Review of Economic Studies (2022) **0**, 1–32 https://doi.org/10.1093/restud/rdac014

Advance access publication 2 March 2022

The Emergence of Market Structure

MARYAM FARBOODI
MIT
GREGOR JAROSCH
Princeton University
and
ROBERT SHIMER
University of Chicago

First version received May 2020*; Editorial decision October* 2021*; Accepted February* 2022 (*Eds.*)

We study a model of over-the-counter trading in which *ex ante* identical traders invest in a contact technology and participate in bilateral trade. We show that a rich market structure emerges both in equilibrium and in an optimal allocation. There is continuous heterogeneity in market access under weak regularity conditions. If the cost per contact is constant, heterogeneity is governed by a power law and there are middlemen, market participants with unboundedly high contact rates who account for a positive fraction of meetings. Externalities lead to overinvestment in equilibrium, and policies that reduce investment in the contact technology can improve welfare. We relate our findings to important features of real-world trading networks.

Key words: Over-the-counter markets; Intermediation; Middlemen; Random matching; Endogenous search intensity; Network formation; Pareto distribution; Welfare.

JEL Codes: D83, D85, G11, G12

1. INTRODUCTION

In many over-the-counter markets, some participants trade much more frequently and with many more partners than others do (see e.g. Bech and Atalay, 2010; Craig and Von Peter, 2014; Fricke and Lux, 2015; Hollifield, Neklyudov and Spatt, 2017). We are interested in understanding why such a market structure exists and in assessing its normative properties.

We do this by examining an over-the-counter market for assets where market participants periodically meet in pairs with the opportunity to trade (Rubinstein and Wolinsky, 1987). We assume that *ex ante* identical market participants make a costly investment which governs how often they contact others. Whenever two participants are in contact, they may trade an asset for an outside good, as in Duffie, Gârleanu and Pedersen (2005) and a large ensuing literature. We find that a market structure with rich heterogeneity emerges naturally in this environment,

The editor in charge of this paper was Adam Szeidl.

creating specialized intermediaries who mitigate trading frictions while economizing on aggregate investment costs.

We verify that if market participants have heterogeneous contact rates, intermediation arises naturally and participants who have more contacts act as intermediaries (Üslü, 2019). By intermediation, we mean that market participants with more contacts buy assets with the sole intent to quickly resell them and sell assets with the sole intent to quickly repurchase them. While there are no static efficiency gains from intermediation, these trades move assets towards traders with more contact opportunities, which improves the economy's future trading opportunities.

Our novel contribution is that we treat participants' contact rates as an investment choice and characterize the resulting market structure in a decentralized equilibrium and a Pareto optimal allocation. In equilibrium, market participants choose their contact rate to maximize future profits net of the investment cost. In an optimal allocation, the contact rate distribution maximizes the social value of the trade net of the investment cost. Although the two allocations are qualitatively similar, we find that equilibrium is inefficient due to congestion and thick-market externalities.

We highlight several key results on the market structure in both the equilibrium and optimal allocations. First, we prove that heterogeneity in contact rates arises endogenously, with *ex ante* homogeneous participants making heterogeneous investments. Specifically, we show that *any* distribution of contact rates can be rationalized through an appropriate investment cost function (Propositions 1 for equilibrium and 1-P for Pareto optimum). Conversely, we show that traders make dispersed investment decisions such that the contact rate distribution is continuous without interior mass when the investment cost function is differentiable (Propositions 2 and 2-P).

The force pushing towards heterogeneity is the gain from intermediation. Each trader benefits both from fundamental and intermediation trades. Assume everyone else chooses the same contact rate. For an individual trader, the gain from fundamental trades is locally linear, while the gain from intermediation is v-shaped around the mass point. In particular, a trader who sets a slightly higher contact rate acts as an intermediary for everyone else. Conversely, a trader who chooses a slightly lower contact rate is intermediated by everyone else. As the value of each intermediation trade is locally linear in the difference in contact rates, the gain to intermediation increases by moving away from the mass point in either direction. This generates a convex kink in the value function at the mass point, which creates an incentive to choose a different contact rate from everyone else.

Second, we characterize the equilibrium and optimal contact rate distribution when there is an exogenous upper bound on contacts, up to which the cost of each contact is constant. If the cost of a contact is neither too high nor too low, we prove that the contact rate distribution is continuous on a convex support extending from a strictly positive endogenous lower bound to the exogenous upper bound (Propositions 3 and 3-P).

We then take the limit of equilibrium and optimal allocations as the exogenous upper bound on contacts goes to infinity. We prove that the limiting distribution of contact rates is Pareto with a tail index of 2 (Propositions 4 and 4-P). Additionally, we call someone a *middleman* if their contact rate is infinite. Although almost no one is a middleman in either an equilibrium or optimal allocation, we prove that middlemen account for a positive fraction of meetings and trades. More precisely, there is a strictly positive probability that a counterparty in any meeting or trade has a contact rate exceeding any finite threshold (still Propositions 4 and 4-P and Corollaries 1 and 1-P).[1]

1. Duffie *et al.* (2005) and the ensuing literature frequently assume the existence of marketmakers with access to a frictionless interdealer market. The middlemen who emerge endogenously in our environment share many features with these marketmakers: other traders stochastically meet middlemen and bargain over the terms of trade, while there

We argue that both the Pareto tail and middlemen connect with evidence. A host of empirical work documents a multi-tiered network structure with a small core and many layers of intermediaries, as well as a power law, namely a Pareto tail in the distribution of the number of trading partners. See, for example, Craig and Von Peter (2014) for the German interbank market and Bech and Atalay (2010) for the federal funds market.

Third, we show that when the cost of each contact is constant but arbitrarily small, the average contact rate is arbitrarily large. Despite this, heterogeneity and intermediation survive within this limit. Each shock to a market participant's idiosyncratic valuation for an asset leads to more than four trades on average as the asset gets reallocated through intermediation chains that typically involve a middleman (Propositions 5 and 5-P). This is consistent with the observation that, despite recent advances in information technology which have reduced search frictions, intermediation remains a hallmark of many decentralized marketplaces (see e.g. Kuprianov, 1993; Philippon, 2015; Biais and Green, 2019).

Finally, we stress that heterogeneity and intermediation are intimately connected. Prior research, e.g., Üslü (2019) and Nekludyov (2019), established that heterogeneity in market access creates a role for intermediation. We show that if intermediation is prohibited, *ex ante* identical market participants choose the same contact rate, both in equilibrium and in the optimum (Proposition 6). That is, without heterogeneity, there is no intermediation, and without intermediation, there is no heterogeneity.

Our main contribution is to endogenize a market structure with rich heterogeneity. After establishing the relevant propositions, we relate the results to an empirical literature that identifies a core of a few highly connected traders, a larger number of peripheral traders that still frequently intermediate, and heterogeneity in market access that is well-described by a Pareto distribution.

Second, our approach offers a natural way to predict the endogenous response of market structure to technological change, such as a reduction in the cost of contacts. We capture this most clearly in our limiting economy where the cost of each contact is arbitrarily small. This allows us to speak to the increasing prominence of financial intermediation, including in decentralized asset markets, despite improving information technologies.

Our approach also offers a direct assessment of the efficiency of endogenous market structure and how those vary with technological parameters. For instance, we show formally that the equilibrium market structure is inefficient. Using Pigouvian taxes, we show that optimal policy taxes frequent traders more per trade than peripheral ones. Numerically, we show that trading volume is excessive in equilibrium and connect these observations with the discussion on financial transaction taxes and regulatory intervention (Tobin, 1978; Burman, Gale, Gault, Kim, Nunns and Rosenthal, 2016).

The rest of the article is organized as follows. Section 2 reviews the related literature. Section 3 describes our model. Section 4 defines equilibrium while Sections 5 and 6 characterize it. Section 7 analyses the Pareto optimal allocation and the nature of search externalities. Section 8 establishes that contact rate dispersion disappears if intermediation is prohibited. Section 9 concludes.

2. RELATED WORK

Our article develops a stochastic network formation model in a frictional trading context. As such, it is related to both the large literature on trade and intermediation in frictional asset markets and

is continuous trade among middlemen. One difference is that our middlemen have the same preferences as other traders and in particular care about their asset position. In contrast, the literature following Rubinstein and Wolinsky (1987) and Duffie *et al.* (2005) assumes that marketmakers only value their trading profits.

to stochastic network formation models in other settings such as disease transmission or social networks. We will review each in turn.

Rubinstein and Wolinsky (1987) were the first to model intermediation in a frictional goods market. We share with them the notion that intermediaries have access to a superior search technology. In two important papers, Duffie *et al.* (2005) and Duffie, Gârleanu and Pedersen (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 the Duffie *et al.* framework to accommodate newly available evidence on trade and intermediation in over-the-counter markets; see Weill (2020) for a recent survey. Üslü (2019) allows for rich heterogeneity in contact rates, pricing, and inventory holdings in a market where traders have continuously distributed flow payoffs.[2] As in our framework, fast dealers are more willing to take on misaligned asset positions, thus emerging as intermediaries. The marketplace features intermediation chains and a core–periphery trading network. Our contribution to this literature is to show that heterogeneous contact rates arise endogenously to leverage the gains from intermediation even with *ex ante* homogeneous traders. We further show how the endogenous choice of contact rates given a cost function disciplines key features of the contact rate distribution. Additionally, our normative analysis shows that both heterogeneous investments in contacts and intermediation by those with a high contact rate are socially desirable.

Hugonnier, Lester and Weill (2020) model a market with separate dealer and costumer sectors, where dispersion in flow payoffs gives rise to intermediation chains among dealers. Afonso and Lagos (2015) similarly have endogenous intermediation because banks with heterogeneous asset positions buy and sell depending on their counterparties' reserve holdings. In contrast to these setups, we offer a theory of endogenous heterogeneity which is rooted in the gains from intermediation.

Farboodi, Jarosch, Menzio and Wiriadinata (2018) model an environment where some traders 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 the allocation since misaligned asset positions are traded toward those who are more efficient at offsetting them. They also study an initial investment stage which 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 assumes the existence of middlemen who facilitate trade through their continuous access to an interdealer market (Duffie *et al.*, 2005; Weill, 2008; Lagos and Rocheteau, 2009). We show that middlemen are a natural outcome when homogeneous traders invest in contact rates and the marginal cost of contacts is constant.

Three recent papers endogenize market structure using a search framework. Hendershott, Li, Livdan and Schürhoff (2020) model a client–dealer network in which clients cannot act as intermediaries themselves but choose the number of dealers they contact. They find that clients choose homogeneous contact rates. In Chang and Zhang (2021), there are gains from concentrating misallocated positions, which in turn gives rise to endogenous intermediation and market structure. Dugast, Üslü and Weill (2019) ask which agents prefer to trade in a centralized and multilateral vs. in a decentralized and bilateral fashion. We assume all

2. A related literature studies the positive and normative consequences of high-frequency trading in centralized financial markets; see, for instance, Pagnotta and Philippon (2018). The decentralized interdealer market in Nekludyov (2019) also features dealers with exogenously given heterogeneous contact rates.

trade takes place in decentralized markets, but endogenize contact rates, which are exogenous in Dugast *et al.* (2019).

Second, the article relates to a literature on stochastic social network formation where agents need to make a costly upfront investment decision in contacts or links. A key difference, however, is that these papers find that agents of the same type make the same choices, and heterogeneity is driven solely by *ex post* shocks and *ex ante* differences in types. That there is no heterogeneity without these forces reflects our result that intermediation is necessary for heterogeneity, as well as the fact that there is no role for intermediation in these social networks.

As an example, Currarini, Jackson and Pin (2009) develop a frictional model of friendship formation, applied to homophily and segregation. Individuals decide on how long to partake in a costly stochastic matching process, which is similar to our assumption of a costly contact rate. Alternatively, Cabrales, Calvó-Armengol and Zenou (2011) consider a framework where agents choose random social interactions and investment simultaneously. The payoff is quadratic, depending on both partners' investments, differing from our linear meeting technology. In both models, identical individuals make identical choices in equilibrium. This is also the case in the remaining papers discussed in this section, with the exception of the last paragraph.

Kremer (1996) is an early contribution to the literature on disease transmission that integrates behaviour into an epidemiological model. Agents choose a rate of partner change that gives utility yet comes with a higher risk of HIV infection. Our paper Farboodi, Jarosch and Shimer (2021) similarly models the choice of social activity in the context of the Covid-19 pandemic. These papers have in common that they model endogenous contact intensity that leads to bilateral transmission (see also Quercioli and Smith, 2006).

Similarly, Duffie, Malamud and Manso (2009) study a search setting where agents make costly effort choices in their search for information that percolates via bilateral meetings. Cross-sectional heterogeneity in search effort arises due to heterogeneity in current information.

In Galeotti and Merlino (2014), workers choose how much information to obtain about job opportunities. They do so by making a costly investment decision into connections with others along the lines of Cabrales *et al.* (2011).

Finally, our results on endogenous heterogeneity in contact rates superficially resemble a literature showing the absence of a pure strategy equilibrium in search models (Butters, 1977; Burdett and Judd, 1983; Burdett and Mortensen, 1998; Duffie, Dworczak and Zhu, 2017). These papers have in common that if all firms charge the same price (or offer the same wage), firms that offer a slightly lower price (higher wage) earn discontinuously higher profits. Our results concern a different object, the contact rate distribution, and we find that the profit function is continuous but not differentiable. More fundamentally, all of the papers referenced in this paragraph show that heterogeneous choices arise in equilibrium. We demonstrate that both the equilibrium and the socially optimal allocation display rich, continuous heterogeneity. Thus, our results do not reflect a particular assumption about price formation but rather demonstrate how the possibility of providing or using intermediation services creates a reason for *ex ante* identical traders to make heterogeneous investments.

