intermediaries along the chain). Even in the single-layer context, when effort input is endogenized, efficient intermediation does not rely on intermediaries to have superior skill which is a novel insight relative to the literature.

More recent papers investigate intermediation chains. Glode and Opp (2016) and Glode, Opp, and Zhang (2019) show that intermediaries reduce the potential informational frictions between parties in each trading layer, therefore making efficient transactions more likely. When the chain is sufficiently long, the information asymmetry within each layer becomes sufficiently small, and fully efficient trades can be restored. He and Li (2022) show that a credit chain formed by intermediaries can facilitate firms with long-term projects to access cheap funding from overlapping-generation households in the form of short-term debt. Intermediation reduces the inefficiency associated with trading debt claims among households but introduces the rollover risks of intermediaries. This trade-off endogenizes the optimal length of the credit chain. From a different perspective, Dasgupta and Maug (2021) argue that excessive layers of inefficient delegation may occur when decision makers have reputation concerns and prefer to pass on the decision to a third party so that their reputation can be better preserved upon failures. In addition to having a different mechanism from this literature, my model provides a novel insight that even short chains can achieve full efficiency. In the context of the long chain, I show that efficiency is typically resolved by the final layers of the chain because motivating effort in initial layers is very difficult.[4]

While most of the aforementioned papers take the participants in intermediary sectors as fixed, Kaniel and Orlov (2020) study how agents with privately observed skill levels decide to join and leave the (single-layer) intermediary sector. Although agents gain a track record while working in the intermediary sector, high-skill agents also face a holdup problem because being fired from the intermediary sector tarnishes their reputation. My work is related in that I also study agents' endogenous participation in the intermediary sectors. However, I abstract from the contracting problem between the agent and intermediary sectors and instead focus on their self-selection into different layers of the intermediation chain.

Berk and Van Binsbergen (2022) study charlatans in high-skill professions similar to the unskilled intermediaries in my model. In their single-layer intermediation model, the existence of charlatans may improve consumer welfare as it increases competition for the skilled agents, making prices more competitive. My work is complementary in that I take the existence of charlatans as given and offer insight into how they are distributed along the intermediation chain.

One of the model's applications in the context of platform and social media influencers contributes to this emerging literature. Cong and Li (2021) provide a nice summary of the recent work.

2 Model Setup and Equilibrium Definition

2.1 Model Setup

The model focuses on the market structure of the intermediary sector: the number of intermediation layers, the size and skill distribution of each layer, and the effort choice of intermediaries. I frame the general model with the terminology "consumers,"

[4]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.

"products," and "intermediaries." In Subsection 2.3, I offer two markets where this model can be applied: the delegated asset management industry (fund of funds and their underlying funds) and the social media influencers' economy (platforms and influencers).

Consider an economy populated with a mass N_C of atomistic consumers and 1 unit of atomistic intermediaries. A fraction n_0 of consumers and a fraction n_S of intermediaries are "skilled" — a term that I describe later. Each consumer needs to consume M different products, one from each market (hence, M markets in total). In each market, a fraction n_P of products are good, generating 1 unit of utility for consumers.[5] Others are bad, generating 0 consumption utility. Products are non-exclusive in that multiple consumers can choose the same product.

Consumers find products through an $I \geq 0$ layer intermediation chain, which is a parameter for now and will be endogenized in Subsection 5.2. The special case $I = 0$ represents that consumers directly access the product markets without intermediaries. Denote by $N_{S,i}$ and N_i the endogenous mass of skilled intermediaries and all intermediaries who choose to enter layer i. The fraction of skilled intermediaries $n_i \equiv \frac{N_{S,i}}{N_i}$ naturally captures the quality of layer i. For notational convenience, denote by $n_{I+1} \equiv n_P$ the quality of the product market.

Each consumer chooses an intermediary from layer 1, who in turn chooses an intermediary from layer 2, and so on. Intermediaries in the final layer I (or consumers if $I = 0$) choose one product from each M product market.[6] Skilled consumers and intermediaries have the following screening technology. By exerting effort $e_i \in [0, \infty)$ $(0 \leq i \leq I)$ and paying the cost $c(e_i)$, an intermediary in layer i (or a consumer if $i = 0$) can find a skilled intermediary in layer $i + 1$ (or a good product if $i = I$) with probability $P(e_i, n_{i+1})$, depending on the transparency in the next layer n_{i+1}. The remaining $1-n_i$ fraction of consumers or intermediaries in each layer are unskilled and do not have this technology (or equivalently, their effort input generates no value). They find a skilled intermediary or a good product with probability $P(0, n_{i+1})$.[7]

[5] For simplicity, I assume the fraction of good products n_P is the same across all M product markets. Allowing for product-specific $n_{P,m}$ for $m = 1, 2, ..., M$ does not affect the results. In addition, the model can accommodate different consumer preferences.

[6] Note that "choosing an intermediary" should not be interpreted literally as picking only one. It is mathematically equivalent to pick a (sub)set of intermediaries from the next layer and the final transaction flowing randomly through one of the selected intermediaries.

[7] The insights and results are robust to layer-dependent screening technologies and cost functions — $P_i(e_i, n_{i+1})$ and $c_i(e_i)$ — and more complex specifications of unskilled intermediaries' technology.

Intermediaries may not have full self-awareness in that they only observe signals about their skill types. Specifically, intermediaries independently receive binary signals (G)ood or (B)ad, with the conditional probabilities

$$P(G|\text{skilled}) = P(B|\text{unskilled}) = \theta \in [\frac{1}{2}, 1]. \tag{1}$$

Bayes' rule implies

$$P(\text{skilled}|G) = \frac{\theta n_S}{\theta n_S + (1-\theta)(1-n_S)} \equiv \hat{p}_G \text{ and } P(\text{skilled}|B) = \frac{(1-\theta)n_S}{(1-\theta)n_S + \theta(1-n_S)} \equiv \hat{p}_B.$$

When $\theta = \frac{1}{2}$, the signal is pure noise and completely uninformative. When $\theta = 1$, intermediaries have full self-awareness in that they know their skill types. For instance, upon entering the asset management industry, fresh graduates may not necessarily know whether they will be good portfolio managers, and their grades in finance courses may be a noisy signal. I generalize the model to contain more than two skill types in Subsection 6.3.

The timing of the game is as follows. After observing the signal, intermediaries can decide in which layer they operate by paying the corresponding entry cost C_i ($i = 1, 2, ..., I$). Once the chain is formed, intermediaries' skill types are realized. Skilled consumers and intermediaries decide their respective effort e_i ($i = 0, 1, ..., I$). Intermediaries receive r for each good product delivered to consumers through them, which can be viewed as a performance pay.[8] Consumers therefore receive $1 - rI$ for each good product consumed or 0 otherwise. Because consumers do not need to choose their location in the chain, knowing their skill types before the chain is formed is irrelevant. Figure 1 summarizes the market structure of this intermediated economy.

[8]In Subsection 6.1, I also show the results are robust to having a fixed fee f that is independent of the quality of the products delivered. In Subsection 6.2, I discuss potential ways to model layer-dependent compensation r_i.

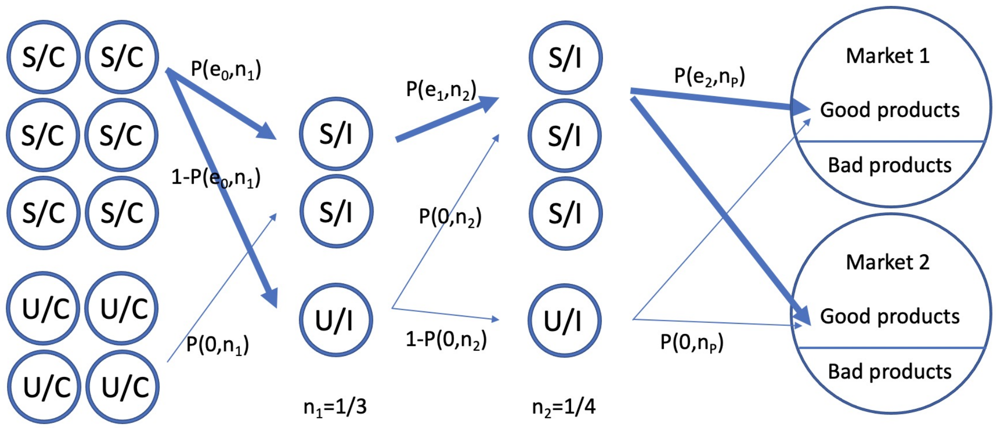

Figure 1: Structure of the economy. The abbreviations "S", "U", "C", and "I" denote skilled, unskilled, consumer, and intermediary, respectively. In this example, 6 out of the 10 consumers are skilled, and hence $n_0 = 60\%$. Similarly, 1 out of the 3 intermediaries in layer 1 is skilled, and therefore $n_1 = 1/3$. The expressions associated with each arrow indicate the probabilities of each possibility. For example, a skilled layer 1 intermediary finds a skilled layer 2 intermediary with probability $P(e_1, n_2)$. Note that many possibilities are omitted in the figure for readability.

I now introduce the payoff to consumers and intermediaries. To simplify the notation, I define the screening matrix

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

which captures the probability that a skilled/unskilled intermediary in layer i (or consumer if $i = 0$) finds a skilled/unskilled intermediary in the next layer (or a good product if $i = I$).

The payoff to skilled ($\Pi_{S,C}$) and unskilled ($\Pi_{U,C}$) consumers is

$$\begin{pmatrix} \Pi_{S,C} \\ \Pi_{U,C} \end{pmatrix} = \max_{e_0} M(1 - rI)\Pi_{i=0}^{I}\mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} [1 + \mathbf{1}_{I=0}(M-1)]\, c(e_0) \\ 0 \end{pmatrix}. \tag{3}$$

Each consumer receives M products. The matrix product $\Pi_{i=0}^{I}\mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ captures the probabilities of a skilled and an unskilled consumer finding a good product, respectively. For each good product, consumers receive 1 unit of utility and pay a total of rI to intermediaries. Skilled consumers also optimally choose the effort level e_0

and pay the effort cost $c(e_0)$. The only complication is that when consumers directly access the product market without intermediaries ($I = 0$), they need to screen each M product markets and therefore pay the effort cost $c(e_0)$ M times.

