NBER WORKING PAPER SERIES

THE EMERGENCE OF MARKET STRUCTURE

Maryam Farboodi
Gregor Jarosch
Robert Shimer

Working Paper 23234
http://www.nber.org/papers/w23234

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2017

This paper was previously circulated under the title “Meeting Technologies in Decentralized Asset Markets”. We are grateful to Fernando Alvarez, Markus Brunnermeier, Xavier Gabaix, Ricardo Lagos, Pierre-Olivier Weill, Randall Wright, and audiences at various seminars and conferences for their thoughts and comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w23234.ack

The Emergence of Market Structure
Maryam Farboodi, Gregor Jarosch, and Robert Shimer
NBER Working Paper No. 23234
March 2017
JEL No. E44,G12,G20

ABSTRACT

What market structure emerges when market participants can choose the rate at which they contact others? We show that traders who choose a higher contact rate emerge as intermediaries, earning profits by taking asset positions that are misaligned with their preferences. Some of them, middlemen, are in constant contact with other traders and so pass on their position immediately. As search costs vanish, traders still make dispersed investments and trade occurs in intermediation chains, so the economy does not converge to a centralized market. When search costs are a differentiable function of the contact rate, the endogenous distribution of contact rates has no mass points. When the function is weakly convex, faster traders are misaligned more frequently than slower traders. When the function is linear, the contact rate distribution has a Pareto tail with parameter 2 and middlemen emerge endogenously. These features arise not only in the (inefficient) equilibrium allocation, but also in the optimal allocation. Moreover, we show that intermediation is key to the emergence of the rest of the properties of this market structure.

Maryam Farboodi
26 Prospect Ave
Bendheim Center for Finance
Princeton University
Princeton, NJ 08540
farboodi@princeton.edu

Gregor Jarosch
Department of Economics
Stanford University
579 Serra Mall
Palo Alto, CA 94305
and NBER
gregorjarosch@gmail.com

Robert Shimer
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
shimer@uchicago.edu

1 Introduction

This paper examines an over-the-counter market for assets where traders periodically meet in pairs with the opportunity to trade (Rubinstein and Wolinsky, 1987). We are interested in understanding the origins and implications of the observed heterogeneity in these markets, whereby we mean that some individuals trade much more frequently and with many more partners than others do. In particular, real world trading networks appear to have a core-periphery structure. Traders at the core of the network act as financial intermediaries, earning profits by taking either side of a trade, while traders in the periphery trade less frequently and their trades are more geared towards obtaining an asset position aligned with their portfolio needs.

We consider a model economy with a unit measure of traders each of whom seeks to trade a single asset for an outside good (money). Following Duffie, Gârleanu and Pedersen (2005) and a large subsequent literature, we assume traders have an intrinsic reason for trade, differences in the flow utility they receive from holding the asset. Moreover, this idiosyncratic valuation changes over time, creating a motive for continual trading and retrading. We add to this a second source of heterogeneity, namely in contact rates. We allow for traders to differ in terms of the frequency at which they meet others. In addition, we assume that the likelihood of contacting any particular trader is proportional to her contact rate.

Under these assumptions, we show that intermediation arises naturally. When two traders who have the same flow valuation for the asset meet, the trader who has a higher contact rate acts as an intermediary, leaving the meeting with holdings that are further from the intrinsically desired one. This occurs in equilibrium because traders with a faster contact rate expect to have more future trading opportunities and so place less weight on their current flow payoff. Intermediation thus moves misaligned asset holdings towards traders with higher contact rates, which improves future trading opportunities. Thus the equilibrium displays a core-periphery structure where the identity of the market participants at the core—fast traders—remains stable over time.[1]

The full model recognizes that traders' contact rates are endogenous. We consider an initial, irreversible investment in this meeting technology. For example, traders may invest

[1]Recent empirical work documents that bilateral asset markets frequently exhibit a core-periphery network structure where few central institutions account for most of the turnover while the majority of market participants remains at the fringe. For the federal funds market, see Bech and Atalay (2010), Allen and Saunders (1986), and Afonso, Kovner and Schoar (2013). For evidence on international interbank lending, see Boss, Elsinger, Summer and Thurner (2004), Chang, Lima, Guerra and Tabak (2008), Craig and Von Peter (2014), and in 't Veld and van Lelyveld (2014). For credit default swaps Peltonen, Scheicher and Vuillemey (2014) and Siriwardane (2015), for the corporate bond market Di Maggio, Kermani and Song (2015), for the municipal bond market Li and Schürhoff (2014), and for asset-backed securities Hollifield, Neklyudov and Spatt (2014).

in having faster communication technologies, better visibility through location choices or advertisement, or relationships with more counterparties. While we assume throughout that traders are ex ante identical, we recognize that they may choose different contact rates, as in a mixed strategy equilibrium. A higher contact rate gives more trading opportunities but we assume it also incurs a higher sunk cost.

We prove that if the cost is a differentiable function of the contact rate, then any equilibrium allocation must have a continuous distribution of contact rates. The force pushing towards heterogeneity is the gains from intermediation. If everyone else chooses the same contact rate, a trader who chooses a slightly faster contact rate acts as an intermediary for everyone else, repeatedly buying and selling irrespective of her intrinsic valuation; while a trader who chooses a slightly slower contact rate never trades once her asset position is aligned with her preferences. The marginal returns to additional meetings thus jump discretely at any mass point, inconsistent with equilibrium under a differentiable cost function. In addition, we show that if the cost function is weakly convex, the equilibrium rate of misalignment is strictly increasing in the contact rate. That is, a higher contact rate comes with an inferior asset position (relative to fundamentals) and derives its benefits from trading profits. In turn, traders on the fringe of the trading network have well-aligned asset positions but pay for the intermediation services provided by the core through bid-ask spreads.

We then turn to the natural assumption that the cost is proportional to the contact rate. We prove that the equilibrium distribution of contact rates has a positive lower bound and is unbounded above, with many traders choosing a high contact rate: First, the right tail of the contact rate distribution is Pareto with tail parameter 2, so the variance in contact rates is infinite. To the best of our knowledge, we are the first to show that a power law is an equilibrium outcome when homogeneous individuals choose their search technology under linear cost. This result carries over to the distribution of trading frequencies and connects our theory tightly with empirical evidence on frictional asset markets.[2] Second, we show that a zero measure of traders chooses an infinite contact rate, giving them continuous contact with the market. These "middlemen" account for a positive fraction of all meetings, earning zero profits in each meeting but making it up on volume. We stress that ex ante there is no difference between middlemen (the core of the network) and the periphery; however, they choose to make different investments and so ultimately play a very different role in the

[2]There is ample empirical evidence on concentration of trade among very few financial institutions. The largest sixteen derivatives dealers intermediate more than 80 percent of the global total notional amount of outstanding derivatives (Mengle, 2010; Heller and Vause, 2012). Bech and Atalay (2010) document that the distribution of trading frequencies in the federal funds markets is well-approximated by a power law, while Peltonen, Scheicher and Vuillemey (2014) find that the degree distribution of the aggregate credit default swap network can be scale-free. For additional financial market variables that are—at least in the tail—well approximated by power laws, see Gabaix, Gopikrishnan, Plerou and Stanley (2006).

market.

We then demonstrate that these forces remain important even in an environment where the cost per contact goes to zero, so search frictions vanish. In this frictionless limit, 27 percent of contacts are with middlemen. Moreover, middlemen account for 41 percent of trading volume and intermediation chains are long: whenever a trader experiences a preference shock, it gives rise to a sequence of trades that end only when two middlemen swap the asset. As a consequence, trading volume far exceeds the minimal amount of reallocation needed to offset preference shocks.

We complement these findings with some numerical results. First, we show that both the distribution of contact rates as well the distribution of trading frequencies are globally well-approximated by a Pareto distribution with tail parameter 2. Second, we develop a numerical proof that the equilibrium exists and is unique.

We then consider optimal trading patterns and investments. We prove that the equilibrium trading pattern—passing misalignment to traders with higher contact rates—is optimal. We also show that all the qualitative features of the equilibrium allocation carry over to the optimum: If costs are a differentiable function of the contact rate, there is optimally no mass points in the distribution of contact rates. If the cost function is weakly convex, faster traders optimally have asset positions that are increasingly detached from their intrinsic preferences. If costs are proportional to contact rates, the optimal contact rate distribution has a Pareto tail with parameter 2 and a zero measure of middlemen account for a positive fraction of all meetings. As costs converge to zero, middlemen optimally account for 37 percent of the meetings and 45 percent of trading volume, so trading volume still far exceeds the minimal amount of reallocation needed to offset preference shocks.

The equilibrium is inefficient due to search externalities. Pigouvian taxes highlight the inefficiencies: traders only capture half the surplus in each meeting, leading to underinvestment in contacts; but they do not internalize a business stealing effect, which induces them to overinvest in contacts. We numerically contrast the equilibrium with the optimum and find systematic overinvestment: The equilibrium has too few slow types and too much intermediation The Pigouvian transfer offsets this by altering equilibrium prices in favor of the buyers of intermediation services, namely slower traders.

Finally, we emphasize the connection between intermediation and dispersion in contact rates. We consider an economy in which traders with the same desired asset holdings never meet, which eliminates the possibility of intermediation in our model economy. Under these conditions, we show that if the cost is a weakly convex function of the contact rate, all traders choose the same contact rate, both in equilibrium and optimally. Thus dispersion in contact rates and intermediation are intimately connected: if there is dispersion in contact

rates, faster traders act as intermediaries; and if intermediation is permitted, contact rates are naturally disperse.

Related Work This paper is closely related to a growing body of work on trade and intermediation in markets with search frictions. Rubinstein and Wolinsky (1987) were the first to model middlemen in a frictional goods market. We share with them the notion that intermediaries have access to a superior search technology.[3] In two important papers Duffie, Gârleanu and Pedersen (2005, 2007) study an over-the-counter asset market where time-varying taste leads to trade. This is also the fundamental force giving rise to gains from trade in our setup. Much of the more recent theoretical work extends their basic framework to accommodate newly available empirical evidence on trade and intermediation in over-the-counter markets.

The decentralized interdealer market in Neklyudov (2014) features dealers with heterogeneous contact rates. The same dimension of heterogeneity is present in Üslü (2016) who also allows for heterogeneity in pricing and inventory holdings.[4] As in our framework, fast dealers in these setups are more willing to take on misaligned asset positions, thus endogenously emerging as intermediaries. The marketplace features intermediation chains and a core-periphery trading network with fast traders at the core. We add to this literature by first showing that heterogeneity in meeting technologies arises naturally to leverage the gains from intermediation even with ex-ante homogeneous agents, and second by showing how the endogenous choice of contact rates disciplines key features of the contact rate distribution. Additionally, our normative analysis shows that both technological heterogeneity and intermediation by those with a high contact rate are socially desirable.

Hugonnier, Lester and Weill (2016) model a market with a continuum of flow valuations which gives rise to intermediation chains; market participants with extreme flow value constitute the periphery and those with moderate flow value constitute the core. Afonso and Lagos (2015) similarly has endogenous intermediation because banks with heterogeneous asset positions buy and sell depending on their counterparties' reserve holdings. In contrast to these setups, ours offers a theory where the identity of the individuals at the center of the intermediation chain remains stable over time, a key empirical feature of many decentralized asset markets (see, for instance, Bech and Atalay (2010) for the federal funds market.)

The identity of the institutions at the core is also stable in Chang and Zhang (2016), where agents differ in terms of the volatility of their taste for an asset and those with less

[3]Nosal, Wong and Wright (2016) extend Rubinstein and Wolinsky (1987) to allow for heterogeneous bargaining power and storage cost but assume homogeneous contact rates.

[4]A related literature studies the positive and normative consequences of high-speed trading in centralized financial markets; see, for instance, Pagnotta and Philippon (2015).

volatile valuation act as intermediaries. The same is true in our framework but heterogeneity in the volatility of an agent's taste arises endogenously since a higher contact rate buffers the impact of the flow value on the net valuation of asset ownership.

Farboodi, Jarosch and Menzio (2016) model an environment where some have superior bargaining power and emerge as middlemen due to dynamic rent extraction motives which are, at best, neutral for welfare. In contrast, intermediation in our setup improves upon the allocation since misaligned asset positions are traded toward those who are more efficient at offsetting them. They also study an initial investment stage that determines the distribution of bargaining power in the population, but restrict the distribution to two points. We allow for a continuous distribution of contact rates and prove that this is consistent with both equilibrium and optimum.

Furthermore, some of the theoretical work on intermediation in over-the-counter markets features exogenously given middlemen who facilitate trade and have access to a frictionless interdealer market (Duffie, Gârleanu and Pedersen, 2005; Weill, 2008; Lagos and Rocheteau, 2009). We show that such middlemen who are in continuous contact with the market are a natural equilibrium outcome when homogeneous agents invest into a search technology.

Other recent work studies the structure of financial markets using explicit network formation models, which also generate core-periphery network structures (Farboodi, 2015; Wang, 2017) or star networks (Babus and Hu, 2015). In this class of models, agents (traders, banks) form explicit links, over which trade can be executed, at either an explicit cost—the cost of maintaining a relationship as in Babus and Hu (2015) and Wang (2017)—or an implicit cost—the counterparty risk in Farboodi (2015). The cost of acquiring a contact rate in our random search setup is closely related to the price of links in this network formation literature following Jackson (2010). In this body of work, multiplicity arises frequently whereas our equilibrium is unique, at least in the case of a linear cost function. In addition, the network models tend to generate a somewhat extreme core-periphery structure, where traders take on one of two roles, the core or the periphery; and traders in the periphery only trade with those in the core. Our model predicts a continuous distribution of trading frequencies and predicts that trades occur both within the periphery and within the core, as well as between core and periphery.

In summary, while the theoretical literature on frictional asset markets has offered a variety of economic mechanisms that give rise to empirically observed intermediation chains and core-periphery trading structures,our analysis offers novel insights along four distinct dimensions: (i) time-invariant heterogeneity arises endogenously to leverage the gains from trade; (ii) middlemen with continuous market contact arise endogenously; (iii) the tail of the endogenous distribution of contact and trading rates is Pareto and our theory hence connects

with the empirical regularities in a very tight way; (iv) our normative analysis shows that both intermediation and heterogeneity in the search technology are closely interrelated and socially desirable.

Finally, the finding that both the equilibrium and optimal allocations have a Pareto tail relates the paper to a large literature in economics that explores theoretical mechanisms which give rise to endogenous power law distributions (Gabaix, 1999; Eeckhout, 2004; Geerolf, 2016). Many other economically important regularities, such as the distributions of city and firm size and the distributions of income and wealth, are empirically well-approximated by power laws. To the best of our knowledge, the mechanism giving rise to the Pareto tail in our environment is novel and unrelated to the ones that are established in the literature (see Gabaix, 2009, 2016, for an overview).

Outline The rest of the paper is organized as follows: Section 2 lays out the model. Section 3 defines and characterizes the equilibrium. Section 4 discusses the socially optimal allocation and how it can be decentralized. Section 5 considers an economy where intermediation is prohibited. Section 6 concludes.

2 Model

We study a marketplace where time is continuous and extends forever. A unit measure of traders have preferences defined over their holdings of an indivisible asset and their consumption or production of an outside good. Traders have rate of time preference r. The supply of the asset is fixed at $\frac{1}{2}$ and individual traders' holdings are restricted to be $m \in \{0, 1\}$, so at any point in time half the traders hold the asset and half do not. Traders have time-varying taste $i \in \{h, l\}$ for the asset and receive flow utility $\delta_{i,m}$ when they are in state (i, m). We assume that $\Delta \equiv \frac{1}{2}\big(\delta_{h,1} + \delta_{l,0} - \delta_{h,0} - \delta_{l,1}\big) > 0$, which implies that traders in the high state are the natural asset owners. Preferences over net consumption of the outside good are linear, so that good effectively serves as transferable utility when trading the asset.

Traders' taste switches between l and h independently at an identical rate $\gamma > 0$. This implies that at any point in time in a stationary distribution, half the traders are in state h and half are in state l. Thus, in a frictionless environment, the supply of assets is exactly enough to satiate the traders in state h. Search frictions prevent this from happening. Instead a typical trader meets another one according to a Poisson process with arrival rate $\lambda \geq 0$. Trade may occur only at those moments. We assume that traders irrevocably commit to a time-invariant contact rate $\lambda \in [0, \infty]$ at time 0. A high contact rate is costly: a trader who chooses a contact rate λ pays a cost $c(\lambda) > 0$ per meeting.

We allow for the possibility that different traders choose different contact rates. Let $G(\lambda)$ denote the cumulative distribution function of contact rates in the population and let Λ denote the average contact rate. Importantly, we allow for the presence of a zero measure of traders who are *middlemen*, choosing $\lambda = \infty$. Middlemen are in continuous contact with the market and may account for a positive fraction of all meetings. That is, we require that $\Lambda \geq \int_0^\infty \lambda dG(\lambda)$ and allow the inequality to be strict, in which case there are middlemen.

Search is random, so whom the trader meets is independent of her current taste and asset holding, but is proportional to the other trader's contact rate. More precisely, conditional on meeting a counterparty, the counterparty's contact rate falls into some interval $[\lambda_1, \lambda_2]$ with probability $\int_{\lambda_1}^{\lambda_2} \frac{\lambda}{\Lambda} dG(\lambda)$. In addition, the probability of meeting a middleman is $1 - \int_0^\infty \frac{\lambda}{\Lambda} dG(\lambda)$. For any function $f : [0, \infty] \to \mathbb{R}$, it will be convenient to define the expected value of f in a meeting:

$$\mathbb{E}(f(\lambda')) \equiv \int_0^\infty \frac{\lambda'}{\Lambda} f(\lambda') dG(\lambda') + \left(1 - \int_0^\infty \frac{\lambda'}{\Lambda} dG(\lambda')\right) f(\infty)$$

This explicitly accounts for the possibility both of meeting a regular trader and of meeting a middleman. When $\Lambda = 0$, so (almost) everyone chooses a zero contact rate, we assume that a trader who chooses a positive contact rate is equally likely to meet any of the other traders and let $\mathbb{E}(f(\lambda')) = f(0)$, the average value of f in the population.

When two traders meet, their asset holdings, preferences, and contact rates are observed by each. If only one trader holds the asset, as will be the case in half of all meetings, the traders may swap the asset for the outside good. Whether trade occurs and what the terms of trade are is determined according to the (symmetric) Nash bargaining solution.

3 Equilibrium

Our analysis of equilibrium is broken into nine subsections. We start by characterizing the value functions and flow of workers between different states. We then turn our focus to a symmetric equilibria, where a trader's behavior only depends on her contact rate and whether her asset holdings are well-aligned with her preferences. We next explain how we make the distribution of contact rates endogenous and define an equilibrium. The remainder of the section develops seven propositions which characterize the equilibrium. Proposition 1 focuses on which trades occur given a contact rate distribution, while Propositions 2–7 characterize the contact rate distribution under different restrictions on the cost function c. The last section characterizes the equilibrium numerically when the cost function is linear.

3.1 Value Functions and Flows

In equilibrium, we need to find two objects.[5] First, let $\mathbf{1}_{\lambda,i,m}^{\lambda',i',m'}$ denote the probability that a trader with contact rate $\lambda \in [0,\infty]$ in preference state $i \in \{h,l\}$ with asset holdings $m \in \{0,1\}$ trades when she contacts a trader with contact rate $\lambda' \in [0,\infty]$ in preference state $i' \in \{l,h\}$ with asset holdings $m' \in \{0,1\}$. Second, let $p_{\lambda,i,m}^{\lambda',i',m'}$ denote the transfer of the outside good from $\{\lambda,i,m\}$ to $\{\lambda',i',m'\}$ when such a trade takes place. Feasibility requires that $\mathbf{1}_{\lambda,i,m}^{\lambda',i',m'} = \mathbf{1}_{\lambda',i',m'}^{\lambda,i,m}$ and $p_{\lambda,i,m}^{\lambda',i',m'} + p_{\lambda',i',m'}^{\lambda,i,m} \geq 0$, where the latter condition ensures that there are no outside resources available to the trading pair. The trading probability and price are determined by Nash bargaining.

