Title
Essays on Models of Decentralized Markets

Author
Lebeau, Lucie

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Essays on Models of Decentralized Markets

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Economics

by

Lucie Lebeau

Dissertation Committee:
Professor Guillaume Rocheteau, Chair
Professor William Branch
Professor John Duffy
Professor Pierre-Olivier Weill

2021

DEDICATION

To my parents, for instilling in me the love of learning and the value of perseverance.

À mes parents, pour m'avoir transmis le goût d'apprendre et de persévérer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Guillaume Rocheteau. His inspired teaching spurred my interest in models of decentralized markets from the very beginning of my first quarter at UC Irvine. I am forever indebted to him for the countless hours he spent workshopping the papers in this dissertation and for the confidence he placed in me, always affirming that I should—above all—follow my interests. His dedication was a source of inspiration throughout the highs and lows of learning to be a researcher. I am certain it will continue to be as I grow in my research.

Many thanks to the rest of my committee, Pierre-Olivier Weill, John Duffy, and William Branch, for their support and insight throughout writing this dissertation. I am especially grateful to Pierre-Olivier Weill for his help with Chapters 1 and 4. His advice and guidance played an extensive role in the publication of Chapter 1 by helping clarify and refine the paper. I also thank two anonymous referees for their contributions.

I am deeply grateful to my coauthors, who took me under their wing and gave me the invaluable opportunity to learn the craft of writing a research paper by following their example: not only Guillaume Rocheteau, Tai-Wei Hu, and Younghwan In—who graciously let me reprint our joint work in Chapter 2, but also John Duffy, Daniela Puzzello, Sébastien Lotz, and Cathy Zhang.

I would also like to thank the colleagues and friends who helped in many ways throughout the Ph.D. program. Francisco Ilabaca and Greta Meggiorini, my office mates, for constant motivation, moral support, and banter—I am especially thankful to Francisco, without whom I would not have enrolled at UC Irvine in the first place. Paul Jackson, for his mentorship and availability. I was lucky to be able to follow in his footsteps. Cristian Frasser, for his insight into the epistemology of monetary theory, always reminding me to look at the larger picture when I get absorbed in the details. Lukas Altermatt, for imparting his passionate approach to research and for always giving me confidence in the value of my work.

I am grateful to the professors who led me to pursue doctoral research in Economics. First, Jonathan Robinson, who reached out to me, as a foreign exchange student at UC Santa Cruz, encouraging me to write an undergraduate thesis when I had never conceived of research as an endeavor I could pursue. Second, Christine Le Clainche, who opened my horizons to this career path when she hired me as her research assistant during my master's, giving me a window into what academic research entails in practice.

Financial support from the UCI Department of Economics, the UCI School of Social Sciences, the Fellowship in Honor of Christian Werner, and the Sheen T. Kassouf Fellowship are gratefully acknowledged and were integral to the completion of this dissertation.

Finally, I thank Elsevier for giving me permission to reprint Chapters 1 and 2.

VITA

Lucie Lebeau

EDUCATION

Ph.D. in Economics **2021**
University of California, Irvine *Irvine, California*

M.A. in Economics **2016**
University of California, Irvine *Irvine, California*

M.A. in Economics and Public Policy **2015**
Sciences Po (Institut d'Études Politiques de Paris) *Paris, France*

B.A. in Social Sciences **2013**
Sciences Po (Institut d'Études Politiques de Paris) *Paris, France*

RESEARCH FIELDS

Macroeconomics, Monetary Economics, Markets with Frictions, Bargaining

REFEREED JOURNAL PUBLICATIONS

Gradual Bargaining in Decentralized Asset Markets **2021**
Review of Economic Dynamics, with G. Rocheteau, T.W. Hu, and Y. In

Credit Frictions and Participation in Over-the-Counter Markets **2020**
Journal of Economic Theory

TEACHING EXPERIENCE

Instructor, Intermediate Economics III **2020**
University of California, Irvine *Irvine, California*

Teaching Assistant **2015-2021**
University of California, Irvine *Irvine, California*

Teaching Assistant **2014**
Université Paris V Descartes *Paris, France*

ABSTRACT OF THE DISSERTATION

Essays on Models of Decentralized Markets

By

Lucie Lebeau

Doctor of Philosophy in Economics

University of California, Irvine, 2021

Professor Guillaume Rocheteau, Chair

This dissertation studies models of decentralized markets with search and bargaining. Chapters 1, 3, and 4 respectively examine applications to over-the-counter asset markets, the labor and credit markets, and the transmission of contagious diseases through social and economic interactions. Chapter 2 develops a novel bargaining framework to model bilateral negotiations in decentralized asset markets.

Chapter 1 formalizes a Nash bargaining game between two players constrained by capacity decisions made prior to entering the negotiation. In equilibrium, strategic interactions drive capacity choices to zero and shut trade down despite the existence of gains from trade. The game is embedded in a general equilibrium model of decentralized asset trade with credit frictions to investigate the interaction between availability of credit and investors' participation, modeled through their choices of inventory and payment capacity. A partial access to credit is sufficient to restore trade. The strategic interactions between payment capacity and inventory generate endogenous heterogeneity in holdings, trade sizes and prices, and complementarity between money and credit.

Chapter 2 develops a new approach to bargaining, with strategic and axiomatic foundations, into models of decentralized asset markets. According to this approach, which encompasses the Nash (1950) solution as a special case, bilateral negotiations follow an agenda that

partitions assets into bundles to be sold sequentially. We construct two alternating-offer games consistent with this approach and characterize their subgame-perfect equilibria. We show the revenue of the asset owner is maximized when assets are sold one infinitesimal unit at a time. In a general equilibrium model with endogenous asset holdings, gradual bargaining reduces asset misallocation and prevents market breakdowns.

Chapter 3 examines the deterioration of credit availability as a novel explanation for the outwards shift of the Beveridge curve in the US following the Great Recession. The model implements a twist in Wasmer and Weil (2004): instead of looking for a loan to finance their vacancy costs, firms borrow to cover a fixed cost of hiring required to convert a match into a hire. This timing allows labor market efficiency to drop following a productivity shock. I build a monthly index of loan approval and conduct an empirical exercise that confirms the relevance of the credit channel.

Chapter 4 studies the endogenous spread of an infectious disease in a random matching model with pairwise meetings, where economic and social gains arise explicitly from person-to-person contacts. When agents can decide whether to engage in interactions, complementarities in the participation decisions of individuals susceptible to contracting the disease generate a large multiplicity of equilibria through adverse selection. The lower the participation of susceptible agents, the higher the prevalence of infection in the pool of participants, further discouraging the participation of susceptible agents. I document a variety of infection dynamics, including plateaus and multiple waves. Adverse selection leads to too much isolation from susceptible agents, and in the calibrated version of the model, the cost of forgone interactions offsets the welfare gains of flattening the curve and mitigating the human toll. When agents cannot opt out of the market but can instead choose whether to wear a mask, the equilibrium is unique. In the calibrated model the human toll is lower than when considering the participation margin, yet at a significantly smaller cost.

Chapter 1

Credit Frictions and Participation in Over-the-Counter Markets

1.1 Introduction

The wide range of purchase arrangements observed across over-the-counter (OTC) markets suggests a significant role for payment capacity frictions in those markets. For example, while in the market for Fed funds, banks lend reserves to each other overnight with unsecured credit, more than half of gross credit exposure in the market for OTC derivatives is collateralized, predominantly with currency, highlighting much more stringent payment and credit frictions.[1] In parallel, inventory considerations may also constitute a notable constraint in OTC markets. For example, Friewald and Nagler (2019) study the impact of inventory costs on spreads in the OTC market for corporate bonds, and estimate it to be

[1]See Duffie (2011) or Afonso and Lagos (2015) for more institutional details regarding the market for Fed funds. Regarding OTC derivatives, according to BIS data, in 2013, 55% of gross credit exposure was collateralized, 80% of which with currency.

greater than that of search frictions.[2] Those two types of capacity constraints, which are endogenous, are likely to interact with each other. In a 2016 speech, then President of the Federal Reserve Bank of New York William C. Dudley tied a sustained decrease in dealers' holdings of corporate bonds to the contemporaneous contraction of funding available in financial markets.[3]

This paper builds a model of decentralized asset trade that explicitly takes into account the endogenous payment and trade capacity constraints that exist in OTC markets. Doing so requires relaxing several assumptions commonly used in the literature concerned with decentralized markets—in particular, the assumption that agents have deep pockets, sparing them from payment frictions, or that they can produce on the spot with no capacity limit, exempting them from inventory frictions.

The paper first demonstrates that two-sided capacity constraints are indeed relevant for OTC trade. Absent credit, when investors optimally choose inventory and payment capacity (real balances) before they enter the OTC market, strategic interactions result in a complete shutdown of the market, as investors prefer not to participate. The paper then examines the extent to which access to credit can restore participation and encourage money holdings. As soon as some access to credit is introduced, investors carry real balances, and trade is restored in the OTC market. Provided a sufficiently low access to credit, the equilibrium then features endogenous heterogeneity in both asset and money holdings, and complementarity between credit and money.

The first part of the paper formalizes the game played by two investors with quasilinear utility who make capacity choices before entering a bilateral trade negotiation. The natural seller picks what can be interpreted as his maximum trade capacity (or, inventory), while the

[2]Randall (2015) and Rapp (2018) also provide evidence of the importance of frictions related to inventories in the market for corporate bonds.

[3]More precisely, he mentions the deterioration "funding liquidity," which he defines as "the ability of a financial entity to raise cash by borrowing on either an unsecured or a secured basis." See Dudley (2006).

natural buyer picks what can be interpreted as her maximum payment capacity. They then negotiate bilaterally, and the terms of trade are determined by the Nash (1950) bargaining solution. Perhaps surprisingly, neither investor chooses to participate in the market—they both pick capacities of zero—despite the existence of gains from trade. Indeed, not only are payment and trade capacity choices strategic complements, but it is always optimal to marginally undercut the other player's capacity so as to obtain better terms of trade, leading to an unraveling to the bottom and a total breakdown of trade. When a seller's trade capacity binds, the fall in trade volume the buyer would experience by marginally decreasing her payment capacity is more than offset by the price impact, resulting in an increase in her bargaining surplus. A similar mechanism operates on the seller's side. When the buyer's payment capacity binds, the price impact of the seller marginally tightening his trade capacity is greater than the impact on volume traded, so that the seller's surplus increases after a marginal inventory cut.

The breakdown of trade hinges on the players' surpluses being non-monotone in capacity choices, which is driven by the bargaining protocol. As a result, other protocols that feature monotone surpluses, for example, Kalai (1977) bargaining, would allow trade to occur in equilibrium. We argue that bargaining protocols that generate monotone surpluses, however, may not necessarily be easy substitutes in an environment with two-sided ex-ante capacity choices, where the monotonicity can generate indeterminacy. Additionally, the non-monotonicity resulting from the Nash solution may in fact be a natural bargaining outcome in the quasilinear environment considered in the present paper. Hu and Rocheteau (2020) describe a strategic game in N rounds, which in our environment, provides strategic foundations for the Nash solution when $N = 1$ and for the Kalai solution when $N \to \infty$. For any $N < \infty$, surpluses are non-monotone and no trade occurs in equilibrium, Kalai ($N \to \infty$) being the only exception.

Note that the capacity-underinvestment spiral, due to strategic interactions between in-

vestors, is to be distinguished from a two-sided holdup problem, as featured for example in Bethune et al. (2019) or Wright et al. (2020). A holdup problem arises when investors do not earn the total return on a costly, ex-ante investment, leading to a socially suboptimal investment level. Here, the unraveling occurs even when the investment in capacities comes at no cost (ex-ante).

Assuming that trade in OTC markets is subject to budget constraints (e.g., absence of perfect credit or deep pockets) and inventory constraints (e.g., no on-the-spot production or imperfect access to a centralized market allowing to quickly readjust inventory), the existence of such strategic interactions in the optimal choice of those capacity constraints bears important implications for the functioning of the market and for the role of credit.

To investigate this, in the second part of the paper, the bargaining game with two-sided capacity choices is embedded into a general equilibrium model of OTC asset market that resembles Lagos and Zhang (2019a,b, 2020). The model bridges the search-theoretic literature of OTC markets a la Duffie et al. (2005) and search-theoretic models of money a la Lagos and Wright (2005), and naturally generates endogenous two-sided capacity constraints. The model features investors with idiosyncratic valuations over an asset. They have access to a decentralized market subject to search and bargaining frictions where low-valuation investors, who are endowed with the asset, can trade with high-valuation investors. Credit and payments frictions are modeled through the constraint that a credit technology is not always available during trade, in which case investors can use another means of payment, money. Before they enter the market, investors make payment and trade (inventory) capacity decisions, interpreted as participation decisions at the intensive margin.

In the pure-currency specification, with no access to credit, the unique equilibrium is non-monetary and it features a complete shutdown of the OTC market, despite the existence of gains from trade between investors. This stark result follows directly from the results obtained in the first part of the paper.

As partial access to credit is introduced, trade is restored in the OTC market. When access to credit is sufficiently high, the equilibrium is non-monetary and all OTC trade occurs through credit. When access to credit is less frequent, equilibria are monetary and feature endogenous heterogeneity in holdings, quantities traded, and prices, as observed empirically in OTC markets. The more infrequent access to credit, the more heterogeneous the participation of investors. Notably, this heterogeneity is not the product of ex-ante heterogeneity, since all investors are identical within their type.

While credit is often neutral in environments where it coexists with money, it is not the case here. By mitigating the strategic interactions between the choices of trade and payment capacities, the availability of credit has strong implications for allocations, prices, and welfare. In fact, an important equilibrium feature is that money and credit do not behave as substitutes but as complements. The model predicts that as access to credit is reduced, inventories are scaled down, diminishing buyers' incentives to hold money and thereby further tightening payment constraints, which in turn amplifies the drop in inventories. The fall in participation goes with a reduction in trade volume, so that the model features a feedback loop between access to funding and market liquidity, amplifying the deterioration of market conditions during credit crunches.

Those predictions stand in stark contrast to those derived under a competitive asset market specification, highlighting the role of market structure. Competitive forces eliminate the strategic complementarities between capacity choices, making participation choices irrelevant. As a result, when the asset market is competitive, the distributions of payment and trade capacities are degenerate, money and credit behave as substitutes, and aggregate trade volume and welfare remain at first best regardless of credit availability.

The paper is organized as follows. After a review of the related literature, Section 1.2 presents and solves for the bargaining game with two-sided capacity choices. Section 1.3 embeds the

game in a general equilibrium model of decentralized asset trade. Section 1.4 derives the implications of the existence of two-sided capacity constraints, and their interaction with credit frictions, for participation, trade volume, prices and welfare.

Related literature. Endogenous, two-sided capacity constraints in bargaining are key to the results derived in this paper. The game theoretic literature typically imposes an exogenous trade capacity constraint, i.e., a fixed pie, and no payment constraint, i.e., transferable utility. Similar assumptions are used in Duffie et al. (2005), who embed bargaining into a model of decentralized asset market. Subsequent papers in the search-theoretic literature on OTC markets release the restriction on asset holdings (e.g., Lagos and Rocheteau (2007), Lagos et al. (2011)), so that the trade capacity is effectively chosen by investors before a negotiation takes place, but still ignore payment constraints, with the assumption that agents have deep pockets.

The present paper formalizes those payment constraints. They are endogenously determined by the agents' choices of real balances holdings, in the spirit of the New Monetarist literature following Lagos and Wright (2005)—a literature, which, on the other hand, typically ignores inventory restrictions (e.g., allowing for on-the-spot production of the good or asset traded). Aruoba et al. (2007) highlight that in a typical New Monetarist economy with Nash bargaining, agents' payment capacity choice is suboptimal, even when it comes at no cost. The present paper highlights much more dramatic outcomes, up to a complete market breakdown, when the capacity choice occurs on both sides of the market.

Other papers do consider the existence of capacity constraints on both sides of a decentralized market: in the context of a decentralized retail market with production in advance, Dutu and Julien (2008), Masters (2013), Anbarci et al. (2019), Baughman and Rabinovich (2021); in the context of decentralized asset trade embedded in the Lagos and Wright (2005) framework, Geromichalos and Herrenbrueck (2016a), Wright et al. (2020), Lagos and Zhang (2019a,b,

2020), among others. In all of these papers, the strategic interactions highlighted here are prevented by one of the following modeling assumptions: (i) no bargaining, use of Kalai bargaining, or take-it-or-leave-it offers, (ii) existence of a perfectly competitive interdealer market that relaxes the inventory constraint, or (iii) perfect coupling between portfolio and participation decisions, effectively preventing agents from storing some inventory even though they would like to do so.[4]

Additionally, relative to this strand of papers, I focus on the impact of varying levels of credit frictions on the investors' capacity decisions. This ties in with the literature on the coexistence of money and credit. Gu et al. (2016) use a search framework to put forward the result that in an economy where credit is easily accessible, money is irrelevant, whereas if credit is difficult to access, money becomes essential, and renders credit irrelevant. In the present paper, the strategic considerations between investors who make their participation decisions provide a counter-example by making credit and money complements when access to credit is not too high.

The complementarity between money and credit creates feedback between the funding liquidity (ease of access to credit) and market liquidity (ease of trading), a phenomenon that has been studied both empirically and theoretically. Rapp (2018) provides evidence that the financing constraints faced by dealers are a large determinant of their inventory costs. Gorton and Metrick (2012) and Copeland et al. (2014) study the role played by the rise of margins in the repo market during the 2008-09 crisis. Theoretically, Gromb and Vayanos (2002), Weill (2007), and Brunnermeier and Pedersen (2009) show the adverse effect of funding constraints on the supply of liquidity.

[4]Anbarci et al. (2019) mention the existence of the strategic interactions studied in this paper, and their potential to prevent trade. However, those interactions are assumed away by simultaneously giving sellers enough incentives to produce early and preventing them from effectively storing some of that production (i.e., not carrying it to the decentralized market, which agents would like to do). The present paper allows to decouple portfolio choice from participation choice, and studies the ensuing strategic interactions in detail. Note that this decoupling can also be found in Berentsen and Rocheteau (2003).

I interpret the intensive margin choice of inventory carried into the OTC market as a participation choice. Other papers that study the participation choices of investors at the extensive margin include Atkeson et al. (2015) and Dugast et al. (2019).

Finally, one of the main results in the paper is that credit frictions result in endogenous heterogeneity in holdings, trade sizes, and prices. Such heterogeneity has been documented empirically in a variety of OTC markets. See Green et al. (2007) and Li and Schürhoff (2019) regarding the interdealer market for municipal bonds, Hendershott et al. (2017), Friewald and Nagler (2019), and Di Maggio et al. (2017) for the corporate bonds market, Bech and Atalay (2010) for the Fed funds market, Hollifield et al. (2017) for the market for asset-backed securities, Arora et al. (2012) and Eisfeldt et al. (2018) for credit default swaps, or Gavazza (2011) for the commercial aircraft market. On the theoretical front, Hugonnier et al. (2020), Atkeson et al. (2015), Dugast et al. (2019), and Üslü (2019), among others, obtain price dispersion in frameworks based off of Duffie et al. (2005). However, this is the result of ex-ante heterogeneity, e.g., in valuations, meeting rates, trade capacities, endowments, etc. In the present paper, homogeneous investors endogenously make different participation choices as the result of strategic interactions.

1.2 Bargaining with two-sided capacity choices

This section introduces and solves for a two-player bargaining game with two-sided capacity choices that will later be embedded in a general equilibrium model of over-the-counter asset trade. After investigating the importance of the bargaining protocol for the results obtained, we explain how the mechanism driving those results differs from a (two-sided) holdup problem.

1.2.1 Game set-up

Consider the following two-stage perfect information game with two players, player ℓ and player h. In stage 1, player ℓ picks $w \in [0, \Omega]$ while player h picks $z \in [0, \infty]$.[5] In stage 2, the two players bargain over a pair $(y \in [0, w], p \in [0, z])$ using generalized Nash bargaining, where $\theta \in (0, 1)$ denotes agent h's bargaining power. Player ℓ's and player h's payoffs are respectively given by $-iz + u(y) - p$ and $-\psi w + p - c(y)$, where $i \geq 0$, $\psi \geq 0$, $u(0) = c(0) = 0$, $u'(0) > c'(0)$, $u'(y) > 0$, $c'(y) > 0$, and $u''(y) \leq 0 \leq c''(y)$. Let $y^* \equiv \arg\max_{y \in \mathbb{R}^+} [u(y) - c(y)]$.

This game formalizes a negotiation between two players with quasilinear utility over the allocation of two goods. One player gets to determine ex-ante the maximum tradable quantity of the first good, while the other player makes a similar decision for the second good, both incurring a cost proportional to the capacity they choose.[6]

Interpretation While the set up is purposely kept general, the quasilinear payoffs make it natural to interpret p as a payment from player h to player ℓ against y units of a commodity or an asset for which there are gains from trade. Then, z can be seen as a payment capacity, and w as a trade or inventory capacity. Later in the paper, when the game is embedded in a general equilibrium setting, player ℓ and player h will be interpreted as investors, the former being endowed with Ω units of an asset and the latter with none. Because the h-investor values the asset more than her counterpart, for example due to liquidity or hedging needs, there exist gains from trade to be realized in pairwise meetings in an OTC market. Before meeting to trade, both investors make participation decisions at the intensive margin. The ℓ-investor decides how much asset to carry in his portfolio subject to a marginal inventory

[5] The upper bound Ω is assumed without loss of generality and will be helpful to simplify notation later on.

[6] While the game is described as a 2-stage game, only in the first stage do the players make decisions. Note that results would go through if the first stage was followed by a Rubinstein (1982) game of alternating offers.

cost ψ. The h-investor picks her payment capacity, interpreted as money holdings, subject to the opportunity cost of carrying real balances.[7]

1.2.2 Subgame-perfect Nash equilibrium

Proposition 1.1 describes the subgame-perfect Nash equilibrium of the sequential game presented in Section 1.2.1.

Proposition 1.1 (Trade breakdown)**.** *There exists a unique subgame-perfect Nash equilibrium and it is such that* $w = z = p = y = 0$.

Despite the existence of gains from trade between players h and ℓ, they pick payment and trade capacities of zero in stage 1, resulting in the inability to generate any surplus from bargaining in stage 2. This outcome is not only socially inefficient, it is the worst feasible outcome.

To prove Proposition 1.1 and understand the intuition behind it, we proceed by backwards induction and first consider the bargaining problem between the two agents in stage 2, taking as given the choices of z and w. Define $S^h = u(y) - p$ the bargaining surplus of player h and $S^\ell = p - c(y)$ that of player ℓ. The Pareto frontier corresponding to this problem is given by

$$S^h = \begin{cases} u\left[\min(y^*, w)\right] - c\left[\min(y^*, w)\right] - S^\ell & \text{if } S^\ell \leq z - c\left[\min(y^*, w)\right], \\ u\left[c^{-1}(z - S^\ell)\right] - z & \text{otherwise,} \end{cases} \tag{1.1}$$

and it is represented in the left panel of Figure 1.1, both for $w \geq y^*$ (outer frontier) and for $w < y^*$ (inner frontier). The Pareto frontier is linear with slope -1 when z is sufficiently large for $\min(y^*, w)$ to be traded. Indeed, in this case, moving along the frontier corresponds

[7] It could also be seen as a maximum loan size or credit line negotiated exogenously with a creditor.

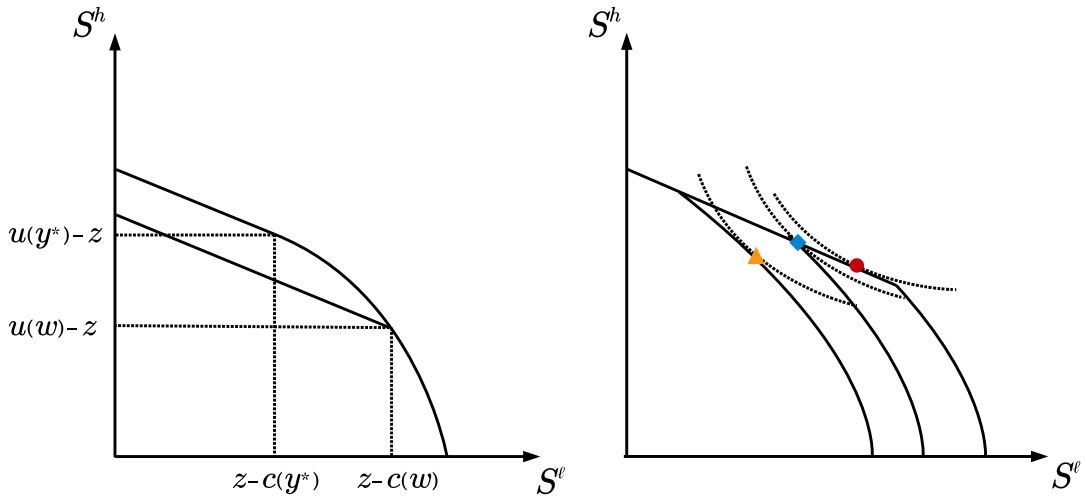

Figure 1.1: Pareto frontiers and equilibrium allocations of the bargaining game. The x-axis corresponds to the ℓ-investor's bargaining surplus, and the y-axis to the h-investor's bargaining surplus. Left panel: Pareto frontiers for $w \geq y^*$ (outer frontier) and $w < y^*$ (inner frontier). Right panel: Pareto frontiers and tangent Nash products for $w < y^*$ and three levels of payment capacity, z. As z increases, the Pareto frontier shifts to the right.

to a pure transfer between the two players. When $\min(y^*, w)$ cannot be achieved, the frontier is strictly concave. In that case, moving along the frontier implies a change in trade volume, which impacts the payoffs of the two players differently. Note that the frontier is not differentiable at $S^\ell = z - c(w)$ provided $w < y^*$.

The Nash problem can be written as

$$max_{y,p} \left[u(y) - p\right]^\theta \left[p - c(y)\right]^{1-\theta} \text{ s.t. } 0 \leq y \leq w \text{ and } 0 \leq z \leq p, \tag{1.2}$$

and its solution is characterized in the following lemma.

Lemma 1.1 (Terms of trade)**.** *Let* $k(y) \equiv (1-\theta)u(y) + \theta c(y)$ *and* $p(y) \equiv [1 - \Theta(y)]\, u(y) + \Theta(y)c(y)$*, where* $\Theta(y) \equiv \theta u'(y)/\left[\theta u'(y) + (1-\theta)c'(y)\right]$*. For* $(z, w) > (0, 0)$*, the terms of*

trade are

$$(y,p) = \begin{cases} (\min(y^*,w), k[\min(y^*,w)]) & \text{if } k[\min(y^*,w)] \leq z, \\ (\min(y^*,w), z) & \text{if } p[\min(y^*,w)] \leq z < k[\min(y^*,w)], \\ (p^{-1}(z), z) & \text{otherwise.} \end{cases}$$

Player h's surplus is

$$S^h(z,w) = \begin{cases} \theta\{u[\min(y^*,w)] - c[\min(y^*,w)]\} & \text{if } k[\min(y^*,w)] \leq z, \\ u[\min(y^*,w)] - z & \text{if } p[\min(y^*,w)] \leq z < k[\min(y^*,w)], \\ u[p^{-1}(z)] - z & \text{otherwise.} \end{cases}$$

Player ℓ's surplus is

$$S^\ell(z,w) = \begin{cases} (1-\theta)\{u[\min(y^*,w)] - c[\min(y^*,w)]\} & \text{if } \min(y^*,w) \leq k^{-1}(z), \\ z - c[\min(y^*,w)] & \text{if } k^{-1}(z) < \min(y^*,w) \leq p^{-1}(z), \\ z - c[p^{-1}(z)] & \text{otherwise.} \end{cases}$$

When $z = 0$ or $w = 0$, the terms of trade are $p = y = 0$ and surpluses are $S^\ell = S^h = 0$.

The terms of trade depend on the relative capacities z and w picked by the players in the first stage. When z is high relative to w, the trade capacity is binding while the payment capacity is not. Player h receives all of w but does not transfer all of z. This is represented by the red dot in the right panel of Figure 1.1. When z is in an intermediate range compared to w, both trade capacities are binding. Player h receives all of w while player ℓ receives all of z. This corresponds to an allocation at the kink of the Pareto frontier (e.g., the blue diamond). Note that this region disappears when $w \geq y^*$, or equivalently, when $z \geq p(y^*)$. Finally, when z is low relative to w, the payment capacity is binding while the trade capacity is not.

Player ℓ receives all of z but only part of w is transferred to player h. This corresponds to an allocation such as the orange triangle.

We now move backwards and solve for the players' optimal choices of z and w. Player h's problem can be written as $\max_{z\in\mathcal{R}^+}\left\{-iz + S^h(z,w)\right\}$ and player ℓ's problem can be written as $\max_{w\in[0,\Omega]}\left\{-\psi w + S^\ell(z,w)\right\}$. Player h's bargaining surplus, $S^h(z,w)$, is non-monotone in z. When $z \leq p[\min(w,y^*)]$, $S^h(z,w)$ is concave in z. It is first increasing and may then be decreasing if $w > \tilde{y}$, where $\tilde{y} = argmax_{y\in[0,\Omega]}\left[u(y) - p(y)\right]$. When $p(w) < z \leq k(w)$, $S^h(z,w)$ is decreasing with slope -1. Finally, when $z > k[\min(w,y^*)]$, it is constant. Player ℓ's bargaining surplus, $S^\ell(z,w)$, can be described similarly. When $\min(y^*,w) < k^{-1}(z)$, $S^\ell(z,w)$ is increasing and concave in w. When $k^{-1}(z) \leq w < p^{-1}(z)$, it is decreasing and concave in w. When $p^{-1}(z) \leq \min(w,y^*)$, it is constant.

We can easily show that player h's objective function is maximized for some $z(w) \leq p\left[\min(\tilde{y},w)\right]$. If $\Omega < y^*$, player ℓ's objective function is maximized for some $w(z) \leq \min[k^{-1}(z),\Omega]$. If $\Omega \geq y^*$, it is maximized for some $w(z) \leq \min\left[k^{-1}(z),y^*\right]$.[8] Given $p(y) < k(y)$ for any $y < y^*$, and $\tilde{y} < y^*$ when $y^* < \Omega$, $z(w)$ and $w(z)$ uniquely intersect at $(z,w) = (0,0)$, concluding the proof of Proposition 1.1.[9] Figure 1.2 sketches the best-response functions for low but non-zero i and ψ, and $y^* < \Omega < \infty$.

Intuitively, a strategic unraveling occurs between the two players, who both attempt to undercut each other so as to obtain better terms of trade. Indeed, because $p(y) < k(y)$ for any $y < y^*$, the payment function $p(y)$ favors player h, while the payment function $k(y)$ benefits player ℓ. Starting from a given trade capacity w_1, player h would pick her payment capacity so as to exhaust the trade capacity at the most favorable price, which is achieved by $z_1 = z(w_1)$. But player ℓ would then adjust his trade capacity down in order to exhaust all of player h's payment capacity at the most favorable price, by picking $w_2 = w(z_1) < w_1$.

[8] The inequalities can be replaced by equality signs for low i and ψ.

[9] Note that the assumption that both sides of the market have some bargaining power, $\theta \in (0,1)$, is necessary for this result to hold.

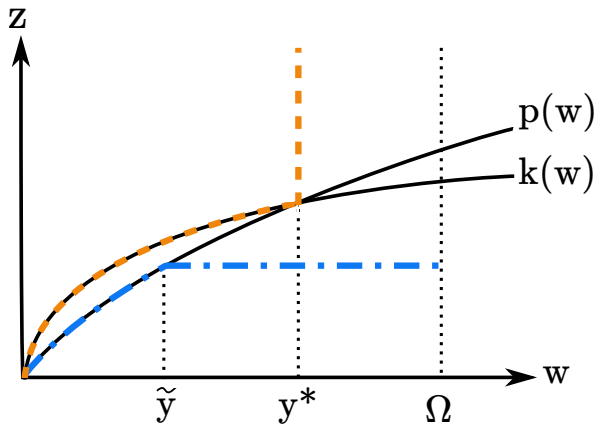

Figure 1.2: Best-response functions for $y^* < \Omega < \infty$: optimal capacity choice of player h, $z(w)$, in dash-dotted blue, and optimal capacity choice of player ℓ, $w(z)$, in dashed orange.

In turn, player h would lower her payment capacity and pick $z_2 = z(w_2) < z_1$, etc. In a sense, capacity choices feature an extreme form of complementarity that leads to a complete unraveling to an autarky equilibrium.

While the argument was made considering only pure strategies, it is easy to check that Proposition 1.1 is robust to mixed strategies. Denote $\bar{z}$ and $\bar{w}$ the highest capacities picked with a positive probability by the players. Following the argument delineated above, we directly obtain that $\bar{z} \leq z(\bar{w})$ and $\bar{w} \leq w(\bar{z})$. These two inequations can only hold jointly for $\bar{w} = \bar{z} = 0$, proving that no mixed-strategy equilibrium with positive capacities exist.

The non-monotonicity of both players' surpluses in the capacities they must pick ex-ante is key for the strategic unraveling to occur and the trade to break down. To understand why the surpluses are not monotone, it is helpful to look graphically at the bargaining outcomes as players vary their capacities, thereby shifting the Pareto frontier. For example, the right panel of Figure 1.1 illustrates the lack of monotonicity of $S^h(z, w)$ in z, in the case when $w < y^*$. Consider an allocation on the steeper part of the frontier (e.g., the orange triangle), where the payment capacity, z, is binding. As player h picks a higher z, the Pareto frontier shifts to the right. Her surplus first increases, as the trade size increases. As she keeps increasing z, she will eventually be able to obtain all w, exhausting player ℓ's trade

capacity. This occurs when the Nash product first hits the kink of the Pareto frontier. At this point, marginally expanding the payment capacity becomes unfavorable. As the Pareto frontier shifts more to the right, the Nash product remains tangent to the kink of the Pareto frontier, i.e., player h transfers a larger payment for the same trade size, w, so that her surplus is decreasing in z. Eventually, the Nash product becomes tangent to the shallower part of the frontier, and the allocation remains the same independently of z.

The ability for players to commit to a maximum payment or trade capacity ex-ante, so as to restrict the bargaining set and impose a more inward Pareto frontier, is therefore the key economic force at play. By doing so, they are able to create market power and impact the trade price in their favor.[10] Were the market perfectly competitive, a single agent's decision to restrict her capacity would not provide her with more market power: the trade price, determined by market clearing, would remain unchanged.

1.2.3 Robustness to other bargaining mechanisms

In the game presented above, the outcome of the bargaining between the two players was assumed to be dictated by the generalized Nash solution. While the Nash solution is arguably one of the most commonly used bargaining mechanisms, there is no definite way to solve for allocations in what amounts to bilateral monopolies. This section investigates the extent to which the trade breakdown described in Proposition 1.1 depends on the mechanism used to determine the terms of trade.

For Proposition 1.1 to stand and an unraveling to the bottom to occur, we need the best-response correspondence of player h, $z(w)$, to always be above that of player ℓ, $w(z)$, in

[10]This relation between capacity constraints and market power is reminiscent of Kreps and Scheinkman (1983), who show that adding capacity constraints to a Bertrand competition leads to the outcomes obtained in a Cournot competition. The incentive to cut one's inventory in order to obtain a better price also has a flavor of the "throw away paradox," highlighted by Gale (1974) and Aumann and Peleg (1974). They show that in a pure-exchange economy, there exist a set of preferences and endowments such that agents would like to throw away some of their endowments in order to obtain higher trade surpluses.

the (w, z) plane, and to uniquely intersect at $(w, z) = (0, 0)$. This requires $S^h(z, w)$ to be strictly decreasing for some range of z and $S^\ell(z, w)$ to be strictly decreasing for some range of w. Otherwise, players would never find themselves better off by decreasing their trade and payment capacities.[11] For that reason, the Kalai and Smorodinsky (1975) proportional solution, which features weakly increasing surpluses, would not lead to a trade breakdown. It may not, however, be a more desirable way of determining the bargaining outcome: applying the Kalai proportional solution to the game described above would lead to equilibrium indeterminacy, with a trade size $y \in [0, y^*]$.[12]

Additionally, in the environment of interest, one could argue that non-monotonicity may in fact be a more generic feature of bargaining outcomes, Kalai being the exception. To do so, we follow an approach developed in Hu and Rocheteau (2020). They describe an N-round extensive-form game which provides strategic foundations to a set of bargaining solutions ranging from the Nash solution to the Kalai solution as N ranges from 1 to ∞. In each round, players negotiate the sale of $\min(w, y^*)/N$ units of the good most valued by player h (i.e., the good that generates gains from trade), according to a Rubinstein (1982) alternating-offer game. When an offer is rejected, the round is over and players move on to the next bundle with some exogenous probability (otherwise, the game ends). The unique subgame-perfect Nash equilibrium of the game can be obtained by applying the Nash bargaining solution iteratively, whereby the constraint on trade size is relaxed by y^*/N units every round, and the disagreement points in round n are given by the outcome of Nash bargaining in round $(n-1)$.

The solution implements the Nash and Kalai solutions as the two extreme cases, $N = 1$

[11] This is not, however, a sufficient condition. The Kalai and Smorodinsky (1975) solution provides an example of non-monotone trade surpluses that do not lead to a trade breakdown. Under this bargaining protocol, the bargaining surplus of player h has a single-valued peak, $z(w)$, that exactly coincides with the single-valued peak of player ℓ's bargaining surplus, $w(z)$, so that strategic undercutting does not occur.

[12] Bargaining surpluses under Kalai bargaining are derived in Appendix A.1. See Baughman and Rabinovich (2021) for a thorough investigation of a similar game under Kalai bargaining. Also, note that the same criticism could be raised against take-it-or-leave-it mechanisms.

and $N \to \infty$, and allows to study bargaining outcomes for any N in between. We show in Appendix A.1 that if players bargain in stage 2 according to this more general specification, for any $N < \infty$ and $\theta \in (0,1)$, the 2-stage game admits a unique equilibrium, $z = w = p = y = 0$. In other words, the trade breakdown is robust as long as players bargain over finitely many bundles of w. Only when players bargain over player ℓ's inventory in infinitely-many rounds, which corresponds to the Kalai solution, are participation and trade restored.

1.2.4 Relation to holdup problems

Inefficiently low trade volumes due to participation, inventory, or investment decisions often result from holdup problems (e.g., see Bethune et al., 2019, Wright at al, 2018, for examples of two-sided holdup problems). A holdup problem is usually defined as a distortion of investment incentives that arises when parties have to make ex-ante, sunk investments before engaging in negotiations. Provided she does not have all of the bargaining power, the agent that bears the investment cost does not receive all of the return on this investment, and therefore she invests a socially inefficiently low amount.

This is not the mechanism operating here. In fact, the unraveling and subsequent trade breakdown described in Proposition 1.1 occur absent any ex-ante investment cost, when $i = \psi = 0$. Players do not pick too little z and w because they enjoy too little of the additional surplus compared to the marginal investment cost. They pick too little z and w because picking any larger capacities would decrease their trading surpluses, regardless of the investment costs born. The inefficiency stems from strategic considerations between the two players, who wish to undercut each other in order to obtain more market power and better terms of trade. They do so by restraining their trade and payment capacities, even when increasing capacity comes at no cost ex-ante.

To provide an example of the difference between the holdup mechanism and the strategic

interactions described in this paper, it is useful to apply the Kalai (1977) proportional solution to our problem. Under Kalai bargaining, trade surpluses are monotonically increasing in capacities, z and w, and maximized for $z \in [k[\min(w, y^*)], \infty]$ and $w \in [k^{-1}(z), \Omega]$ (see Appendix A.1). Players maximize their surpluses by carrying enough capacity to trade as much as possible subject to the other player's capacity and the payment function $k(.)$. Any extra capacity is left unused and has no impact on the price and quantity traded. Assuming $i = \psi = 0$, the best-responses are $z(w) \geq k[\min(w, y^*)]$ and $k^{-1}(z) \leq w(z) \leq \Omega$, and any $y^* \in [0, \min(y^*, \Omega)]$ can be traded in equilibrium. As mentioned in Section 1.2.3, there exists no strategic interactions between investors' capacity choices, and the trade breakdown in Proposition 1.1 does not occur. However, a trade breakdown purely due to a severe two-sided holdup problem could still occur conditional on high enough ex-ante costs associated with the players' capacities choices. For example, for $\psi > 0$ and i high enough, player h's best-response function, $z(w)$, is strictly less than $k(\min(w, y^*))$, while player ℓ's best-response function, $w(z)$, is equal to or less than $k^{-1}(z)$. The only intersection is $(w, z) = (0, 0)$, so that no trade occurs, as a result of a pure holdup problem.

1.3 Application to an OTC asset market with credit frictions

We now follow the interpretation proposed in Section 1.2.1 and embed the bargaining game with two-sided capacity choices into a general equilibrium model of OTC asset trade. The model can be seen as bridging Lagos and Wright (2005) and Duffie et al. (2005), and thus closely resembles Lagos and Zhang (2019a), with the addition of credit frictions and participation decisions at the intensive margin. After presenting the environment, we solve for the general equilibrium and propose an algorithm to construct asymmetric monetary equilibria when no symmetric monetary equilibrium exists.

1.3.1 Environment

Time is discrete and continues forever. Each time period is divided into two stages where different activities take place. The economy is populated by a continuum of infinitely-lived investors of measure two. They are evenly split into two types, low valuation (ℓ) and high valuation (h), which determines both their endowment and the utility they derive from holding a one-period-lived and perfectly-divisible asset in the first stage. The asset comes in fixed supply, Ω, and is endowed to ℓ-investors at the beginning of each period.[13] In the second stage of each period, all investors have access to a production technology that transforms labor into a perishable consumption good at unit cost. This good is used as the numéraire and provides unit utility to all investors.

The expected lifetime utility of a χ-investor, where $\chi \in \{h, \ell\}$, can be written as

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t (\varepsilon_\chi y_t + c_t), \tag{1.3}$$

where y_t corresponds to the investor's holdings of the asset at the end of the first stage, $c_t > 0$ ($c_t < 0$) is his consumption (production) of the numéraire good in the second stage and $\beta \in (0, 1)$ is a discount factor. We normalize $\varepsilon_\ell = 1$ and let $\varepsilon \equiv \varepsilon_h > 1$.[14,15]

In the first stage, there exists an OTC market for investors to reallocate the asset among themselves. Due to search frictions, an investor gets to access this market with probability γ, in which case she is randomly matched with an investor of the opposite type. With probability $(1-\gamma)$, the investor cannot trade in the first stage. Payments in the OTC market

[13] In Appendix A.2, investors' types are randomly drawn each period, the asset becomes long lived and no longer endowed.

[14] The linearity in the preferences of investors over asset holdings is not necessary to obtain the main results presented in the paper, which follow from Proposition 1.1. It is preferred, however, for simplicity of exposition and for ease of comparison with closely-related papers who use similar specifications—in particular, Lagos and Zhang (2019a,b, 2020)

[15] Describing $\varepsilon_\chi y_t$ as an investor's utility from holding y_t units of asset is not to be interpreted literally, but as a reduced-form formalization of heterogeneity in the liquidity or hedging needs experienced by investors during the first stage, generating heterogeneity in how much they value the asset.

are subject to frictions. A pair of matched investors have access to credit with probability α. In that case, the asset buyer can issue an IOU to the seller, and full repayment can be enforced in the second stage. With probability $(1-\alpha)$, monitoring, record-keeping, and commitment issues prevent the use of IOUs. This generates the need for another asset, money, to be used as means of payment. Money is intrinsically useless. Its supply is exogenous and grows at a net rate π through lump-sum transfers to all investors at the beginning of the second stage.

Before they enter the first stage, all investors must make participation (or capacity) decisions at the intensive margin. They choose how much of their asset and money holdings to carry into the OTC market.[16] In case of a match, investors bargain over the terms of trade, subject to constraints due to their participation decisions—ℓ-investors can only sell up to the size of their inventory (no short-selling), and h-investors can only use the real balances they carry unless they have access to credit.

In the second stage, all investors can trade the numéraire good, produced on the spot, as well as money, in a Walrasian market. The price of money in terms of the numéraire good is denoted ϕ_t^m.

1.3.2 Equilibrium

I focus on stationary equilibria with constant aggregate real balances, such that the gross rate of return on money, ϕ_{t+1}^m/ϕ_t^m, is constant and equal to $1/(1+\pi)$.