3. MODEL

We study an economy where time is continuous and extends forever. We focus throughout on an aggregate steady state. There is a unit measure of market participants, hereafter traders, who each have preferences defined over their holdings of an indivisible asset in fixed supply and their consumption or production of an outside good. Traders exit the market when hit by an

idiosyncratic shock with arrival rate $r>0$. When a trader exits, she is replaced with a newborn trader so as to keep the population fixed at 1.[3]

3.1. *Asset holdings and preferences*

Traders' asset holdings and preferences follow Duffie *et al.* (2005). An individual trader's asset holding is restricted to be $b\in\{0,1\}$. Traders have time-varying taste $i\in\{h,l\}$ for the asset and receive flow utility $\delta_{i,b}$ when they are in state (i,b). We assume that $\Delta\equiv\frac{1}{2}\left(\delta_{h,1}+\delta_{l,0}-\delta_{h,0}-\delta_{l,1}\right)$ is strictly positive, which implies that traders in the high state are the natural asset owners.

Half of all traders are born in state $(h,1)$ and half in state $(l,0)$.[4] Thereafter, a trader's taste switches from l to h when hit by an idiosyncratic shock with arrival rate $\gamma>0$ and back again at the same rate. Since this shock is idiosyncratic, half the traders are in state h and half are in state l in the stationary distribution. We similarly fix the supply of the asset at $\frac{1}{2}$, so at any point in time half the traders hold the asset and half do not. Thus, in a frictionless environment, the supply of assets is exactly enough to satiate the demand from traders with taste h.

Preferences over net consumption of the outside good are linear, so the outside good effectively serves as transferable utility when trading the asset. We assume that whether trade occurs and what the terms of trade are is determined according to the symmetric Nash bargaining solution. Traders discount the future only because of the exit probability r. When a trader exits holding the asset, it is transferred to a newborn trader with taste h, and the dying trader is not compensated.

3.2. *Contact technology*

Asset trades occur pairwise in a frictional asset market. Newborn traders choose a time-invariant rate $\lambda\in\mathcal{X}\equiv[0,\bar{\lambda}]$ at which they make contact with another trader, where $\bar{\lambda}$ is an exogenous upper bound. A high contact rate is costly: a trader who chooses a contact rate λ pays an *ex ante* cost $C(\lambda)$, where $C:\mathcal{X}\to\mathbb{R}$ is non-decreasing.[5] We allow different traders to choose different contact rates.[6]

A trader who chooses a contact rate λ meets a counterparty at rate λ. Search is random, so whom a trader meets is independent of her contact rate, taste, and asset holding. Let $\mathcal{B}$ be the Borel σ-algebra generated by $\mathcal{X}$ and μ_F be a probability measure on the measurable space $(\mathcal{B},\mathcal{X})$ which gives the probability that, conditional on a meeting, the counterparty's contact rate is some $\lambda'\in S\subset\mathcal{B}$, with $\mu_F(\mathcal{X})=1$. This probability measure is a key equilibrium object in our environment. For notational convenience, let F with $F(\lambda)\equiv\mu_F([0,\lambda])$ denote the cumulative distribution function associated with the measure μ_F. Conditional on meeting a counterparty with a given contact rate, the counterparty's taste and asset holding are drawn from

3. Vayanos and Wang (2007) is an early example of an asset market with stochastic exit and entry.

4. We can relax the assumption that all newborn traders are born in one of these two states, but it is convenient to assume that they do not know their state when they choose λ. We view this as reasonable because we do not think that the short-run desire to trade is an important determinant of the irreversible investment in λ. Preference shifts occur at a much higher frequency than exit, while trading opportunities in many markets occur at a higher frequency still. This implies that the initial state of new entrants will have little impact on the steady-state distribution of asset holdings and tastes.

5. In Section 6, we focus on the linear cost case, $C(\lambda)=c\lambda$. There we take a limit with $\bar{\lambda}\to\infty$.

6. One possible interpretation of this assumption is that participants may increase their contact rate by investing in their communication capacity, either through improved information technology or by simply hiring more or more able individuals to staff their trading desk. An alternative interpretation is that market participants may invest into relationships with more counterparties. That is, they may invest time and resources to increase the length of their contact list.

the population distribution with that contact rate, independent of the trader's contact rate, taste, and asset holding.

One can think of this contact technology as the continuous-time, continuous-type limit of the following physical environment with n traders and discrete time periods of length $dt < 1/\bar{\lambda}$: each period, each type λ trader accesses a market with i.i.d. probability λdt. If this process results in an odd number of traders accessing the market, one additional trader, chosen uniformly at random, is selected to also access the market. The traders who thus access the market are then randomly matched in pairs (see e.g. Shimer and Smith, 2001). In a large economy, the chance of any trader matching because an odd number of traders accessed the market becomes negligible, and so a trader matches each period with probability λdt. Moreover, for the same reason, traders meet counterparties in proportion to the counterparties' contact rate.[7] We return to this at the end of Section 4 and then offer additional detail in Appendix A.2.

We will show that individual behaviour depends only on the counterparty measure μ_F. Nevertheless, to connect the model to data, we also characterize the distribution of contact rates in the population. Let $\mu_G(S)$ denote the measure of traders whose contact rate is some $\lambda' \in S \subset \mathcal{B}$, with $\mu_G(\mathcal{X}) = 1$. Again, let G with $G(\lambda) \equiv \mu_G([0,\lambda])$ denote the associated cumulative distribution function. The measures μ_F and μ_G are related through the following transformation,[8]

$$\mu_F(S) \equiv \frac{\int_S \lambda d\mu_G(\lambda)}{\int_{\mathcal{X}} \lambda d\mu_G(\lambda)}. \tag{1}$$

This captures our assumption that traders meet counterparties in proportion to their contact rate. In words, the conditional probability of drawing a counterparty from a particular group of traders is given by the fraction of meetings that accrues to that group. Finally, we let the average contact rate in the population be denoted by $\Lambda \equiv \int_{\mathcal{X}} \lambda d\mu_G(\lambda)$.

3.3. *Individual trader behaviour*

We next turn to the behaviour of individual traders. To do so, we first need some notation. Let $p^{\lambda',i',1-b}_{\lambda,i,b} = p^{\lambda,i,b}_{\lambda',i',1-b}$ denote the endogenous probability that a trader with contact rate $\lambda \in [0,\bar{\lambda}]$, taste $i \in \{h,l\}$, and asset holdings $b \in \{0,1\}$ trades when she contacts a trader with contact rate $\lambda' \in [0,\bar{\lambda}]$, taste $i' \in \{l,h\}$, and asset holdings $1-b$. Let $t^{\lambda',i',1-b}_{\lambda,i,b} = -t^{\lambda,i,b}_{\lambda',i',1-b}$ denote the endogenous transfer of the outside good from (λ,i,b) to $(\lambda',i',1-b)$ when such a trade takes place. The trading probability and price jointly constitute the terms of trade and are determined by Nash bargaining, as we discuss further in Section 3.4. Finally, let $\sigma_{\lambda,i,b}$ denote the endogenous fraction of traders with contact rate λ who have taste i and asset holding b. This is determined from a steady state condition which we discuss in Section 3.5.

Now let $v_{\lambda,i,b}$ denote the present value of a trader (λ,i,b). Given μ_F, p, t, and σ, the value function satisfies

7. This is the natural counterpart to a frictional labour market where an employer meets job seekers in proportion to the job seekers' search intensity (see Petrongolo and Pissarides, 2001). An alternative is a "telephone-line" matching function where traders initiate contacts at some chosen rate λ and can also be contacted otherwise. The distribution of λ' among the counterparties then depends on who initiated the contact. Such a technology is inconsistent with our assumption that the distribution of whom a trader meets is independent of her contact rate.

8. Appendix A.1 discusses technical details on how we move between the two probability measures, in particular how we deal with cases where equation (1) is not invertible. Here and throughout the article, $\int_S \Omega(\lambda) d\mu_G(\lambda) = \int_S \Omega d\mu_G$ is the integral of Ω on $S \subset \mathcal{B}$ with respect to the measure μ_G. We use $\int_{\lambda_1}^{\lambda_2} \Omega(\lambda) d\lambda$ to denote the integral on $[\lambda_1,\lambda_2]$ with respect to the Lebesgue measure.

$$rv_{\lambda,i,b} = \delta_{i,b} + \gamma\left(v_{\lambda,\sim i,b} - v_{\lambda,i,b}\right) + \lambda \int_{\mathcal{X}} \sum_{i' \in \{h,l\}} \sigma_{\lambda',i',1-b} p_{\lambda,i,b}^{\lambda',i',1-b} \left(v_{\lambda,i,1-b} - v_{\lambda,i,b} - t_{\lambda,i,b}^{\lambda',i',1-b}\right) d\mu_F(\lambda'). \quad (2)$$

The left-hand side of equation (2) is the flow value of the trader, where discounting reflects the exit rate. The value comes from three sources, listed in order on the right hand side. First, she receives a flow payoff $\delta_{i,b}$ which depends on her tastes and asset holdings. Second, her tastes shift from i to $\sim i$ at rate γ. Third, she meets another trader at rate λ with type λ' drawn from the counterparty measure μ_F, in which case they may swap asset holdings in return for a payment. Conditional on λ', the counterparty's state is $(i', 1-b)$ with probability $\sigma_{\lambda',i',1-b}$. If there is trade, the trader has a capital gain from swapping assets and transferring the outside good, $v_{\lambda,i,1-b} - v_{\lambda,i,b} - t_{\lambda,i,b}^{\lambda',i',1-b}$. If the counterparty has the same asset holdings, there is no scope for trade.

Armed with the value function, we can formally describe the restriction which the model imposes on the contact rate distribution. Recall that traders choose their contact rate when they enter the market in order to maximize their expected value. This implies that any contact rate chosen in equilibrium must maximize

$$\pi_\lambda \equiv \tfrac{1}{2}(v_{\lambda,l,0} + v_{\lambda,h,1}) - C(\lambda). \quad (3)$$

This formulation reflects the fact that traders are equally likely to enter in the low taste state without the asset or the high taste state with the asset, and they choose λ accordingly subject to the investment cost C.

3.4. *Terms of trade*

The terms of trade are set in accordance with a symmetric Nash bargaining solution. This means that trade occurs whenever doing so can make both parties better off, and that transfers equate the gains from trade without throwing away any resources. That is, if there is a transfer $t_{\lambda,i,b}^{\lambda',i',1-b} = -t_{\lambda',i',1-b}^{\lambda,i,b}$ satisfying $v_{\lambda,i,1-b} - v_{\lambda,i,b} - t_{\lambda,i,b}^{\lambda',i',1-b}$ and $v_{\lambda',i',b} - v_{\lambda',i',1-b} - t_{\lambda',i',1-b}^{\lambda,i,b}$ both positive, then trade occurs with a transfer such that $v_{\lambda,i,1-b} - v_{\lambda,i,b} - t_{\lambda,i,b}^{\lambda',i',1-b} = v_{\lambda',i',b} - v_{\lambda',i',1-b} - t_{\lambda',i',1-b}^{\lambda,i,b}$. If any feasible transfer implies a strict loss from trade, there is no trade.

With a little bit of algebra, this means that the trading probability is given by

$$p_{\lambda,i,b}^{\lambda',i',1-b} = \begin{cases} 1 & \\ 0 & \end{cases} \quad \text{if } v_{\lambda,i,1-b} + v_{\lambda',i',b} \gtrless v_{\lambda,i,b} + v_{\lambda',i',1-b}; \quad (4)$$

and that when there is trade, the transfer satisfies

$$t_{\lambda,i,b}^{\lambda',i',1-b} = \frac{1}{2}\left(v_{\lambda,i,1-b} + v_{\lambda',i',1-b} - v_{\lambda,i,b} - v_{\lambda',i',b}\right). \quad (5)$$

When $v_{\lambda,i,1-b} + v_{\lambda',i',b} = v_{\lambda,i,b} + v_{\lambda',i',1-b}$, trade may be probabilistic. If trade does occur, the transfer is still given by equation (5).

3.5. *Stationary distribution*

The steady state fraction of type λ traders in different states, $\sigma_{\lambda,i,b}$, depends on the trading probabilities through the balance of inflows and outflows:

$$\left(r+\gamma+\lambda\int_{\mathcal{X}}\sum_{i'\in\{h,l\}}\sigma_{\lambda',i',1-b}p_{\lambda,i,b}^{\lambda',i',1-b}d\mu_F(\lambda')\right)\sigma_{\lambda,i,b}$$
$$=\gamma\sigma_{\lambda,\sim i,b}+\lambda\left(\int_{\mathcal{X}}\sum_{i'\in\{h,l\}}\sigma_{\lambda',i',b}p_{\lambda,i,1-b}^{\lambda',i',b}d\mu_F(\lambda')\right)\sigma_{\lambda,i,1-b}+\frac{r}{2}\mathbb{I}_{(i,b)\in\{(h,1),(l,0)\}}. \quad (6)$$

The left-hand side of equation (6) measures the outflows from state (i,b) for traders with contact rate λ. A trader leaves the state either when she exits, when she has a taste shock, or when she trades with another trader with the opposite asset holding. The right-hand side measures the inflows. A trader with contact rate λ enters state (i,b) when she has the opposite taste and has a taste shock, when she has the opposite asset holding and trades, or, if (i,b) is equal to either $(h,1)$ or $(l,0)$, half the time when she is newborn. Here, the indicator function $\mathbb{I}$ is equal to 1 if the condition in the subscript holds and is zero otherwise.

3.6. *Summary*

A natural notion of equilibrium is given by the following primitive objects: a counterparty measure μ_F, value function v, trading probability function p, price function t, and steady state shares σ which jointly satisfy equations (2)–(6).

4. EQUILIBRIUM

In this section, we define a symmetric equilibrium, which we prove is a special case of the primitive notion of equilibrium that we just defined. To do so, we impose a restriction on the trading pattern which we keep in place for the remainder of the article. We relegate formal derivations and details to Supplementary Appendix C.

To begin with, we call traders' asset holding positions *misaligned* either when they hold the asset and have taste l, or when they do not hold the asset and have taste h. We call traders' asset holding positions *well-aligned* in the other two states. Let $m_\lambda \equiv \sigma_{\lambda,l,1}+\sigma_{\lambda,h,0}$ denote the fraction of traders with contact rate λ who are misaligned.