Let $v_{\lambda,i,m}$ denote the present value of the profits of a trader $\{\lambda,i,m\}$. This is defined recursively by

$$rv_{\lambda,i,m} = \delta_{i,m} + \gamma\big(v_{\lambda,\sim i,m} - v_{\lambda,i,m}\big) + \lambda \sum_{i' \in \{h,l\}} \sum_{m' \in \{0,1\}} \mathbb{E}\left(\mathbf{1}_{\lambda,i,m}^{\lambda',i',m'} \mu_{\lambda',i',m'} \big(v_{\lambda,i,m'} - v_{\lambda,i,m} - p_{\lambda,i,m}^{\lambda',i',m'}\big)\right) - \lambda c(\lambda). \quad (1)$$

The left hand side of equation (1) is the flow value of the trader. This comes from four sources, listed sequentially on the right hand side. First, she receives a flow payoff $\delta_{i,m}$ that depends on her preferences and asset holdings. Second, her preferences shift from i to $\sim i$ at rate γ, in which case the trader has a capital gain $v_{\lambda,\sim i,m} - v_{\lambda,i,m}$. Third, she meets another trader at rate λ, in which case they may swap asset holdings in return for a payment. Here $\mu_{\lambda',i',m'}$ denotes the endogenous fraction of traders with contact rate λ' who are in preference state i' and have asset holding m'. If the two agree to trade, with probability $\mathbf{1}_{\lambda,i,m}^{\lambda',i',m'}$, the trader has a capital gain from swapping assets and transferring the outside good, $v_{\lambda,i,m'} - v_{\lambda,i,m} - p_{\lambda,i,m}^{\lambda',i',m'}$. Finally, the trader pays a cost $c(\lambda)$ in each meeting.

The fraction of type λ traders in different states, $\mu_{\lambda,i,m}$, also depends on the trading probabilities through balanced inflows and outflows:

$$\left(\gamma + \lambda \sum_{i' \in \{h,l\}} \mathbb{E}\left(\mathbf{1}_{\lambda,i,m}^{\lambda',i',1-m} \mu_{\lambda',i',1-m}\right)\right) \mu_{\lambda,i,m} = \gamma \mu_{\lambda,\sim i,m} + \lambda \sum_{i',\in \{h,l\}} \mathbb{E}\left(\mathbf{1}_{\lambda,i,1-m}^{\lambda',i',m} \mu_{\lambda',i',m}\right) \mu_{\lambda,i,1-m}. \quad (2)$$

A trader exits the state $\{\lambda,i,m\}$ either when she has a preference shock, at rate γ, or when she meets and succeeds in trading with another trader with the opposite asset holding. A

[5] We focus throughout on steady states.

trader enters this state when she is in the opposite preference state and has a preference shock or she is in the opposite asset holding state and trades.

Nash bargaining imposes that trade occurs whenever it makes both parties better off, and that that trading prices equate the gains from trade without throwing away any resources. That is, if there are prices $p_{\lambda,i,m}^{\lambda',i',m'} + p_{\lambda',i',m'}^{\lambda,i,m} = 0$ such that $v_{\lambda,i,m'} - v_{\lambda,i,m} \geq p_{\lambda,i,m}^{\lambda',i',m'}$ and $v_{\lambda',i',m} - v_{\lambda',i',m'} \geq p_{\lambda',i',m'}^{\lambda,i,m}$, trade occurs at a prices such that $v_{\lambda,i,m'} - v_{\lambda,i,m} - p_{\lambda,i,m}^{\lambda',i',m'} = v_{\lambda',i',m} - v_{\lambda',i',m'} - p_{\lambda',i',m'}^{\lambda,i,m}$. Otherwise there is no trade. It follows immediately that

$$\mathbf{1}_{\lambda,i,m}^{\lambda',i',m'} = \begin{cases} 1 \\ 0 \end{cases} \quad \text{if } v_{\lambda,i,m'} + v_{\lambda',i',m} \gtrless v_{\lambda,i,m} + v_{\lambda',i',m'}; \tag{3}$$

and that the trading prices satisfy

$$p_{\lambda,i,m}^{\lambda',i',m'} = \frac{1}{2}\left(v_{\lambda,i,m'} + v_{\lambda',i',m'} - v_{\lambda,i,m} - v_{\lambda',i',m}\right)$$

Of course, if $m = m'$, there is no possibility for gains from trade. In the remainder of our analysis, we ignore such meetings.

3.2 Symmetry

We call traders' asset holding positions *misaligned* both when they hold the asset and are in preference state l and when they do not hold the asset and are in preference state h. We call traders' asset holding positions *well-aligned* both when they hold the asset and are in preference state h and when they do not hold the asset and are in preference state l. We focus on allocations in which the two misaligned states and the two well-aligned states are treated symmetrically. That is, we look only at equilibria where $\mathbf{1}_{\lambda,i,m}^{\lambda',i',m'} = \mathbf{1}_{\lambda,\sim i,1-m}^{\lambda',\sim i',1-m'}$.[6] That such trading patterns may be consistent with equilibrium is a consequence of our symmetric market structure, where half the traders are in each preference state and half of the traders hold the asset.

In a symmetric equilibrium, equation (2) implies $\mu_{\lambda,i,m} = \mu_{\lambda,\sim i,1-m}$ for all $\{\lambda, i, m\}$. That is, the fraction of traders with contact rate λ in the high state, $i = h$, who hold the asset, $m = 1$, is equal to the fraction of traders with the same contact rate who are in the low state, $i = l$, and do not hold the asset $m = 0$. That is, the fraction of type-λ traders in either well-aligned state are symmetric. The remaining traders are misaligned, and again there are

[6] To be specific, this implies the following requirement: If a type λ trader sells the asset to a type λ' trader when both are in state h then it must be that they trade in the opposite direction when both are in state l.

equal shares of the two misaligned states for each λ.[7]

It is mathematically convenient to refer to traders only by their alignment status, where $a = 0$ indicates misaligned and $a = 1$ indicates well-aligned. Let $\mathbf{1}_{\lambda,a}^{\lambda',a'}$ indicate the trading probability between traders (λ, a) and (λ', a') conditional on them having the opposite asset holdings. Let $m_\lambda \equiv \mu_{\lambda,l,1} + \mu_{\lambda,h,0}$ denote the fraction of traders with contact rate λ who are misaligned. Equation (2) reduces to

$$\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,0}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,0}^{\lambda',1}(1 - m_{\lambda'})\right)\right) m_\lambda \\ = \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,1}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,1}^{\lambda',1}(1 - m_{\lambda'})\right)\right)(1 - m_\lambda). \quad (4)$$

The left hand side is the outflow rate from the misaligned states. This occurs either following a preference shock or a meeting with a trader who has the opposite asset holdings where trade occurs. The right hand side is the inflow rate, again following the same events.

Let the average value of a misaligned and well-aligned trader be denoted by $v_{\lambda,0} \equiv \frac{1}{2}(v_{\lambda,l,1} + v_{\lambda,h,0})$ and $v_{\lambda,1} \equiv \frac{1}{2}(v_{\lambda,l,0} + v_{\lambda,h,1})$, respectively. Also define $s_\lambda \equiv v_{\lambda,1} - v_{\lambda,0}$, the surplus from being well-aligned rather than misaligned. Knowing the surplus function is sufficient to tell whether trade occurs. To see this, note that the net value of alignment when in state h, $v_{\lambda,h,1} - v_{\lambda,h,0}$ equals the net value of alignment when in state l, $v_{\lambda,l,0} - v_{\lambda,l,1}$ up to a constant common to all types λ. It then follows from condition (3) that the surplus function s_λ governs the patterns of trade since we have that

$$\begin{aligned} v_{\lambda,h,1} - v_{\lambda,h,0} + v_{\lambda',l,0} - v_{\lambda',l,1} &= s_\lambda + s_{\lambda'} \\ v_{\lambda,l,1} - v_{\lambda,l,0} + v_{\lambda',l,0} - v_{\lambda',l,1} &= -s_\lambda + s_{\lambda'} \\ v_{\lambda,h,1} - v_{\lambda,h,0} + v_{\lambda',h,0} - v_{\lambda',h,1} &= s_\lambda - s_{\lambda'} \\ v_{\lambda,l,1} - v_{\lambda,l,0} + v_{\lambda',h,0} - v_{\lambda',h,1} &= -s_\lambda - s_{\lambda'}. \end{aligned}$$

Taking advantage of symmetry in the misalignment rates and the Nash bargaining solution, equation (1) reduces to

$$rv_{\lambda,0} = \delta_0 + \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left((s_\lambda + s_{\lambda'})^+ m_{\lambda'} + (s_\lambda - s_{\lambda'})^+(1 - m_{\lambda'})\right) - \lambda c(\lambda) \quad (5)$$

$$\text{and } rv_{\lambda,1} = \delta_1 - \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left((-s_\lambda + s_{\lambda'})^+ m_{\lambda'} + (-s_\lambda - s_{\lambda'})^+(1 - m_{\lambda'})\right) - \lambda c(\lambda), \quad (6)$$

[7] We assume that if a positive measure of traders choose to live in autarky, $\lambda = 0$, half of them are initially endowed with the asset and half are not. Thus they do not affect the share of traders with $\lambda > 0$ who hold the asset.

where $\delta_0 \equiv \frac{1}{2}(\delta_{l,1} + \delta_{h,0})$ and $\delta_1 \equiv \frac{1}{2}(\delta_{l,0} + \delta_{h,1}) = \Delta + \delta_0$. Again, both equations reflect the sum of four terms. The first is the average flow payoff of a misaligned or well-aligned trader. The second is the gain or loss from a preference shock that switches the alignment status. The third is the gain from meetings, reflecting that only half of all meetings are with traders who hold the opposite asset; and in these events each trader walks away with half of the joint surplus, if there is any. The $+$-superscript is shorthand notation for the $\max\{\cdot, 0\}$ and reflects that meetings result in trade if and only if doing so is bilaterally efficient. The final term is the search cost.

Finally, we can simplify equation (4) using the Nash bargaining solution as well, since trades occur if and only if doing so is bilaterally efficient:

$$\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{s_\lambda + s_{\lambda'} > 0} m_{\lambda'} + \mathbb{I}_{s_\lambda > s_{\lambda'}}(1 - m_{\lambda'})\right)\right) m_\lambda$$
$$= \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{s_\lambda < s_{\lambda'}} m_{\lambda'} + \mathbb{I}_{s_\lambda + s_{\lambda'} < 0}(1 - m_{\lambda'})\right)\right)(1 - m_\lambda). \quad (7)$$

Here the indicator function $\mathbb{I}$ is equal to 1 if the inequality in the subscript holds and is zero otherwise.

3.3 Endogenizing the Distribution of Contact Rates

At an initial date 0, all traders choose their contact rate λ in order to maximize their value. If traders are impatient, that means that their choice will depend on their alignment status at date 0. This would make it necessary to solve for transitional dynamics from this initial condition. We circumvent this issue by focusing on the no-discounting limit of this economy, $r \to 0$. The surplus $s_\lambda = v_{\lambda,1} - v_{\lambda,0}$ is finite in this limit, while the present value of the gain from switching alignment status, $r(v_{\lambda,1} - v_{\lambda,0})$ converges to zero. It follows that the trader's initial asset holdings does not affect their incentive to invest and we may ignore the transitional dynamics.

The focus on the no-discounting limit reflects our expectation that the short-run desire to trade is not an important determinant of the irreversible investment in meeting technologies. We think of the preference shifts as occurring at a much higher frequency than discounting, while trading opportunities may occur at a higher frequency still. This implies that the importance of holding the asset at the correct time, Δ, is likely to be a much more important determinant of this investment.

3.4 Definition of Equilibrium

We define a steady state equilibrium in the limiting economy with $r \to 0$. The definition relies only on objects that are well-behaved in this limit.

Definition 1 *A steady state equilibrium is a distribution of contact rates $G(\lambda)$, an average contact rate Λ, an allocation of misalignment m_λ, and undiscounted surplus function s_λ, satisfying the following conditions:*

1. *Balanced inflows and outflows into misalignment as given by equation (7);*
2. *Consistency of s_λ with the value functions (5) and (6),*

$$\Delta = 2\gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\Big(\big((s_\lambda + s_{\lambda'})^+ - (s_{\lambda'} - s_\lambda)^+\big)m_{\lambda'} + \big((s_\lambda - s_{\lambda'})^+ - (-s_\lambda - s_{\lambda'})^+\big)(1 - m_{\lambda'})\Big); \quad (8)$$

3. *Optimality of the ex-ante investment decision:*
   (a) *$dG(\lambda) > 0$ only if it maximizes*

$$\delta_1 - \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left(\big(-s_\lambda + s_{\lambda'}\big)^+ m_{\lambda'} + \big(-s_\lambda - s_{\lambda'}\big)^+(1 - m_{\lambda'})\right) - \lambda c(\lambda);$$

   (b) *Middlemen make finite profits: $\Lambda \geq \int_0^\infty \lambda dG(\lambda)$ and*

$$\lim_{\lambda\to\infty}\left(\frac{1}{4}\mathbb{E}\left((-s_\lambda + s_{\lambda'})^+ m_{\lambda'} + (-s_\lambda - s_{\lambda'})^+(1 - m_{\lambda'})\right) - c(\lambda)\right) \leq 0,$$

   with complementary slackness.

We have already explained the first two conditions. Condition 3(a) ensures that if traders choose a contact rate λ, it maximizes their average payoff $\lim_{r\to 0} rv_{\lambda,1} = \lim_{r\to 0} rv_{\lambda,0}$. Condition 3(b) ensures that if there are middlemen, they earn zero profits in each meeting. Middlemen earn profit by taking half the surplus from meetings where they change the alignment status of their trading partners. If middlemen earned more profit in an average meeting than the cost of a meeting, being a middleman would be arbitrarily profitable, inconsistent with equilibrium. If they earned less, there would be no middlemen, which implies $\Lambda = \int_0^\infty \lambda' dG(\lambda')$.

3.5 Equilibrium Trading Patterns

We now begin our characterization of the equilibrium, starting with trading patterns given any distribution $G(\lambda)$.

Proposition 1 *In equilibrium, when two traders with opposite asset positions meet they*

1. *always trade the asset if both are misaligned;*
2. *never trade the asset if both are well-aligned;*
3. *trade the asset if one is misaligned and the other is well-aligned and the well-aligned trader has the higher contact rate.*

The Appendix contains proofs of all our propositions. The proof shows that the surplus function s_λ is non-negative and decreasing. The result then follows immediately.

The first two parts of the proposition reflect fundamentals. Trade between two misaligned traders turns both into well-aligned traders, thus creating gains in a direct, static fashion. Trade between two well-aligned traders turns both misaligned and never happens for the same static reason. The third part of the proposition reflects option value considerations and is the key feature of the endogenous trading pattern which arises in this environment, namely intermediation. It states that a faster trader buys the asset from a slower trader if both are in preference state l; and she sells the asset to the slower trader if both are in preference state h. These trades do not immediately increase the number of well-aligned traders, but they move misalignment towards traders who expect more future trading opportunities. These trades occur in equilibrium because traders with low contact rates are able to compensate traders with high contact rates for taking the misaligned positions.

The possibility of intermediation implies that a trader's buying and selling decisions become increasingly detached from her idiosyncratic preferences as her contact rate increases. In other words, a high contact rate moderates the impact of the idiosyncratic taste component on a trader's valuation of the asset. It follows that those who become intermediaries, in the center of the valuation chain, are traders with frequent meetings. Figure 1 shows the intermediation chain which follows from Proposition 1. Slow traders are at the periphery of the trading chain, not trading once their asset position is aligned with their preferences. In turn, the fast traders constitute the endogenous core of the trading network, buying and selling largely irrespective of their preference state. In doing so, they take on misaligned asset positions from types with lower contact rates and are compensated through the bid-ask spread. This also implies that faster traders not only meet other traders more frequently but also trade more frequently conditional on a meeting because they take on the misalignment from traders with lower search efficiency.

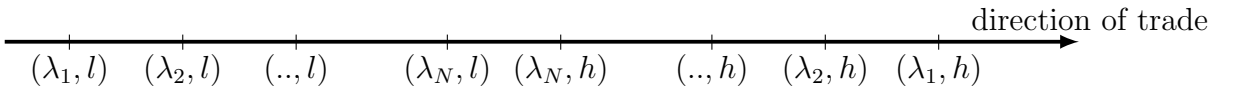

Figure 1: Direction of trade across traders with contact rate $\lambda \in \{\lambda_1, .., \lambda_N\}$ with $\lambda_1 < \lambda_2 < .. < \lambda_N$ and current taste $i \in \{l, h\}$.

3.6 Distribution of Contact and Misalignment Rates

We next show that a non-degenerate distribution $G(\lambda)$, that is the coexistence of traders with different λ, arises in equilibrium even when market participants are ex-ante homogeneous. To do so, we first solve explicitly for the surplus function, taking advantage of the trading patterns determined in Proposition 1. We prove in the Appendix that the surplus function satisfies

$$s_\lambda = \frac{\Delta}{2\gamma}\left(1 - e^{-\int_\lambda^\infty \phi_{\lambda'} d\lambda'}\right). \tag{9}$$

where

$$\phi_\lambda \equiv \frac{8\gamma}{\lambda\Big(8\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)} \tag{10}$$

and $H(\lambda) \equiv \mathbb{E}(\mathbb{I}_{\lambda'>\lambda})$ denotes the fraction of meetings with a trader who has a higher contact rate and $L(\lambda) \equiv \mathbb{E}(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'})$ denotes the fraction of meetings with a misaligned trader who has a lower contact rate. In addition, part 3(a) of the definition of equilibrium implies that $G(\lambda)$ is increasing at λ only if λ maximizes

$$\delta_1 - \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\Big(\mathbb{I}_{\lambda'<\lambda}\big(s_{\lambda'} - s_\lambda\big)m_{\lambda'}\Big) - \lambda c(\lambda). \tag{11}$$

We use this to prove the following result:

Proposition 2 *Assume $c(\lambda)$ is continuously differentiable. Then the equilibrium distribution of search efficiency $G(\lambda)$ has no mass points, except possibly at $\lambda = 0$.*

This proposition implies that although all traders are ex-ante identical, there is no symmetric equilibrium in which all traders choose identical actions. Even stronger, almost all traders choose different types.[8] The proof shows that gross flow profits have a convex kink at any mass point. Given a differentiable cost function we hence conclude that mass points are inconsistent with optimality of the ex-ante investment decision.

[8]The one caveat is that a positive fraction of traders may choose to live in autarky, setting $\lambda = 0$. If this is optimal, all traders must get the same payoff, and so the value of participating in this market must be zero. This is the case when the cost function c is too high.

To develop an understanding for the result consider an environment where everyone has contact rate $\bar{\lambda}$. This turns out to create a convex kink in the value function at $\bar{\lambda}$. To understand why, consider the marginal impact of an increase in the contact rate for a trader with a contact rate $\bar{\lambda}$. This allows the trader to act as an intermediary for all the other traders. Although the gains from intermediation are small, on the order of the difference between the contact rates, the opportunities to intermediate are frequent, whenever she meets a misaligned trader with the opposite asset holdings. Thus a higher contact rate creates a first order gain. Conversely, consider the marginal impact of a decrease in the contact rate for a trader with contact rate $\bar{\lambda}$. This allows all other traders to intermediate for her, dramatically reducing her misalignment probability. Of course, this doesn't come for free; she pays for these trades using the outside good. Nevertheless, the benefits from the trades are again linear in the difference in contact rates. Thus a lower contact rate also creates a first order gain. This creates a convex kink in the value function at the mass point $\bar{\lambda}$. This logic carries over to any mass point.[9]

A different way to see this is in terms of the number of trades. Setting the same contact rate as everyone else, trades only occur with both traders are misaligned and holding the opposite asset position. Choosing a slightly different contact rate allows for gains from a host of other trades, where one party is misaligned and the other well-aligned. In other words, the nature and frequency of trades depends starkly on the contact rate compared with other market participants. One gets intermediated by faster traders and intermediates slower traders. As soon as a positive measure of traders has the same contact rate, this discretely changes the marginal returns to λ and is hence inconsistent with equilibrium under a differentiable cost function.

The absence of a pure strategy equilibrium is a common feature of search models (Butters, 1977; Burdett and Judd, 1983; Burdett and Mortensen, 1998; Duffie, Dworczak and Zhu, 2016). These papers have in common that if all firms charge the same price (or offer the same wage), firms that offering a slightly lower price (higher wage) earn discontinuously higher profits. Our results concern a different object, the contact rate, and we find that the profit function is continuous but not differentiable. We therefore believe that our finding is distinct from those in the existing literature.