Denote z_t an investor's real balances. The maximum discounted utility of a χ-investor who

[16]The amount of assets and real balances that they choose not to carry can be stored until the second stage at no cost. One can interpret this as investors being able to commit to a maximum trade capacity, potentially smaller than their actual inventory. Not allowing them to commit to such trade capacity would give investors an incentive to lie about their inventory size, or "hide" part of their assets in equilibrium.

enters the second stage is

$$W_t^\chi(z_t) = \max_{c_t, z_{t+1}} \left\{ c_t + \beta V_{t+1}^\chi(z_{t+1}) \right\} \text{ s.t. } 0 \leq z_{t+1} = \frac{z_t + T_t - c_t}{1+\pi}, \tag{1.4}$$

where V_{t+1}^χ is her maximum expected discounted utility in the upcoming first stage, T_t the lump-sum transfer, and z_{t+1} her choice of real balances for the next period. Plugging in for the budget constraint into the objective function, we can rewrite

$$W_t^\chi(z_t) = z_t + T_t + \max_{z_{t+1} \geq 0} \left\{ -(1+\pi) z_{t+1} + \beta V_{t+1}^\chi(z_{t+1}) \right\}. \tag{1.5}$$

As is standard in this class of models, W^χ is linear in wealth.

Denote S_i^χ the surplus from trade of a χ-investor in the first stage, where $i \in \{m, c\}$ respectively refer to money-only and credit trades. Let y_i denote the amount of asset traded against a payment p_i. Making use of the linearity of W, we have

$$\begin{aligned} S_i^h &= \varepsilon y_i + W^h(z - p_i) - W^h(z) = \varepsilon y_i - p_i, \\ S_i^\ell &= -y_i + W^\ell(z + p_i) - W^\ell(z) = p_i - y_i. \end{aligned} \tag{1.6}$$

Investors' surpluses from trading in the OTC market are equivalent to the surpluses described in Section 1.2.1, with $u(y) \equiv \varepsilon y$ and $c(y) = y$.

Consider the negotiation between an h-investor with a payment capacity of z and an ℓ-investor with a trade capacity of w.[17] The following lemma characterizes the allocations and surpluses in both money-only and credit meetings, as determined by the generalized Nash solution, where $\theta \in (0,1)$ corresponds to the h-investor's bargaining power.

Lemma 1.2 (Terms of trade and surpluses in the OTC market)**.** *Let* $\delta_1 \equiv \varepsilon / [\theta\varepsilon + 1 - \theta]$

[17]The amount of real balances carried by an ℓ-investor bargaining with a h-investor does not impact the negotiation, so that we do not need to keep track of it.

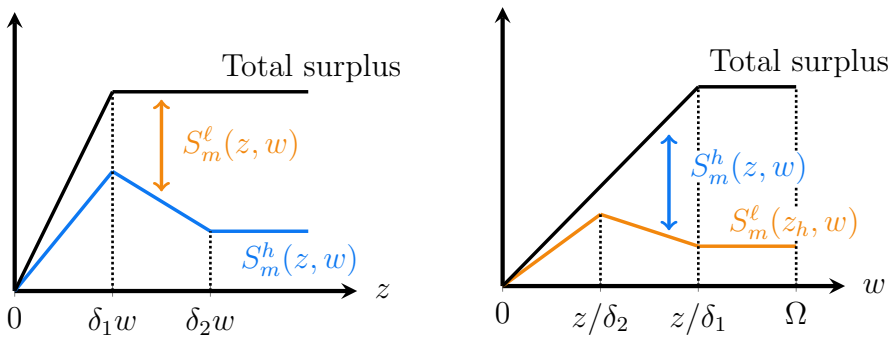

Figure 1.3: Left: Bargaining surpluses as a function of the h-investor's payment capacity (real balances). Right: Bargaining surpluses as a function of the ℓ-investor's trade capacity (asset holdings), where $\Omega \geq z/\delta_1$.

and $\delta_2 \equiv (1-\theta)\varepsilon + \theta$. *(i) In a match where credit is available, the terms of trade are given by* $(y_c, p_c) = (w, \delta_2 w)$. *The h-investor's surplus is* $S_c^h(w) = \theta(\varepsilon - 1)w$, *and the* ℓ*-investor's surplus is* $S_c^\ell(w) = (1-\theta)(\varepsilon - 1)w$. *(ii) In a match where only money can be used, the terms of trade,* (y_m, p_m), *and the investors' trade surpluses,* $S_m^h(z,w)$ *and* $S_m^\ell(z,w)$, *are given by Lemma 1.1 with* $k(y) = \delta_2 y$ *and* $p(y) = \delta_1 y$.

In matches where only money can be used for payments, the investors' surpluses are similar to those described in Section 1.2.2. In particular, S_m^h and S_m^ℓ are non-monotone in the investors' capacity choices. As can be seen in Figure 1.3, they are first increasing then decreasing in z and w respectively, and are maximized at $z = \delta_1 w$ and $w = z/\delta_2$. As a result, the strategic incentives to undercut a trade partner by reducing one's capacity, described in Section 1.2, will be present in those matches. This is not the case in matches where credit is available, where both agents' surpluses are monotonically increasing in the trade capacity of the ℓ-investor, and independent of the payment capacity of the h-investor.

The maximum expected discounted utility of an h-investor who entered the first stage with

z real balances is

$$V^h(z) = \max_{z^p} \gamma \left[\alpha \mathbb{E} S_c^h(\tilde{w}) + (1-\alpha)\mathbb{E} S_m^h(z, \tilde{w})\right] + W^h(z) \text{ s.t } 0 \leq z^p \leq z, \tag{1.7}$$

where z^p corresponds to her payment capacity choice. The investor faces uncertainty with respect to the possibility of payment by credit and with respect to the trade capacity of a potential trading partner, $\tilde{w}$. Similarly, the maximum expected discounted utility of an ℓ-investor who enters the first stage with z real balances is

$$\begin{aligned} V^\ell(z) = \max_{w, z^p} & \gamma \left[\alpha S_c^\ell(w) + (1-\alpha)\mathbb{E} S_m^h(\tilde{z}, w)\right] + \Omega + W^\ell(z) \\ & \text{s.t } 0 \leq z^p \leq z \text{ and } 0 \leq w \leq \Omega, \end{aligned} \tag{1.8}$$

where $z^p \leq z$ is his payment capacity decision, and $w \leq \Omega$ is his trade capacity decision.

Denote $F^z(z)$ the equilibrium distribution of real balances held by h-investors from a period to another, and $F^w(w)$ the equilibrium distribution of the trade capacities of ℓ-investors in the OTC market. They are respectively supported by $\mathbb{Z}$ and $\mathbb{W}$, with $\underline{z} \equiv \min(\mathbb{Z})$, $\bar{z} \equiv \max(\mathbb{Z})$, $\underline{w} \equiv \min(\mathbb{W})$, and $\bar{w} \equiv \max(\mathbb{W})$.

Lemma 1.3 (Distributions of holdings and participation)**.** *Let*

$$v^h(z) \equiv -iz + \gamma(1-\alpha) \int S_m^h(z, w) dF^w(w), \text{ where } i \equiv \frac{1+\pi}{\beta}, \tag{1.9}$$

and

$$v^\ell(w) \equiv \gamma \left[\alpha S_c^\ell(w) + (1-\alpha) \int S_m^\ell(z, w) dF^z(z)\right]. \tag{1.10}$$

(i) For i¿0, $z = z^p = 0$ for all ℓ-investors and $z = z^p$ for all h-investors.

(ii) For all $z \in \mathbb{Z}$,

$$z \in \textit{argmax } v^h(z) \textit{ s.t. } z \geq 0. \tag{1.11}$$

For all $w \in \mathbb{W}$,

$$w \in \textit{argmax } v^\ell(w) \textit{ s.t. } w \in [0, \Omega]. \tag{1.12}$$

Carrying real balances from a period to another exclusively benefits the h-investor in the first stage. Because doing so is costly as long as $i > 0$, in equilibrium, the ℓ-investor does not carry real balances from a period to another, and the h-investor only accumulates the amount corresponding to the payment capacity she would like to have in the OTC market. As a result, $F^z(z)$ corresponds to the equilibrium distribution of payment capacities chosen by the h-investor, and from now on we refer to z as the h-investor's choice of payment capacity.

We are now ready to define our notion of equilibrium.

Definition 1.1 (Equilibrium)**.** *An equilibrium consists in a list*

$$\{F^z(z), F^w(w), p_j(z,w), y_j(z,w)\}_{j\in\{m,c\}} \textit{ for all } w \in [0,\Omega] \textit{ and } z \in \mathbb{R}^+,$$

such that (i) $\{p_j(z,w), y_j(z,w)\}_{j\in\{m,c\}}$ *satisfy Lemma 1.2, (ii) all* $z \in \mathbb{Z}$ *satisfy* (1.11) *taking* $F^w(w)$ *as given and all* $w \in \mathbb{W}$ *satisfy* (1.12) *taking* $F^z(z)$ *as given, (iii)* $F^z(z) = F^w(w) = 0$ *for all* $z < \underline{z}$, $w < \underline{w}$, *and* $F^{z\prime}(z) = F^{w\prime}(w) = 0$ *for all* $z \notin \mathbb{Z}$, $w \notin \mathbb{W}$, *(iv) the competitive market open in the second stage of each period clears.*

When they exist, I will focus on symmetric equilibria, where all ℓ-investors make the same participation decision, and all h-investors carry the same amount of real balances. We denote

$\mathcal{W}$ the welfare in the first stage of each period, computed as the expected sum of all investors' utility in that stage.

Proposition 1.2 (Equilibrium regimes)**.** *(i) High access to credit,* $\alpha > \min\left(1/\delta_2, 1 - i/[\gamma\theta(\varepsilon - 1)]\right)$.
(i.i) When $\alpha > 1 - i/[\gamma\theta(\varepsilon - 1)]$*, there exists a unique symmetric equilibrium. It is non-monetary, and OTC trade only occurs in credit meetings, with* $z = p_m = y_m = 0$*,* $w = y_c = \Omega$*,* $p_c = \delta_2\Omega$ *and* $\mathcal{W} = \Omega + \gamma\alpha(\varepsilon - 1)\Omega$.
(i.ii) When $1/\delta_2 < \alpha \leq 1 - i/[\gamma\theta(\varepsilon - 1)]$*, there exists a symmetric monetary equilibrium. Trade occurs in the OTC market both in credit and money meetings. When the inequality is strict, the monetary equilibrium is unique and such that* $z = p_m = p_c = \delta_1\Omega$*,* $w = y_m = y_c = \Omega$*, and* $\mathcal{W} = \Omega + \gamma(\varepsilon - 1)\Omega$.
(ii) Low access to credit, $\alpha \leq \min\left(1/\delta_2, 1 - i/[\gamma\theta(\varepsilon - 1)]\right)$
(ii.i) When $0 < \alpha \leq \min\left(1/\delta_2, 1 - i/[\gamma\theta(\varepsilon - 1)]\right)$*, there exists no symmetric monetary equilibrium, but there exist multiple monetary equilibria. OTC trade occurs in both credit and money meetings, with* $\Omega < \mathcal{W} < \Omega + \gamma(\varepsilon - 1)\Omega$.
(ii.ii) When $\alpha = 0$*, there exists a unique equilibrium. It is non-monetary and the OTC market shuts down, with* $z = w = p_m = y_m = 0$*, and* $\mathcal{W} = \Omega$.

This proposition describes equilibrium regimes as a function of the prevalence of credit in the OTC market. There are two main regions, each divided in two sub-regions. In region (i), credit is abundant enough to negate the strategic interactions described in Section 1.2, and there exists a symmetric equilibrium with an operative OTC market. In region (ii), credit is not abundant enough for this to happen, and capacity constraints remain relevant. In this region, as expected following Proposition 1.1, the OTC market shuts down when credit is completely absent. However, it remains operative as long as credit is available in a strictly positive measure of matches. Then, monetary equilibria exist, but are asymmetric.

To understand these results, consider first the subset of symmetric equilibria, solving for (1.11) and (1.12) assuming $F^w(w)$ and $F^z(z)$ are degenerate. The h-investor's optimal

payment capacity is given by

$$z = \begin{cases} \delta_1 w & \text{if } i < (1-\alpha)\gamma\theta(\varepsilon-1) \\ z \in [0, \delta_1 w] & \text{if } i = (1-\alpha)\gamma\theta(\varepsilon-1) \\ 0 & \text{otherwise.} \end{cases} \tag{1.13}$$

If access to credit is too high relative to the cost of money, $\alpha > 1 - i/[\gamma\theta(\varepsilon-1)]$, the h-investor prefers not to carry any money. If access to credit is low enough relative to the cost of holding money, she accumulates exactly enough to buy all of the ℓ-investor's assets at the lowest price, $z = \delta_1 w$. The ℓ-investor's participation choice amounts to maximizing his expected bargaining surplus,

$$\mathbb{E}S^{\ell} = \begin{cases} (1-\theta)(\varepsilon-1)w & \text{if } w \leq z/\delta_2 \\ \alpha(1-\theta)(\varepsilon-1)w + (1-\alpha)(z-w) & \text{if } z/\delta_2 < w \leq z/\delta_1 \\ \alpha(1-\theta)(\varepsilon-1)w & \text{otherwise.} \end{cases} \tag{1.14}$$

Note that $\mathbb{E}S^{\ell}$ is piecewise linear. The first and third pieces are increasing in w. The middle piece is increasing if $\alpha > 1/\delta_2$, constant if $\alpha = 1/\delta_2$ and decreasing otherwise. When picking his trade capacity, the ℓ-investor faces the following trade-off: in credit meetings, his bargaining surplus is strictly monotone in his trade capacity—encouraging him to bring all of his endowment—while in money meetings, it is maximized at $w = z/\delta_2$—encouraging him to restrict his trade capacity. When the probability of a credit meeting is high, the monotone part of the weighted average dominates and the expected surplus is monotone as well. When the probability of a credit meeting is low, the non-monotone part dominates, and the ℓ-investor's expected surplus is non-monotone. As a result, the ℓ-investor's optimal trade capacity is $w = \Omega$ if $\Omega \leq z/\delta_2$. If $\Omega \in (z/\delta_2, z/\delta_1)$ then $w = \Omega$ if $\alpha \geq 1/\delta_2$ and $w = z/\delta_2$ otherwise. Finally, if $\Omega \geq z/\delta_1$, $w = \Omega$ if $z \leq \alpha\delta_2\Omega$ and $w = z/\delta_2$ otherwise.

In the region considered in (i.i), $\alpha > 1 - i/[\gamma\theta(\varepsilon - 1)]$, so that real balances are too costly for the h-investor relative to access to credit, and she prefers to not carry any. Then, the ℓ-investor does not face the trade-off between maximizing surplus in money meetings versus maximizing surplus in credit meetings—there is not surplus to be had in money meetings. As a result, he simply picks his trade capacity w to maximize his surplus in credit meetings, that is, $w = \Omega$. This result highlights the typical substitutability between money and credit: high access to credit renders money useless, so that the latter cannot be positively valued in equilibrium. The OTC market is open but only functions through credit trades, so that welfare cannot reach first best unless $\alpha = 1$.

When $\alpha < 1 - i/[\gamma\theta(\varepsilon - 1)]$, in any symmetric equilibrium, the optimal payment capacity for the h-investor is to carry exactly $z = \delta_1 w$. This revives the trade-off faced by the h-investor's, whose capacity choice now depends on the prevalence of credit. If he expects credit to be available often, $1/\delta_2 < \alpha < 1 - i/[\gamma\theta(\varepsilon - 1)]$, the losses incurred by obtaining a relatively lower price in money meetings when carrying $w = \Omega$ are offset by the gains enjoyed by trading a relatively higher volume in credit meetings. In this case, there exists a unique symmetric monetary equilibrium where the ℓ-investor picks a trade capacity of $w = \Omega$, the h-investor a payment capacity of $z = \delta_1\Omega$, and welfare achieves first best conditional to the search frictions, $\mathcal{W} = \Omega + \gamma(\varepsilon - 1)\Omega$.[18] This corresponds to region (i.ii). In short, in both regions (i.i) and (i.ii), credit negates the strategic complementarity between capacity choices, and it does so by partially decoupling the investors' capacity choices from their trade partner's capacity choice. In region (i.i) the prevalence of credit matches encourages the h-investor to carry no money holdings regardless of the ℓ-investor's capacity, while in region (i.ii), it provides incentives for the ℓ-investor to carry all of his endowment regardless of the h-investor's money holdings.

[18] In the knife-edge case where $\alpha = 1 - i/[\gamma\theta(\varepsilon - 1)]$, the equilibrium just described still exists but it is not the only monetary equilibrium. There exists a continuum of equilibria with $w = \Omega$, $z \in [0, \delta_1 w]$, and $\mathcal{W} = \Omega + \gamma\left[\alpha(\varepsilon - 1)\Omega + (1 - \alpha)(\varepsilon - 1)z/\delta_1\right]$.

Now, when $\alpha \leq \min\left(1/\delta_2, 1 - i/[\gamma\theta(\varepsilon - 1)]\right)$, the optimal payment capacity for the h-investor is still to carry exactly $z = \delta_1 w$, but money meetings are too prevalent for the ℓ-investor to accept the losses incurred in those meetings by carrying $w = \Omega$. Instead, he would rather carry $w = z/\delta_2$. The best responses are then identical to the best responses described in Section 1.2, where the two investors always want to undercut each other by restricting their respective capacities. The only symmetric equilibrium is $z = w = 0$, so that no symmetric monetary equilibrium exists. When $\alpha = 0$, corresponding to region (ii.ii), this is the unique equilibrium. The OTC market breaks down and there is not trade in equilibrium, leading to the lowest feasible total welfare.[19] Money is not valued, and ℓ-investors do not participate to the OTC market (i.e., they choose a trade capacity of 0). This result follows directly from Proposition 1.1. As soon as credit is introduced in a strictly positive measure of meetings, however, money can again be valued and OTC restored, leading to welfare gains relative to the pure-currency equilibrium. This corresponds to regime (ii.i). The existence of monetary equilibria in this regime is proven in Proposition 1.3, where we propose an example of such an equilibrium.

The breakdown of the OTC market in region (ii.ii) highlights the importance of both the bargaining solution and the assumptions made regarding capacity and payment constraints in models of OTC markets such as Lagos and Zhang (2019a,b, 2020) and Duffie et al. (2005). In Lagos and Zhang (2019a), even though there is not credit, the strategic interactions described here are circumvented by the use of a take-it-or-leave-it bargaining mechanism. In Lagos and Zhang (2019b, 2020), where perfect credit is absent as well, trade in the OTC market can be facilitated by dealers, who have access a competitive interdealer market. By giving them the opportunity to locate assets ex-post, upon demand from the customer, the interdealer markets takes away the inventory capacity constraint.[20] In Duffie et al. (2005)

[19]The same result, obtained in an environment with infinitely-lived assets and stochastic preference shocks (rather than fixed types), is derived in Appendix A.2.

[20]By extension, providing investors access to a competitive interdealer market (through dealers) in the present model would play a role very similar to that played by access to credit, described in details in the remaining of the paper. While the interdealer market would relax the seller-side capacity constraint

and the literature that follows, it is the assumption of investors having deep pockets, freeing them from payment constraints, which makes those strategic choices disappear.

We now turn to the construction of monetary equilibria in region (ii.i).

Asymmetric equilibrium construction When $\alpha \leq \min(1/\delta_2, 1 - i/[\gamma\theta(\varepsilon - 1)])$, the main hurdle in the way of monetary equilibria is the incentive faced by the ℓ-investor to restrict his trade capacity and enter the undercutting spiral described previously, instead of carrying all of his endowment regardless of the h-investor's payment capacity. The following general results, proven in Appendix A.1, will be helpful in constructing equilibria that circumvent this problem.

Lemma 1.4 (Minimum and maximum capacities supported in equilibrium)**.** *In any monetary equilibrium,* $\bar{w} = \Omega$, $\bar{z} \leq \delta_1 \bar{w}$, $\underline{w} \geq \min(\Omega, \underline{z}/\delta_2)$ *and* $\underline{z} \geq \delta_1 \underline{w}$.

Let $\alpha \leq \min(1/\delta_2, 1 - i/[\gamma\theta(\varepsilon - 1)])$, $w = \Omega$ and $z = \delta_1 w = \delta_1 \Omega$. We know that no such symmetric equilibrium exists since the ℓ-investor would deviate to $w = z/\delta_2 = \delta_1\Omega/\delta_2$. To sustain a monetary equilibrium, which requires $\bar{w} = \Omega$ according to Lemma 1.4, we then have two options: (1) increase the ℓ-investor's payoff from carrying $w = \Omega$, (2) decrease the ℓ-investor's payoff from carrying $w = \delta_1\Omega/\delta_2$, so that it is equal to the payoff from carrying $w = \Omega$. As shown by the dashed line on the left panel of Figure 1.4, option (1) is not feasible, as it would require $z > \delta_1\Omega$, which is ruled out by Lemma 1.4. Thus, we must follow the second option. As displayed on the right panel of Figure 1.4, decreasing the payoff from a deviation to $w = \delta_1\Omega/\delta_2$ can be made possible by the addition of a second mass point in $\mathbb{Z}$, z_2, represented by the dashed line. The lower z_2, the lower the benefit of carrying $w = \delta_1\Omega/\delta_2$. To maximize the range of α where the equilibrium we construct exists, we pick the lowest feasible z_2 according to Lemma 1.4, $z_2 = \delta_1^2\Omega/\delta_2$. The last step consists in solving

(inventory) by allowing assets to be located ex-post, access to credit relaxes the buyer-side capacity constraint (real balances) by allowing them to locate funds ex-post.

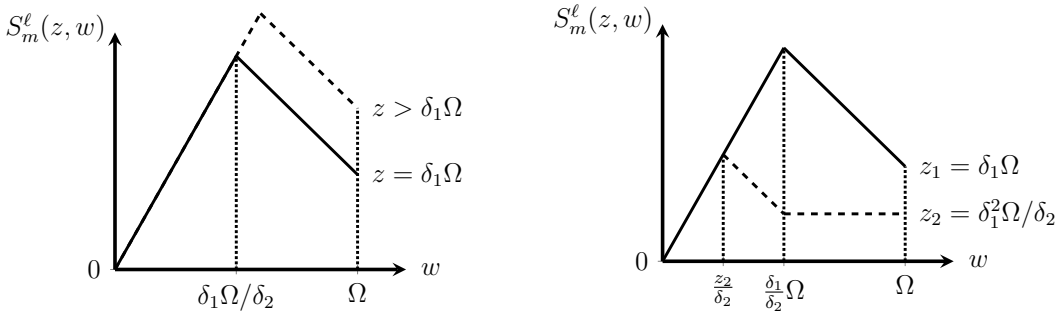

Figure 1.4: Construction of a two-mass-point equilibrium. The ℓ-investor's strategies are on the x-axis, while the y-axis represents his surplus in money-only meetings, depending on the payment capacity of his trade partner, z.

for the distributions $F^w(w)$ and $F^z(z)$ that make each type of investor indifferent between their two equilibrium strategies.

As α decreases and money meetings become more and more prevalent, preventing the ℓ-investor from deviating to $w = z_2/\delta_2$ requires a higher and higher probability of meetings with h-investors who carry z_1 rather than z_2. When the required probability exceeds 1, the two-mass-point equilibrium cannot be sustained anymore. We then repeat the same process, and make the ℓ-investor indifferent between $w = \Omega$, $w = \delta_1\Omega/\delta_2$, and $w = \delta_1^2\Omega/\delta_2^2$ by adding a third z, $z_3 = \delta_1^3\Omega/\delta_2^2$. The equilibrium obtained by following this algorithm is formalized in Proposition 1.3. It features the coexistence of money and credit for any $0 < \alpha \leq 1 - i/[\gamma\theta(\varepsilon - 1)]$.[21]

Proposition 1.3 (Coexistence of money and credit)**.** *Consider the following equilibrium supports for $F^w(w)$ and $F^z(z)$,*

$$\tilde{\mathbb{W}} = \{w_n\}_{n=1}^N = \left\{\left(\frac{\delta_1}{\delta_2}\right)^{n-1}\Omega\right\}_{n=1}^N \text{ and } \tilde{\mathbb{Z}} = \{z_n\}_{n=1}^N = \{\delta_1 w_n\}_{n=1}^N. \tag{1.15}$$

For any $\alpha \in (0, 1 - i/[\gamma\theta(\varepsilon - 1)]]$, there exists a unique monetary equilibrium with $\mathbb{W} = \tilde{\mathbb{W}}$

[21] Another proof of the existence of a monetary equilibrium for any α in this parameter region, which makes use of Glicksberg (1952) fixed-point theorem, is provided in Appendix A.1.

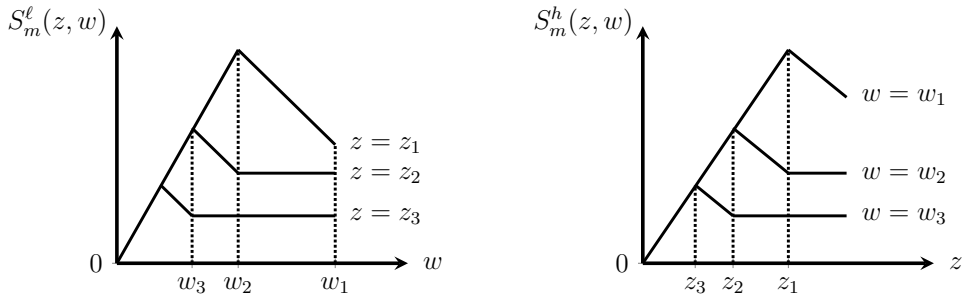

Figure 1.5: Structure of equilibrium trade and payment capacities when $N = 3$.

and $\mathbb{Z} = \tilde{\mathbb{Z}}$. *It is such that*

$$N = \left\{ \lceil x \rceil : \frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\frac{\delta_2^x - 1}{\delta_2 - 1} = 1 \right\}, \tag{1.16}$$

$$Pr(z = z_n) = \frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\delta_2^{n-1} \text{ for } n = 1, 2, ..., N-1, \tag{1.17}$$

$$Pr(w = w_n) = \frac{(1-\alpha)\theta(\varepsilon-1) - i}{(1-\alpha)[\theta(\varepsilon-1)+1]} \left(\frac{\varepsilon}{\delta_1}\right)^{n-N} \text{ for } n = 2, 3, ..., N, \tag{1.18}$$

with $Pr(z = z_N) = 1 - Pr(z > z_N)$, *and* $Pr(w = w_1) = 1 - Pr(w < w_1)$. *When* $N = 1$, $Pr(w = \Omega) = Pr(z = \delta_1 \Omega) = 1$.

Figure 1.5 provides an example of the equilibrium structure given by (1.15) when $N = 3$. The equilibrium is constructed so that the unit prices in the OTC market always belong to $\{\delta_1, \delta_2\}$. In money meetings, investors trade at the low price, δ_1, when the payment constraint of the h-investor is exactly satisfied or binds, and at the high price, δ_2, otherwise. To see this, note that for any (z_n, w_m) on the support, we have $z_n/w_m = \delta_1 (\delta_1/\delta_2)^{n-m}$. When $n = m$, z_n is the best response to w_m. The h-investor obtains all of the trade capacity, w_m, at the

low price, δ_1. When $n > m$, the h-investor's payment constraint is binding. She purchases as much as she can at the low price, so that the quantity traded is z_n/δ_1. When $n < m$, the h-investor's payment capacity does not bind, but the ℓ-investor's trade capacity does. He sells all of his inventory at the high price. In credit meetings, the trade price is always the high price, δ_2. Investors are then made indifferent between the different capacities in the equilibrium supports by balancing the differences in expected prices received in bilateral meetings by differences in expected quantities traded.

In order to prove Proposition 1.3, first note that due to the linearity of payoffs, the h-investor's expected payoff is linear between two consecutive mass points in $\mathbb{Z}$, as is the ℓ-investor's payoff between two consecutive mass points in $\mathbb{W}$. Because investors must be indifferent between all of the equilibrium strategies, it must be that in equilibrium, the h-investor's expected payoff is constant between z_N and z_1, while the ℓ-investor's expected payoff is constant between w_N and w_1.

We can then solve for the distributions $F^w(w)$ and $F^z(z)$ that make the ℓ- and h- indifferent by recursion. For $w_1 = \Omega$ to be an equilibrium mass point, the slope of the ℓ-investor's expected surplus when w tends to $w_1^- = \Omega^-$ must be 0, or

$$\alpha \frac{\partial S_c^\ell(w)}{\partial w}_{|w=\Omega} + (1-\alpha)\Pr(z = z_1)\frac{\partial S_m^\ell(z_1, w)}{\partial w^-}_{|w=\Omega} = 0,$$

from which we obtain $\Pr(z = z_1) = \alpha(1-\theta)(\varepsilon-1)/(1-\alpha)$. We can then solve for $\Pr(z = z_2)$, which again must make the slope of the ℓ-investor's expected surplus equal to 0 as w tends to w_2^-,

$$\alpha \frac{\partial S_c^\ell(w)}{\partial w}_{|w=w_2} + (1-\alpha)\left[\Pr(z_1)\frac{\partial S_m^\ell(z_1, w)}{\partial w^-}_{|w=w_2} + \Pr(z_2)\frac{\partial S_m^\ell(z_2, w)}{\partial w^-}_{|w=w_2}\right] = 0.$$

We get $\Pr(z = z_2) = \alpha(1-\theta)(\varepsilon-1)\delta_2/(1-\alpha)$. Iterating this process up to z_{N-1} gives (1.17), and $\Pr(z = z_N) = 1 - \sum_{j=1}^{N-1} Pr(z = z_j)$. A similar method can be used to obtain (1.18),

however starting from the indifference condition for z_N, iterating up towards the indifference condition of z_2, and with $\Pr(w = w_1) = 1 - \sum_{j=2}^{N} \Pr(w = w_j)$. Note that the probabilities given by (1.17) and (1.18) are always positive in the parameter region we are considering. Also note that the h-investor's expected surplus is increasing up to z_N and decreasing past z_1, so that she would not want to deviate either way.

Three conditions remain to be checked to ensure that the equilibrium exists. First, we must make sure that $\sum_{j=2}^{N} Pr(w = w_j) < 1$. Second, we must also have $\sum_{j=1}^{N-1} \Pr(z = z_j) < 1$. Third, we must check that the slope of the ℓ-investor's surplus when w tends to w_N^- is positive, so that he does not have an interest to deviate to $w = z_N/\delta_2$. The first condition is satisfied as long as $i < (1-\alpha)(1-\theta)(\varepsilon - 1)$, which holds. The second condition requires $N \leq \tilde{N}$, where $\tilde{N}$ satisfies (1.16). This also holds since equilibrium requires $N = \tilde{N}$. The third condition requires that $\alpha \leq 1/\delta_2$ when $N > 1$ and $\alpha \geq 1/\delta_2$ when $N = 1$. We show in Appendix A.1 that (1.16) directly implies $\alpha \geq 1/\delta_2$ when $N = 1$ and $1/\delta_2^N \leq \alpha \leq 1/\delta_2^{N-1}$ for $N > 1$, so that the third condition is indeed satisfied. This proves that the monetary equilibrium proposed in Proposition 1.3 exists and is unique.[22]

1.4 Implications

We now derive the implications of the model for the impact of access to credit on investors' participation, trade volume, and prices. To highlight the role of market structure, we contrast the model's predictions to those we would obtain if the asset market was a perfectly competitive Walrasian market instead of an OTC market subject to bargaining frictions. When making those comparisons, the search and credit frictions are kept identical—investors access the asset market with probability γ, and can use credit with probability α. In the competi-

[22]It is not claimed that the equilibrium described in Proposition 1.3 is the unique monetary equilibrium. It is unique conditional on $\mathbb{W}$ and $\mathbb{Z}$ given by (1.15). I focus on this kind of equilibrium because their recursive structure is intuitive and they can be solved analytically.

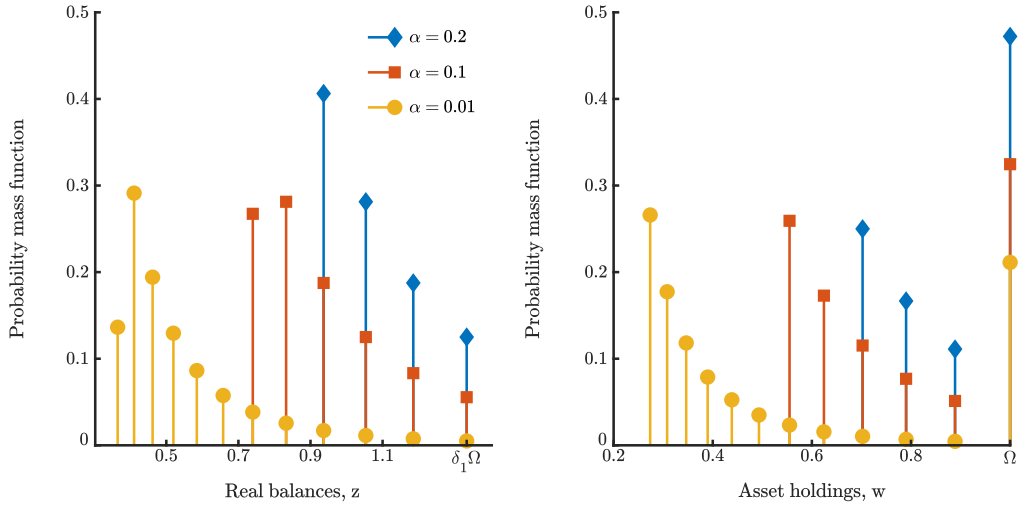

Figure 1.6: Left panel: Distributions of payment capacities (i.e., real balances) across h-investors for different α. Right panel: Distributions of asset inventories (i.e., participation) across ℓ-investors for different α.

tive market, the trade surplus of an h-investor who purchases y^d units of assets is $(\varepsilon - q)y^d$, while the trade surplus of an ℓ-investors who sells y^s units of assets is $(q-1)y^s$, where q is the unit price of the asset. It is taken as given by the investors and determined by market clearing.

We first derive implications related to the heterogeneity in participation, trade volume, and prices in the OTC market. From Proposition 1.3, when access to credit is low (but not inexistent) in the OTC market, all monetary equilibria are asymmetric. The strategic interactions that occur when investors make capacity choices generate heterogeneity in inventories and real balances endogenously. Figure 1.6 provides examples of the distributions of z and w held by investors for different levels of α.[23]

Result 1.1 (Endogenous degree of heterogeneity)**.** *As α decreases, N increases.*

This follows directly from (1.16) and highlights that the degree of heterogeneity in the

[23]The parameters used for this numerical example and the following ones are $\varepsilon = 2$, $\theta = 0.5$, $\Omega = 1$, $\gamma = 1$ and $i = 0.1$. The latter was chosen small enough so that $1/\delta_2 < 1 - i/[\gamma\theta(\varepsilon - 1)]$, so as to ensure that there exists a symmetric monetary equilibrium for α high enough.

economy, as measured by N (the number of mass points in $F^w(w)$ and $F^z(z)$), increases as access to credit, α, diminishes. At the limit when α tends to 0, N tends to ∞. We thus expect a market with more stringent credit frictions to feature more heterogeneity in the positions of investors.

Heterogeneity in trade and payment capacities translates into heterogeneity in the terms of trade in the OTC market. For example, Figure 1.7 focuses on money-only meetings and shows the distributions of quantity traded, y_m, and asset price, p_m/y_m, for three levels of α.[24] The existence of a price discrepancy between two sales of the same asset, executed between pairs of investors with identical payoffs and identical bargaining powers, at the same time period, is a notable result. This heterogeneity is endogenous and driven by differences in the positions of investors. Asset sellers with a relatively lower position negotiate higher prices in meetings subject to payment constraints (or equivalently, asset buyers with a relatively lower payment capacity negotiate lower prices).

A competitive asset market, by shutting down the strategic interactions between the capacity choices of investors, would only feature symmetric equilibria. This paper therefore highlights a new theoretical channel through which heterogeneity in participation, portfolios, and terms of trade, largely documented empirically, can arise in OTC markets subject to bargaining frictions. Notably, this heterogeneity is not due to ex-ante heterogeneity among investors, nor is it due to multiplicity of equilibria.

Lemma 1.5. *Let $0 < \alpha' < \alpha < 1 - i/[\gamma\theta(\varepsilon - 1)]$. Then $F^w_\alpha(w)$ first-order stochastically dominates $F^w_{\alpha'}(w)$ and $F^z_\alpha(z)$ first-order stochastically dominates $F^z_{\alpha'}(z)$.*

The proof is provided in Appendix A.1. We make use of this lemma to derive the following

[24] The fact that the distribution of prices is supported by the same two mass points as α varies is due to the equilibrium structure described in Proposition 1.3. Indeed, it restricts the different combinations of p_m and y_m across matches to two possibilities: either the payment capacity is binding ($z < \delta_2 w$), in which case the price is "low", $p_m = \delta_1$, or the payment capacity is slack ($z \geq \delta_2 w$), and the price is "high", $p_m = \delta_2$, and equal to the price in credit meetings.

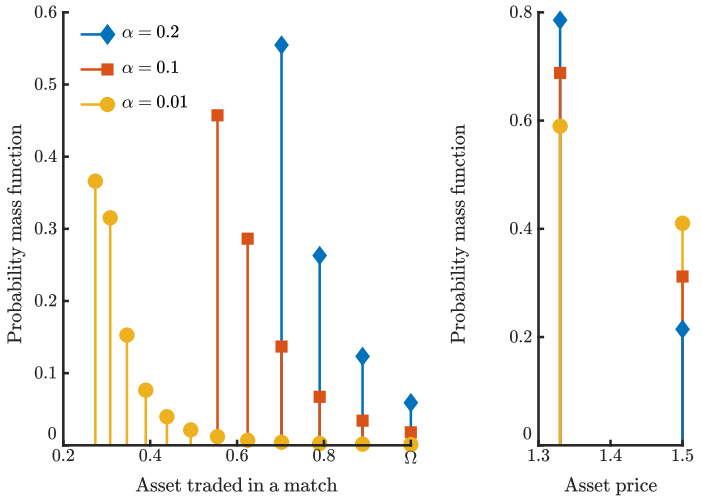

Figure 1.7: Distributions of quantities of asset traded and prices in money-only matches for different levels of α.

three results.

Result 1.2 (Participation, trade volume, and welfare)**.** *When the equilibrium is monetary, the average participation of ℓ-investors (trade capacity), the average trade volume in money meetings, the aggregate trade volume in the OTC market, and welfare decrease as α decreases.*

As credit is less accessible, the average participation of ℓ-investors decreases. Indeed, recall that ℓ-investors face the following trade-off: the larger their inventory, the larger their gains from trade in credit meetings, but the smaller the possibility of trading at a high price in money meetings. As credit becomes scarcer, incentives to reduce inventory increase, pushing average inventories down. As a result, aggregate trade volume in the OTC market decreases as well, which brings welfare down.[25] At the limit when α tends to 0, aggregate trade tends to 0 and welfare tends to Ω, coinciding with the no-credit equilibrium.[26] In the perfectly competitive case, on other hand, the equilibrium features full participation from ℓ-investors

[25]The positive relation between access to funding and inventories is documented, for example, by Macchiavelli and Zhou (2019) for the corporate bonds market. As for the positive relation between access to credit and trade volume, some evidence can be found in event studies that analyze the behavior of OTC-traded stocks before and after regulatory changes that make them marginable, surveyed in Fortune et al. (2001) and Pruitt and Tse (1996). Trade volume consistently increases after the addition of a stock to the list.

[26]The aggregate participation of ℓ-investors, however, does not tend to 0 as long as $i > 0$. Indeed, in order

($w = \Omega$) and first-best trade size ($y = \Omega$) for any $\alpha > 0$.[27] As a result, when the market is competitive, the participation decision of ℓ-investors is irrelevant. Not only can the asset market withstand the complete absence of credit, aggregate trade volume and welfare remain at first best.

Result 1.3 (Complementarity between money and credit)**.** *When the equilibrium is monetary, average real balances holdings (payment capacities) decrease as α decreases.*

We saw in Proposition 1.2 that money cannot be valued when $\alpha > 1 - i/[\gamma\theta(\varepsilon - 1)]$. In this region, money and credit act as substitutes, in that the high availability of credit renders money useless, and the two do not coexist. We also observe this outcome when the asset market is competitive—only when $\alpha = 0$ can the equilibrium be monetary, i.e., money and credit never coexist.

More generally, a milder form of substitutability between money and credit is usually a robust outcome of models in which money and credit coexist. For example, with the competitive asset market structure, one can easily make money and credit coexist by adding concavity to the payoffs.[28] Then, when the equilibrium is monetary, the amount of real balances carried by h-investors into the competitive asset market increases as α decreases. Gu et al. (2016) investigate the coexistence of money and credit in decentralized markets and find that "the economy does not need both: if credit is easy, money is irrelevant; if credit is tight, money is essential, but credit becomes irrelevant. Changes in credit conditions are neutral because real balances respond endogenously to keep total liquidity constant."

This is not the case here when $\alpha < 1 - i/[\gamma\theta(\varepsilon - 1)]$. As underlined by Result 1.3, in this region, as the probability of access to credit diminishes, the average amount of real

to sustain a monetary equilibrium even when α is very low, h-investors must expect a high enough payoff from money-only meetings.

[27] And also when $\alpha = 0$ as long as $i < \gamma(\varepsilon - 1)$. Otherwise, no trade occurs since credit is not available and the cost of money holdings is too high to depart from $z = 0$.

[28] E.g., assuming satiation point, such that the utility received by an h-investor who holds y units of assets at the end of the first stage is $\varepsilon \min(y, \bar{y})$, where $\bar{y} > \Omega$.

balances held by h-investors diminishes as well, making money and credit complements. An example of this complementarity is visible in the left panel of Figure 1.8. This result is, once again, due to the strategic interactions between investors' when they choose their trade and payment capacities. The fall in average trade capacity that ensues a decrease in credit in turn encourages h-investors to decrease their holdings of real balances in order to regain market power and favorable terms of trade.[29] It is easy to show that this complementarity would be overturned, making credit indeed irrelevant, were those strategic interactions shut down. For example, assume that ℓ-investors produce assets on the spot (e.g., derivatives), up to an exogenously determined capacity Ω, and that they cannot restrict this capacity. In this case, in any monetary equilibrium, h-investors would pick a payment capacity of $z = \delta_1 \Omega$. Trade size would be $y = \Omega$ both in credit and money meetings, making the type of payment used irrelevant. As a result, a drop in the availability of credit would have no impact on trade volume, prices, nor welfare as long as money is valued. The two would be perfect substitutes.

One implication of the complementarity between money and credit is that it generates a feedback loop between "funding liquidity" and "market liquidity" reminiscent of that described by Brunnermeier and Pedersen (2009). When credit gets tighter, aggregate trade volume decreases due to ℓ-investors reducing their inventories, independently of h-investors' payment capacities. The drop in inventories, however, leads h-investors to reduce their payment capacities, worsening the funding conditions in the market, and aggravating the drop

[29]Note that Gu et al. (2016) acknowledge that their results do not always hold, and that "it is important to know what kinds of assumptions may or may not make credit matter. If economists want to argue that credit conditions are important, they should be able to articulate how the assumptions in the models [Gu et al. (2016) study] are violated, and they might check what happens in the models they use once money is introduced." This paper provides one example of a natural environment where money and credit coexist and where credit does matter when both payment and trade capacity constraints are introduced. Lagos and Zhang (2019b) provide a different example of complementarity between money and credit. In their model, assets purchases can be financed via money and collateralized credit, where the asset traded can be used as collateral. The degree of credit financing in the economy, as measured by the loan-to-value ratio, is exogenous. The higher the loan-to-value ratio, the higher the value of money, as assets purchased with money can get more leverage and further increase the investor's position.