Explicitly noting some special cases is useful. For example, when $I = 0$, consumer payoff (3) simply becomes

$$\begin{pmatrix} \Pi_{S,C} \\ \Pi_{U,C} \end{pmatrix} = \max_{e_0} M \begin{pmatrix} P(e_0, n_P) \\ P(0, n_P) \end{pmatrix} - \begin{pmatrix} Mc(e_0) \\ 0 \end{pmatrix} = \begin{pmatrix} \max_{e_0} M\left[P(e_0, n_P) - c(e_0)\right] \\ MP(0, n_P) \end{pmatrix}. \tag{4}$$

When there is a single layer of intermediary $I = 1$, consumer payoff (3) becomes

$$\begin{aligned} \begin{pmatrix} \Pi_{S,C} \\ \Pi_{U,C} \end{pmatrix} &= \max\nolimits_{e_0} M(1-r) \begin{pmatrix} P(e_0, n_S) & 1 - P(e_0, n_S) \\ P(0, n_S) & 1 - P(0, n_S) \end{pmatrix} \begin{pmatrix} P(e_1, n_P) \\ P(0, n_P) \end{pmatrix} - \begin{pmatrix} c(e_0) \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \max_{e_0} M(1-r)\left\{P(e_0, n_S)P(e_1, n_P) + \left[1 - P(e_0, n_S)\right] P(0, n_P)\right\} - c(e_0) \\ M(1-r)\left\{P(0, n_S)P(e_1, n_P) + \left[1 - P(0, n_S)\right] P(0, n_P)\right\} \end{pmatrix}, \end{aligned} \tag{5}$$

which captures the two ways that a consumer can find a good product: through a skilled intermediary who can find a good product with probability $P(e_1, n_P)$ or through an unskilled intermediary with the corresponding probability $P(0, n_P)$.

Now consider the payoff to intermediaries. An unskilled intermediary in layer i receives

$$\begin{aligned} \Pi_{U,i} = \quad & \underbrace{N_C \begin{pmatrix} n_0 & 1 - n_0 \end{pmatrix} \Pi_{j=0}^{i-1} \mathbb{P}_j \begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\text{mass of consumers finding an unskilled intermediary}} \quad \frac{1}{N_i - N_{S,i}} Mr \\ & \underbrace{\begin{pmatrix} 0 & 1 \end{pmatrix} \Pi_{j=i}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of an unskilled intermediary finding a good product}} \qquad -C_i \end{aligned} \tag{6}$$

The matrix product in the first row of (6) captures the probability that a consumer goes through an unskilled intermediary in layer i. There are a total of N_C consumers, and the total consumer flow is shared by the $N_i - N_{S,i}$ unskilled intermediaries in layer i. The matrix product in the second row of (6) captures the probability that an unskilled intermediary can find a good product through the subsequent chain. Each consumer needs M products, and for each good product delivered to each consumer, the intermediary receives r. Finally, the intermediary pays an entry cost C_i.

Similarly, a skilled intermediary in layer $i = 1, 2, ..., I$ receives

$$\Pi_{S,i} = \max_{e_i} \underbrace{N_C \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{i-1} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{mass of consumers finding a skilled intermediary}} \frac{1}{N_{S,i}} Mr \underbrace{\begin{pmatrix} 1 & 0 \end{pmatrix} \Pi_{j=i}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{probability of a skilled intermediary finding a good product}} - c(e_i) \left[1 + \mathbf{1}_{i=I} (M-1)\right] - C_i. \tag{7}$$

As in (6), the two matrix products in (7) capture the probability that a consumer goes through a skilled intermediary in layer i and the probability that this skilled intermediary ultimately finds a good product. Unlike (6), the skilled intermediary can affect the second probability by choosing effort level e_i, which shows up in $\mathbb{P}_i$. A final-layer intermediary ($i = I$) needs to screen M different product markets and therefore needs to exert effort cost $c(e_I)$ repeatedly M times (as a consumer would if $I = 0$). Those in earlier layers screen the intermediaries in the next layer and therefore incur the effort cost $c(e_i)$ only once. They also need to pay the entry cost C_i. Note that skilled intermediaries can always choose zero effort $e_i = 0$ to receive the payoff to an unskilled intermediary, hence $\Pi_{S,i} \geq \Pi_{U,i}$.

Explicitly noting intermediaries' payoff functions for $I = 1$ as the simplest example is again useful. In this case, the total mass of unskilled intermediaries is $1 - n_S$. They receive $[1 - P(e_0, n_S)]$ fraction of skilled consumers and $[1 - P(0, n_S)]$ fraction of unskilled consumers. In addition, unskilled intermediaries find a good product with probability $P(0, n_P)$. Hence, their payoff (6) can be equivalently written as

$$\Pi_{U,1} = M \frac{[1 - P(e_0, n_S)]\, n_0 + [1 - P(0, n_S)]\, (1 - n_0)}{1 - n_S} r N_C P(0, n_P) - C_1. \tag{8}$$

Similarly, the mass of n_S skilled intermediaries receive $P(e_0, n_S)$ fraction of skilled consumers and $P(0, n_S)$ fraction of unskilled consumers. Hence, their payoff (7) becomes

$$\Pi_{S,1} = \max_{e_1} M \left[\frac{P(e_0, n_S) n_0 + P(0, n_S)\, (1 - n_0)}{n_S} r N_C P(e_1, n_P) - c(e_1) \right] - C_1. \tag{9}$$

2.2 Equilibrium Definition and Two Benchmarks

The exogenous parameters of the model include the number of consumers N_C, the fraction of skilled consumers, intermediaries, and good products n_0, n_S, and n_P, respectively, the number of layers in intermediation chain I, and the entry cost for each layer C_i. An equilibrium consists of a set of effort levels $\{e_i^*|i = 0, 1, ..., I\}$, the distribution of skilled and unskilled (or equivalently, the total number of) intermediaries in each layer $\left\{N_{S,i}^*, N_i^*|i = 1, 2, ..., I\right\}$, and the price that each intermediary charges for a good product r^*. The $3I + 2$ 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, determined by skilled consumers or intermediaries (3) and (7).

2. [I conditions] Intermediaries with bad signals break even whenever they operate in a layer $i = 1, 2, ..., I$:

$$(1 - \hat{p}_B)\,\Pi_{U,i} + \hat{p}_B\Pi_{S,i} = 0. \tag{10}$$

Alternatively, if a layer does not include intermediaries with bad signals, then $N_{S,i}^* = \hat{p}_G N_i^*$ and (10) holds with inequality $\leq$.

3. [$I - 1$ independent conditions] Intermediaries with good signals are indifferent between different layers that they operate in

$$(1 - \hat{p}_G)\,\Pi_{U,i} + \hat{p}_G\Pi_{S,i} = \Pi_G^*, \tag{11}$$

for some constant $\Pi_G^* \geq 0$ independent of i. Alternatively, if a layer does not include intermediaries with good signals, then $N_{S,i}^* = \hat{p}_B N_i^*$ and (11) holds with inequality $\leq$. Conditions (10) and (11) together imply that

$$\Pi_{S,i} = (1 - \hat{p}_B)\frac{\Pi_G^*}{\hat{p}_G - \hat{p}_B}, \quad \Pi_{U,i} = -\hat{p}_B\frac{\Pi_G^*}{\hat{p}_G - \hat{p}_B} \tag{12}$$

are constants for all i. It is worth noting that instead of having the $2I - 1$ independent equilibrium conditions in (10) and (11), one might alternatively assume that the unskilled intermediaries break even ex-post, and intermediaries with good or bad signals ex-ante are indifferent between the layers in which they operate.[9] This alter-

[9] Specifically, under the alternative specification, condition (10) becomes $(1 - \hat{p}_B)\,\Pi_{U,i} + \hat{p}_B\Pi_{S,i} = \Pi_B^*$, giving $I - 1$ independent conditions. Condition (11) remains the same. As a result, (12) becomes $\Pi_{S,i} = \frac{\Pi_G^* - \Pi_B^*}{\hat{p}_G - \hat{p}_B}$ and $\Pi_{U,i} = 0$.

native specification is a different way to endogenize the equilibrium fee r^*, but does not qualitatively affect the market structure of the intermediary sectors.[10]

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

$$\sum_{i=1}^{I} N_{S,i}^* = n_S \text{ and } \sum_{i=1}^{I} N_i^* = 1.$$

Throughout the paper, I also impose three standard assumptions on the screening technology and the cost function $P(e,n)$ and $c(e)$.

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

Without effort input, the probability of finding a skilled intermediary or good product is simply the transparency of the respective markets. As long as good intermediaries or product exists, sufficiently large effort input results in finding them almost certainly. Finally, if an intermediary layer or the product market contains only one type (all good $n = 1$ or all bad $n = 0$), then effort input does not affect the outcome.

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 product, but its marginal effect diminishes. Higher transparency n naturally also helps the screening outcome. A simple example is

$$P(e,n) = \frac{(1+e)\,n}{(1+e)\,n + 1 - n} = \frac{(1+e)\,n}{1+en}. \tag{13}$$

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

$$c(e) = \frac{1}{2}e^2. \tag{14}$$

[10] Throughout the paper, I focus on the limiting cases (for example, $N_C \to \infty$). In equilibrium, the intermediary fee $r^* \to 0$. Different specifications are essentially different ways to microfound $r^* \to 0$, and therefore do not affect the qualitative insights.

Welfare in the economy is defined as

$$W = \underbrace{MN_C \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{i=0}^{I} \mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{total consumer utility}} - \underbrace{c(e_0)n_0N_C - \sum_{i=1}^{I-1} c(e_i)N_{S,i} - c(e_I)N_{S,I}M}_{\text{total costs of screening efforts}}, \tag{15}$$

which is the total utility generated by consumers net of all screening costs incurred by skilled consumers and intermediaries. Note that the payment from consumers to intermediaries r is purely transferral and therefore does not affect welfare.