Whenever the contact rate distribution has no mass points, we can use the equilibrium trading pattern to simplify the inflow-outflow equality (7):

$$\big(\gamma + \tfrac{1}{2}\lambda(H(\lambda) + L(\lambda))\big)m_\lambda = \big(\gamma + \tfrac{1}{2}\lambda L(\lambda)\big)(1 - m_\lambda). \tag{12}$$

[9]Section 5 studies an environment where the only admissible trades are between asset holders in state l and non-holders in state h. We show that the equilibrium distribution of contact rates then collapses to a single mass point. It follows that heterogeneity fundamentally arises to leverage the gains from intermediation.

Misaligned traders become well-aligned when they experience a preference shock (rate γ), meet a faster trader with the opposite asset position (rate $\frac{1}{2}\lambda H(\lambda)$), or meet a slower misaligned trader with the opposite asset position (rate $\frac{1}{2}\lambda L(\lambda)$). Well-aligned traders become misaligned following the first or third events.

Proposition 2 shows that there is dispersion in equilibrium contact rates under a relatively weak condition. This naturally leads us to ask how much dispersion. If the cost function is weakly convex, we find that dispersion is not too extreme, in the sense that the support of the contact rate distribution is convex:

Proposition 3 *Assume $\lambda c(\lambda)$ is weakly convex. Then the equilibrium distribution of search efficiency $G(\lambda)$ has a convex support. Moreover, if there are middlemen ($\Lambda > \int_0^\infty \lambda dG(\lambda)$), the support of $G(\lambda)$ is unbounded above.*

The proof shows that if there is a "hole" in the support of $G(\lambda)$, traders' value function must be strictly concave on the hole. This is inconsistent with both extreme points of the hole maximizing the traders' value. The open tail in the presence of middlemen reflects the same argument. If there were an upper bound in the support of G, any individual above the bound would conduct the same types of trades as those at the upper bound. This would just linearly scale her revenues from trade with a misalignment rate identical to those at the upper bound. If, given the weakly convex cost function, no individual finds it optimal to do so, then it cannot be optimal to acquire an infinite contact rate either.

Proposition 3 rules out the possibility that most traders choose a low contact rate, while a few traders choose a very high contact rate acting as intermediaries. This would be the case in a star network. This is a consequence of the complementarity in the matching technology. If a trader chooses a very low contact rate, she only meets intermediaries infrequently, and so cannot take advantage of their intermediation services.

We next turn to the connection between contact rates and misalignment rates. A higher contact rate has two opposing effects on an individual's misalignment rate. First, an individual is more frequently able to offset a misaligned position. However, a trader with a higher contact rate also intermediates more frequently, taking on misalignment from slower traders. Proposition 4 states that the latter force dominates everywhere on the support of $G(\lambda)$ if the cost function is weakly convex:

Proposition 4 *Assume $\lambda c(\lambda)$ is weakly convex and continuously differentiable. Then the equilibrium misalignment rate m_λ is strictly increasing on the support of $G(\lambda)$.*

We stress that Proposition 4 imposes that $G(\lambda)$ is the equilibrium contact rate distribution.

The result is not true for an arbitrary distribution of contact rates.[10] A faster contact rate deteriorates a trader's allocation, because she is more likely to serve as an intermediary. The reason some traders invest in a faster contact rate then must come from trading profits, the returns to frequently buying and selling the asset with a favorable bid-ask spread.

A corollary of Proposition 4 considers an extension to our model where traders differ ex ante in how much they care about having a well-aligned asset position, Δ. Proposition 4 suggests that faster traders at the core of the network will naturally be those with smaller Δ, while slower traders at the periphery will be those with larger Δ. This is the opposite of what one would expect to see without intermediation.

Finally, we point to a related observation: The proof of Proposition 6 below shows that the trading probability conditional on a meeting is strictly increasing in λ. This is likewise a consequence of the fact that faster traders trade with slower traders whenever the slower trader is misaligned. When the slowest trader in the economy is well aligned, she never trades, but the fastest trader trades regardless of her alignment status.

3.7 Linear Cost Function: Analytical Results

In this section, we restrict the cost function to be linear, so the cost per meeting is constant, $c(\lambda) = c$. We start by showing how to express the equilibrium conditions in a pair of equations. Let $\underline{\lambda}$ denote the lowest contact rate in the support of G. Then following the arguments in the proof of Proposition 4, we can reduce the model to a pair of first order ordinary differential equations in H and L:

$$4\gamma(2H(\lambda) + \lambda H'(\lambda)) = \lambda^2\big(H(\lambda)L'(\lambda) - L(\lambda)H'(\lambda)\big), \tag{13a}$$

$$\big(4\gamma + \lambda(H(\lambda) + 2L(\lambda))\big)L'(\lambda) = -(2\gamma + \lambda L(\lambda))H'(\lambda), \tag{13b}$$

with $H(\underline{\lambda}) = 1$ and $L(\underline{\lambda}) = 0$. The first equation represents the optimality condition that all traders earn the same profit in equilibrium, derived in the appendix (equation 32 with $c(\lambda) = c$), while the second represents the steady state misalignment rate condition (equation 12). In this section, we manipulate these equations to partially characterize the equilibrium analytically. Our main result is the following:

Proposition 5 *Assume $c(\lambda) = c$. If $c < \Delta/16\gamma$, the equilibrium distribution of contact rates $G(\lambda)$ has a strictly positive lower bound $\underline{\lambda}$, has a Pareto tail with tail parameter 2, and has a*

[10] To see this, consider a distribution with a "hole" in the support. $m(\lambda)$ would be strictly decreasing over such an interval with a zero measure of traders for two reasons. First, the types of trades an individual would engage in would be identical anywhere on the interval. Second, all individuals, except possibly the fastest, trade more frequently into alignment then out of it, that is $m(\lambda) < \frac{1}{2}$. It follows that just scaling up the contact rate without altering the types of trade reduces the misalignment rate.

zero measure of middlemen who are in continuous contact with the market, accounting for a strictly positive fraction of meetings, $\Lambda > \int_0^\infty \lambda dG(\lambda)$. *If* $c \geq \Delta/16\gamma$, $\Lambda = 0$ *in equilibrium.*

If meetings are too costly the market place collapses into autarky. In turn, if there is trade then the equilibrium $G(\lambda)$ has rich features which we describe in turn.

The strictly positive lower bound $\underline{\lambda}$ reflects that the value of a trader smoothly converges towards its autarky level as $\lambda \to 0$. If the cost are strictly below the minimum level which leads to autarky traders, fare strictly better than under autarky, which allows us to rule out contact rates close to zero.

To develop an understanding for the Pareto tail, note that we have already established that the distribution has no mass points and an open right tail in the presence of middlemen. It thus follows that gross flow values (ignoring the linear cost $c\lambda$) must be linear above the lowest contact rate $\underline{\lambda}$. A Pareto distribution with tail parameter 2 implies that increasing a trader's contact rate leaves the frequency at which she meets a faster trading partner unchanged. That is, the reduction in the fraction of individuals who are faster is exactly offset by the increase in the contact rate. Furthermore, the relative contact rate conditional on meeting a faster individual is also independent of λ. On the other hand, increasing λ linearly scales the frequency at which she meets a slower trader; while the partner's expected contact rate converges to a constant, namely just the average contact rate among the finitely fast traders (once λ is in the tail). Finally, increasing λ linearly increases meetings with a middleman. Thus, the Pareto tail parameter 2 guarantees that as λ increases in the tail, it linearly increases the rate of contacting slower traders and middlemen while leaving trading opportunities with faster traders unchanged; jointly, these features deliver linear gross flow values in the tail. The endogenous Pareto tail seems unrelated to well-known mechanisms that give rise to Pareto distributions in various contexts (Gabaix, 2016).

To understand the emergence of middlemen, we first note that an individual trader's value can be decomposed into pure trading profits and the returns from having a well-aligned asset position. Furthermore, as λ becomes increasingly large, the types of trades a trader conducts become increasingly independent of λ, since in most trades she acts as an intermediary. This first implies that the misalignment rate converges to a constant, and second that the total trading profits are linear in λ. It follows that, for fast traders with different contact rates to be equally well off, the net trading profits have to converge to zero as λ grows large. From this observation it also follows that whenever traders in equilibrium are better off than under autarky it has to be that the misalignment rate of fast types falls strictly below its autarky value of $\frac{1}{2}$. This is what middlemen guarantee since they allow even very fast types to offload misaligned asset positions.In summary, an open right tail requires net trading profits to converge to zero; middlemen then allow fast intermediaries to

offload misalignment, restoring the incentives for intermediation.

A large body of empirical work documents that the degree distribution of various financial markets is often well described by a power law (see footnote 2). However, what is mapped out empirically is the distribution of trading rates, the product of the contact rate λ and the probability of trading in a meeting, p_λ. The next proposition then connects the results describing the distribution of contact rates to the distribution of trading rates, allowing us to directly relate to the empirical literature.

Proposition 6 *Assume $c(\lambda) = c < \Delta/16\gamma$. The equilibrium distribution of trading rates inherits the tail properties of the contact rate distribution, i.e. it has a Pareto tail with tail parameter 2 and a zero measure of middlemen account for a strictly positive fraction of trades.*

Intuitively, the trading rate inherits the Pareto tail of the contact rate distribution, since the trading probability conditional on a meeting converges to a positive constant in the tail. This also ensures that middlemen account for a positive fraction of trades.

Thus, with linear cost, our setup gives rise to a distribution of trading rates that looks like its empirical counterpart. We highlight that the empirical literature frequently finds power law coefficients close to 2. For instance, Gabaix, Gopikrishnan, Plerou and Stanley (2006) report a tail parameter of 2.5 for the distribution of trading volume in the stock market. While the exact Pareto result holds only in the tail, our numerical results in Section 3.9 show that the entire distribution closely resembles a Pareto with tail parameter 2.

3.8 Linear Cost Function: Frictionless Limit

For many real world markets, frictions are small and so a natural question is whether intermediation retains its prominent role in the frictionless limit and whether we obtain additional insights from studying frictions in markets where frictions are small. This section therefore focuses on the model with a linear cost when the cost of a meeting is negligible, $c \to 0$. In this limit, the lower bound on contact rates converges to infinity, $\underline{\lambda} \to \infty$, so in some sense everyone can trade instantaneously. Still, this hides important heterogeneity in contact rates in the limiting economy. To emphasize this point, in an economy with $0 < c < \Delta/16\gamma$ and hence $\underline{\lambda} < \infty$, we define a trader's relative contact rate as $\rho \equiv \lambda/\underline{\lambda}$ and call a trader "finite" whenever $\rho < \infty$ and a middleman if $\rho = \infty$. We find that the distribution of ρ and the fraction of meetings that are with middlemen have well-behaved limits as $c \to 0$ and hence $\underline{\lambda} \to \infty$. This limit then also allows us to obtain a sharp characterization of volume, the rate

at which assets are traded:[11]

Proposition 7 *Assume $c(\lambda) = c$. In the limit of equilibrium allocations as $c \to 0$, a strictly positive fraction of meetings are with middlemen and a strictly positive fraction are with finite traders. Volume lies between 2γ and 2.5γ and can be decomposed as follows: finite traders' purchases from middlemen account for a volume of $\frac{1}{2}\gamma$; middlemens' purchases from middlemen account for a volume of $\frac{1}{2}\gamma$; middlemens' purchases from finite traders account for a volume of $\frac{1}{2}\gamma$; and finite traders' purchases from finite traders account for the remaining volume, which lies between $\frac{1}{2}\gamma$ and γ.*

The proof of Proposition 7 provides exact expressions for the fraction of meetings with middlemen (approximately 26.9 percent), volume (approximately 2.46γ, of which 40.7 percent is accounted for by middlemens' purchases), and the ratio between the average contact rate Λ and the lower bound on contact rates $\underline{\lambda}$ (both of which converge to infinity in the limit). It also provides an implicit equation for the distribution of the relative contact rate ρ.

We contrast Proposition 7 with a naïve view of a market without frictions: all traders can trade instantaneously upon receiving a preference shock and only trade with other traders who receive the opposite preference shock at the same instant. That means that volume equals the share of traders in the low preference state times the rate at which they are hit by preference shocks, $\frac{1}{2}\gamma$. Note that this view leaves no role for intermediation or middlemen. In contrast, we obtain more than four times as much trading volume. Furthermore, the proposition highlights that a meaningful role for heterogeneity in contact rates and intermediation by both finite traders and middlemen pertains to the limiting economy.

To understand this result note that we are looking at a frictionless limit where almost no one is misaligned. Whenever a trader (who is almost surely a finite trader) suffers a preference shock, she is very likely to be well-aligned and very unlikely to contact another misaligned trader, i.e. $\lim_{\lambda\to\infty} L(\lambda) \to 0$. Instead, the market passes the asset towards faster traders whenever possible. Since the faster trader is still very unlikely to be misaligned, this trade does not reduce misalignment, but simply moves it towards the core traders. This piece of the process stops once the misalignment is passed to a middleman. What the volume decomposition shows is that, in the frictionless limit, the reallocation of the asset in response to taste shocks runs through an intermediation chain that always involves middlemen. Every time there is a preference shock, there is a trade between a middleman and a finite trader. Half these trades are asset purchases and half are asset sales.

When middlemen take on the asset from a finite trader they, too, move into misalignment. Afterwards, they quickly trade away from misalignment, but only by meeting another

[11] We make the natural assumption that two middlemen trade if and only if both are misaligned and have opposite asset holdings.

middleman. As a consequence trade from middlemen to middlemen accounts for a volume of $\frac{1}{2}\gamma$. The reason is that, when c is small, the misalignment rate of finite traders is proportional to the square of the misalignment rate of middlemen. Thus, as misalignment converges to zero, traders are far more likely to contact misaligned middlemen than misaligned finite traders. Taken together, whenever a trader experiences a taste shock the market rapidly reallocates her asset position to a misaligned finite type with opposite preferences. But instead of doing so directly, the position gets traded through an intermediation chain. This chain runs through faster finite types towards middlemen who then first reallocate the position internally before passing it back to misaligned finite types with opposite preferences.

3.9 Linear Cost Function: Numerical Results

This section focuses on numerical results when the cost function is linear, $c(\lambda) = c$ for all λ. Throughout we normalize $\gamma = 1$. We first offer a numerical result that strongly suggest existence and uniqueness of the equilibrium allocation. The sole parameter governing the equilibrium allocation is the cost of contacts relative to the gains from alignment, c/Δ. For a given value of $\underline{\lambda} > 0$, we solve equation (13) to find the functions $H(\lambda)$ and $L(\lambda)$, which completely characterize the equilibrium. We then back out the implied ratio c/Δ from the requirement that any value of λ in the support of the distribution G, and in particular $\underline{\lambda}$, is a profit maximizing choice. Inverting this gives us the set of equilibrium allocations corresponding to each value of c/Δ.

For $c/\Delta \geq 1/16$, Proposition 5 proves that autarky is the unique equilibrium. Otherwise, the red line in the top panel in Figure 2 suggests that each value of $c/\Delta \in (0, 1/16)$ corresponds to a unique value of $\underline{\lambda}$.[12] When costs are close to $1/16$, the lower bound is very small and it grows without bound as costs converge to 0. Since the rest of the equilibrium allocation is determined from $\underline{\lambda}$ using equation (13), this implies that each value of $c/\Delta \in (0, 1/16)$ corresponds to a unique equilibrium allocation. If this mapping is indeed one-to-one, this would prove existence and uniqueness of equilibrium. An analytical proof of this claim eludes us.

The middle panel in the figure plots the average contact rate Λ relative to the lower bound $\underline{\lambda}$ as a function of the relative cost c/Δ.[13] The third panel in Figure 2 plots the fraction of meetings that are with middlemen, $1 - \int_0^\infty \lambda dG(\lambda)/\Lambda$. This too is strictly decreasing in the cost, converging to 0 when relative costs converge to $1/16$ and approximately 0.27 when costs converge to zero. Finally, the fourth panel in Figure 2 plots volume per unit of time. This rises from 0 when relative costs converge to $1/16$, exceeds $\frac{1}{2}$ (the frictionless benchmark)

[12] The blue lines in Figures 2 and 3 refer to the optimal allocation, which we define and analyze in Section 4.

[13] Exact expressions for all limiting expressions in this paragraph are given in the proof of Proposition 7.

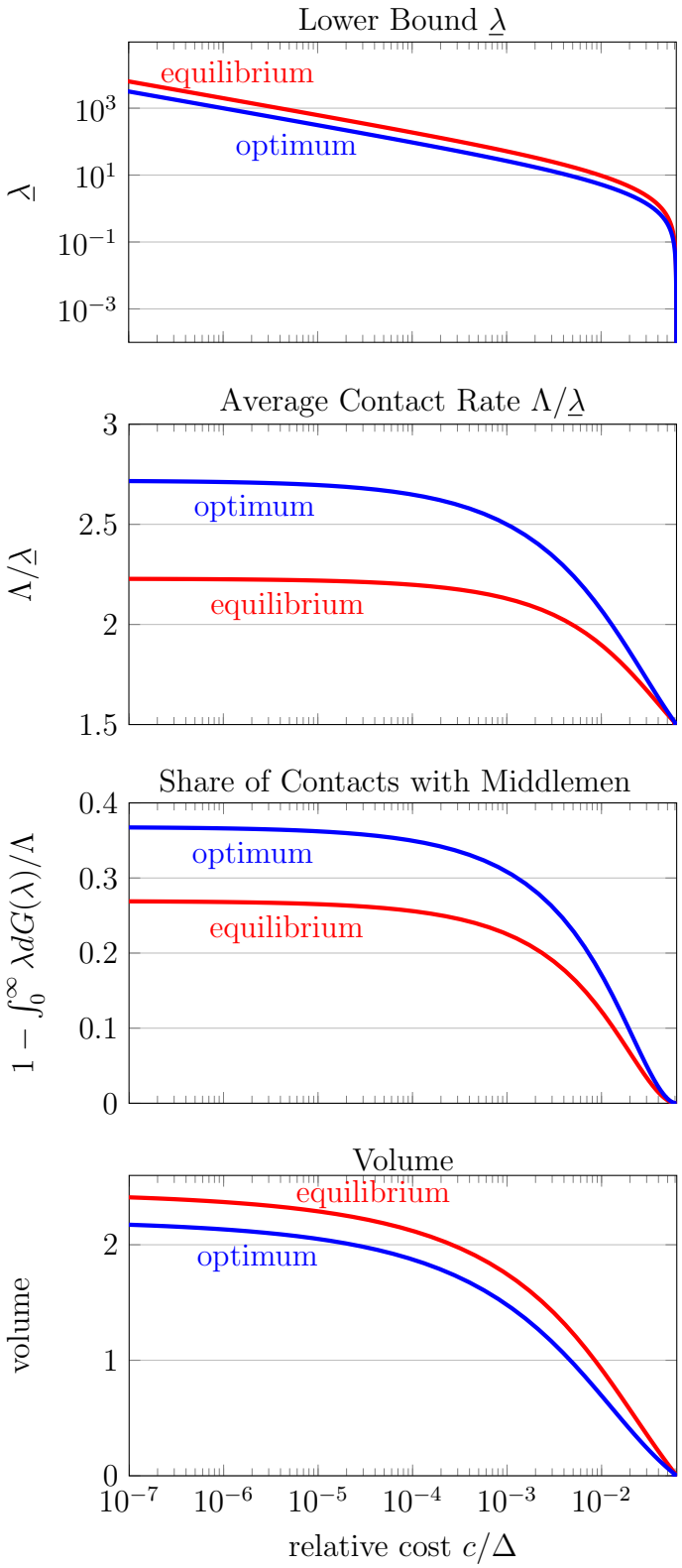

Figure 2: Features of the equilibrium and optimal contact rate distribution given relative costs c/Δ. We normalize $\gamma = 1$.

when $c/\Delta < 0.02$, and converges to 2.46 in the limit as costs converge to zero.

Figure 3 hones in on the equilibrium allocation for a particular value of the relative cost c/Δ. The top panel plots the distribution of contact rates. The survivor function $1 - G(\lambda)$ is nearly linear in log-log space and has slope -2. That is, it is close to an exact Pareto distribution with tail parameter 2, although we can prove that it is not an exact Pareto distribution. In line with proposition 6, the middle panel shows that the same holds for the equilibrium distribution of trading rates, the subject of much attention in the empirical literature as discussed above. That is, the figure suggests that the entire distribution of trading rates is well-approximated by a power law. Moreover, most traders have a trading rate that exceeds γ, which means that intermediation is important for understanding the vast majority of trades in this marketplace. Finally, the bottom panel shows the gradual increase in the misalignment rate experienced by traders with higher contact rates, a corollary of Proposition 4.