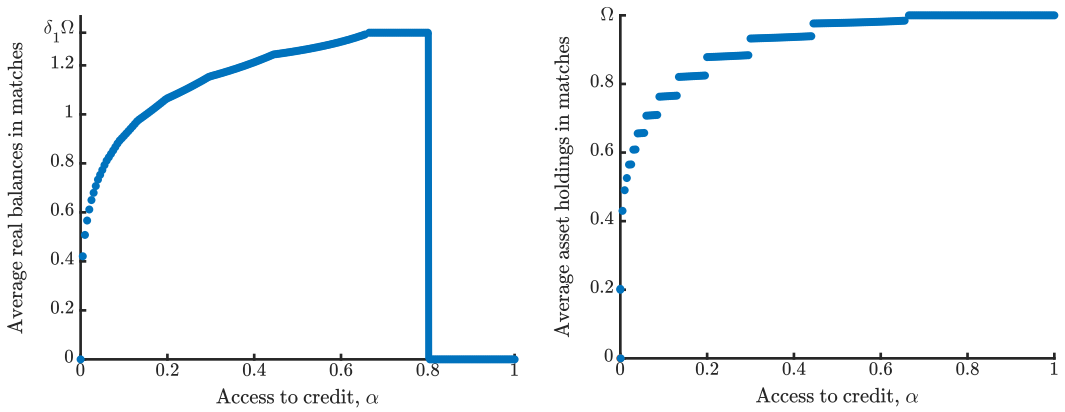

Figure 1.8: Left panel: Average payment capacity (i.e., real balances) of h-investors as a function of α. Right panel: Average inventories (i.e., participation) of ℓ-investors as a function of α.

in trade volume.[30] In contrast, were money and credit substitutes, the tightening of credit conditions would be mitigated by an increase in real balances holdings.

While the discussion above focused on the negative feedback due to the complementarity between money and credit, it is also interesting to look at it from the opposite perspective. When credit is completely absent, money cannot be valued and no trade occurs. As soon as some credit is introduced, money gets valued and trade is restored: credit does not render money useless, it helps it get off the ground. Indeed, as soon as credit is available, some ℓ-investors will readily carry assets in their inventories, which in turn encourages h-investors to carry money for no-credit meetings.

Result 1.4 (Prices)**.** *The impact of α on the average price in money meetings and on the average price in the OTC market is ambiguous, and may be non-monotone.*

We described earlier how trade surplus in money matches is shared depending on the ratio of trade to payment capacities. When this ratio is low enough for payment constraints to

[30]In Brunnermeier and Pedersen (2009), the mechanism is different. In their model, tight funding liquidity encourages traders to reduce their positions, reducing market liquidity and increasing volatility. This in turn increases the risk of financing a trade, thus increasing margins and decreasing funding liquidity further. In this paper, access to credit is exogenous and therefore cannot react to investors' reduced inventories, but overall funding liquidity does react (negatively) through the decrease in real balances holdings.

be slack, the surplus is split in proportion to the agents' bargaining powers, and the price is $p_m = \delta_2$. When the ratio is high, such that payment constraints are binding, the surplus sharing rule favors the h-investor more, and the price is lower, $p_m = \delta_1$.[31] When access to credit diminishes, both the average trade capacity and the average payment capacity decrease. The total impact on prices is ambiguous: if average trade capacities decrease more than average payment capacities, then the average price in money meetings increases, and vice versa. In the numerical example presented earlier, the former occurs. As access to credit is tightened, the ℓ-investor is compensated for lower trade volumes by better terms of trade in money meetings, i.e., a larger share of bargaining surplus and higher prices. Note that the ambiguity in the impact of credit on prices is due to the complementarity of money and credit. Were holdings of real balances not pushed down in reaction to a credit crunch, prices would only be impacted by the inventory reduction, and would unambiguously increase.[32] This is what happens, for example, when the asset market is competitive (assuming preferences with a satiation point).

The second part of the result has to do with the average asset price across all meetings in the OTC market. Note that the asset price in credit meetings is $p_c = \delta_2$, and it is always greater than the average price in money meetings. As access to credit diminishes, more and more trades occur at the money-only price, putting downwards pressure on the aggregate price. However, the price in those money-only meetings can concurrently increase, putting upwards pressure on the aggregate price. The aggregate impact is ambiguous, and may be monotone. In our numerical example, the asset price increases in α conditional on N remaining the same, however there exists a downwards trend overall, as each decrease in N generates a

[31] This implies a negative relation between inventory size and prices, which conforms with what can be observed in OTC asset markets. For example, Friewald and Nagler (2019) highlight that a reduction in inventory goes with an increase in returns for dealers in the corporate bond markets, and Li and Schürhoff (2019) document a similar phenomenon in the municipal bonds market.

[32] Were money and credit substitutes, the increase in real balances would amplify the price increase.

price drop.[33,34]

1.5 Conclusion

This paper studies the role of credit constraints in a textbook model of over-the-counter asset market, when both payment and inventory constraints are made explicit and endogenous. Investors choose the amount of real balances they wish to carry, as well as their trade capacity, which is interpreted as a participation decision at the intensive margin.

In this environment, credit is not neutral, and can severely impact the well-functioning of the asset market even when money is available and valued. At the limit when there is no access to credit, there is no participation—the market completely shuts down despite the existence of gains from trade. This is due to strategic interactions between the investors, who internalize the impact of their participation decision on the bilateral terms of trade and undercut each other to the bottom so as to try and obtain better prices.

The introduction of credit in some positive measure of meetings allows trade to be revived and money to be valued. Indeed, the existence of credit encourages the participation of sellers to the market, since they can liquidate all of their inventory at a high price in credit meetings, which in turn encourages buyers to carry money to make payments in case credit is not available. Those strategic interactions between investors generate endogenous heterogeneity in participation, money holdings, trade sizes, and prices, akin to empirical regularities observed in OTC markets. In this light, the present paper adds to the literature that studies dispersion in asset markets and suggests strategic interactions between investors who

[33]The non-monotonicity seems to be a feature of the multiplicity of equilibria in this parameter region, and it is likely that other predictions could be made by studying other equilibrium structures.

[34]Empirical evidence of the relation between access to funding and prices in OTC markets is mixed. Event studies surveyed by Fortune et al. (2001) and Pruitt and Tse (1996) look at OTC-traded stocks after they were added to the list of marginable OTC stocks, thereby easing their financing, find no consensus. Macchiavelli and Zhou (2019) find that in the corporate bond market, dealers increase their bid-ask spread when they have less access to repo funding for their inventories.

make participation (or inventory) decisions as a novel channel to generate such dispersion endogenously.

An important implication of the model is that money and credit are behave as complements. While it prevents the shutdown of the market as long as some credit exists, this complementarity can also worsen the impact of a credit crunch. Indeed, following a tightening of credit, sellers reduce their inventories and trade volume diminishes. This fall in market liquidity drives buyers to hold fewer real balances, which aggravates the reduction in funding liquidity in the market.

These outcomes stand in contrast to those derived in an economy where the asset market is competitive, in which participation is irrelevant, money and credit are substitutes, and changes in the availability of credit are fully absorbed by variations in prices.

While the implications of two-sided capacity constraints and their interaction with credit were studied in the context of an asset market, these modeling devices are applicable to a wide range of environments. One such example is the decentralized retail market described in Berentsen et al. (2011), where buyers are subject to payment constraints a la Lagos and Wright (2005), and sellers are subject to inventory constraints due to production being done ex-ante in a frictional labor market. Allowing for participation decisions in this framework would likely give access to credit in the retail market a significant role for equilibrium unemployment.

Another avenue for research lies in exploring further how market structure impacts the mechanism described in this paper. For example, one could explore further how the addition of dealers with access to a competitive interdealer market, as described in Lagos and Zhang (2020), may play a role similar to that played by credit, however relaxing the capacity constraint on the other side of the market.

Chapter 2

Gradual Bargaining in Decentralized Asset Markets

with G. Rocheteau, T.-W. Hu and Y. In

2.1 Introduction

Modern monetary theory and financial economics formalize asset trades in the context of decentralized markets with explicit game-theoretic foundations (e.g., Duffie et al. (2005), Lagos and Wright (2005)). These models replace the elusive Walrasian auctioneer by a market structure with two core components: a technology to form pairwise meetings and a strategic or axiomatic mechanism to determine prices and trade sizes. This paper focuses on the latter: the negotiation of asset prices and trade sizes in pairwise meetings.

Going back to Diamond (1982), the search-theoretic literature has placed stark restrictions on individual asset inventories, typically $a \in \{0, 1\}$. As a result, in versions of the model with bargaining (e.g., Shi (1995) , Trejos and Wright (1995), Duffie et al. (2005)), the only

item to negotiate in pairwise meetings — the *agenda* of the negotiation — is the price of an indivisible asset in terms of a divisible commodity.[1] Recent incarnations of the model (surveyed in Lagos et al. (2017)) allow for unrestricted portfolios of divisible assets, $\mathbf{a} \in \mathbb{R}_+^J$ with $J \in \mathbb{N}$. A key conceptual difference when $\mathbf{a} \in \mathbb{R}_+^J$ is that the agenda of the negotiation is not unique. Any ordered partition of $\mathbf{a} \in \mathbb{R}_+^J$ constitutes an agenda, where the elements of this partition correspond to items to be negotiated sequentially. For instance, agents can sell their whole portfolio at once, as a large block, or they can partition their portfolio into bundles of varying compositions and sizes to be added to the negotiation table one after another.

The possibility of negotiating asset sales according to different agendas raises several questions regarding trading strategies and price formation in decentralized asset markets. Do agendas matter for asset prices and trade sizes when agents have perfect foresight and information is complete? What is the optimal strategy of the asset owner to partition his portfolio, e.g., should the portfolio be negotiated as a whole or divided into smaller bundles? What is the relation between the bargaining problem with an agenda and the bargaining solution of Nash (1950)?

Our contribution is to introduce a new and generalized approach to bargaining over portfolios of assets in models of decentralized asset markets with the notion of agenda at the forefront, under both strategic and axiomatic foundations. The paper is composed of two parts. The first part provides a detailed description of bargaining games with an agenda and derives a series of methodological results that will be useful to incorporate these bargaining games into a general market structure. The second part focuses on the general equilibrium and derives some implications of the agenda of the negotiation for asset prices, allocations, and

[1] A thorough treatment of the axiomatic and strategic solutions for such bargaining problems is provided by Osborne and Rubinstein (1990). In Osborne and Rubinstein (1990) agents trade an indivisible consumption good and pay with transferable utility. The interpretation is reversed in Shi (1995) and Trejos and Wright (1995)where the indivisible good is fiat money and agents negotiate over a divisible consumption good. In Duffie et al. (2005) the indivisible good is a consol and agents pay with transferable utility.

welfare.

We start with a simple agenda that partitions a portfolio of homogeneous assets into N bundles of equal sizes. This agenda is a natural extension of the negotiation in Shi (1995) and Trejos and Wright (1995), where the indivisible asset is now interpreted as a bundle of divisible assets. The extensive-form bargaining game, called the alternating-ultimatum-offer game, is composed of N rounds. In each round, one asset bundle is up for negotiation. One player makes an ultimatum offer, and the identity of the proposer alternates across rounds. Agents are forward-looking and can anticipate the outcomes of future rounds. In contrast to the Rubinstein game, our game is nonstationary, since the amount of assets left for negotiation decreases over time, and it admits a unique subgame-perfect equilibrium (SPE) characterized by a system of difference equations with initial condition allowing us to compute the terminal allocation in closed-form for all N.

The limit as N goes to infinity is called the *gradual solution*. It gives a simple and intuitive relationship between asset prices and trade sizes, and it has properties distinct from the Nash (1950) solution that make it tractable for general equilibrium analysis, including monotonicity and concavity of trade surpluses with respect to trade size. Moreover, it coincides with the axiomatic ordinal solution of O'Neill et al. (2004) where an agenda is defined as a collection of Pareto frontiers indexed by time.

In order to relate our approach to the Nash solution, commonly used in the asset market literature (e.g., Duffie et al. (2005), Lagos and Wright (2005)), we extend our N-round game by assuming that in each round agents play an alternating-offer game with risk of breakdown, as in Rubinstein (1982). The equilibrium allocation of this N-round game is obtained by applying the Nash solution consecutively N times, where the solution in one iteration becomes the disagreement point of the next iteration. We characterize the outcome in closed form for all N and show it coincides with the Nash solution and the gradual solution in the two limiting cases $N = 1$ and $N = +\infty$, respectively. We endogenize the agenda of

the negotiation by letting asset owners choose N to maximize their surplus from trade. The optimal choice is $N = +\infty$, i.e., it is optimal for the owner to add assets on the bargaining table gradually, one infinitesimal unit at a time.

The second part of the paper incorporates bargaining solutions with an agenda into a general equilibrium model of decentralized asset markets with endogenous portfolios along the lines of Lagos and Wright (2005) and Lagos and Zhang (2020). The equilibrium under Nash bargaining ($N = 1$) features asset misallocation: a fraction of the asset supply ends up being held by agents with no liquidity needs. In contrast, under gradual bargaining ($N = +\infty$), the first best is implemented as long as the asset supply is sufficiently abundant. In the case of fiat money, the optimal policy, the Friedman rule, generates the first best under gradual bargaining for all bargaining powers whereas it fails to do so under generalized Nash bargaining as long as producers have some bargaining power. Using the same calibrated parameter values as in Lagos and Wright (2005), going from $N = 1$ to $N = +\infty$ increases output and consumption at the optimal policy by 76%. Even a moderate increase from $N = 1$ to $N = 5$ raises output by 39%.

This finding is especially stark in a monetary version of the model where agents trade short-lived assets that they value according to linear preferences, e.g., as in the model of OTC market of Lagos and Zhang (2020). We allow agents to choose how much of their short-lived assets to bring into a match. Under Nash bargaining, the OTC market shuts down for all interest rates and the equilibrium achieves its worst allocation. This result is a direct consequence of the non-monotonicity of agents' surpluses with respect to the quantity of goods or assets that they bring to the negotiation table. Under gradual bargaining, the OTC market is active and the equilibrium achieves first best for all interest rates below a positive threshold.

Finally, we extend our environment to allow for any arbitrary number of assets. All assets, except fiat money, generate the same stream of dividends. The notion of agenda allows us to

introduce a new asset characteristic – negotiability – defined as the inverse of the amount of time required for the sale of each unit of the asset to be finalized, e.g., each asset added to the negotiation table needs to be authenticated and ownership rights take time to transfer.[2] Our model generates an endogenous pecking order: assets that are more negotiable are put on the negotiating table before the less negotiable ones. In equilibrium, the most negotiable assets have lower rates of return and higher velocities. Hence, our model explains rate-of-return differences of seemingly identical assets. As the time horizon of the negotiation becomes arbitrarily large, differences in rates of return vanish but differences in velocities persist. We discuss the potential of our model to address two puzzles in monetary theory, the rate-of-return dominance puzzle and the indeterminacy of the exchange between two fiat currencies.

Related literature

Models of decentralized markets adopting a strategic approach to the bargaining problem in pairwise meetings were pioneered by Rubinstein and Wolinsky (1985). Bargaining with an agenda composed of multiple issues was first studied by Fershtman (1990). The axiomatic formulation with a continuous agenda was developed by O'Neill et al. (2004). We provide both its first application in the context of decentralized asset market models and strategic foundations with two extensive-form games that admit as limiting outcomes the ordinal solution of O'Neill et al. (2004). Another important distinction relative to the work of O'Neill et al. is the fact that we specify the agenda in terms of the agents' initial endowments or asset holdings and not only in terms of utility space. Our approach allows us to identify agendas that are meaningful in the context of decentralized markets.

[2]The concept of negotiability dates back to the 17th century and referred to institutional arrangements aiming at enhancing liquidity by "centralizing all rights to the underlying asset in a single physical document, [...] reducing the costs a prospective purchaser incurs in acquiring [...] information about the asset" (Mann (1996)). The concept of blockchains - immutable, decentralized ledgers that can record ownership and transfer of intangible assets - can be seen as a digital incarnation of the original idea of negotiability.

Our contribution on the strategic foundations was influenced by an earlier working paper by Wiener and Winter (1998) where they assert in Section 8 that the relevant limits of three distinct bargaining games with alternating offers should generate the same outcome as the ordinal solution of O'Neill et al. (2004).[3] We provide a complete and rigorous proof of this statement in the context of an over-the-counter bargaining game with forward-looking agents and liquidity constraints.

Our second game based on a "repeated" Stahl-Rubinstein game is related to the Stole and Zwiebel (1996) game in the literature on intra-firm wage bargaining. See Brügemann et al. (2019) for a recent re-examination of this game. Some of the key differences are as follows. In the intra-firm bargaining literature workers sell an indivisible unit of labor, whereas in models of asset markets agents sell divisible assets. Moreover, we let agents choose both the quantity of assets to sell and the number of rounds of the negotiation. The extensive form of the game is also different. In our game, if agents fail to reach an agreement in one round, they move to the next round, but the agreements of earlier rounds are preserved. In the Stole-Zwiebel game, all previous agreements are erased.

The "repeated" Rubinstein game is also used in Hu and Rocheteau (2020) to establish strategic foundations for the Kalai (1977) solution in a bargaining game with liquidity constraints. To that end, however, they use the agenda according to which bundles of goods are negotiated sequentially. In contrast, here we are interested in sequential sales of asset bundles and we obtain a new solution concept, the gradual Nash solution, in the context of decentralized asset markets.

The general equilibrium framework into which we incorporate bargaining games with an agenda corresponds to a version of the Lagos and Wright (2005) model with divisible Lucas

[3]The relevant results from Wiener and Winter (1998) are contained in their Propositions 5 and 7. Because their Appendix 4 only contains sketches of proofs, it is unclear whether those results would apply to our setting with forward-looking agents and liquidity constraints. Hence, we set up precise extensive-form games and prove equivalence results in the context of our model.

trees, as in Geromichalos et al. (2007) and Lagos (2010).[4] We also consider a variant where agents trade assets because of idiosyncratic valuations, as in Duffie et al. (2005). See also Lagos and Rocheteau (2009) and Üslü (2019) with unrestricted portfolios; Geromichalos and Herrenbrueck (2016a), Lagos and Zhang (2020), and Wright et al. (2020), with asset trades financed with money.[5]. We are the first ones to point out the importance of the agenda of the bargaining game for qualitative and quantitative results.

Our extension with multiple assets contributes to the literature on asset price puzzles in markets with search frictions, e.g., Vayanos and Weill (2008) based on increasing-returns-to-scale matching technologies; Rocheteau (2011), Li et al. (2012) and Hu and Rocheteau (2013) based on informational asymmetries; Lagos (2013) based on self-fulfilling beliefs in the presence of assets' extrinsic characteristics; and Geromichalos and Herrenbrueck (2016b) based on matching and bargaining friction differentials across the secondary markets where each asset is traded. Closer to what we do, Zhu and Wallace (2007) explain the coexistence of money and interest-bearing bonds using a bargaining protocol with a two-item agenda where bargaining powers vary with the item under negotiation. In contrast to their approach, our bargaining solution has both axiomatic and strategic foundations and we do not make bargaining powers specific to the asset being negotiated.

2.2 The gradual bargaining game

In this section we describe an OTC bargaining game whereby two players negotiate the sale of divisible assets in exchange for consumption goods. We set up the game and its payoffs so that it can easily be embedded into an off-the-shelf general equilibrium model of decentralized asset markets in Section 2.4. In this section and the next we provide a series of

[4] In those models, the asset owner has all the bargaining power. Rocheteau and Wright (2013) adopt the proportional bargaining solution, endogenize participation, and consider non-stationary equilibria.

[5] See Trejos and Wright (2016) for a model that nests Shi (1995), Trejos and Wright (1995) and Duffie et al. (2005)

methodological results regarding OTC bargaining games with an agenda, their axiomatic and strategic foundations, and their positive and normative implications. This section focuses on a simple extensive-form game, called the alternating-ultimatum-offer bargaining game, and its relationship to an axiomatic solution provided by O'Neill et al. (2004). Section 2.3 generalizes this extensive-form game to establish a connection with the Nash bargaining solution, commonly used in the literature, and endogenizes the choice of the agenda.

The bargaining game is composed of two players, called *consumer* and *producer*, who negotiate the sale of z units of an asset in exchange for units of a commodity labeled decentralized market (DM) good.[6] See left panel of Figure 2.1. The labels *consumer* and *producer* refer to agents' roles regarding the DM good. The consumer is the buyer of the DM good and the seller of the asset while the producer is the seller of the DM good and hence the buyer of the asset. The DM good is produced on the spot once an agreement is reached. We interpret $z > 0$ as the total asset holdings of the consumer that are up for sale. This quantity will be endogenized in Section 2.4 by allowing agents to make a portfolio choice. An outcome of the negotiation is a pair $(y, p) \in \mathbb{R}_+ \times [0, z]$ where p is the amount of assets sold for y units of the DM goods. Preferences over outcomes are represented by the following payoff functions:

$$\begin{aligned} u^b &= u(y) - p + u_0^b \\ u^s &= -\upsilon(y) + p + u_0^s, \end{aligned}$$

where u_0^b and u_0^s are the payoffs in case of disagreement (endogenized in general equilibrium later) and the superscripts b and s stand for buyer and seller of the DM good. As is standard in the search-theoretic literature on asset markets, payoffs are linear in p, hence the asset transfers utility perfectly across players up to the amount z. In contrast to p, the DM

[6]The DM good has been given different interpretations in the New Monetarist literature: a perishable consumption good or service (e.g., Lagos and Wright, 2005), physical capital (e.g., Wright et al., 2020), or an illiquid consol that is valued differently by different players (e.g., Duffie et al., 2005).

good does not transfer utility perfectly across players, i.e., in general $u'(y) \neq v'(y)$. More specifically, we assume $u'(y) > 0$, $u''(y) < 0$, $u'(0) = +\infty$, $u(0) = v(0) = v'(0) = 0$, $v'(y) > 0$, $v''(y) > 0$, and $u'(y^*) = v'(y^*)$ for some $y^* > 0$.[7] Preferences and asset holdings are common knowledge. We illustrate the determination of the players' payoffs from a trade (y^e, p^e) in the right panel of Figure 2.1 where disagreement points are normalized to $u_0^b = u_0^s = 0$.

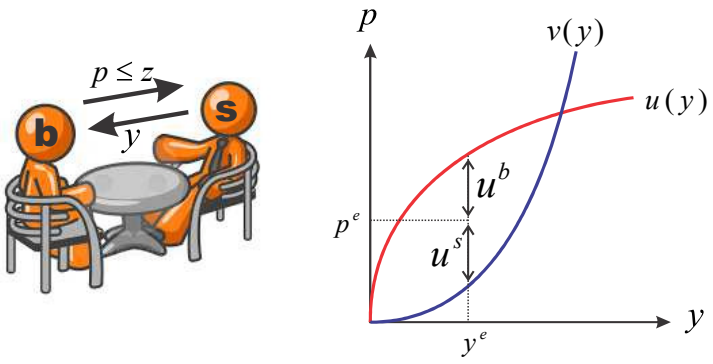

Figure 2.1: Left: Bilateral negotiation between consumer (b) and producer (s). Right: Payoffs of the gradual bargaining game.

In the following we first propose an extensive-form game to determine (y, p) and then we adopt an axiomatic approach to show the robustness of the solution.

2.2.1 The alternating-ultimatum-offer bargaining game

The game has N rounds. In each round, the consumer can negotiate at most z/N units of assets for some DM output. The round-game corresponds to a two-stage ultimatum game: in the first stage an offer is made; in the second stage the offer is accepted or rejected.[8] In order to maintain some symmetry between the two players (when N is large), the identity

[7] The Inada condition on $u(y) - v(y)$ is only needed when we incorporate the bargaining game into a general equilibrium structure. The concavity assumption makes the set of feasible utilities convex and it will allow us to obtain uniqueness of the general equilibrium later.

[8] A feature of our game is that if an offer is rejected, the z/N units of assets that are unsold cannot be renegotiated later in the game. While this assumption is no different from the one in standard ultimatum games (i.e., agents are committed to the rules of the game), the solution to our game, however, is robust to this feature, i.e., the game could include more than N rounds to allow for some amount of renegotiation. See Appendix B.2.

of the proposer alternates across rounds.[9] We assume N is even and the producer is the one making the first offer. These assumptions will be inconsequential when we consider the limit as N becomes large. The game tree is represented in Figure 2.2.

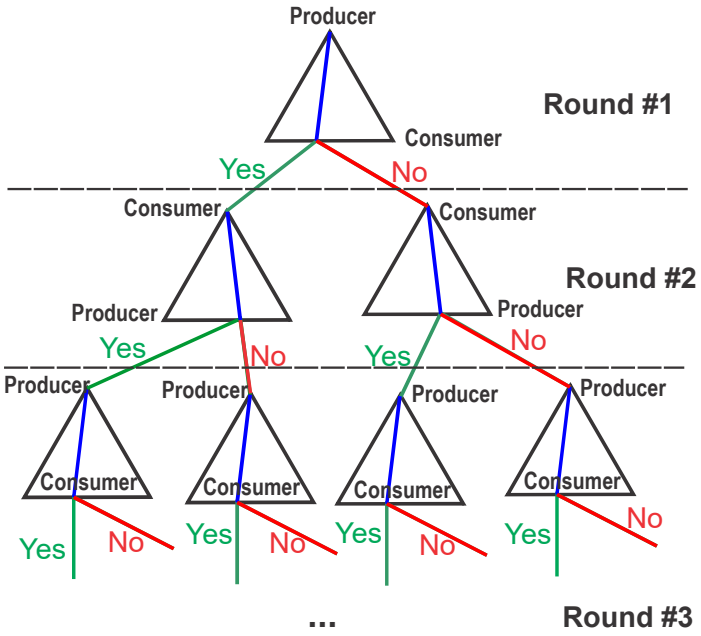

Figure 2.2: Game tree of the alternating-ultimatum-offer game.

In order to solve for the equilibrium, it is useful to introduce an explicit notion of time in the negotiation, denoted by τ. We map asset holdings into time by assuming that $\delta > 0$ units of asset can be negotiated per unit of time. Hence, $\tau \equiv nz/(\delta N)$ is the time at the end of the n^{th} round of the negotiation (in each of the n rounds, z/N assets are up for negotiation, and each asset takes $1/\delta$ units of time to be negotiated). We will rely heavily on δ in our general equilibrium model with multiple assets of Section 2.5. The utility accumulated by the consumer up to time τ is

$$u^b(\tau) = u\left[y(\tau)\right] - p(\tau) + u_0^b, \tag{2.1}$$

where $y(\tau)$ is the consumer's cumulative consumption at time τ, $p(\tau)$ is his cumulative

[9] Our game resembles the finite bargaining game with alternating offers of Ståhl (1972). It differs from it in that players are negotiating different items in each round.

payment with the asset. The utility accumulated by the producer up to τ is

$$u^s(\tau) = -\upsilon\left[y(\tau)\right] + p(\tau) + u_0^s. \tag{2.2}$$

Given the feasibility constraint $p(\tau) \leq \delta\tau$, we can define a Pareto frontier for each τ, i.e.,

$$u^b = \max_{y,p\leq\delta\tau} \left\{u\left(y\right) - p + u_0^b\right\} \quad \text{s.t.} \quad -\upsilon\left(y\right) + p + u_0^s \geq u^s.$$

These Pareto frontiers play a key role to solve for the SPE of the game by backward induction.

Lemma 2.1. ***(Pareto frontiers)*** *The Pareto frontier at time τ satisfies $H(u^b, u^s, \tau) = 0$ where*

$$H(u^b, u^s, \tau) = \begin{cases} u(y^*) - \upsilon(y^*) - (u^b - u_0^b) - (u^s - u_0^s) & \text{if } \ u^s - u_0^s \leq \delta\tau - \upsilon(y^*) \\ \delta\tau - \upsilon[u^{-1}(\delta\tau + u^b - u_0^b)] - (u^s - u_0^s) & \text{otherwise.} \end{cases} \tag{2.3}$$

The function H is continuously differentiable, increasing in τ (strictly so if $y < y^*$), decreasing in u^b and u^s. Consequently, each Pareto frontier has a negative slope:

$$\left.\frac{\partial u^s}{\partial u^b}\right|_{H(u^b,u^s,\tau)=0} = \begin{cases} -1 & \text{if } \ u^s - u_0^s \leq \delta\tau - \upsilon(y^*) \\ -\frac{\upsilon'(y)}{u'(y)} & \text{otherwise.} \end{cases}$$

The Pareto frontier is linear when $y = y^*$. When $y < y^*$, it is strictly concave.

We call a bargaining round an *active round* if there is trade. We say that a SPE is *simple* if in each active round the consumer offers z/N units of assets, except possibly for the last active round, and active rounds are followed by inactive rounds (if any).

Proposition 2.1. ***(SPE of the alternating-ultimatum-offer game.)*** *All SPE of the alternating-ultimatum-offer game share the same final payoffs, $(\widetilde{u}_N^b, \widetilde{u}_N^s)$, corresponding to*

the last term of the sequence, $\{(\widetilde{u}_j^b, \widetilde{u}_j^s)\}_{j=0}^N$*, with* $(\widetilde{u}_0^b, \widetilde{u}_0^s) = (u_0^b, u_0^s)$*, and*

$$H(\widetilde{u}_j^b, \widetilde{u}_{j-1}^s, jz/N) = 0 \text{ and } \widetilde{u}_j^s = \widetilde{u}_{j-1}^s, \text{ for } j \geq 1 \text{ odd}, \tag{2.4}$$

$$H(\widetilde{u}_{j-1}^b, \widetilde{u}_j^s, jz/N) = 0 \text{ and } \widetilde{u}_j^b = \widetilde{u}_{j-1}^b, \text{ for } j \geq 2 \text{ even}. \tag{2.5}$$

If the final y *is less than* y^**, then the SPE is unique and simple; otherwise, there is a unique simple SPE. Moreover, in any simple SPE, the intermediate payoffs,* $\{(u_n^b, u_n^s)\}_{n=1,2,\ldots,N}$*, converge to the solution,* $\left\langle u^b(\tau), u^s(\tau)\right\rangle$*, to the following differential equations as* N *approaches* $+\infty$*:*

$$u^{b\prime}(\tau) = -\frac{1}{2}\frac{\partial H(u^b, u^s, \tau)/\partial \tau}{\partial H(u^b, u^s, \tau)/\partial u^b} \tag{2.6}$$

$$u^{s\prime}(\tau) = -\frac{1}{2}\frac{\partial H(u^b, u^s, \tau)/\partial \tau}{\partial H(u^b, u^s, \tau)/\partial u^s}. \tag{2.7}$$

Proposition 2.1 (proved in Appendix B.2) establishes that the SPE of the alternating-ultimatum-offer game is essentially unique — any multiplicity when $y = y^*$ is due to differences in the timing of asset sales that are payoff-irrelevant. According to (2.4)-(2.5) the terminal payoffs are obtained as the final terms of a simple recursion whereby in any odd periods j the producer's utility is unchanged relative to the previous round, $j-1$, and the consumer's payoff is chosen so that the pair of utilities belong to the Pareto frontier corresponding to the asset holdings jz/N. Thus, the consumer gets the full surplus in odd rounds while the producer receives full surplus in even rounds.

When N approaches $+\infty$, i.e., when bargaining becomes gradual, equilibrium payoffs are characterized by the system of differential equations, (2.6)-(2.7). The interpretation of this solution is as follows. An increase in τ by one unit expands the bargaining set by

$\partial H/\partial\tau$. The maximum utility gain that the consumer could enjoy from this expansion is $-\left(\partial H/\partial\tau\right)/\left(\partial H/\partial u^b\right)$, as illustrated by the horizontal arrow in Figure 2.3. According to (2.6), the consumer enjoys half of this gain. The same holds true for the producer. By combining (2.6) and (2.7), the slope of the gradual agreement path is:

$$\frac{\partial u^s}{\partial u^b} = \frac{\partial H(u^b, u^s, \tau)/\partial u^b}{\partial H(u^b, u^s, \tau)/\partial u^s}. \tag{2.8}$$

According to (2.8), the slope of the gradual bargaining path is equal to the opposite of the slope of the Pareto frontier.

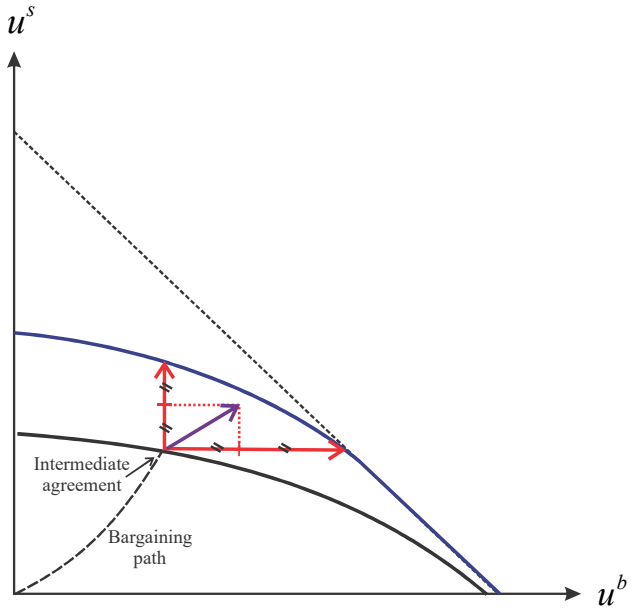

Figure 2.3: Solution to a gradual bargaining problem.

The proof of Proposition 2.1 consists of two steps: first, we characterize the SPE for any (sub)game with an arbitrary number of remaining rounds, J. In the second part, we establish that the sequence of intermediate payoffs of the SPE converges to the solution to the system of differential equations, (2.6) and (2.7), as N approaches $+\infty$. The intuition goes as follows. Suppose the negotiation enters its last round, N, and the two agents have agreed upon some intermediate payoffs (u^b_{N-1}, u^s_{N-1}). The consumer makes the last take-it-or-leave offer, which maximizes his payoff by keeping the producer's payoff unchanged at u^s_{N-1}. Graphi-

cally, the final payoffs are constructed from the intermediate payoffs by moving horizontally from the lower Pareto frontier, to which (u^b_{N-1}, u^s_{N-1}) belongs, to the upper Pareto frontier corresponding to an increase in assets of z/N, as shown in the left panel of Figure 2.4.

We now move backward in the game by one round. Suppose that the negotiation enters round $N-1$ with some intermediate payoffs, (u^b_{N-2}, u^s_{N-2}), with the producer making the offer. Now, if the consumer rejects the producer's offer, the negotiation enters its last round and the consumer's payoff is obtained as before, i.e., by moving horizontally from the lower frontier to the upper frontier. Given the consumer's payoff, the producer's payoff is obtained such that the pair of payoffs is located on the last Pareto frontier. Graphically, there is first a horizontal move from the initial payoff, (u^b_{N-2}, u^s_{N-2}), to the next Pareto frontier that determines the consumer's terminal payoff, and then a vertical move to the following frontier that determines the producer's payoff, u^s_N, as shown in the right panel of Figure 2.4. We iterate this procedure backward until we reach the start of the game with initial payoffs (u^b_0, u^s_0).

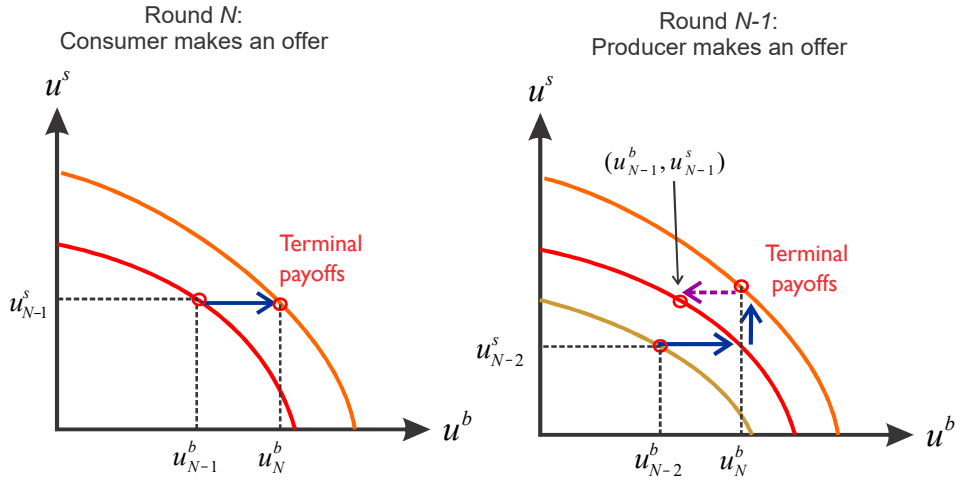

Figure 2.4: Left panel: offer in last round; Right panel: offer in $(N-1)^{\text{th}}$ round

Once we have the terminal payoffs, we use another backward induction to determine the sequence of intermediate payoffs. The intermediate payoffs at the end of the $(N-1)^{\text{th}}$ round lie on the $(N-1)^{\text{th}}$ frontier and are obtained by moving horizontally from the N^{th} frontier to

the $(N-1)^{\text{th}}$ frontier since the consumer is making the last offer. The intermediate payoffs on the $(N-2)^{\text{th}}$ frontier are obtained by moving first vertically, from the N^{th} frontier to the $(N-1)^{\text{th}}$ frontier, and then horizontally from the $(N-1)^{\text{th}}$ frontier to the $(N-2)^{\text{th}}$ frontier by using the same reasoning as above. It turns out that the two sequences constructed above get closer to one another as N becomes large, and, both converge to the gradual bargaining path according to (2.8).

We now characterize in closed form the final allocations. From (2.4)-(2.5) the final outcome corresponds to the last term of the sequence $\{(y_n, p_n)\}_{n=0}^{N}$ computed recursively from $(y_0, p_0) = (0, 0)$ as follows:

$$(y_n, p_n) \in \arg\max \{u(y_n) - u(y_{n-1}) - (p_n - p_{n-1})\} \quad \text{if } n \text{ odd} \tag{2.9}$$

$$(y_n, p_n) \in \arg\max \{(p_n - p_{n-1}) - [v(y_n) - v(y_{n-1})]\} \quad \text{if } n \text{ even}, \tag{2.10}$$

where each maximization problem is subject to the participation constraints,

$$v(y_n) - v(y_{n-1}) \leq p_n - p_{n-1} \leq u(y_n) - u(y_{n-1}), \tag{2.11}$$

and the feasibility condition,

$$p_n - p_{n-1} \leq \frac{z}{N}. \tag{2.12}$$

In odd periods (y_n, p_n) corresponds to a take-it-or-leave-it offer by the consumer whereas in even periods it coincides with a take-it-or-leave-it offer by the producer. If (2.12) binds in

each round, then the solution is:

$$y_n = \upsilon^{-1}\left[\frac{z}{N} + \upsilon(y_{n-1})\right] \quad \text{if } n \text{ odd} \tag{2.13}$$

$$y_n = u^{-1}\left[\frac{z}{N} + u(y_{n-1})\right] \quad \text{if } n \text{ even.} \tag{2.14}$$

So, the solution y_N can easily be computed given the initial condition, $y_0 = 0$.

2.2.2 Negotiated price and trade size

We now turn to the implications of the gradual bargaining solution for asset prices and trade sizes and focus on the limit case where N approaches infinity. We derive in closed form a payment function, $p(y)$, that specifies the quantity of assets required to purchase y units of goods, and that plays a critical role in models of asset liquidity (e.g., Lagos et al., 2017). From the definition of H in (2.3), the solution to the bargaining game, (2.6)-(2.7), can be reexpressed as

$$u^{b\prime}(\tau) = \delta\frac{u'(y) - \upsilon'(y)}{2\upsilon'(y)} \tag{2.15}$$

$$u^{s\prime}(\tau) = \delta\frac{u'(y) - \upsilon'(y)}{2u'(y)}, \tag{2.16}$$

if $\delta\tau < u^s - u_0^s + \upsilon(y^*)$, and $u^{b\prime}(\tau) = u^{s\prime}(\tau) = 0$ otherwise. From (2.15) and (2.16) the slope of the gradual bargaining path is $\partial u^s/\partial u^b = \upsilon'(y)/u'(y)$, which is increasing in y, i.e., it becomes steeper as the negotiation progresses. The producer's share in the match surplus increases throughout the negotiation as the gap between $u'(y)$ and $\upsilon'(y)$ shrinks over time.

Proposition 2.2. ***(Prices and trade sizes)*** *Along the gradual bargaining path, the price*

of the asset in terms of DM goods is

$$\frac{y'(\tau)}{\delta} = \frac{1}{2}\left(\overbrace{\frac{1}{\upsilon'(y)}}^{bid\ price} + \overbrace{\frac{1}{u'(y)}}^{ask\ price}\right) \quad for\ all\ y < y^*. \tag{2.17}$$

The overall payment for y units of consumption is

$$p(y) = \int_0^y \frac{2\upsilon'(x)u'(x)}{u'(x)+\upsilon'(x)}dx. \tag{2.18}$$

If $z \geq p(y^*)$ then $y = y^*$ and $y = p^{-1}(z)$ otherwise.

According to (2.17), the negotiated price is the arithmetic average of the bid and ask prices. The bid price of one unit of asset at time τ, i.e., the maximum price in terms of DM goods that the producer is willing to pay to acquire it, is equal to $1/\upsilon'(y)$. The ask price at time τ, i.e., the minimum price in terms of DM goods that the consumer is willing to accept to give up the asset, is $1/u'(y)$. The bid price decreases with y because the producer incurs a convex cost to finance an additional unit of asset. The ask price increases with y because the consumer enjoys a decreasing marginal utility in exchange of an additional unit of asset. So the negotiated price can be non-monotone with the size of the trade.

The payment function, (2.18), can be rewritten as

$$p(y) = \int_0^y \frac{\upsilon'(x)}{u'(x)+\upsilon'(x)}u'(x)dx + \int_0^y \frac{u'(x)}{u'(x)+\upsilon'(x)}\upsilon'(x)dx.$$

It is reminiscent of the payment function obtained from the Nash solution (e.g., Lagos and Wright, 2005) where $p^{\text{Nash}}(y) = [1-\Theta(y)]\,u(y) + \Theta(y)\upsilon(y)$ and $\Theta(x) \equiv u'(x)/\left[u'(x)+\upsilon'(x)\right]$ is interpreted as the consumer's share in the surplus of the match. In order to make the

connection clearer, we integrate $p(y)$ by parts to obtain:

$$p(y) = \overbrace{[1-\Theta(y)]\,u(y) + \Theta(y)\upsilon(y)}^{\text{Nash}} + \int_0^y \Theta'(x)\,[u(x)-\upsilon(x)]\,dx.$$

So the payment function under gradual bargaining is the sum of the payment function under Nash bargaining and an additional term that is negative since $\Theta'(x) < 0$. This additional term takes into account the change in the consumer's share over the gradual negotiation, i.e., as the negotiation advances the consumer's share decreases. We will come back to this comparison in the next section. It is worth noticing that $p(y)$ is independent of δ, and hence the outcome of the negotiation does not depend on the time it takes to negotiate assets sequentially: only N matters for the outcome. We will make δ relevant in Section 2.5 by assuming that the negotiation has a stochastic time horizon.

From (2.18) we can compute the consumer's surplus from a trade:

$$u(y) - p(y) = \int_0^y \frac{u'(x)\,[u'(x)-\upsilon'(x)]}{u'(x)+\upsilon'(x)}dx, \quad \text{for all } y \leq y^*.$$

The surplus increases with y, is strictly concave for all $y < y^*$, and is maximized at $y = y^*$. We will emphasize the importance of the monotonicity of the surplus for individual choices of asset holdings and asset prices later when we turn to the general equilibrium.

2.2.3 Asymmetric agenda

So far the agenda of the negotiation corresponds to a uniform partition of the portfolio, $[0, z]$, where each asset bundle has the same size, z/N. In the following we modify the agenda to provide a non-cooperative foundation for asymmetric bargaining powers. Such asymmetric solutions are useful in many applications to decentralized asset markets with endogenous participation and investment decisions. We still assume that N is even. In each

round where the consumer is making the offer, the amount of assets that can be negotiated is $2\theta z/N$ where $\theta \in [0,1]$. In rounds where the producer is making the offer, the amount of assets up for negotiation is $2(1-\theta)z/N$. Note that $\theta = 1/2$ corresponds to the bargaining game studied earlier. We show in Appendix B.2 that the solution to this bargaining game generalizes (2.6)-(2.7) as follows:

$$u^{b\prime}(\tau) = -\theta \frac{\partial H(u^b, u^s, \tau)/\partial \tau}{\partial H(u^b, u^s, \tau)/\partial u^b} \tag{2.19}$$

$$u^{s\prime}(\tau) = -(1-\theta) \frac{\partial H(u^b, u^s, \tau)/\partial \tau}{\partial H(u^b, u^s, \tau)/\partial u^s}, \tag{2.20}$$

where $\theta \in [0,1]$ is interpreted as the consumer's bargaining power.[10] By the same reasoning as above, the DM price of assets evolves according to

$$\frac{y'(\tau)}{\delta} = \left(\theta \overbrace{\frac{1}{\upsilon'(y)}}^{\text{bid price}} + (1-\theta) \overbrace{\frac{1}{u'(y)}}^{\text{ask price}} \right). \tag{2.21}$$

It is now a weighted average of the bid and ask prices where the weights are given by the relative bargaining powers of the consumer and the producer. From (2.21) the DM price of the asset is increasing in θ. The payment for y units of DM consumption is

$$p(y) = \int_0^y \frac{u'(x)\,\upsilon'(x)}{\theta u'(x) + (1-\theta)\upsilon'(x)} dx \quad \text{for all } y \leq y^*. \tag{2.22}$$

[10] This solution coincides with the axiomatic solution of Wiener and Winter (1998). One could make the bargaining power a function of time, τ, or output traded, y, without affecting the results significantly.