I analyze two straightforward benchmarks. The first-best outcome of this economy is simply

$$W_{FB} = N_C M.$$

That is, all consumers receive good products in every M market with no one incurring any screening costs.

The second benchmark is that consumers directly access the product market without intermediaries ($I = 0$). The optimal e_0^* solves (4) and satisfies the first-order condition

$$\frac{\partial P(e_0^*, n_P)}{\partial e_0} = c'(e_0^*). \tag{16}$$

In this case, the probability $P(e_0^*, n_P)$ is clearly bounded away from 1, and the cost $c(e_0^*)$ is positive. Hence, total welfare is bounded away from the first-best outcome $N_C M$. Intermediation will simultaneously improve the probability of finding good products and save the total screening costs in the economy.

The formal analysis of the model is organized as follows. Section 3 focuses on the case of a single-layer intermediation and highlights that superior skill or quality of the intermediary sector is not necessary for efficient intermediation to occur. Section 4 studies the structure with two layers of intermediaries — perhaps also the most empirically relevant case. The key message is that when intermediaries have full self-awareness, skilled and unskilled intermediaries can self-select into different layers, thereby creating a two-layer chain with a heterogeneous skill distribution. This type of structures can approximate the first-best outcome. When intermediaries do not have full self-awareness, Section 5 shows that a long intermediation chain can restore the first-best outcome. I also endogenize the optimal chain length I^* by introducing a small dissipative cost per layer.

2.3 Two Applications

I offer two applications of the model. The first one is the asset management industry, which is an intermediary sector between investors and the capital market. Investors need to invest in the capital market, with some objectives such as finding portfolios with the desired risk-return trade-off. Portfolios that satisfy these objectives can be labeled as the good products in the model. Instead of researching different assets and carrying out self-directed investments, investors can delegate the task to fund of funds or pension fund managers (layer 1 intermediaries). These managers in turn choose the underlying funds — for example, private equity funds (PE) or hedge funds (layer 2 intermediaries) — who ultimately choose portfolios from various asset markets. Investors can exert effort in selecting fund-of-funds managers, who can in turn exert effort in selecting funds, which in turn exert effort in selecting assets. Fund managers often receive performance-based pay. For instance, PE and hedge funds often receive "carried interest," which is typically 20% of the profits above some benchmark hurdle rate. Even when the compensation is a fixed percentage fees based on total assets under management (AUM), it is well documented that fund flows and consequently the AUM are sensitive to past performance (Berk and Green 2004). In this sense, one can broadly view fund managers as receiving performance based compensation. In this context, the model offers predictions on the size of each sector (number of fund of funds, PE or hedge funds), the skill distribution of their managers, and their effort choices. These insights will be discussed in Subsection 4.2.

The second application is social media platforms and influencers, which can be interpreted as intermediaries between consumers and products. Instead of spending time and effort to try various products in person, consumers may choose to visit social media platforms such as YouTube, Instagram, and Tiktok (layer 1 intermediaries). These platforms in turn choose from millions of channels and influencers (layer 2 intermediaries) and narrow down the choice set to several hundred ones, tailored for individual consumers. The success of a platform relies on its sophisticated algorithm to analyze individual consumer's preferences and recommend suitable channels/influencers. As noted in footnote 6, narrowing down the choice set by platforms is a form of screening in the model and does not exclude consumer's ultimate selection of influencers or channels. Influencers choose and recommend products to their followers, and in return, receive compensation based on the viewer counts or net purchase they generate. This compensation is arguably performance-based because those who can better attract consumer attention, satisfy their needs, and generate purchases typically are more successful and receive larger compensation. The model offers predictions on the number of platforms and influencers as well as their quality and effort choices. Subsection 4.1 contains the discussion that is more relevant to this interpretation.

3 Single Layer of Intermediation $I = 1$

I begin the formal analysis with a single layer of intermediation ($I = 1$). In this case, the sector size is exogenously given by the fixed supply of intermediaries $N_{S,1}^* = n_S$ and $N_1^* = 1$. The remaining equilibrium variables, price r^* and the effort choices of skilled consumers and intermediaries e_i^* ($i = 0, 1$), are jointly determined by the break-even condition of the bad intermediaries (10) and effort optimization problems (3) and (7).

A comparison of consumer payoffs between the two simplest cases — direct market access (4) and one layer of intermediaries (5) — reveals three frictions associated with intermediation. First, consumers need to exert costly effort to find a skilled intermediary, which is only successful with some probability. Second, the intermediary still needs to exert another costly effort to find a good product. Finally, consumers need to compensate and incentivize intermediaries for their effort. In contrast, when consumers directly screen in the product market, they only incur the screening effort once. As such, it is not obvious whether efficient intermediation can emerge.

The main takeaway from this section is that not only can a single layer of intermediation ($I = 1$) improve consumer welfare, but the payoff to skilled consumers can also approach the first best when there are large numbers of consumers and markets $N_C, M \to \infty$.[11] Furthermore, the result holds for an intermediary sector with arbitrary low quality $n_S > 0$. In particular, the intermediary sector on average can have worse skills than consumers ($n_S < n_0$) and lower transparency than the product markets ($n_S < n_P$), but efficient intermediation can still occur.

Proposition 1 *If the entry cost C_1 is set such that $C_1 = aM\left(N_C\right)^b$ for any constants $a > 0$, $b \in (0, 1)$, then for any $n_S > 0$, when $M, N_C \to \infty$, skilled consumers' payoff per product approaches the first best ($\frac{\Pi_{S,C}}{M} \to 1$).*

[11]The large numbers are relative to the number of intermediaries which is normalized to 1.

The unskilled consumers' payoff per product approaches

$$\frac{\Pi_{U,C}}{M} \to P(0, n_S) + [1 - P(0, n_S)]\, P(0, n_P), \tag{17}$$

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 consumer delegations N_C and screening many product markets M. First, the large consumer delegation motivates intermediaries' effort. Specifically, there are MN_C transactions in total, and intermediaries (with bad signals) need to break even on the entry cost C_1. Hence, the equilibrium fee r^* must be vanishingly small ($r^* \sim \frac{C_1}{MN_C} = (N_C)^{b-1}$), due to large consumer delegation ($N_C \to \infty$). However, the amount of consumers that skilled intermediaries receive is in the order of N_C.[12] Therefore, despite the small fee r^*, skilled intermediaries have large effort incentive ($r^* N_C \to \infty$) and almost certainly find good products in each market ($P(e_1^*, n_P) \to 1$).[13] Second, the many products M a skilled intermediary can screen on behalf of consumers motivates skilled consumers to exert high effort ($e_0^* \to \infty$) and find a skilled intermediary ($P(e_0^*, n_S) \to 1$). Therefore, skilled consumers can almost certainly receive a good product from each market by paying a negligible fee r^*. The unskilled consumers lack the ability to find skilled intermediaries, and their payoff per product is therefore bounded away from 1, but they still fare better than with direct market access. Intuitively, if they find an unskilled intermediary, the outcome would be the same as their direct search in the product market. However, if they find a skilled intermediary by chance, they will be able to receive good products with higher probability.

The two-sided economies of scale may remind readers of Diamond (1984)'s theory on 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) also receive deposits from many depositors and monitor many underlying firms in order to generate efficiency. However, unlike Diamond (1984), intermediaries' entry decision and effort choice in my paper provide new insights which are

[12]A skilled intermediary at least receives $\frac{N_C}{1} = N_C$ consumers if consumers do not screen and at most receives $\frac{N_C}{n_S}$ consumers if all consumers find skilled intermediaries. Hence, consumer delegation is in the order of N_C.

[13]The average effort cost $\frac{c(e_1^*)}{N_C} \to 0$, as discussed before.

discussed below. Perhaps more importantly, single-layer intermediation in my paper serves as a benchmark case in the sense that the market structure of the intermediary sector is exogenous. In Sections 4 and 5, I study longer intermediation chains show that having many consumers (N_C) and products (M) is no longer necessary for efficiency, among other novel insights.

It is worth pointing out that efficient intermediation does not rely on the quality of the intermediary sector n_S, because the high effort incentive of the skilled intermediaries is sufficient to compensate the possibility that they may have worse skill than consumers or worse quality than the product market.

Another interesting observation is that when the entry cost C_1 is either too high or too low, consumer payoff decreases. On the low end — for example when $C_1 \to 0$ — the equilibrium price $r^* \to 0$ because of intermediaries' breakeven condition. This in turn implies that skilled intermediaries have little effort incentive. Therefore, the consumer payoff approaches $P(0, n_P) < 1$, bounded away from first best. On the high end, when C_1 is excessively high, the break even fee r^* also becomes large, reducing consumer payoff. In fact, consumer payoff may even become negative. In both cases, consumers fare better with direct market access instead of relying on intermediaries. Hence, we have the following corollary.

Corollary 1 *Consumer payoffs per product* $\frac{\Pi_{S,C}}{M}$ *and* $\frac{\Pi_{U,C}}{M}$ *are non-monotonic in the intermediary entry cost* C_1*. In addition, when* C_1 *is either sufficiently large or close to* 0*, the consumer payoff is weakly higher with direct screening in the product market without intermediaries.*

Figure 2 summarizes the results in this section with a numerical example.