4 Optimal Allocation

This section examines which trading patterns and contact rate distributions maximizes average utility. We imagine a hypothetical social planner who can instruct traders both on their choice of λ at the initial date and on whether to trade in each future meeting, but who cannot directly alleviate the search frictions in the economy. Since all traders are identical at time 0, the solution to the social planner's problem gives us the Pareto optimal allocation.

4.1 Planner Problem

We work directly with the undiscounted problem where the planner wishes to maximize steady state average utility.[14] We also impose that the planner must use a symmetric trading pattern, as in the equilibrium. Thus we simply need to keep track of the number of misaligned traders at each contact rate λ.

The planner's objective is to maximize

$$\delta_0 + \Delta \int_0^\infty (1 - m_\lambda) dG(\lambda) - \Lambda \mathbb{E}(c(\lambda)) \tag{14}$$

Each misaligned trader gets a flow payoff of δ_0, while each well-aligned trader gets a flow payoff of $\delta_1 = \Delta + \delta_0$, expressed in the first two terms. In addition, the planner must pay the

[14] It is also possible to write down the discounted planner's problem, take the limit as the discount rate converges to zero, and focus on steady states. The results are the same.

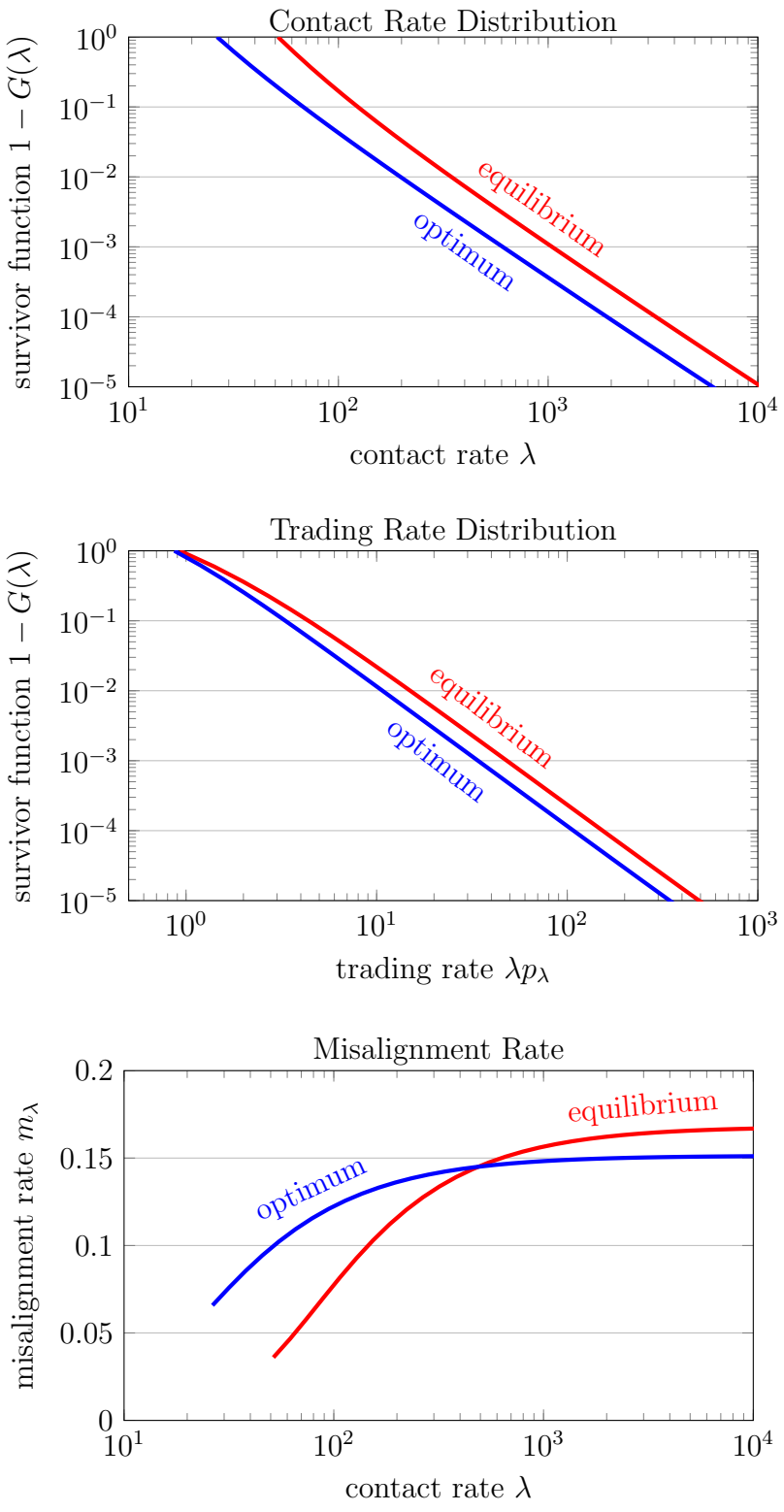

Figure 3: Equilibrium and optimal contact rate distribution, trading rate distribution, and misalignment rates, $c/\Delta = 0.001$ and $\gamma = 1$.

search costs, equal to the product of the average search intensity Λ and the cost per meeting in the average meeting.

The planner has two instruments. The first is that she chooses the set of admissible trades. When a trader with contact rate λ and alignment status $a \in \{0,1\}$ meets a trader with contact rate λ' and alignment status $a' \in \{0,1\}$ and the opposite asset position, they trade with probability $\mathbf{1}_{\lambda,a}^{\lambda',a'} = \mathbf{1}_{\lambda',a'}^{\lambda,a}$. This implies that the steady state misalignment rate satisfies

$$\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,0}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,0}^{\lambda',1}(1-m_{\lambda'})\right)\right) m_\lambda = \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,1}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,1}^{\lambda',1}(1-m_{\lambda'})\right)\right)(1-m_\lambda). \quad (15)$$

Second, the planner chooses the distribution of contact rates $G(\lambda)$ and the average contact rate Λ. We allow the planner to set $\Lambda > \int_0^\infty \lambda' dG(\lambda')$, i.e. to use middlemen.

The optimal allocation is a contact rate distribution $G(\lambda)$ and Λ and trade indicator functions $\mathbf{1}_{\lambda,a}^{\lambda',a'} \in [0,1]$ that maximize (14) subject to (15). Our main result is that the equilibrium and optimum allocations are qualitatively the same:

Proposition 8 *Propositions 1–7 hold with the word "equilibrium" replaced by "optimum."*

The proof proceeds by solving the planner's constrained optimization problem. The first order conditions yield an expression for the social net value of asset ownership, S_λ, which is the Lagrange multiplier on the misalignment constraint (15). We then show that the planner requires trade whenever the joint social surplus of the transaction is positive. The social surplus function S_λ is closely related to the private surplus s_λ, and in particular is decreasing and nonnegative:

$$S_\lambda = \frac{\Delta}{2\gamma}\left(1 - e^{-\int_\lambda^\infty \Phi_{\lambda'} d\lambda'}\right), \quad (16)$$

where

$$\Phi_\lambda \equiv \frac{4\gamma}{\lambda\Big(4\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)}. \quad (17)$$

This is scarcely changed from equations (9) and (10) in the decentralized equilibrium. Moreover, the planner sets $dG(\lambda) > 0$ only if λ maximizes

$$-\gamma S_\lambda + \frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda}\big(S_{\lambda'} - S_\lambda\big)m_{\lambda'}\right) - \lambda\left(c(\lambda) + \frac{\gamma}{\Lambda}\int_0^\infty S_{\lambda'}(1-2m_{\lambda'})dG(\lambda')\right) \quad (18)$$

and sets $\Lambda \geq \int_0^\infty \lambda dG(\lambda)$ and

$$\lim_{\lambda \to \infty} \left(\frac{1}{2} \mathbb{E} \left((S_{\lambda'} - S_\lambda) m_{\lambda'} \right) - c(\lambda) \right) \leq \frac{\gamma}{\Lambda} \int_0^\infty S_{\lambda'} (1 - 2m_{\lambda'}) dG(\lambda'), \tag{19}$$

with complementary slackness. This is analogous to part 3 of the definition of equilibrium.

When the cost of contacts is constant, $c(\lambda) = c$, the appendix also characterizes the planner's problem as the first order differential equation in H and L, analogous to equation (13a) governing equilibrium:

$$2\gamma \left(2H(\lambda) + \lambda H'(\lambda) \right) = \lambda^2 \left(H(\lambda) L'(\lambda) - L(\lambda) H'(\lambda) \right). \tag{20}$$

Equation (13b) is unchanged. We use this pair of equations for the planner's analog of Propositions 5, 6, and 7.

The proposition first implies that the equilibrium trading pattern is optimal. The intuition for the result is straightforward. The planner's objective function boils down to minimizing the average rate of misalignment. The planner therefore demands trade if it reduces static misalignment and rejects it if it raises static misalignment. In the case where only one trader is misaligned, the planner moves the misalignment towards the trader with more future trading opportunities. That is, the planner uses faster traders as intermediaries. This does not affect the current misalignment rate, but improves future trading possibilities. It follows that, given a non-degenerate distribution of contact rates $G(\lambda)$, the intermediated trading pattern governing equilibrium is optimal. This implies that the allocation would be strictly inferior if traders could only engage in fundamental trades.

The atomless feature of the optimal distribution allows the planner to leverage the gains from meetings through intermediation. That is, any meeting between two individuals with identical contact rates λ is gainful solely in the double-coincidence case. In contrast, when two individuals with different contact rates meet each other, the meeting is socially gainful not only in the double-coincidence case but also when misalignment can be traded towards the faster individual. An atomless distribution maximizes the fraction of meetings in which there are gains from trade.

Furthermore, the proposition implies that the optimal distribution has a Pareto tail and features middlemen, a zero measure of individuals with infinite contact rate. The Pareto tail with parameter 2 arises from exactly the same logic discussed above for the equilibrium case. Given the linear cost function it equates the marginal social returns to meetings created by different types λ. The planner introduces middlemen for reasons that likewise mimic the equilibrium case. A trader's social value consists of the direct flow valuation she derives

from her asset holdings and from the impact she has on the overall allocation net of the cost of meetings. The latter component again becomes linearly scalable as λ becomes very high; when a fast trader becomes even faster, she conducts the same types of trade, just more frequently. It thus follows that the net social value that arises from trade by fast types needs converges to zero when fast types coexist, as is the case with the Pareto tail. But in the absence of middlemen fast traders have a misalignment rate converging to the autarky level of $\frac{1}{2}$, implying that those traders have zero social value in excess of autarky. Clearly, this cannot be part of the optimal solution where the marginal social returns to meetings are equated across traders as long as the optimal solution dominates autarky—which is the case whenever $c < \Delta/16\gamma$.

In the frictionless limit, as the constant meeting cost c approaches zero, the optimal allocation shares the same qualitative features as equilibrium. However, the exact magnitudes are different. In the optimal allocation, the fraction of meetings that go to middlemen is e^{-1}, more than in equilibrium. Volume is $(e - \frac{1}{2})\gamma \approx 2.22\gamma$, which is smaller than the equilibrium trading rate. As in equilibrium, middlemen buy from middlemen at rate $\frac{1}{2}\gamma$ and from finite traders at rate $\frac{1}{2}\gamma$ while finite traders buy from middlemen at rate $\frac{1}{2}\gamma$. That is, just like in equilibrium, the reallocation of the asset following a taste shock happens through an intermediation chain which always involves two middlemen. The only difference from the equilibrium allocation is that the optimal volume of purchases by one finite trader of the asset held by another is slightly lower, namely $(e - 2)\gamma$.

We can also characterize the limiting behavior of the entire optimal search intensity distribution in closed form. Let Ψ denote the distribution of $\rho \equiv \lambda/\underline{\lambda}$, so $\Psi(\rho) = G(\rho\underline{\lambda})$ for all $\rho \geq 1$. When costs converge zero, $\underline{\lambda}$ grows without bound but $\Psi(\rho)$ converges to $(1-\rho^{-1})e^{\rho^{-1}}$. Compared to the equilibrium meeting distribution provided in the appendix, one can see that the equilibrium distribution is more concentrated than the optimal solution. These discrepancies between the optimal and the equilibrium allocation in the frictionless limit are consistent with the more general discussion of inefficiency offered in the next subsection.

4.2 Pigouvian Taxes

Although the qualitative features of the equilibrium and optimal allocations are identical, the equilibrium distribution is still inefficient. There are two sources of inefficiency. The first comes from bargaining. In the decentralized equilibrium, each trader keeps only half of the surplus from every meeting, while the social planner recognizes the value of the entire surplus. This force induces traders to underinvest in meetings in equilibrium. Formally, this is readily observable by contrasting equations (11) and (18): The planner puts twice the

weight on the option value of trade compared to the individual trader.

The second source of inefficiency is a business stealing effect. When one trader invests more in meetings, she diverts meetings towards herself. That is, investing more in meetings does not affect the contact rate of the other market participants, but it changes the distribution of whom they meet. The failure to internalize this effect causes traders to overinvest in meetings in equilibrium. This business stealing effect is represented by the last (negative) term in (18). Notice that no corresponding term exists in the equilibrium counterpart (11).

We can directly correct for each of these externalities. First, assume that whenever two traders meet, an outside agent doubles the surplus from the meeting. Second, assume that whenever a trader meets anyone, the outside agent charges her a tax which is independent of her contact rate, $\bar{\tau} \equiv \frac{\gamma}{\Lambda} \int_0^\infty S_{\lambda'}(1-2m_{\lambda'})dG(\lambda')$. Then the equilibrium Bellman equations for misaligned and well-aligned traders become

$$rv_{\lambda,0} = \delta_0 + \gamma s_\lambda + \frac{\lambda}{2}\mathbb{E}\left((s_\lambda + s_{\lambda'})^+ m_{\lambda'} + (s_\lambda - s_{\lambda'})^+(1-m_{\lambda'})\right) - \lambda(c(\lambda)+\bar{\tau})$$
$$\text{and } rv_{\lambda,1} = \delta_1 - \gamma s_\lambda + \frac{\lambda}{2}\mathbb{E}\left((-s_\lambda + s_{\lambda'})^+ m_{\lambda'} + (-s_\lambda - s_{\lambda'})^+(1-m_{\lambda'})\right) - \lambda(c(\lambda)+\bar{\tau}).$$

Taking the difference between these equations, we confirm that this reduces to equation (53), and hence the equilibrium and optimal surplus functions are equal, $s_\lambda = S_\lambda$. Moreover, the equilibrium choice of λ, which must maximize the right hand side of the Bellman equation for the well-aligned trader, coincides with the optimal choice from conditions (18) and (19). It follows that the optimal allocation is an equilibrium with these Pigouvian taxes.

In the case of a constant cost per meeting, the tax and subsidy scheme that decentralizes these forces nets out to zero (in expectation) for all traders λ. That is, traders receive a subsidy when they are misaligned and pay a tax when they are well-aligned, but the weighted average of these transfers is zero:

Proposition 9 *When $c(\lambda) = c$, the tax and subsidy system that decentralizes the optimal allocation is such that $\bar{\tau} = c$ and the expected per worker subsidy in each meeting also equals c for all types λ.*

To understand this finding, first note that the social value of an average meeting equals in an optimal allocation is the cost of the meeting, $2c$ for all traders λ. The subsidy we propose doubles the joint surplus in each meeting, hence adds on average an additional $2c$. Since those are split symmetrically, the average expected subsidy per worker and meeting is c. To understand the tax, it is easiest to think about the business stealing externality in the following way: When increasing her contact rate, a trader effectively prevents others from meeting each other. In particular, for each two additional meetings a trader "acquires,"

she replaces one average meetings between two other traders. Under the optimal allocation, that average meeting has value $2c$. Thus, the tax $\bar{\tau}$ that internalizes the business stealing externality equals c.

It may seem paradoxical that a tax and subsidy system which does not transfer resources across risk-neutral traders could affect the equilibrium allocation. The reason it has a real effect is that the taxes and transfers alter payoffs conditional on alignment status, hence the threat points in bargaining and the equilibrium price of assets. This in turn affects investment decisions. To see this, we consider an equivalent representation of the tax and subsidy scheme, which transfers are independent of who a trader contacts. Under this scheme, misaligned trader pays a per-meeting tax

$$\tau_{\lambda,0} = \frac{\gamma}{\Lambda}\int_0^\infty S_{\lambda'}(1-2m_{\lambda'})dG(\lambda') - \frac{1}{4}\mathbb{E}\big((S_\lambda + S_{\lambda'})m_{\lambda'} + \mathbb{I}_{\lambda'>\lambda}(S_\lambda - S_{\lambda'})(1-m_{\lambda'})\big),$$

while a well aligned trader pays a tax

$$\tau_{\lambda,1} = \frac{\gamma}{\Lambda}\int_0^\infty S_{\lambda'}(1-2m_{\lambda'})dG(\lambda') - \frac{1}{4}\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda}(S_{\lambda'} - S_\lambda)m_{\lambda'}\big).$$

It is immediate that $\tau_{\lambda,1} > \tau_{\lambda,0}$ for all λ and proposition 9 implies they average out to zero. This means that the taxes affect the threat points in an intermediated trade. Intermediaries take on misaligned asset positions from slower types. The taxes on the slower trader reduce her gains from trade, which in turn reduces the profitability of intermediation under Nash bargaining. Thus, while the direct payments net out to zero, their indirect effect on asset prices offset the inefficiencies that lead to overinvestment in the equilibrium without taxes.

4.3 Equilibrium and Optimum: Numerical Illustration

This section numerically contrasts the equilibrium with its optimal counterpart along several dimensions. First, the blue lines in Figure 2 summarize the optimal allocation with γ normalized to 1. For any cost function, the top panel shows that the lower bound on the optimal contact rate distribution is lower than under the equilibrium contact rate distribution.[15] The bottom panel shows that volume is also inefficiently high in equilibrium. We view this as numerical proof that the equilibrium displays systematic overinvestment.

To develop an intuition for this finding, we note that the business stealing externality is identical for all types λ. In turn, the positive externality that arises from bargaining is larger for those traders with higher average surplus meetings which are those with lower λ.

[15]Furthermore, there appears to be a one-to-one mapping between contact rates $c/\Delta < 1/16$ and lower bounds $\underline{\lambda}$. Just as in equilibrium, this suggests that the optimal allocation exists and is unique.

As a consequence, there are too few slow types and too many fast types and the undistorted equilibrium displays excessive trading volume.

On the other hand, given the lower bound on the contact rate, the rest of the distribution is more compressed in equilibrium than the optimum. The second panel in Figure 2 shows that the optimal average contact rate relative to the lower bound is higher than the equilibrium counterparts. Likewise, the third panel shows that the optimal share of middlemen is higher than the equilibrium share.

The blue lines in Figure 3 illustrate more details of the optimal allocation when $c/\Delta = 0.001$. The top panel shows that the optimal distribution of contact rates is "close" to a Pareto with tail parameter 2, similar to the equilibrium, although we can again prove it is not an exact Pareto. The middle panel shows that this carries over to the distribution of trading rates, implying that the empirically documented scale-free nature of many financial networks is also a feature of a market that optimally leverages the gains from intermediation when the cost per meeting is constant. Once again most traders have an optimal trading rate that exceeds γ, the natural benchmark for an economy without intermediation. Intermediation is key to understanding the amount of trades in the optimal allocation.

Finally, the blue line in the bottom panel in Figure 3 plots the misalignment rate against λ in both equilibrium and the optimal allocation. As expected from Propositions 4 and 8, traders with a higher contact rate have a higher misalignment rate in both equilibrium and optimum. The equilibrium distribution of contact rates first order stochastically dominates the optimal one; as a consequence, all traders with finite λ are more likely to meet with a faster individual when comparing equilibrium with the optimal case. The reason the misalignment rate crosses is that the planner allocates a larger fraction of meetings to middlemen; see the bottom panel of Figure 2. This disproportionately benefits fast traders since they can offload misalignment.

5 Constrained Economy: The Role of Intermediation

To understand the role of intermediation and its connection with heterogeneity, consider an economy in which intermediation is not allowed. To be concrete, suppose meetings between two traders with the same preference state simply do not occur. It follows that whenever a misaligned trader meets a well-aligned trader, they have opposite preference states and hence the same asset holdings, and so there is no scope for trade. We show in this section that without intermediation, the equilibrium and optimal distribution of contact rates are degenerate as long as the cost function $\lambda c(\lambda)$ is weakly convex.

Constrained Equilibrium We start by defining an equilibrium without intermediation analogously to the definition of equilibrium with intermediation.