2.2.4 An axiomatic approach

An axiomatic approach, by abstracting from the details of the bargaining game, provides a sense of the robustness of our solution.[11] Nash (1950)'s definition of a bargaining problem, which does not include the notion of agenda, was extended by O'Neill et al. (2004). The agenda takes the form of a family of feasible sets indexed by time. The difficulty is to identify the relevant agenda for the problem at hand. In the context of our model where agents negotiate gradually the sale of assets, a gradual bargaining problem between a consumer holding z units of asset and a producer is a collection of Pareto frontiers, $\left\langle H(u^b, u^s, \tau) = 0, \tau \in [0, z/\delta]\right\rangle$ and a pair of disagreement points, (u_0^b, u_0^s).

A gradual agreement path is a function, $o : [0, z/\delta] \to \mathbb{R}_+ \times [0, z]$, that specifies an allocation (y, p) for all $\tau \in [0, z/\delta]$ and associated utility levels, $\left\langle u^b(\tau), u^s(\tau)\right\rangle$. The gradual solution of O'Neill et al. (2004) is the unique solution to satisfy five axioms: Pareto optimality, scale invariance, symmetry, directional continuity, and time consistency. The first three axioms are axioms imposed by Nash (1950) and are required to hold along the gradual agreement path. Formally, Pareto optimality means that $H\left[u^b(\tau), u^s(\tau), \tau\right] = 0$ for all τ. Scale invariance means that if $(\tilde{H}, \tilde{u}_0)$ is obtained by positive linear transformations from (H, u_0), then the gradual agreement path $\left\langle \tilde{u}^b(\tau), \tilde{u}^s(\tau)\right\rangle$ is obtained by the same linear transformations from $\left\langle u^b(\tau), u^s(\tau)\right\rangle$ and the underlying real allocations are unaffected. The axiom of symmetry requires that if (H, u_0) is symmetric, then $u^b(\tau) = u^s(\tau)$ for all τ. The last two axioms are specific to the new definition of the bargaining problem. The requirement of time consistency specifies that if the negotiation were to start with the agreement reached at

[11] As written by Serrano (2008) in his description of the Nash program:

> The non-cooperative approach to game theory provides a rich language and develops useful tools to analyze strategic situations. One clear advantage of the approach is that it is able to model how specific details of the interaction may impact the final outcome. One limitation, however, is that its predictions may be highly sensitive to those details. For this reason it is worth also analyzing more abstract approaches that attempt to obtain conclusions that are independent of such details. The cooperative approach is one such attempt.

time τ_0 as the new disagreement point, $(\tilde{u}_0^b, \tilde{u}_0^s) = \left[u^b(\tau_0), u_0^s(\tau_0)\right]$, then the bargaining path onward would be unchanged, i.e., $\left[\tilde{u}^b(\tau), \tilde{u}^s(\tau)\right] = \left[u^b(\tau+\tau_0), u_0^s(\tau+\tau_0)\right]$. The last axiom of directional continuity is more technical and imposes the following notion of continuity for the gradual agreement path. If two agendas H and $\tilde{H}$ are close in a neighborhood of (u_0, τ_0), then the rates of utility gains at u_0 for these two problems are also close. Formally, for a bounded neighborhood B of u_0, the proximity between two agendas H and $\tilde{H}$ is measured by $\sup_{u\in B}\left\|\nabla H(u,\tau_0) - \nabla\tilde{H}(u,\tau_0)\right\|$. Directional continuity requires that the utility gains, $\left[\partial u^b(\tau;H)/\partial\tau, \partial u^s(\tau;H)/\partial\tau\right]$, be continuous in H with respect to the metric above.

Theorem 1 of O'Neill et al. (2004) applied to our bargaining problem above shows that there is a unique solution that satisfies the five axioms of Pareto optimality, scale invariance, symmetry, directional continuity, and time consistency, and this solution coincides with (2.6)-(2.7). It means that the equilibrium payoffs of the alternating-ultimatum-offer bargaining game coincide with the axiomatic solution from O'Neill et al. (2004). While scale invariance was imposed as an axiom, O'Neill et al. (2004) show that the solution exhibits ordinality endogenously: the solution is covariant with respect to any order-preserving transformation. This result is noteworthy because Shapley (1969) shows that for standard Nash problems with two players, no single-valued solution can satisfy Pareto efficiency, symmetry, and ordinality. Finally, if the axiom of symmetry is dropped, then the generalized ordinal solutions solve (2.19)-(2.20).

2.3 Relation to Nash bargaining

The game studied in Section 2.2.1 is extended so that each round, $n \in \{1, ..., N\}$, is composed of an unbounded number of stages during which the two players bargain over z/N units of assets following an alternating-offer protocol as in Rubinstein (1982). The consumer is the first proposer if n is odd, and the producer is the first proposer otherwise. The round-game,

illustrated in Figure 2.5, is as follows. In the initial stage, the first proposer makes an offer and the other agent either accepts it or rejects it. If the offer is accepted, round n ends and agents move to round $n+1$. If the offer is rejected then there are two cases. With probability $(1-\xi_n)$ round n is terminated and the players move to round $n+1$ without having reached an agreement. With probability ξ_n the negotiation continues and the responder becomes the proposer in the following stage. We focus on the limit case where ξ_n converges to one, and the order of convergence is from ξ_N to ξ_1.

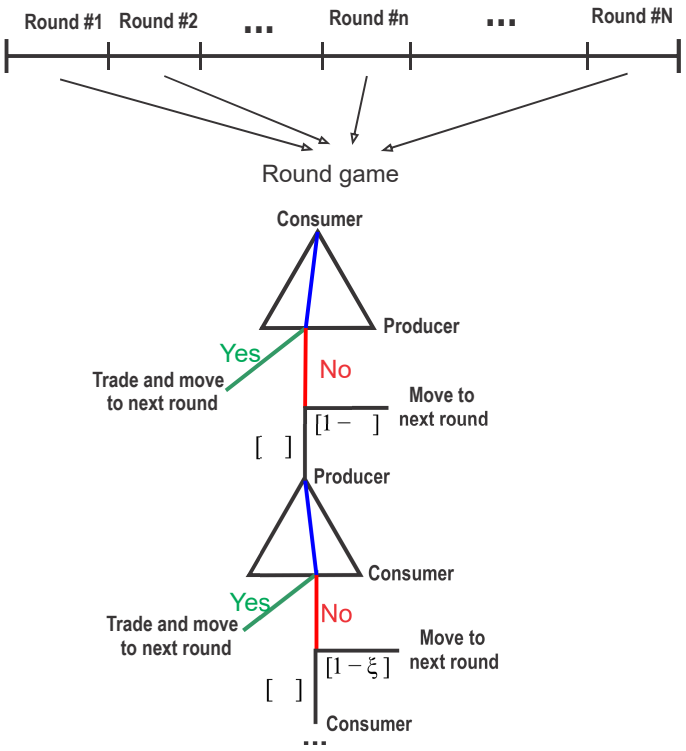

Figure 2.5: Game tree with alternating offers in each round.

Proposition 2.3. ***(Repeated Rubinstein game.)*** *There exists a SPE of the repeated Rubinstein game when taking limits according to the order* $\xi_N \to 1$, $\xi_{N-1} \to 1$, ..., $\xi_1 \to 1$, *characterized by a sequence of intermediate allocations,* $\{(y_n, p_n)\}_{n=0}^{N}$, *solution to:*

$$(y_n, p_n) \in \arg\max_{y,p} \left[u(y) - p - u(y_{n-1}) + p_{n-1}\right]\left[-v(y) + p + v(y_{n-1}) - p_{n-1}\right]$$
$$s.t. \quad p \leq \frac{nz}{N}, \qquad (2.23)$$

for all $n \in \{1, ..., N\}$ *with* $(y_0, p_0) = (0, 0)$. *As* $N \to +\infty$ *the solution converges to the solution of the alternating-ultimatum-offer game characterized in Proposition 2.2.*

The intermediate allocation at the end of each round, given by (2.23), maximizes the Nash product of agents' surpluses where the endogenous disagreement points are the intermediate payoffs of the previous round. The proof (in Appendix B.3) is based on backward induction. Consider the last round with some intermediate agreement (u^b_{N-1}, u^s_{N-1}). The outcome of the Rubinstein game as the risk of breakdown goes to zero corresponds to the Nash solution with disagreement point (u^b_{N-1}, u^s_{N-1}). Next, consider round $N-1$ with intermediate payoffs (u^b_{N-2}, u^s_{N-2}). The relevant disagreement points, $(\tilde{u}^b_{N-1}, \tilde{u}^s_{N-1})$, are given by the outcome of the negotiation in round N if there is no agreement in round $N-1$, i.e., $(\tilde{u}^b_{N-1}, \tilde{u}^s_{N-1})$ maximizes the Nash product $\left(\tilde{u}^b_{N-1} - u^b_{N-2}\right)\left(\tilde{u}^s_{N-1} - u^s_{N-2}\right)$. Given $(\tilde{u}^b_{N-1}, \tilde{u}^s_{N-1})$, the negotiation in round $N-1$, in which players are forward looking, determines the final payoffs. As the risk of breakdown vanishes, these payoffs, (u^b_N, u^s_N), coincide with the Nash solution, i.e., they maximize $\left(u^b_N - \tilde{u}^b_{N-1}\right)\left(u^s_N - \tilde{u}^s_{N-1}\right)$. For any given initial condition (u^b_0, u^s_0), this iterative procedure pins down the terminal payoffs. Once terminal payoffs are determined, we use a second backward induction to find the sequence of intermediate payoffs. Intermediate payoffs in round $N-1$, (u^b_{N-1}, u^s_{N-1}), correspond to the disagreement points of the Nash solution that generates the terminal payoffs, i.e., $(u^b_{N-1}, u^s_{N-1}) = (\tilde{u}^b_{N-1}, \tilde{u}^s_{N-1})$. And so on. The determination of payoffs is illustrated in Figure 2.6.

In order to fix the intuition, suppose $N = 2$. In the first round of the negotiation agents negotiate (y_1, p_1) taking into account that the outcome of the second round, (y_2, p_2), is a function of the interim agreement, (y_1, p_1). In case of disagreement, the players negotiate in a single round the remaining $z/2$ assets so that the allocation is given by

$$(\tilde{y}_1, \tilde{p}_1) \in \arg\max_{y,p} \left[u(y) - p\right]\left[-\upsilon(y) + p\right] \quad \text{s.t.} \quad p \leq \frac{z}{2}. \tag{2.24}$$

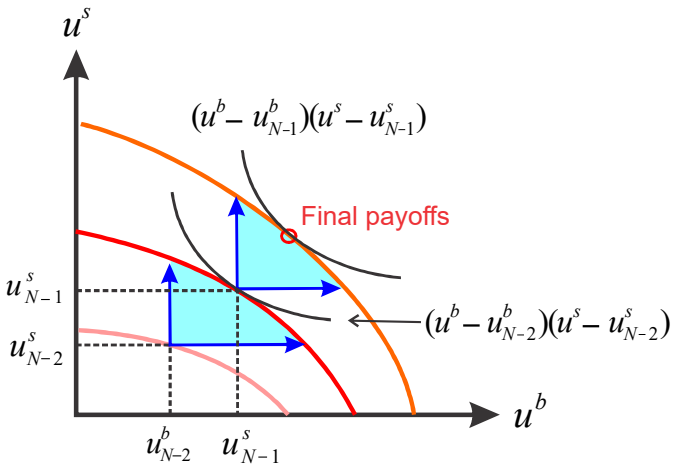

Figure 2.6: Computing terminal payoffs from round $N-1$.

Because agents in the first round can anticipate the outcome of the second round, they are negotiating the final outcome, (y, p) where $y = y_1 + y_2(y_1, p_1)$ and $p = p_1 + p_2(y_1, p_1)$. Hence, the final outcome is given by

$$(y,p) \in \arg\max_{y,p} \left[u(y) - p - u(\tilde{y}_1) + \tilde{p}_1\right]\left[-v(y) + p + v(\tilde{y}_1) - \tilde{p}_1\right] \quad \text{s.t.} \quad p \leq z. \tag{2.25}$$

In the second round, agents solve the same problem where the disagreement point correspond to the actual trade in the first round, (y_1, p_1), i.e.,

$$(y,p) \in \arg\max_{y,p} \left[u(y) - p - u(y_1) + p_1\right]\left[-v(y) + p + v(y_1) - p_1\right] \quad \text{s.t.} \quad p - p_1 \leq \frac{z}{2}. \tag{2.26}$$

The first round outcome is chosen so that (2.25) and (2.26) hold and it turns out that $(y_1, p_1) = (\tilde{y}_1, \tilde{p}_1)$. The interim trade in the first round corresponds to the trade that would take place in case of disagreement. The total surplus in both rounds is equal to $[u(y) - v(y)] - [u(\tilde{y}_1) - v(\tilde{y}_1)]$, i.e., agents are negotiating the marginal surplus in each round. It is easy to generalize the logic to $N > 2$. For instance, if $N = 3$ then the final outcome is negotiated in round 1 by forward-looking agents according to the Nash solution with disagreement points corresponding to the solution of (2.25).

From (2.23) the intermediate allocations, $\{(y_n, p_n)\}_{n=0}^{N}$, solve:

$$\int_{y_{n-1}}^{y_n} \frac{\upsilon'(y_n)u'(x) + u'(y_n)\upsilon'(x)}{u'(y_n) + \upsilon'(y_n)} dx \leq \frac{z}{N} \quad \text{" = " if } y_n < y^*, \tag{2.27}$$
$$p_n - p_{n-1} = \min\left\{ \frac{[u(y^*) - u(y_{n-1})] + [\upsilon(y^*) - \upsilon(y_{n-1})]}{2}, \frac{z}{N} \right\},$$

with $y_0 = 0$. From (2.27), when the liquidity constraint, $p_n \leq nz/N$, binds, then the payment for $y_n - y_{n-1}$ units of DM goods is equal to a weighted sum of the marginal utilities of consumption and the marginal disutilities of production. If $N = 1$ then (2.27) corresponds to symmetric Nash.[12] Summing (2.27) across n and taking the limit as N goes to $+\infty$ gives the gradual solution.

In the following proposition we let consumers (asset owners) choose the number of rounds of the negotiation, N. The key observation from (2.27) is that the consumer's share in the surplus of the n^{th} round, $u'(y_n)/\left[u'(y_n) + \upsilon'(y_n)\right]$, decreases with y_n.

Proposition 2.4. ***(Optimal gradualism)*** *Consumers obtain their highest surplus by negotiating the sale of their assets one infinitesimal unit at a time, $N = +\infty$.*

The agenda underlying the Nash solution ($N = 1$) is suboptimal from the standpoint of asset owners. They strictly prefer to sell their assets gradually over time. The consumer's gain from bargaining gradually is

$$p_1(y) - p_\infty(y) = \int_0^y \left[\frac{\upsilon'(y)}{u'(y) + \upsilon'(y)} - \frac{\upsilon'(x)}{u'(x) + \upsilon'(x)} \right] \left[u'(x) - \upsilon'(x)\right] dx,$$

where $p_1(y)$ is the amount of assets in exchange for y units of DM goods if the negotiation takes place in a single round, which implements the Nash solution. Under Nash bargaining the producer's share in each increment of the match surplus is constant and equal to $\upsilon'(y)/\left[u'(y) + \upsilon'(y)\right]$, which is larger than the variable share, $\upsilon'(x)/\left[u'(x) + \upsilon'(x)\right]$ for all

[12] In the Appendix B.4, we study a version of the game that implements the generalized Nash solution.

$x < y$, under gradual bargaining. Intuitively, selling all the assets at once has a negative impact on the consumer's surplus share that can be reduced by selling them through small quantities — a form of dynamic price discrimination.

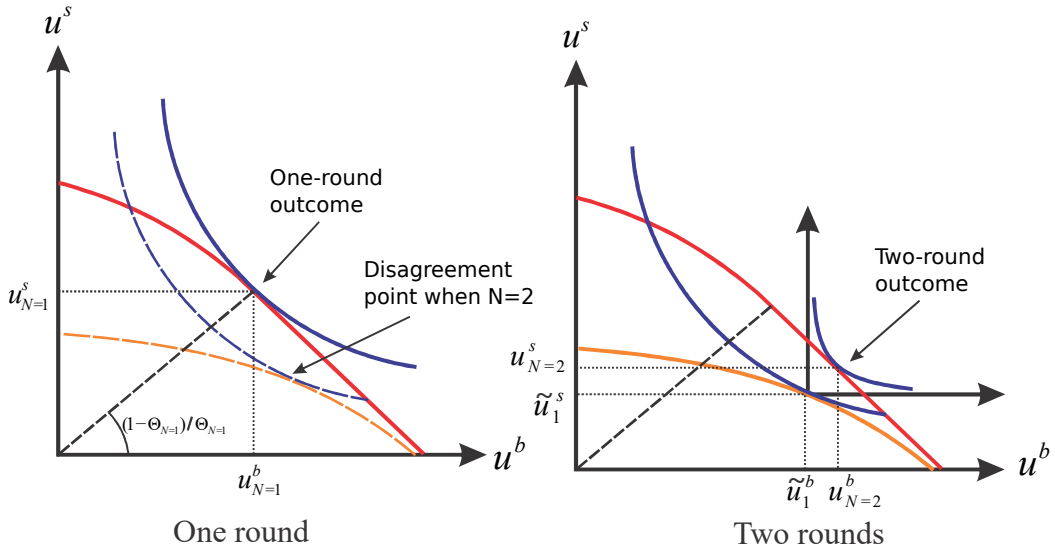

Figure 2.7: Comparison of one-round (left) vs two-round (right) bargaining

In order to deepen our intuition for why the consumer prefers to bargain gradually, let us compare the consumer's surplus in a negotiation with $N = 2$ rounds and the consumer's surplus in a negotiation with $N = 1$ round. Recall that irrespective of N, in each round of the negotiation agents are forward looking and are bargaining over the final outcome. By changing the number of rounds, N, one changes agents' disagreement points in the first round of the negotiation. The question is: how do disagreement points change with N, and how do these changes affect the final outcome? From (2.25) if the number of rounds increases from $N = 1$ to $N = 2$, the disagreement points in the initial round increase from $(0, 0)$ to $[u(\tilde{y}_1) - \tilde{p}_1, v(\tilde{y}_1) - \tilde{p}_1]$ where $(\tilde{y}_1, \tilde{p}_1)$ given by (2.24) is the outcome of a one-round negotiation when the consumer holds $z/2$ assets. In the panels of Figure 2.7, the disagreement point denoted $(\tilde{u}^b_1, \tilde{u}^s_1)$ is located at the intersection of the blue curve representing the Nash product and the orange Pareto frontier corresponding to $z/2$ units of asset. The Nash solution with disagreement point $(\tilde{u}^b_1, \tilde{u}^s_1)$ generates the same outcome as the Nash solution

with disagreement point $(0,0)$ if and only if

$$\frac{u(\tilde{y}_1)-\tilde{p}_1}{u(\tilde{y}_1)-v(\tilde{y}_1)}=\Theta(y_{N=1})\equiv\frac{u'(y_{N=1})}{u'(y_{N=1})+v'(y_{N=1})},$$

where $\Theta(y_{N=1})$ is consumer's share in the match surplus if the negotiation takes place in one round only and $y_{N=1}$ is the output level. If the players' shares of the surplus when negotiating over $z/2$ units of assets are equal to the shares when negotiating over z units, then the consumer does not gain from negotiating in multiple rounds. Graphically, in Figure 2.7, this condition requires $(\tilde{u}_1^b,\tilde{u}_1^s)$ to be located on the dashed line joining $(0,0)$ to $(u_{N=1}^b,u_{N=1}^s)$. Provided that $\tilde{y}_1<y^*$, i.e., $z<u(y^*)+v(y^*)$, the consumer's share when negotiating over $z/2$ is

$$\Theta(\tilde{y}_1)=\frac{u'(\tilde{y}_1)}{u'(\tilde{y}_1)+v'(\tilde{y}_1)}>\Theta(y_{N=1}),$$

because y is increasing in z, i.e., $\tilde{y}_1<y_{N=1}$, and $u'(y)/v'(y)$ is decreasing in y. Thus, the consumer receives a larger share of the surplus in $(\tilde{u}_1^b,\tilde{u}_1^s)$ than in $(u_{N=1}^b,u_{N=1}^s)$, and hence, in Figure 2.7, $(\tilde{u}_1^b,\tilde{u}_1^s)$ is located below the dashed line joining the origin to $(u_{N=1}^b,u_{N=1}^s)$. The quantity $u'(y)/v'(y)>1$ in the expression for Θ represents the marginal gain from trade. The tigther the consumer's liquidity constraint, the larger the marginal gain from trade, and the larger the consumer's share in the surplus.

How does the relationship between Θ and $u'(y)/v'(y)$ emerge from an alternating-offer game, i.e., why is Θ decreasing in y? Suppose the players adopt stationary strategies whereby the consumer offers y^b when it is her turn to make an offer and the producer offers $y^s<y^b$ when it is her turn, and suppose y^b and y^s are in the neighborhood of $y<y^*$. The consumer's gain from rejecting a producer's offer in order to make a counteroffer is approximately equal to $u'(y)\left(y^b-y^s\right)$ while the producer's gain from rejecting a consumer's offer in order to make a counteroffer is approximately equal to $v'(y)\left(y^b-y^s\right)$. Given that $y<y^*$, $u'(y)\left(y^b-y^s\right)>$

$v'(y)\left(y^b - y^s\right)$. If the surplus is divided evenly, the cost from missing on a trade in the event of a termination is equal for both players. It means that the consumer has a bigger incentive to delay the agreement while producer is more eager to trade and is willing to accept a lower share of the surplus. As a result, in equilibrium the consumer receives a larger share of the surplus. The tighter the consumer's liquidity constraint, the lower the y and the stronger this effect. Gradual bargaining effectively makes the liquidity constraint binding in every round of negotiation and hence allows for the greatest advantage to the consumer.

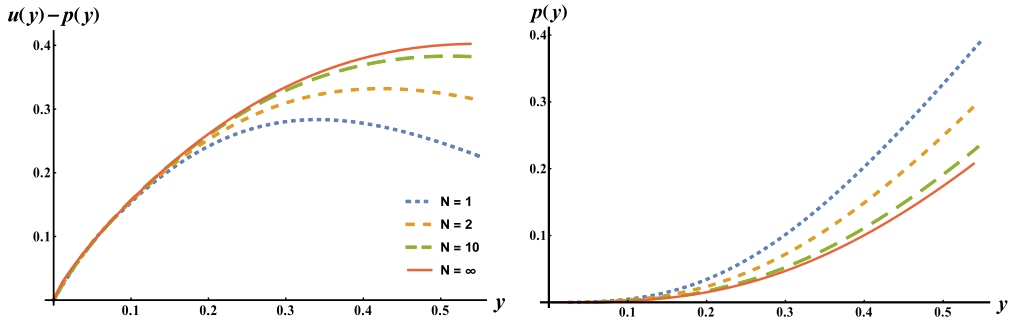

Figure 2.8: Consumer surplus and payment as a function of trade size for different N.

Figure 2.8 plots the final payment and the consumer's surplus as a function of trade size for games with $N \in \{1, 2, 10, +\infty\}$. The larger the number of rounds, the lower the payment, and the higher the consumer's surplus for any trade size, y. For finite values of N the consumer's surplus is non-monotone in y. Moreover, the value of y that maximizes the consumer's surplus increases with N. The monotonicity (or lack thereof) of the consumer's surplus will have important normative implications when we study the general equilibrium of the economy. To explore these implications, in what follows whenever we refer to N-round games we are using this repeated Rubinstein game.

Finally, it should be clear that if the consumer is better off when the negotiation takes place gradually, the opposite is true for the producer. Indeed, provided that the outcome of the negotiation is on the Pareto frontier, the consumer and the producer have opposite views on how to order the different outcomes. It means that the producer prefers the protocol in

which asset holdings are negotiated all at once. In Hu and Rocheteau (2020) we complement this result by showing that the producer would prefer to bargain gradually over the output, y, instead of bargaining gradually over z, or bargaining in a single round. In that sense, gradual bargaining always dominates a one-round negotiation provided that the right agenda is chosen.

2.4 Gradual bargaining in general equilibrium

Sections 2.2 and 2.3 provided the methodological tools to analyze OTC bargaining games with an agenda. The games we studied took as given the asset holdings that were up for negotiation (z) and omitted intertemporal considerations, such as the opportunity cost of holding assets across periods, that are critical for portfolio choices and allocations in decentralized markets. We now move to the general equilibrium analysis of decentralized asset markets and provide a user-friendly guide of bargaining solutions with an agenda in this context.

In terms of economic insights, we study the implications of gradualism to determine terms of trade for asset prices, allocations, and welfare. While Proposition 2.4 established that it is optimal for asset owners to sell their assets gradually, we will now demonstrate that gradual negotiations lead to allocations that are superior from a social welfare perspective. We will provide a stark example of an asset market based on a simplified version of Lagos and Zhang (2020) where Nash bargaining generates the worst possible allocation whereas gradual bargaining generates the first best.

2.4.1 General equilibrium setting

The environment is based on the workhorse model of monetary theory of Lagos and Wright (2005).[13] The population of agents is divided evenly between a unit measure of consumers and a unit measure of producers. There is an infinite (countable) number of periods, where each period is divided into two stages. The first stage is the decentralized market studied earlier where agents trade goods and assets in pairwise meetings formed at random. The measure of bilateral matches is $\alpha \in (0, 1]$. The second stage, labeled CM (for centralized market), features a centralized Walrasian market. It is in this second stage that agents choose their asset holdings, z, by taking prices and rates of return parametrically. There is one good in each stage and we take the CM good as numéraire. The timing within a representative period is illustrated in Figure 2.9.

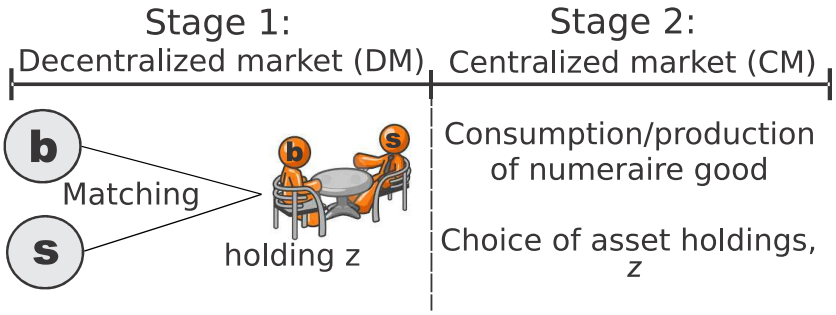

Figure 2.9: Timing of a representative period.

Consumers' preferences are represented by the period utility function, $u(y) - h$, where y is the DM good traded in pairwise meetings in stage 1 and h is the disutility of producing h units of numéraire in stage 2. Producers' preferences are represented by $-v(y)+c$, where c is the consumption of the numéraire in stage 2. Recall that y^* is the quantity that maximizes gains from trade in pairwise meetings, $u'(y^*) = v'(y^*)$. Note also that all agents' utilities are linear in the numéraire good, which is consistent with the quasi-linear payoffs of the

[13] We adopt the version with two distinct types of agents as in Lagos and Rocheteau (2005) and Rocheteau and Wright (2005). For various treatments of the New Monetarist model, see Rocheteau and Nosal (2017) and Lagos et al. (2017).

bargaining game in Section 2.2. All agents share the same discount factor across periods, $\beta \equiv (1+\rho)^{-1} \in (0,1)$.

Agents, who are anonymous, cannot issue private IOUs. This assumption creates a need for liquid assets. As in Lagos (2010) and Geromichalos et al. (2007) there is an exogenous measure A of long-lived Lucas trees that are perfectly durable, storable at no cost, and non-counterfeitable. All trees are identical and one unit of tree pays off $d \geq 0$ units of numéraire at the start of the CM. Fiat money is a special case where $d = 0$. For that special case we allow the supply of the asset to grow at a constant rate, π, through lump-sum transfers or taxes to either consumers or producers. We denote ϕ_t the competitive (ex dividend) price of Lucas trees in the CM in terms of the numéraire Hence, if an agent holds a units of Lucas trees at the beginning of a period, his asset holdings expressed in terms of the numéraire are $z = a(\phi_t + d)$. In pairwise meetings, agents bargain gradually according to the strategic game or axiomatic solution described in Section 2.2.3 where the consumer's bargaining power is $\theta \in [0,1]$.

In order to fix ideas, a preview of trade patterns in equilibrium is as follows. Consumers in pairwise meetings consume some endogenous quantity y in exchange for some endogenous quantity of assets. Producers in pairwise meetings produce y in exchange for assets. In the second stage, roles are reversed: consumers replenish their asset holdings by producing the numéraire good with their own labor while producers sell the assets received in the first stage in exchange for the numéraire good to consume.

2.4.2 Asset prices and welfare

We restrict our attention to stationary equilibria where the price of Lucas trees is constant at ϕ and hence their gross rate of return is also constant and equal to $R = 1 + r = (\phi + d)/\phi$. In the case of fiat money, $R = \phi_{t+1}/\phi_t$, is equal to the inverse of the gross growth rate of the

money supply, $1/(1+\pi)$.

Value functions The lifetime expected utility of a consumer (i.e., buyer of DM goods) with wealth z in the CM is

$$W^b(z) = \max_{z',h} \left\{ -h + \beta V^b(z') \right\} \quad \text{s.t.} \quad z' = R\left(z + h\right), \tag{2.28}$$

where z' are next-period asset holdings, and $V^b(z')$ is the value function at the start of the DM. From (2.28) the consumer chooses his production of numéraire and future asset holdings in order to maximize his discounted continuation value net of the disutility of production. According to the budget constraint, next-period asset holdings are equal to current asset holdings plus output from production, everything multiplied by the gross rate of return of assets. Substituting h by its expression coming from the budget identity into the objective, we obtain

$$W^b(z) = z + \max_{z' \geq 0} \left\{ -\frac{z'}{R} + \beta V^b(z') \right\}. \tag{2.29}$$

As is standard, W^b is linear in wealth. Hence, the payoff to a consumer who brought z units of trees in a pairwise meeting in the DM is $u^b = u(y) + W^b(z-p) = u(y) - p + u_0^b$ where $u_0^b = W^b(z)$, as specified in Section 2.2. There is a similar equation defining the value function of a producer (seller of the DM goods), $W^s(z)$.

Bargaining with an agenda The terms of trade in pairwise meetings are determined according to the gradual bargaining solution described in Sections 2.2 and 2.3. An intuitive and tractable way to solve this bargaining game in a general equilibrium model is as follows. Suppose an interim agreement, (y, p), has been reached where $y < y^*$ and the consumer adds an infinitesimal quantity ∂z of assets to the bargaining table. The outcome, $(\partial y, \partial p)$, for

this new round of negotiation is given by the generalized Nash solution, i.e.,

$$(\partial y, \partial p) \in \arg\max \left[u'(y)\partial y - \partial p\right]^{\theta} \left[\partial p - \upsilon'(y)\partial y\right]^{1-\theta} \quad \text{s.t.} \quad \partial p \leq \partial z. \tag{2.30}$$

Given that the consumer has already secured a consumption level y, the surplus from the agreement $(\partial y, \partial p)$ is $u'(y)\partial y - \partial p$ where the additional amount of consumption is valued at the marginal utility, $u'(y)$. Similarly, the cost to the seller to produce an additional ∂y is $\upsilon'(y)\partial y$ and hence his surplus is $\partial p - \upsilon'(y)\partial y$. Note that the total surplus is positive as long as $y < y^*$. Provided ∂z is small, the solution to (2.30) is such that $\partial p = \partial z$. So the bargaining problem is the same as the one in Shi (1995) and Trejos and Wright (1995) where agents bargain over the output in exchange for an indivisible unit of money, here ∂z, according to the Nash solution. The problem is even easier in that the surpluses are linear in ∂y. It also resembles the use of the generalized Nash solution in the Lagos and Wright (2005) model except that now the negotiation takes place at the margin. The first-order condition of the maximization problem in (2.30) with respect to ∂y gives

$$\frac{\partial y}{\partial z} = \frac{\theta u'(y) + (1-\theta)\upsilon'(y)}{u'(y)\upsilon'(y)}. \tag{2.31}$$

This solution coincides with (2.21) by substituting $\tau = p/\delta$. It gives the marginal value of real balances in terms of DM consumption. We can then compute the payment function, $p(y)$, by integrating $\partial p/\partial y = \partial z/\partial y$ over $[0, y]$ for all $y < y^*$, i.e.,

$$p(y) = \int_0^y \frac{u'(x)\upsilon'(x)}{\theta u'(x) + (1-\theta)\upsilon'(x)} dx, \ \ \forall y < y^*. \tag{2.32}$$

We denote $z^* = p(y^*)$ as the wealth required to purchase y^*. The total consumption of a buyer holding $z \leq z^*$ is then

$$y(z) = \int_0^z \frac{\theta u'(x) + (1-\theta)\upsilon'(x)}{u'(x)\upsilon'(x)} dx. \tag{2.33}$$

Given the payment and consumption functions, $p(y)$ and $y(z)$, we compute the lifetime expected utility of a consumer bringing z assets to the DM:

$$V^b(z) = \alpha \left\{ u\left[y(z)\right] + W^b\{z - p\left[y\left(z\right)\right]\}\right\} + (1-\alpha)\, W^b(z), \tag{2.34}$$

According to (2.34) a consumer meets a producer with probability α. The consumer enjoys y units of DM consumption in exchange for p units of assets. With probability $1-\alpha$ the consumer is unmatched and enters the CM with z units of asset.

Choice of asset holdings Substituting $V^b(z)$ with its expression given by (2.34), and using the linearity of $W^b(z)$, the consumer's choice of asset holdings solves

$$\max_{z\geq 0} \left\{ -sz + \alpha \left\{ u\left[y(z)\right] - p\left[y(z)\right]\right\}\right\}, \tag{2.35}$$

where s is the spread between the rate of time preference and the real rate on liquid Lucas trees,

$$s = \frac{\rho - r}{R} \geq 0. \tag{2.36}$$

We rewrite the portfolio problem, (2.35), as a choice of DM consumption, taking into account that the payment function, $p(y)$, is given by (2.22). It becomes:

$$\max_{y\in[0,y^*]} \left\{ -sp(y) + \alpha \int_0^y \frac{\theta u'(x)\left[u'(x) - \upsilon'(x)\right]}{\theta u'(x) + (1-\theta)\upsilon'(x)} dx \right\}. \tag{2.37}$$

Note that we can restrict the choice of y to $[0, y^*]$ since the second term is maximum when $y = y^*$. The objective function is continuous and strictly concave for all $y \in (0, y^*]$. The

first-order condition gives

$$s = \alpha\theta \left(\frac{u'(y)}{v'(y)} - 1 \right). \tag{2.38}$$

From (2.38) the interest rate spread has a simple expression as the product of three components: the search friction, α, the bargaining power, θ, and the marginal value of wealth in the DM, $u'(y)/v'(y) - 1$. Interestingly, the expression for the liquidity premium on the right side of (2.38) is much simpler than the one obtained from the Nash solution that involves the second derivatives of u and v and that is not necessarily monotone in y.

By market clearing,

$$p(y) \leq \left(\frac{1+\rho}{\rho - s} \right) Ad, \;\; " = " \;\; \text{if} \;\; s > 0, \tag{2.39}$$

where we have used that the cum-dividend price of the asset is $\phi + d = (1+\rho)d/(\rho - s)$. When $s > 0$, consumers hold exactly $p(y) = (\phi + d)A$. If $s = 0$, then from (2.38) $y = y^*$. The total supply of the asset, $(\phi + d)A$, is no less than $p(y^*)$ since assets can also be held as a pure store of value. An equilibrium can be reduced to a pair (s, y) that solves (2.38) and (2.39). We measure social welfare as the sum of surpluses in pairwise meetings, $\mathcal{W} = \alpha \left[u(y) - v(y) \right]$, but we do not include the output from Lucas trees, Ad.

Proposition 2.5. ***(Asset prices and welfare.)*** *An equilibrium exists and is unique.*

1. ***(Lucas trees,*** $d > 0$***).*** *If* $Ad \geq \rho p(y^*)/(1+\rho)$ *then* $s = 0$ *and* $y = y^*$ *in all matches. If* $Ad < \rho p(y^*)/(1+\rho)$ *then* $s > 0$ *and* $y < y^*$.

2. ***(Comparison to Nash.)*** *Suppose* $\theta < 1$. *The equilibrium under Nash bargaining never implements the first best, i.e.,* $y < y^*$ *for all* $A > 0$.

3. ***(Fiat money,*** $d = 0$***).*** *For all* $s > 0$, $y < y^*$. *As* s *approaches 0,* y *tends to* y^* *for all* $\theta \in (0, 1]$.

The first part of Proposition 2.5 shows that the first best output in pairwise meetings is achieved for all bargaining powers provided that the asset supply is sufficiently abundant. While intuitive, the second part of Proposition 2.5 shows that this result does not hold under Nash bargaining. If agents bargain all at once ($N = 1$ in the repeated Rubinstein game) according to Nash, then for all $\theta < 1$, the equilibrium never achieves first best irrespective of the supply of assets. The non-monotonicity of the Nash solution generates asset misallocation by preventing the market from clearing if all the asset supply is held by consumers. As a result, a fraction of A is held by producers even though they have no liquidity needs while consumers are liquidity-constrained. This result shows that gradual bargaining is not only desirable for asset owners to increase their surpluses (Proposition 2.4), it is also socially desirable to avoid the misallocation of assets.

The last part of Proposition 2.5 is a corollary of the first part in the case of fiat money. The spread s is now taken as a policy parameter. As is standard in monetary models, as long as $s > 0$ the output is inefficiently low. However, if $s = 0$, which corresponds to the Friedman rule, then the equilibrium implements the first best for all bargaining powers. Again, it is in sharp contrast with the inability of the Friedman rule to generate the first best under Nash bargaining (Lagos and Wright, 2005).[14]

Using the same calibrated parameter values as Lagos and Wright (2005), $u(y) = y^{0.61}/0.61$, $v(y) = y$, and $\theta = 0.343$, we compare the output traded at the Friedman rule by playing the game described in Section 2.3 for some arbitrary N relative to the first-best output, y^*, which is obtained at the limit when $N = +\infty$. Increasing N from 1 to 5 raises output in bilateral matches by about 39%, and increasing N to infinity raises it by 76%. If consumers divide their asset holdings into 5 bundles, they raise their surplus by 34%. Taking N to

[14]The gradual solution, $N = +\infty$, is not the only bargaining solution able to implement the first best at the Friedman rule when producers have some bargaining power. A case in point is the proportional solution proposed by Kalai (1977). See Aruoba et al. (2007). However, the Kalai solution is not scale invariant and does not have strategic foundations. Interestingly, in Hu and Rocheteau (2020) we show that in the context of quasi-linear environments the same extensive-form games we described earlier provide foundations for the proportional solution when the agenda consists in bargaining over output gradually.

infinity expands their surplus by 95%.

2.4.3 An OTC market with linear payoffs

In order to illustrate the last part of Proposition 2.5, we provide a stark example of an OTC market where Nash bargaining delivers the worst possible allocation while gradual bargaining delivers the first best. We adopt a specification with linear payoffs, similar to Lagos and Zhang (2020) and consider an endowment economy. At the beginning of each period sellers (previously labeled producers) are endowed with Ω units of DM goods interpreted as short-lived assets and have a linear technology to transform each unit of the DM good into $\varepsilon_\ell > 0$ units of numéraire. Buyers (previously labeled consumers) receive no endowment but can transform the DM good into $\varepsilon_h > \varepsilon_\ell$ units of numéraire. Hence, $u(y) = \varepsilon_h y$ and $v(y) = \varepsilon_\ell y$. Sellers choose the quantity of DM goods, $\omega \leq \Omega$, to bring into a bilateral match and consume the rest.[15] We set $d = 0$ so that purchases of DM goods are financed with fiat money. The spread s given by (2.36) is the difference between the rate of return of money, $r = -\pi/(1+\pi)$ where π is the money growth rate implemented through lump-sum transfers, and the rate of time preference. It can also be interpreted as a nominal interest rate on an illiquid bond.

Suppose first that agents negotiate according to Nash. The outcome in a match where the buyer holds z and the seller holds ω is given by:

$$\max_{y,p}(\varepsilon_h y - p)(p - \varepsilon_\ell y) \quad \text{s.t.} \quad p \leq z \quad \text{and } y \leq \omega. \tag{2.40}$$

If the liquidity constraint, $p \leq z$, does not bind, then the solution is $y = \omega$ and $p = (\varepsilon_h + \varepsilon_\ell)\omega/2$. Buyers purchase all the DM goods, which is socially efficient, and a payment

[15] The assumption according to which agents can choose to bring only a fraction of their asset holdings in a match was introduced by Berentsen and Rocheteau (2003), Lagos and Rocheteau (2008), and Lagos (2010). This assumption addresses the fact that under Nash bargaining agents might have incentives to hide some of their assets.

is made to divide the match surplus evenly. This trade is feasible if $(\varepsilon_h + \varepsilon_\ell)\omega/2 \leq z$. If $p \leq z$ binds then there are two cases to distinguish. If $(\varepsilon_h + \varepsilon_\ell)\, z \geq 2\varepsilon_h\varepsilon_\ell\omega$, then agents swap their inventories, $y = \omega$ and $p = z$. Otherwise, if $(\varepsilon_h + \varepsilon_\ell)\, z < 2\varepsilon_h\varepsilon_\ell\omega$, the buyer spends all his real balances, $p = z$, in order to purchase $y = (\varepsilon_h + \varepsilon_\ell)\, z/(2\varepsilon_\ell\varepsilon_h)$. The seller's surplus, $u^s(\omega, z) \equiv p(\omega, z) - \varepsilon_\ell y(\omega, z)$, is piecewise linear and non-monotone in ω. It reaches a maximum for $\omega = 2z/\left(\varepsilon_h + \varepsilon_\ell\right)$. Similarly, the buyer's surplus, $u^b(\omega, z) \equiv \varepsilon_h y(\omega, z) - p(\omega, z)$, is piecewise linear, non-monotone in z, and reaches a maximum when $z = 2\varepsilon_h\varepsilon_\ell\omega/(\varepsilon_h + \varepsilon_\ell)$.

Suppose, alternatively, that agents bargain gradually over real balances. The outcome of the negotiation for the marginal unit of money is

$$\max_{\partial y, \partial p} (\varepsilon_h \partial y - \partial p)(\partial p - \varepsilon_\ell \partial y) \quad \text{s.t.} \quad \partial p \leq \partial z \text{ and } \partial y < \omega - y. \tag{2.41}$$

The first feasibility condition, $\partial p \leq \partial z$, states that the buyer cannot spend more than the ∂z units of real balances that have been added to the negotiation table. The second feasibility condition, $\partial y < \omega - y$, states that the seller cannot deliver more than his remaining inventories. As long as $\omega > y$ and provided that ∂z is infinitesimal, the constraint $\partial p \leq \partial z$ binds and the solution of (2.41) takes the form

$$\partial y = \left(\frac{\varepsilon_h + \varepsilon_\ell}{2\varepsilon_\ell\varepsilon_h}\right) \partial z.$$

It follows immediately that the change in the buyer's surplus is $u^{b\prime}(z) = \varepsilon_h \partial y/\partial z - 1 = (\varepsilon_h - \varepsilon_\ell)/(2\varepsilon_\ell)$ if $y \leq \omega$ does not bind. The buyer's surplus is monotone increasing in his real balances. By integrating $\partial z/\partial y$ we obtain the following linear payment function,

$$p(y) = \frac{2\varepsilon_h\varepsilon_\ell}{\varepsilon_h + \varepsilon_\ell} y. \tag{2.42}$$

It follows that the seller's surplus from selling ω units of DM goods, assuming the buyer has enough real balances to do so, is $p(\omega) - \varepsilon_\ell\omega = \varepsilon_\ell\left(\varepsilon_h - \varepsilon_\ell\right)\omega/\left(\varepsilon_h + \varepsilon_\ell\right)$, which is monotone

increasing in ω.