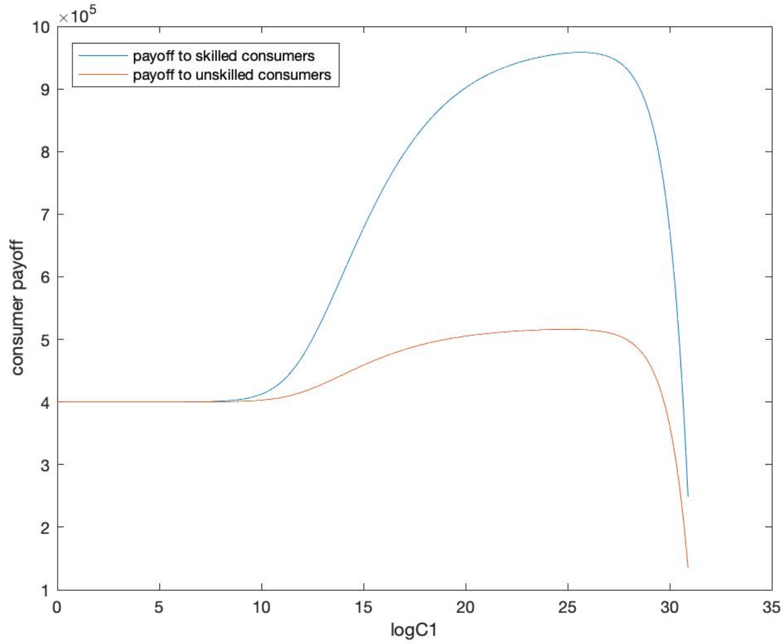

Figure 2: Consumer payoff as a function of the (log) entry cost C_1 under $I = 1$ layer of intermediation. The parameters are $N_C = 10^8$, $M = 10^6$, $n_0 = 0.6$, $n_P = 0.4$, $n_S = 0.2$, and $\theta = 0.8$. The entry cost C_1 ranges from 1 to 2.63×10^{13}. The functional forms of $P(e, n)$ and $c(e)$ are given by (13) and (14). The hump-shaped payoff curves illustrate Corollary 1. Mapping to Proposition 1, the skilled consumers can reach 96% of the first-best payoff $M = 10^5$. The unskilled consumers can receive up to 99% of the theoretical upper bound $(0.2+0.8*0.4 = 0.52)M$ in (17).

4 Two Layers and Intermediaries' Self-Selection

In this section, I consider two layers of intermediation ($I = 2$), which opens up the possibility that intermediaries with different signals (skill types) can self-select into different layers. This self-selection in turn generates endogenous sector sizes (N_i), skill distribution (n_i), and effort choice e_i.

To highlight the self-selection effect, I consider the extreme case where intermediaries perfectly know their skill types ($\theta = 1$) in Subsections 4.1 and 4.2. The key takeaway is that even with a single product market $M = 1$ (or, more generally, for any given M), all consumers can receive the first-best payoffs asymptotically. This outcome features a significant welfare improvement over the single-layer structure (Proposition 1) in the sense that unskilled consumers can also achieve the first-best payoff, and the result does not rely on having large M and N_C, which is one of the key differences from Diamond (1984). The asymptotic first-best outcome can be achieved in two ways, featuring heterogeneous skill distributions across the two layers: Either the first layer (Subsection 4.1) or the second layer (Subsection 4.2) is fully skilled and the other layer contains mixed skill types. The former structure resembles the social media platform and influencer chain, whereas the latter resembles the asset management industry.

In Subsection 4.3, I show that imperfect self-awareness of the intermediaries reduces the effectiveness of their self-selection. The consumer payoff decreases and in general is bounded away from the first best. This analysis also motivates Section 5, where I show that a long intermediation chain can restore full efficiency even with unknown skill types ($\theta = \frac{1}{2}$).

4.1 High Quality First Layer

As discussed, I assume perfect self-awareness $\theta = 1$ and consider a two-layer structure featuring a high-quality first layer ($n_1 = 1$) that can implement the first best asymptotically for any given M. This structure can be implemented by having a significantly higher entry cost to the first layer $C_1 \gg C_2$.

Proposition 2 *When $\theta = 1$, $I = 2$, and N_C is sufficiently large, there exists an equilibrium with the following features:*

1. All unskilled intermediaries stay in the second layer $N^*_{U,2} = 1 - n_S$, and only a small amount of skilled ones join the first layer $N^*_{S,2} \gg N^*_{S,1} \to 0$.

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

3. All consumer payoffs approach the first best: $\Pi_{S,C} = \Pi_{U,C} \to M$.

4. Finally, the entry cost to the first layer dominates the second layer $C_1 \gg C_2$.

In this equilibrium, because of the high entry cost in layer 1, unskilled intermediaries self-select into layer 2. As a result, the first layer only contains a relatively small number of skilled intermediaries, each receiving a large amount of consumer delegation. In other words, the first layer is small and high quality. The second layer is mixed with all unskilled intermediaries and the remaining skilled ones. The high quality of the first layer means that consumers do not need to exert any effort

$e_0^* = 0$. The large amount of consumer delegation incentivizes intermediaries in layer 1 to exert high effort and almost certainly find a skilled intermediary in layer 2. For the same reason, skilled intermediaries in layer 2 also exert high effort in finding good products. Hence, consumers can almost certainly receive good products from all markets ($P(e_1^*, n_2^*)P(e_2^*, n_P) \to 1$).

Why consumers can receive the first-best payoff M is not obvious because the equilibrium price both needs to be vanishingly small ($r^* \to 0$) and needs to provide sufficient effort incentives in both layers ($e_i^* \to \infty$). The main technical complication is the strategic complementarity of effort across layers. For example, if skilled intermediaries in layer 2 exert higher effort, then finding them becomes more valuable, which in turn motivates the skilled ones in layer 1 to exert higher effort. Conversely, higher effort by skilled intermediaries in layer 1 results the skilled ones in layer 2 receiving more consumer delegation, which in turn motivates their higher effort. However, if intermediaries in layer 1 screen too harshly such that the unskilled intermediaries in layer 2 receive few consumers — for example $\frac{1-P(e_1^*, n_2^*)}{1-n_S} N_C < \infty$ even as $N_C \to \infty$ — then the equilibrium price r^* needs to be bounded away from 0 for the unskilled intermediaries to break even on a given level of C_2. This self-fulfilling nature threatens the existence of an equilibrium with the desired property $r^* \to 0$. The proof in the appendix details the construction method.[14]

Proposition 2 explains the market structure of social media platforms (first layer) and influencers (second layer). Clearly, the entry cost to developing a platform is much higher than the cost of becoming an influencer on these platforms ($C_1 \gg C_2$). The first layer is therefore "small" ($N_{S,1}^* \to 0$) in the sense there are only a handful of social media platforms that consumers frequently visit. These platforms are also very skilled ($n_1^* = 1$) in analyzing consumers' preferences and delivering the relevant channels and influencers to them: Consumers in practice receive customized recommendations and search outcomes. Platforms invest heavily in developing algorithms ($P(e_1^*, n_2^*) \to 1$), which is arguably the key to a platform's success in practice. Because of the high quality of the platform sector, consumers do not need to spend time choosing which platforms are good ($e_0^* = 0$). In comparison, the influencer sector is much bigger

[14]To establish the existence, I first construct a sequence of equilibrium variables $N_{S,1}^*$ and e_2^* as $N_C \to \infty$, and solve for the corresponding r^* and e_1^* from skilled intermediaries' incentive compatibility conditions. Next, I show that with appropriately designed $N_{S,1}^*$ and e_2^*, the desired equilibrium properties $r^* \to 0$ and $P(e_1^*, n_2^*) \to 1$ are satisfied. Finally, I show that the entry costs C_1 and C_2 can be calculated from the equilibrium variables, and $C_1 \gg C_2$ indeed follows.

($N_2^* \approx 1$) and has mixed quality ($n_2^* \approx n_S < 1$). However, the "best" channels and influencers (skilled intermediaries) receive lots of viewers and followers and are frequently recommended by platforms to consumers. Because of the large amount of consumer delegation, they are incentivized to produce good contents and promote good products ($P(e_2^*, n_P) \to 1$). Finally, the compensation that individual consumers pay to the intermediaries is vanishingly small ($r^* \to 0$). For example, platforms often show consumers several seconds of advertisements for each visit. Even though the cost is negligible to individual consumers, the collective advertisement income is often sizable for platforms ($r^* N_C \to \infty$).

4.2 High Quality Second Layer

I now move on to the second possibility that layer 2 is high quality that only consists of skilled intermediaries. This outcome can be implemented by a high entry cost to the second layer: $C_2 \gg C_1$.

Proposition 3 *When $\theta = 1$ and $I = 2$, if C_1 is sufficiently small and C_2 is sufficiently large, then there is an equilibrium in which $N_{U,2}^* = 0$, $N_{S,2}^* \leq N_S$, and $e_0^* = e_1^* = 0$. In this equilibrium, for any $M \geq 1$ and N_C, both skilled and unskilled consumers receive payoff $\Pi_{S,C} = \Pi_{U,C} \to M$.*

In this equilibrium, because of the high entry cost C_2 in layer 2, unskilled intermediaries stay away, and only a small amount of skilled intermediaries operate in this layer $n_2^* = 1$. Unskilled intermediaries, together with the remaining skilled ones, enter layer 1. The fact that $n_2^* = 1$ also implies that layer 1 intermediaries as well as consumers, regardless of their skills, do not need to make any effort ($e_0^* = e_1^* = 0$). Finally, the skilled intermediaries in layer 2 exert high effort in finding good products and $P(e_2^*, n_P) \to 1$ because of the large amount of consumer delegation they receive $\frac{N_C}{N_{S,2}^*}$. Intuitively, unskilled intermediaries do not wish to enter layer 2 because they do not have the ability to select products. Instead, they are better off operating in layer 1 and free riding on layer 2 intermediaries' skill.

This outcome is interesting because it does not rely on a large pool of consumers ($N_C \to \infty$) or products ($M \to \infty$), which is another significant improvement over the previous structure in Proposition 2. This improvement comes from the fact that as the entry cost in layer 2 increases $C_2 \to \infty$, its sector size becomes arbitrarily small

$N_{S,2}^* \to 0$. Consequently, even for a fixed finite number of consumers N_C, each layer 2 intermediary can receive a large delegation $\frac{N_C}{N_{S,2}^*} \to \infty$, creating economies of scale.

This structure maps particularly well to the asset management industry, where 42% of index-tracking mutual funds track just one index: the S&P 500 (Jiang, Vayanos, and Zheng 2020). It is arguably more costly to produce and maintain an index (second layer) than to register as a passive mutual fund manager (first layer). Consequently, only a handful of indices, such as the S&P 500, Russell 1000/2000, FTSE 100, and so on are widely followed. Each one receives a lot of delegations by passive index mutual funds, free riding on the index providers' screening effort and portfolio construction. Terminal investors do not need to spend a lot of effort in selecting which S&P 500 mutual fund to invest in, once they decide to invest in the S&P 500, as all such funds provide similar services.