Definition 2 *A steady state equilibrium is a distribution of contact rates $G(\lambda)$, an average contact rate Λ, an allocation of misalignment m_λ, and undiscounted surplus function s_λ, satisfying the following conditions:*

1. *Balanced inflows and outflows into misalignment as given by*

$$\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{s_\lambda + s_{\lambda'} > 0} m_{\lambda'}\right)\right) m_\lambda = \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{s_\lambda + s_{\lambda'} < 0}(1 - m_{\lambda'})\right)\right)(1 - m_\lambda). \tag{21}$$

2. *Consistency of s_λ with the value functions (5) and (6) adjusted for the no-intermediation case,*

$$\Delta = 2\gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\Big((s_\lambda + s_{\lambda'})^+ m_{\lambda'} - (-s_\lambda - s_{\lambda'})^+(1 - m_{\lambda'})\Big) \tag{22}$$

3. *Optimality of the ex-ante investment decision:*

   (a) $dG(\lambda) > 0$ only if it maximizes

$$\delta_1 - \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left(\big(-s_\lambda - s_{\lambda'}\big)^+(1 - m_{\lambda'})\right) - \lambda c(\lambda)$$

   (b) Middlemen make finite profits: $\Lambda \geq \int_0^\infty \lambda dG(\lambda)$ and

$$\lim_{\lambda \to \infty}\left(\frac{1}{4}\mathbb{E}\left((-s_\lambda - s_{\lambda'})^+(1 - m_{\lambda'})\right) - c(\lambda)\right) \leq 0,$$

   with complementary slackness.

Note that there are two relevant types of meetings in this constrained economy, those between two misaligned traders with the opposite asset holdings, and those between two well-aligned traders with the opposite asset holdings. We can extend our earlier results to prove that the first type of meeting results in trade while the second does not. That is, in equilibrium, two well-aligned agents never find it optimal to jointly trade into misalignment.

Constrained Planner's Problem We turn next to the planner's problem. The planner again chooses the distribution of contact rates along with the admissible set of trades; as in equilibrium the planner is subject to the constraint that intermediation is not allowed.

The objective of the planner is unchanged, given by equation (14). Since we are interested in the case where intermediation is not allowed the planner is subject to an adjusted constraint on the evolution of the misalignment rate,

$$\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,0}^{\lambda',0} m_{\lambda'}\right)\right) m_\lambda = \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,1}^{\lambda',1}(1-m_{\lambda'})\right)\right)(1-m_\lambda). \tag{23}$$

The following proposition then summarizes our key findings for the environment where solely fundamental trades between traders with different preferences are allowed.

Proposition 10 *Consider an economy with no intermediation and a weakly convex cost $\lambda c(\lambda)$. In equilibrium all traders choose a common value $\lambda = \Lambda$. The same holds in the solution to the planner's problem. With a continuously differentiable cost function and interior solution, the equilibrium contact rate is inefficiently high.*

The second part of the proposition highlights that the overinvestment result we numerically document for the unconstrained case holds in the constrained case. Individuals do not internalize that investing in additional meetings lowers misalignment in the marketplace thereby reducing the marginal value of meetings acquired by others. This force leads to overinvestment in the aggregate.

More importantly, the first part of the proposition highlights that the heterogeneity that arises in the full economy is an immediate, and socially desirable, consequence of intermediation. When individuals are restricted to trades driven by static fundamentals, there is no gains from heterogeneity in the contact rate.

This mass point result reflects that, without intermediation, there is effectively decreasing returns to contacts at the individual level. The misalignment rate is strictly decreasing in meetings; as an individual becomes increasingly well-aligned fewer meetings lead to gainful trading opportunities. As a consequence, an unequal distribution of meetings comes with first-order losses and the optimal distribution is degenerate. The same is true in equilibrium; with a weakly convex cost function but decreasing returns on the individual level the only distribution where ex-ante homogeneous traders have identical value is a degenerate one.

In summary, intermediation and heterogeneity are closely interconnected in a market with search frictions. Without heterogeneity there is no intermediation, and without intermediation there is no heterogeneity. Heterogeneity is useful because in meetings where both sides have identical preferences misalignment can be traded towards the faster trader using intermediation to facilitate the transmission of the asset to those with a desire for it.

6 Conclusions

We study a model of over-the-counter trading in which ex-ante identical traders invest in a meeting technology and participate in bilateral trade. We show that when traders have heterogeneous search efficiencies, fast traders intermediate for slow traders: they trade against

their desired position and take on misalignment from slower traders. Moreover, we characterize how, starting with ex ante homogeneous traders, the distribution of contact rates is determined endogenously in equilibrium, and how it compares with the corresponding socially optimal distribution. We argue that an economy with homogeneous contact rates is neither an equilibrium nor socially desirable when the cost of meetings is differentiable. Under a linear cost function the endogenous and optimal distribution of trading rates is governed by a power law, an empirical hallmark of various financial markets. Moreover, middlemen with an infinite contact rate account for a positive fraction of meetings. We also characterize the transfer scheme which decentralizes the optimal allocation, offsetting the forces that lead to overinvestment in the undistorted equilibrium. Finally, we argue that when intermediation is prohibited, dispersion in contact rates disappears both in equilibrium and in the optimal allocation, which illustrates the interplay between heterogeneity and intermediation in a frictional marketplace.

We have kept our model as simple as possible in order to show how intermediation and middlemen naturally arise in over-the-counter markets. It would be natural to extend our model to a more complex environment, for example one in which the two misaligned states are not symmetric, or one in which the binary restriction on asset holdings is relaxed. We believe that the basic forces we highlight in this paper will be robust to such extensions. Likewise, we believe that the random matching model with endogenous contact rates may be useful for understanding other issues in financial markets, such as the percolation of information (Duffie and Manso, 2007). We hypothesize that middlemen may serve a useful role in this process as well.

References

Afonso, Gara and Ricardo Lagos, "Trade Dynamics in the Market for Federal Funds," *Econometrica*, 2015, *83* (1), 263–313.

_ **, Anna Kovner, and Antoinette Schoar**, "Trading Partners in the Interbank Lending Market," 2013.

Allen, Linda and Anthony Saunders, "The Large-Small Bank Dichotomy in the Federal Funds Market," *Journal of Banking & Finance*, 1986, *10* (2), 219–230.

Babus, Ana and Tai-Wei Hu, "Endogenous Intermediation in Over-the-Counter Markets," 2015.

Bech, Morten L. and Enghin Atalay, "The Topology of the Federal Funds Market," *Physica A: Statistical Mechanics and its Applications*, 2010, *389* (22), 5223–5246.

Boss, Michael, Helmut Elsinger, Martin Summer, and Stefan Thurner, "Network Topology of the Interbank Market," *Quantitative Finance*, 2004, *4* (6), 677–684.

Burdett, Kenneth and Dale T Mortensen, "Wage Differentials, Employer Size, and Unemployment," *International Economic Review*, 1998, pp. 257–273.

_ **and Kenneth L Judd**, "Equilibrium Price Dispersion," *Econometrica*, 1983, *51* (4), 955–969.

Butters, Gerard R, "Equilibrium Distributions of Sales and Advertising Prices," *The Review of Economic Studies*, 1977, *44* (3), 465–491.

Chang, Briana and Shengxing Zhang, "Endogenous Market Making and Network Formation," 2016.

Chang, Eui Jung, Eduardo José Araujo Lima, Solange M Guerra, and Benjamin M Tabak, "Measures of Interbank Market Structure: An Application to Brazil," *Brazilian Review of Econometrics*, 2008, *28* (2), 163–190.

Craig, Ben and Goetz Von Peter, "Interbank Tiering and Money Center Banks," *Journal of Financial Intermediation*, 2014, *23* (3), 322–347.

Duffie, Darrell and Gustavo Manso, "Information Percolation in Large Markets," *American Economic Review*, 2007, *97* (2), 203–209.

_ , **Nicolae Gârleanu, and Lasse Heje Pedersen**, "Over-the-Counter Markets," *Econometrica*, 2005, *73* (6), 1815–1847.

_ , _ , **and** _ , "Valuation in Over-the-Counter Markets," *Review of Financial Studies*, 2007, *20* (6), 1865–1900.

_ , **Piotr Dworczak, and Haoxiang Zhu**, "Benchmarks in Search Markets," 2016.

Eeckhout, Jan, "Gibrat's Law for (all) Cities," *American Economic Review*, 2004, pp. 1429–1451.

Farboodi, Maryam, "Intermediation and Voluntary Exposure to Counterparty Risk," 2015.

_ , **Gregor Jarosch, and Guido Menzio**, "Tough Middlemen - A Theory of Intermediation as Pure Rent Extraction," 2016.

Gabaix, Xavier, "Zipf's Law for Cities: An Explanation," *Quarterly Journal of Economics*, 1999, pp. 739–767.

_ , "Power laws in economics and finance," *Annu. Rev. Econ.*, 2009, *1* (1), 255–294.

_ , "Power Laws in Economics: An Introduction," *The Journal of Economic Perspectives*, 2016, *30* (1), 185–205.

_ , **Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley**, "Institutional investors and stock market volatility," *The Quarterly Journal of Economics*, 2006, *121* (2), 461–504.

Geerolf, Francois, "A Theory of Pareto Distributions," 2016.

Heller, Daniel and Nicholas Vause, "Collateral Requirements for Mandatory Central Clearing of Over-the-Counter Derivatives," 2012.

Hollifield, Burton, Artem Neklyudov, and Chester S Spatt, "Bid-Ask Spreads, Trading Networks and the Pricing of Securitizations: 144a vs. Registered Securitizations," 2014.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, "Heterogeneity in Decentralized Asset Markets," 2016.

in 't Veld, Daan and Iman van Lelyveld, "Finding the Core: Network structure in interbank markets," *Journal of Banking & Finance*, 2014, *49*, 27–40.

Jackson, Matthew O., *Social and Economic Networks*, Princeton University Press, 2010.

Lagos, Ricardo and Guillaume Rocheteau, "Liquidity in Asset Markets with Search Frictions," *Econometrica*, 2009, *77*, 403–426.

Li, Dan and Norman Schürhoff, "Dealer Networks," 2014.

Maggio, Marco Di, Amir Kermani, and Zhaogang Song, "The Value of Trading Relationships in Turbulent Times," 2015.

Mengle, David, "Concentration of OTC Derivatives among Major Dealers," *ISDA Research Notes*, 2010, *4*.

Neklyudov, Artem V, "Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers," 2014.

Nosal, Ed, Yuet-Yee Wong, and Randall Wright, "Who Wants to be a Middleman?," 2016.

Pagnotta, Emiliano and Thomas Philippon, "Competing on Speed," 2015.

Peltonen, Tuomas A, Martin Scheicher, and Guillaume Vuillemey, "The Network Structure of the CDS Market and its Determinants," *Journal of Financial Stability*, 2014, *13*, 118–133.

Rubinstein, Ariel and Asher Wolinsky, "Middlemen," *The Quarterly Journal of Economics*, 1987, pp. 581–594.

Siriwardane, Emil N, "Concentrated Capital Losses and the Pricing of Corporate Credit Risk," 2015.

Üslü, Semih, "Pricing and Liquidity in Decentralized Asset Markets," 2016.

Wang, Chaojun, "Core-Periphery Trading Networks," 2017.

Weill, Pierre-Olivier, "Liquidity Premia in Dynamic Bargaining Markets," *Journal of Economic Theory*, 2008, *140* (1), 66–96.

A Appendix

Proof of Proposition 1. We show that s_λ is non-negative and strictly decreasing for all $\lambda > 0$. The proposition immediately follows.

If $s_\lambda < 0$, $(-s_\lambda + s_{\lambda'})^+ \geq (s_\lambda + s_{\lambda'})^+$ and $(-s_\lambda - s_{\lambda'})^+ \geq (s_\lambda - s_{\lambda'})^+$ for all λ'. Equation (8) then implies $s_\lambda \geq 2/\gamma$, a contradiction. This proves s is non-negative. Likewise, equation (7) implies m is non-negative. It follows immediately that for all λ and λ', $\mathbb{I}_{s_\lambda + s_{\lambda'} > 0} m_{\lambda'} \geq \mathbb{I}_{s_\lambda < s_{\lambda'}} m_{\lambda'}$ and $\mathbb{I}_{s_\lambda > s_{\lambda'}}(1 - m_{\lambda'}) \geq \mathbb{I}_{s_\lambda + s_{\lambda'} < 0}(1 - m_{\lambda'})$. Equation (7) then implies $m_\lambda \leq 1/2$ for all λ.

Next, use s nonnegative to rewrite equation (8) as

$$2\gamma s_\lambda = \Delta + \frac{\lambda}{4}\mathbb{E}\left(\big((-s_\lambda + s_{\lambda'})^+ - (s_\lambda + s_{\lambda'})\big)m_{\lambda'} - (s_\lambda - s_{\lambda'})^+(1 - m_{\lambda'})\right).$$

Use $(-s_\lambda + s_{\lambda'})^+ = s_{\lambda'} - \min\{s_\lambda, s_{\lambda'}\}$ and $(s_\lambda - s_{\lambda'})^+ = s_\lambda) - \min\{s_\lambda, s_{\lambda'}\}$ to rewrite this as

$$2\gamma s_\lambda = \Delta + \frac{\lambda}{4}\mathbb{E}\left(\big(-\min\{s_\lambda, s_{\lambda'}\} - s_\lambda\big)m_{\lambda'} - \big(s_\lambda - \min\{s_\lambda, s_{\lambda'}\}\big)(1 - m_{\lambda'})\right).$$

Grouping terms, this gives

$$s_\lambda = \frac{4\Delta + \lambda\mathbb{E}\big(\min\{s_\lambda, s_{\lambda'}\}(1 - 2m_{\lambda'})\big)}{8\gamma + \lambda}. \tag{24}$$

View this as a mapping $s = T(s)$. We claim that for any cumulative distribution function G and misalignment function m with range $[0, 1/2]$, T is a contraction, mapping continuous functions on $[0, \Delta/2\gamma]$ into the same set of functions. Continuity is immediate. Similarly, if s is nonnegative, $T(s)$ is nonnegative. If $s \leq \Delta/2\gamma$,

$$T(s)_\lambda \leq \left(\frac{8\gamma + \lambda\mathbb{E}(1 - 2m_{\lambda'})}{8\gamma + \lambda}\right)\left(\frac{\Delta}{2\gamma}\right).$$

Since the misalignment rate is nonnegative, the result follows.

Finally, we prove T is a contraction. If $|s^1_\lambda - s^2_\lambda| \leq \varepsilon$ for all λ,

$$|T(s^1)_\lambda - T(s^2)_\lambda| \leq \frac{\lambda\varepsilon\mathbb{E}(1 - 2m_{\lambda'})\big)}{8\gamma + \lambda} \leq \varepsilon\mathbb{E}(1 - 2m_{\lambda'}).$$

Note that the second inequality uses the fact that the fraction is increasing in λ and hence evaluates it at the limit as λ converges to infinity. Since $\mathbb{E}(1 - 2m_{\lambda'}) < 1$, this proves that T is a contraction in the sup-norm, with modulus $\int_0^\infty \frac{\lambda'}{\Lambda}(1 - 2m_{\lambda'})dG_{\lambda'}$.

Next we prove that the mapping T takes nonincreasing functions s and maps them into

decreasing functions. Take $\lambda_1 < \lambda_2$ and let $E(\lambda) \equiv \mathbb{E}\big(\min\{s_\lambda, s_{\lambda'}\}(1-2m_{\lambda'})\big)$. Note that m nonnegative and $s_\lambda \leq \Delta/2\gamma$ implies $E(\lambda) \leq \Delta/2\gamma$ as well. Similarly, s nonincreasing implies E is nonincreasing as well. Then

$$\begin{aligned} T(s)(\lambda_1) - T(s)(\lambda_2) &= \frac{4\Delta + \lambda_1 E(\lambda_1)}{8\gamma + \lambda_1} - \frac{4\Delta + \lambda_2 E(\lambda_2)}{8\gamma + \lambda_2} \\ &\geq \frac{4\Delta + \lambda_1 E(\lambda_1)}{8\gamma + \lambda_1} - \frac{4\Delta + \lambda_2 E(\lambda_1)}{8\gamma + \lambda_2} \\ &= \frac{4(\lambda_2 - \lambda_1)\big(\Delta - 2\gamma E(\lambda_1)\big)}{(8\gamma + \lambda_1)(8\gamma + \lambda_2)} > 0, \end{aligned}$$

The first equality is the definition of T. The first inequality uses $E(\lambda_2) \leq E(\lambda_1)$. The second equality groups the two fractions over a common denominator. And the second equality uses $E(\lambda) < \Delta/2\gamma$. This proves the result. It follows that the equilibrium surplus function is decreasing. ■

Deriving Equation (9)**.** Since the surplus function is nonnegative and nonincreasing, we can rewrite equation (8) as

$$\Delta = 2\gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\bigg(\Big(\big(s_\lambda + s_{\lambda'}\big) - \mathbb{I}_{\lambda'<\lambda}\big(s_{\lambda'} - s_\lambda\big)\Big)m_{\lambda'} + \mathbb{I}_{\lambda'>\lambda}\big(s_\lambda - s_{\lambda'}\big)(1 - m_{\lambda'})\bigg)$$

Regroup terms:

$$4\Delta = \big(8\gamma + 2\lambda\mathbb{E}m_{\lambda'}\big)s_\lambda + \lambda\mathbb{E}\big(\mathbb{I}_{\lambda'\geq\lambda}\big(s_\lambda - s_{\lambda'}\big)(1 - 2m_{\lambda'})\big) \tag{25}$$

Differentiate with respect to λ:

$$0 = \big(8\gamma + \lambda\mathbb{E}(\mathbb{I}_{\lambda'>\lambda} + 2\lambda\mathbb{E}(\mathbb{I}_{\lambda'<\lambda}m_{\lambda'})\big)s'_\lambda + 2\mathbb{E}m_{\lambda'}s_\lambda + \mathbb{E}\big(\mathbb{I}_{\lambda'>\lambda}\big(s_\lambda - s_{\lambda'}\big)(1 - 2m_{\lambda'})\big)$$

Replace the last two terms using equation (25) and simplify using the definitions of H and L to get

$$s'_\lambda = \frac{8\gamma s_\lambda - 4\Delta}{\lambda\Big(8\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)} = \phi_\lambda\left(s_\lambda - \frac{\Delta}{2\gamma}\right),$$

where ϕ_λ is defined in equation (10). The general solution to this differential equation is

$$s_\lambda = \frac{\Delta}{2\gamma} - ke^{\int_{\underline{\lambda}}^{\lambda}\phi_{\lambda'}d\lambda'}$$

for fixed $\underline{\lambda}$ and some constant of integration k. Use the fact that $\lim_{\lambda\to\infty} s_\lambda = 0$ to pin down the constant of integration, yielding equation (9).

Proof of Proposition 2. We proceed by contradiction. Assume $G(\lambda)$ has a mass point at some $\bar{\lambda} > 0$. From equation (10), we have that ϕ_λ jumps up at $\bar{\lambda}$ since $2m_\lambda < 1$ for all λ. Differentiating equation (9), we have that

$$s'_\lambda = -\frac{\Delta}{2\gamma}\phi_\lambda e^{-\int_\lambda^\infty \phi_{\lambda'} d\lambda'},$$

and so this jumps down at $\bar{\lambda}$. It follows that s_λ has a concave kink at $\bar{\lambda}$.