Figure 2.10 provides a graphical representation of the two solutions. Under linear preferences, the Pareto frontiers are piecewise-linear with a kink. The part of each frontier with unit slope in absolute value corresponds to outcomes where $\omega = \Omega$ and the liquidity constraint does not bind. In the flatter portion, the liquidity constraint binds. In Figure 2.10 the outcome of the Nash solution is such that $\omega = \Omega$ and $p \leq z$. If the negotiation takes place gradually, the interim outcomes are located on the flatter part of the Pareto frontiers until the upper frontier is reached. By the same reasoning as in Section 2.3, the buyer's surplus is larger under gradual bargaining because the binding liquidity constraint means that the buyer is effectively more patient in each round of the negotiation, thereby shifting the bargaining share in her favor.

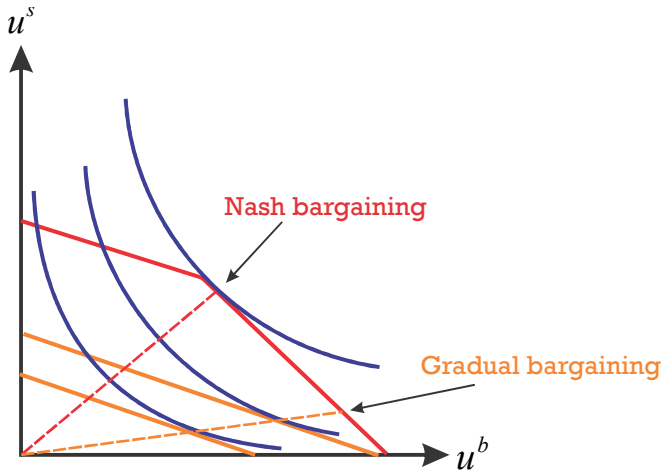

Figure 2.10: Nash versus gradual bargaining under linear preferences.

In the CM, the seller chooses $\omega \leq \Omega$ to maximize $\int u^s(\omega, z) dF^b(z)$, where $F^b(z)$ is the distribution of real balances across buyers. The buyer's problem consists in choosing z in order to maximize $-sz + \alpha \int u^b(z, \omega) dF^s(\omega)$ where $F^s(\omega)$ is the distribution of inventories held by sellers in DM matches. We characterize equilibrium allocations in the following proposition.

Proposition 2.6. ***(Allocations in OTC markets under liquidity constraints.)*** *Sup-*

pose sellers are endowed with Ω units of DM goods and preferences are given by $u(y) = \varepsilon_h y$ and $v(y) = \varepsilon_\ell y$. The liquid asset takes the form of fiat money, $d = 0$

1. ***Nash bargaining**. For all $s \geq 0$, there exists no monetary equilibrium and the OTC market is inactive.*

2. ***Gradual bargaining**. If $s \leq \frac{\alpha(\varepsilon_h - \varepsilon_\ell)}{2\varepsilon_\ell}$ then there exists a monetary equilibrium implementing the first best.*

Proposition 2.6 provides a stark illustration of the importance of the agendas of the bilateral negotiations in OTC markets. If agents bargain according to Nash, then the OTC market is inactive and money is not valued for all $s \geq 0$, even at the Friedman rule. All DM goods are held by the least productive agents, which corresponds to the worst allocation.[16] We represent the seller's and buyer's best-response functions, ω^{BR} and z^{BR}, for symmetric equilibria in the left panel of Figure 2.11. The only intersection is when $z = \omega = 0$. If sellers bring Ω in the match, then buyers bring at most $z = 2\varepsilon_h\varepsilon_\ell\Omega/(\varepsilon_h + \varepsilon_\ell)$ real balances in order to maximize their surplus. But if sellers anticipate this amount of real balances, they will bring at most $\omega = 4\varepsilon_h\varepsilon_\ell\Omega/(\varepsilon_h + \varepsilon_\ell)^2 < \Omega$. And so on. The process unravels until neither the buyer nor the seller brings anything to trade.

In contrast, if agents bargain gradually, then the first-best trades are implemented in all matches provided that s is not too high. We represent the best-response correspondences under gradual bargaining and assuming symmetry across agents in the right panel of Figure 2.11. Note that sellers bring at the minimum the amount of DM goods corresponding to what buyers can pay for and they can bring up to their full endowment Ω (i.e., their best response is an interval). For low spreads, there exists a Nash equilibrium where $\omega = \Omega$ and $z = 2\varepsilon_h\varepsilon_\ell\Omega/(\varepsilon_h + \varepsilon_\ell)$. The OTC market is active and it achieves the first best where in all

[16] This result does not rely on preferences being linear and is robust to various alternative assumptions. See Lebeau (2020) for details.

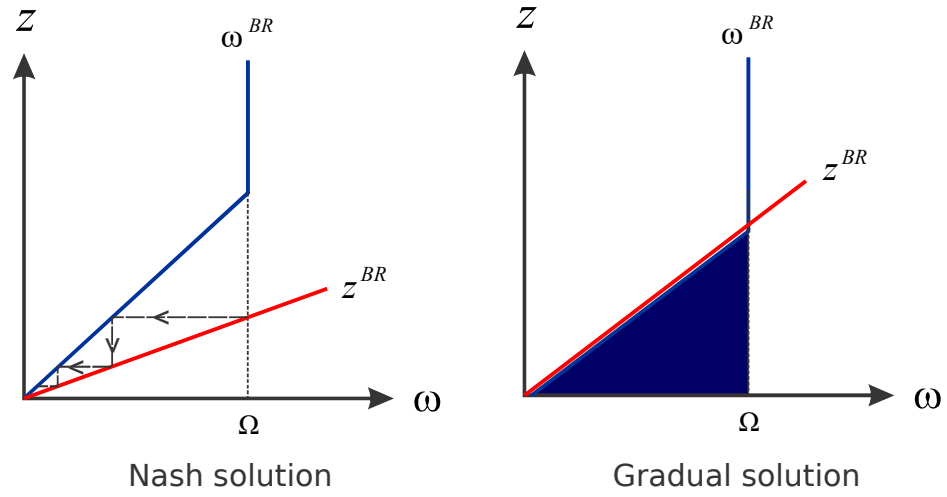

Figure 2.11: Symmetric best-response correspondences under Nash (left) and gradual (right) bargaining

matches sellers transfer all their endowments of DM goods to buyers. The unraveling that occurs under the Nash solution is avoided precisely because agents' surpluses are monotone increasing in the goods or assets they bring in a match.

2.5 Gradual bargaining with multiple assets

So far, we assumed that there is a single asset. We now relax this assumption and introduce multiple assets. Following Zhu and Wallace (2007), we explain rate-of-return and liquidity differences among assets from the bargaining protocol. Besides the bargaining protocol, we also need some fundamental features to distinguish assets. The feature we exploit is that it takes time to negotiate assets sequentially. In order to make this time dimension relevant, we will assume that the duration of the negotiation is stochastic and exponentially distributed. In contrast to Zhu and Wallace (2007), we do not impose an arbitrary order according to which assets are negotiated and we do not change bargaining powers across stages of the negotiation. The fundamental difference between assets will be their negotiability, δ, which is the amount of asset that can be negotiated per unit of time.

There are now J types of Lucas trees indexed by $j \in \{1, ..., J\}$. For simplicity, we assume that the Lucas trees fully depreciate in one period for all types, and that each Lucas tree born in $t-1$ pays off one unit of numéraire in the CM of t, i.e., $d = 1$. The supply of type-j Lucas trees is fixed at A_j and the new Lucas trees are received by consumers in a lump-sum fashion at the beginning of each CM. The CM price of Lucas tree j is ϕ_j, their gross real rate of return is $R_j = 1 + r_j = 1/\phi_j$, and the interest-rate spread relative to an illiquid asset is

$$s_j = \frac{\rho - r_j}{R_j}. \tag{2.43}$$

We index fiat money by $j = 0$, and hence for asset 0, $d = 0$. It is the only long-lived asset with gross real rate of return equal to $R_0 = 1/(1+\pi)$. The spread, $s_0 = i = (1+\rho)(1+\pi)-1$, is interpreted as the nominal interest rate of an illiquid asset.

In order to differentiate these assets, we take seriously the notion that it takes time to negotiate the sale of a portfolio of assets gradually over time. This notion is embedded into the concept of agenda according to which different items are negotiated sequentially. We take this sequential negotiation as a primitive, i.e., a technological constraint imposed on the negotiation. More specifically, over a small time interval of length $\Delta > 0$, agents can negotiate the sale of $\delta_j \Delta$ units of asset j (expressed in terms of numéraire), where $\delta_j > 0$ is a measure of the speed of the negotiation that captures the process of negotiating, authenticating assets, and transferring asset ownership (e.g., physical transfer, a ledger, a blockchain technology).[17] We focus on the case where N is large and Δ is small. Note that here we do not think of N as a choice variable. Instead, we take gradual bargaining as a physical constraint on how assets can be traded. However, the asset owner can choose the order according to which different assets are negotiated.

[17] In that regard, our theory complies with the Wallace et al. (1998) dictum in that it specifies assets by how their physical properties determine the technology to transfer their ownership, which permits the assets' role in exchange to be endogenous.

We rank assets according to their negotiability, $\delta_0 \geq \delta_1 \geq \delta_2 \geq ... \geq \delta_J$. We assume that fiat money is the most negotiable asset because it is a tangible object whose ownership is asserted by simply carrying it and it can be authenticated with relatively small effort. It takes more time to transfer and verify the ownership of non-tangible assets (e.g., crypto-currencies), making them less negotiable. Complex financial securities take even more time to be authenticated and evaluated. In Figure 2.12 we provide some evidence based on Pagnotta and Philippon (2018) and O'Keeffe (2018) that transaction times vary for different classes of assets.[18]

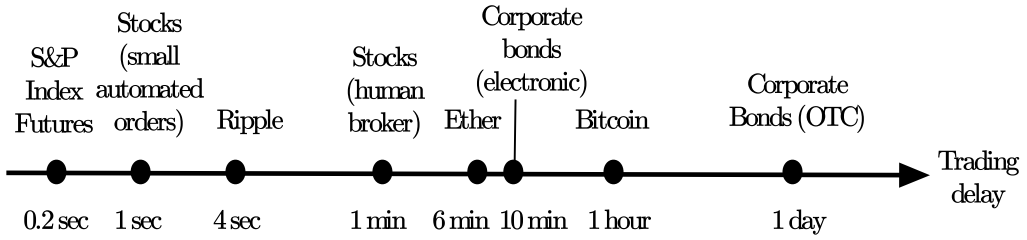

Figure 2.12: Trading delays by asset classes. Sources: Pagnotta and Philippon (2018), O'Keeffe (2018).

Without any additional assumption, δ_j is irrelevant for the final outcome of the negotiation. In order to make time relevant, we assume that the total amount of time allocated to the negotiation, $\bar{\tau}$, is a random variable exponentially distributed with mean $1/\lambda$ and realized at the beginning of a match. The assumption of a random duration of the negotiation is commonly used in models with alternating offers (e.g., Binmore et al., 1986). The consumer's bargaining power is θ.[19]

We let consumers choose the order according to which assets are sold (after $\bar{\tau}$ has been realized). The cumulative amount of asset of type j that has been up for negotiation at time

[18] As mentioned earlier, it is hard to disentangle the different sources of delays in asset transactions (see, e.g., Duffie (2011)) but there is strong evidence that those delays vary across assets. In our model, we keep search frictions the same across assets and attribute all the differences to the negotiation process and the time to transfer ownership.

[19] One could allow θ to be a function of τ, which would not affect our results qualitatively. One could also assume that θ varies with the type of asset that is currently under negotiation. Such extension would allow our theory to encompass the explanations for rate-of-return differences across assets by Zhu and Wallace (2007) and Rocheteau and Nosal (2017).

τ is denoted $\omega_j(\tau)$ and $\omega(\tau) = \sum_{j=0}^{J} \omega_j(\tau)$ is the value of the asset portfolio that has been negotiated up to τ. It obeys the following law of motion:

$$\omega_j'(\tau) = \delta_j \sigma_j(\tau) \quad \text{for all } j \in \{0, 1, ..., J\}, \tag{2.44}$$

where $\sigma_j(\tau) \in [0, 1]$ is the fraction of time devoted to the sale of asset j at time τ and $\sum_{j=0}^{J} \sigma_j(\tau) = 1$. So, the amount of asset j added to the negotiation table at time τ is the product of the negotiability of asset j, δ_j, and the fraction of time that the consumer dedicates to the negotiation of asset j, σ_j. Moreover, feasibility implies $\sigma_j(\tau) \in [0, 1]$ if $\omega_j(\tau) < a_j$ and $\sigma_j(\tau) = 0$ otherwise. In words, an agent can add asset j on the negotiating table at time τ only if he has not sold all his holdings of asset j prior to τ. Replacing δ by ω' in (2.21), the change in the consumer's consumption over time is

$$y'(\tau) = \frac{\theta u'(y) + (1 - \theta) v'(y)}{u'(y) v'(y)} \omega'(\tau), \tag{2.45}$$

if $y(\tau) < y^*$ and $y'(\tau) = 0$ otherwise. The left side is the output purchased over an infinitesimal amount of time. The right side is composed of two terms: the amount of output that a marginal unit of wealth buys times the amount of wealth that can be negotiated over a small time interval.

The surplus of a consumer in a DM match with portfolio $\mathbf{a} = [a_j]_{j=0}^{J}$, agenda $\sigma = [\sigma_j]_{j=0}^{J}$, and time to negotiate $\bar{\tau}$ is:

$$S(\mathbf{a}, \sigma, \bar{\tau}) = \theta \int_0^{\bar{\tau}} \ell\left[y(\tau)\right] \omega'(\tau) d\tau = \theta \int_0^{\omega(\bar{\tau})} \ell\left[p^{-1}(\omega)\right] d\omega, \tag{2.46}$$

where $\ell(y) \equiv u'(y)/v'(y) - 1$ is the marginal surplus, and $p^{-1}(\omega) = y^*$ whenever $\omega > p(y^*)$. Over a small time interval of length $d\tau$ the consumer sells $\omega'(\tau)$ units of assets where each unit generates a marginal surplus to the consumer equal to $\theta \ell(y)$. The right side of (2.46) is obtained by adopting the change of variable $\omega = \omega(\tau)$. It follows that the consumer surplus

depends on the agenda σ only through the amount of assets that can be negotiated up to $\bar{\tau}$, $\omega(\bar{\tau})$. The right side of (2.46) has a simple interpretation. The consumer receives a fraction θ of the sum of the marginal surpluses, $\ell(y)$, negotiated over the time interval $[0, \bar{\tau}]$. From (2.44),

$$\omega(\bar{\tau}) = \int_0^{\bar{\tau}} \sum_{j=0}^{J} \delta_j \sigma_j(\tau) d\tau.$$

The total wealth negotiated over $[0, \bar{\tau}]$ is the sum over all asset types and all infinitesimal time intervals of the marginal quantities of asset added to the negotiation table. In order to characterize the optimal strategy to maximize $\omega(\bar{\tau})$ we denote $T_0 = 0$ and

$$T_j(\mathbf{a}) = \sum_{k=0}^{j-1} \frac{a_k}{\delta_k} \text{ for all } j \in \{1, 2, ..., J+1\}. \tag{2.47}$$

That is, T_j is the time that it takes to sell the first $j-1$ most negotiable assets.

Lemma 2.2. ***(Pecking order)*** *For any portfolio* $\mathbf{a}$ *and any realization of* $\bar{\tau}$*, the optimal choice* $\sigma^* = [\sigma_j^*]$ *is given by*

$$\sigma_j^*(\tau) = \begin{cases} 1 & \text{if } T_j < \tau \leq T_{j+1} \\ 0 & \text{otherwise} \end{cases}.$$

Lemma 2.2 shows that it is optimal to adopt a pecking order to sell assets.[20] Consumers start paying with money. When their money holdings are exhausted, they start selling asset 1, etc. Hence, in a fraction $1 - e^{-\lambda T_1}$ of matches only money is used to finance consumption, where T_1 is endogenous. In a fraction $e^{-\lambda T_1} - e^{-\lambda T_2}$ of matches both money and type-1 Lucas trees serve as means of payments. And so on. Given this pecking order, the expected

[20]For a pecking-order theory of payments based on informational asymmetries between consumers and producers, see Rocheteau (2011).

maximized surplus of the consumer is:

$$S(\mathbf{a}) = \int_0^{+\infty} \lambda e^{-\lambda\tau} S(\mathbf{a}, \sigma^*, \tau) d\tau = \theta \sum_{j=0}^{J} \delta_j \int_{T_j}^{T_{j+1}} e^{-\lambda\tau} \ell[y(\tau)] d\tau. \tag{2.48}$$

Over the time interval $[T_j, T_{j+1}]$ agents negotiate asset j where the speed of the negotiation is given by δ_j.

We now turn to the asset pricing implications of this pecking order. The portfolio problem in the CM is given by

$$\max_{\mathbf{a}\geq\mathbf{0}} \{-\mathbf{sa} + \alpha S(\mathbf{a})\}, \tag{2.49}$$

where $\mathbf{s} = [s_j]$ is the vector of asset spreads. According to (2.49) the consumer maximizes his expected DM surplus net of the costs of holding assets as measured by the spreads $[s_j]$. The FOCs of the maximization problem (2.49) are:

$$s_j = \alpha \frac{\partial S(\mathbf{a})}{\partial a_j}. \tag{2.50}$$

The left side of (2.50) is the opportunity cost of holding asset j. The right side is the probability α that the consumer receives an opportunity to spend, α, multiplied by the marginal liquidity value from holding asset j. The expression of this last term is given in the following lemma.

Lemma 2.3. *The marginal value of asset j to a consumer with portfolio* $\mathbf{a}$ *is*

$$\frac{\partial S(\mathbf{a})}{\partial a_j} = \overbrace{\theta\lambda \sum_{k=j+1}^{J} \int_{T_k}^{T_{k+1}} \frac{(\delta_j - \delta_k)}{\delta_j} e^{-\lambda\tau} \ell[y(\tau)] d\tau}^{\textit{negotiability value}} + \overbrace{\theta e^{-\lambda T_{J+1}} \ell[y(T_{J+1})]}^{\textit{liquidity value}}. \tag{2.51}$$

From (2.51), holding an additional unit of a_j has two benefits to the consumer. First, there

is a liquidity benefit according to which the consumer has more wealth, which relaxes his liquidity constraint and allows him to consume more if the negotiation is not terminated before the whole portfolio has been sold. This effect is captured by the last term on the right side and is analogous to (2.38). Second, there is a negotiability benefit according to which asset j speeds up the negotiation relative to less negotiable assets of types $k > j$. This first term on the right side of (2.51) is asset specific, as it depends on δ_j.

By market clearing $a_j = A_j$ for all $j \geq 1$. Hence, an equilibrium can be reduced to a list $\left\langle a_0, \{s_j\}_{j=1}^J \right\rangle$ that solves (2.50). In the following proposition we measure the liquidity of an asset by its velocity or turnover defined as

$$\mathcal{V}_j \equiv \frac{\alpha \int_0^{+\infty} \lambda e^{-\lambda x} \int_0^x \omega_j'(\tau) 1_{\{\omega(\tau) < p(y^*)\}} d\tau dx}{A_j}. \tag{2.52}$$

The numerator corresponds to the aggregate quantity of asset j sold in pairwise meetings while the denominator is the supply of the asset.

In order to fix ideas, suppose there is a single asset, fiat money. From (2.50) and (2.51), y solves

$$i = \alpha\theta e^{-\lambda \frac{p(y)}{\delta_0}} \ell(y). \tag{2.53}$$

From (2.53) the nominal interest rate is equal to the product of four components: the search friction, α, the bargaining power, θ, the negotiability friction, $e^{-\frac{\lambda}{\delta_0} p(y)}$, and the marginal value of wealth in the DM, $\ell(y)$. So, bargaining frictions affect the liquidity services of money through two channels: traders' bargaining powers and the time to negotiate real balances. The negotiability term is akin to a pledgeability coefficient but it is endogenous and depends on the time it takes to negotiate assets, the stochastic time horizon of the negotiation, and the bargaining protocol as represented by $p(y)$. From (2.52) the velocity of

money is

$$\mathcal{V}_0 = \frac{\alpha\delta_0\left[1 - e^{-\lambda\frac{p(y)}{\delta_0}}\right]}{\lambda p(y)}.$$

As the negotiability of money tends to infinity, its velocity approaches α.

The next proposition studies the implications of our model for asset prices and liquidity in the general case with J assets.

Proposition 2.7. ***(The negotiability structure of asset yields.)*** *For all $\{A_j\}_{j=1}^J$, if $\delta_0 > \delta_1$ then there is a $\bar{\iota} > 0$ such that for all $i < \bar{\iota}$ there exists a unique steady-state monetary equilibrium with aggregate real balances $A_0(i) > 0$. Let $\Omega_1 = A_0(i)$ and for each $j = 2, .., J$, let $\Omega_j = A_0(i) + \sum_{k=1}^{j-1} A_k$.*

1. *If $\Omega_{j+1} < p(y^*)$ and $\delta_j > \delta_{j+1}$, then $s_j > s_{j+1}$. If $\Omega_{j+1} \geq p(y^*)$, then $s_{j+k} = 0$ for all $k \geq 0$.*

2. *If $\delta_j > \delta_{j+1}$ and $p(y^*) > \Omega_j$, then $\mathcal{V}_j > \mathcal{V}_{j+1}$. If $p(y^*) \leq \Omega_j$ then $\mathcal{V}_j = 0$.*

3. *As λ approaches 0, $|s_j - s_{j'}|$ approaches 0 for all $j, j' \in \{0, ..., J\}$. Asset velocity, $\mathcal{V}_j$, approaches α for all j such that $\Omega_j \leq p(y^*)$, 0 for all j such that $\Omega_j \geq p(y^*)$, and $\alpha\left[p(y^*) - \Omega_j\right]/A_j$ for j such that $p(y^*) \in (\Omega_j, \Omega_{j+1})$.*

Proposition 2.7 has several implications. First, fiat money is valued for low i irrespective of the supply of Lucas trees. Even if the capitalization of all Lucas trees, $\sum_{k=1}^J A_k$, is larger than liquidity needs, $p(y^*)$, money is useful because it can be negotiated faster, thereby allowing agents to finance a larger consumption when $\bar{\tau}$ is low.

Second, even though all Lucas trees yield identical dividend streams, the equilibrium features rate-of-return differences across assets. Provided that asset supplies are not too large, assets

with a high negotiability command a lower rate of return than assets with a low negotiability, i.e., $r_j < r_{j+1}$ if $\delta_j > \delta_{j+1}$. The key components of our theory is that negotiation takes time as assets are sold gradually, and not all assets can be sold at equal speed due to technological differences to authenticate and transfer assets. Part 2 of Proposition 2.7 shows that assets that are more negotiable have a higher velocity, which is a consequence of the endogenous pecking order. As a result, there is a positive correlation between velocity and asset prices.

Finally, Part 3 of Proposition 2.7 considers the limit when the expected time horizon of the negotiation becomes arbitrarily large. If the risk that the negotiation ends before the portfolio of assets has been sold goes to zero, then the rates of return of all assets converge to the same value, i.e., there is rate-of-return equality. In that case the negotiability of assets, and the order according to which they are negotiated, does not affect their rates of return. The order in which assets are sold, however, matters for velocities. Indeed, only a fraction of assets are used for transactions and those assets have a maximum velocity equal to α.

In our working paper (Rocheteau et al., 2018), we consider two applications of our model for dual asset economies. In the first application, we interpret asset 1 as a short-term government bond and study its coexistence with fiat money and the implications for open market operations. If the time horizon of the negotiation, $\bar{\tau}$, is deterministic, there is a monetary equilibrium with $T_2 = \bar{\tau}$ provided that the supply of bonds, A_1, is not too low or too high and $\bar{\tau}$ is in some intermediate range. Output and the interest rate spread are determined recursively according to:

$$y = p^{-1}\left[\delta_0\bar{\tau} - \left(\frac{\delta_0 - \delta_1}{\delta_1}\right)A_1\right] \tag{2.54}$$

$$s_1 = \frac{\delta_0}{\delta_1}i - \alpha\theta\left(\frac{\delta_0 - \delta_1}{\delta_1}\right)\ell(y). \tag{2.55}$$

Relative to the existing literature, our model generates a new effect captured by the presence

of A_1 in the right side of (2.54), which we call negotiability effect, according to which a reduction in the supply of bonds, A_1, reduces output and spreads. This effect requires $\delta_0 > \delta_1$. An open market sale of bonds decreases output by crowding out a highly negotiable asset, money, with a less negotiable asset, bonds. This effect, we believe, captures the common wisdom regarding the transmission of open market operations on output. When $\bar{\tau}$ is stochastic, we distinguish this negotiability effect from the standard liquidity effect in New Monetarist models according to which an increase in A_1 raises the aggregate liquidity of the economy and hence output. The negotiability effect dominates for realizations of $\bar{\tau}$ in some intermediate range while the liquidity effect dominates for large realizations of $\bar{\tau}$.

The second application studies a dual currency economy. The supply of currency 0 (e.g., the domestic currency) grows at rate π_0 and the supply of currency 1 (e.g., the foreign currency) grows at rate π_1. Currency 0 is easier to authenticate and can be transferred faster than currency 1, i.e., $\delta_0 > \delta_1$. In the context of cryptocurrencies, currency 0 has lower confirmation times than currency 1. However, the supply of currency 0 grows faster than the supply of currency 1, $\pi_0 > \pi_1$. If $\bar{\tau}$ is in some intermediate range, there exists a unique steady-state equilibrium where both currencies 0 and 1 are valued and output solves

$$\frac{i_0\delta_0 - i_1\delta_1}{\delta_0 - \delta_1} = \alpha\theta\ell(y). \tag{2.56}$$

Inflation rates affect output according to $\partial y/\partial\pi_0 < 0$ and $\partial y/\partial\pi_1 > 0$. Moreover, currency 0 appreciates vis-a-vis currency 1 as α or θ increases or as $\bar{\tau}$ decreases. If the inflation rate of the most negotiable currency increases, then output decreases, in accordance with textbook comparative statics. However, as π_1 increases, agents find it optimal to reduce their holdings of currency 1 and raise their holdings of currency 0. As a result, they can buy more output over the time horizon $\bar{\tau}$. In the context of a dollarization equilibrium this would mean that an increase of the inflation rate of the foreign currency raises output by reverting the dollarization process.

2.6 Conclusion

The objective of this paper was to introduce a new approach to bargaining into models of decentralized asset markets. More than a new solution, we advocate for a new definition of the bargaining problem for negotiations over unrestricted asset portfolios. This new definition is a natural extension of existing bargaining theories (e.g., Osborne and Rubinstein, 1990) for a new class of models of decentralized markets with richer asset holdings. It includes as a primitive the agenda of the negotiation, i.e., a partition of the portfolio into asset bundles to be sold sequentially.

Our approach complies with the Nash program: it has (multiple) strategic foundations, in the form of alternating-offer games, and axiomatic foundations. It encompasses existing bargaining solutions, such as Nash, for specific agendas. We showed through several examples that the choice of the agenda is crucial for allocations and welfare. For instance, the choice of the agenda can have dramatic implications for the functioning of OTC markets with outcomes varying from a complete break-down to the implementation of first-best trades.

In our working paper, we provide many additional results and applications. In the companion paper of Hu and Rocheteau (2020) we show that the proportional solution of Kalai (1977) can be interpreted as a gradual solution for the agenda that consists in bargaining gradually over the output. This result is significant because it shows that one can provide strategic foundations for the Kalai solution in quasi-linear environments commonly used in search-theoretic models. In addition, while Kalai (1977) does not impose the scale invariance axiom of Nash (1950), the gradual solution is both scale invariant and ordinal. As another extension we endogenized the time it takes to negotiate assets through some costly investment before the negotiation starts. Much more can be done with this novel approach to bargaining in decentralized asset markets.

Chapter 3

Do Financial Frictions Shift the Beveridge curve? Theory and Evidence

3.1 Introduction

The United States has witnessed a marked outwards shift of its empirical Beveridge curve, a well-established negative relation between the unemployment rate and the vacancy rate, starting in December 2008 and up to the end of 2016. Depicted in Figure 3.1, such a shift means that a higher vacancy rate is required to sustain any given unemployment rate. Several such instances have occurred in the past fifty years. What makes this one occurrence of particular interest, however, is its cause. Elsby et al. (2015) demonstrate that while an increase in the flow into unemployment had always been the force driving these phenomena, this recent shift has been triggered by a decrease in the flow out of unemployment. Conditional on market tightness, it has become more difficult for jobseekers to be hired, and for

firms to fill their vacancies. In other words, the efficiency with which the labor market is able to produce jobs seems to have been impaired.

The Mortensen Pissarides (MP) model, which has become the canonical framework to explain the frictions that give rise to unemployment, provides us with a good theory of the Beveridge curve. However, it does not give as compelling a theory to account for shifts of the curve. In its most basic form, the model predicts a job-finding rate thirty percent higher than that realized up to four years into the recovery. To reconcile theory with empirics, one would need to assume an exogenous decrease in the efficiency parameter of the matching function, a device intended to proxy for the complex processes underlying the pairing of job seekers with vacant positions. Pissarides (2000) warns the readers of his textbook that "if [such empirical] changes are frequent, the usefulness of the matching function is reduced." He then develops a theoretical extension that could provide endogenous Beveridge curve shifts: mismatch. However, empirical studies show that this channel could at most account up to thirty percent of the shift.[1]

In this paper, I explore whether credit availability could have contributed to shifting the Beveridge curve. Indeed, one of the peculiar features of the Great Recession was the turmoil undergone by the financial and banking systems, leading to a sharp decline in credit provision. For example, in the first quarter of 2009, 64% of loan officers reported tightening credit standards, whereas none of them had reported so in the first quarter of 2007.[2] As shown in Figure 3.2, this sharp tightening was not followed by a clear easing of the lending standards after the Recession. One can infer that the standards remained relatively high throughout the recovery and up to now. The labor market impact of such turbulence in the credit market, both during and after the Great Recession, has been demonstrated by several recent contributions. For example, Chodorow-Reich (2014) shows that "employment at precrisis clients of lenders at the 10th percentile of bank health fell by roughly 4 to 5 percentage

[1] See Şahin et al. (2014) for example.
[2] FRB: Senior Loan Officer Opinion Survey

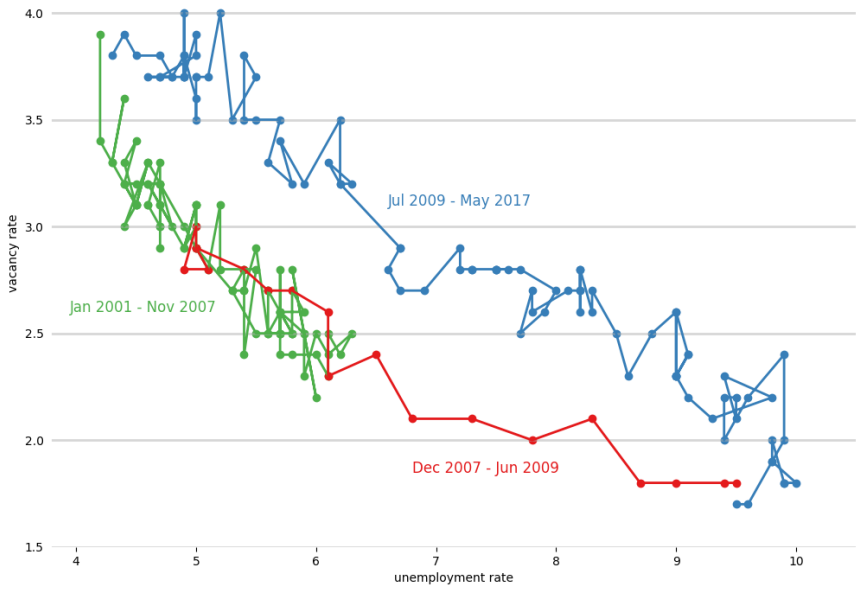

Figure 3.1: Beveridge curve in the US, January 2001 to May 2017, JOLTS.

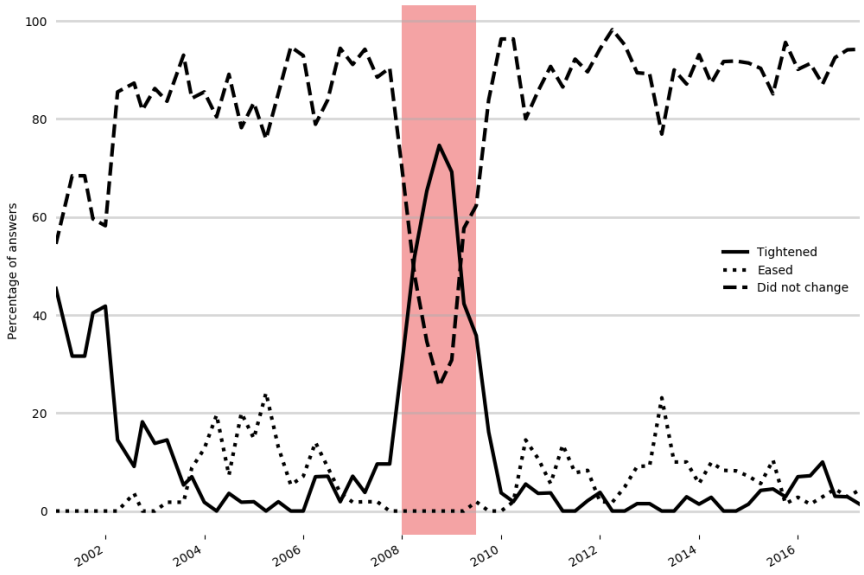

Figure 3.2: Lending standards for small firms, Quarter 1 2001 to Quarter 2 2017. Plotted is the percentage of each response to "Over the past three months, how have your bank's credit standards for approving applications for C&I loans to small firms changed?", FRB Senior Loan Officer opinion survey.

points more than at clients of lenders at the 90th percentile" and that "the credit channel can explain between one-third and one-half of the employment decline at small and medium-sized firms in the year following Lehman." It seems natural to ask whether the marked deterioration of credit conditions could have made it harder for firms to hire, independently of their willingness to do so; and for that matter, contributed to the shift of the Beveridge curve.

While it is difficult to directly observe instances of firms failing to conclude an employment contract because of their inability to finance the hire, there exists empirical evidence of firms having to forego investment opportunities due to the lack of funding. For example, Campello et al. (2011) use data from the CFO survey to document that in the fourth quarter of 2008, difficulties in raising external finance caused 86% of credit-constrained US firms to bypass interesting investment projects, and 56% of them to cancel planned investment projects. Although I exclusively focus on hires, interpreted as one particular kind of investment project, this is exactly the channel I try to capture in this paper. A simple departure from Wasmer and Weil (2004) (WW thereafter) allows me to implement this idea. Instead of looking for funds to finance their vacancy search costs, firms need loans to finance a fixed hiring cost paid after they have matched with a worker. I motivate this timing, and highlight the paper's contributions in relation to the existing literature on the shift of the Beveridge curve and on the interaction between labor and credit markets in section 3.2.

Section 3.3 lays out a baseline model to expose as simply as possible the impact of post-match credit frictions on the Beveridge curve. In that section, I abstract away from WW and take the financing process as exogenous. Firms' loan applications are accepted with a fixed probability and, as expected, a decrease in the aggregate loan-approval rate decreases equilibrium labor market tightness. This induces a move along the Beveridge curve towards a higher unemployment rate. More interestingly, the loan-acceptance rate also acts as a Beveridge curve shifter, such that a decrease in credit availability triggers an outwards shift of

the curve. This second effect is confounded with that of the aggregate efficiency parameter of the matching function. Section 3.4 endogenizes the loan-approval rate in the spirit of Wasmer and Weil (2004), with the twist described earlier. In equilibrium, both a productivity shock and an increase in the fixed cost of hiring decrease banks' entry, and the loan-finding rate declines. This results in the Beveridge curve dynamics described in the baseline model, which is not the case in the standard WW model. These theoretical results were obtained taking wage determination exogenously. To check their robustness, I implement wage bargaining into the model in section 3.5. The negotiations between workers and firms and between firms and banks occur sequentially, and I study the two possible orders: wage negotiated first, or loan negotiated first. In section 3.6, the baseline model is used to assess the extent to which the observed variation in credit availability in the US could account for the Beveridge curve shift following December 2008. To do so, I build an index that intends to capture the aggregate loan-finding probability of US firms from 1986 to 2017 and allows me to carry out a counterfactual exercise: how would the vacancy rate have evolved after the recession, had access to credit remained constant? I show that access to credit explains 7% of the Beveridge curve shift when using the most restrictive series, and 69% when using the least restrictive one. While these estimates are very imprecise, they do seem to confirm that credit plays a significant role in a firm's ability to hire. Section 3.7 concludes.

3.2 Related literature

Shift of the Beveridge curve and matching efficiency. Referring to a so-called "matching efficiency puzzle," Barnichon and Figura (2015) and Hall and Schulhofer-Wohl (2018) attempt to fix the matching function by allowing heterogeneity in inputs and sectoral matching technologies. The idea underlying these generalized matching functions is twofold. First, the recession may have shifted the composition of the pools of firms and workers,

with groups displaying structurally lower matching efficiencies becoming more prominent. A convincing example of this phenomenon is the documented increase in the share of long-duration unemployed workers. Second, the recession might also have increased dispersion in the inputs, which would mechanically create inefficiency if we assume the matching process to be convex. While the generalized matching functions allow for a much better fit of the model to the data after the recession, Hornstein and Kudlyak (2016) argue that the documented countercyclical search intensity offsets the composition effect, "leaving the entire drop in match efficiency to be explained." For this reason, the dispersion effect may be more promising. The idea of a rise in dispersion is indeed related to the notion of mismatch, which could explain up to 30% of the Beveridge curve shift according to Şahin et al. (2014). Another caveat, however, is that the generalized matching functions described above provide purely mechanical answers: they do not provide endogenous explanations for changes in dispersion or in composition. A second strand of literature filled this gap by focusing on the determinants of workers' search efficiency through the search behavior of jobseekers. However, no determinant seems to be able to solely explain the stark deterioration of the job-finding rate. In particular, as mentioned earlier, taking into account search intensity does not help because it behaves in a countercyclical fashion. Only recently has the literature turned towards the recruiting behavior of firms, following an important contribution by Davis et al. (2013). Using firm and establishment-level microdata, they are able to show that firms do not rely uniquely on the number of openings in order to achieve their hiring targets, but instead modulate their recruiting intensity through many different margins, such as screening effort. Gavazza et al. (2018) combine Davis et al. (2013)'s generalized matching function with a firm life cycle as well as a collateral constraint to study how productivity and financial shocks transmit to the labor market through recruiting intensity. Leduc and Liu (2016) relate recruiting intensity to uncertainty. Kaas and Kircher (2015) endogenize recruiting intensity through wage-posting, which they find can account for two-thirds of the variation in vacancy yield. Sedláček (2014) endogenizes it through yet another channel: vari-

able hiring standards. Elsby et al. (2015), however, show that recruiting intensity may not be the answer. A counterfactual exercise shows that the index of recruiting intensity devised by Davis et al. (2013) cannot account for the shift in a persistent fashion. Motivated by these works, my project is rooted in the idea that giving more attention to the determinants of vacancy yields on the firms' side certainly constitutes a promising step towards a better understanding of the matching function. I depart from the papers previously cited in that I do not try to capture a fall in the recruiting intensity of firms, but rather in their ability to recruit. These two channels bear very different policy implications. Indeed, in the first case, the fall in recruiting efficiency stems from the firm's optimal recruiting intensity decision. In the second case, it stems from the firm being constrained. Note that this paper does not claim to provide a stand-alone explanation of the Beveridge curve shift. Instead, my goal is to contribute to the existing literature by studying a new mechanism and assessing its importance.

Impact of financial frictions on the labor market. Wasmer and Weil (2004) were pioneers in integrating financial frictions to a labor search framework. Petrosky-Nadeau and Wasmer (2013) provide a dynamic extension. Petrosky-Nadeau (2014) attempts to model the financial frictions more realistically, embedding them in a costly state verification problem rather than a search and matching problem. A common feature of these studies is the focus on the propagation of shocks in the economy, but little attention is given to the efficiency of matching nor, for this matter, to shifts of the Beveridge curve. Boeri et al. (2018) also add financial frictions to the MP model, using limited pledgeability a la Holmstrom and Tirole (1997). While their research question is closer to mine, their focus is primarily turned towards the firm's liquidity accumulation decision. I follow the original WW approach to implement financial frictions to my model. First, modeling financial frictions as search frictions still seems very relevant empirically. Mills and McCarthy (2014) report that small firms (less than 500 employees) take on average 24 hours to search for banks and apply for

loans. Greenstone et al. (2020) provide evidence of relationship lending and of the costs associated with finding a lender: "one standard deviation reduction in the 2009 measure of local credit supply shocks is associated with a 17% reduction in total county level small business loan originations from the end of 2008. through the end of 2010." Second, WW's formalization allows me to keep a very tractable representative agent model, solvable in analytical form. It is also worth mentioning another branch of literature that flourished after the Great Recession, in the wake of Jermann and Quadrini (2012), taking a narrower focus on labor market outcomes. The common denominator of these models, surveyed in Boeri et al. (2018), is the idea that short-term liquidity constraints may impact firms hiring decisions as they require liquidity to pay worker salaries. For example, Monacelli et al. (2011) suggest that the deterioration of credit conditions worsens firms' bargaining position when negotiating wages, and therefore reduces incentives to hire.[3] The fixed cost of hiring I introduce in WW is very similar to Pissarides (2009). However, this work is embedded in the large literature attempting to solve the unemployment volatility puzzle, and does not directly relate to the Beveridge curve. Another difference is that in contrast with Pissarides (2009), I do not model the fixed cost as sunk. This consideration does not matter in my standard specification given equilibrium wage is set exogenously. Later in the paper, when I extend the model to allow for wage bargaining, the timing is such that the fixed cost is to be paid after the wage has been negotiated.

Differences with Wasmer and Weil (2004). WW formalize the general equilibrium interaction between financial and labor markets by adding a decentralized, frictional credit market to the MP model. Firms do not have wealth of their own and need financing from banks in order to cover their search costs in the labor market. For this reason, firms must successfully search for a bank before they can open a vacancy. This set up allows WW to provide an elegant formalization of the feedback loop between the two markets, and

[3] A channel I can explore in section 3.5, where wage bargaining is endogenized.

to highlight the accelerator effect played by credit frictions following to aggregate shocks. However, because the search for financing occurs before the firm searches for a worker, shocks to credit cannot impact aggregate matching efficiency. In other words, they can only result in moves along the Beveridge curve. For credit availability to generate shifts of the Beveridge curve, the need for credit must enter the model after the firm and the worker have already matched. I implement this in my model by assuming that once matched with a worker, a firm needs to pay a fixed cost in order to effectively perform the hire. The fixed cost must be financed by bank credit, such that the hire cannot take place if the firm fails to find financing. While he is not concerned with credit frictions, Pissarides (2009) introduces a very similar fixed component to the matching costs born by the firm in the MP model. He argues that "the assumption that there are fixed costs to job creation is in itself a realistic assumption. These costs include the costs of negotiating with the successful job applicant, putting her on the firm's payroll, and training her."[4] Exactly as in his model, the crucial element for my results to go through is the timing of the fixed cost. It must be paid after the worker has been found, and thus be independent of the duration of vacancies. Focusing on hiring costs post-match rather than vacancy search costs pre-match seems natural in an economy where online job postings have become prevalent, a phenomenon likely to decrease the cost of publicizing vacancies. Another supportive statistic comes from Davis et al. (2013). They estimate that 40% of hires come from establishments that did not report a vacancy in the preceding month, because the hire was opportunistic or resulted from a very short search. This backs the idea that credit may impact the hiring success of firms more strongly at the time of hiring than at the time of searching for workers.