In a symmetric equilibrium, we restrict attention to equilibria in which the two misaligned states and the two well-aligned states are treated symmetrically. That is, we impose $p_{\lambda,i,0}^{\lambda',i',1}=p_{\lambda,\sim i,1}^{\lambda',\sim i',0}$ for all λ, $i\neq\sim i$, and $i'\neq\sim i'$. This means that if a type λ trader with taste i would buy the asset from a type λ' trader with taste i', then a type λ trader with the opposite taste $\sim i$ would sell the asset to a type λ' trader with the opposite taste $\sim i'$.

In Supplementary Appendix C.1, we prove that when trading probabilities are symmetric in this sense, misalignment rates are also symmetric, $\sigma_{\lambda,l,b}=\sigma_{\lambda,h,1-b}$ for all λ and b. Moreover, we prove the existence of a *surplus function* $s(\lambda)=v_{\lambda,h,1}-v_{\lambda,h,0}-q=v_{\lambda,l,0}-v_{\lambda,l,1}+q$, where

$$q\equiv\frac{\delta_{h,1}+\delta_{l,1}-\delta_{h,0}-\delta_{l,0}}{2r}, \quad (7)$$

independent of λ. The surplus function tells us the value of being well aligned, up to the additive constant q, which equals the price of the asset in the frictionless limit. Manipulating the value

function (2) as well as the Nash bargaining solution (4) and (5), we derive in Supplementary Appendix C.1 a Bellman equation for the surplus function $s(\lambda)$:

$$(r+2\gamma)s(\lambda)=\Delta+\frac{\lambda}{4}\int_{\mathcal{X}}\Big(\big((s(\lambda')-s(\lambda))^{+}-(s(\lambda)+s(\lambda'))^{+}\big)m_{\lambda'} \\ +\big((-s(\lambda)-s(\lambda'))^{+}-(s(\lambda)-s(\lambda'))^{+}\big)(1-m_{\lambda'})\Big)d\mu_F(\lambda'), \tag{8}$$

where $z^{+}\equiv\max\{z,0\}$ and reflects that meetings result in trade if and only if trade is bilaterally efficient. The $r+2\gamma$ on the left-hand side reflects discounting due to exit and taste shocks. Δ is the average difference in flow payoffs between the well-aligned and misaligned states. The remaining terms capture how the option value of trade changes through alignment. At rate $\frac{\lambda}{2}$, traders meet others with opposite asset holdings. When trade occurs, the gains are split equally, giving us $\frac{\lambda}{4}$. The trader meets both misaligned (fraction $m_{\lambda'}$) and well-aligned (fraction $1-m_{\lambda'}$) counterparties of type λ'. Whenever a trade makes this trader well aligned (misaligned), there is a gain (loss) $s(\lambda)$. Similarly, whenever a trade makes the partner λ' well aligned (misaligned), there is a gain (loss) $s(\lambda')$. Trade occurs only if the sum of the two gains is positive.

We also use the steady state equation (6) and the Nash bargaining solution (4) to derive in Supplementary Appendix C.1 the symmetric flow balance equation governing the stationary distribution of misalignment:

$$\left(r+\gamma+\frac{\lambda}{2}\int_{\mathcal{X}}\big(\mathbb{I}_{s(\lambda)+s(\lambda')>0}m_{\lambda'}+\mathbb{I}_{s(\lambda)>s(\lambda')}(1-m_{\lambda'})\big)d\mu_F(\lambda')\right)m_{\lambda} \\ =\left(\gamma+\frac{\lambda}{2}\int_{\mathcal{X}}\big(\mathbb{I}_{s(\lambda)<s(\lambda')}m_{\lambda'}+\mathbb{I}_{s(\lambda)+s(\lambda')<0}(1-m_{\lambda'})\big)d\mu_F(\lambda')\right)(1-m_{\lambda}). \tag{9}$$

The indicator function $\mathbb{I}$ is equal to 1 if the inequality in the subscript holds and is zero otherwise. The left-hand side captures the outflow from misalignment. Misaligned traders become well aligned if they exit and are replaced, if they have a taste shock, or if they trade. The latter event can occur with both well-aligned and misaligned counterparties, so long as the joint surplus of the transaction is positive. The right hand side captures the inflow into misalignment from well-aligned traders who experience a taste shock or who trade.

Finally, building on equation (3), we prove in Supplementary Appendix C.1 that symmetry and Nash bargaining imply that the choice of contact rate maximizes

$$r\pi_{\lambda}=\delta_1-\gamma s(\lambda)+\frac{\lambda}{4}\int_{\mathcal{X}}\big((s(\lambda')-s(\lambda))^{+}m_{\lambda'}+(-s(\lambda)-s(\lambda'))^{+}(1-m_{\lambda'})\big)d\mu_F(\lambda')-rC(\lambda), \tag{10}$$

where $\delta_1\equiv\frac{1}{2}(\delta_{h,1}+\delta_{l,0})$ is the average flow payoff in the well-aligned state. The structure of this equation is similar to equation (8). Equation (10) reflects each trader's cost–benefit analysis when she chooses her contact rate. The first term is the flow payoff from being well aligned. The second term is the cost of losing her well-aligned status following a taste shock. This cost turns out to be decreasing in the trader's contact rate, reflecting the fact that a higher contact rate enables a trader to more quickly realign her asset position with her tastes. The third term captures the profit that a trader earns from her contacts when she is in the well-aligned state. We show below that this comes from providing intermediation services. Finally, the last term captures the

exogenous cost associated with choosing a contact rate, parallel to the cost of forming a link in Jackson and Wolinsky (1996) and the subsequent literature on network formation.[9]

We are now able to define a (symmetric) equilibrium.

Definition 1. *An equilibrium is a counterparty measure μ_F, a misalignment rate function m: $\mathcal{X} \to [0,1]$, and a surplus function $s: \mathcal{X} \to \mathbb{R}$, satisfying:*

1. *the surplus equation (8);*
2. *the flow balance equation (9); and*
3. *optimal investment: $\mu_F(\mathcal{Y}) = 1$, where $\mathcal{Y} = \arg\max_{\lambda \in \mathcal{X}} \pi_\lambda$ and π_λ is defined given μ_F, m, and s in equation (10).*

We have already explained all three conditions.

Given an equilibrium tuple of reduced-form objects (μ^F, m, s), we can recover the primitive objects described in Section 3; see Supplementary Appendix C.2 for more details. First, we use all three objects, most notably the surplus function, to recover the value function $v_{\lambda,i,b}$. From that, we get the trading probability $p_{\lambda,i,b}^{\lambda',i',1-b}$ and the transfer $t_{\lambda,i,b}^{\lambda',i',1-b}$ via equations (4) and (5). Next, using symmetry, the misalignment rate encodes the fraction of type λ traders in different states, $\sigma_{\lambda,i,b}$. Putting this together, we obtain a value function satisfying equation (2) and steady state shares satisfying equation (6). Finally, the choice of contact rates is equivalent to what is in equation (3).

To understand the definition of equilibrium, it may help to comment briefly on the possibility of an *autarky equilibrium*, where the average contact rate is $\Lambda = 0$, or equivalently $\mu_G(\{0\}) = 1$. We emphasize that such an equilibrium does not necessarily exist in this environment. This is because our contact technology allows each trader to choose any contact rate $\lambda \in \mathcal{X}$, regardless of what others are doing.[10] In particular, a trader who chooses a contact rate λ will meet some other trader drawn from the counterparty measure μ_F at rate λ. In Section 3.2, we already discussed a physical matching process which justifies this assumption, as we show in more detail in Appendix A.2. There, we explain how a single trader can contact others at an arbitrary strictly positive rate, even in the case where everyone else chooses a zero contact rate. We also show that this matching process is consistent with autarky, $\mu_G(\{0\}) = 1$, and *simultaneously* with $\mu_F(\{0\})$ taking an arbitrary value between 0 and 1, depending on details.

The flexibility of the counterparty distribution in an autarky equilibrium reflects the fact that we cannot use equation (1), i.e., Bayes rule, to recover μ_F from μ_G when $\mu_G(\{0\}) = 1$. But, we stress that our definition of equilibrium insists that μ_F puts all its weight on value-maximizing contact rates, including possibly a zero contact rate. If 0 is the unique profit-maximizing contact rate, this implies $\mu_F(\{0\}) = 1$, but we find that an autarky equilibrium sometimes requires a different counterparty measure; see the proof of Proposition 3 for an example.

5. CHARACTERIZATION WITH GENERAL COST FUNCTIONS

This section develops two main results characterizing equilibrium. As a stepping stone, Lemma 1 establishes which trades occur for an arbitrary counterparty distribution. Proposition 1 then shows that any counterparty distribution is an equilibrium for some cost function and shows how to

9. The assumption that $C(\lambda)$ is paid upfront is isomorphic to one where traders pay $rC(\lambda)$ per unit of time, or one where traders pay $rC(\lambda)/\lambda$ per meeting.

10. One could instead impose that when the contact rate distribution is degenerate at zero, $\mu_G(\{0\}) = 1$, it is impossible to meet other traders. Such an alternative assumption would imply that an autarky equilibrium always exists.

construct such a cost function; and Proposition 2 shows that rich dispersion in contact rates arises under general conditions. See Supplementary Appendix C.3 for technical details, including formal proofs of these results.

5.1. *Equilibrium trading patterns*

We start by characterizing equilibrium trading patterns given any counterparty measure μ_F.

Lemma 1. *In any equilibrium, the surplus function $s(\lambda)$ is positive-valued and strictly decreasing. 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 Nash bargaining solution (4) and the definition of the surplus function $s(\lambda)=v_{\lambda,h,1}-v_{\lambda,h,0}-q=v_{\lambda,l,0}-v_{\lambda,l,1}+q$ jointly imply

$$p_{\lambda,h,0}^{\lambda',l,1}=\begin{cases}1\\0\end{cases}\quad\text{if } s(\lambda)+s(\lambda')>0,\qquad p_{\lambda,h,0}^{\lambda',h,1}=\begin{cases}1\\0\end{cases}\quad\text{if } s(\lambda)>s(\lambda'),$$

$$p_{\lambda,l,0}^{\lambda',h,1}=\begin{cases}1\\0\end{cases}\quad\text{if } 0>s(\lambda)+s(\lambda'),\qquad p_{\lambda,l,0}^{\lambda',l,1}=\begin{cases}1\\0\end{cases}\quad\text{if } s(\lambda')>s(\lambda).$$

The bulk of the proof establishes that the surplus function is positive-valued and strictly decreasing, from which the trading patterns follow immediately.

Üslü (2019) and Nekludyov (2019) both derive similar results in richer settings with an exogenous counterparty distribution. Thus, Lemma 1 is a special case of the key findings in these papers, a stepping stone to our novel result on equilibrium dispersion in contact rates.

The first two parts of Lemma 1 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 lemma reflects option value considerations and is the key endogenous trading pattern that arises in this environment. It states that a faster trader buys the asset from a slower trader if both have taste l; and she sells the asset to the slower trader if both have taste h. We label trades *intermediation* when both traders have the same taste for the asset. Intermediation does not immediately increase the number of well-aligned traders, but it moves misalignment towards traders who expect more future trading opportunities. Intermediation yields gains in equilibrium because traders with higher contact rates are faster at offloading misaligned positions in future trades.

The possibility of intermediation implies that a trader's buying and selling decisions become increasingly detached from her idiosyncratic tastes 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, positioned at the centre of the trading chain, are traders with a high contact rate. Figure 1 shows the intermediation chain which follows from Lemma 1. Slow traders are at the periphery of the trading chain, not trading once their asset position is aligned with their tastes. In turn, fast traders constitute the endogenous core of the trading network, buying and selling largely irrespective of their tastes. In doing so, they take on

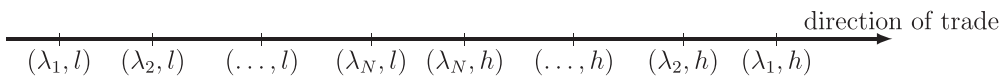

FIGURE 1

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

misaligned asset positions from types with lower contact rates simply because they are better at locating other traders. That is, they intermediate.

5.2. *Recovering the cost function*

We next prove that our model can rationalize any observed counterparty measure μ_F:

Proposition 1. *For any counterparty measure μ_F, there exists a cost function C such that μ_F is an equilibrium. Moreover, C is unique on support of μ_F, up to an additive constant.*

The formal proof proceeds in three steps. First, we show that the counterparty distribution uniquely determines the misalignment rate. Then, we derive the functional form for the surplus equation, thereby proving that the counterparty distribution uniquely determines the surplus function. Finally, we show how to recover the cost function from these three objects.

Since every contact rate measure μ_G is associated with some counterparty measure μ_F, a corollary is that our model can rationalize any contact rate measure through the choice of an appropriate cost function C. As such, our framework offers a novel way of modelling and rationalizing real-world trading networks, complementing the existing literature on network formation. In particular, we can rationalize the coexistence of traders with very different rates of market access despite them being *ex ante* identical. In the next subsection as well as Section 6, we invert the analysis in Proposition 1 to characterize the counterparty distribution under some natural assumptions on the shape of the cost function.

5.3. *Heterogeneity in contact rates*

We next show our second main result, that the coexistence of traders with different contact rates arises naturally in equilibrium even when market participants are *ex ante* homogeneous.

Proposition 2. *Assume C is differentiable and C' is Lipschitz continuous. Then any equilibrium counterparty distribution F and contact rate distribution G are absolutely continuous on $[0, \bar{\lambda})$. If additionally C is weakly convex, $C'(0) < \frac{\Delta\gamma^2}{2r(r+2\gamma)^3}$, and $C'(\bar{\lambda}) \geq \frac{4\gamma\Delta}{r\bar{\lambda}^2}$, then a positive measure of traders choose a contact rate in the interval $(0, \bar{\lambda})$ in any equilibrium.*

Recall that $F(\lambda) \equiv \mu_F([0, \lambda])$ is the counterparty distribution and $G(\lambda) \equiv \mu_G([0, \lambda])$ is the contact rate distribution.