Now consider part 3(a) of the definition of equilibrium. Using the monotonicity and nonnegativity of s_λ, the choice of λ must maximize

$$\delta_1 - \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda}\big(s_{\lambda'} - s_\lambda\big)m_{\lambda'}\right) - \lambda c(\lambda)$$

The first derivative with respect to λ is

$$-\left(\gamma + \frac{\lambda}{4}\mathbb{E}(\mathbb{I}_{\lambda'<\lambda}m_{\lambda'})\right)s'_\lambda + \frac{1}{4}\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda}\big(s_{\lambda'} - s_\lambda\big)m_{\lambda'}\right) - c(\lambda) - \lambda c'(\lambda).$$

Note that s'_λ jumps down at $\bar{\lambda}$ and the other expressions are continuous in λ. Therefore the slope of the objective function jumps up at $\bar{\lambda}$. That is, $\bar{\lambda}$ represents a local minimum in traders' objective function, contradicting the assumption that $G(\lambda)$ has a mass point at $\bar{\lambda}$. ■

Proof of Proposition 3. To find a contradiction, we suppose that there is a hole in the distribution of contact rates. That is, there are contact rates $\underline{\lambda} < \bar{\lambda} < \infty$ such that $dG(\lambda) = 0$ for all $\lambda \in (\underline{\lambda}, \bar{\lambda})$. Moreover, for this to be a hole in the distribution, we require that $\mathbb{E}(\mathbb{I}_{\lambda'\leq\underline{\lambda}}) > 0$ and $\mathbb{E}(\mathbb{I}_{\lambda'\geq\bar{\lambda}}) > 0$, so individuals choose contact rates on either side of the hole. Finally, without loss of generality, we choose the thresholds $\underline{\lambda}$ and $\bar{\lambda}$ such that some individuals choose the extreme values of this set in equilibrium. Using part 3(a) of the definition of equilibrium, this implies $\underline{\lambda}$ and $\bar{\lambda}$ each maximize

$$v_\lambda \equiv -\gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}(\mathbb{I}_{\lambda'<\lambda}(s_{\lambda'} - s_\lambda)m_{\lambda'}) - \lambda c(\lambda).$$

Note that v_λ is continuous in λ, even if there are mass points in the contact rate distribution. We prove that v_λ is strictly concave for $\lambda \in (\underline{\lambda}, \bar{\lambda})$, which contradicts the requirement that $\underline{\lambda}$ and $\bar{\lambda}$ both maximize v.

To prove concavity of v, first rewrite v in this interval, using the fact that no one has a contact rate in the interval $(\underline{\lambda}, \bar{\lambda})$:

$$v_\lambda \equiv -\left(\gamma + \frac{\lambda}{4}\mathbb{E}(\mathbb{I}_{\lambda' \leq \underline{\lambda}} m_{\lambda'})\right) s_\lambda + \frac{\lambda}{4}\mathbb{E}(\mathbb{I}_{\lambda' \leq \underline{\lambda}} s_{\lambda'} m_{\lambda'}) - \lambda c(\lambda).$$

The second term is linear in λ and the third is weakly concave, so the result follows if the first term is strictly concave, i.e. if $\left(\gamma + \frac{\lambda}{4}\mathbb{E}(\mathbb{I}_{\lambda' \leq \underline{\lambda}} m_{\lambda'})\right) s_\lambda$ is strictly convex. To establish this, we solve equation (9) for s_λ:

$$s_\lambda = \frac{\Delta}{2\gamma}\left(1 - \frac{\lambda\Big(8\gamma + \bar{\lambda}\big(H(\bar{\lambda}) + 2L(\underline{\lambda})\big)\Big)}{\bar{\lambda}\Big(8\gamma + \lambda\big(H(\bar{\lambda}) + 2L(\underline{\lambda})\big)\Big)} e^{-\int_{\bar{\lambda}}^{\infty} \phi_{\lambda'} d\lambda'}\right).$$

The result immediately follows from twice differentiating with respect to λ.

Finally, by setting $\bar{\lambda} = \infty$ and using part 3(b) of the definition of equilibrium, a similar argument proves that $\Lambda = \int_0^\infty \lambda dG(\lambda)$ whenever the support of $G(\lambda)$ is bounded above which implies the second part of the proposition. ■

Proof of Proposition 4. Use equation (12) to get

$$m_\lambda = \frac{2\gamma + \lambda L(\lambda)}{4\gamma + 2\lambda L(\lambda) + \lambda H(\lambda)}. \tag{26}$$

We are trying to prove that

$$m'_\lambda = \frac{-\lambda(2\gamma + \lambda L(\lambda)H'(\lambda) + H(\lambda)(\lambda^2 L'(\lambda) - 2\gamma)}{\big(4\gamma + 2\lambda L(\lambda) + \lambda H(\lambda)\big)^2} \geq 0 \tag{27}$$

on the support of G.

To prove this, we exploit restrictions on $H'(\lambda)$ and $L'(\lambda)$ implied by equilibrium. The first restriction simply recognizes from the definitions of H and L that $m_\lambda = -L'(\lambda)/H'(\lambda)$ and so equation (26) implies

$$\big(4\gamma + 2\lambda L(\lambda) + \lambda H(\lambda)\big)L'(\lambda) = -\big(2\gamma + \lambda L(\lambda)\big)H'(\lambda). \tag{28}$$

The second equation uses our equilibrium conditions. By Proposition 3, we know that the support of G is convex. For any value of λ in the support, part 3(a) of the definition of

equilibrium states that the profit must be the same, some $\bar{v}$:

$$\bar{v} = \delta_1 - \gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda}\big(s_{\lambda'} - s_\lambda\big)m_{\lambda'}\right) - \lambda c(\lambda), \tag{29}$$

where we simplify the expression slightly using the results from Proposition 1. This implies the first order condition

$$\begin{aligned} 0 &= -\left(\gamma + \frac{\lambda}{4}L(\lambda)\right)s'_\lambda + \frac{1}{4}\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda}(s_{\lambda'} - s_\lambda)m'_\lambda\big) - c(\lambda) - \lambda c'(\lambda) \\ &= -\left(\gamma + \frac{\lambda}{4}L(\lambda)\right)s'_\lambda + \frac{\bar{v} - \delta_1 + \gamma s_\lambda}{\lambda} - \lambda c'(\lambda) \\ &= \frac{\Delta\left(4\gamma + \lambda L(\lambda)\right)e^{-\int_\lambda^\infty \phi_{\lambda'}d\lambda'}}{\lambda\Big(8\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)} + \frac{\bar{v} - \delta_1 + \frac{\Delta}{2}\left(1 - e^{-\int_\lambda^\infty \phi_{\lambda'}d\lambda'}\right)}{\lambda} - \lambda c'(\lambda) \\ &= \frac{-(\delta_1 - \delta_0)H(\lambda)e^{-\int_\lambda^\infty \phi_{\lambda'}d\lambda'}}{2\Big(8\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)} + \frac{\bar{v} - \frac{\delta_1+\delta_0}{2} - \lambda^2 c'(\lambda)}{\lambda} \end{aligned} \tag{30}$$

where the second line simplifies the using first equation (29), the third line replaces s_λ and its derivative using equation (9) and replaces ϕ_λ using equation (10), and the fourth line groups terms and recognizes that $\Delta = \delta_1 - \delta_0$. Rewrite this as

$$\frac{\lambda H(\lambda)e^{-\int_\lambda^\infty \phi_{\lambda'}d\lambda'}}{8\gamma + \lambda(H(\lambda) + 2L(\lambda))} + \frac{2\lambda^2 c'(\lambda)}{\delta_1 - \delta_0} = \frac{2\bar{v} - \delta_1 - \delta_0}{\delta_1 - \delta_0} \tag{31}$$

for all λ in the support of $G(\lambda)$. Differentiating this expression and replacing ϕ_λ gives us a second equation relating $H'(\lambda)$ and $L'(\lambda)$:

$$\frac{2e^{-\int_\lambda^\infty \phi_{\lambda'}d\lambda'}\big(\lambda(4\gamma + \lambda L(\lambda))H'(\lambda) + H(\lambda)(8\gamma - \lambda^2 L'(\lambda))\big)}{\big(8\gamma + \lambda(H(\lambda) + 2L(\lambda))\big)^2} + \frac{2\lambda}{\Delta}(2c'(\lambda) + \lambda c''(\lambda)) = 0. \tag{32}$$

Solve equations (28) and (32) for $H'(\lambda)$ and $L'(\lambda)$ and substitute into equation (27). The resulting expression is cumbersome, but it is easy to verify that it is strictly positive if $2c'(\lambda) + \lambda c''(\lambda) \geq 0$, i.e. if $\lambda c(\lambda)$ is weakly convex. ■

Proof of Proposition 5. We start by proving that there is an equilibrium with $\Lambda = 0$ if and only if $c \geq \Delta/16\gamma$. To do this, we suppose that such an equilibrium exists. Equation (7) implies $m_0 = 1/2$. Since the surplus function is nonnegative and nonincreasing

(Proposition 1), equation (8) implies that for all $\lambda > 0$,

$$s_\lambda = \frac{4\Delta}{8\gamma + \lambda},$$

with $s_0 = \Delta/2\gamma$. Then part 3 of the definition of equilibrium implies that there is an equilibrium with $\lambda = 0$ if and only if 0 maximizes

$$-\gamma s_\lambda + \frac{\lambda}{8}(s_0 - s_\lambda) - \lambda c.$$

Substituting for s_λ, this reduces to the necessary and sufficient condition

$$0 \in \arg\max \frac{\Delta(\lambda - 8\gamma)}{16\gamma} - \lambda c.$$

The objective function is linear in λ, maximized at 0 if and only if $c \geq \Delta/16\gamma$, completing this step of the proof.

We next suppose that $c \geq \Delta/16\gamma$ and prove that there is no equilibrium with $\Lambda > 0$. To find a contradiction, suppose that such an equilibrium exists and let $\bar{\lambda} > 0$, possibly infinite, denote the fastest contact rate in the population. As in the first step of the proof, part 3 of the definition of equilibrium implies that $\bar{\lambda}$ must deliver at least as high profits as setting $\lambda = 0$:

$$-\gamma s_{\bar{\lambda}} + \frac{\bar{\lambda}}{4}\mathbb{E}\left(\left(s_{\lambda'} - s_{\bar{\lambda}}\right)m_{\lambda'}\right) - \bar{\lambda}c \geq -\gamma s_0$$

Now use the fact that s is nonincreasing and that $m_{\lambda'} \leq 1/2$. This implies

$$\frac{1}{2}(s_0 - s_{\bar{\lambda}}) \geq \mathbb{E}\left(\left(s_{\lambda'} - s_{\bar{\lambda}}\right)m_{\lambda'}\right).$$

Thus a necessary condition to have an equilibrium with $\Lambda > 0$ is

$$\frac{8\gamma + \bar{\lambda}}{8}(s_0 - s_{\bar{\lambda}}) - \bar{\lambda}c \geq 0$$

Equations (9) and (10) imply

$$s_{\bar{\lambda}} = \frac{2\Delta}{4\gamma + \bar{\lambda}\mathbb{E}(m_{\lambda'})},$$

while $s_0 = \Delta/2\gamma$. Therefore the necessary condition to have an equilibrium with $\Lambda > 0$ reduces

$$\frac{(8\gamma + \bar{\lambda})\bar{\lambda}\Delta\mathbb{E}(m_{\lambda'})}{16\gamma\big(4\gamma + \bar{\lambda}\mathbb{E}(m_{\lambda'})\big)} - \bar{\lambda}c \geq 0$$

Using the fact that $\mathbb{E}(m_{\lambda'}) \in (0, 1/2)$, the left hand side is a strictly convex function of $\bar{\lambda}$.

Moreover, evaluate the first derivative of the left hand side in the limit as $\bar{\lambda}$ converges to infinity. It is negative if and only $c \geq \Delta/16\gamma$. Since the left hand side is nonincreasing for large $\bar{\lambda}$ and globally convex, it is strictly decreasing for all $\bar{\lambda}$. Finally, since the left hand side is 0 at $\bar{\lambda} = 0$, it is strictly negative at all positive values of $\bar{\lambda}$, a contradiction. This proves that if $c \geq \Delta/16\gamma$, there is a unique equilibrium and it has $\Lambda = 0$.

The remainder of the proof focuses on the case where $c < \Delta/16\gamma$. We have already proved that there is no equilibrium with $\Lambda = 0$ and now characterize an equilibrium with $\Lambda > 0$. For all λ in the support of G, equation (31) with $c(\lambda) = c$ reduces to

$$\frac{\lambda H(\lambda) e^{-\int_{\lambda}^{\infty} \phi_{\lambda'} d\lambda'}}{8\gamma + \lambda(H(\lambda) + 2L(\lambda))} = \frac{2\bar{v} - \delta_1 - \delta_0}{\delta_1 - \delta_0}. \tag{33}$$

If 0 is in the support of G, this implies $\bar{v} = \frac{\delta_1 + \delta_0}{2}$, the value of autarky. That is, the right hand side of this equation is 0. The left hand side is then also equal to 0 at all λ in the support of G. This is a contradiction for all strictly positive λ that are smaller than the highest value in the population, a contradiction. This proves that 0 is not in the support of G, but instead the lower bound of the support of G is $\underline{\lambda} > 0$.

Replicating the proof of Proposition 3, we can show that the value function v_λ is strictly concave at all $\lambda < \underline{\lambda}$. Since it must also be weakly higher at $\underline{\lambda}$ than at any lower value of λ, this implies it is strictly increasing on this interval. It follows that $\bar{v} > v_0 = \frac{\delta_1 + \delta_0}{2}$.

Turn now to the upper bound $\bar{\lambda}$ and, to find a contradiction, suppose $\bar{\lambda} < \infty$. That is, $H(\bar{\lambda}) = 0$. Then for all $\lambda \geq \bar{\lambda}$, equation (10) implies

$$\phi_\lambda \equiv \frac{4\gamma}{\lambda\Big(4\gamma + \lambda L(\bar{\lambda})\Big)}.$$

Then equation (9) implies

$$s_\lambda = \frac{2\Delta}{4\gamma + \lambda L(\bar{\lambda})} \qquad \forall \lambda \geq \bar{\lambda}. \tag{34}$$

Since no one chooses these high values of λ, part 3(a) of the definition implies that for all $\lambda > \bar{\lambda}$, the profits must be lower than choosing $\bar{\lambda}$:

$$-\gamma s_\lambda + \frac{\lambda}{4}\mathbb{E}\left(\big(s_{\lambda'} - s_\lambda\big)m_{\lambda'}\right) - \lambda c \leq -\gamma s_{\bar{\lambda}} + \frac{\bar{\lambda}}{4}\mathbb{E}\left(\big(s_{\lambda'} - s_{\bar{\lambda}}\big)m_{\lambda'}\right) - \bar{\lambda} c.$$

Rearranging terms, replacing s_λ using the previous expression, and using $\lambda > \bar{\lambda}$, this reduces to

$$\frac{1}{4}\mathbb{E}(s_{\lambda'} m_{\lambda'}) \leq c. \tag{35}$$

Now compare the value of choosing $\bar{\lambda}$ to the value of autarky, $\lambda = 0$. Using $\bar{v} > v_0$, we have

$$-\gamma s_{\bar{\lambda}} + \frac{\bar{\lambda}}{4}\mathbb{E}\left(\left(s_{\lambda'} - s_{\bar{\lambda}}\right)m_{\lambda'}\right) - \bar{\lambda}c > -\gamma s_0.$$

Use equation (8) to eliminate s_0 and equation (34) to eliminate $s_{\bar{\lambda}}$:

$$\bar{\lambda}\left(\frac{1}{4}\mathbb{E}(s_{\lambda'}m_{\lambda'}) - c\right) > 0.$$

This contradicts inequality (35), establishing the contradiction.

We next prove that middlemen account for a positive fraction of meetings in equilibrium. Our starting point is equation (33). Take the limit as λ converges to ∞, noting that that $e^{-\int_{\lambda}^{\infty}\phi_{\lambda'}d\lambda'} \to 1$ and $8\gamma/\lambda \to 0$. The equation reduces to

$$\lim_{\lambda\to\infty} H(\lambda) = \frac{2\bar{v} - \delta_1 - \delta_0}{\delta_1 - \bar{v}} \lim_{\lambda\to\infty} L(\lambda), \tag{36}$$

where $\lim_{\lambda\to\infty} H(\lambda)$ is the fraction of meetings with middlemen and $\lim_{\lambda\to\infty} L(\lambda)$ is the fraction of misaligned traders. Since we have already proved that $\bar{v} > \frac{\delta_1+\delta_0}{2}$ and there are misaligned traders in equilibrium, the result follows immediately.

To prove the Pareto tail, differentiate equation (33) with respect to λ and evaluate at $\lambda \to \infty$ to get

$$\lim_{\lambda\to\infty} \lambda^3 dG(\lambda) = \frac{24(2\bar{v} - \delta_1 - \delta_0)\gamma\Lambda}{\Delta - (2\bar{v} - \delta_1 - \delta_0)(1 - 2\lim_{\lambda\to\infty} m_\lambda)}.$$

Using the steady state expressions, we find that

$$\lim_{\lambda\to\infty} m_\lambda = \frac{\int_0^\infty \frac{\lambda'}{\Lambda} m_{\lambda'} dG(\lambda')}{1 - \int_0^\infty \frac{\lambda'}{\Lambda} dG(\lambda') + 2\int_0^\infty \frac{\lambda'}{\Lambda} m_{\lambda'} dG(\lambda')} = \frac{\delta_1 - \bar{v}}{\delta_1 - \delta_0} \tag{37}$$

where the second equation simplifies the first using equation (36). Substituting this into the tail parameter expression, we get

$$\lim_{\lambda\to\infty} \lambda^3 dG(\lambda) = \frac{24\gamma\Lambda}{\frac{\delta_1-\delta_0}{2\bar{v}-\delta_1-\delta_0} - \frac{2\bar{v}-\delta_1-\delta_0}{\delta_1-\delta_0}}. \tag{38}$$

Note that this is positive since traders' value is less than the first best, $\bar{v} < \delta_1$. This implies that in the right tail, the G distribution is well-approximated by the Pareto distribution

$$G(\lambda) = 1 - \frac{12\gamma\Lambda}{\frac{\delta_1-\delta_0}{2\bar{v}-\delta_1-\delta_0} - \frac{2\bar{v}-\delta_1-\delta_0}{\delta_1-\delta_0}}\lambda^{-2}. \tag{39}$$

This completes the proof. ■

Proof of Proposition 6. Consider a λ-agent. When misaligned, which happens a fraction m_λ of the time, she trades with any faster trader with opposite asset holding, as well as slower misaligned traders with opposite asset holding. So her trading probability conditional on a meeting is $\frac{1}{2}(H(\lambda) + L(\lambda))$. When she is well aligned, a fraction $(1 - m_\lambda)$ of the time, then she takes on misalignment from slower traders with opposite asset holding, so she trades with probability $\frac{1}{2}L(\lambda)$. Combining this, the probability that a type λ trader actually trades in a meeting is

$$p_\lambda = \frac{1}{2}(m_\lambda H(\lambda) + L(\lambda)).$$

Replace $m_\lambda = -L'(\lambda)/H'(\lambda)$ and eliminate $L'(\lambda)$ and $H'(\lambda)$ using equation (13):

$$p_\lambda = \frac{\gamma(H(\lambda) + 2L(\lambda)) + \lambda L(\lambda)(H(\lambda) + L(\lambda)}{4\gamma + \lambda(H(\lambda) + 2L(\lambda)}.$$

We are interested in the tail behavior of the trading rate λp_λ.

First observe that λp_λ is increasing. To show this, differentiate wtih respect to λ and replace $L'(\lambda)$ and $H'(\lambda)$ using equation (13). The result follows algebraically.