Another example of this structure is the financial advisory industry. There are close to 200,000 CFA charter holders, many working as financial advisors (first layer). They often recommend similar financial products, such as funds, bonds, equity, and insurance plans (second layer) to consumers. If creating one's own financial product is more difficult than becoming a financial adviser, then Proposition 3 predicts that a large amount of financial advisors provide similar services to consumers.

The structure characterized in Proposition 3 continue to hold even when the two entry costs are similar $C_1 \approx C_2$: Only skilled intermediaries in layer 2, although its sector size may no longer be vanishingly small. When the number of consumers is large enough $N_C \to \infty$, which is arguably the more relevant case in practice, the first-best outcome can still be attained. The numerical example in Figure 3 offers such an example.

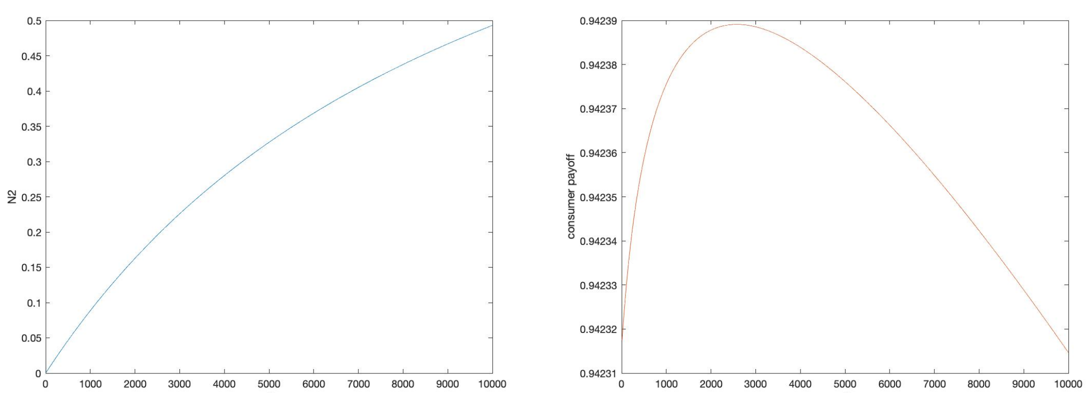

Figure 3: Size of the second layer N_2 (left panel) and consumer payoff (right

panel) as functions of the first-layer entry cost $C_1 \in [10, 10^4]$. Other parameters in this numerical example are $N_C = 10^9$, $M = 1$, $n_0 = 0.6$, $n_P = 0.4$, $n_S = 0.5$, $\theta = 0.999$, and $C_2 = 10^4$. The consumer payoff can reach 94% of the first-best outcome. As the entry cost C_1 increases to the level of C_2, the size of the second layer increases to include all skilled intermediaries: $N_2 \to n_S$.

The case with similar entry costs $C_1 \approx C_2$ maps the structure to another example of the asset management industry: fund of funds or pension funds (first layer) and underlying PE and hedge funds (second layer). The entry costs to becoming a fund manager or a fund of funds manager are likely to be similar ($C_1 \approx C_2$). In this case, both intermediary sectors can have significant representation in the economy, with the more skilled intermediaries — the PE and hedge funds — staying in the second layer. This observation is consistent with the anecdotal evidence that practitioners typically view PE and hedge funds as top career choices and these sectors often attract the top talent in the financial sector.

As a concluding remark of this section, it is useful to briefly discuss the potential deviation of consumers to skip layer 1 and directly purchase with layer 2 intermediaries. Doing so saves a layer of fees r^* paid to the first-layer intermediaries. In the case of Proposition 2 where the first layer is high quality, this deviation is clearly not optimal as consumers rely on the first layer intermediaries' effort to screen the low quality second layer. Intuitively, consumers do need the platforms' expertise to recommend influencers, rather than viewing millions of channels themselves. This structure is therefore robust to this deviation threat. However, in the case of Proposition 3, all layer 2 intermediaries are skilled, and investors therefore do have the incentive to deviate. Therefore, this equilibrium requires commitments to the structure of the chain. In practice, this commitment is achieved by restricted access. In the index-tracking mutual fund and index provider example, terminal investors typically do not have the capacity to directly invest in hundreds of stocks to replicate the index. In the financial advisor or fund-of-funds example, layer 1 intermediaries (financial advisors and fund-of-funds managers) often provide exclusive access to the underlying funds that investors do not have or are not aware of. However, because of the deviation incentives, this structure is less robust. For example, the rise of index ETFs, allowing investors to cheaply access the indices, significantly disrupts the business model of the traditional index-tracking mutual funds.

4.3 Imperfect Self-Awareness and Welfare Loss

In what follows, I revert back to the more general case that intermediaries do not know their skill types perfectly but only observe a noisy signal $\theta < 1$. An immediate consequence of this imperfection is that no intermediary sector can consists only of skilled intermediaries. In fact, the fraction of skilled intermediaries in any layer is bounded by

$$\hat{p}_B \leq n_i^* \leq \hat{p}_G,$$

that is, the probability of being skilled conditional on good or bad signals. Hence, given any fixed chain length I, there is a positive probability (bounded below by $(1-\hat{p}_G)^I(1-n_P)$) that unskilled consumers will find unskilled intermediaries in all layers and eventually receive a bad product. In addition, since all layers have mixed skill types, both skilled consumers and intermediaries have effort incentives and therefore incur effort costs in equilibrium. These two forces prevent the consumer payoff from reaching the first best. These observations are summarized as follows.

Proposition 4 *When intermediaries receive noisy signals about their skill types $\theta <$ 1, for any given $M < \infty$ and $I < \infty$, consumer payoffs are bounded away from the first best for any intermediary structure $\{N_{S,i}, N_i | i = 1, 2, ..., I\}$.*

5 Long Chain: Efficiency without Self-Awareness

5.1 A Convergence Property and Full Efficiency

As discussed in Subsection 4.3, the imperfect self-awareness of intermediaries negatively affects the consumer payoff because skilled and unskilled intermediaries start to mix in all layers. In this section, I show that a sufficiently long chain $I \to \infty$ can overcome this problem and restore the first-best outcome asymptotically for all consumers.

To highlight the economics, I analyze the extreme case where intermediaries do not know their skill type ($\theta = \frac{1}{2}$). Consequently, there is no feasible self-selection by skill types, and all intermediary layers share the same quality $n_i^* \equiv \frac{N_{S,i}^*}{N_i^*} = n_S$. The remaining $2I+2$ equilibrium variables — including effort choice $\{e_i^* | 0 \leq i \leq I\}$, sector size $\{N_i^* | 1 \leq i \leq I\}$, and intermediary fee r^* — satisfy the following equilibrium conditions, simplified from those in Subsection 2.2.

1. [$I+1$ conditions] Skilled consumers' and intermediaries' effort optimization problems (7) and (3), with the simplification that $n_i^* = n_S$ for all $1 \le i \le I$.

2. [I conditions] Because the signals are completely uninformative, all intermediaries are ex-ante homogeneous ($\hat{p}_G = \hat{p}_B = n_S$) and therefore break even. Conditions (10) and (11) reduce to

$$n_S \Pi_{S,i} + (1 - n_S)\, \Pi_{U,i} = 0, \tag{18}$$

where $\Pi_{S,i}$ and $\Pi_{U,i}$ are given by (7) and (6), respectively.

3. [1 condition] Market clearing $\sum_i N_i^* = 1$.

As a key step in the equilibrium construction and to help with intuition, it is useful to characterize the probability of finding a good product (or a skilled intermediary in layer i) when all intermediary layers are homogeneous; that is, all skilled intermediaries exert the same effort level e. To simplify notation, I also assume $n_P = n_S$ in the product market and denote $p_1 = P(e, n_S)$ and $p_0 = P(0, n_S)$. In this case, all screening matrices are identical

$$\mathbb{P} \equiv \begin{pmatrix} p_1 & 1-p_1 \\ p_0 & 1-p_0 \end{pmatrix}.$$

The (left) invariant distribution under $\mathbb{P}$ is given by $\begin{pmatrix} p^\dagger & 1-p^\dagger \end{pmatrix}$, where

$$p^\dagger = \frac{p_0}{1 - p_1 + p_0}. \tag{19}$$

It is also easy to check that $p_0 < p^\dagger < p_1$.[15]

Proposition 5 *Suppose a skilled intermediary is identified in layer j with probability $p^{(j)}$, then after i more layers of homogeneous intermediation with screening matrix $\mathbb{P}$, the probability of finding a skilled intermediary in layer $j+i$ (or a good product in the final layer) is given by*

$$p^{(j+i)} = p^\dagger + \left(p^{(j)} - p^\dagger\right)(p_1 - p_0)^i\,. \tag{20}$$

Furthermore, $p^{(j+i)} \to p^\dagger$ as $i \to \infty$ for any initial value $p^{(j)}$.

As the chain becomes sufficiently long, the success probability $p^{(j+i)}$ converges to

[15]It is easy to verify that $\begin{pmatrix} p^\dagger & 1-p^\dagger \end{pmatrix}\mathbb{P} = \begin{pmatrix} p^\dagger & 1-p^\dagger \end{pmatrix}$.

the invariant probability $p^\dagger$. Expression (20) reveals that every layer of intermediation reduces the distance between the success probability $p^{(j+i)}$ and the invariant distribution $p^\dagger$ by a factor of $p_1 - p_0$. Therefore, higher effort input, which increases p_1, has two effects. First, it increases the limiting probability $p^\dagger$. Second, it accelerates the convergence.

A theoretical point emerging from (20) is that effort incentive decays exponentially in earlier layers of the chain. Consequently, it is much more difficult to incentivize upstream intermediaries than downstream ones. Specifically, an intermediary in layer $j-1$ can choose $p^{(j)}$ through effort input. The sensitivity of the eventual success probability with respect to the layer-$j-1$ intermediary's screening outcome $\frac{\partial p^{(I)}}{\partial p^{(j)}} = (p_1 - p_0)^{I-j}$ shrinks exponentially as j becomes smaller. Intuitively, upstream intermediaries can rely on subsequent layers to bring the success probability to the invariant probability $p^\dagger$, which in turn erodes the marginal value of effort by upstream intermediaries.