Proposition 2 implies that although all traders are *ex ante* identical, there is no symmetric equilibrium in which all traders choose identical actions, except possibly at the boundaries of the choice set 0 and $\bar{\lambda}$. In particular, under mild restrictions on the cost function, *ex ante* identical traders choose to be continuously heterogeneous, meaning $\mu_F((0, \bar{\lambda})) > 0 = \mu_F(\{\lambda\})$ for all $\lambda \in (0, \bar{\lambda})$. That is, a positive measure of traders choose an interior contact rate, but a zero measure choose the same interior contact rate.

The critical force underlying Proposition 2 is intermediation. To develop an intuition, we argue that if everyone has a common contact rate λ, there is a convex kink in the profit function π_λ at λ, reflecting the gains from intermediation. To see this, consider the marginal return to a change in the contact rate at the mass point λ. A trader with contact rate λ only engages in fundamental trades, but this changes discretely at slightly different contact rates. A trader who chooses a contact rate $\lambda+\varepsilon$, $\varepsilon>0$, intermediates for the entire marketplace, trading independently of her intrinsic valuation whenever the counterparty is misaligned and trading is feasible (Lemma 1). The profits from intermediation are proportional to the difference in surplus functions, $s(\lambda)-s(\lambda+\varepsilon)$. Since s is strictly decreasing, this is locally linearly increasing in the trader's contact rate when it exceeds λ.

Conversely, consider the intermediation returns of a trader who chooses a contact rate $\lambda+\varepsilon$, $\varepsilon<0$. Now the trader benefits from others intermediating for her, with profits still proportional to the difference in surplus functions, $s(\lambda+\varepsilon)-s(\lambda)$. We still have s strictly decreasing, so now as ε increases to zero from below, the trading profits again shrink to zero from above. That is, intermediation profits are locally linearly decreasing in the trader's contact rate when it is smaller than λ.

The intermediation benefits of moving away from the mass points are hence positive in both directions. On top of that, there are fundamental trading benefits to a higher contact rate, which are locally linear. This leads to a convex kink, which in turn means that choosing the mass point cannot be optimal if the cost function is differentiable. This logic carries over to any interior mass point. As soon as a positive measure of traders has the same contact rate, there is a discrete jump up in the marginal return to contacts at this mass point, inconsistent with equilibrium under a differentiable cost function. A similar intuition rules out counterparty distributions that are not absolutely continuous, but for this we require a stronger condition, that marginal cost is Lipschitz.

The theoretical finding that intermediation in this frictional setting leads to a market structure with rich heterogeneity connects closely with empirical evidence. Fricke and Lux (2015) estimate a network model that allows for a continuous notion of "coreness" for the Italian interbank market. Their Figure 9 strongly suggest a vast amount of heterogeneity, inconsistent with a simple binary classification of market participants. in 't Veld and van Lelyveld (2014) find very similar results for the Dutch interbank market (see their Figure A.13). Di Maggio, Kermani and Song (2017) similarly show that various measures of market participation in the US corporate bond market are continuously distributed (see their Figure 1).[11] These markets thus have a tiered network structure with rich heterogeneity.

Our theory further posits that almost all market participants at least occasionally intermediate. This likewise connects with data. Craig and Von Peter (2014) only classify 2.4% of the banks in their German interbank market as the core, yet find that 92.7% of banks occasionally intermediate in that market. Bech and Atalay (2010) group banks in the US federal funds market into multiple tiers and show that there are many links within tiers, so intermediation is frequently done in a decentralized fashion (see their Figure 4). For similar evidence from the US market for asset-backed securities, see Hollifield *et al.* (2017), Figure 5. These authors also show that peripheral dealers are frequently part of long intermediation chains that involve multiple dealers; see their Table 4. Table 2 in Coen and Coen (2021) shows that 23% of all trades in the sterling corporate bond market are directly between customers.

11. Similarly, see Boss, Elsinger, Summer and Thurner (2004) for the Austrian interbank market, Cont, Moussa and Santos (2010) for the Brazilian interbank network, Martinez-Jaramillo, Alexandrova-Kabadjova, Bravo-Benitez and Solórzano-Margain (2014) for the Mexican interbank network, and Coen and Coen (2021) for the sterling corporate bond market.

Jointly, this suggests that frictional asset markets frequently feature traders with various degrees of market access, most of whom at least occasionally engage in intermediation activities. This closely aligns with the endogenous market structure which arises in our setting.

5.4. *Ordinary differential equation system*

Absolute continuity, as established in Proposition 2, implies that the cumulative distribution function is described by its derivative, $F(\lambda)-F(0)=\int_0^\lambda F'(\lambda')d\lambda'$ for all λ, and similarly for G. In Supplementary Appendix C.3, we define $M(\lambda)\equiv\int_0^\lambda m_{\lambda'}dF(\lambda')$ to be the fraction of meetings that are with a misaligned trader with contact rate below some $\lambda\in\mathcal{X}$. We then characterize equilibrium using a first-order ordinary differential equation system in F, M, and s with known boundary values. We show the differential equations in Appendix B. The simplicity of this differential equation system enables us to solve the model numerically for arbitrary cost functions that satisfy our smoothness requirement. We also use the system throughout Section 6 to obtain an analytical characterization with a linear cost function.

5.5. *Robustness*

Proposition 2 is the most general result in this article, showing that dispersion in contact rates arises under weak conditions. To underscore its generality, we next offer some conjectures on the robustness of the result to our modelling assumptions.

5.5.1. Symmetry. We have assumed that the asset endowment is equal to $1/2$, exactly equal to the measure of traders in the high state at any point in time. This makes the restriction to symmetric equilibrium natural. Without symmetry, we would have to deal with four value functions and with the share of traders in each of four taste-asset holding states, making notation more cumbersome. Moreover, it is no longer *ex ante* obvious which trades take place. For example, if the asset is scarce, the market may shut out the slowest traders, never selling the asset to them, even if they are in the high state. Still, there will always be a role for intermediation among the traders fast enough to hold the asset. Since intermediation is what drives heterogeneity, as we discussed above, we expect a version of Proposition 2 to be robust to such an extension.

5.5.2. Restricted asset holdings. We have restricted asset holdings to be either zero or one, which we view as the limit of an extremely convex inventory cost. Allowing for unrestricted asset holdings, as for instance in Üslü (2019), would preserve the connection between intermediation and heterogeneous contact rates. In fact, we believe it would amplify the force creating a kink in the profit function at a mass point, since a slightly faster trader would be unrestricted in her ability to intermediate for the mass of traders.

5.5.3. Contact technology. We have assumed that a trader's contact rate does not depend on the choices others make, but who she meets depends on these choices. We could have made other choices. For example, in footnote 7, we discuss the telephone matching function, where a trader chooses how often to call others but also receives calls from others at a rate that is independent of her choice of contact rate. Thus, even a trader who chooses not to contact anyone will be able to buy and sell assets. We believe that this does not affect the forces pushing towards dispersion in contact rates. In particular, if everyone else chooses a common contact rate λ, one trader's choice of contact rate does not affect whom she meets in either model. It just leads to

gains from acting as an intermediary if the trader chooses a faster contact rate, or to gains from being intermediated for if the trader chooses a slower contact rate.

5.5.4. Time-invariant λ. If they could do so costlessly, traders would want to adjust their contact rate in response to their time-varying taste and asset holding. They do not do so because we have assumed that the choice of contact rate is irreversible. If we allow identical traders to do so at no cost, then a trader's current contact rate no longer affects future trading opportunities and so the motive for intermediation disappears. Thus, some irreversibility in contact rates is important for our results. But, as long as changing the contact rate either takes time or incurs an irreversible cost, the current contact rate is relevant for future trading opportunities and so some trades will involve intermediation. And because there is intermediation, there is an incentive to choose a different contact rate than others. That is, each trader's contact rate may move around over time, but the associated stationary distribution of contact rates will still not feature any mass points.

6. CHARACTERIZATION WITH A LINEAR COST FUNCTION

This section characterizes equilibrium under the assumption that the cost function C is linear, $C(\lambda)=c\lambda$. After analysing the baseline model, we extend our analysis to a limiting economy with no upper bound on contacts, $\bar{\lambda}\to\infty$. For this case, we also consider what happens in the frictionless limit, when the marginal cost of contacts c converges to zero. We relegate technical details, including all proofs, to Supplementary Appendix C.4.

6.1. *Equilibrium characterization*

We start by proving existence of equilibrium and characterizing its properties when the cost function is linear:

Proposition 3. *Assume $C(\lambda)=c\lambda$. Fix r, γ, Δ, and $\bar{\lambda}$. There exist thresholds $\bar{c}>\underline{c}>0$ such that*

$$\text{if } \begin{cases} c\geq\bar{c} \\ c\in(\underline{c},\bar{c}) \\ c\leq\underline{c}, \end{cases} \quad \text{then there is a } \begin{cases} \textit{autarky equilibrium} \\ \textit{intermediated trade equilibrium} \\ \textit{degenerate trade equilibrium,} \end{cases}$$

and any equilibrium takes one of these three forms. In an autarky equilibrium, the average contact rate is $\Lambda=0$. In an intermediated trade equilibrium, the average contact rate is $\Lambda\in(0,\bar{\lambda})$; the support of the counterparty distribution is a convex interval $[\underline{\lambda},\bar{\lambda}]$ with $\underline{\lambda}\in(0,\bar{\lambda})$ and $dF(\bar{\lambda})>0$; and the misalignment rate m_λ is increasing on $[\underline{\lambda},\bar{\lambda}]$. In a degenerate trade equilibrium, the average contact rate is $\Lambda=\bar{\lambda}$.

Recall that $\Lambda\equiv\int_{\mathcal{X}}\lambda d\mu_G(\lambda)$. The proof gives explicit expressions for the thresholds $\bar{c}$ and $\underline{c}$ and characterizes the contact rate and counterparty distributions for any value of c.

We do not claim uniqueness of the equilibrium and indeed can construct examples in which an autarky equilibrium and an intermediated trade equilibrium coexist for the same parameter values. However, any equilibrium must lie in one of the three classes described in the proposition.

Proposition 3 states that for $c\geq\bar{c}$, there exists an equilibrium where all trading activity collapses, while for $c\leq\underline{c}$, there exists an equilibrium without intermediation since all traders choose the highest contact rate $\bar{\lambda}$. More interestingly, for a non-empty interval of costs $(\underline{c},\bar{c})$,

there exists an equilibrium where a non-degenerate contact rate distribution G and intermediation emerge endogenously. Such an equilibrium has four key properties: first, no trader has a contact rate below a strictly positive lower bound $\underline{\lambda}$. Second, a strictly positive fraction of traders choose $\bar{\lambda}$. Third, the remaining counterparties have a continuously distributed contact rate on $[\underline{\lambda}, \bar{\lambda})$. And finally, traders who choose a faster contact rate are misaligned more often.

The strictly positive lower bound $\underline{\lambda}$ in the intermediated trade equilibrium reflects the fact that the profits of a trader are a continuous function, converging to the autarky value as λ converges to 0. With $c < \bar{c}$, traders in the non-degenerate equilibrium do strictly better than autarky, and so it must be the case that no one chooses a contact rate too close to zero.

We postpone the discussion of the second feature, mass at $\bar{\lambda}$, to the next subsection. To understand the third finding, suppose there were a "hole" in the support, with no trader choosing a contact rate inside a nontrivial interval on $[0, \bar{\lambda}]$. In this case, the proof shows that the profits must be unequal at the two endpoints. Why? Because trading profits over that range would be linear in λ since trading opportunities would not be changing, while improvements in a trader's asset position show diminishing returns to scale. This implies the profit function must be concave on the interval, which is inconsistent with both extreme points yielding a higher value than any intermediate point.

The finding that faster contact rates are associated with higher misalignment rates might be counterintuitive. A higher contact rate has two opposing effects on a trader's misalignment rate. On the one hand, a trader is more frequently able to offset a misaligned position. On the other hand, a trader with a higher contact rate intermediates more frequently, taking on misalignment from slower traders. The proposition states that the latter force dominates everywhere on the support of F. That is, traders do not invest in a faster contact rate to reduce their misalignment, but rather to trade more frequently.

Intuitively, doubling a trader's contact rate from λ to 2λ more than doubles his opportunities for intermediation, since he can also intermediate for traders with contact rate $\lambda' \in [\lambda, 2\lambda)$. An increase in the misalignment rate is then needed to offset this increase in intermediation profits, leaving the trader's profits unchanged. We stress that this result holds in equilibrium, not for an arbitrary distribution of contact rates. Still, we find that this result is more general than the linear cost case. For example, in Farboodi, Jarosch and Shimer (2017), we show that misalignment rate is increasing in the contact rate when the marginal cost of contacts is increasing and convex.

6.2. *Middlemen and Pareto tail*

Proposition 3 implies that whenever there is trade, a positive fraction of contacts are with traders who choose the maximum permissible contact rate, $dF(\bar{\lambda}) > 0$, and so the choice of $\bar{\lambda}$ affects equilibrium. We next examine what happens when $\bar{\lambda}$ is large. To do this, we define a *limiting equilibrium* as the limit of equilibria of a sequence of economies n which are identical except for their upper bounds $\bar{\lambda}_n$, with $\bar{\lambda}_n \to \infty$:

Definition 2. *Assume $C(\lambda) = c\lambda$. Fix r, γ, Δ, and c. For any $\bar{\lambda}$, let $(\mu_{F,\bar{\lambda}}, m_{\bar{\lambda}}, s_{\bar{\lambda}})$ be an equilibrium when the maximum contact rate is $\bar{\lambda}$ and as usual let $F_{\bar{\lambda}}(\lambda) = \mu_{F,\bar{\lambda}}([0,\lambda])$ for all $\lambda \leq \bar{\lambda}$. Also extend the definition of $(F_{\bar{\lambda}}, m_{\bar{\lambda}}, s_{\bar{\lambda}})$ to the positive reals in an arbitrary way. (F, m, s) with domain $[0,\infty)^3$ is a* limiting equilibrium *if there exists an increasing unbounded sequence $\{\bar{\lambda}_n\}$ with associated $(F_{\bar{\lambda}_n}, m_{\bar{\lambda}_n}, s_{\bar{\lambda}_n})$ which converges pointwise to (F, m, s).*

Intuitively, a limiting equilibrium is the limit of a sequence of equilibria as we increase $\bar{\lambda}$. The only subtle point is that we need to extend the range of the functions $(F_{\bar{\lambda}_n}, m_{\bar{\lambda}_n}, s_{\bar{\lambda}_n})$ above

the upper bound $\bar{\lambda}_n$. A natural, but not necessary, way to do this is to impose that for $\lambda > \bar{\lambda}$, $F_{\bar{\lambda}}(\lambda)=1$ (reflecting that $F_{\bar{\lambda}}$ is a cumulative distribution function), $m_{\bar{\lambda}}(\lambda)=\frac{2\gamma+\lambda M(\bar{\lambda})}{2(r+2\gamma+\lambda M(\bar{\lambda}))}$ (reflecting equation 9), and $s_{\bar{\lambda}}(\lambda)=\frac{2\Delta}{2(r+2\gamma)+\lambda M(\bar{\lambda})}$ (reflecting equation 8).