Now let F denote the cumulative distribution of trading rates. Since λp_λ is increasing, $F(\lambda p_\lambda) = G(\lambda)$ for all λ. Differentiating this gives

$$dF(\lambda p_\lambda) = \frac{dG(\lambda)}{\lambda p'_\lambda + p_\lambda}.$$

Equation (39) shows us the behavior of $dG(\lambda)$ in the tail, a Pareto. To see the behavior of the denominator, again differentiate λp_λ with respect to λ and replace $L'(\lambda)$ and $H'(\lambda)$ using equation (13). This time take the limit as λ converges to infinity:

$$(\lambda p_\lambda)^3 \lim_{\lambda\to\infty} dF(\lambda p_\lambda) = \lambda^3 \lim_{\lambda\to\infty} dG(\lambda) \left(\frac{\lim_{\lambda\to\infty} L(\lambda)(\bar{v} - \delta_0)}{\delta_1 - \delta_0}\right)^2.$$

Since $\lambda^3 \lim_{\lambda\to\infty} dG(\lambda) > 0$ by equation (38), this is positive and hence the tail of F is Pareto with tail parameter 2. ■

Proof of Proposition 7. For a fixed value of $c \in (0, 16\Delta/\gamma)$, equilibrium is characterized by a lower bound $\underline{\lambda}$ and the differential equations (13) as well as the requirement that equilibrium costs are in fact c. We get this last condition by evaluating equation (30) at

$\lambda = \underline{\lambda}$ and evaluating $s'_{\underline{\lambda}}$ using equations (9) and (10):

$$\frac{c}{\Delta} = \frac{4\gamma \exp\left(-\int_{\underline{\lambda}}^{\infty} \frac{8\gamma}{\lambda(8\gamma+\lambda H(\lambda)+2\lambda L(\lambda))} d\lambda\right)}{\underline{\lambda}(8\gamma+\underline{\lambda})}. \tag{40}$$

We are interested in evaluating these equations when $c \to 0$, but there are two difficulties with that limit. First, we prove below that $\underline{\lambda} \to \infty$, which makes the limiting system of equations problematic. Second, we find that $L(\lambda) \to 0$ in the same limit. We propose a change in variables to deal with both of these issues. Let $\rho \equiv \lambda/\underline{\lambda}$ denote a worker's contact rate *relative* to the lower bound. Let $h(\rho) \equiv H(\lambda)$ and $\ell(\rho) \equiv L(\lambda)\lambda/\gamma$. Rewrite equations (13) and (40) in terms of ℓ and h:

$$\begin{aligned}
4(2h(\rho) + \rho h'(\rho)) &= \rho\big(h(\rho)\ell'(\rho) - \ell(\rho)h'(\rho)\big) - h(\rho)\ell(\rho), \\
\big(4\gamma + \underline{\lambda}\rho h(\rho) + 2\gamma\ell(\rho)\big)(\rho\ell'(\rho) - \ell(\rho)) &= -\underline{\lambda}\rho^2(2+\ell(\rho))h'(\rho), \\
\frac{c}{\Delta} &= \frac{4\gamma \exp\left(-\int_{1}^{\infty} \frac{8\gamma}{\rho(8\gamma+\rho\underline{\lambda}h(\rho)+2\gamma\ell(\rho))} d\rho\right)}{\underline{\lambda}(8\gamma+\underline{\lambda})}.
\end{aligned}$$

The first two equations, together with the terminal condition $h(1) = 1$ and $\ell(1) = 0$, determine h and ℓ for any $\underline{\lambda}$. If $\underline{\lambda} < \infty$, they imply that the right hand side of the third equation is positive and hence this is inconsistent with the third equation. On the other hand, we prove below that the first two equations are well-behaved (and indeed solve them) when $\underline{\lambda} = \infty$. In this limit, the third equation holds when $c = 0$. Therefore this must describe the limiting equilibrium with $c = 0$.

We now turn to the characterization of these equations when $\underline{\lambda} \to \infty$. Solve the first two equations for h' and ℓ' in this limit:

$$\ell'(\rho) = \frac{8 + 7\ell(\rho) + \ell(\rho)^2}{\rho(3+\ell(\rho))} \text{ and } \frac{h'(\rho)}{h(\rho)} = \frac{-4}{\rho(3+\ell(\rho))}, \tag{41}$$

with terminal conditions are $\ell(1) = 0$ and $h(1) = 1$. An equilibrium allocation in the limiting economy is characterized by the solution to these to ordinary differential equations.

The differential equation for ℓ is autonomous and so we start with that one. To solve it, define the inverse functions $f(\ell(\rho)) = \rho$ and rewrite the differential equation as

$$\frac{f'(\ell)}{f(\ell)} = \frac{(3+\ell)}{8 + 7\ell + \ell^2}$$

with terminal condition $f(0) = 1$. The solution is

$$f(\ell) = \sqrt{1 + \tfrac{7}{8}\ell + \tfrac{1}{8}\ell^2} \left(\frac{16 + (7 - \sqrt{17})\ell}{16 + (7 + \sqrt{17})\ell} \right)^{\frac{1}{2\sqrt{17}}} = \rho. \tag{42}$$

This is a continuous, increasing mapping from positive reals to positive reals. That implies that the mapping is invertible, i.e. for any $\ell(\rho) \geq 0$, there is a unique value of $\rho \geq 1$ that is consistent with the equation of interest. This implies $\ell(\rho)$ is a well-behaved function.

Next turn to the equation for $h(\rho)$. To solve it, define $\eta(\ell) \equiv h(f(\ell))$. Using the same inverse function and its derivative, the solution is

$$\begin{aligned} \eta(\ell) &= \exp\left(- \int_0^\ell \frac{4f'(x)}{f(x)(3+x)} dx \right) \\ &= \exp\left(- \int_0^\ell \frac{4}{8 + 7\ell' + \ell'^2} d\ell' \right) \\ &= \left(\frac{16 + (7 - \sqrt{17})\ell}{16 + (7 + \sqrt{17})\ell} \right)^{4/\sqrt{17}}, \end{aligned} \tag{43}$$

where the second equation replaces f using equation (42) and the third calculates the integral. Now since $\ell(\rho)$ increases without bound when ρ is large, we get

$$\lim_{\rho \to \infty} h(\rho) = \lim_{\ell \to \infty} \eta(\ell) = \left(\frac{7 - \sqrt{17}}{7 + \sqrt{17}} \right)^{4/\sqrt{17}} = 0.269.$$

This is the probability of contacting a middleman in the zero cost limit, the first result in the Proposition.

We turn next to volume, which we measure as the rate that a trader buys the asset (and hence also the rate that a trader sells the asset). In an economy with finite $\underline{\lambda}$, we write this as

$$\mathcal{V} \equiv \int_{\underline{\lambda}}^{\infty} \frac{\lambda}{4} (L(\lambda) + H(\lambda) m_\lambda) dG(\lambda) + \frac{\Lambda H(\infty) L(\infty)}{4} + \frac{\Lambda H(\infty)^2 m_\infty^2}{4}, \tag{44}$$

where $L(\infty) \equiv \lim_{\lambda \to \infty} L(\lambda)$, $H(\infty) \equiv \lim_{\lambda \to \infty} H(\lambda)$, and m_∞ is the misalignment rate of middlemen. The integrand in the first term is the rate that a finite trader with contact rate λ purchases the asset. This happens when the trader does not hold the asset (probability $\frac{1}{2}$) and meets someone (at rate λ), and either (i) the other trader is slower, hold the asset, but doesn't want it (probability $\frac{1}{2}L(\lambda)$) or (ii) the other trader is faster (probability $H(\lambda)$ and the type-λ trader wants the asset (probability m_∞). Integrating this over the contact rate distribution gives the rate that finite traders buy the asset. The second term is the

rate that middlemen buy the asset from finite traders. Middlemen make contacts at rate $\Lambda H(\infty)$. With probability $\frac{1}{2}L(\infty)$, the contact is with a finite trader who wants to sell the asst, and then with probability $\frac{1}{2}$ the middleman is able to buy it. The final term is the rate that middlemen buy the asset from other middlemen. Again, middlemen make contacts at rate $\Lambda H(\infty)$, with probability $\frac{1}{2}m_\infty$, the middleman wants to buy the asset, and with probability $\frac{1}{2}H(\infty)m_\infty$, she meets another middleman who wants to sell the asset. Summing these terms gives us trading volume.

We cannot directly evaluate this expression in the limiting economy because L, H, and G are all poorly behaved. Instead, define $\hat{\ell}(\rho) \equiv L(\rho\underline{\lambda})\underline{\lambda}$, and let $\Psi(\rho) \equiv G(\rho\underline{\lambda})$ denote the cumulative distribution of relative contact rates. Noting that $m_\lambda = -L'(\lambda)/H'(\lambda)$, we can rewrite volume as

$$\mathcal{V} = \frac{\gamma}{4}\int_1^\infty \left(\ell(\rho) - \frac{\rho\ell'(\rho) - \ell(\rho)}{\rho h'(\rho)}h(\rho)\right) d\Psi(\rho) + \frac{\Lambda h(\infty)\hat{\ell}(\infty)}{4\underline{\lambda}} + \frac{\Lambda h(\infty)^2 m_\infty^2}{4}, \tag{45}$$

where the terms have the exact same interpretation as equation (44). We compute the value of each term in turn.

Start with the integral. To measure this, we first need to characterize the contact rate distribution G in the limiting economy. And as a preliminary step towards that goal, we find the mean contact rate relative to the lower bound $\Lambda/\underline{\lambda}$. For any finite $\underline{\lambda}$, $H'(\lambda) = -\lambda dG(\lambda)/\Lambda$ and also $H'(\lambda) = h'(\rho)/\underline{\lambda}$. Combining these gives

$$\frac{\underline{\lambda}}{\Lambda}d\Psi(\rho) = -\frac{h'(\rho)}{\rho}d\rho, \tag{46}$$

Since $\int_1^\infty d\Psi(\rho) = 1$, we can integrate both sides of this to get

$$\frac{\Lambda}{\underline{\lambda}} = -\frac{1}{\int_1^\infty \frac{h'(\rho)}{\rho}d\rho} = -\frac{1}{\int_0^\infty \frac{\eta'(\ell)}{f(\ell)}d\ell} = \frac{1}{\sqrt{2}}\left(\frac{7+\sqrt{17}}{7-\sqrt{17}}\right)^{\frac{7}{2\sqrt{17}}} \approx 2.23, \tag{47}$$

where again the second equation is a transformation of variables and the third calculates the integral using the functional forms for f and η in equations (42) and (43).

Turn next to the probability distribution over ℓ, a monotonic transformation of the probability distribution over ρ. Its density $\Gamma'(\ell)$ satisfies

$$\Gamma'(\ell) = \Psi'(f(\ell))f'(\ell) = -\frac{\Lambda h'(f(\ell))f'(\ell)}{\underline{\lambda}f(\ell)} = -\frac{\Lambda\eta'(\ell)}{\underline{\lambda}f(\ell)},$$

where the first equation is the definition of Γ, the second uses equation (46), and the third

uses the definition of η. Again using the functional forms for f and η in equations (42) and (43), we can integrate this to get

$$\Gamma(\ell) = \frac{\ell}{\sqrt{8+7\ell+\ell^2}} \left(\frac{2\ell+7+\sqrt{17}}{2\ell+7-\sqrt{17}} \right)^{\frac{7}{2\sqrt{17}}}. \tag{48}$$

The fraction of traders with a contact rate less than ρ is then $\Gamma(\ell(\rho))$.

We return now to the first term in the trading rate (45), the rate that a finite trader buys the asset from either another finite trader or a middleman:

$$\begin{aligned} \frac{\gamma}{4} \int_1^\infty \left(\ell(\rho) - \frac{\rho\ell'(\rho) - \ell(\rho)}{\rho h'(\rho)} h(\rho) \right) d\Psi(\rho) &= \frac{1}{2} \left(1 + \int_1^\infty \ell d\Gamma(\ell) \right) \gamma \\ &= \left(\sqrt{2} \left(\frac{7+\sqrt{17}}{7-\sqrt{17}} \right)^{\frac{7}{2\sqrt{17}}} - 3 \right) \gamma \approx 1.46\gamma \end{aligned} \tag{49}$$

The first equation simplifies the expressions with the differential equations (41) and changes variables to write everything in terms of ℓ. The second equation solves the integral using the known functional form for Γ in equation (48).

Next turn tot the rate that middlemen buy from finite traders. This uses the steady state misalignment rate equation (13b), rewritten in terms of h and $\hat{\ell}$:

$$\big(4\gamma + \rho\big(\underline{\lambda} h(\rho) + 2\hat{\ell}(\rho)\big)\big)\hat{\ell}'(\rho) = -\underline{\lambda}(2\gamma + \rho\hat{\ell}(\rho))h'(\rho),$$

with $h(1) = 1$ and $\hat{\ell}(1) = 0$. Take the limit as $\underline{\lambda} \to \infty$ and solve the differential equation to get

$$\frac{\Lambda h(\rho)\hat{\ell}(\rho)}{4\underline{\lambda}} = \frac{-\gamma\Lambda}{2\underline{\lambda}} \int_1^\rho \frac{h'(\rho')}{\rho'} d\rho' \to \frac{\gamma}{2}, \tag{50}$$

where the limiting result follows immediately from equation (47). This proves, middlemen buy from finite traders at rate $\frac{1}{2}\gamma$.

The same argument implies middlemen sell to finite traders at rate $\frac{1}{2}\gamma$ and an immediate corollary is that finite traders buy from middlemen at the same rate $\frac{1}{2}\gamma$. Subtracting this from equation (49), we get that finite traders buy from finite traders at rate

$$\left(\sqrt{2} \left(\frac{7+\sqrt{17}}{7-\sqrt{17}} \right)^{\frac{7}{2\sqrt{17}}} - \frac{7}{2} \right) \gamma \approx 0.96\gamma,$$

indeed a number between $\frac{1}{2}\gamma$ and γ.

The last step of the proof is finding the rate that middlemen trade with other middlemen. To find this, we need to compute the misalignment rate of middlemen. We start in the case where $\underline{\lambda}$ is finite. Middlemen meet other traders at an infinite rate, and so the key is to compute the fraction of those meetings that switch their alignment status in either direction. First, if they are misaligned, they trade in meetings where the other trader is a misaligned, finite speed trader (probability $L(\infty)$) or in meetings where the other trader is a misaligned middleman (probability $H(\infty)m_\infty$). If they are well-aligned, they trade only in meetings where the other trader is a misaligned, finite speed trader. Thus m_∞ solves

$$\left(L(\infty) + H(\infty)m_\infty\right) m_\infty = L(\infty)(1 - m_\infty).$$

Middlemen's own preference shocks do not appear in this expression because middlemen have infinitely more trading opportunities than preference shocks. Again rewrite this in terms of h and $\hat{l}$ to get

$$\left(\frac{\hat{\ell}(\infty)}{\underline{\lambda}} + h(\infty)m_\infty\right) m_\infty = \frac{\hat{\ell}(\infty)}{\underline{\lambda}}(1 - m_\infty).$$

Multiplying both sides of the equation by $\Lambda h(\infty)/4$, the previous equation implies

$$\frac{\Lambda h(\infty)\hat{\ell}(\infty)}{4\underline{\lambda}} - \frac{\Lambda h(\infty)^2 m_\infty^2}{4} = \frac{\Lambda h(\infty)\hat{\ell}(\infty)}{2\underline{\lambda}} m_\infty.$$

Now in the limit as $\underline{\lambda}$ converges to infinity, equation (50) showed that $\frac{\Lambda h(\rho)\hat{\ell}(\rho)}{2\underline{\lambda}} \to \gamma$, while $m_\infty \to 0$. Thus the right hand side converges to 0 and so must the left hand side. Since again $\frac{\Lambda h(\infty)\hat{\ell}(\infty)}{4\underline{\lambda}} \to \frac{1}{2}\gamma$, this implies $\frac{\Lambda h(\infty)^2 m_\infty^2}{4}$ has the same limit. Thus middlemen buy from middlemen at rate $\frac{1}{2}\gamma$. This completes the proof. ■

Proof of Proposition 8. We start by solving the problem of maximizing (14) subject to the constraint (15) for all λ. We do this by writing the Lagrangian, placing a multiplier $S_\lambda dG(\lambda)$ on constraint (15), a multiplier θ_0 on the constraint that $\int_0^\infty dG(\lambda) = 1$, and a multiplier θ_1 on the constraint that $\mathbb{E}(1) = 1$. The Lagrangian is

$$\begin{aligned}\mathcal{L} = \Delta \int_0^\infty (1 - m_\lambda) dG(\lambda) - \Lambda \mathbb{E}(c(\lambda)) + \theta_0 \left(1 - \int_0^\infty dG(\lambda)\right) + \theta_1 \left(1 - \mathbb{E}(1)\right) \\ + \int_0^\infty S_\lambda \Bigg(\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,0}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,0}^{\lambda',1}(1 - m_{\lambda'})\right)\right) m_\lambda \\ - \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,1}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,1}^{\lambda',1}(1 - m_{\lambda'})\right)\right)(1 - m_\lambda)\Bigg) dG(\lambda). \quad (51)\end{aligned}$$

Take the first order condition with respect to m_λ. Suppressing the multiplicative constant $dG(\lambda)$, this is

$$\Delta = 2\gamma S_\lambda + \frac{\lambda}{2}\mathbb{E}\Big(\big(\mathbf{1}_{\lambda,0}^{\lambda',0}(S_\lambda + S_{\lambda'}) + \mathbf{1}_{\lambda,1}^{\lambda',0}(S_\lambda - S_{\lambda'})\big)m_{\lambda'} + \big(\mathbf{1}_{\lambda,0}^{\lambda',1}(S_\lambda - S_{\lambda'}) + \mathbf{1}_{\lambda,1}^{\lambda',1}(S_\lambda + S_{\lambda'})\big)(1 - m_{\lambda'})\Big) \quad (52)$$

Next take the first order conditions for the indicator functions $\mathbf{1}_{\lambda,a}^{\lambda',a'}$. In doing this, we implicitly use the unstated constraints $0 \leq \mathbf{1}_{\lambda,a}^{\lambda',a'} = \mathbf{1}_{\lambda',a'}^{\lambda,a} \leq 1$. We get

$$S_\lambda + S_{\lambda'} \gtrless 0 \Rightarrow \mathbf{1}_{\lambda,0}^{\lambda',0} = \begin{cases} 1 \\ 0 \end{cases}, \qquad S_\lambda \gtrless S_{\lambda'} \Rightarrow \mathbf{1}_{\lambda,0}^{\lambda',1} = \begin{cases} 1 \\ 0 \end{cases},$$

$$S_\lambda + S_{\lambda'} \lessgtr 0 \Rightarrow \mathbf{1}_{\lambda,1}^{\lambda',1} = \begin{cases} 1 \\ 0 \end{cases}, \qquad S_{\lambda'} \gtrless S_\lambda \Rightarrow \mathbf{1}_{\lambda,1}^{\lambda',0} = \begin{cases} 1 \\ 0 \end{cases}.$$

Using this, rewrite the first order condition (52) as

$$\Delta = 2\gamma S_\lambda + \frac{\lambda}{2}\mathbb{E}\Big((S_\lambda + S_{\lambda'})^+ - (S_{\lambda'} - S_\lambda)^+\big)m_{\lambda'} + \big((S_\lambda - S_{\lambda'})^+ - (-S_\lambda - S_{\lambda'})^+\big)(1 - m_{\lambda'})\Big) \quad (53)$$

This is identical to equation (8) for the surplus in the decentralized economy, except that the terms multiplying λ are twice as large for the planner. The proof of Proposition 1 implies the surplus function is uniquely defined by this equation and moreover is decreasing and nonnegative. Optimal trading patterns (the planner's version of Proposition 1) follow immediately.

Next, using monotonicity of S, rewrite equation (53) as

$$\big(4\gamma + \lambda\mathbb{E}\big(1 - \mathbb{I}_{\lambda' \leq \lambda}(1 - 2m_{\lambda'})\big)\big)\, S_\lambda = 2\Delta + \lambda\mathbb{E}\Big(\mathbb{I}_{\lambda' > \lambda} S_{\lambda'}(1 - 2m_{\lambda'})\Big) \quad (54)$$

Differentiate with respect to λ to get

$$\Big(4\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)S'_\lambda + (H(\lambda) + 2L(\lambda))S_\lambda = \mathbb{E}\big(\mathbb{I}_{\lambda' > \lambda} S_{\lambda'}(1 - 2m_{\lambda'})\big) \quad (55)$$

Replace the right hand side using equation (54) and simplify to get

$$S'_\lambda = \frac{4\gamma S_\lambda - 2\Delta}{\lambda\Big(4\gamma + \lambda(H(\lambda) + 2L(\lambda))\Big)} = \Phi_\lambda \left(S_\lambda - \frac{\Delta}{2\gamma}\right),$$

where Φ_λ is given in equation (17). The general solution to this differential equation is

$$S_\lambda = \frac{\Delta}{2\gamma} - Ke^{\int_{\underline{\lambda}}^{\lambda} \Phi_{\lambda'} d\lambda'}$$

for fixed $\underline{\lambda}$ and some constant of integration K. Use the fact that $\lim_{\lambda\to\infty} S_\lambda = 0$ to pin down the constant of integration, equation (16).