As a simple application of Proposition 5, I can easily calculate an upper bound on welfare associated with any I-layer intermediation chain. The success probability of skilled intermediaries $P(e, n_i)$ is less than 1. Hence, the probability that a consumer finds a good product is bounded above by a chain associated with the following screening matrix:

$$\hat{\mathbb{P}} \equiv \begin{pmatrix} 1 & 0 \\ p_0 & 1 - p_0 \end{pmatrix}.$$

Ignoring all effort costs, the total welfare is bounded above by

$$MN_C \begin{pmatrix} n_0 & 1 - n_0 \end{pmatrix} \hat{\mathbb{P}}^{I+1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = MN_C \left[1 - (1 - n_0)(1 - p_0)^{I+1}\right]. \qquad (21)$$

For any given I, this upper bound is bounded away from the first-best level MN_C as long as $n_0, p_0 < 1$. When I is small, the distance from the first best can be wide. This upper bound also substantiates Proposition 4.

I now construct an equilibrium market structure such that a long intermediation chain can restore the first-best outcome.

Proposition 6 *Suppose $I \geq 3$, for any given N_C and M, there exists a sequence of entry costs C_i $(1 \leq i \leq I)$ such that the associated equilibrium has the following features:*

1. *The effort choices of skilled consumers and layer 1 intermediaries* e_0^* *and* e_1^* *are arbitrarily close to* 0.
2. *Skilled intermediaries in subsequent layers* $i \geq 2$ *all choose* $e_i^* = e^*$ *for some arbitrarily large* e^*.
3. *The size of the first layer* N_1^* *is arbitrarily close to* 1*, and all* N_i^* *for* $i \geq 2$ *are arbitrarily close to 0.*
4. *The equilibrium fee* r^* *is arbitrarily close to* 0.
5. *As* $I \to \infty$*, consumers' payoff as well as total welfare approach the first-best level* MN_C.

I first summarize the key steps of the proof. Rather than specifying the I entry costs $\{C_i\}$, I start by conjecturing I equilibrium effort levels of intermediaries e_i^*: the effort in the first layer e_1^* is arbitrarily small (feature 1), and the effort in all subsequent layers $e_i^* = e^*$, $i \geq 2$, is sufficiently large (feature 2). I then show that the conjectured effort sequence can be implemented by a set of entry costs C_i and establish the remaining features.

In this equilibrium, consumers have negligible effort incentives (feature 1) because the skilled intermediaries in layer 1 make a small effort, making them only marginally better than the unskilled ones. To rationalize the small effort choice of layer 1 intermediaries, the equilibrium r^* needs to be vanishingly small too (feature 4). To rationalize the high effort choices of the subsequent layers (feature 2), their sector size must be small (feature 3), generating a considerable amount of consumer delegation. Finally, the probability of a consumer finding a good product converges to the invariant distribution $p^\dagger$, which is arbitrarily close to 1 ($p_1 = P(e^*, n_S) \to 1$ in condition (19)). Hence, the consumer payoff converges to the first best (feature 5). Since all intermediaries break even, the consumer payoff is also a welfare measure.

Intuitively, a long chain allows many small intermediary sectors to exist in the downstream of the chain. Their high effort incentive essentially creates an absorbing state: Once skilled intermediaries are found in one layer, they keep finding skilled ones in subsequently layers until they find a good product. A sufficiently long chain provides consumers with many opportunities to reach the absorbing state without their own effort input, thereby approximating first best.

Recent literature on intermediation chain (Glode and Opp 2016, Glode, Opp, and Zhang 2019, and He and Li 2022) makes the observation that each layer of the chain gradually solves the inefficiency in the economy. An important difference and a novel theoretical observation here is that when effort is considered, the inefficiencies are mostly resolved by the final layers of the chain and it is difficult to incentivize effort early on.

It is also worth noting that Proposition 6 is a strong result in that it does not rely on M or N_C being large or any assumptions on intermediary quality n_S (similar to Proposition 3). Furthermore, it is also robust to the deviation incentive of skipping over layers due to constant quality across layers and the fact that intermediaries' effort is weakly increasing over layers.

Mapping to reality, it is sometimes puzzling why bond trades are intermediated many times (see He and Li 2022 for details). Through the lens of this model, one explanation could be that the intermediary sectors along the bond-trading chain have largely homogeneous quality ($n_i^* = n_S$). Having a longer chain allows intermediaries in each layer to process a larger volume of transactions, creating higher effort incentives.

5.2 Dissipative Costs and Optimal Chain Length

The convergence property associated with a long chain (Propositions 5 and 6 and expression (21) as $I \to \infty$) implies that the marginal benefit associated with one additional layer diminishes in the limit. Hence, if there is an arbitrarily small dissipative cost associated with each layer, such as communication, contracting, or enforcement costs between intermediaries, then the optimal chain length is finite. Formally, I study the welfare maximization problem (slightly modified from (15)) by a social planner who can decide the entry costs C_i and chain length I.

$$\max_{C_i, I} W = \underbrace{MN_C \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{i=0}^{I} \mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{total consumer utility}} - \underbrace{c(e_0)n_0 N_C - \sum_{i=1}^{I-1} c(e_i) N_{S,i} - c(e_I) N_{S,I} M}_{\text{total costs of screening efforts}} - \psi I M N_C, \tag{22}$$

where ψ is the aforementioned dissipative cost per intermediation layer (I) for each consumer-product transaction (MN_C).

To preserve space, I study an illustrative case based on the upper bound on welfare as in (21), noting without providing details that the full social planner's problem bears a similar conclusion. Ignoring all effort costs, the welfare in (22) is bounded above by

$$MN_C\left[1-(1-n_0)(1-p_0)^{I+1}\right]-\psi IMN_C.$$

Ignoring the integer constraint, the optimal chain length I^* solves

$$-(1-n_0)(1-p_0)^{I^*+1}\ln(1-p_0)=\psi,$$

or equivalently

$$I^*=\frac{\ln\frac{\psi}{-(1-n_0)\ln(1-p_0)}}{\ln(1-p_0)}<\infty.$$

6 Robustness and Extensions

6.1 Fixed Fee and Performance Fee

This paper focuses on the market structure of the intermediary sector. To minimize other economic channels, I have reduced the compensation of intermediaries to a simple performance pay (i.e., a payment of r when a good product is ultimately delivered to a consumer). This is, in my view, the simplest approach to incorporate endogenous price and effort choice. One natural extension is to introduce a fixed pay based on the flow of consumers through an intermediary, independent of the ultimate performance.

Formally, the intermediary payoff now has two components. Denote by f the fee that an intermediary can collect whenever a product (regardless of its quality) is delivered to a consumer through this intermediary. In addition and as before, if the product is good, the intermediary also charges a performance-based fee r, and the total compensation is $f+r$.

For simplicity, I consider the case where intermediaries do not know their skills $\theta=\frac{1}{2}$. Expression (18) gives the break-even condition of an intermediary in layer i, which can be simplified to

$$\frac{MN_C}{N_i}\underbrace{\left[f+r\begin{pmatrix} n_0 & 1-n_0\end{pmatrix}\Pi_{j=0}^{I}\mathbb{P}_j\begin{pmatrix}1\\0\end{pmatrix}\right]}_{\text{total compensation}}-n_Sc(e_i)\left[1+\mathbf{1}_{i=I}(M-1)\right]-C_i=0. \tag{23}$$

Intuitively, the N_i intermediaries in layer i share all consumer transactions MN_C ex-ante even though ex-post skilled and unskilled ones receive a different amount of consumer delegation. Also from an ex-ante perspective, the probability that a consumer receives a good product is $\begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{I} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, hence, the total compensation in (23). With probability n_S the intermediary is skilled, thereby incurring the effort cost. Finally, intermediaries pay the unconditional entry costs C_i.

The main finding is that the introduction of a fixed fee f as a parameter does not generate new equilibrium outcomes.

Proposition 7 *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* $\hat{C}_i = C_i - f\frac{MN_C}{N_i}$ *without a fixed fee* $\hat{f} = 0$.

Proposition 7 is easy to check as the intermediaries' break-even condition (23) is unaltered by the entry cost modification and elimination of f. In addition, everyone's effort incentive is also independent of the fixed fee and entry costs. Intuitively, the fixed pay f can be interpreted as a subsidy to the entry cost C_i and therefore does not affect the outcome.

With additional model ingredients, one can potentially endogenize f. For example, in the model where intermediaries can pay a cost K to become skilled, there will be one more equilibrium condition that

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

which can pin down the equilibrium fixed fee f^*. As determining a compensation structure is not the focus of this paper, I leave this question for future research.

6.2 Heterogeneous Compensation across Layers

Another natural extension is to allow intermediaries operating in different layers to receive different compensation r_i^*. Similar to the fixed fee f^* extension discussed in Subsection 6.1, the simplicity of the current model does not provide enough equilibrium conditions to pin down the performance fee r_i^* for each layer. Intuitively, the fixed supply of intermediaries ($\sum_{i=1}^{I} N_i = 1$) determines the aggregate level of performance fee r^* in the current model. To individually determine the layer-dependent performance fee r_i^*, one needs to have a fixed supply of intermediaries N_i for each layer. However, the size of each layer N_i itself is a variable of interest in this model and therefore cannot be simultaneously determined with r_i.

One possible way to endogenize r_i^* is to introduce competition and the market power of intermediaries in each layer, depending on the sector size N_i. To keep the focus of this paper, I again leave this possibility to future research.