To get some intuition for how the limit of equilibria behaves, it is useful for us to briefly discuss the mathematical structure we use to characterize equilibrium. To begin, recall that if the marginal cost function is Lipschitz continuous, we can represent any equilibrium as the solution to a system of three ordinary differential equations in F, M, and s. In the linear cost case, we prove that we can reduce this to a pair of ordinary differential equations, $(F',M')=X_1(\lambda,F,M)$ on $[\underline{\lambda},\bar{\lambda})$, with boundary condition $F(\underline{\lambda})=M(\underline{\lambda})=0$. The function X_1 depends on the parameters r and γ but not on c, Δ, or $\bar{\lambda}$. We show that there is a discontinuous increase in F and M at $\bar{\lambda}$, the $dF(\bar{\lambda})>0$ result in Proposition 3. In a limiting equilibrium, we simply solve the same differential equations on $[\underline{\lambda},\infty)$.

If we knew the lower bound on contact rates $\underline{\lambda}$, we would be done, but $\underline{\lambda}$ is endogenous. To find it, we use a second equation, expressed succinctly as $c=X_2(\underline{\lambda})$, telling us the cost c which makes $\underline{\lambda}$ the lower bound on contact rates. The function X_2 depends on r, γ, and Δ both directly and indirectly through the solution to the ordinary differential equation system $(F',M')=X_1(\lambda,F,M)$, and on $\bar{\lambda}$ indirectly through the discontinuity in F and M at $\bar{\lambda}$. Moreover, if c is not too big, we can always find a value of $\underline{\lambda}>0$ that makes this equation hold. As $\bar{\lambda}$ grows without bound, the indirect effect of $\bar{\lambda}$ on $\underline{\lambda}$ vanishes, which ensures that $\underline{\lambda}$ has a well-behaved limit when we take the upper bound on contacts to infinity. This value of $\underline{\lambda}$, combined with the solution to the differential equation system on the unbounded interval $[\underline{\lambda},\infty)$, yields our characterization of limiting equilibrium.

Using this approach, we obtain the following characterization of limiting equilibrium:

Proposition 4. *Assume* $C(\lambda)=c\lambda$ *with* $c<\frac{\gamma\Delta}{8r(r+\gamma)(r+2\gamma)}$. *Then in a limiting equilibrium, there are* middlemen, *meaning* $\lim_{\lambda\to\infty}F(\lambda)<1$*; and the contact rate distribution has a* Pareto tail with tail index 2, *meaning* $\lim_{\lambda\to\infty}\lambda^2(1-G(\lambda))$ *is positive and finite.*

Proposition 3 showed that a strictly positive fraction of meetings are with traders at any finite upper bound $\bar{\lambda}$. The existence of middlemen is a stronger result, because it implies that this fraction does not vanish as $\bar{\lambda}$ goes to infinity. In order to show the existence of middlemen, we show that the counterparty distribution remains bounded away from one as $\lambda\to\infty$. In the limiting economy, almost every trader has a finite contact rate ($\lim_{\lambda\to\infty}G(\lambda)=1$), yet a positive fraction of their counterparties has a higher contact rate. We refer to these counterparties as *middlemen*.

Why do middlemen emerge? Suppose there were none, so $\lim_{\lambda\to\infty}F(\lambda)=1$. Then a trader with a high contact rate has almost all her meetings with slower traders. Such a trader will therefore trade irrespective of her intrinsic valuation and so will have a misalignment rate close to $\frac{1}{2}$. In particular, her misalignment rate would be higher than that of a trader living in autarky, for whom equation (9) implies $m_0=\frac{\gamma}{r+2\gamma}$. To justify the choice of a large value of λ, it must then be the case that the fast trader earns strictly positive profits from trading. But trading profits scale linearly in the tail of the distribution, since trading opportunities are effectively the same and a trader can choose any multiple of λ. This is inconsistent with a large finite value of λ being optimal, contradicting the hypothesis that $\lim_{\lambda\to\infty}F(\lambda)=1$. In short, what middlemen do is ensure that even very fast traders have a misalignment rate strictly below $\frac{1}{2}$, obviating the need for them to earn profits from intermediation.[12]

12. Correspondingly, with a finite $\bar{\lambda}$, mass at the upper bound guarantees that traders remain indifferent across λ as $\lambda\to\bar{\lambda}$.

We next note that the contact rate distribution has support $[\underline{\lambda}, \infty)$, a natural extension of the support $[\underline{\lambda}, \bar{\lambda}]$ in Proposition 3. The reason for the unbounded support with a linear cost function is that a trader's misalignment rate converges to a constant, while her trading profits scale linearly with her contact rate, since her trading opportunities no longer change. In fact, linear scaling of the benefits of contacts tells us that the cost function must be asymptotically linear for an open tail to emerge. While we have strong intuition for the open tail, we do not have a clear understanding of why the tail turns out to be a Pareto tail with a tail index of 2. Nonetheless, this finding connects closely with empirical evidence as we document below.

Before discussing the evidence, we note that what is mapped out empirically is the distribution of trading rates $\alpha \equiv \lambda p_\lambda$, the product of the contact rate λ and the probability of trading in a meeting, p_λ. Assuming trades occur only if there are strict gains, p_λ is uniquely defined. In particular, trade only occurs when a misaligned trader meets a trader with a strictly higher contact rate or a misaligned trader with the same contact rate; or when a trader, well aligned or misaligned, meets a misaligned trader with a lower contact rate. Let $\hat{G}(\alpha)$ denote the population distribution of trading rates in a limiting equilibrium. Then a corollary to Proposition 4 connects the results describing the distribution of contact rates to the distribution of trading rates:

Corollary 1. *Assume* $C(\lambda) = c\lambda$ *with* $c < \frac{\gamma \Delta}{8r(r+\gamma)(r+2\gamma)}$. *In a limiting equilibrium, the fraction of trades with middlemen is strictly positive; and the trading rate distribution has a Pareto tail with tail index* 2, *meaning* $\lim_{\alpha \to \infty} \alpha^2 (1 - \hat{G}(\alpha))$ *is positive and finite.*

Since a positive fraction of meetings are with middlemen and there is a positive probability of trade in one of these meetings, the first part of the result is immediate. The trading rate inherits the tail properties of the contact rate distribution, because the trading probability conditional on a meeting converges to a positive constant at high contact rates.

We turn now to the empirical content of these results. Our finding that there are middlemen is an extension of the result in Proposition 3 that there is a mass of traders at the upper bound $\bar{\lambda}$, showing that this is not an artefact of a finite upper bound. We view both results as indicating the existence of a "core" of the market, traders who are highly connected both to each other and to the rest of the market, and who intermediate for all other traders. The empirical literature frequently identifies a core of highly connected entities. Craig and Von Peter (2014) define the core as the top tier of banks which constitute a complete graph among themselves. Applying this to the German interbank market, they classify 2.7% of banks as the core. in 't Veld and van Lelyveld (2014), applying the same methodology to the Dutch interbank market, group 13% of banks in the core. Hollifield *et al.* (2017) identify a core of 6–10% of dealers in the inter-dealer derivatives market, accounting for 60–70% of trades (see their Table 4). Di Maggio *et al.* (2017), for the corporate bond market, think of the core as the top 50 dealers, who account for 80% of transactions.

We turn next to the Pareto tail in $\hat{G}$. The empirical literature, motivated by network models, typically measures a trader's *degree*, i.e., the number of counterparties during some interval of time. In our model, a trader's expected degree in a unit time interval is simply equal to the number of trading partners α per unit of time, and hence the degree distribution inherits the Pareto tail of the trading rate distribution.[13] Many papers document that the degree distribution has a Pareto tail in different over-the-counter markets. Examples include Li and Schürhoff (2019) for the municipal bonds market, Hollifield *et al.* (2017) for derivatives, Peltonen, Scheicher and Vuillemey (2014)

13. For a trader with trading rate α, the realized degree during a unit time interval is a Poisson random variable with mean α. Since the standard deviation of a Poisson distribution is equal to the square root of the mean, uncertainty about the realized degree becomes irrelevant when α is large. As a result, a Pareto mixture of Poisson distributions inherits the tail properties of the Pareto distribution.

for the credit default swap market, Bech and Atalay (2010) for the out-degree of banks in the federal funds market, and Boss *et al.* (2004), De Masi, Iori, Precup, Gabbi and Caldarelli (2008), and De Masi, Iori and Caldarelli (2006) for different European interbank markets. We also note that, in the numerical illustration in Section 7.2, we show that the entire trading rate distribution, not just the tail, is well-approximated by a Pareto distribution.

Overall, we thus argue that the key features of the endogenous market structure which arises in our setting connect tightly with a set of stylized facts on over-the-counter markets. It features traders with vastly different amounts of activity, many of whom at least occasionally intermediate for others. It also features a core of a few, highly connected traders who account for a substantial amount of overall activity. And it features a Pareto tail in the degree distribution.

6.3. *Frictionless limit:* $c \to 0$

In many real-world markets, trading frictions are small and so one might question the value of modelling frictions in such markets. Furthermore, advances in information technologies are likely to reduce frictions over time so one might wonder whether this will lead to a diminished role of intermediation and heterogeneity.

This section uses our model to show that intermediation retains its prominent role in the frictionless limit, which we capture through an assumption that the marginal cost of contacts becomes negligible, $c \to 0$. In this case, everyone chooses a fast contact rate and so the aggregate misalignment rate converges to zero. Still, we demonstrate a clear sense in which heterogeneity and intermediation are preserved in the limit.[14] In particular, we obtain a sharp characterization of trading volume, measured as the amount of asset purchases per unit of time.[15]

The following proposition characterizes the overall trading volume along with its decomposition in the frictionless limit.

Proposition 5. *Assume* $C(\lambda) = c\lambda$. *Consider a sequence of limiting equilibria as c converges to zero. The aggregate trading volume* $\mathcal{V}$ *converges to approximately* 2.46γ *and can be decomposed as follows: middlemen's purchases from other middlemen account for a volume* $\mathcal{V}_{mm} = \frac{1}{2}\gamma$; *middlemen's purchases from non-middlemen account for a volume of* $\mathcal{V}_{mn} = \frac{1}{2}\gamma$; *non-middlemen's purchases from middlemen account for a volume* $\mathcal{V}_{nm} = \frac{1}{2}\gamma$; *and non-middlemen's purchases from non-middlemen account for a volume* $\mathcal{V}_{nn} \approx 0.96\gamma$.

To prove this proposition, we first compute trading probabilities in an economy with finite $\bar{\lambda}$, then construct a limiting equilibrium, and then take the limit of trade volume as c converges to zero. See the proof of Proposition 5 for details. The proof also provides the exact expression for volume and the fraction of meetings with middlemen.

We contrast Proposition 5 with a naïve view of a market without frictions: all traders can trade instantaneously upon receiving a taste shock and only trade with other traders who receive the opposite taste shock at the same instant. That means that volume equals the share of traders with taste l times the rate at which they are hit by taste shocks, $\frac{1}{2}\gamma$. Note that this view leaves no role for intermediation or middlemen. In contrast, we obtain nearly five times as much trading

14. We first take a limiting equilibrium, where $\bar{\lambda}$ grows without bound, and then take the limit as c converges to zero. The order of limits is important. With the opposite order of limits, there is no intermediation when c is small. We find this order of limits to be more interesting since in our view the upper bound $\bar{\lambda}$ is present only for technical reasons.

15. We maintain that agents with identical λ and different misalignment status do not trade since the transaction has zero value. With minimum curvature in the utility function, they might well do so and volume would be higher then. In this sense, these results can be viewed as a lower bound on volume.

volume in the frictionless limit. Furthermore, the proposition highlights that a meaningful role for heterogeneity in contact rates and intermediation is preserved in the limiting economy.

To understand this result, note that we are looking at a frictionless limit so almost no one is misaligned. Whenever a trader (who is almost surely not a middleman) suffers a taste shock, she is very likely to become misaligned and very unlikely to contact another misaligned trader. As a consequence, "fundamental" trades between two misaligned traders become exceedingly rare. 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. The volume decomposition shows that, in the frictionless limit, the reallocation of the asset in response to taste shocks runs through an intermediation chain that always involves middlemen. That is, middlemen purchase the asset from (sell the asset to) non-middlemen at exactly the same rate at which asset owners (non-owners) get moved into misalignment by a taste shock.

When a middleman purchases the asset from a slower trader, they too move into misalignment. Afterwards, they find another misaligned middleman with the opposite asset position, both becoming well aligned. As a consequence, trade between middlemen accounts for a volume of $\frac{1}{2}\gamma$. The reason that reallocation always involves middlemen when c is small is that the average misalignment rate of traders with a finite contact rate is proportional to the square of the misalignment rate of middlemen. Thus, as misalignment converges to zero, a misaligned counterparty is almost surely a misaligned middleman, although even misaligned middlemen are scarce.

Taken together, whenever a trader experiences a taste shock, the market rapidly reallocates her asset position. But instead of doing so directly, the position gets traded through an intermediation chain. This chain runs through increasingly faster types towards middlemen, who then reallocate the position internally.