Now return to the Lagrangian (51). The first order condition with respect to $dG(\lambda)$ implies $dG(\lambda) > 0$ only if λ maximizes

$$\begin{aligned}
&\Delta(1 - m_\lambda) - \lambda c(\lambda) - \theta_0 - \theta_1 \frac{\lambda}{\Lambda} \\
&+\frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}^{\lambda',0}_{\lambda,0} m_{\lambda'} S_{\lambda'} m_\lambda + \mathbf{1}^{\lambda',0}_{\lambda,1} m_{\lambda'} S_{\lambda'}(1 - m_\lambda) - \mathbf{1}^{\lambda',1}_{\lambda,0} S_{\lambda'}(1 - m_{\lambda'}) m_\lambda - \mathbf{1}^{\lambda',1}_{\lambda,1} S_{\lambda'}(1 - m_{\lambda'})(1 - m_\lambda)\right).
\end{aligned}$$

Using the optimal trading patterns, this reduces to

$$\begin{aligned}
&\Delta(1 - m_\lambda) - \lambda c(\lambda) - \theta_0 - \theta_1 \frac{\lambda}{\Lambda} \\
&\qquad + \frac{\lambda}{2} m_\lambda \mathbb{E}\left(m_{\lambda'} S_{\lambda'}\right) + \frac{\lambda}{2}(1 - m_\lambda)\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'} S_{\lambda'}\right) - \frac{\lambda}{2} m_\lambda \mathbb{E}\left(\mathbb{I}_{\lambda'>\lambda} S_{\lambda'}(1 - m_{\lambda'})\right).
\end{aligned}$$

Multiply equation (54) by $\frac{1}{2} m_\lambda$ and add to the previous expression. $dG(\lambda) > 0$ only if λ maximizes

$$\Delta + \frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'} S_{\lambda'}\right) - \left(2\gamma + \frac{\lambda}{2}\mathbb{E}\big(1 - \mathbb{I}_{\lambda'<\lambda}(1 - 2m_{\lambda'})\big)\right) S_\lambda m_\lambda - \lambda c(\lambda) - \theta_0 - \theta_1 \frac{\lambda}{\Lambda}.$$

Using equation (15) under the optimal trading pattern and dropping irrelevant constants, this simplifies to

$$dG(\lambda) > 0 \Rightarrow \lambda \in \arg\max \left(-\gamma S_\lambda + \frac{\lambda}{2}\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'}(S_{\lambda'} - S_\lambda)\big) - \lambda c(\lambda) - \theta_1 \frac{\lambda}{\Lambda}\right) \tag{56}$$

Next, consider the first order condition with respect to Λ, which appears implicitly inside

each of the expectations operators. Using the efficient trading patterns, this gives

$$\begin{aligned}\theta_1 &= \int_0^\infty S_\lambda \left(\left(\frac{\lambda}{2} \mathbb{E}\left(m_{\lambda'} + \mathbb{I}_{\lambda'>\lambda}(1 - m_{\lambda'}) \right) \right) m_\lambda - \left(\frac{\lambda}{2} \mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'} \right) \right) (1 - m_\lambda) \right) dG(\lambda) \\ &= \gamma \int_0^\infty S_\lambda (1 - 2m_\lambda) dG(\lambda) \end{aligned} \tag{57}$$

where the second line again uses equation (15) under the efficient trading pattern. Substitute this into equation (56) to get that $dG(\lambda) > 0$ only if λ maximizes the expression in (18).

Finally, consider the behavior of (18) for large λ. If the inequality (19) is violated, the planner would set λ unboundedly large for everyone. This incurs infinite costs and so cannot be optimal. If the inequality (19) is slack, large values of λ would be inconsistent with condition (18), which implies $\Lambda = \int_0^\infty \lambda dG(\lambda)$. To have middlemen, the inequality (19) must be binding.

Once we have established the close link between the mathematical structures of the equilibrium and optimal allocations and surplus functions, it is straightforward to replicate the proofs of Propositions 2–6. Here we present one key piece of the result, the analog to equation (13a), which is critical for characterizing the optimal allocation when the cost function is linear.

For any value of λ in the support of G, equation (56) states that the planner marginal values must be the same, some $\bar{V}$:

$$\bar{V} = -\gamma S_\lambda + \frac{\lambda}{2} \mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'} (S_{\lambda'} - S_\lambda) \big) - \lambda c(\lambda) - \theta_1 \frac{\lambda}{\Lambda}, \tag{58}$$

where we simplify the expression slightly using the results from the planner's analog of to Proposition 1. This implies the first order condition

$$\begin{aligned} 0 &= -\left(\gamma + \frac{\lambda}{2} L(\lambda) \right) S'_\lambda + \frac{1}{2} \mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda} (S_{\lambda'} - S_\lambda) m'_\lambda \big) - c(\lambda) - \lambda c'(\lambda) - \frac{\theta_1}{\Lambda} \\ &= -\left(\gamma + \frac{\lambda}{2} L(\lambda) \right) S'_\lambda + \frac{\bar{V} + \gamma S_\lambda}{\lambda} - \lambda c'(\lambda) \\ &= \frac{\Delta \left(2\gamma + \lambda L(\lambda) \right) e^{-\int_\lambda^\infty \Phi_{\lambda'} d\lambda'}}{\lambda \Big(4\gamma + \lambda (H(\lambda) + 2L(\lambda)) \Big)} + \frac{\bar{V} + \frac{\Delta}{2} \left(1 - e^{-\int_\lambda^\infty \Phi_{\lambda'} d\lambda'} \right)}{\lambda} - \lambda c'(\lambda) \\ &= \frac{-(\delta_1 - \delta_0) H(\lambda) e^{-\int_\lambda^\infty \Phi_{\lambda'} d\lambda'}}{2 \Big(4\gamma + \lambda (H(\lambda) + 2L(\lambda)) \Big)} + \frac{\bar{V} + \frac{\delta_1 - \delta_0}{2} - \lambda^2 c'(\lambda)}{\lambda} \end{aligned} \tag{59}$$

where the second line simplifies the first using equation (58), the third line replaces S_λ and

its derivative using equation (16) and replaces Φ_λ using equation (17), and the fourth line groups terms and recognizes that $\Delta = \delta_1 - \delta_0$. Rewrite this as

$$\frac{\lambda H(\lambda) e^{-\int_\lambda^\infty \Phi_{\lambda'} d\lambda'}}{4\gamma + \lambda(H(\lambda) + 2L(\lambda))} + \frac{2\lambda^2 c'(\lambda)}{\delta_1 - \delta_0} = \frac{2\bar{V} + \delta_1 - \delta_0}{\delta_1 - \delta_0} \tag{60}$$

for all λ in the support of $G(\lambda)$. Differentiating this expression and replacing Φ_λ gives us an second equation relating $H'(\lambda)$ and $L'(\lambda)$:

$$\frac{2e^{-\int_\lambda^\infty \Phi_{\lambda'} d\lambda'} \big(\lambda(2\gamma + \lambda L(\lambda))H'(\lambda) + H(\lambda)(4\gamma - \lambda^2 L'(\lambda))\big)}{\big(4\gamma + \lambda(H(\lambda) + 2L(\lambda))\big)^2} + \frac{2\lambda}{\Delta}(2c'(\lambda) + \lambda c''(\lambda)) = 0. \tag{61}$$

For the case of linear cost, this reduces to equation (20).

Next, we consider the limit of the optimal allocation as $c \to 0$, i.e. the planner's analog to Proposition 7. We start with some preliminaries. First, with a linear cost function, $c(\lambda) = c$, the Lagrangian (51) can be simplified as

$$\begin{aligned}\mathcal{L} = \Delta \int_0^\infty (1 - m_\lambda) dG(\lambda) - c\Lambda + \theta_0 \left(1 - \int_0^\infty dG(\lambda)\right) \\ + \int_0^\infty S_\lambda \Bigg(\left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,0}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,0}^{\lambda',1}(1 - m_{\lambda'})\right)\right) m_\lambda \\ - \left(\gamma + \frac{\lambda}{2}\mathbb{E}\left(\mathbf{1}_{\lambda,1}^{\lambda',0} m_{\lambda'} + \mathbf{1}_{\lambda,1}^{\lambda',1}(1 - m_{\lambda'})\right)\right)(1 - m_\lambda)\Bigg) dG(\lambda).\end{aligned}$$

The first order condition with respect to Λ implies

$$\begin{aligned} c &= \frac{1}{\Lambda}\int_0^\infty S_\lambda \Bigg(\left(\frac{\lambda}{2}\mathbb{E}\left(m_{\lambda'} + \mathbb{I}_{\lambda'>\lambda}(1 - m_{\lambda'})\right)\right) m_\lambda - \left(\frac{\lambda}{2}\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda} m_{\lambda'}\right)\right)(1 - m_\lambda)\Bigg) dG(\lambda) \\ &= \frac{\gamma}{\Lambda}\int_0^\infty S_\lambda (1 - 2m_\lambda) dG(\lambda), \end{aligned} \tag{62}$$

where the second line uses equation (15) under the efficient trading pattern. Comparing this with equation (57), we get

$$\theta_1 = \Lambda c \tag{63}$$

Second, evaluate equation (59) at $\lambda = \underline{\lambda}$, eliminate $S'_{\underline{\lambda}}$ using equations (16) and (17), and θ_1 using equation (63). This gives

$$\frac{c}{\Delta} = \frac{\gamma \exp\left(-\int_{\underline{\lambda}}^\infty \frac{4\gamma}{\lambda(4\gamma + \lambda H(\lambda) + 2\lambda L(\lambda))} d\lambda\right)}{\underline{\lambda}(4\gamma + \underline{\lambda})}. \tag{64}$$

The functions H and L satisfy the differential equations (13b) and (20) with boundary conditions $H(\underline{\lambda}) = 1$ and $L(\underline{\lambda}) = 0$.

As in equilibrium, we can prove that the optimal allocation must have $\underline{\lambda} \to \infty$ as $c \to 0$. To handle this, we do the same transformation of variables as in equilibrium, letting $h(\rho) \equiv H(\lambda)$ and $\ell(\rho) \equiv L(\lambda)\lambda/\gamma$ where $\rho \equiv \lambda/\underline{\lambda}$. Rewrite equations (13b), (20) and (64) in terms of ℓ and h:

$$\begin{aligned}
\big(4\gamma + \underline{\lambda}\rho h(\rho) + 2\gamma\ell(\rho)\big)(\rho\ell'(\rho) - \ell(\rho)) &= -\underline{\lambda}\rho^2(2 + \ell(\rho))h'(\rho), \\
2(2h(\rho) + \rho h'(\rho)) &= \rho\big(h(\rho)\ell'(\rho) - \ell(\rho)h'(\rho)\big) - h(\rho)\ell(\rho), \\
\frac{c}{\Delta} &= \frac{\gamma \exp\left(-\int_1^\infty \frac{4\gamma}{\rho(4\gamma + \rho\underline{\lambda}h(\rho) + 2\gamma\ell(\rho))}d\rho\right)}{\underline{\lambda}(4\gamma + \underline{\lambda})}.
\end{aligned}$$

with $h(1) = 1$ and $\ell(1) = 0$. Next, solve for $\ell'(\rho)$ and $h'(\rho)$ in the limit with $\underline{\lambda} \to \infty$:

$$\ell'(\rho) = \frac{2 + \ell(\rho)}{\rho} \text{ and } h'(\rho) = \frac{-2h(\rho)}{\rho(2 + \ell(\rho))}.$$

These differential equations can easily be solved in closed form:

$$\ell(\rho) = 2(\rho - 1) \text{ and } h(\rho) = e^{\rho^{-1}-1}. \tag{65}$$

Using this, we can calculate the fraction of middlemen as

$$\lim_{\rho\to\infty} h(\rho) = e^{-1} \approx 0.368.$$

This is the optimal allocation analog of one of the key results in Proposition 7.

Next, let $\Psi(\rho) \equiv G(\rho\underline{\lambda})$ denote the cumulative distribution of relative contact rates. Using equation (46) and $\int_1^\infty d\Psi(\rho) = 1$, this implies

$$\frac{\underline{\lambda}}{\Lambda} = -\int_1^\infty \frac{h'(\rho)}{\rho}d\rho = e^{-1},$$

so $\Lambda/\underline{\lambda} = e \approx 2.72$ in the economy without search frictions. Then substitute this back into equation (46) gives

$$\Psi'(\rho) = \rho^{-3}e^{\rho^{-1}} \Rightarrow \Psi(\rho) = (1 - \rho^{-1})e^{\rho^{-1}}, \tag{66}$$

where the result follows by integrating the density function. This is an explicit solution for the distribution of relative contact rates in the limiting economy.

Finally, we turn to volume. We start with the rate that finite traders buy, the first term

in equation (45). We use the transformation in terms of h, ℓ, and Ψ to get

$$\frac{\gamma}{4}\int_1^\infty \left(\ell(\rho) - \frac{\rho\ell'(\rho)-\ell(\rho)}{\rho h'(\rho)}h(\rho)\right) d\Psi(\rho) = \frac{\gamma}{2}\int_1^\infty (2\rho-1)\rho^{-3}e^{\rho^{-1}}d\rho = (e-3/2)\gamma \approx 1.22\gamma.$$

This compares to equation (49) in equilibrium. Finally, the buying rate of middlemen is unchanged. This proves that finite traders buy from finite traders at rate $(e-2)\gamma \approx 0.72\gamma \in (\frac{1}{2}\gamma, \gamma)$, they buy from middlemen at rate $\frac{1}{2}\gamma$, and middlemen buy from finite traders at rate $\frac{1}{2}\gamma$, and sell to middlemen at rate $\frac{1}{2}\gamma$. ■

Proof of Proposition 9. First, the definition of $\bar{\tau}$ and equation (57) implies $\bar{\tau} = \theta_1/\Lambda$. Then equation (63) implies $\bar{\tau} = c$.

We next show that the expected subsidy likewise equals c. Denote the expected subsidy for an agent with contact rate λ by $\tau^+(\lambda)$. This is given by

$$\begin{aligned}\tau^+(\lambda) = \frac{1}{4}\Big(&(1-m_\lambda)\,\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda}\,(S_{\lambda'}-S_\lambda)\,m_{\lambda'}\big) + m_\lambda\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda}\,(S_{\lambda'}+S_\lambda)\,m_{\lambda'}\big)\\ &+ m_\lambda\mathbb{E}\big(\mathbb{I}_{\lambda'>\lambda}\,(S_\lambda - S_{\lambda'})\,(1-m_{\lambda'})\,\big) + m_\lambda\mathbb{E}\big(\mathbb{I}_{\lambda'>\lambda}\,(S_\lambda+S_{\lambda'})\,m_{\lambda'}\big)\Big).\end{aligned}$$

Rearrange the second line to get

$$\begin{aligned}\tau^+(\lambda) = \frac{1}{4}\Big(&(1-m_\lambda)\,\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda}\,(S_{\lambda'}-S_\lambda)\,m_{\lambda'}\big) + m_\lambda\mathbb{E}\big(\mathbb{I}_{\lambda'<\lambda}\,(S_{\lambda'}-S_\lambda)\,m_{\lambda'}\big)\\ &+ 2m_\lambda S_\lambda L(\lambda) + m_\lambda S_\lambda H(\lambda) - m_\lambda\mathbb{E}\big(\mathbb{I}_{\lambda'>\lambda}S_{\lambda'}\,(1-2m_{\lambda'})\,\big)\Big).\end{aligned}$$

Differentiating the expression in (56) with respect to λ, we get that

$$\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda}\,(S_{\lambda'}-S_\lambda)\,m_{\lambda'}\right) = 2\left(c + \frac{\theta_1}{\Lambda} + \left(\gamma + \frac{\lambda}{2}L(\lambda)\right)S'_\lambda\right)$$

It follows that

$$\begin{aligned}\tau^+(\lambda) =& \frac{1}{4}\Big(2\left(c+\frac{\theta_1}{\Lambda}\right) + (2\gamma+\lambda L(\lambda))\,S'_\lambda + m_\lambda S_\lambda\,(2L(\lambda)+H(\lambda))\\ &- m_\lambda\Big(4\gamma+\lambda(H(\lambda)+2L(\lambda))\Big)S'_\lambda - m_\lambda S_\lambda(H(\lambda)+2L(\lambda))\Big)\\ =& \frac{1}{2}\left(c+\frac{\theta_1}{\Lambda}\right) + \frac{1}{4}S'_\lambda\Big((1-2m_\lambda)\,(2\gamma+\lambda L(\lambda)) - \lambda m_\lambda H(\lambda)\Big)\\ =& \frac{1}{2}\left(c+\frac{\theta_1}{\Lambda}\right) = c\end{aligned}$$

where the first equality uses the expression for $\mathbb{E}\left(\mathbb{I}_{\lambda'<\lambda}\left(S_{\lambda'}-S_\lambda\right)m_{\lambda'}\right)$ and equation (55), the second regroups terms, and the third uses the inflow-outflow equation (12). The final equality follows from equating equations (57) and (62). This proves the result. ■

Proof of Proposition 10. We begin with equilibrium. Condition 3b in definition 2 immediately implies that there are no middlemen whenever their per-meeting cost is strictly positive. Solve equation (22) to get

$$s_\lambda = \frac{4\Delta - \lambda\mathbb{E}\left(s'_\lambda m'_\lambda\right)}{8\gamma + \lambda\mathbb{E}(m'_\lambda)}, \tag{67}$$

a decreasing and convex function. Condition 3a then implies that if $\lambda c(\lambda)$ is weakly concave, all traders choose the same value of $\lambda = \Lambda$. That is, $\mathbb{E}\left(s'_\lambda m'_\lambda\right) = s_\Lambda m_\Lambda$ and $\mathbb{E}(m'_\lambda) = m_\Lambda$.

To find the equilibrium contact rate, we solve explicitly for s_λ. First, simplify (67) when $\lambda = \Lambda$:

$$s_\Lambda = \frac{2\Delta}{4\gamma + \Lambda m_\Lambda}.$$

Then rewrite the expression for a general value of λ:

$$s_\lambda = \frac{\Delta\left(16\gamma + 4\Lambda m_\Lambda - 2\lambda m_\Lambda\right)}{(4\gamma + \Lambda m_\Lambda)(8\gamma + \lambda m_\Lambda)}. \tag{68}$$

Finally, using condition 3a, it follows that if $\lambda c(\lambda)$ is weakly convex, the equilibrium choice of Λ solves

$$\frac{4\gamma\Delta m_\Lambda}{(4\gamma + \Lambda m_\Lambda)(8\gamma + \Lambda m_\Lambda)} = c(\Lambda) + \Lambda c'(\Lambda). \tag{69}$$

We next turn to the planner's problem. As in equilibrium, it is straightforward to show that only meetings between two misaligned traders result in trade. Replicating the proof of Proposition 8, we get that the optimal surplus function satisfies

$$\Delta = 2\gamma S_\lambda + \frac{\lambda}{2}\mathbb{E}\left((S_\lambda + S_{\lambda'})m_{\lambda'}\right) \Rightarrow S_\lambda = \frac{2\Delta - \lambda\mathbb{E}\left(S_{\lambda'}m_{\lambda'}\right)}{4\gamma + \lambda\mathbb{E}(m_{\lambda'})},$$

decreasing and convex. We also obtain that the planner has $dG(\lambda) > 0$ only if λ maximizes

$$-\gamma S_\lambda - \lambda\left(c(\lambda) + \frac{\gamma}{\Lambda}\int_0^\infty S_{\lambda'}(1 - 2m_{\lambda'})dG(\lambda')\right),$$

analogous to condition (18). Convexity of S implies that if the cost function is convex, the planner places all weight on a single value of λ:

Replicating the arguments for the equilibrium, we then use this to prove that

$$S_\lambda = \frac{\Delta\big(4\gamma + 2\Lambda m_\Lambda - \lambda m_\Lambda\big)}{(4\gamma + \lambda m_\Lambda)(2\gamma + \Lambda m_\Lambda)}. \tag{70}$$

It follows that the optimal choice of Λ satisfies the first order condition

$$\frac{\gamma \Delta m_\Lambda}{(2\gamma + \Lambda m_\Lambda)(4\gamma + \Lambda m_\Lambda)} = c(\Lambda) + \Lambda c'(\Lambda), \tag{71}$$

where we simplify the expression slightly using the steady state relationship (23).

To prove that the equilibrium contact rate is inefficiently high, use the steady state misalignment rate to eliminate Λ from the left hand side of equations (69) and (71). This gives, respectively,

$$\frac{\Delta m_\Lambda^3}{\gamma(1 + 2m_\Lambda)} = c(\Lambda) + \Lambda c'(\Lambda)$$
$$\frac{\Delta m_\Lambda^3}{4\gamma(1 - m_\Lambda)} = c(\Lambda) + \Lambda c'(\Lambda)$$

For any $\Lambda > 0$, $0 \leq m_\Lambda < 1/2$ and therefore $1 + 2m_\Lambda < 4(1 - m_\lambda)$ It follows that the left hand side of equation (69) is always bigger than the left hand side of equation (71) at a fixed value of $\Lambda > 0$. Since both left hand sides are increasing in m_Λ, it follows that the equilibrium misalignment rate must be weakly lower than the optimum, strictly so if the equilibrium rate is less than $1/2$ and the cost function is continuously differentiable. Again using the steady state misalignment equation, the same condition ensures that the equilibrium contact rate weakly exceeds the optimum whenever the rate is positive. ■