[4]One could also include screening costs to this list, depending on the interpretation made of the standard vacancy costs that firms have to pay while they are searching for workers. These are most usually seen as encompassing not only the search but also the selection process, in which case screening costs would not be an adequate interpretation of the fixed hiring cost featured in this model. I do not have data breaking down costs pre- and post-match. Including search costs, estimates of the total cost of hiring a new employee range from $4000 to $8000. See Blatter et al. (2012) and or Bersin Talent Acquisition Factbook by Deloitte for example.

3.3 Baseline model: exogenous loan-approval rate

The goal of this section is to show how introducing credit frictions at the time of hiring can impact the Beveridge curve and could be mistaken for a drop in matching efficiency.

3.3.1 Environment

Time is discrete and continues forever. There are three types of agents in the model: entrepreneurs, workers, and banks. All types enjoy linear utility from the consumption of a unique good, and discount from a period to another at rate β. Production occurs in pairs of one entrepreneur and one worker, referred to as firms, and results in an output y per period. Any entrepreneur is free to open a vacancy to attract workers, at a cost c per period. Vacancies and unemployed workers match according to a matching function $M(U_t, V_t)$, where U_t is the stock of jobseekers and V_t the stock of open vacancies at time t. The matching technology is increasing and concave in both of its arguments, and displays constant returns to scale. Let $V_t/U_t \equiv \theta_t$ be the tightness of the labor market, then entrepreneurs match with workers with a probability $M(U_t, V_t)/V_t = q(\theta_t)$ every period, while jobseekers match with a vacancy with a probability $M(U_t, V_t)/U_t = \theta_t q(\theta_t)$ every period. The properties of the matching function imply that $q'(\theta) < 0$ and $[\theta q(\theta)]' > 0$. The model differs from the standard MP model in that in order to turn a match into a hire, the entrepreneur must pay a one-time fixed cost, F. This cost has to be financed by bank credit. For now, I will assume that upon matching with a worker, entrepreneurs are able to send out applications and get accepted immediately with an exogenous probability p. This approval rate will later be endogenized as a "loan-finding" rate in the spirit of Wasmer and Weil (2004). In case of a rejection, the entrepreneur loses its match with the worker and returns to having an open vacancy. In case of an approval, the firm and the bank must negotiate the repayment plan, which consists in a per period payment, ρ, until the job is destroyed. Upon agreement,

the bank pays the fixed cost F, the the worker is hired and production starts immediately. Job destruction is exogenous and occurs with probability s every period. Productivity y and wage w are exogenous. I assume that the wage is set such that it is consistent with a worker's decision to accept the job offer.

3.3.2 Bellman equations

Let the present value, for an entrepreneur, of having a vacancy open and trying to hire be

$$V_E = -c + \beta q(\theta) p J_E + \beta(1 - q(\theta)p)V_E. \tag{3.1}$$

Every period the entrepreneur has a vacancy open, he has to pay a cost c. This allows him to be matched with a jobseeker with probability $q(\theta)$. If this search is successful, the entrepreneur then sends loan applications and gets approved instantly with probability p. In case of an approval, the entrepreneur can hire the worker and enters the production stage at the next period. If either the search for a worker or for a loan fails, then the entrepreneur goes back to having an open vacancy. The present-discounted value of a job is

$$J_E = y - w - \rho + \beta s V_E + \beta(1 - s)J_E. \tag{3.2}$$

While a job is running, the entrepreneur earns per-period profits equal to the job productivity net of wages and repayments to the bank. Every period, the job can be destroyed with probability s, in which case the entrepreneur goes back to the search stage.

3.3.3 Bargaining of the loan repayment

The loan contract between a entrepreneur and a bank is one-dimensional. The bank finances the fixed cost of hiring, F, upfront, in exchange of which the entrepreneur commits to repay ρ for as long as the job is running. Note that ρ combines both the principal and the interest payments. It is determined by generalized Nash bargaining, such that

$$\rho = \text{argmax } (J_E - V_E)^{1-\alpha} \left(\frac{\rho}{1-\beta(1-s)} - F \right)^{\alpha}, \tag{3.3}$$

where $\alpha \in (0,1)$ is the bargaining power of the bank and $(1-\alpha)$ that of the entrepreneur. In case of an agreement, the entrepreneur will be able to enter the production phase, with present value J_E, while the bank will earn the expected present-discounted sum of repayments net of the fixed cost. The disagreement point is the value of having an open vacancy for the entrepreneur, and zero for the bank.

3.3.4 Equilibrium labor market tightness

I only consider stationary equilibria in which the expected profits from a job investment, $\Pi \equiv (y-w)/(1-\beta(1-s)) - F$, are non-negative.

Vacancy supply. Profit maximization drives the present-discounted value of a vacancy to zero, $V_E = 0$. From (3.1), we get

$$J_E = \frac{c}{\beta q(\theta) p}. \tag{3.4}$$

In equilibrium, the value for an entrepreneur of having a job running must be equal to the average search cost the firm has to incur to find a worker and a loan. Imposing the free-entry

condition on (3.2) yields

$$J_E = \frac{y - w - \rho}{1 - \beta(1 - s)}. \tag{3.5}$$

In equilibrium, the value of a job is also equal to the expected present-discounted sum of future profits net of the loan repayment. Finally we can get the vacancy-supply condition by equating (3.4) and (3.5):

$$\frac{c}{\beta q(\theta)} = \frac{y - w - \rho}{1 - \beta(1 - s)} p. \tag{3.6}$$

One last step is required to obtain a closed-form vacancy-supply condition: solving for ρ.

Loan repayment. In case of an agreement, the firm's surplus is $J_E - V_E = (y - w - \rho) = (1 - \beta(1 - s))$. Using this into the bargaining rule (3.3) allows us to the equilibrium per-period loan repayment,

$$\rho = \alpha(y - w) + (1 - \alpha)F[1 - \beta(1 - s)]. \tag{3.7}$$

The loan repayment is a weighted average of the ow output net of wages and the equivalent per-period cost of the fixed hiring cost. It is independent of labor market tightness. The higher the bargaining power of the bank, the more it can extract from the job's net profits. Note that given the Pareto frontier is linear, the generalized Nash solution coincides with Kalai's proportional solution. The bank's expected surplus is $\alpha\Pi$ and the firm's expected surplus is $(1 - \alpha)\Pi$.

Labor market tightness. A closed-form solution for equilibrium labor market tightness is given by the following vacancy-supply condition (VS), obtained by plugging (3.7) into

(3.6):

$$\frac{1}{q(\theta)} = \frac{\beta(1-\alpha)\Pi p}{c}. \tag{3.8}$$

Compared to the standard MP vacancy-supply condition, three additional parameters impact equilibrium market tightness: p, F and α. Everything else equal, a higher probability of loan approval increases labor market tightness, $\partial\theta/\partial p > 0$, while a higher fixed cost of hiring or a higher bargaining power for banks decrease labor market tightness, $\partial\theta/\partial F < 0$ and$\partial\theta/\partial\alpha < 0$. The intuition behind these results is straight-forward. A higher probability of loan approval increases the value for a firm of matching with a worker. This encourages firms to enter the labor market, which increases market tightness. On the other hand, both a higher fixed cost of hiring and a higher bargaining power for banks increase the per-period repayment the firm has to make to the bank in case of a loan agreement. Facing a lower expected surplus from job creation, entrepreneurs exit the market and labor market tightness increases.

3.3.5 Beveridge curve

I can now study the impact of credit frictions at time of hiring on the equilibrium relation between unemployment and vacancies. At steady-state unemployment rate, the flows in and out of unemployment must be equal, $s(1-u) = \theta q(\theta) p u$, so that

$$u = \frac{s}{s + \theta q(\theta) p}. \tag{3.9}$$

The impact of credit frictions on unemployment is twofold. First, p acts as Beveridge curve shifter. For a given market tightness, a decrease in p increases u: the Beveridge curve shifts outwards. Second, p impacts θ, as highlighted in the vacancy-supply condition (3.8). This causes moves along the Beveridge curve. An increase in p increases θ, and implies sliding

down the Beveridge curve, towards a higher unemployment rate.

Using the standard Cobb-Douglas matching function, $M(U,V) = A\sqrt{UV}$, the Beveridge curve can be written as $= s = (s + pA\sqrt{\theta})$. For a given θ, the shifting effect of credit frictions is confounded with the efficiency parameter A.

3.4 Endogenizing credit frictions

Until now, credit frictions were imposed exogenously. In order to study how productivity shocks may impact the firms' ability to find credit, one needs to endogenize the loan-approval rate, p. Following WW, I add a decentralized credit market to the baseline framework described in section 3.3.1. Banks are free to enter the market and to search for entrepreneurs to finance, at a cost k per period. In this market, banks meet the entrepreneurs already matched with a worker according to the matching function $H(B_t, M_t)$, which features the same properties as $M(U_t, V_t)$. Let $\phi_t = M_t/B_t$. A matched firm meets a bank with probability $p(\phi_t) = H(B_t, M_t)/M_t$, decreasing in ϕ_t. A bank meets a matched firm with probability $\phi_t p(\phi_t) = H(B_t, M_t)/B_t$, increasing in ϕ_t. Note that contrary to WW, loans are not intended to finance the vacancy costs (i.e., the ongoing cost of searching for a worker), but a fixed hiring cost. Because of this, a firm enters the credit market only after it has successfully matched with a worker in the labor market. The timing of events is summarized in Figure 3.3.

3.4.1 Bellman equations

On the firm's side, the Bellman equations corresponding to the worker-search and to the production stage are identical to the setting with exogenous p, (3.1) and (3.2). the only difference being that loan-acceptance probability p now depends on credit market tightness

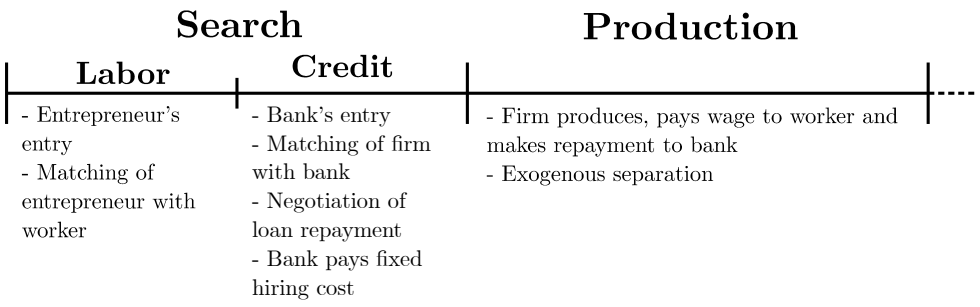

Figure 3.3: Timeline of job creation with endogenous bank entry.

ϕ.

I now describe the bank's problem. A bank goes through two stages: search for a firm to finance, and production. Let the present value, for a bank, of searching for a firm to finance be

$$V_B = -k + \beta\phi p(\phi)(J_B - F) + \beta(1 - \phi p(\phi))V_B. \tag{3.10}$$

A bank has to pay a cost k to acquire the screening technology necessary to screen applicants and process applications. This allows the bank to meet a firm the following period with probability $\phi p(\phi)$. In that event, the firm and the bank negotiate the terms of the loan and the bank pays the fixed cost of hiring upfront. This allows production to start immediately. The present-value of a financed job, for a bank, is

$$J_B = \rho + \beta s V_B + \beta(1 - s)J_B. \tag{3.11}$$

The bank earns the loan repayment as long as the job is running. If the job is destroyed, the bank goes back to looking for loan applicants.

3.4.2 Bargaining of the loan repayment

The Nash bargaining program is now

$$\rho = \text{argmax}(J_E - V_E)^{1-\alpha}(J_B - F - V_B)^{\alpha}. \tag{3.12}$$

The Pareto frontier is still linear and the first-order condition requires

$$\alpha(J_E - V_E) = (1-\alpha)(J_B - F - V_B). \tag{3.13}$$

3.4.3 Equilibrium market tightness and unemployment

Free entry of banks in the credit market and of entrepreneurs in the labor market drive V_E and V_B to zero. Following the same steps as in section 3.3.4, we get that the equilibrium value of a matched firm still is given by (3.4), and the vacancy-supply condition is

$$\frac{c}{\beta q(\theta)} = \frac{y - w - \rho}{1 - \beta(1-s)} p(\phi). \tag{3.14}$$

From (3.10), we can then get that the value of a job, for a bank, must equal the total costs associated with financing a job,

$$J_B = \frac{k}{\beta \phi p(\phi)} + F. \tag{3.15}$$

From (3.11), we also get that the equilibrium value of a job, for a bank, is equal to the expected present-discounted sum of future repayments,

$$J_B = \frac{\rho}{1 - \beta(1-s)}. \tag{3.16}$$

Combining (3.15) with (3.16) gives the credit-market analogous of the vacancy-supply condition, which I will name the loan-supply condition (LS),

$$\frac{k}{\phi p(\phi)} = -F + \frac{\rho}{1-\beta(1-s)}. \tag{3.17}$$

Making use of (3.5) and (3.16) into (3.12), we can directly see that the program is exactly the same as in the exogenous framework, such that the equilibrium repayment ρ is still defined by (3.7). The independence result is now particularly interesting as it contrasts with WW. Written in discrete time, WW's repayment schedule satisfies $\rho^{WW} = \alpha(y-w) + (1-\alpha)(1-\beta(1-s))c/(\beta q(\theta))$, and thus positively depends on market tightness. In their framework, because banks finance vacancy search costs, the total cost born by the bank eventually depends on the time it takes for the firm to find a worker. This is why their contract takes into account average search duration, $q(\theta)^{-1}$. In my framework, there is no uncertainty regarding the bank's financing costs at the time the contract is written, and those do not depend labor market tightness. Additionally, conditional on having found a worker, the surplus enjoyed by the firm in case a loan agreement is reached does not depend on the (sunk) search costs. Hence, total surplus is independent of labor market tightness, and θ is absent from the loan-repayment equation. Plugging for equilibrium ρ into (3.14) and (3.17), we can pin down equilibrium (θ, ϕ).

Proposition 3.1. *In equilibrium, the pair (θ, ϕ) must satisfy the following two equations, respectively denoted the vacancy-supply (VS) and the loan-supply (LS) conditions:*[5]

$$\frac{c}{\beta q(\theta)} = (1-\alpha)\Pi p(\phi), \tag{3.18}$$

and

$$\frac{k}{\phi p(\phi)} = \alpha\Pi. \tag{3.19}$$

[5]These are denoted the EE and BB 0-profit loci in WW.

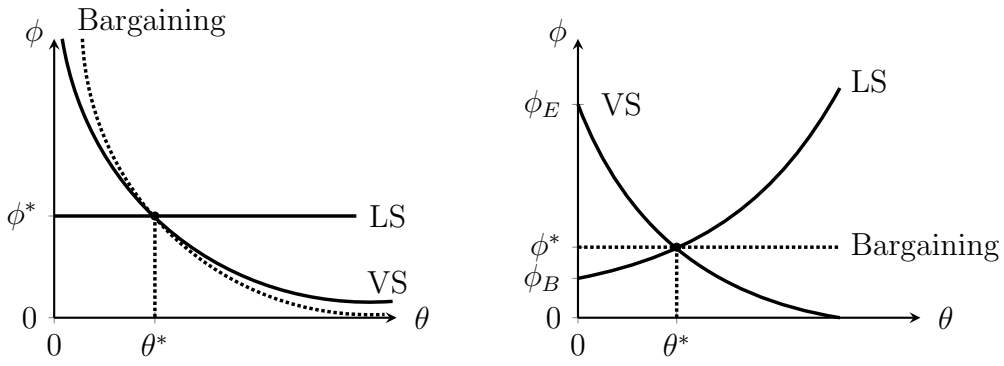

Figure 3.4: Determination of equilibrium (θ, ϕ) in LL (left panel) and WW (right panel).

The VS condition states that in a zero-profit equilibrium, labor and credit markets tightnesses are inversely related. Indeed, if labor market tightness increases, it takes more time on average for firms to find a worker, and therefore the expected costs of search are higher. For profits to stay null, firms must be compensated by lower costs of looking for a bank. Another way to look at it is that as the credit market tightness increases, it becomes less likely for a firm to get the financing necessary to hire. This reduces firms' incentives to enter the labor market, and therefore labor market tightness decreases. The LS condition states that in a zero-profit equilibrium, credit market tightness is inversely related to the bank's expected profits and positively related to the bank's search costs. From the VS equation we can see that $\lim_{\theta\to 0}\phi = \infty$, and $\lim_{\theta\to\infty}\phi = 0$. Hence, an equilibrium exists, and is unique, as long as $\alpha[(y-w)/(1-\beta(1-s))-F] > 0$. This is true by definition. The determination of equilibrium (θ, ϕ) is represented in the left panel of Figure 3.4. On that panel, the bargaining curve is obtained by plugging (3.4) and (3.15) into the Nash bargaining rule (3.13), which gives $\theta = (1-\alpha)kq(\theta)/\alpha c$.

Note that equilibrium ϕ is uniquely pinned down by the loan-supply condition. It is independent from θ, and increases in the expected value of a job, Π. In other words, banks' entry is not influenced by labor market θ, and only depends on fundamental parameters of the economy (bargaining power of the bank and expected value of a job). This is an interesting

equilibrium result to contrast to WW. Recast in discrete time, equilibrium in WW is defined by the VS, LS and bargaining conditions:

$$\frac{\kappa}{\beta p(\phi)} = (1-\alpha)\frac{\beta q(\theta)}{1-\beta[1-q(\theta)]}\left[\frac{y-w}{1-\beta(1-s)} - \frac{c}{\beta q(\theta)}\right] \tag{3.20}$$

$$\frac{k}{\beta\phi p(\phi)} = \alpha\frac{\beta q(\theta)}{1-\beta[1-q(\theta)]}\left[\frac{y-w}{1-\beta(1-s)} - \frac{c}{\beta q(\theta)}\right] \tag{3.21}$$

$$\phi = \frac{1-\alpha}{\alpha}\frac{k}{c} \tag{3.22}$$

where κ is the per-period search cost born by firms in the credit market (set to 0 in my model). Equilibrium determination in WW is represented in the right panel of Figure 3.4, where ϕ_B satisfies $k/[\beta\phi_B p(\phi_B)] = \alpha[(y-w)/(1-\beta(1-s))]$ and ϕ_E satisfies $\kappa/[\beta p(\phi_E)] = (1-\alpha)[(y-w)/(1-\beta(1-s))]$.

In their model, ϕ is also uniquely determined by fundamental parameters, through the bargaining rule (3.22). However, those parameters only relate to search costs, such that equilibrium credit market tightness does not vary over the business cycle, nor following shocks to the matching technology. This matters because the unresponsiveness of ϕ to aggregate productivity shocks implies that the deterioration of credit conditions during economic downturns cannot account for a decline in labor market efficiency.

Finally, the Beveridge curve can be written as

$$u = \frac{s}{s+\theta q(\theta)p(\phi)}. \tag{3.23}$$

Once again, notice that credit market conditions impact the Beveridge curve through two channels: an indirect amplification channel, with θ, and a direct channel, with $p(\phi)$. In WW, the Beveridge curve does not include this latter term, and financial frictions only play an amplification role.

Response to shocks. The impact of a productivity shock on equilibrium tightnesses and unemployment rate is represented in Figure 3.5. Note this figure could also represent the impact of any negative shock to the value of a job, Π (e.g., an increase in wage or an increase in the fixed cost of hiring). The mechanism is as follows. A negative shock to productivity y decreases expected profits from a job, Π, which results in the VS curve shifting down. Everything else equal, labor market tightness is now lower for any given level of credit market tightness. Indeed, if the firm expects lower gains from creating a job, it has to be compensated by lower search costs—either through a higher probability of finding a worker, or either through a higher probability of finding financing. Absent a feedback between the credit and the labor markets, this is all what would happen: equilibrium labor market tightness decreases, and unemployment increases, displayed by a move along the Beveridge curve. This counterfactual situation is represented by θ_c and u_c on the graph. Adding credit frictions has two effects. First, a drop in productivity now shifts the loan-supply curve to the right. The adjustment margin for the credit market is driven by banks' entry (matched firms are a state variable), and lower expected profits decrease banks' incentives to enter the market. As a result, equilibrium credit market tightness increases, which magnifies the drop in labor market tightness along the VS curve, as well as the increase in the unemployment rate along the Beveridge Curve. This result corresponds to the amplification effect documented in WW. There is now an additional step. Indeed, since θ is lower for any given ϕ, it means that the Beveridge curve must have shifted outwards. This takes equilibrium unemployment even further, to u_2.

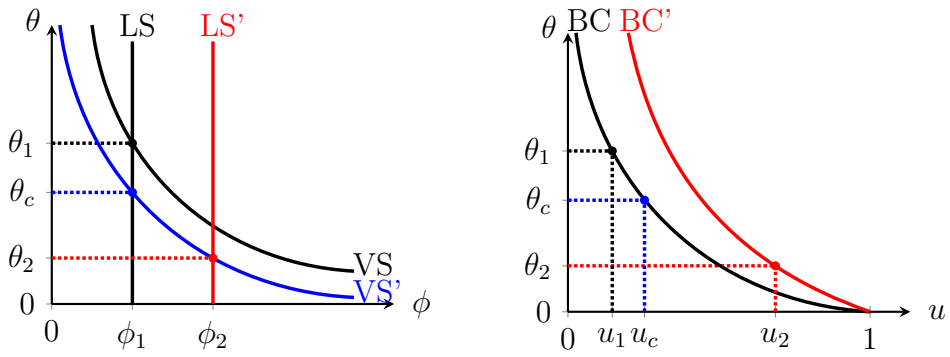

Figure 3.5: Impact of a negative shock to y on eq. (θ, ϕ) (left panel) and on the Beveridge curve (right panel)

3.5 Extension: endogenous wage bargaining

Would the impact of financial frictions on equilibrium market tightnesses, unemployment, and the Beveridge curve be different if we allowed firms and workers to negotiate the wage? The difficulty here, as highlighted by WW, is that we now have to consider bargaining between not two but three agents. Firms have to bargain on two different fronts: on one end, they negotiate with banks over the loan-repayment contract; on the other, they negotiate with the jobseeker over the wage contract. Because the order of negotiations is likely to impact both of these outcomes, I will explore the two possible scenarios - wage first, or loan contract first.

3.5.1 Worker's value functions

In order to solve the bargaining problem between a firm and a worker, we need to specify the worker's value functions. Let the present-discounted lifetime utility from being unemployed be

$$U = z + \beta\theta q(\theta)p(\phi)W + \beta[1 - \theta q(\theta)p(\phi)]U. \tag{3.24}$$

When unemployed, a worker earns a real compensation z (e.g., benefits, leisure) and can expect to be matched with a firm the following period with probability $\theta q(\theta)$. If she has matched with a firm, the unemployed worker's status instantaneously depends on the outcome of the firm's search for a bank. If the firm does not succeed in finding financing, the match with the worker breaks, and she goes back to unemployment. If, however, the loan search is successful, then the job can start. The present-discounted value of having a job, for a worker, is

$$W = w + \beta(1 - s)W + \beta sU. \tag{3.25}$$

While employed, the worker earns a wage w, and faces the risk of a job destruction with probability s. The worker's problem can be simplified into a system of two equations linear in W_0 and W_2,

$$U = \frac{z + \beta\theta q(\theta)p(\phi)W}{1 - \beta[1 - \theta q(\theta)p(\phi)]}, \tag{3.26}$$

and

$$W = \frac{w + \beta sU}{1 - \beta(1 - s)}. \tag{3.27}$$

3.5.2 Loan contract first, wage second

This order of negotiations is similar to the one described in WW. Applied to my model, it means that once a firm has matched with a worker, it waits until having found a bank and specified the loan contract before negotiating with the worker. The sequence of bargaining problems is solved by backwards induction, using generalized Nash bargaining at each stage. I start by working out the bargaining problem between a firm and a worker, taking the loan repayment, ρ, as given. I then proceed to solve for ρ, using the wage the firm expects to

bargain later on with the worker.

Let $\epsilon \in (0,1)$ be the bargaining power of the firm. The Nash problem can be written as

$$w = \text{argmax} \left(J_E - V_E\right)^{\epsilon} \left(W - U\right)^{1-\epsilon}. \tag{3.28}$$

Imposing the free-entry condition on the firms' side and taking the first-order condition yields

$$\epsilon(W - U) = (1 - \epsilon)J_E. \tag{3.29}$$

Plugging in for (3.5) and (3.27), we get

$$w = \epsilon(1 - \beta)U + (1 - \epsilon)(y - \rho). \tag{3.30}$$

The wage is a weighted average of the worker's per-period expected value of unemployment, and the output net of the loan repayment. As the worker gets more bargaining power (lower ϵ), she can extract more of the job's net profits. On the other hand, as the firm gains more bargaining power, it can bring the wage closer to the worker's reservation wage. We can also re-express the wage as a function of ρ only by solving for U. We get

$$w = (1 - \bar{\epsilon})(y - \rho) + \bar{\epsilon}z \tag{3.31}$$

where

$$\bar{\epsilon} = \frac{\epsilon[1 - \beta(1 - s)]}{(1 - \epsilon)[1 - \beta(1 - \theta q(\theta)p(\phi) - s)] + \epsilon[1 - \beta(1 - s)]} \in (0,1). \tag{3.32}$$

As the firm's bargaining power tends to one, $\bar{\epsilon}$ tends to one as well, and can drive the wage down up until the worker is indifferent between accepting or rejecting the offer. When the

firm's bargaining power tends to zero, $\bar{\epsilon}$ also tends to zero and the worker can extract up to the totality of the net profits.

The next step consists in solving for the loan contract, using the wage rate derived in (3.31). The Nash problem can be written as in (3.12). After imposing the free entry condition as well as equilibrium wage, we have

$$\rho = \text{argmax} \left(\bar{\epsilon} \frac{y - \rho - z}{1 - \beta(1-s)} \right)^{1-\alpha} + \left(\frac{\rho}{1 - \beta(1-s)} - F \right)^{\alpha}. \tag{3.33}$$

Note that because of the introduction of $\bar{\epsilon}$, the frontier of the bargaining set is not linear anymore. Solving the maximization problem gives us

$$\rho = \alpha(y - z) + (1 - \alpha)F[1 - \beta(1 - s)]. \tag{3.34}$$

This equation bears close resemblance to that obtained when the wage was exogenous. The only difference is that the maximum repayment the bank can negotiate, with a bargaining power of one, would be the output net of the worker's reservation wage, $y - z$, instead of $y - w$. Because $z \leq w$, the firm is now worse off. Indeed, it looks like the negotiation happens as if the firm were later be able to extract all of the surplus of its match with the worker. It can only do so when $\epsilon = 1$. Another interesting feature is that the repayment still does not depend on market tightnesses.

Finally we can plug back for ρ into our wage equation, and get

$$w = (1 - \bar{\epsilon})(1 - \alpha)[y - F(1 - \beta(1 - s))] + [1 - (1 - \bar{\epsilon})(1 - \alpha)]z. \tag{3.35}$$

The impact of $\bar{\epsilon}$ on the wage is straightforward: as the firm enjoys a higher discounted bargaining power against the worker, it is able to bring the wage closer and closer to the reservation wage z. More interestingly, when the firm enjoys a higher bargaining power

against the bank, it is able to extract less of the surplus when negotiating with a worker. Equivalently, workers are harmed by stronger banks through a decrease in wages. As for firms, we can show that unless the firm can extract all the surplus against a worker, the increase in loan repayments triggered by a higher bargaining power for banks is stronger than the decrease in wages, such that overall firms are made worse off as well.

Plugging (3.34) and (3.35) into the LS and VS conditions, (3.17) and (3.14), we can characterize equilibrium (θ^*, ϕ^*).

Proposition 3.2. *When the firm negotiates the loan contract with the bank before it negotiates the wage with the worker, equilibrium* (θ, ϕ) *satisfies*

$$\frac{c}{\beta q(\theta)} = \bar{\epsilon}(1-\alpha)\left(\frac{y-z}{1-\beta(1-s)} - F\right)p(\phi) \tag{3.36}$$

and

$$\frac{k}{\phi p(\phi)} = \alpha\left(\frac{y-z}{1-\beta(1-s)} - F\right). \tag{3.37}$$

Because $z \leq w$, the right-hand side of (3.37) is higher than when the wage was set exogenously. Hence, equilibrium credit market tightness is lower, as is the case in WW. Studying the vacancy-supply condition requires more attention. We can rewrite (3.36) as

$$\frac{c}{\beta q(\theta)} = p(\phi)\left[\epsilon(1-\alpha)\left(\frac{y-z}{1-\beta(1-s)} - F\right) - (1-\epsilon)\frac{c\theta}{1-\beta(1-s)}\right], \tag{3.38}$$

which still defines a positive relation between labor and credit market tightnesses. We can see from the VS and LS conditions that qualitatively, the equilibrium response to shocks is identical to the setting with exogenous wage. The magnitude of responses however depends on the slope of the VS condition, which may be shallower or steeper than the exogenous wage setting depending on the model's parameters.

3.5.3 Wage first, loan contract second

In this section I assume that the firm and the worker negotiate directly after matching, such that the wage is already known when the firm subsequently negotiates with the bank. As before, I proceed by backwards induction and start by solving for the loan repayment contract. The set up here is exactly the same as when the wage was exogenous, thus we directly get

$$\rho = \alpha(y - w) + (1 - \alpha)F[1 - \beta(1 - s)]. \tag{3.39}$$

Plugging for ρ into (3.28), we get

$$w = \text{argmax} \left((1 - \alpha) \left(\frac{y - w}{1 - \beta(1 - s)} - F \right) \right)^{\epsilon} (W_2 - W_0)^{1-\epsilon}. \tag{3.40}$$

Using the first-order condition, we can write

$$\epsilon(W_2 - W_0) = (1 - \epsilon)\frac{y - w - F[1 - \beta(1 - s)]}{1 - \beta(1 - s)}. \tag{3.41}$$

Notice the only difference with (3.29) is the presence of $F[1 - \beta(1 - s)]$ in the numerator instead of ρ. Thus we directly get the equilibrium wage,

$$w = (1 - \bar{\epsilon}) \left[y - F(1 - \beta(1 - s)) \right] + \bar{\epsilon} z. \tag{3.42}$$

When the wage is negotiated before the loan repayment contract is, the outcome is "as if" the firm was able to extract all of the surplus from its relationship with a bank later on. Notice that compared to the reverse order of negotiation, equilibrium wage is higher and does not depend on the distribution of bargaining powers between the bank and the worker.

Plugging back for equilibrium w into the repayment equation, we get

$$\rho = \alpha\bar{\epsilon}(y - z) + (1 - \alpha\bar{\epsilon})F[1 - \beta(1 - s)]. \tag{3.43}$$

In terms of loan repayment, the firm is, again, worse off when the wage is endogenized than when it is not. However, negotiating the loan contract first results in a lower loan repayment than when the order of negotiations is reversed. In particular, the lower the bargaining power of the firm against the worker, the lower the repayment. Overall, the lower repayment cancels out the higher wage, and in equilibrium, the firm's expected profits, $(y-w-\rho)/[1-\beta(1-s)]$, are the same regardless of the order of negotiation.

Plugging (3.43) and (3.42) into the LS and VS conditions, (3.17) and (3.14), we can characterize equilibrium (θ^*, ϕ^*).

Proposition 3.3. *When the firm negotiates the wage with the worker before it negotiates the loan contract with the bank, equilibrium (θ, ϕ) satisfies*

$$\frac{c}{\beta q(\theta)} = \bar{\epsilon}(1 - \alpha)\left(\frac{y - z}{1 - \beta(1 - s)} - F\right)p(\phi) \tag{3.44}$$

and

$$\frac{k}{\phi p(\phi)} = \bar{\epsilon}\alpha\left(\frac{y - z}{1 - \beta(1 - s)} - F\right). \tag{3.45}$$

The VS condition is exactly the same as in the previous section: the order of negotiation does not influence firms' entry. This is intuitive, as firms' profits are not impacted by the order of negotiation. The LS condition is slightly different, as the bargaining power of the bank, α, is now multiplied by the discounted bargaining power the firm enjoys against a worker, $\bar{\epsilon}$. Because $\bar{\epsilon}$ depends on both ϕ and θ, equilibrium credit market tightness cannot be pinned down by the LS condition only. In the (θ, ϕ) plane, the LS condition is not

represented by a horizontal line anymore. Instead, it resembles the LS condition from WW (see right panel of Figure 3.4). In this setting, the response to a productivity shock may differ qualitatively from the previous results. A decrease in y would trigger upwards shifts both for the VS and the LS conditions. While this unambiguously raises equilibrium labor market tightness, equilibrium credit market tightness may go in either direction. For that matter, the Beveridge curve may shift inwards rather than outwards.

3.6 Empirical exercise: assessing the contribution of the credit channel

In this section, I conduct an empirical exercise aimed at assessing the empirical relevance of the credit channel formalized in section 3.3 using an index of loan approval I construct. In the spirit of Elsby et al. (2015), I derive a counterfactual vacancy series that gives the vacancy rate had the index not changed since 2005. This can be used to plot a p-constant counterfactual Beveridge curve and to estimate how much of the shift may be accounted for by changes in p.

3.6.1 Data

Lacking access to monthly data on credit availability for the whole economy, I build a loan-approval rate series using data from the NFIB Small Business Survey. This survey provides monthly data going back to 1986 up to now, with a sample size of around 1500 small businesses. The survey question of interest is the following: "During the last three months, was your firm able to satisfy its borrowing needs?" Over the time frame considered, January 2005 to June 2017, the average number of respondents to this question is 1031. The four possible answers are "yes" (31.6%), "no" (5.9%), "NA" (49.2%) and "no reply" (13.3%).

The breakdown of answers from January 2005 to June 2017 is plotted in Appendix C, Figure C.1.

Small firm loan-approval rate. I generate two series directly from the survey's answers : $p_{s,1} = pr(\text{yes}|\text{answered})$ and $p_{s,2} = pr(\text{yes}|\text{answered and not NA})$. The two series are plotted in light shade in Figure 3.6. They both follow a similar path up to the beginning of 2012. However, while $p_{s,2}$ has completely recovered by 2015, this is not the case of $p_{s,1}$. This goes together with a steady rise in the share of "NA" answers following the recession (see Figure C.1). "NA" can be interpreted as firms who did not apply for a loan during the three months preceding the survey. Even though it might seem reasonable to exclude them, as in $p_{s,2}$, there is evidence that firms self-exclude from asking for loans because they expect to be denied. Rather than a decrease in the demand for credit, a rise in "NA" answers could then be due to an expected decrease in credit supply.[6] For this reason, running the analysis using both series and comparing the results obtained might shed interesting insights.

Aggregate loan-approval index. To extrapolate the small-business approval rates to the whole economy, an estimate of the fractions of matches corresponding to small and large firms is necessary. To do so, I use two approximations. First, by lack of a better estimate, the loan-approval rate for large firms is assumed to be 1. Second, the share of total hires due to small firms is set at 60%.[7] The implications of these assumptions are discussed later. We can now write:

$$hires_t = p_t \cdot matches_t = p_{s,t} \cdot matches_{s,t} + 1 \cdot matches_{b,t}$$

[6]Using the Kauffman Firm Survey, Zarutskie and Yang (2016) highlight that compared to 2007, between 2008 and 2010, young firms were between 4 to 5 percentage point more likely to not apply for loans because they anticipated being denied. These same firms were then more likely subsequently have more employees, and owners working longer hours. These facts support the hypothesis of a drop in applications due to a shock to credit supply.

[7]Historical JOLTS data, Mills and McCarthy (2014)

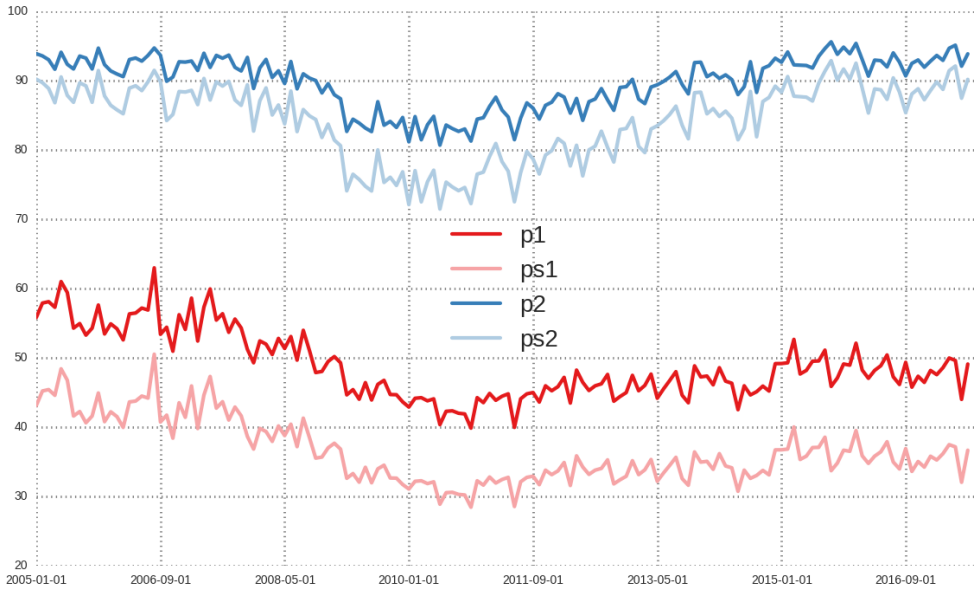

Figure 3.6: Indices of loan approval. Light shades correspond to small firms only, dark shades correspond to the whole economy. p_1 and $p_{s,1}$ are built including "NA" answers, p_2 and $p_{s,2}$ discarding them.

and

$$p_{s,t} \cdot matches_{s,t} = 0.60 \cdot hires_t,$$

where p_s is the loan approval rate for small firms, the loan approval rate for large firms is set to one, and $matches_{s/b,t}$ are the matches corresponding to small/big firms. Using these two equations, we easily get the aggregate series $p_t = p_{s,t}/(0.4p_{s,t} + 0.6)$, plotted in Figure 3.6 in darker shades. Recall that for simplicity, the contribution of small firms to hiring was set as constant over time. This could lead to an upward bias on the index if the contribution of small firms actually decreased during the Great Recession. However, this effect is mitigated by a conservative choice of 1 for the loan-approval rate faced by big firms. This issue could be solved by using JOLTS data on hiring at the firm-size level, in which case the share of hiring attributable to small firms could be set accurately over time. How does the index

compare with other measures of loan-finding success? In the 2015 Small Business Credit Survey, 62% of small firms declared having received all or most of their requested financing. In 2014, 51% of them did. These numbers are much closer in magnitude to the corresponding yearly average loan-approval rates given by $p_{s,1}$ (34% and 37%) than to $p_{s,2}$ (85% and 90%). This suggests that exclusively taking into account the "yes" answer from the NFIB survey, as in p_2, may overestimate firms' ability to obtain lending.

3.6.2 Counterfactual Beveridge curve

Counterfactual vacancies series. In steady-state, $s(L_t - U_t) = p_t M(U_t, V_t)$ must hold. Because the matching function displays constant returns to scale, we can write $s(1 - u_t) = p_t M(u_t, v_t)$. Taking the log of both sides and differentiating with respected to p, keeping u constant, yields

$$\frac{dlnv}{dlnp}\Big|_{u,\dot{u}=0} = -\frac{1}{1-\eta}, \tag{3.46}$$

with η the elasticity of the labor market matching function with respect to vacancies. Following Elsby et al. (2015), I can now derive a p-constant counterfactual vacancies series by netting the shifts implied by (3.46) from the realized vacancies series:

$$\tilde{v_t} = v_t \left(\frac{p_t}{p_0}\right)^{\frac{1}{1-\eta}}. \tag{3.47}$$

The initial value for the aggregate loan-approval rate, p_0, is set to January 2005. In light of Petrongolo and Pissarides (2001), η is set to 0.5. The counterfactual vacancies series $\tilde{v}$ are represented in Figure 3.7. Figure 3.8 represents the counterfactual Beveridge curves. Note these are not the steady-state vacancy-unemployment pairs that would have emerged in equilibrium absent fluctuations in the loan-approval rate. Instead, they represent the path of vacancies derived by imposing a constant availability of credit, and by taking the

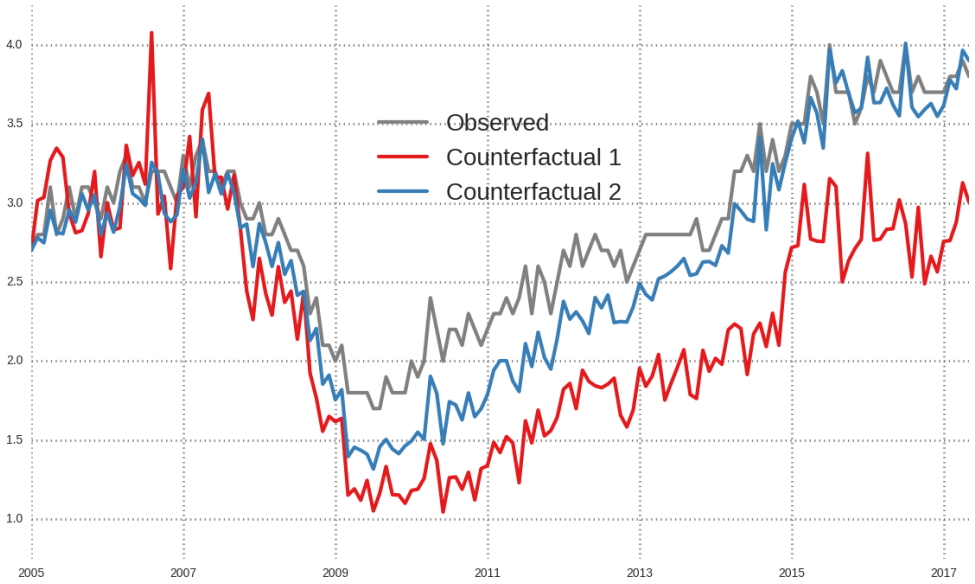

Figure 3.7: p-constant counterfactual vacancy series.

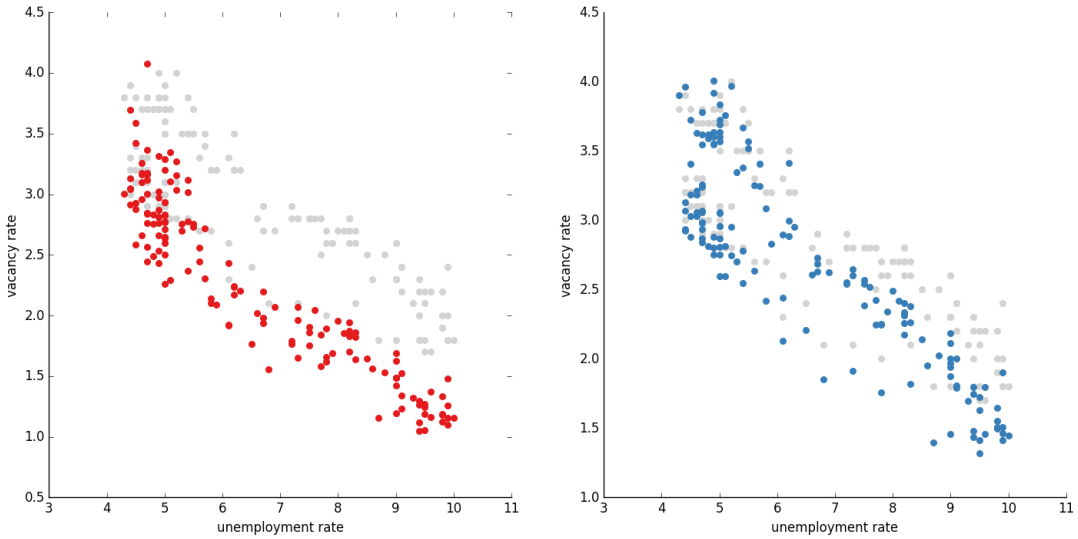

Figure 3.8: p-constant counterfactual Beveridge curve in color, observed Beveridge curve in gray. The left panel uses p_1, the right panel uses p_2.

path of the unemployment rate as given. The observed outwards shift almost disappears on the left panel, in which the counterfactual vacancies series is derived from p_1 (that is, the index that includes "NA" responses). In the right panel, however, the outwards shifts seems much less alleviated, and therefore the role of credit availability of lesser magnitude. This confirms that more work ought to be done to understand better whether the decrease in the share of respondents applying to credit is due to a lack of investments opportunities, or to self-selection.