Taken together, intermediaries retain their prominent role in an almost-frictionless setting. This finding connects naturally with the ever-increasing prominence of financial intermediation services despite the massive advances in information technologies in recent decades (Kuprianov, 1993; Philippon, 2015; Biais and Green, 2019).

7. OPTIMAL ALLOCATION

This section examines which trading patterns and contact rate distributions are Pareto optimal. We imagine a hypothetical social planner who can instruct traders both on their choice of λ at birth and on whether to trade in each future meeting but who cannot directly alleviate the search frictions in the economy.

The section first sets up the planner problem and establishes several formal result that parallel the equilibrium characterization. We then contrast equilibrium and optimum numerically and then close with two exercises that document that equilibrium displays excessive trade.

7.1. *Planner problem*

The hypothetical social planner chooses a contact rate measure μ_G as well as symmetric trading patterns in order to maximize steady state utility net of the future cost of meetings:[16]

$$\delta_1 - \Delta \int_{\mathcal{X}} m_\lambda d\mu_G(\lambda) - r \int_{\mathcal{X}} C(\lambda) d\mu_G(\lambda). \tag{11}$$

16. With transferable utility, any Pareto optimal allocation also solves the problem of a utilitarian planner who weights all traders' welfare equally. Since traders do not discount the future (except through exit, in which case they get replaced by another trader), this is equivalent to maximizing undiscounted, i.e., steady state, utility. Thus, the problem we solve here characterizes any allocation that is Pareto optimal among the set of symmetric allocations.

The first two terms gives the flow payoffs from alignment. The average well-aligned trader has a flow payoff of $\delta_1 \equiv \frac{1}{2}(\delta_{h,1}+\delta_{l,0})$, the average flow payoff in the well-aligned state. The average misaligned trader has a flow payoff of $\delta_1-\Delta$. In addition, the planner must pay the search costs when a trader exits and is replaced by a newborn one. These are integrated using the contact rate measure μ_G. The planner also recognizes the misalignment m_λ is endogenous and depends both on the contact rate measure and on the choice of who trades with whom, both of which are under the planner's control.

In Supplementary Appendix D.1, we use calculus of variations to derive necessary conditions characterizing optimality. First, we show that there is a social surplus function $S(\lambda)$, which tells us the gain the planner enjoys by moving a trader from the misaligned state to the well-aligned state. We prove that this satisfies

$$\begin{aligned}(r+2\gamma)S(\lambda) = \Delta + \frac{\lambda}{2}\int_{\mathcal{X}} \Big(&\big((S(\lambda')-S(\lambda))^+ - (S(\lambda)+S(\lambda'))^+\big)m_{\lambda'} \\ &+ \big((-S(\lambda)-S(\lambda'))^+ - (S(\lambda)-S(\lambda'))^+\big)(1-m_{\lambda'})\Big) d\mu_F(\lambda'). \end{aligned} \tag{12}$$

This is the direct counterpart to equation (8). The only difference is that the planner internalizes the full gains from trade in each transaction, whereas an individual trader only internalizes her own half.

Second, we prove that trade occurs if and only if it results in an increase in total social surplus. This gives us the flow balance equation governing the stationary distribution of misalignment:

$$\begin{aligned}&\left(r+\gamma+\frac{\lambda}{2}\int_{\mathcal{X}} \big(\mathbb{I}_{S(\lambda)+S(\lambda')>0}m_{\lambda'} + \mathbb{I}_{S(\lambda)>S(\lambda')}(1-m_{\lambda'})\big) d\mu_F(\lambda')\right) m_\lambda \\ &= \left(\gamma+\frac{\lambda}{2}\int_{\mathcal{X}} \big(\mathbb{I}_{S(\lambda)<S(\lambda')}m_{\lambda'} + \mathbb{I}_{S(\lambda)+S(\lambda')<0}(1-m_{\lambda'})\big) d\mu_F(\lambda')\right)(1-m_\lambda). \end{aligned} \tag{13}$$

This is exactly analogous to equation (9), with social surplus playing the role of private surplus.

Third, we define the social value function

$$\begin{aligned} r\Pi_\lambda = &-\gamma S(\lambda) \\ &+\frac{\lambda}{2}\int_{\mathcal{X}} \big((S(\lambda')-S(\lambda))^+ m_{\lambda'} + (-S(\lambda)-S(\lambda'))^+(1-m_{\lambda'})\big) d\mu_F(\lambda') - \bar{\theta}\lambda - rC(\lambda), \end{aligned} \tag{14}$$

where

$$\bar{\theta} = \int_{\mathcal{X}} \frac{\gamma-(r+2\gamma)m_\lambda}{\lambda} S(\lambda) d\mu_F(\lambda). \tag{15}$$

Equation (14) is analogous to the equilibrium condition (10), with two important differences. First, the planner values the whole surplus from a trade, while in equilibrium each party only gets half the surplus. This is similar to the difference between equations (8) and (12). And second, the planner internalizes that there is a cost $\bar{\theta}$ from each meeting, approximately equal to the annuitized marginal cost of meetings for an average counterparty, $\bar{\theta} \approx r\int_{\mathcal{X}} C'(\lambda) d\mu_F(\lambda)$.[17] This reflects the fact that an increase in one trader's contact rate diverts meetings from other traders, something individuals do not internalize in equilibrium. We return to these differences when discussing efficiency in Section 7.3.

17. See Supplementary Appendix D.2 for details on this approximation.

We prove that $\mu_F(\mathcal{Y}^P)=1$, where $\mathcal{Y}^P=\text{argmax}_{\lambda\in\mathcal{X}}\Pi_\lambda$ is the set of contact rates that maximize the social value function. Thus the planner only utilizes contact rates that maximize this measure of social value.

Putting this together, any symmetric Pareto optimal allocation can be characterized using a surplus function S, a misalignment rate m, and a counterparty distribution μ_F such that equations (12) and (13) are satisfied and $\mu_F(\mathcal{Y}^P)=1$, where $\mathcal{Y}^P=\text{argmax}_{\lambda\in\mathcal{X}}\Pi_\lambda$ and Π_λ is defined in equation (14). This has an almost-identical mathematical structure to the definition of equilibrium. In the remainder of Supplementary Appendix D, we leverage this to show through a series of propositions, structured to mimic the equilibrium Propositions 1–5, that the equilibrium and optimum allocations are qualitatively similar. We summarize those results here.

We first show in Lemma 1-P that the equilibrium trading pattern is optimal. Trade occurs whenever two misaligned traders with the opposite asset holdings meet. It also occurs whenever a slower misaligned trader meets a faster well-aligned trader with the opposite asset holding. The intuition is straightforward. The planner's objective function boils down to minimizing the average rate of misalignment for a given distribution of contact rates. 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, since this does not affect the current misalignment rate but improves future trading possibilities. That is, the planner uses faster traders as intermediaries.

Next, in Proposition 1-P, we show that any counterparty distribution satisfies the necessary conditions for optimality for some cost function C. This means that we cannot make any inference about the efficiency of the observed contact rate distribution unless we have independent knowledge of the cost function.

In Proposition 2-P, we prove that when marginal cost is continuous, the optimal contact rate distribution is continuous on $[0,\bar{\lambda})$. The atomless feature of the optimal contact rate distribution allows the planner to leverage the gains from meetings through intermediation. Any meeting between two traders with identical contact rates λ is beneficial solely when both are misaligned. In contrast, when two traders with different contact rates meet each other, the meeting is socially beneficial even if only the slower trader is misaligned, since there are gains from intermediation. An atomless distribution maximizes the fraction of meetings in which there are gains from trade. Similarly, if marginal cost is Lipschitz, the planner does not want to place too much mass in a neighbourhood of any interior contact rate, ensuring that the counterparty distribution is absolutely continuous on $[0,\bar{\lambda})$.

Proposition 3-P shows that with a linear cost function, it is optimal to have one of three configurations, depending on the level of marginal cost. Autarky is optimal when marginal cost is high. A degenerate trade allocation, with everyone at $\bar{\lambda}$, is optimal when marginal cost is low. And an intermediated trade allocation is optimal for intermediate values. In this allocation, the support of the contact rate distribution is an interval $[\underline{\lambda},\bar{\lambda}]$ and the misalignment rate is increasing on this support.

We then extend this to the limiting case with $\bar{\lambda}\to\infty$. In Proposition 4-P, we prove that with a linear cost function, the optimal distribution has a Pareto tail with parameter 2 and features middlemen. The planner introduces middlemen for reasons that mimic the equilibrium case. If there were no middlemen, then the misalignment rate of the fastest traders would be $\frac{1}{2}$, higher than the misalignment rate in autarky. To compensate, it would have to be the case that fast traders' meetings generate social value in excess of their cost rc. But that would imply that the planner would want to increase their contact rate, thereby creating middlemen.

Finally, we consider the frictionless limit, where the marginal cost of contacts converges to zero. In Proposition 5-P, we calculate trading volume. While the volume of trades involving middlemen is unchanged from equilibrium, we prove that there are optimally fewer trades between

pairs of non-middlemen than occur in equilibrium. That is, the planner relies to a greater extent on middlemen, making them account for a larger fraction of meetings than would occur in equilibrium.

7.2. *Equilibrium vs. optimum: numerical illustration*

We next contrast limiting equilibria with the optimal allocation in the linear cost case. The resulting model only has four parameters, the exit rate r, the arrival rate of preference shocks γ, the cost of contacts c, and the average benefit from alignment Δ. It is straightforward to prove that both equilibrium and optimal allocations are homogeneous of degree zero in c and Δ. Doubling both doubles the surplus s and S but affects neither the counterparty distribution nor the misalignment rate. Effectively these determine the size of a unit of payoff. Similarly, the equilibrium and optimal allocations are homogeneous of degree one in r, γ, Δ, and $1/c$. Doubling the first three parameters and cutting c in half leads to new equilibrium and optimal allocations in which everyone chooses twice as high a contact rate without changing their surplus or misalignment. Effectively, this scaling determines the length of a unit of time. Putting these two observations together, we conclude that only two of the four parameters can qualitatively affect equilibrium or optimal allocations.

With this in mind, we normalize the length of a time period to 1 year and impose $r=0.05$ and $\gamma=2.75$, consistent with Duffie *et al.* (2007). To start, we fix $c=0.001\Delta$, but later consider the robustness of our results to other values of the cost. We take the limit of equilibria as $\bar{\lambda}\to\infty$.[18]

The red lines in Figure 2 summarize the optimal allocation, while the blue lines show the equilibrium. The top left panel shows that the equilibrium contact rate distribution first order stochastically dominates the optimal one. That is, the equilibrium displays excessive investment in contacts across the board. We revisit this observation in the next subsection where we complement it with two theoretical exercises that relate overinvestment to the model's externalities. We also note that equilibrium and optimal contact rate distributions are both well-approximated by a Pareto distribution with tail parameter 2, a line with slope -2 in the figure.

The bottom left panel shows that these features carry over to the distribution of trading rates. In particular, although equilibrium trading rates are too high, both the equilibrium and optimal trading rate distributions have a Pareto tail with parameter 2. This implies 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. We also note in the bottom left panel that 97% of traders in the equilibrium allocation and 93% in the optimal allocation have a trading rate that exceeds the arrival rate of preference shocks γ, a natural upper bound for an economy without intermediation. The explanation for these additional trades is intermediation.

If the socially optimal contact rate distribution were just a proportionately shifted version of the equilibrium contact rate distribution (e.g. everyone had half as many contacts), the counterparty distributions would be shifted by the same proportion. The top right panel in Figure 2 shows that this is not the case. The equilibrium counterparty distribution does not first order stochastically

18. As mentioned previously, we do not have a uniqueness proof. Nevertheless, our numerical calculations are consistent with the depicted allocations, both in equilibrium and optimum, being unique. In particular, recall that with a linear cost function, the equilibrium and optimal allocations are both described by a lower bound $\underline{\lambda}$ and a pair of well-behaved ordinary differential equations in F and M. To verify that this is an equilibrium or optimal allocation, we then compute the implied marginal cost of contacts c. That is, we have a mapping from $\underline{\lambda}$ to c. Numerically, this relationship appears to be monotonically decreasing, so higher cost is associated with a smaller lower bound on contacts; see the top left panel of Figure 3. This would imply that the equilibrium and optimal allocations are unique for arbitrary c. We therefore refer to *the* equilibrium/optimum in this section.

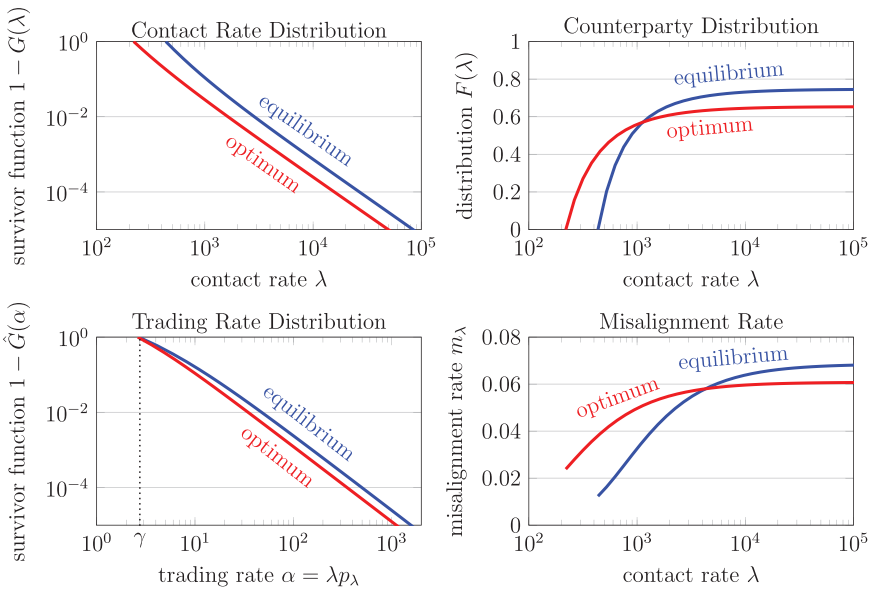

FIGURE 2

Equilibrium and optimal counterparty distribution, contact rate distribution, trading rate distribution, and misalignment rates. We set $c/\Delta=0.001$, $r=0.05$, and $\gamma=2.75$. The dotted line in the bottom left panel indicates the value of γ, the maximum trading rate needed for fundamental trades.

dominate the optimal one. Instead, we find that in the socially optimal allocation, a larger fraction of meetings are with middlemen and other fast traders compared with the equilibrium. This ensures that fast traders more often encounter other fast traders. Since faster traders are more frequently misaligned, this facilitates fundamental trades, reducing their misalignment rate.