6.3 Multiple Skill Types

For expositional clarity, I assume binary skill types in the main model, which I now relax. Suppose there are T discrete types with different screening technologies

$$P_t(e_{i,t}, \delta_{i+1}) \in \Delta^T \equiv \{x \in \mathbb{R}^T | x_i \in [0,1], \sum x_i = 1\}, \tag{24}$$

where $\delta_{i+1} \in \Delta^T$ is the distribution of types in layer $i+1$. Type $t = 1, 2, ..., T$ intermediary in layer i can input effort $e_{i,t}$ and, based on the prior distribution δ_{i+1}, can find an intermediary in layer $i+1$ with distribution $P_t(e_{i,t}, \delta_{i+1})$ as in (24). Clearly, this setup nests the main model as a special case. The 2-dimensional screening matrix in (2) is replaced by a T-dimensional matrix

$$\mathbb{P}_i \equiv \begin{pmatrix} P_1(e_{i,1}, \delta_{i+1}) \\ ... \\ P_t(e_{i,t}, \delta_{i+1}) \\ ... \\ P_T(e_{i,T}, \delta_{i+1}) \end{pmatrix}, \quad i = 0, 1, 2, ..., I. \tag{25}$$

I intuitively argue, without providing the technical details, that the insights from Sections 3 to 5 remain robust.

Revisiting the single-layer specification $I = 1$ as in Proposition 1, when both N_C and $M \to \infty$, all types of intermediaries have large effort incentive due to delegation as long as their effort inputs result in first-order stochastic dominance in P_t. As such, consumer payoff improves due to better screening outcome 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 product market. As in Proposition 2, consumers here do not need to screen. In the latter case, similar to Proposition 3, consumers 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 product market. In both cases, consumer payoff can approach first best.

Finally, when skill types are unknown, all intermediary layers share the same prior skill distribution, denoted by δ. In this case, 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 5. As the chain becomes sufficiently long $I \to \infty$, the limiting matrix $\mathbb{P}^I$ depends on the converging properties of $\mathbb{P}$. If a sufficiently high effort e_1 results in a type-1 intermediary almost certainly finding another type-1 intermediary or a good product ($P_1(e_1, \delta) \to \begin{pmatrix} 1 & 0 & ... & 0 \end{pmatrix}$ 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 a good product approaches 1. The construction in Proposition 6 remains valid as an implementation of the first-best outcome.

7 Conclusion

Skilled intermediaries can help many consumers find good products. However, consumers still need to find skilled intermediaries and compensate them for their effort. I construct a model of an intermediary market structure focusing on the number of layers, sector size, and skill distribution along the intermediation chain, as well as the effort choices of the intermediaries. I show that having superior quality or skill is not necessary for efficient intermediation to occur. Having a short (two-layer) intermediation chain makes it possible for intermediaries to self-select into different layers. When intermediaries can self-select based on their skill types, structures featuring heterogeneous skill distributions across the two layers approximate the first-best outcome. Finally, when intermediaries do not perfectly know their skill types, thereby reducing the effectiveness of their self-selection, a long intermediation chain can again restore the first best. The inefficiencies are resolved mostly by the final layers of the chain.

The model has various applications across different markets, including, for example, social media platforms and influencers and the delegated asset management industry. Many interesting details of these markets are missing on purpose, as the model is designed to focus on the common feature of these intermediary sectors: they help their clients select products or next layer intermediaries. The model can be enriched to include rebates from product manufacturers to influencers, a quality choice by manufacturers, pricing of products or assets, competition among intermediaries in each layer, and so on. I look forward to future research offering insights into these topics.

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] Cong, Lin William, and Siguang Li. "A Model of Influencer Economy." Available at SSRN 3974661 (2021).

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

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

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

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

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

[13] 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.

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

[15] 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.

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

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

[18] 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.

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

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:

$$\max_p A_i p - \hat{c}(p), \tag{26}$$

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 (26) is identical to the original problem, their solutions must also coincide $p_i^* = P(e_i^*, n)$. The first-order condition of (26) with respect to p is therefore given by

$$A_i = \hat{c}'(p_i^*). \tag{27}$$

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

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

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,

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

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: I first show that the effort inputs of both skilled consumers and intermediaries approach infinity: $e_0^*, e_1^* \to \infty$. Because $\Pi_{U,1} \leq \Pi_{S,1}$, the breakeven condition (10) implies that $\Pi_{S,1} \geq 0$ and $\Pi_{U,1} \leq 0$. 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 (9) together with the fact that $\Pi_{S,1} \geq 0$ imply that

$$r^* \geq \frac{C_1}{MN_C P(e_1, n_P)\frac{P(e_0,n_S)n_0 + P(0,n_S)(1-n_0)}{n_S}} = a'\left(N_C\right)^{b-1}, \tag{28}$$

for some constant $a' > 0$. Condition (9) also implies that the intermediary's effort choice $e_1^* \to \infty$, as $N_C \to \infty$.

Furthermore, Lemma 1 implies that the average cost of screening effort in condition (9) vanishes

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

Condition (8) together with the fact that $\Pi_{U,1} \leq 0$ imply that

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

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 (28) can be strengthened to

$$r^* \sim \left(N_C\right)^{b-1} \to 0, \tag{29}$$

where the convergence to 0 is independent of M and holds whenever $N_C \to \infty$. Consequently, skilled consumers' effort incentive in (5) $M(1-r^*)\,[P(e_1^*, n_P) - P(0, n_P)] \to \infty$ as $M \to \infty$, and therefore, $e_0^* \to \infty$.

Next, apply Lemma 1 to the skilled consumer's optimization problem in (5), and it implies that $\frac{c(e_0^*)}{M} \to 0$.

We can now calculate consumer payoffs in the limiting case as $M, N_C \to \infty$. For skilled consumers,

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

For unskilled consumers,

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

completing the proof. ■

Proof of Proposition 2: In this conjectured equilibrium, consumers (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

$$M \frac{1}{N_{S,1}^*} r^* N_C [P(e_2^*, n_P) - P(0, n_P)] \frac{\partial P(e_1^*, n_2^*)}{\partial e_1^*} = c'(e_1^*), \tag{30}$$

2. first-order condition of skilled layer 2 intermediaries

$$\frac{P(e_1^*, n_2^*)}{n_S - N_{S,1}^*} r^* N_C \frac{\partial P(e_2^*, n_P)}{\partial e_2^*} = c'(e_2^*), \tag{31}$$

3. skilled intermediaries indifference condition between the two layers

$$\begin{aligned} &M \tfrac{1}{N_{S,1}^*} r^* N_C [P(e_1^*, n_2^*)P(e_2^*, n_P) + (1 - P(e_1^*, n_2^*)) P(0, n_P)] - c(e_1^*) - C_1 \\ &= M \left[\tfrac{P(e_1^*, n_2^*)}{n_S - N_{S,1}^*} r^* N_C P(e_2^*, n_P) - c(e_2^*) \right] - C_2 \geq 0, \end{aligned} \tag{32}$$

4. unskilled intermediaries break even in layer 2

$$M \frac{1 - P(e_1^*, n_2^*)}{1 - n_S} r^* N_C P(0, n_P) = C_2, \tag{33}$$

5. and finally, unskilled intermediaries do not wish to enter layer 1 (an inequality

constraint that needs to satisfy in equilibrium)

$$M\frac{1}{N_{S,1}^*}r^*N_C\left[P(0,n_1)P(e_2^*,n_P)+(1-P(0,n_1))\,P(0,n_P)\right]\leq C_1. \tag{34}$$

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 $N_C\to\infty$, such that $\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}=N_C^{x_1}$ and $N_{S,1}^*=N_C^{x_2}$, 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 (30) and (31), we have

$$\frac{\frac{\partial P(e_2^*,n_P)}{\partial e_2^*}}{c'(e_2^*)}=M\frac{n_S-N_{S,1}^*}{N_{S,1}^*}\frac{P(e_2^*,n_P)-P(0,n_P)}{P(e_1^*,n_2^*)}\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}. \tag{35}$$

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

$$\lim_{N_C\to\infty}\frac{\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}}{N_{S,1}^*}=\lim_{N_C\to\infty}N_C^{x_1-x_2}=0 \text{ and } \lim_{N_C\to\infty}\frac{\frac{\frac{\partial P(e_1^*,n_2^*)}{\partial e_1^*}}{c'(e_1^*)}}{N_{S,1}^*}N_C=\lim_{N_C\to\infty}N_C^{1+x_1-x_2}=\infty, \tag{36}$$

then (35) implies that its solution $e_2^*\to\infty$. In addition, (30) implies that $r^*\to 0$ and $r^*N_C\to\infty$. Conditions (32) and (33) jointly determine $C_2>0$ and C_1.

Next, I show that $C_1\gg C_2$. Manipulate (32), we have

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

where the last inequality is due to the optimality of e_1^*. Let $N_C\to\infty$, 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

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

Note from (36) that $r^* N_C \to \infty$ and $N^*_{S,1} \to 0$, so $C_1 \gg C_2$.

Furthermore, I verify the inequality condition (34). Using (32) and (33), we know that (34) is equivalent to

$$\begin{aligned} &\tfrac{1}{N^*_{S,1}} \left\{ \left[P(e_1^*, n_2^*) - P(0, n_1)\right] \left[P(e_2^*, n_P) - P(0, n_P)\right] - \tfrac{c(e_1^*) N^*_{S,1}}{M r^* N_C} \right\} \\ &\geq \left[\tfrac{P(e_1^*, n_2^*)}{n_S - N^*_{S,1}} P(e_2^*, n_P) - \tfrac{1 - P(e_1^*, n_2^*)}{1 - n_S} P(0, n_P) \right] - \tfrac{c(e_2^*)}{r^* N_C} \end{aligned} \tag{37}$$

Apply Lemma 1 to condition (30) implies that $\frac{c(e_1^*) N^*_{S,1}}{M r^* N_C} \to 0$, as $N^*_{S,1} \to 0$. It immediately follows that the left-hand side of (37) approaches ∞ as $N^*_{S,1} \to 0$, whereas the right-hand side is bounded. Hence, the inequality condition (34) holds.