Contribution to the BC shift. It is interesting to compare the vertical shifts undergone by the observed and counterfactual Beveridge curves for dates that share unemployment rates. For example, the unemployment rate was 7.3% both in December 2008, starting point of the Beveridge curve shift, and in August 2013. However, the vacancy rate was 0.7 percentage points higher in August 2013. Using p_1, the change in loan approval could explain 0.4 percentage points of that increase, that is, around 40%. Using p_2, however, none of the shift could be explained by changes in credit availability. Carrying out the same exercise for February 2009 and January 2012, we get that credit can explain around 69% of the shift using p_1, and around 7% using p_2. While these estimates are imprecise, they point in the direction that credit availability may account for a significant fraction of the Beveridge curve shift. Constructing a better index of loan availability would help evaluate the contribution of credit more precisely.

3.7 Conclusion

In this paper, I add to the growing literature on Beveridge curve shifters by proposing credit frictions as a novel channel. This focus follows the path opened by recent contributions, who call for studying more closely the role played by recruiting efficiency in the context of the

Great Recession shift. By changing the timing of credit frictions in Wasmer and Weil (2004), I show that the effective vacancy yield of a firm can be negatively impacted by productivity shocks when firms are credit-constrained at the time of turning a match into a hire.

The key mechanism in this paper is the existence of an event that occurs after a firm and a worker have met, and can, given certain conditions, prevent the hire. Under this perspective, one can easily think of other stories worthy to explore. For example, firm and worker may discover the productivity of their match only after they meet, in which case the hire would only occur if the realized productivity is above some reservation threshold. This idea of a "stochastic matching" is developed in chapter 6 of Pissarides (2000). More anecdotally, the rise in the use of opioid-based pain medication among the American prime working age population, which has received a lot of media attention recently, could be grounds for another story.[8] Interviewed in the New York Times, a business owner claims that he cannot fill his vacancies, as "at least 25 percent [of adequate candidates] fail the drug tests." Such anecdotal evidence is also reported in the Federal Reserve Board's 2017 Beige Book. While the workers fit the job requirements, positive drug testing prevent the firms from hiring, which results in lower vacancy yields.

Disentangling these different stories and assessing which are most relevant requires more empirical work. In the context of this paper, constructing an index of loan-availability more tightly related to the process of hiring could help the precision of the estimates obtained in the empirical exercise carried out in section 6. This would allow to better identify the actual contribution of the credit channel to the shift of Beveridge curve observed after the recession.

[8]See Krueger (2017).

Chapter 4

Social Engagement and the Spread of Infectious Diseases

4.1 Introduction

This paper investigates the transmission of an infectious disease in a random matching model where economic and social gains from trade directly stem from person-to-person contacts. In typical epidemiological models of disease transmission, infection dynamics are mechanically driven by exogenous, reduced-form parameters meant to encompass both the fundamentals of the disease and individuals' behaviors. Modeling individuals as rational and forward-looking agents, who face a trade-off between the benefits from engaging in social interactions and the infection risk it carries, allows us to endogenize the infection rate by letting agents' behavior react over time as the epidemic develops.

I consider two response margins, both of which have been at the forefront of public health recommendations since the onset of the COVID-19 pandemic: social engagement and mask-

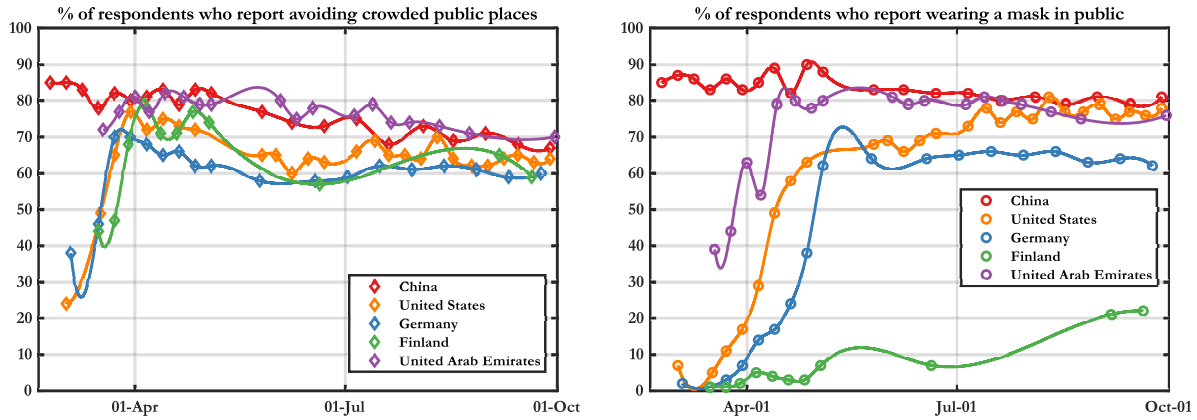

Figure 4.1: Protective behaviors in a sample of countries during the COVID-19 pandemic. Left Panel - Percentage of survey respondents who report avoiding crowded places. Right panel - Percentage of survey respondents who report wearing a mask in public. Source: YouGov.

wearing.[1] Figure 4.1 shows that since the beginning of 2020, and throughout the world, a significant share of individuals have taken action along both of those margins. While policy mandates have certainly largely contributed to those reactions, they do not account for all of it. For example, although Finland never required mask-wearing in public, the right panel shows that by the end of September 2020, more than 20% of the Finns surveyed reported taking that precaution.[2]

The paper aims to answer the following questions. First, when and to which extent do individuals modify their behaviors in response to the outbreak of a contagious disease? Second, how are epidemic dynamics and economic outcomes impacted by these rational changes in behavior? In particular, can the equilibrium path still be precisely predicted, as is the case in epidemiological models that do not account for endogenous changes in behaviors? Third, how do these dynamics and outcomes differ when agents modulate the

[1]For example, in an interview with ABC News, Dr. Fauci, director of the National Institute of Allergy and Infectious Diseases in the United States, stated that "[the] best way that you can avoid — either acquiring or transmitting infection — is to avoid crowded places, to wear a mask whenever you're outside." The Centers for Disease Control and Prevention (CDC) makes similar recommendations.

[2]Additional evidence can be found in Farboodi et al. (2020), who use data based on cellphone tracking to show that everywhere across the United States individuals started to reduce their social activity before any policy measures were enacted.

frequency of their social engagement relative to when they modulate the precautions they take during social interactions?

To answer these questions, I build a model that retains a structure similar to that of compartmental models of disease transmission, developed in the wake of Kermack and McKendrick (1927). The population is divided into three groups respectively labeled S, I, and R: susceptible individuals, who can contract the disease; infectious individuals, who carry the disease and can transmit it; resistant individuals, who can neither contract nor transmit the disease.

Agents get to engage in bilateral interactions with randomly-chosen partners. Each interaction may or may not generate utility, reflecting the fact that some contacts may not be desired. When a susceptible agent enters in contact with an infectious agent, the former contracts the virus with some probability. The measure of infectious individuals grows from the flow of previously-susceptible individuals newly infected, and shrinks from the flow of individuals who recover from the disease. Infection dynamics are then driven by the effective reproduction number: the number of susceptible individuals expected to contract the disease from the same infectious agent while she is a carrier.

I first analyze equilibrium outcomes when agents are given the opportunity to stay home instead of engaging in social and economic interactions. Second, I assume that agents cannot opt out of the market but can wear a mask, which diminishes the probability of contracting (or transmitting) the virus during social interactions. The two decisions significantly differ in the trade-offs they imply: self-isolation offers absolute safety, but comes at the relatively large opportunity cost of forgoing social engagement; mask-wearing only offers a partial protection and carries an inconvenience cost, but still allows the wearer to engage in social and economic activity.

Throughout the paper, I study two limiting cases regarding the status of agents who recover from the disease. In a first specification, labeled SIS, it is assumed that agents do not gain

immunity, so that they are again susceptible after they recover. Under this assumption, the equilibrium reduces to a system of two ordinary differential equations that can be analyzed with phase diagrams. A second specification where recovered agents become resistant is also studied. This adds a third differential equation to the model, which is then calibrated to US data and solved numerically.[3]

In terms of calibration, while I mostly follow the method developed by Farboodi et al. (2020), I depart from them by making use of micro-founded data to calibrate two important parameters: the meeting rate and the transmissibility of the virus during an interaction. The former is calibrated using survey studies that document the number of interpersonal contacts experienced by respondents on a daily basis. The latter is calibrated following contact tracing studies, which track the contacts of individuals who have tested positive to the virus, and record whether those contacts, who have been asked to isolate, contracted the virus.

A first important result is that in a world where utility is directly derived from contacts between individuals, there exist complementarities between the participation decisions of susceptible agents. Since matching among agents is random, the risk of infection in a given contact depends on the composition of the pool of participants. More specifically, as infectious and resistant agents always participate, the probability of a susceptible agent contracting the virus in a given contact decreases the more susceptible peers participate. Equilibrium participation then becomes the outcome of a game between susceptible agents, whereby multiple Nash equilibria may coexist. For example, it could be rational for a susceptible individual to participate if all other susceptible agents participate, and to stay home if all other susceptible agents stay home. This gives rise to adverse selection: the fewer

[3]As of October 2020, while there is still disagreement regarding a definitive immunity in individuals who have recovered from COVID-19, there exists a body of evidence pointing at temporary immunity for the majority of cases. See references in https://www.nytimes.com/2020/08/16/health/coronavirus-immunity-antibodies.html, retrived on October 5, 2020.

agents not carrying the virus participate, the higher the prevalence of infection in the pool of participants (i.e., the lower the "quality" of the pool), and the lower the net benefit of participating, which further drives non-carriers out of the market and increases the prevalence of the disease among participants.

The complementarities between the decisions of susceptible agents translate to a multiplicity of equilibrium paths in both the SIS and SIR specifications as long as the cost suffered by infected agents is in a medium range, a condition satisfied in the calibrated model. I restrict my attention to classes of equilibria that satisfy some specified coordination rules for each instant along the equilibrium path where multiple Nash equilibria coexist.[4] I first consider two extreme rules, where susceptible agents either always coordinate to participate or to stay home whenever both could hold in equilibrium. The paths obtained provide bounds to all other equilibrium paths in the (S, I) phase plane. Following either rule, the infection curve is considerably flatter than in the benchmark case with no behavioral response, where the infection curve reaches a peak with around 40% of the population infected at the height of the epidemic. When agents coordinate to go out and engage in interactions whenever possible, the measure of infected agents never surpasses 7% of the population. At the other extreme, when agents coordinate on staying home as much as possible, it never goes past 2%.

I then consider other coordination rules. One specifies that in the multiplicity region, susceptible agents coordinate on going out with probability $x \in (0, 1)$. The lower x, the flatter the infection curve and the more delayed the development of the epidemic. Another rule is based on the idea that individuals may be more likely to coordinate on staying home when the epidemic seems more severe, so that coordination is determined by comparing the number of active cases to a set threshold. Using this rule, infection curves feature plateaus. The last rule I impose has to do with "isolation fatigue." After coordinating on the safe behavior of

[4]These can be seen as restricting agents' beliefs.

staying home for a long time, susceptible agents may get fatigued and switch to coordinate on participating. This coordination rule allows the equilibrium path to feature multiple waves of infections. These results highlight that countries or regions with similar fundamentals can still experience significantly different infection dynamics, driven by equilibrium beliefs.

Interestingly, across all equilibrium paths explored, as time goes to infinity, a relatively similar measure of agents will have been infected—between 78% and 80% of the population. In comparison, in the benchmark model, 96% of the population would eventually have been infected. A policy implication is that the shape of the infection curve is not necessarily, in itself, a good measure of how well a population is faring in terms of long-term outcomes: widely different shapes could eventually lead to similar steady states. Coordination is nevertheless extremely relevant when it comes to welfare, as the economic and social costs of forgone social contacts vary largely across equilibria.

When agents cannot opt out of the market but can decide to wear a mask, equilibrium analysis is much simplified. First, infectious agents have no incentives to take this costly precaution. Second, there are no complementarities (in a static sense) between the decisions of different susceptible agents: taking as given the future course of the epidemic, the net benefit of wearing a mask for a given susceptible agent is independent of the behavior of other susceptible peers. As a result, the equilibrium is unique. In the SIR simulations, the cost of wearing a mask is calibrated as a fixed percent of the utility received by agents when they engage in interactions. For reasonable calibrations, the equilibrium path is such that susceptible agents do wear a mask once the epidemic has gained enough ground, and they stop doing so once it has sufficiently subsided. As expected, the costlier the masks, the shorter the amount of time during which they are worn. The infection curve is again considerably flattened, and a bit delayed.

Across all specifications, the number of active cases never gets past 12%. The cumulative

measure of agents that has been infected by the end of the epidemic remains between 72% and 75% for the different cost specifications. In terms of welfare, the mask-wearing margin yields better outcomes than the participation margin. Not only is the long-run cumulative number of cases even lower than for the participation model, the associated costs are extremely low. This translates to a total welfare loss between 5.7 and 6.1 trillion dollars for the model with mask-wearing, compared to 7.6 trillion dollars in the benchmark with no behavior response.

Relation to the literature A large body of economic literature aimed at endogenizing the dynamics predicted by epidemiological models quickly developed in the wake of the COVID-19 outbreak.[5] This paper is most-closely related to a subset of those papers, which endogenize individual-level participation in a SIR model, using forward-looking rational agents who maximize their lifetime utility: Bethune and Korinek (2020), Farboodi et al. (2020), Garibaldi et al. (2020), McAdams (2020b), and Toxvaerd (2020).[6]

In those five papers, and different from the present paper, an agent's utility is not directly derived from each social interaction but from her level of "social activity" (a continuous variable). Additionally, in all but McAdams (2020b), that utility is independent of the "social activity" of other agents. The main implication is that there are no complementarities between the participation decisions of different agents in the economy, which are essential to generate the infection dynamics obtained in the present paper. Indeed, in my paper, because utility stems from each individual contact, it inherently requires meeting other agents, and thus directly depends on the participation of other agents. Similar complementarities are highlighted by McAdams (2020b), written concurrently. In that paper, utility is specified as depending on the participation of other agents in reduced form, with a utility function

[5]There did already exist a small economic literature related to infectious disease, spurred by the HIV outbreak in the 1990s. McAdams (2020a) provides a comprehensive review of economic epidemiology, with a focus on recent developments but also going back to those seminal papers.

[6]Note that Bethune and Korinek (2020) also studies an SIS specification. Other papers that endogenize economic activity in a SIR model at a more aggregate level include for example Eichenbaum et al. (2020) and Krueger et al. (2020).

that increases in aggregate participation. In other words, complementarities in participation decisions are directly built in. Both models predict equilibrium multiplicity, and highlight the role of coordination. What differs is that while McAdams focuses on the theory, I calibrate the model and explore the form that multiplicity takes when applied to COVID-19 in the US. I simulate paths at the two extremes of participation, when agents participate as much and as least as possible, and quantify the corresponding range of human and economic costs. I also explore additional coordination rules and show how they can give rise to infection dynamics such as plateaus and multiples waves, which can be observed in the data but are absent from Bethune and Korinek (2020) and Farboodi et al. (2020), two models that are also calibrated to the COVID-19 epidemic in the US.

The information structure also differs across the aforementioned papers. In the first two, it is assumed that agents do not know whether they are susceptible or infectious. In Garibaldi et al. (2020) and Toxvaerd (2020), like in my paper, agents know their status. In McAdams (2020b), agents who contract the virus originally do not know it, but may eventually learn it. When applied to the COVID-19 pandemic, the latter specification certainly seems the most appropriate, as 40% of infections are asymptomatic Oran and Topol (2020)). Due to the structure of my model, however, uncertainty would require keeping track of the distribution of beliefs over time, depending on the exact history of matches encountered, a challenging problem left for future work.

The second part of the paper focuses on mask-wearing, a decision absent from the five papers mentioned above. This decision can be seen as a specific example of behavioral reaction along a "vigilance" margin, which is for example present in Engle et al. (2020). Vigilance is costly, but decreases one's risk of infection (as well as others'). A major difference is that in Engle et al. (2020), agents are partially myopic: they maximize an objective function that only depends on the state of the epidemic at that time. In the present paper, agents are perfectly forward-looking. Salanié and Treich (2020) combine the two margins studied in this paper,

mask-wearing and isolation, and focus on the impact of a mandatory policy related to mask-wearing on agents' isolation behavior, however in a static setting.

The adverse selection that occurs in the model with participation is similar to a mechanism highlighted by Kremer (1996) in the context of HIV, where participation decisions impact not only the number of matches but also the composition of the pool: as more individuals choose abstinence, the prevalence of infection may increase.

The rest of the paper is organized as follows. Section 4.2 presents the model environment. Section 4.3 describes the infection dynamics predicted by standard SIS and SIR models absent the participation and mask-wearing margins, thus serving as a benchmark against which to compare subsequent results: outcomes when adding the participation margin, in Section 4.4, and outcomes when adding the mask-wearing margin, in Section 4.5.

4.2 Environment

Time t is continuous and goes on forever. The economy is populated by a measure P of infinitely-lived agents who discount the future at rate $r > 0$. At all points in time, agents can choose to engage in a meeting process. Meetings, or "social contacts," are bilateral and occur at random with a Poisson arrival rate $\alpha(N)$, where $N \leq P$ represents the measure of participating agents. When an agent enters in contact with another agent, she enjoys $y > 0$ utils with probability p and 0 otherwise—not all meetings may be desirable.[7]

We consider the existence of a virus that can spread in the population. Agents can be in one of three states: susceptible, infected, or resistant.[8] The measures of agents in each state are respectively denoted $S(t)$, $I(t)$, and $R(t)$, with $S(0)$, $I(0)$ and $R(0)$ taken as given. An

[7] Garibaldi et al. (2020) highlight the existence of those "unintended contacts" and explain their relation to the matching technology.

[8] In this model, being infected and infectious are strictly equivalent.

agent's state is private information, exclusively known by that agent.

Agents who participate to the meeting process can take a protective measure, interpreted as wearing a mask, at a flow cost $k > 0$. The decision is made before any contact is realized. Upon contact with an infectious agent j, a susceptible agent i becomes infectious with probability $\tau^{ij} \in (0, 1)$, with $\{i, j\} \in \{m, n\}^2$. The superscripts denote whether each agent in the meeting is wearing a mask (m) or not (n). It is assumed that $\tau^{mm} < \tau^{nm} < \tau^{mn} < \tau^{nn}$. While masks are most effective when both agents wear them, the second best occurs when only the infected agent wears one.

Infectious agents, who recover at Poisson rate $\gamma > 0$, suffer a flow cost $\psi > 0$ as long as they are infected. The flow cost of infection represents both the direct, physical cost of being sick, and the indirect cost associated with the prospect of dying from the disease.[9]

Throughout the paper, I study and compare two assumptions regarding the status of agents who recover from the virus. In one case, it is assumed that they gain immunity and become resistant, an absorbing state. In the other case, it is assumed that they transition back to being susceptible. The first case will be referred to as the SIR model and the second as the SIS model.

Preliminary results: equilibrium in a virus-free economy When $I(t) = 0$, the population is free from the virus. In this case, there are no incentives to refrain from participating in the meeting process, and no incentives to wear a mask during those meetings. The lifetime discounted utility of an agent in state j, where $j \in \{S, I, R\}$, is denoted V_j and is given by the Hamilton-Jacobi-Bellman (HJB) equation,

$$rV_j(t) = \alpha(P)\tilde{y} + \dot{V}_j(t), \tag{4.1}$$

[9]Note that death is not formally modeled otherwise. Agents always recover at time goes to infinity, so that the population remains constant.

where the dot represents a time derivative. Because the whole population participates to the meeting process, $N = P$, and agents meet at Poisson arrival rate $\alpha(P)$. Each meeting provides an expected benefit of $\tilde{y} \equiv py$ utils. The equilibrium path is such that V^j is constant and equal to the present value of all future meetings, $\alpha(P)\tilde{y}/r$.

4.3 A quick primer on epidemiological models

It will be helpful to first review the dynamics of the model when neither the participation nor the mask-wearing decision are taken into account, which will later serve as a benchmark. As such, the model looks like an off-the-shelves SIS/SIR model, where the number of participants is equal to the population size, $N = P$, and the transmission probability corresponds to that when neither agent wears a mask, τ^{nn}. At any point in time, agents are either susceptible, infectious, or resistant, so that the following identity must hold,

$$P = S + I + R, \tag{4.2}$$

where, to simplify the exposition, I suppressed the explicit dependence of S, I and R on time. The measure of infected agents evolves according to

$$\dot{I} = \alpha(P)\tau^{nn} S \frac{I}{P} - \gamma I. \tag{4.3}$$

The first term on the right-hand side corresponds to the inflow of susceptible agents newly infected. S agents enters in contact with $\alpha(P)$ other agents, a proportion I/P of which is infected. For each of these contacts with infectious agents, the probability for the susceptible agent to catch the virus is τ^{nn}. The second term on the right-hand side corresponds to the outflow of previously-infected agents that recover. In a SIS model, $\dot{R} = 0$, so that R is constant. Assuming that the original stock of resistant individuals is null, in the SIS

specification, $R = R(0) = 0$. The law of motion for S is equal to the negative of (4.3). In a SIR model, $\dot{R} = \gamma I$ and $\dot{S} = -\alpha(P)\tau^{nn}SI/P$.

We denote $\sigma \equiv \alpha(P)\tau^{nn}/\gamma$ the basic reproduction number. This number is often referred to as $\mathcal{R}_0$, but the notation σ is preferred here to avoid any confusion with the initial measure of resistant agents, $R(0)$. It corresponds to the number of people that an infectious individual would be expected to infect before recovering, *assuming that the whole population is susceptible*. It has to be distinguished from the effective reproduction number, $\sigma_e(t) \equiv \sigma S(t)/P$, a time-dependent variable which measures the number of people an infected individual would be expected to infect *given the actual measure of susceptible agents in the population at time* t.

Steady states In the SIS and SIR models, the system is at steady state when $\dot{I} = \dot{S} = \dot{R} = 0$. Plugging $\dot{I} = I = 0$ into (4.3), we directly obtain that in the SIS model, there always exists a virus-free steady state, with $I^* = 0$, $S^* = P$ and $R^* = 0$. There also exists an endemic steady state, with $I^* > 0$, as long as $\sigma > 1$. It is such that $I^* = P\left[1 - 1/\sigma\right] > 0$, $S^* = P/\sigma$ and $R^* = 0$. In this steady state, the flow of new infections exactly offsets the flow of recovered agents at each instant. In other words, in an endemic steady state, the effective reproduction number must be exactly equal to one. Because the effective reproduction number is never greater than the basic reproduction number, this requires the basic reproduction to be greater than one.[10]

In contrast, in the SIR model, there exists no endemic steady state. Intuitively, the SIR system cannot be at steady state as long as $I > 0$, as there would be a strictly positive increase in the measure of recovered individuals at each instant. As a result, in the SIR model, steady state is only achieved when the virus has entirely been eradicated, $I^* = 0$.[11]

[10] Mathematically, for $\sigma_e = \sigma S/P = 1$ to hold given $S/P \in [0, 1)$, we need $\sigma > 1$.

[11] This result can be overturned by introducing birth and death dynamics to the system, in which case an endemic steady state can be sustained.

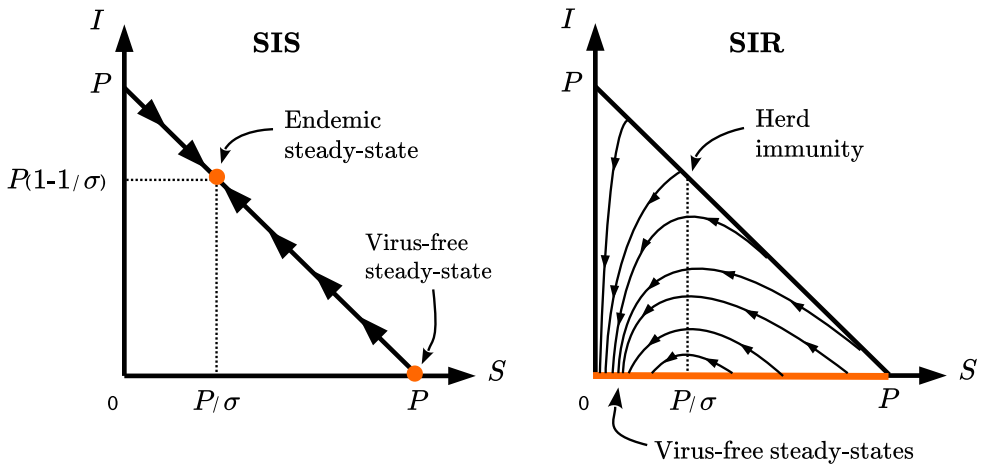

Figure 4.2: Dynamics of the SIS and SIR model with no participation nor mask-wearing decision, where $\sigma \equiv \alpha(P)\tau^{nn}/\gamma > 1$ is the basic reproduction number.

There exists a continuum of such virus-free steady states, indexed by $R^* \in [0, P]$, with $S^* = P - R^*$. In sum, any combination of S^* and I^* adding up to the total population can be sustained as a steady state.

Dynamics The dynamics of the SIS and SIR systems can be represented in phase diagrams. They are drawn in Figure 4.2 with S on the x-axis and I on the y-axis.

The left panel represents the dynamics under the SIS specification when $\sigma > 1$. Because we assumed $R = 0$, the pair (I, S) must always be located on the hypotenuse of the triangle. The number of infections increases (decreases) as long as the effective reproduction rate number is greater (smaller) than one, or equivalently, $S > (<) P/\sigma$. Once S reaches the threshold P/σ, the system has reached the endemic steady state. As a result, when the endemic steady state exists, it is globally stable. Starting from any arbitrarily low $I(0)$, the infection grows until it reaches the steady state. When the endemic steady state does not exist, i.e., when $\sigma < 1$, the effective reproduction number is mechanically smaller than one, and the virus-free steady state is globally stable. Starting from any $I(0)$ the number of infections decreases until the virus is eradicated.

The dynamics under the SIR specification are represented in the right panel. Whether the measure of infected agents increases or decreases is still determined by the effective reproduction number: I increases (decreases) when $S > (<) P/\sigma$. However, S can only decrease, as contrary to the SIS model, there is no inflow of previously-infected susceptible agents. As a result, the effective reproduction number is strictly decreasing over time and must eventually drop below one regardless of its initial value. At this point, we say that the population has reached "herd immunity," i.e., there are not enough susceptible individuals left for each infected individual to infect more than one susceptible agent before recovering. Starting from an arbitrarily low $I(0)$ and along the hypotenuse ($R = 0$), the initial effective reproduction number must be greater than one, so that the stock of infected first increases. Once herd immunity has been reached, the measure of infected agents then steadily decreases until the system reaches the (virus-free) steady state. Steady states with $S > P/\sigma$ are not locally stable, and therefore could never be reached starting from any $I(0) > 0$, because the effective reproduction number as we approach $I = 0$ would be greater than one. Equivalently, only steady states such that $R^* \in [P(1 - 1/\sigma), P]$ can be reached. Because R^* indicates the aggregate measure of agents that have been infected during the epidemic, it means that at the minimum, a share $(1 - 1/\sigma)$ of the population would be infected over the course of the epidemic. Finally, it can be shown that R^* and S^* are uniquely determined by $I(0)$. The higher $I(0)$, the higher R^* and the lower S^*.

In the remaining of the paper, we relax the assumption that the whole population seeks to engage in social interactions at all times and allow participants to wear masks, so as to study the implications of those endogenous responses on both the dynamics of the epidemic and long-run outcomes. We will focus on the parameter region such that the basic reproduction number, σ, is greater than one: absent any behavioral response, the virus would become endemic in the SIS model, and starting from low enough $I(0)$, there would be an epidemic in the SIR model.

4.4 To go or not to go

In this section we focus on the participation decision of agents in the market. It is assumed that there is no protective measure, i.e., agents do not wear masks, and we let $\tau \equiv \tau^{nn}$ to simplify the notation. We first set up the SIS model and derive a few key analytical insights in Section 4.4.1, before moving on to the SIR model in Section 4.4.2. The SIR model is then calibrated and solved numerically in Sections 4.4.3 and 4.4.4.

4.4.1 SIS model

The HJB for an infectious agent is

$$rV_I = \alpha(N)\tilde{y} - \psi + \gamma(V_S - V_I) + \dot{V}_I, \tag{4.4}$$

where V_I is the expected lifetime discounted utility of being infectious and V_S that of being susceptible. The first term makes use of the equilibrium result that infectious agents would always choose to participate to the meeting process. While they could choose to stay home, it would never be rational for them to do so, as it carries an opportunity cost but no benefit.[12] The cost of being infected is captured by the second term, while the third term captures the potential upside of recovery. The HJB equation for susceptible agents is

$$rV_S = \max\left\{0, \alpha(N)\left[\tilde{y} + \tau\frac{I}{N}(V_I - V_S)\right]\right\} + \dot{V}_S. \tag{4.5}$$

[12]This would be different if preferences favored altruism, e.g., infected individuals could suffer a cost from endangering or infecting susceptible individuals, or if infectious individuals did not know their own state, in which case they may also fear getting infected. Additionally, for completeness, note that it could in fact be rational for an infectious agent to stay home if all other agents stay home as well. Going out would bring no benefit, so that the infectious agent would be indifferent. This is a pure coordination problem and we will ignore this type of equilibrium in the remaining of the paper.

The maximization represents the participation decision of the susceptible agent. She can decide to stay home, in which case her utility is normalized to zero, or she can decide to go out, in which case she gets utility from a fraction of the social contacts she will encounter, but also faces the downside of potentially getting infected. Due to random matching and full participation from infectious agents, the probability of a partner being infectious is I/N. Conditional on meeting an infectious agent, the probability of contracting the virus is τ.

Equations (4.5) and (4.4) can then be combined to obtain a single differential equation in $\omega \equiv V_S - V_I$,

$$\dot{\omega} = (r+\gamma)\,\omega + \alpha(S^p + I)\min\left\{\tilde{y}, \tau\frac{I}{N}\omega\right\} - \psi, \tag{4.6}$$

where S^p denotes the measure of susceptible agents who participate. A second differential equation comes from the law of motion for the measure of infected individuals,

$$\dot{I} = \alpha(N)\tau S^p\frac{I}{N} - \gamma I. \tag{4.7}$$

It is almost identical to (4.3), derived in the benchmark SIS model. The difference is that is now depends on the measure of participating susceptible agents, S^p, rather than the measure of susceptible agents, S (both directly and through N).

To close the model, we need to solve for the aggregate participation of susceptible agents, S^p. To do so, we first solve for the participation decision problem of a single susceptible agent, j. She engages in social interactions with probability a_j, given by

$$a_j\begin{cases} = 0 & < \\ \in [0,1] \text{ if } \tilde{y} & = \ \tau\frac{I}{N}\omega, \\ = 1 & > \end{cases} \tag{4.8}$$

where the aggregate participation of the whole population, N, is given by

$$N = S^p + I^p = \int_{i \in S} a_i di + I. \tag{4.9}$$

To decide whether to participate, the susceptible agent compares the expected utility from a social contact, $\tilde{y}$, to the expected chance of receiving the virus multiplied by the cost of getting infected, $\tau(I/N)\omega$. Because $N = S^p + I$, the chance of contracting the virus during a social contact depends on the participation decision of all other susceptible agents as well. Ceteris paribus, a higher number of susceptible agents shifts the composition of the pool of participants in such a way that I/N decreases, thereby reducing the riskiness of the pool and making any given interaction safer. In other words, an increase in the number of susceptible agents participating, keeping the number of infected participants the same, reduces the marginal cost of a contact without impacting its marginal benefit. This generates complementarities between the participation decisions of susceptible agents: more participation from susceptible agents encourages the participation of other susceptible agents. Note that this feedback loop is independent of the assumptions made regarding the matching technology. In particular, it does not require increasing returns to scale.

Solving for S^p for a given pair (I, ω) is now akin to solving a Nash equilibrium, whereby the participation decision of each individual, driven by (4.8), must be the best response given the participation decisions of all other individuals. There are two dominance regions and one multiplicity region. When $\tilde{y} > \tau\omega$, there exists a unique Nash equilibrium, where all susceptible individuals participate, $S^p(I, \omega) = S$. Intuitively, in this region, the expected utility from any given social contact would be higher than the expected cost of infection even if all participants were infectious, so that the only possible equilibrium outcome is for everyone to participate. When $\tilde{y} < \tau(I/P)\omega$, there also exists a unique Nash equilibrium, although with no participation from susceptible individuals, $S^p(I, \omega) = 0$. In this region, the expected utility earned from a contact would be lower than the expected cost of infection,

even if everyone were to participate, so that it can never be worth it for susceptible agents to participate. When $\tau(I/P)\omega \leq \tilde{y} \leq \tau\omega$, there exist multiple Nash equilibria: the two previous corner equilibria, as well as an equilibrium with partial participation, $S^p(I,\omega) = (\tau\omega/y - 1)\, I$. By construction, in the latter type of equilibrium, susceptible agents are indifferent between staying in or going out.

Definition 4.1 (Equilibrium definition)**.** *An equilibrium consists in a list of time paths for the two state variables* $\{S(t), I(t)\}$ *and the two control variables* $\{S^p(t), \omega(t)\}$*, such that* (4.6) *and* (4.7) *are satisfied, where* S^p *is a Nash equilibrium consistent with the individual participation decision rule given by* (4.8) *and* (4.9)*,* $S(t) = P - I(t)$ *and* $I(0)$ *is given.*

Steady states There always exists a virus-free steady state, where, by definition, $I^* = 0$ and $S^* = P$. It features full participation, $S^{p*} = N = S$, and is therefore identical to the virus-free steady state described in the benchmark SIS model. Because the virus is eradicated, there is no risk in participating. As long as other people are participating, it is therefore worthwhile to join.[13] In this steady state, $\omega^* = \psi/(r + \gamma)$.

There may also exist endemic steady states, which require $\alpha(P)\tau S^{p*}/(S^{p*} + I^*) = \gamma$. Now denoting the basic reproduction number $\sigma^p(S^p) \equiv \alpha(S^P + I)\tau/\gamma$ and the effective reproduction number $\sigma_e^p(S^p) \equiv \sigma^p(S^P)S^p/(S^p + I)$, this condition can be rewritten as $\sigma_e^p(S^p) = 1$. As was the case for the standard SIS model, the effective reproduction number must be equal to one for the system to be in a endemic steady state. What differs is that the effective reproduction number is now a jump variable and an increasing function of the participation of susceptible agents.

We obtain that $S^p = 0$ cannot hold in an endemic steady state, since the effective reproduction number would be equal to zero. An endemic steady state with full participation,

[13] As mentioned earlier, there could also be an equilibrium with no participation at all, $S^{p*} = N^* = 0$, but that would purely due to a coordination problem, not to the virus. In this scenario, susceptible agents would be indifferent between participating or not, so we assume that they do participate.

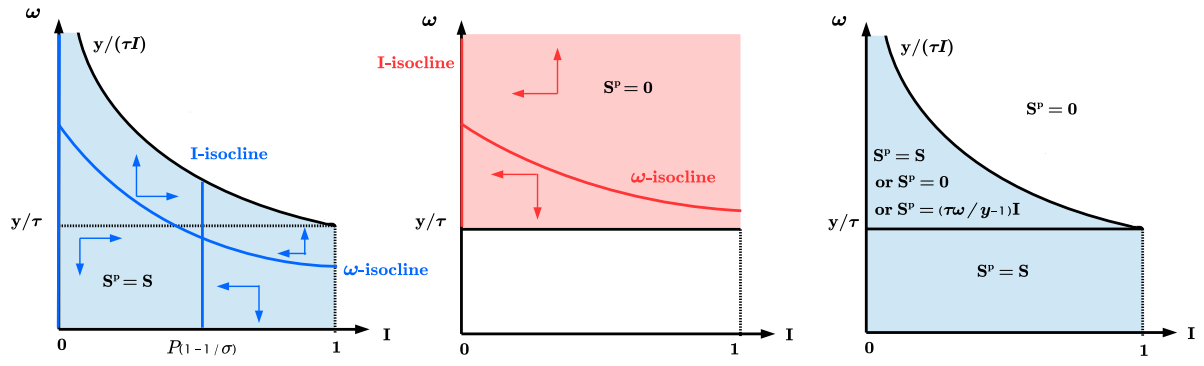

Figure 4.3: Construction of the phase diagram in the SIS model with participation.

$S^{p*} = S$, implies $S^* = P/\sigma$, $I^* = P(1 - 1/\sigma)$ and $N^* = P$. This steady state is identical to the endemic steady state in the standard SIS model. Note that $\omega^* = \psi/[r + \alpha(P)\tau]$, so that this steady state exists as long as $\tilde{y} \geq \tau \{1 - \gamma/[\alpha(P)\tau]\} \psi/[r + \alpha(P)\tau]$, i.e., the cost of infection ψ and the transmission probability τ are not too high relative to the expected utility from a social contact, $\tilde{y}$. Finally, there may also exist an endemic steady state with partial participation, $S^{p*} = (\tau\omega^*/\tilde{y} - 1)I^*$, as long as the cost of infection, ψ, is not too high nor too low (although it requires $\alpha'(.) > 0$).

Dynamics We can now study the dynamics of the SIS model with participation in a phase diagram. Figure 4.3 displays its construction, with I on the x-axis and ω on the y-axis (the second state variable, S, can be obtained from the identity $S + I = P$).

As depicted in the right panel, three regions can be delineated. In the blue region, $\tilde{y} > \tau(I/P)\omega$, so that there exists a Nash equilibrium with full participation. The blue arrows in the left panel depict the direction of motion in that region under full participation, while the blue curves represent the isoclines, $\dot{I} = 0$ and $\dot{\omega} = 0$. The virus-free steady state, denoted by the blue dot on the y-axis, is unstable. This is intuitive: with full participation, the model is akin to the standard SIS model, and we saw that in a standard SIS model the virus-free steady state is unstable as long as the basic reproduction number is greater than one, which

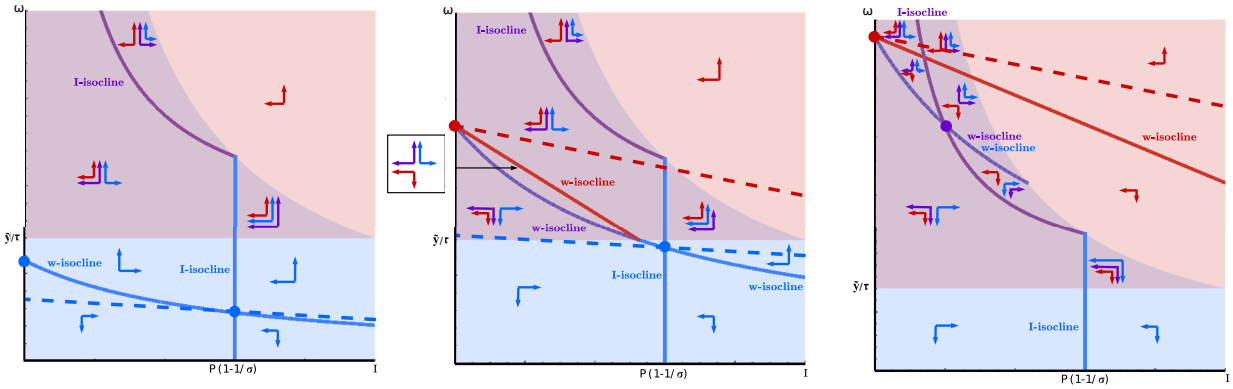

Figure 4.4: Dynamics of the SIS model with participation for three levels of the flow cost of infection, ψ. Left panel: low ψ. Middle panel: medium ψ. Right panel: high ψ.

we assume. The endemic steady state, denoted by the blue circle at the intersection of the isoclines in the center of the region, is saddle-path stable. The saddle path is represented by the dashed blue curve.

In the red region, $\tilde{y} < \tau\omega$, so that there exists a Nash equilibrium with no participation. The red arrows in the middle panel represent the direction of motion in this region with no participation. Intuitively, without participation from susceptible agents, the measure of infectious agents always decreases. The virus-free steady state, represented by the red dot, is saddle-path stable. The saddle path is given by the dash red curve.

In the purple region, located exactly where the blue and red regions overlap, $\tau(I/P)\omega \leq \tilde{y} \leq \tau\omega$, so that the two corner Nash equilibria coexist with the partial participation equilibrium. In this region, the direction of motion therefore depends on whether susceptible agents coordinate on full participation, no participation, or partial participation.

To study equilibrium dynamics, we must combine the three graphs presented in Figure 4.3. The regions in which isoclines intersect, which depends on parameter values, will determine the set of equilibrium paths. Figure 4.4 displays the complete phase diagrams for three different values of the flow cost of infection, ψ.

As represented in the left panel, when the cost is low enough, the endemic steady state features full participation and is the only stable steady state.[14] The equilibrium path is unique and follows the saddle path. Dynamics are then exactly identical to those described under the benchmark model. Because getting sick only carries a small cost, agents do not react by adjusting their participation margin despite having the opportunity to do so.

When the cost of infection is in a medium range, as shown in the middle panel, both an endemic steady state with full participation and a virus-free steady state may be stable, depending on the coordination of susceptible agents. In this parameter region, there exists a large multiplicity of equilibria. For example, agents can still coordinate on participating throughout, in which case the dynamics are identical to those described in the left panel. At the other extreme, susceptible agents can coordinate on staying home as the number of infected asymptotically converges to zero. Agents can also switch coordination along the way, in which case the equilibrium paths for S and I may no longer be monotone.

Finally, when the cost of infection is very high, so that there does not exist an endemic steady state with full participation anymore, the only stable steady state is the virus-free steady state, as visible in the right panel.

Takeaways and discussion We can now highlight a few takeaways and discuss some of the key underlying assumptions.

1) Endogenous, policy-free equilibrium lock-down. When the cost of infection is high enough, the participation of susceptible individuals endogenously drops in equilibrium. This can be interpreted as an endogenous lock-down, which does not stem from any policy interaction but purely from individuals' fear of being infected. Ignoring this margin would then lead to overestimating the rate at which the virus is transmitted. This echoes results from other

[14] More precisely, this occurs as long as the intersection of the blue saddle path with the y-axis occurs for $\omega \leq \tilde{y}/P$.

papers that also add a participation margin to SIR/SIS models, e.g. Bethune and Korinek (2020), Farboodi et al. (2020), Garibaldi et al. (2020), McAdams (2020b), and Toxvaerd (2020).

2) Impact of individual behaviors on the effective reproduction number. In the standard SIS model, the effective reproduction number is a state variable that evolves with the measure of susceptible agents. When allowing for the participation margin, it becomes a jump variable that positively depends on the participation of susceptible individuals through two channels. First, the participation margin impacts the composition of the pool of agents an infectious individual is expected to meet: the lower the participation of susceptible agents, the higher the chance of a random encounter leading to virus transmission. Second, the participation of susceptible agents may also impact the number of meetings if the matching technology does not feature constant returns to scale, i.e., $\alpha'(.) \neq 0$. With increasing returns to scale, a drop in the participation of susceptible agents leads to a smaller arrival rate of meetings, and therefore fewer opportunities for the virus to be transmitted.

3) *Complementarities in participation and social utility.* The participation decisions of susceptible agents are complementary to each other, a phenomenon that is absent from Bethune and Korinek (2020), Farboodi et al. (2020), Garibaldi et al. (2020), and Toxvaerd (2020). In those papers, an increase in the participation of other susceptible agents is irrelevant for the participation decision of a susceptible agent. This difference is rooted in the assumptions made regarding the way individuals receive "social utility." In the three papers just mentioned, social utility is not directly derived from individual contacts, but more so derived from what could be interpreted as "time spent outside." Agents compare the utility from spending a given amount of time outside to the overall probability of getting infected during that time. The former is independent of the number of people outside. The latter increases in the participation of infectious agents, but is independent of the participation of susceptible agents. In the present paper, agents directly gain utility from each contact—contacts are

exactly what provides utility. Therefore, when susceptible agents decide whether to engage in those contacts or to stay home, they compare the expected utility from each given contact to the chance of being infected in each of those contacts. It seems that both specifications may be appropriate for different real-world activities. The former seems to appropriately describe activities such as walking one's dog or going to the theater. The benefits from engaging in those activities do not directly depend on meeting other individuals, but the more people in the street or in the theater, the higher the chance of being infected. Agents would always be at least weakly better off with less participation from other agents. In contrast, the specification used in the present paper may be more appropriate to describe gregarious behaviors that have been at the heart of the COVID-19 spread: family and religious gatherings, parties, etc. In those events, utility directly comes from meeting with people, so that there would be no benefit from engaging in those activities alone. When deciding whether to attend or not, one must compare the utility from being around people to the chance of being infected.[15]

4) *Equilibrium indeterminacy and variety of infection curve shapes.* When the cost of infection is not too high nor too low, the complementarities described above renders the equilibrium indeterminate. Coordination then plays an important role both for short-run dynamics and for long-run outcomes. In the short-run, because the coordination of susceptible agents can switch at any time when in the purple region, we can observe a variety of infection curves, for example featuring one or several peaks, as the number of infected may get up and down successively. This is very different than the dynamics obtained in the typical SIS model, where the infection curve is always monotone increasing or decreasing. In the long run, both the virus-free and the endemic steady states may be reached. This highlights the role of norms, or coordination rules, in the population.