Finally, the bottom right panel in Figure 2 plots the misalignment rate as a function of the contact rate in both equilibrium and the optimal allocation. As expected from Propositions 3 and 3-P, traders with a higher contact rate have a higher misalignment rate. Moreover, we can see that the equilibrium misalignment rate of fast traders is too high. This reflects the relative scarcity of meetings with middlemen and other fast traders in the decentralized equilibrium.

Figure 3 shows the robustness of these results to the level of cost c/Δ. The top left panel shows that the lower bound on the optimal contact rate distribution is lower than under the equilibrium contact rate distribution. The top right panel shows that the average contact rate is higher relative to the lower bound in the optimum than the equilibrium. The bottom left panel shows that middlemen play a more prominent role in the optimum, as we found in the example with $c/\Delta=0.001$. And the bottom right panel shows that both total and intermediation volume is inefficiently high in equilibrium. These last three results are all consistent with our analytical results with vanishing costs. We show here that they hold more generally. We find qualitatively similar results with other values of r and γ.

7.3. *Equilibrium vs. optimum: excessive trade*

Although the qualitative features of the equilibrium and optimal allocations are nearly identical, the equilibrium allocation is still inefficient. This section discusses the externalities and subsequently shows, using two separate formal arguments, that the planner discourages investment in contacts, consistent with our numerical results.

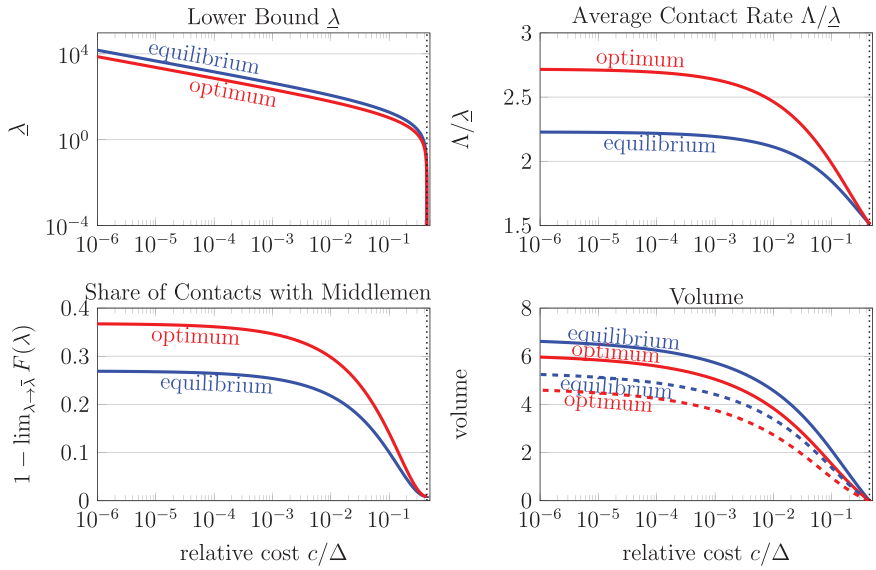

FIGURE 3

Features of the equilibrium and optimal allocation given relative costs c/Δ with $r=0.05$ and $\gamma=2.75$. The vertical dotted line indicates the value of c/Δ, where the market shuts down, $\bar{c}^*$, the same in both equilibrium and the optimal allocation. The solid lines in the bottom right panel show total volume, while the dashed lines show intermediation volume.

The inefficiency is rooted in externalities in the decentralized contact technology. When a trader invests in more contacts, she diverts contacts towards herself and away from other traders. This reflects the fact that investing more in meetings does not affect the contact rate of the other traders, but it changes the distribution of whom they meet.

The externalities can be seen by comparing equation (8) with (12) and equation (10) with (14). The external cost of one trader increasing her contact rate is that doing so reduces the rate that other traders meet each other. We capture this *congestion externality* in our characterization of the social optimum by imposing a constant cost $\bar{\theta}$ on meetings in equation (14). The external benefit of the trader increasing her contact rate is that other traders value meeting her and she only captures half of this due to Nash bargaining. We capture this *thick-market externality* in our characterization by doubling the value of trade in equations (12) and (14). Hence, the social surplus captures the full joint value of these exchanges. This is in contrast to the private surplus, which disregards the half that accrues to the counterparty.

The surplus in a typical trade, and hence the thick-market externality, is generally higher for slower traders, reflecting the fact that the social surplus function is globally decreasing (Lemma 1-P). Conversely, the congestion externality is constant for all traders. We can directly correct for each of these externalities. In Supplementary Appendix E.1, we show that a simple tax and subsidy scheme, where traders get type-specific payments which depend on their alignment status, decentralizes the planning allocation. We show that the scheme subsidizes the misaligned state relative to the well-aligned state. Interestingly, when the cost function C is linear, the marginal subsidy, averaging across the misaligned and well-aligned states, is zero, so Pigouvian taxes do not distort investment choices by transferring resources directly between traders. Instead, the tax and subsidy scheme works by manipulating the threat points in bargaining through subsidies to misaligned traders and taxes on well-aligned ones. This shifts the terms of trade in favour of slower traders, which in turn discourages investment.

We offer a complementary perspective on overinvestment in Supplementary Appendix E.2. We again study the linear cost case and consider a situation where the counterparty distribution F is exogenously given at its socially optimal level, but prices are set through decentralized bargaining without taxes or subsidies. In other words, we drop the third part of the definition of equilibrium. We then examine the incentives of a single trader who enters such a marketplace and can choose her contact rate to maximize her expected revenue net of investment costs, π_λ in equation (10). While the social planner is indifferent across all values $\lambda \geq \underline{\lambda}$, we show that private payoffs are not constant. Instead, an individual trader confronted with the optimal distribution F has a strictly increasing and unbounded profit function π. Aligning with the intuition above, this shows that private incentives lead to excessive investment in equilibrium. Put differently, optimal policy must discourage private incentives to overinvest.

In closing, we connect the difference between the equilibrium and optimal allocation with ongoing debates about financial or securities transaction taxes (FTT or STT) (Tobin, 1978; Burman *et al.*, 2016; Hemmelgarn, Nicodème, Tasnadi and Vermote, 2016). The US currently implements an FTT, set at roughly 2 cents per $1,000 traded (SEC, 2019; Klein, 2020). Many European countries impose FTTs at varying levels covering stocks, bonds, and derivatives.

Our results show that there is overinvestment and excessive trade in equilibrium, which broadly makes the case for policies that discourage trade. More specifically, our model implies that low-value transactions, e.g., intermediation trades between two fast traders, should be taxed more heavily than high-value transactions, e.g., fundamental trades between two slow misaligned traders. Trades between a slow trader and a middleman should face an intermediate tax. Thus, our model offers a novel and natural rationale for policies that selectively tax traders in the core of the financial network while going easy on infrequent market participants with low volume. We stress, however, that the goal of the tax should not be the elimination of intermediation. On the contrary, we have seen that an optimal policy reduces the number of meetings across the board but, in relative terms, redirects meetings towards middlemen and other fast traders.

8. CONSTRAINED ECONOMY: THE ROLE OF INTERMEDIATION

Without intermediation, our model would not generate dispersed contact rates. To prove this, we consider an economy in which meetings between two traders with the same tastes do not occur. It follows that whenever a misaligned trader meets a well-aligned trader, they have opposite tastes and hence the same asset holdings, and so there is no possibility of intermediation. We show in this section that without intermediation, the equilibrium and optimal distributions of contact rates are degenerate as long as the cost function $C(\lambda)$ is weakly convex.

In Supplementary Appendix F, we first offer the adjusted definition of equilibrium along with the adjusted planner problem that correspond to this setting. We then prove the following result:

Proposition 6. *Consider an economy with no intermediation and a weakly convex cost function $C:\mathcal{X}\to\mathbb{R}$. In equilibrium, all traders choose a common value λ. The same holds in the solution to the planner's problem.*

The proposition highlights that the heterogeneity that arises in the full economy is an immediate, and socially desirable, consequence of intermediation. When traders are restricted to trades driven by static fundamentals, there is no gain from heterogeneity in the contact rate. This 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 a trader becomes increasingly well aligned, fewer meetings lead to gainful trading opportunities. As a consequence, the optimal

distribution is degenerate. The same is true in equilibrium; with a weakly convex cost function but decreasing returns on the individual level, all individuals choose the same contact rate.

In summary, intermediation and heterogeneity are 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 tastes, misalignment can be transmitted towards the faster trader to facilitate the transfer of the asset to those who desire it.

9. CONCLUSIONS

We study a model of over-the-counter trading in which *ex ante* identical traders invest in a contact technology and participate in bilateral trade. We show that when traders have heterogeneous contact rates, 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 Pareto optimal distribution. We argue that an economy with mass points in the interior of the contact rate distribution is neither an equilibrium nor socially desirable when the cost of meetings is differentiable. Under a linear cost function, the equilibrium and optimal distributions of trading rates are governed by a power law, an empirical feature of various financial markets. Moreover, middlemen with the highest possible 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 close by highlighting areas for future research. An important one is a general existence result, beyond the linear cost case. Furthermore, 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 interesting to extend our model to a more complex environment, for example, one in which traders differ *ex ante* in how much they care about having a well-aligned asset position. This might "purify" the mixed strategy equilibrium we study here. Recall that under natural restrictions on the cost function, slow traders' asset positions are well aligned with their taste more often than the faster traders who intermediate for them. We therefore conjecture that traders who care the least about their alignment status are the natural intermediaries and have the highest incentives to invest in a high contact rate. 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.

Acknowledgments. This article was previously circulated under the title "Meeting Technologies in Decentralized Asset Markets." We are grateful to Rebekah Dix for research assistance, and to Fernando Alvarez, Markus Brunnermeier, Xavier Gabaix, Ricardo Lagos, Moritz Lenel, Hugo Lhuillier, Pierre-Olivier Weill, four anonymous referees and the editor, and audiences at various seminars and conferences for their thoughts and comments.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online. And the replication packages are available at https://dx.doi.org/10.5281/zenodo.5951869.

Data Availability Statement

The data underlying this article and the numerical programs which generate it are available at https://doi.org/10.5281/zenodo.5951869.

APPENDIX

A. CONTACT RATE DISTRIBUTION

A.1. *Additional details*

We would like to invert equation (1) to recover μ_G from μ_F, but unfortunately this is impossible without further restrictions. To see why, fix a measure μ_G and a number $\alpha \in (0,1)$ and then define a new measure $\tilde{\mu}_G \equiv \alpha\mu_G(S)+(1-\alpha)\mathbb{I}_{0\in S}$, where the indicator function is 1 if $0 \in S$ and 0 otherwise. Since $\int_S \lambda d\tilde{\mu}_G(\lambda) = \alpha \int_S \lambda d\mu_G(\lambda)$ for all $S \subset \mathcal{B}$, equation (1) implies the two contact rate distributions have the same counterparty distribution. Intuitively, they differ only in the fraction of the population with a zero contact rate. Since these traders never meet anyone, the fraction does not affect the counterparty distribution.

We mitigate this issue by imposing a natural restriction, beyond equation (1), on the share of traders with a zero contact rate, $\mu_G(\{0\})$. Recall that $\mathcal{Y}$ is the set of utility maximizing contact rates. If $0 \notin \mathcal{Y}$, we impose $\mu_G(\{0\}) = 0$. When this is the case, we have $\Lambda \equiv \int_{\mathcal{X}} \lambda d\mu_G(\lambda) > 0$. Then, we can use the Radon–Nikodym theorem and equation (1) to move between the probability measures $\mu_G(\lambda)$ and $\mu_F(\lambda)$:

$$\frac{d\mu_F(\lambda)}{d\mu_G(\lambda)} = \frac{\lambda}{\Lambda}. \tag{A.1}$$

Multiply both sides by $\frac{1}{\lambda}$ and integrate both sides under the measure μ_G to get

$$\int_{\mathcal{X}} \frac{1}{\lambda} \frac{d\mu_F(\lambda)}{d\mu_G(\lambda)} d\mu_G(\lambda) = \int_{\mathcal{X}} \frac{1}{\Lambda} d\mu_G(\lambda) = \frac{1}{\Lambda}$$

or

$$\Lambda = \frac{1}{\int_{\mathcal{X}} \frac{1}{\lambda} d\mu_F(\lambda)}. \tag{A.2}$$

Together equations (A.1) and (A.2) define Λ and μ_G given μ_F whenever $\frac{1}{\lambda}$ is Lebesgue integrable under the measure μ_F and $0 \notin \mathcal{Y}$.

If $\frac{1}{\lambda}$ is Lebesgue integrable under the measure μ_F but $0 \in \mathcal{Y}$, then the right-hand side of equation (A.2) is an upper bound on Λ. For any value of $\Lambda \leq 1/\int_{\mathcal{X}} \frac{1}{\lambda} d\mu_F(\lambda)$ and any set S with $0 \notin S$, we can then find $\mu_G(S)$ using equation (A.1). We then set $\mu_G(\{0\}) = 1 - \mu_G((0, \bar{\lambda}])$. This is a valid contact rate measure associated with the counterparty measure μ_F.

Finally, if $\frac{1}{\lambda}$ is not Lebesgue integrable under the measure μ_F, then $\Lambda = 0$ and $\mu_G(\{0\}) = 1$. This is the case whenever $\mu_F(\{0\}) > 0$.