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

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

completing the proof. ■

Proof of Proposition 3: 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 consumers 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

$$\Pi_{S,1} = \Pi_{U,1} = M \frac{1}{1 - N^*_{S,2}} r^* N_C P(e_2^*, n_P) - C_1 = 0, \tag{38}$$

$$\Pi_{S,2} = M \left[\frac{1}{N^*_{S,2}} r^* N_C P(e_2^*, n_P) - c(e_2^*) \right] - C_2 = 0, \tag{39}$$

and

$$M r^* N_C \frac{\partial P(e_2^*, n_P)}{\partial e_2^*} = N^*_{S,2} c'(e_2^*). \tag{40}$$

Using (38) and (39) to eliminate r^*, we have

$$\frac{N^*_{S,2}}{1 - N^*_{S,2}} = \frac{C_1}{M c(e_2^*) + C_2}. \tag{41}$$

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 (38) and (40), we have

$$\frac{\frac{\partial P(e^*_2, n_P)}{\partial e^*_2}}{P(e^*_2, n_P)} = \frac{c'(e^*_2)}{Mc(e^*_2) + C_2}. \tag{42}$$

To see that (42) 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 (42).

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 (42) 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 (38), $N^*_{S,2} \to 0$, and $C_1 \to 0$, and it quickly follows $r^* \to 0$. Therefore, the consumer payoff

$$\Pi_C = M(1 - 2r^*)P(e^*_2, n_P) \to M,$$

completing the proof. ■

Proof of Proposition 5: The screening matrix can be diagonalized as follows

$$\mathbb{P} = \begin{pmatrix} 1 & 1-p_1 \\ 1 & -p_0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & p_1 - p_0 \end{pmatrix} \begin{pmatrix} \frac{p_0}{1-p_1+p_0} & \frac{1-p_1}{1-p_1+p_0} \\ \frac{1}{1-p_1+p_0} & -\frac{1}{1-p_1+p_0} \end{pmatrix},$$

where

$$\begin{pmatrix} \frac{p_0}{1-p_1+p_0} & \frac{1-p_1}{1-p_1+p_0} \\ \frac{1}{1-p_1+p_0} & -\frac{1}{1-p_1+p_0} \end{pmatrix} = \begin{pmatrix} 1 & 1-p_1 \\ 1 & -p_0 \end{pmatrix}^{-1}.$$

Hence,

$$\begin{aligned}
\mathbb{P}^i &= \begin{pmatrix} 1 & 1-p_1 \\ 1 & -p_0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (p_1-p_0)^i \end{pmatrix} \begin{pmatrix} \frac{p_0}{1-p_1+p_0} & \frac{1-p_1}{1-p_1+p_0} \\ \frac{1}{1-p_1+p_0} & -\frac{1}{1-p_1+p_0} \end{pmatrix} \\
&= \begin{pmatrix} 1 & (1-p_1)(p_1-p_0)^i \\ 1 & -p_0(p_1-p_0)^i \end{pmatrix} \begin{pmatrix} \frac{p_0}{1-p_1+p_0} & \frac{1-p_1}{1-p_1+p_0} \\ \frac{1}{1-p_1+p_0} & -\frac{1}{1-p_1+p_0} \end{pmatrix} \\
&= \begin{pmatrix} \frac{p_0+(1-p_1)(p_1-p_0)^i}{1-p_1+p_0} & \frac{(1-p_1)-(1-p_1)(p_1-p_0)^i}{1-p_1+p_0} \\ \frac{p_0-p_0(p_1-p_0)^i}{1-p_1+p_0} & \frac{1-p_1+p_0(p_1-p_0)^i}{1-p_1+p_0} \end{pmatrix} \\
&= \begin{pmatrix} p^\dagger + \left(1-p^\dagger\right)(p_1-p_0)^i & 1-p^\dagger - \left(1-p^\dagger\right)(p_1-p_0)^i \\ p^\dagger - p^\dagger (p_1-p_0)^i & 1-p^\dagger + p^\dagger (p_1-p_0)^i \end{pmatrix}.
\end{aligned}$$

The probability that a skilled intermediary in layer $j+i$ can be identified is given by

$$\begin{aligned}
p^{j+i} &= \begin{pmatrix} p^j & 1-p^j \end{pmatrix} \mathbb{P}^i \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p^j & 1-p^j \end{pmatrix} \begin{pmatrix} p^\dagger + \left(1-p^\dagger\right)(p_1-p_0)^i \\ p^\dagger - p^\dagger (p_1-p_0)^i \end{pmatrix} \\
&= p^\dagger + \left(p^j - p^\dagger\right)(p_1-p_0)^i.
\end{aligned}$$

■

Proof of Proposition (6): The proof is constructive. Fix an arbitrarily small ϵ and a sufficiently large e^*. I will construct a set of entry costs C_i such that the conjectured features in the statement hold in equilibrium. The effort choices e_i^* $(1 \leq i \leq I)$ must satisfy the following I first-order conditions.

$$\begin{aligned}
&\begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{j=0}^{i-1} \mathbb{P}_j \begin{pmatrix} 1 \\ 0 \end{pmatrix} M N_C r \Pi_{j=i+1}^{I} \left(P(e_j^*, n_S) - P(0, n_S)\right) \frac{\partial P(e_i^*, n_S)}{\partial e_i^*} \\
&= n_S N_i c'(e_i^*) \left[1 + \mathbf{1}_{i=I}(M-1)\right].
\end{aligned} \tag{43}$$

Plugging in $e_1 = \epsilon$ and $e_i = e^*$ for $i \geq 2$, and together with

$$\sum_{i=1}^{I} N_i = 1,$$

there are $I+1$ linear conditions for a total of $I+1$ variables $\{N_i, r\}$. Hence, a solution exists. The entry costs C_i are then determined by the break-even conditions of the intermediaries in each layer.

Next, I show that features 1-4 hold in this constructed equilibrium. For feature 1, I only need to verify that the consumer effort e_0 can be arbitrarily small. Consider the first-order condition with respect to e_0:

$$M(1-rI)\Pi_{j=1}^{I}\left(P(e_j^*, n_S) - P(0, n_S)\right)\frac{\partial P(e_0^*, n_S)}{\partial e_0^*} = c'(e_0^*).$$

Since $P(e_j^*, n_S) \leq 1$ for any $j \geq 2$ and $e_1^* = \epsilon$ is arbitrarily small, the above condition implies

$$M\left(P(\epsilon, n_S) - P(0, n_S)\right)\frac{\partial P(e_0^*, n_S)}{\partial e_0^*} \geq c'(e_0^*),$$

which in turn implies that $\frac{c'(e_0^*)}{\frac{\partial P(e_0^*, n_S)}{\partial e_0^*}}$ can be arbitrarily small and so is e_0^*. This establishes feature 1.

Feature 2 is automatic by construction. Next, I show feature 3. First, consider the first-order condition (43) with respect to e_1^*:

$$\begin{pmatrix} n_0 & 1-n_0 \end{pmatrix}\mathbb{P}_0\begin{pmatrix} 1 \\ 0 \end{pmatrix} MN_C r\left(P(e^*, n_S) - P(0, n_S)\right)^{I-1}\frac{\partial P(e_1^*, n_S)}{\partial e_1^*} = n_S N_1 c'(e_1^*).$$

Note that

$$\begin{pmatrix} n_0 & 1-n_0 \end{pmatrix}\mathbb{P}_0\begin{pmatrix} 1 \\ 0 \end{pmatrix} \geq P(0, n_S) = n_S.$$

Hence,

$$N_1 \geq \frac{\frac{\partial P(e_1^*, n_S)}{\partial e_1^*}}{c'(e_1^*)} MN_C r\left(P(e^*, n_S) - P(0, n_S)\right)^{I-1}. \tag{44}$$

Similarly, because

$$\begin{pmatrix} n_0 & 1-n_0 \end{pmatrix}\Pi_{j=0}^{i-1}\mathbb{P}_j\begin{pmatrix} 1 \\ 0 \end{pmatrix} \leq 1$$

and

$$\Pi_{j=i+1}^{I}\left(P(e_j^*, n_S) - P(0, n_S)\right) \leq 1,$$

the first-order conditions (43) with respect to e_i, $i \geq 2$ implies that

$$MN_C r\frac{\partial P(e^*, n_S)}{\partial e^*} \geq n_S N_i c'(e^*)\left[1 + \mathbf{1}_{i=I}(M-1)\right].$$

Hence,

$$N_i \leq \frac{\frac{\partial P(e^*,n_S)}{\partial e^*}}{c'(e^*)} \frac{MN_C r}{n_S \left[1 + \mathbf{1}_{i=I}(M-1)\right]}. \tag{45}$$

Taking the ratio between (44) and (45), we have

$$\frac{N_1}{\sum_{i=2}^{I} N_i} \geq \frac{\frac{\frac{\partial P(e_1^*,n_S)}{\partial e_1^*}}{c'(e_1^*)}}{\frac{\frac{\partial P(e^*,n_S)}{\partial e^*}}{c'(e^*)}} \frac{\left(P(e^*,n_S) - P(0,n_S)\right)^{I-1} n_S \left[1 + \mathbf{1}_{i=I}(M-1)\right]}{I-1}.$$

The first fraction approaches ∞ whenever $e_1^* = \epsilon \to 0$ and $e^* \to \infty$. Since $\sum_{i=1}^{I} N_i = 1$, it follows that $N_1 \to 1$ and $N_i \to 0$ for all $i \geq 2$, establishing feature 3.

Since $N_1 < 1$ and $\frac{\frac{\partial P(e_1^*,n_S)}{\partial e_1^*}}{c'(e_1^*)} \to \infty$ as $e_1^* \to 0$, condition (44) implies that $r \to 0$. This establishes feature 4.

Finally, to show feature 5, I calculate consumer payoff. For any I, as long as the chosen ϵ is sufficiently small and e^* is sufficiently large, we have $e_0 \to 0$ (feature 1), $r = \frac{1}{I^2} \to 0$ (feature 4) and $P(e^*, n_S) \to 1$. Proposition 5 implies that the probability that a consumer finds a good product converges to

$$\frac{P(0,n_S)}{1 - P(e^*,n_S) + P(0,n_S)} \to 1.$$

Hence, the total welfare given by (15) is

$$W = \underbrace{M(1-rI)N_C \begin{pmatrix} n_0 & 1-n_0 \end{pmatrix} \Pi_{i=0}^{I} \mathbb{P}_i \begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\text{total consumer utility}} \to MN_C,$$

completing the proof. ■