A.2. *Physical matching process*

Consider an economy with a (large) finite number of traders $n \geq 2$ and (short) finite periods of length dt. Each trader has a type λ satisfying $0 \leq \lambda \leq \bar{\lambda}$, with $dt < 1/\bar{\lambda}$. The type determines the probability of matching during each period. Matching proceeds in two stages. First, a trader with type λ draws a binary outcome, 0 or 1, with probability $1-\lambda dt$ and λdt, respectively. Let $k \leq n$ denote the number of traders who draw 1. If k is even, all traders who drew 1 match in pairs, with all such pairings equally likely. If k is odd, uniformly randomly select one trader who drew 0 and switch that outcome to 1. Then match all 1's in pairs, with all such pairings equally likely.

We are interested in the behaviour of μ_G and μ_F when $n \to \infty$. In such an economy, the per-period matching probability for a type λ trader converges to λdt. This is because $\bar{\lambda} dt < 1$ ensures that many traders draw 0. Since at most one trader draws 0 and gets switched to 1, the likelihood of this happening to any particular trader is infinitesimal.

Now consider an economy in autarky, where everyone sets $\lambda = 0$. A single trader can deviate and set a strictly positive value for λ, in which case he matches with a counterparty with contact rate 0 with probability λdt in each period. In this case, $\mu_F(\{0\}) = \mu_G(\{0\}) = 1$.

However, this environment can also accommodate a limit where $\mu_G(\{0\}) = 1$ but $\mu_F(\{0\}) < 1$. That is, autarky does not imply a degenerate counterparty distribution. To see how, fix a number $p > 0$ and consider an economy with $n > p$ traders. Assume each trader has contact rate $\lambda > 0$ with probability p/n and 0 otherwise. The traders then draw 0 or 1 as described above. For finite n, the number of traders who draw a 1 each period is a Binomial random variable with parameters $(n, p\lambda dt/n)$. As n grows without bound, this converges to a Poisson random variable with mean $p\lambda dt$ and hence density $e^{-p\lambda dt}(p\lambda dt)^k/k!$ for $k = 0, 1, 2, \ldots$.

Using this, one can compute the counterparty distribution for those who have chosen a contact rate $\lambda > 0$. (For those who chose a zero contact rate, all counterparties have a contact rate λ, but those matches almost never occur.) For example, if in a given period $k = 1$, then that trader matches with a zero-contact-rate trader. If $k = 2$, the two traders match with each other, i.e., with a λ-contact-rate trader. If $k = 3$, it is a mix of the previous cases, so the counterparty distribution for

a λ-contact-rate trader puts weight $1/3$ on a zero-contact-rate counterparty. Summing across k and weighting implies that the fraction of traders who match with a zero contact rate partner is

$$\frac{\sum_{k \text{ odd}} \frac{e^{-p\lambda dt}(p\lambda dt)^k}{k(k!)}}{p\lambda dt},$$

where $e^{-p\lambda dt}(p\lambda dt)^k/k!$ is the density of the Poisson distribution, $1/k$ is the odds of matching with a zero partner when k is odd, and $p\lambda dt$ is the expected number of traders who choose $\lambda > 0$ and match. Through an appropriate choice of p, this sum can take any value between 0 and 1. In particular, for $p > 0$, $\mu_F(\{0\}) < 1$.

As such, this example describes a physical matching process where as $n \to \infty$, $\mu_G(\{0\}) = 1 > \mu_F(\{0\})$. The key is that in a large economy, almost everyone has a zero contact rate, but counterparties are very different than the typical trader.

B. DIFFERENTIAL EQUATION SYSTEM

In Supplementary Appendix C.3, we prove that on the interior of its support, (F, m, s) solves the following differential equation system:

$$F'(\lambda) = \frac{\big(2r+4\gamma+\lambda(1-F(\lambda)+2M(\lambda))\big)\big(8\gamma(1-F(\lambda))-8rM(\lambda)+\zeta(\lambda)\big)}{2\lambda\Big(\gamma(8r+8\gamma+3\lambda(1-F(\lambda)))+\lambda M(\lambda)(3r+6\gamma+\lambda(1-F(\lambda)+M(\lambda)))\Big)}, \tag{B.1}$$

$$M'(\lambda) = \frac{\big(2\gamma+\lambda M(\lambda)\big)\big(8\gamma(1-F(\lambda))-8rM(\lambda)+\zeta(\lambda)\big)}{2\lambda\Big(\gamma(8r+8\gamma+3\lambda(1-F(\lambda)))+\lambda M(\lambda)(3r+6\gamma+\lambda(1-F(\lambda)+M(\lambda)))\Big)}, \tag{B.2}$$

$$s'(\lambda) = \frac{4((r+2\gamma)s(\lambda)-\Delta)}{\lambda\big(4(r+2\gamma)+\lambda(1-F(\lambda)+2M(\lambda))\big)}, \tag{B.3}$$

where

$$\zeta(\lambda) \equiv \frac{r\lambda\big(4(r+2\gamma)+\lambda(1-F(\lambda)+2M(\lambda))\big)^2 C''(\lambda)}{\Delta-(r+2\gamma)s(\lambda)}; \tag{B.4}$$

see equations (34) and (50)–(52). Moreover, we have terminal conditions $F(\underline{\lambda}) = M(\underline{\lambda}) = 0$ and $s(\bar{\lambda}) = \frac{2\Delta}{2r+4\gamma+\bar{\lambda}M(\bar{\lambda})}$, as well as the requirement that F and M are non-decreasing.

If the cost function is linear, $C''(\lambda) = \zeta(\lambda) = 0$ and so we can solve the differential equations (B.1) and (B.2) for F and M alone.

REFERENCES

AFONSO, G. and LAGOS, R. (2015), "Trade Dynamics in the Market for Federal Funds", *Econometrica*, **83**, 263–313.

BECH, M. L. and ATALAY, E. (2010), "The Topology of the Federal Funds Market", *Physica A: Statistical Mechanics and its Applications*, **389**, 5223–5246.

BIAIS, B. and GREEN, R. (2019), "The Microstructure of the Bond Market in the 20th Century", *Review of Economic Dynamics*, **33**, 250–271.

BOSS, M., ELSINGER, H., SUMMER, M. and THURNER, S. (2004), "Network Topology of the Interbank Market", *Quantitative Finance*, **4**, 677–684.

BURDETT, K. and MORTENSEN, D. T. (1998), "Wage Differentials, Employer Size, and Unemployment", *International Economic Review*, **39**, 257–273.

_____ AND JUDD, K. L. (1983), "Equilibrium Price Dispersion", *Econometrica*, **51**, 955–969.

BURMAN, L. E., GALE, W. G., GAULT, S., KIM, B., NUNNS, J. and ROSENTHAL, S. (2016), "Financial Transaction Taxes in Theory and Practice", *National Tax Journal*, **69**, 171–216.

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

CABRALES, A., CALVÓ-ARMENGOL, A. and ZENOU, Y. (2011), "Social Interactions and Spillovers", *Games and Economic Behavior*, **72**, 339–360.

CHANG, B. and ZHANG, S. (2021), "Endogenous Market Making and Network Formation". Available at SSRN 2600242.

COEN, J. and COEN, P. (2021), "A Structural Model of Liquidity in Over-the-Counter Markets" (Mimeo, LSE).

CONT, R., MOUSSA, A. and SANTOS, E. B. (2010), "Network Structure and Systemic Risk in Banking Systems", *Edson Bastos e, Network Structure and Systemic Risk in Banking Systems*, 327–368.

CRAIG, B. and VON PETER, G. (2014), "Interbank Tiering and Money Center Banks", *Journal of Financial Intermediation*, **23**, 322–347.

CURRARINI, S., JACKSON, M. O. and PIN, P. (2009), "An Economic Model of Friendship: Homophily, Minorities, and Segregation", *Econometrica*, **77**, 1003–1045.

DE MASI, G., IORI, G. and CALDARELLI, G. (2006), "Fitness Model for the Italian Interbank Money Market", *Physical Review E*, **74**, 066112.

———, ———, O. V. PRECUP, GABBI, G. and CALDARELLI, G. (2008), "A Network Analysis of the Italian Overnight Money Market", *Journal of Economic Dynamics and Control*, **32**, 259–278.
DI MAGGIO, M., KERMANI, A. and SONG, Z. (2017), "The Value of Trading Relations in Turbulent Times", *Journal of Financial Economics*, **124**, 266–284.
DUFFIE, D. and MANSO, G. (2007), "Information Percolation in Large Markets", *American Economic Review*, **97**, 203–209.
———, GÂRLEANU, N. and PEDERSEN, L. H. (2005), "Over-the-Counter Markets", *Econometrica*, **73**, 1815–1847.
———, ———, AND ———, (2007), "Valuation in Over-the-Counter Markets", *Review of Financial Studies*, **20**, 1865–1900.
———, DWORCZAK, P. and ZHU, H. (2017), "Benchmarks in Search Markets", *The Journal of Finance*, **72**, 1983–2044.
———, MALAMUD, S. and MANSO, G. (2009), "Information Percolation with Equilibrium Search Dynamics", *Econometrica*, **77**, 1513–1574.
DUGAST, J., ÜSLÜ, S. and WEILL, P.-O. (2019), "A Theory of Participation in OTC and Centralized Markets" (Technical Report, National Bureau of Economic Research).
FARBOODI, M., JAROSCH, G. and SHIMER, R. (2017), "The Emergence of Market Structure" (NBER WP 23234).
———, ———, AND ———, (2021), "Internal and External Effects of Social Distancing in a Pandemic", *Journal of Economic Theory*, **196**, 105293.
———, ———, MENZIO, G. and WIRIADINATA, U. (2018), "Intermediation as Rent Extraction" (NBER WP 24171).
FRICKE, D. and LUX, T. (2015), "Core–Periphery Structure in the Overnight Money Market: Evidence from the E-mid Trading Platform", *Computational Economics*, **45**, 359–395.
GALEOTTI, A. and MERLINO, L. P. (2014), "Endogenous Job Contact Networks", *International Economic Review*, **55**, 1201–1226.
HEMMELGARN, T., NICODÈME, G., TASNADI, B. and VERMOTE, P. (2016), "Financial Transaction Taxes in the European Union", *National Tax Journal*, **69**, 217.
HENDERSHOTT, T., LI, D., LIVDAN, D. and SCHÜRHOFF, N. (2020), "Relationship Trading in Over-the-Counter Markets", *The Journal of Finance*, **75**, 683–734.
HOLLIFIELD, B., NEKLYUDOV, A. and SPATT, C. (2017), "Bid-Ask Spreads, Trading Networks, and the Pricing of Securitizations", *The Review of Financial Studies*, **30**, 3048–3085.
HUGONNIER, J., LESTER, B. and WEILL, P.-O. (2020), "Frictional Intermediation in Over-the-Counter Markets", *The Review of Economic Studies*, **87**, 1432–1469.
IN 'T VELD, D. and VAN LELYVELD, I. (2014), "Finding the Core: Network Structure in Interbank Markets", *Journal of Banking & Finance*, **49**, 27–40.
JACKSON, M. O. and WOLINSKY, A. (1996), "A Strategic Model of Social and Economic Networks", *Journal of Economic Theory*, **71**, 44–74.
KLEIN, A. (2020), "What is a Financial Transaction Tax?", *Brookings Policy*. https://www.brookings.edu/policy2020/votervital/what-is-a-financial-transaction-tax-2/
KREMER, M. (1996), "Integrating Behavioral Choice into Epidemiological Models of AIDS", *The Quarterly Journal of Economics*, **111**, 549–573.
KUPRIANOV, A. (1993), "Over-the-Counter Interest Rate Derivatives", *FRB Richmond Economic Quarterly*, **79**, 65–94.
LAGOS, R. and ROCHETEAU, G. (2009), "Liquidity in Asset Markets with Search Frictions", *Econometrica*, **77**, 403–426.
LI, D. and SCHÜRHOFF, N. (2019), "Dealer Networks", *The Journal of Finance*, **74**, 91–144.
MARTINEZ-JARAMILLO, S., ALEXANDROVA-KABADJOVA, B., BRAVO-BENITEZ, B. and SOLÓRZANO-MARGAIN, J. P. (2014), "An Empirical Study of the Mexican Banking System's Network and Its Implications for Systemic Risk", *Journal of Economic Dynamics and Control*, **40**, 242–265.
NEKLUDYOV, A. (2019), "Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers", *Review of Economic Dynamics*, **33**, 57–84.
PAGNOTTA, E. S. and PHILIPPON, T. (2018), "Competing on Speed", *Econometrica*, **86**, 1067–1115.
PELTONEN, T. A., SCHEICHER, M. and VUILLEMEY, G. (2014), "The Network Structure of the CDS Market and Its Determinants", *Journal of Financial Stability*, **13**, 118–133.
PETRONGOLO, B. and PISSARIDES, C. A. (2001), "Looking into the Black Box: A Survey of the Matching Function", *Journal of Economic Literature*, **39**, 390–431.
PHILIPPON, T.(2015), "Has the US Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation", *American Economic Review*, **105**, 1408–1438.
QUERCIOLI, E. and SMITH, L. (2006), "Contagious Matching Games" (Technical Report, Working Paper).
RUBINSTEIN, A. and WOLINSKY, A. (1987), "Middlemen", *The Quarterly Journal of Economics*, **102**, 581–594.
SEC (2019), "Fee Rate Advisory #2 for Fiscal Year 2019".
SHIMER, R. and SMITH, L. (2001), "Matching, Search, and Heterogeneity", *Advances in Macroeconomics*, **1**, 5.
SIDERIS, T. C. (2013), *Ordinary Differential Equations and Dynamical Systems*, Vol. 2 of *Atlantis Studies in Differential Equations* (Paris: Atlantis Press).

TOBIN, J. (1978), "A Proposal for International Monetary Reform", *Eastern Economic Journal*, **4**, 153–159.
ÜSLÜ, S. (2019), "Pricing and Liquidity in Decentralized Asset Markets", *Econometrica*, **87**, 2079–2140.
VAYANOS, D. and WANG, T. (2007) "Search and Endogenous Concentration of Liquidity in Asset Markets", *Journal of Economic Theory*, **136**, 66–104.
WEILL, P.-O. (2008), "Liquidity Premia in Dynamic Bargaining Markets", *Journal of Economic Theory*, **140**, 66–96.
_____, (2020), "The Search Theory of Over-the-Counter Markets", *Annual Review of Economics*, **12**, 747–773.