[15]Note that complementarities in economic activity are also present in McAdams (2020b), where aggregate activity is an argument of agents' utility function. This can be seen as a reduced-form representation of the meeting mechanism described in the present paper.

5) *Adverse selection and its welfare cost.* As long as the cost of infection is not too low, it is easy to see that optimally, a planner would like to quarantine the small measure of infected individuals from the very beginning of the epidemic, and keep that rule in force forever. Isolated infectious agents would gradually recover, such that in the long run, only a very small measure of infected agents would remain infected and miss out on trade. In equilibrium, if anything happens, it is the exact opposite: infected agents always participate, and susceptible agents are forced to stay home. This is very costly, since susceptible agents initially form the wide majority of the population, and when they are confined, their measure only increases as infectious agents recover. This outcome is especially unfortunate when it occurs in the region where a Nash equilibrium with full participation could also occur, as it then resembles an extreme case of adverse selection. While it would be rational for a susceptible agent to participate if all other susceptible agents were also participating, as susceptible agents drop out, the risk of infection grows relative to the benefit of going out, which encourages participants to drop out further more. The equilibrium features a full unraveling, where only infectious agents are left in the market. It is worth noting that even though as currently described, this mechanism relies on the assumption that agents know their epidemiological status, the emergence of adverse selection would robust to different information specifications, as long as agents infrequently get informed about their status (through the development of symptoms or through testing, for example).

4.4.2 SIR model

We now assume that infected agents who recover become permanently resistant. The HJB equation for susceptible agents remains identical to (4.5). The HJB for infected agents is similar to (4.4) with the third term replaced by $\gamma(V_R - V_I)$, where V_R is the expected lifetime

discounted utility of a resistant agent. The HJB for resistant agents is

$$rV_R = \alpha(N)py + \dot{V}_R, \tag{4.10}$$

where we anticipate the result that in equilibrium, resistant agents have no incentives to stay out. We can easily show that $V_R - V_I = V_R^* - V_I^* = \psi/(r+\gamma)$, which is then used to again obtain a differential equation in ω,

$$\dot{\omega} = r\omega + \alpha(N)\min\left\{\tilde{y}, \frac{I}{N}\tau\omega\right\} - \frac{r}{r+\gamma}\psi. \tag{4.11}$$

The laws of motion for I is identical to (4.7) and the law of motion for R is given by

$$\dot{R} = \gamma I. \tag{4.12}$$

Finally, the participation decision of a susceptible agent is still given by (4.8), where $N = S^p + I + R$.

Definition 4.2 (Equilibrium definition)**.** *An equilibrium consists in a list of time paths for the three state variables* $\{I(t), R(t), S(t)\}$ *and the two control variables* $\left\{S^P(t), \omega(t)\right\}$ *such that* (4.11)*,* (4.7) *and* (4.12) *are satisfied, where* S^p *is a Nash equilibrium consistent with the individual participation decision rule given by* (4.8) *where* $N(t) = S^p(t) + I(t) + R(t)$*,* $S(t) = P - I(t) - R(t)$*,* $R(0)$ *and* $I(0)$ *are given.*

Steady states As was the case for the benchmark SIR model, any steady state must be virus-free, since we would otherwise have $\dot{I} > 0$. Absent any infection risk, the steady state participation of susceptible agents must be full, $S^{p*} = S$. This implies $\omega^* = \psi/(r+\gamma)$. There is a continuum of such steady states, indexed by $S^* \in [0, P]$, with $R^* = P - S^*$, as in the standard SIR model. Contrary to the SIS specification, adding a participation margin does not impact the set of steady states in the SIR model.

We now calibrate the model so as to study equilibrium dynamics.

4.4.3 Calibration

The SIR model is calibrated to the COVID-19 epidemic in the United States, at a daily frequency. There are three types of parameters to be calibrated: epidemiological parameters, that depend on the characteristics of the virus; parameters related to the meeting and matching of agents; and parameters related to costs and preferences. Calibrated values are summarized in Table 4.1.

Epidemiological parameters. The baseline probability of transmission of the virus in a given contact between an infected and a susceptible agent, $\tau \equiv \tau^{nn}$, is calibrated based on medical studies. A meta-analysis of 172 observational studies by Chu et al. (2020) predicts an infection probability of 3% after a contact between two individuals at a distance of three feet. Contact tracing studies in China by Bi et al. (2020) and Luo et al. (2020) track the secondary infections stemming from contacts with positive individuals, and respectively find infection probabilities of 6.6% and 3.7%. An intermediate value of 5% is chosen for the calibration, so that $\tau^{nn} = 0.05$. The recovery parameter is calibrated following Farboodi et al. (2020), who assume an expected recovery time of 7 days, implying that infectious agents recover at a Poisson arrival rate of 1/7. While many patients take more than 7 days to recover, this value is closer to the expected time during which infected agents can transmit the virus. Finally, to match the model's timeline to the real-world timeline, we need to calibrate time 0, when the first infection occurs (i.e., $I(0) = 1$). We follow a study by Worobey et al. (2020), who pin down the first case of coronavirus that eventually led to an outbreak in the US to mid-February, and set $t = 0$ to February 14, 2020.[16] For simplicity, it is assumed that

[16]Robustness checks with a later start date, similar to that used in Farboodi et al. (2020), are presented in Appendix D.2.

at this time, no one was resistant to the virus, so that $R(0) = 0$, and $S(0) = P - I(0)$.

Matching parameters The matching rate is assumed to be $\alpha(N) = \alpha N$, implying that the flow number of meetings, αN^2, displays increasing returns in the number of participants. Population P is calibrated to the 2019 estimate of the US population from the Census Bureau, around 328.24 million individuals. The matching parameter α is calibrated to match the average number of in-person social contacts experienced by individuals on a daily basis when there is no epidemic. A telephone survey of four counties in North Carolina by DeStefano et al. (2011) reports an average of 10 contacts per day. Feehan and Cobb (2019) survey Facebook users and find an average of 12 contacts per day. We pick the conservative value of 10, implying $\alpha P = 10$.[17]

Costs and preferences The daily discount rate is calibrated to match an annual discount rate of 5%, hence $r = 0.05/365$. The expected utility from a meeting, $\tilde{y} \equiv py$ is normalized to 1. The cost of infection is calibrated relative to this unit utility, following the method used in Farboodi et al. (2020). Because the expected cost of death for an infected individual trumps all other costs associated with the infection, they calibrate the cost of infection to reflect the cost of potentially losing one's life. Formally, $\psi/(r + \gamma) = \pi\nu$, where π is the infection fatality rate and ν the value of statistical life. The left-hand side represents the discounted expected cost of the disease: infected agents pay a flow cost ψ as long as they are sick. The infection fatality rate is calibrated to 0.0062. To calibrate the value of statistical life, Farboodi et al. (2020) follow Hall et al. (2020), who estimate that each remaining year of life is worth \$270,000, and that COVID-19 victims would on average expect 14.5 remaining years of life. This implies $\nu = \$3,915,000$. To convert this value to utils, first note that it is equivalent to say that individuals are willing to pay a lump sum of \$3,915, or a daily stream

[17] This is also in line with studies run in Europe, e.g. Mossong et al. (2008) find a lower bound of 7.8 daily contacts in Germany and 19.8 daily contacts in Italy. In a similar study run in Hong-Kong, Leung et al. (2017) find an average number of contacts of 8.

Param.	Definition	Target/Sources	Value
P	Population	US population	328.24M
α	Matching parameter	Average number of daily contacts	10/P
τ	Baseline transmissibility	Medical studies	0.05
γ	Recovery rate	7 days of infection on average	0.14
ψ	Cost of infection	Expected cost of death	272
$\tilde{y}$	Individual benefit from interaction	Normalisation	1
r	Discount rate	5% yearly discount rate	0.05/365
$t = 0$	Initial infection date	Epidemiological studies	Feb. 14, 2020
σ	Basic reproduction number	Implied	3.5
\$ per util	Exchange rate	Implied	12.33

Table 4.1: Calibrated parameters for the SIR model with participation.

of \$3,915r, to avoid a 0.1% probability of death. With a median yearly consumption of \$45,000 (Hall et al., 2020), this means that individuals would be indifferent between a 0.1% probability of death and forgoing $3,915r \cdot 365 \cdot 100/45,000 = 0.435\%$ of their consumption, or

$$\frac{\alpha P \tilde{y}}{r} - 0.001\nu = \frac{(1 - 0.435/100)\alpha P \tilde{y}}{r}. \tag{4.13}$$

We can now solve for ν and obtain $\nu = 317,550$ utils, from which we get $\psi = (r + \gamma)\pi\nu \approx 272$.[18] Note that the previous analysis implies an exchange rate between utils and dollars of $3,915,000/317,550 = 12.33$ dollars per util.

It is worth noting that Farboodi et al. (2020) and Bethune and Korinek (2020) use different approaches than the present paper to calibrate the timeline and the infection rate $\alpha\tau$. I explain those differences and explore the robustness of my results to those different approaches in Appendix D.2.

[18]This analysis implies that the entirety of an individual's consumption requires person-to-person contacts. Robustness checks where individuals still obtain some baseline utility when they stay home are presented in Table D.4 in Appendix D.2.

4.4.4 Results and discussion

Due to the coordination problem among susceptible agents, there exists a large multiplicity of equilibria. I consider a subset of equilibria that obey some coordination rules—more precisely, paths that exogenously dictates whether the susceptible population coordinates in or out of the market for each instant where both can be Nash equilibria. Key equilibrium results are described below. More detailed results are available in Table D.1 in Appendix D.1. A description of the algorithm used to solve the model can be found in Appendix D.3.

Randomized coordination rule I first focus on rules that specify that susceptible agents coordinate to participate with probability x, where $x \in \{0, 0.1, 0.2, \ldots, 1\}$. The two extreme cases, $x = 0$ and $x = 1$, correspond to equilibria where susceptible agents respectively always coordinate to stay home, and always coordinate to go out. Those two equilibrium paths are represented in a phase diagram in the left panel of Figure 4.5 in blue and in red, alongside the path that would be observed in a standard SIR model with no participation decision. Labeled "No behavioral response," it is represented by a dashed black line and will serve as a benchmark. Although uninformative about the time dimension, this phase diagram is helpful to understand the relation between the two state variables S and I. Starting from the bottom right corner, with $I(0) = 1$ and $S(0) = 0$, all three path originally feature an increase in the number of infected agents (and thus, mechanically, a decrease in the number of susceptible agents).

While the benchmark path and the path with maximum participation ($x = 1$) originally overlap, they diverge while I is still relatively low. In the benchmark case, the number of infected agents only starts to go down once the herd immunity threshold of $S/P = \gamma/(\alpha\tau P) = 0.2857$ has been reached. It then steadily declines and reaches the virus-free steady state with $S^*/P = 0.037$. Adding a participation margin allows for the number of infected agents to decrease much earlier, even when agents participate as much as possible.

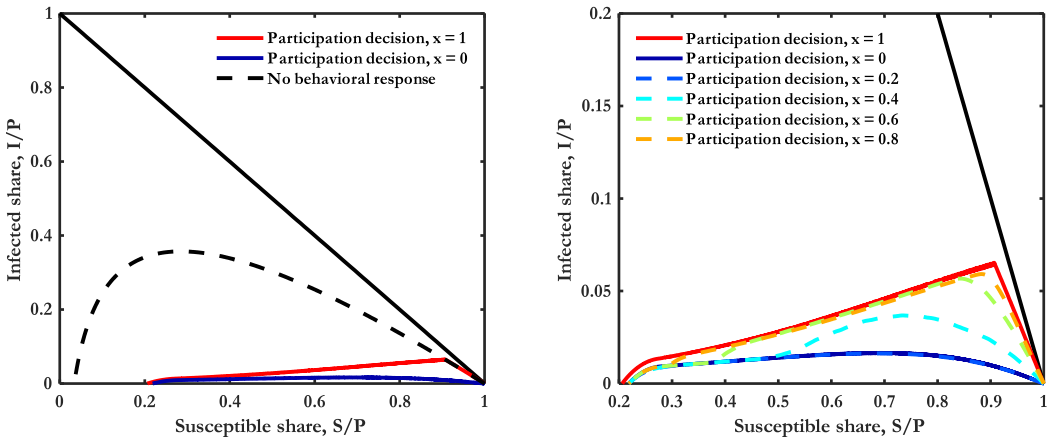

Figure 4.5: Equilibrium paths of the SIR model with participation decision, under the randomized coordination rule, represented in a phase plane. Left panel - Paths for the two extremes cases: in blue, agents always coordinate to stay home ($x = 0$), in red, agents always coordinate to go out ($x = 1$). The path with no participation decision, in dashed black, is provided for comparison. Right panel - Paths for the two extreme cases and intermediate cases. The graph is a magnified view of the bottom-right part of the full-size graph.

The virus-free steady state is reached with $S^*/P = 0.2061$. In other words, in this scenario, 79.39% of the population would have been infected by the end of the epidemic, compared to 96.63% in the benchmark case. Now looking at the other extreme case, where $x = 0$, we can observe slightly different dynamics. First, the relation between S and I is very flat, so that the measure of infected agents remains very low throughout the epidemic. However, it starts to declines for a lower S than was the case with $x = 1$, and eventually converges towards a virus-free steady state remarkably close to that resulting from the maximum equilibrium participation, with $S^*/P = 0.2185$, i.e., 78.15% of the population has been infected by the end of the epidemic.

To understand how those dynamics play out over time and how they relate to the participation decision, we can look at the four panels of time series in Figure 4.6. They run from February 14, 2020, when the first case is assumed to occur, to February 14, 2026. From the top-right panel, which displays the share of population infected over time, we can confirm that the addition of a participation margin allows to considerably flatten the curve, even

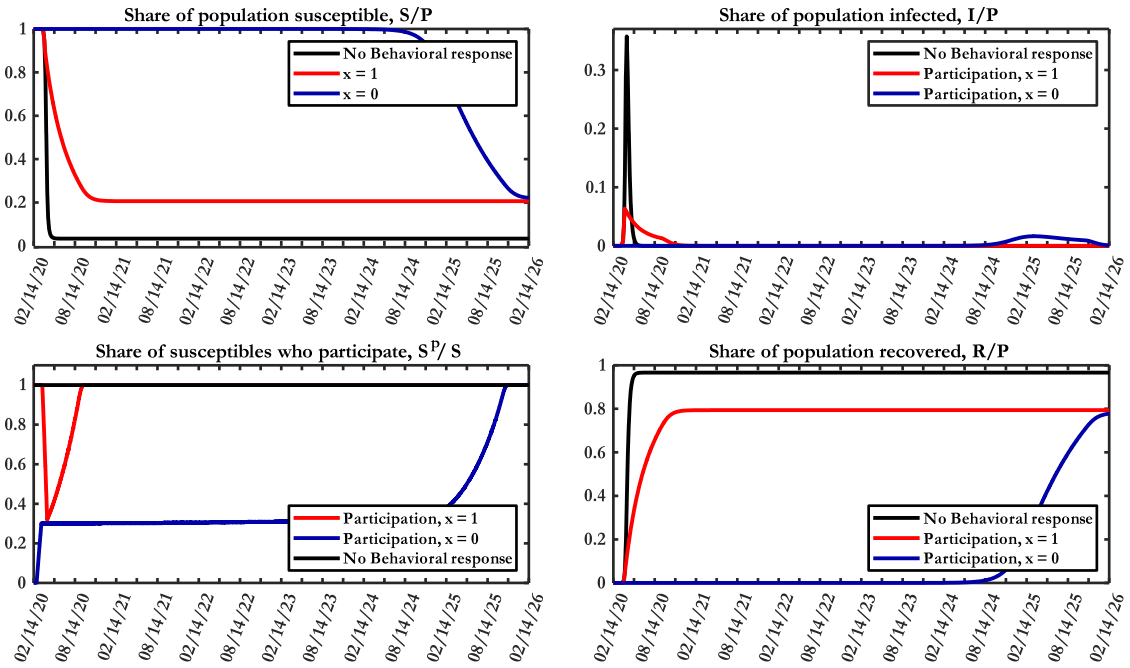

Figure 4.6: Time paths of epidemiological measures for the SIR model with participation, under the randomized coordination rule: susceptible agents coordinate to participate with probability x, for $x = 0$ and $x = 1$. Note that 3 months lapse between each tick on the x-axis, and that labels are plotted every 6 months.

when $x = 1$. In this case, the peak of infections only slightly goes over 6% over the population, against a peak at almost 40% of the population for the benchmark case. When $x = 0$, the curve is even flatter, and the measure of infectious agents remains below 2% of the population at all times.

The flattening of the curve is what allows for a much smaller number of individuals to be infected over the course of the epidemic. This can be seen directly by looking at the bottom-right panel, which shows the share of population recovered as a function of time, since the share of population recovered as time goes to infinity must be equal to the share of population that has been infected in total. We can confirm that it is considerably smaller when the participation margin is introduced.

The evolution of participation over time is displayed in the bottom-left panel. Because of the discrete nature of the algorithm used to solve the model, values for S^p often jump between

0 and S when in the multiplicity region. As a result, participation is plotted as a rolling average with ten-day windows.[19] In the equilibrium path with the highest participation, $x = 1$, we can see that originally, all susceptible agents participate. By mid-March, around one month after the first infection, participation sharply drops. It reaches a low point at 30% of its regular level by mid-April. During this time, the peak of the epidemic is reached, and the number of infections starts declining. Participation then gets back up, and the economy reaches full participation again by September 2020. In the equilibrium path with the lowest participation, $x = 0$, participation drops to zero from day one. It then climbs back to reach a bit less than 30% of the baseline participation level by the beginning of March, and remains stable at this level for many months. Only around May of 2024 does it start to slowly increase again, the economy getting back to full participation by November 2025. As a result, when the coordination rules is such that susceptible individuals stay home, the epidemic is not only much flatter, it is considerably delayed. This is visible in the top-right panel of Figure 4.6.

We now turn our attention to welfare measures. The expected human toll of the epidemic is displayed in the left panel of Figure 4.7.[20] As expected, because the coordination rule bears little impact on the total number of agents infected by the end of the epidemic, its impact on the expected number of fatalities is also limited. The model predicts around 1.6 million fatalities for both $x = 0$ and $x = 1$, compared with almost 2 million without the participation margin. However, whether agents coordinate in or out does have a very large impact on the welfare costs of the epidemic, displayed in the right panel of Figure 4.7. The total costs, represented by full lines, can be broken down between the direct costs borne by infected agents and the opportunity costs due to susceptible agents staying home. When

[19] This explains why interior values can be obtained even though, in each period, there is either full or no participation.

[20] Even though the model does not feature deaths per se, in that the population always remains constant, the cost of infection ψ takes into account the expected cost of death for infectious agents. The expected number of fatalities is computed by multiplying the infection fatality rate, π, by the number of recovered agents.

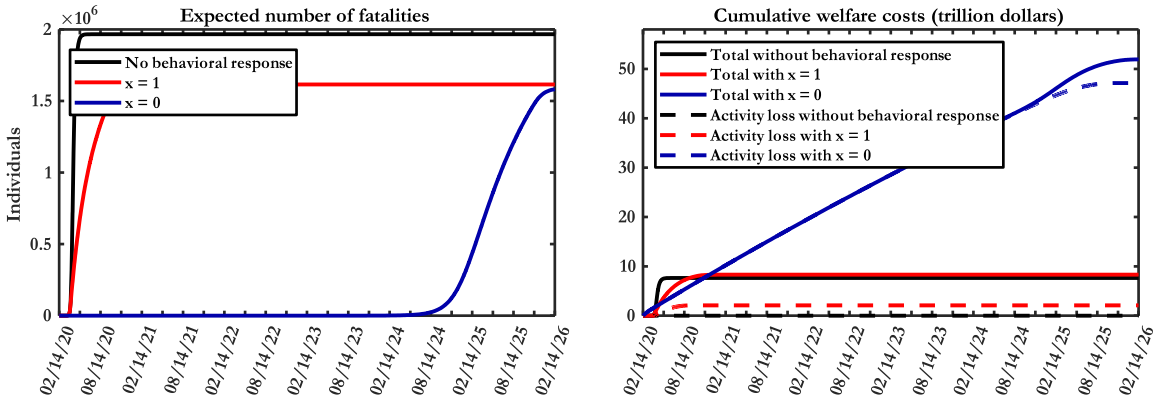

Figure 4.7: Time paths of expected number of fatalities and cumulative welfare losses for the SIR model with participation, under the randomized coordination rule: agents coordinate to participate with probability x, for $x = 0$ and $x = 1$. Total welfare losses correspond to the discounted cumulative sum of sickness costs borne by infected agents and activity losses due to susceptible agents staying home. Note that 3 months lapse between each tick on the x-axis, and that labels are plotted every 6 months.

agents coordinate on going out, $x = 1$, total costs amount to 8.3 trillions dollars, roughly 40% of US GDP.[21] Three quarters of this cost are due to sickness costs. When agents coordinate on staying in, $x = 0$, total costs skyrocket to around 51 trillion dollars. While sickness costs are 1.5 trillion dollars lower than when $x = 1$, 90% of of the total losses stem from the loss in activity. Recall that in that scenario, the participation of susceptible agents drops to 30% of its usual level for months on end. While this delays the epidemic, it does little to reduce the total number of infections. As a result, the losses due to the reduction in activity trump the gains in sickness costs by far, and the equilibrium with $x = 0$ is much worse, from a welfare perspective, than the equilibrium with $x = 1$. Interestingly, the benchmark path is even slightly better in terms of welfare than the equilibrium path with $x = 1$. While sickness costs are only a bit larger, there is no activity loss whatsoever, leading to a better net outcome. This is due to the adverse selection problem described earlier, and further discussed in *takeaway 5* at the end of the section.

Until now, we focused on the two extremes cases of the coordination rule, $x = 0$ and $x = 1$.

[21] All welfare numbers are computed as the discounted sums of welfare losses from time 0.

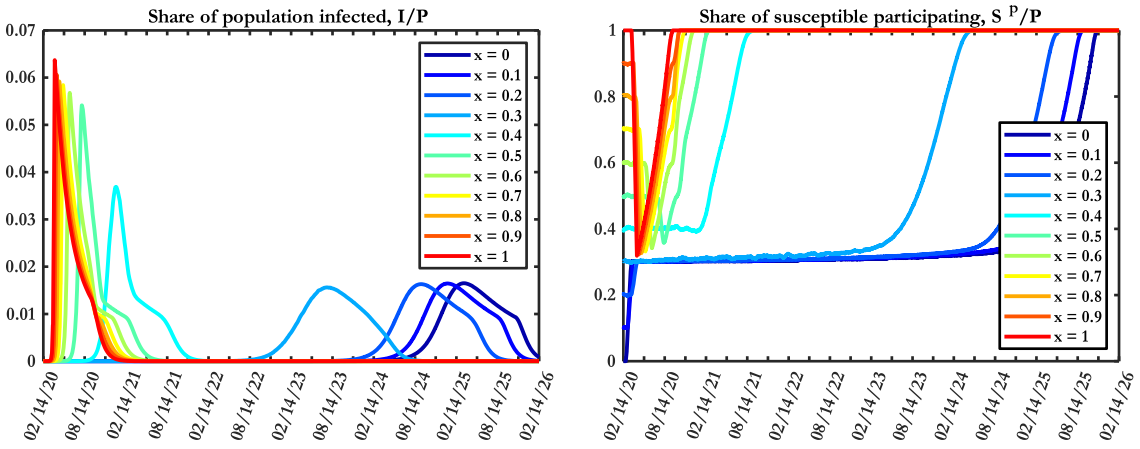

Figure 4.8: Time paths of share of infected population and share of susceptible population participating, under the randomized coordination rule: agents coordinate to participate with probability x, for $x = 0, 0.1, ..., 1$. Note that 3 months lapse between each tick on the x-axis, and that labels are plotted every 6 months.

We can now look at intermediate cases. Note that for those cases, the realized equilibrium path depends on the random draws made over time. For this reason, for each intermediate case, 50 simulations were run and then averaged. Going back to a phase diagram representation on the right panel of Figure 4.5, we can see that intermediate cases populate the phase plane in between the two cases studied earlier. Because all paths converge towards virus-free steady states very close to those described for $x = 0$ and $x = 1$, we can already see that the long term epidemiological outcomes will remain similar across those intermediates cases, with around 78% of the population infected. Mechanically, the expected number of fatalities remains around 1.6 million across cases. Similarly, because those paths are "bound" by the two extreme paths, welfare outcomes will lie in between the two welfare outcomes described earlier.

Figure 4.8 shows the particular shapes taken by the infection curves (in the left panel), as well as the associated participation decisions (on the right panel), for all x. As the probability of coordination to go out, x, decreases, the infection curve flattens and shifts to the right. In terms of participation, for $x \leq 0.3$, participation first starts at $100x\%$, then increases to 30%, and finally gradually goes back up to 100% after some time. The lower x, the longer

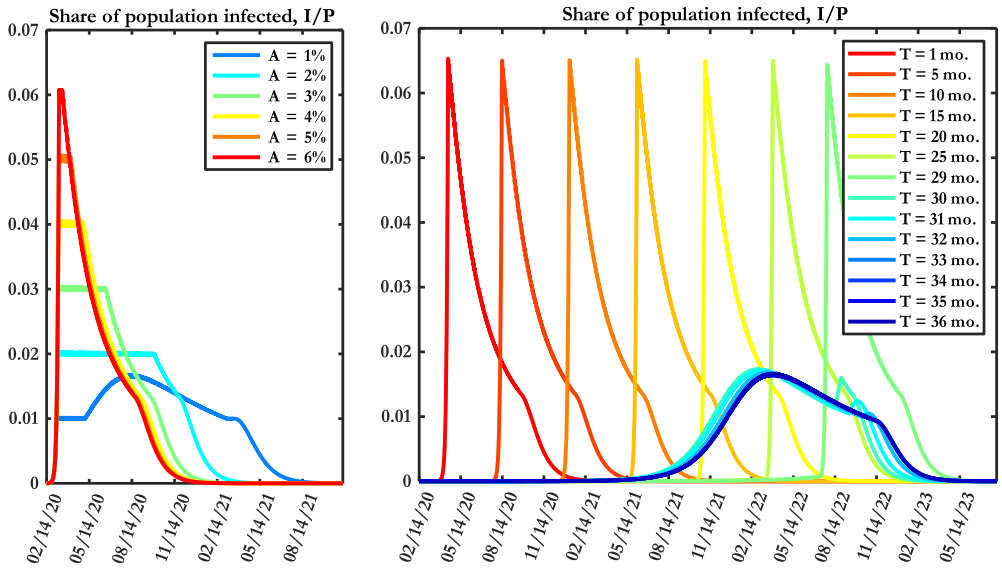

Figure 4.9: Time paths of expected number of fatalities and cumulative welfare losses for the SIR model with participation, under alternative coordination rules. Left panel - Susceptible agents coordinate to participate as long as $I < A$, and coordinate on staying in otherwise. Right panel - Susceptible agents coordinate to stay home as long as $t < T$, and coordinate to participate afterwards (fatigue specification).

the time during which participation remains at 30%. For $x \geq 0.4$, participation starts at $100x\%$, drops below 40%, then gets back up to 100% relatively quickly. The higher x, the earlier and the deeper the drop, and the faster the recovery.

Alternative coordination rules While the probability-based coordination rule was picked as a straightforward rule allowing to populate the phase diagram between the two extreme rules of always or never coordinating to go out, an infinite number of alternative rules could be used. I present two such rules in this section.

The first alternative rule assumes that agents coordinate on staying in or going out depending on the measure of active infections cases, I. When this number is lower than some threshold A, agents feel safe and coordinate out. When it is higher, agents coordinate on staying home. Infections curves for $A \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}\%$ of the population are displayed in the left panel of Figure 4.9. As expected, when A is high, the infection curve resembles

the one obtained when agents always coordinate to go out. Indeed, in this case, the measure of infected agents never reaches the threshold, so that the rule is never binding. In contrast, for low enough A, while the infection curve initially follows that of the $x = 1$ scenario, the number of infections then perfectly stabilizes once A active infections have been reached. Note that this does not mean that susceptible agents entirely stop going out. In that case, the number of infections would decline. Instead, susceptible agents go out exactly enough to maintain the effective reproduction number at one. It then eventually decreases, except for very low A. When the threshold is equal to 0.1% of the population, the number of infections initially stabilizes, but infections start growing again after a few months.

The second alternative rule assumes that individuals initially coordinate on staying home, which could be seen as the "good behavior" due to social norms, but eventually get fatigued and coordinate on going out from time T onward for $T \in \{1, 5, 10, 15, 20, 25, 29, 30, 31, 32, 33, 34, 35, 36\}$ months are displayed in the right panel of Figure 4.9. Up to $T = 29$ months, fatigue arising later and later only shifts the infection curve to the right. For $T = 30$, that is, assuming that fatigue develops two and half year into the epidemic, an interesting phenomenon occurs. The start of the epidemic shifts back, in between the starts of the path with $T = 15$ and $T = 20$. Instead of a sharp increase in the number of active cases, the curve is much flatter. While it then resembles the curves observed earlier with low x, there are two big differences: first, the epidemic is not as delayed; second, there can be a second wave of infections. For example, for $T = 30$, while the number of infected individuals starts to go down around January 2022, it goes up again, this time more steeply, roughly six months later. While left for further exploration, it is easy to see that the existence of more than two infection waves can easily be obtained as equilibrium outcomes, conditional on picking the coordination rule appropriately.

Takeaways and discussion Key insights from the calibrated SIR model with participation are summarized below.

1) Equilibrium indeterminacy and multiplicity of infection dynamics. The calibrated model features an infinite number of equilibrium paths. The infection curve can be single-peaked like in the standard SIR model, but can also feature several peaks or even remain flat for some time. It can develop quickly, and disappear before the end of 2020, or be delayed for months and even years. Those dynamics are driven by the coordination of susceptible agents in the economy. While not explored in the paper at the moment, it could be interesting to consider factors that could impact the coordination of agents: country-wide experience with previous infectious disease and associated norms, policy messages and recommendations, etc.

2) Endogenous flattening of the infection curve. The idea of "flattening the curve" has consistently been put forward by policy makers during the COVID-19 pandemic, often with two main justifications: first, it would help lessen the load on hospitals, thereby increasing chances of recovery for infected individuals; second, it would help minimize the number of victims until a vaccine is found. Neither of these two incentives is present in the current model, since the recovery rate γ is independent of the number of infections, and the possibility of a vaccine is not modeled. Nevertheless, we can observe a significant flattening of the curve across all of the examples presented, relative to the infection curve in a world with no participation margin. Individuals' expected cost of infection provides them with incentives to stay home, which ends up diminishing the death toll of the epidemic even in a world where herd immunity is the only way out. It is interesting to note that due to the complementarities between the participation decisions of susceptible agents, the endogenous "lock-down" can happen from the very beginning of the outbreak (e.g., in the case with $x = 0$), when the number of infected agents is still very low. This differs from Bethune and Korinek (2020), Farboodi et al. (2020), Garibaldi et al. (2020), and Toxvaerd (2020), where behaviors only start to change after the epidemic has gained more ground.

3) Participation and herd immunity. In a standard SIR model, the epidemic can only start declining once the number of susceptible individuals is low enough, which can be directly mapped to a minimum number of people having been infected and having gained immunity. This herd immunity threshold is very high for COVID-19: using the calibration presented in Section 4.4.3, we saw that it would require more than 71% of the population to have been infected for the outbreak to start declining. The simulations presented above show that the epidemic can start to wane much before that many people have been infected. For example, when $x = 1$, the number of active cases starts to decline after only 2.8% of the population has been infected, much before herd immunity has been achieved. It is sufficient for the *participating population* to have reached herd immunity, a sort of qualified herd immunity. As long as $\sigma_e^p = \alpha(S^p + I + R)\tau S^p/(S^p + I + R) < 1$, the measure of infected individuals decreases. Because this is a jump variable, it cannot be mapped into a corresponding measure of people having recovered, and it does not necessarily last forever (as visible in Figure 4.9, where some equilibrium paths go down then up again). Eventually, the entire population does gain herd immunity, and participation can get back to full participation.[22] Hence, while herd immunity is eventually reached by the population in both the standard SIR model and the SIR model with participation, the important difference is that in the latter model, the epidemic can start declining much before the threshold has been reached, eventually leading to a much smaller number of total infections.

4) Invariance of long-run epidemiological outcomes. One striking result is that despite the large multiplicity of equilibria, long-run epidemiological outcomes do not differ very much across those equilibrium paths. In all specifications reported in the analysis, the steady-state number of resistant agents lies between 78% and 81%. This implies that two regions

[22] As I gets close to 0, all susceptible agents participate because the risk of infection is very small compared to the benefits from interacting with resistant agents. As a result, the effective reproduction number is greater than one when I is close to zero and $S > P/\sigma$. Thus, as was the case in the benchmark SIR model, only steady states where $S^* \in [0, P/\sigma]$ can be reached, which implies that there will always be at least a share $(1 - 1/\sigma)$ of the population eventually infected.

with widely different infection curves may not necessarily be headed towards widely different long-run outcomes, and that infection curves by themselves may not be sufficient indicators of how well a region is doing.

5) Negative welfare outcomes. The issue of adverse selection was highlighted in section 4.4.1. Simulations with the SIR model confirm that it comes at a great cost. From the point of view of society, diminishing participation is beneficial: it allows the number of total infections to drop. However, in equilibrium, the drop of participation comes from susceptible agents exclusively. It would be much better both for susceptible agents and for society if infectious agents, who are the large minority, were the ones staying home. Even in the case where agents participate as much as possible, the losses due to missed interactions remain extremely large, and the gains due to fewer infections are not sufficient to offset those losses. Equilibria that simply delay the epidemic, in such a way that the infection curve is shifted to the right, are even worse. As discussed in *takeaway 4*, while they feature a delayed epidemic, those equilibrium paths do not come with significantly fewer infections, so that the diminished participation is a pure loss for society. It is worth noting that this negative welfare result largely depends on the way infections (and by proxy, deaths), are valued both privately and by society. Robustness checks show that either increasing the private cost of infection by 50%, or the social cost of infection by 45%, would make the equilibrium path with $x = 1$ reach lower welfare losses than the benchmark path.

4.5 To mask or not to mask

We now consider a model where agents cannot opt out from social interactions, so that $N(t) = P$, but can choose to wear a mask as a precaution. After solving for the equilibrium analytically in a SIS specification, I set up, calibrate and solve numerically for the SIR equilibrium outcomes. In a last part, the participation decision is reintroduced in order to

shed light on the interaction between the two reaction margins.

4.5.1 SIS model

The HJB equation for an infected agent is similar to (4.4), derived in the participation model, where we saw that infected agents have no incentives to restrict their participation. Likewise, in the SIR model with mask-wearing, infected agents prefer not to wear masks. As a result, the only relevant transmission rates are τ^{nn} and τ^{mn}. The second superscript being superfluous, notations for those two parameters are simplified to τ^n and τ^m. The HJB equation for a susceptible agent is

$$rV_S = \alpha(P)\tilde{y} + \max\left\{-k + \tau^m \frac{I}{P}(V_I - V_S), \tau^n \frac{I}{P}(V_I - V_S)\right\} + \dot{V}_S. \tag{4.14}$$

As in the benchmark model, the agent matches with other agents with a Poisson arrival rate $\alpha(P)$, in which case she can expect to earn $\tilde{y}$ utils. The second term of the maximization represents the expected cost of social engagement when not wearing a mask. In this case, neither she nor the infected trade partners she may meet wear masks, and the probability of infection when in contact with an infectious agent is τ^n. The first term in brackets represents the cost of social engagement when wearing a mask. The agent would suffers a flow disutility k, but would contract the disease with a smaller probability, $\tau^m < \tau^n$, in case of contact with an infectious agent.

Again, we can combine the two HJB equations to obtain one differential equation in $\omega \equiv V_S - V_I$,

$$\dot{\omega}(t) = (r+\gamma)\omega(t) - \psi + \min\left\{k + \alpha(P)\frac{I(t)}{P}\tau^m\omega(t), \alpha(P)\frac{I(t)}{P}\tau^n\omega(t)\right\}. \tag{4.15}$$

The measure of infected agents evolves according to the law of motion

$$\dot{I} = S^m \alpha(P) \frac{I}{P} \tau^m + (S - S^m) \alpha(P) \frac{I}{P} \tau^n - \gamma I, \tag{4.16}$$

where S^m denotes the measure of susceptible agents who wear a mask. The first term corresponds to the inflow of previously-susceptible agents who wore masks but still contracted the virus. The second term corresponds to susceptible agents who did not wear masks and got infected. The third term corresponds to previously-infected agents who recovered (and become susceptible again).

To close the model, we now need to solve for the measure of susceptible agents wearing a mask. To decide whether to wear a mask, a susceptible agent weighs the cost, k, against the benefit, $\alpha(P)\frac{I}{P}\omega(\tau^n - \tau^m)$. We directly get that in aggregate, the measure of susceptible agents who wear masks, S^m, is given by

$$S^m \begin{cases} = S & < \\ \in [0, S] \text{ if } k & = \quad \alpha(P)(\tau^n - \tau^m) I\omega/P \\ = 0. & > \end{cases} \tag{4.17}$$

Definition 4.3 (Equilibrium definition)**.** *An equilibrium consists in a list of time paths for the two state variables* $\{S(t), I(t)\}$ *and the two control variables* $\{S^m(t), \omega(t)\}$*, such that* (4.15) *and* (4.16) *are satisfied, where* S^m *is given by* (4.17)*,* $S(t) = P - I(t)$*, and* $I(0)$ *is given.*

Steady-states As in the benchmark and participation-based SIS models, an endemic and a virus-free steady state coexist. The virus-free steady state is such that $I^* = 0$, $S^* = P$, $S^m = 0$ and $\omega = \psi/(r + \gamma)$. Absent any risk of infection, there is no incentive for susceptible agents to be wearing masks.

In the endemic steady state, the distribution of infectious and susceptible agents as well as the prevalence of masks depend on the magnitude of k relative to $(\tau^n - \tau^m)$. Let $\underline{k} \equiv [\alpha(P)\tau^m - \gamma](\tau^n - \tau^m)\psi / \{\tau^m(r+\gamma) + \tau^n[\alpha(P)\tau^m - \gamma]\}$ and $\bar{k} \equiv [\alpha(P)\tau^n - \gamma](\tau^n - \tau^m)\psi / \{\tau^n(r+\gamma) + \tau^n[\alpha(P)\tau^n - \gamma]\}$. There are three cases.

When $k \geq \bar{k}$, $I^* = P(1 - 1/\sigma) \equiv I^{*n}$, $S^* = P/\sigma$, $S^{m^*} = 0$ and $w^* = \psi/[r + \alpha(P)\tau^n]$. Intuitively, if the cost of wearing a mask is high enough compared to the benefit, even the endemic steady state features no mask-wearing. In that case, it is identical to the endemic steady state from the benchmark SIS model.

When $k < \underline{k}$, the endemic steady state is such that $I^* = P\{1 - \gamma/[\alpha(P)\tau^m]\} \equiv I^{*m} < I^{*n}$, $S^* = P\gamma/[\alpha(P)\tau^m]$, $S^{m*} = S$, and $\omega^* = \psi/[r + \alpha(P)\tau^m]$. For a low enough k, all susceptible agents wear a mask in the endemic steady state. Recall that in an endemic steady state, the effective reproduction number must be equal to one. For the two steady states just described, it is given by $\sigma_e^m = \alpha(P)\tau^j S/(\gamma P)$, with j respectively equal to n and m. Given $\tau^m < \tau^n$, masks wearing allows the endemic steady state to occur with a lower proportion of infectious agents in the population.

Finally, when $k \in (\underline{k}, \bar{k})$, the endemic steady state is given by $I^* = P(r+\gamma)k/[\alpha(P)(\tau^n - \tau^m)\psi - \tau^n k] \in (I^{*m}, I^{*n})$, $\omega^* = \psi/(r+\gamma) - k\tau^n/[(r+\gamma)(\tau^n - \tau^m)]$. We can show that I^* is increasing in k, so that $S^* = P - I^*$ is decreasing in k, and $S^{m^*} = (\alpha\tau^n S^* - \gamma)/(\tau^n - \tau^m)$ is decreasing in k as well. The costlier masks, the smaller the prevalence of masks at steady state, so that a higher proportion of the population must be infected.

Dynamics System dynamics can be analyzed with the help of the phase diagrams displayed in Figure 4.10. The left panel represents the case when $k < \underline{k}$, the middle panel the case when $k \in [\underline{k}, \bar{k}]$, and the right panel the case where $k > \bar{k}$. Each diagram is split into two zones by the mask-wearing indifference curve, plotted in black and given by $\omega = Pk/[(\tau^n - \tau^m)I]$.

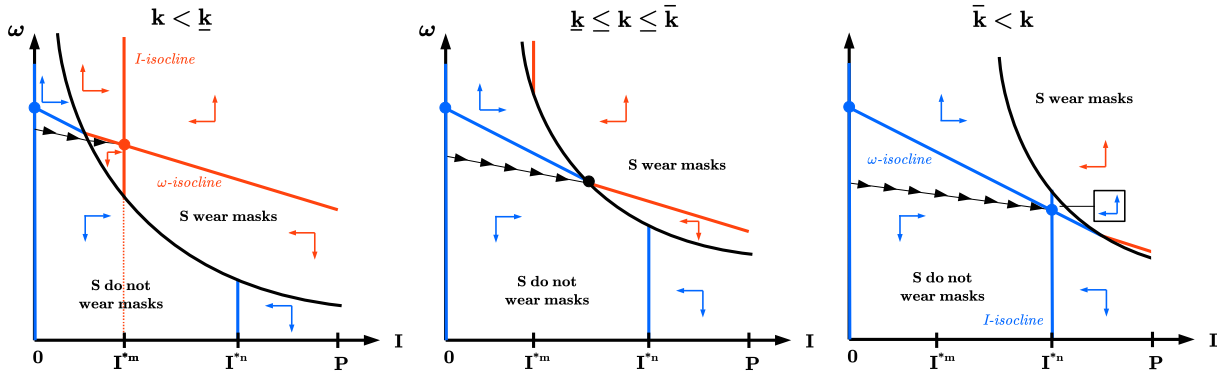

Figure 4.10: Dynamics of the SIS model with mask-wearing decision, as a function of the flow cost of wearing a mask, k.

Above this curve, all susceptible agents wear masks. Below, none do. Along the curve, when I goes to zero, ω goes to infinity. As a result, at the limit, agents never wear masks because the cost is too high relative to the infrequent benefit. Then, the reasoning explained in Section 4.3 applies: as long as the basic reproduction number is greater than one, the measure of infected agents would always increase when I is close to zero, making the virus-free steady state unstable. This result is different than what had been obtained in the participation model. With the participation margin, it was possible for susceptible agents to coordinate on staying home even when I got close to zero because of the complementarities between their individual decisions. Here, there is no such complementarities. A corollary is that agents do not mask at the very onset of the epidemic either, when I is still low. Instead, dynamics can be described as follows.

When k is high, starting from a low $I(0)$, the unique equilibrium path is such that $\omega(0)$ jumps onto the saddle path, represented by the black line with arrows, and the system remains on that path, with I growing and S going down, until it converges to the endemic steady state represented by the blue dot. There is no mask-wearing at any point in time. Thus, equilibrium dynamics are exactly identical to the equilibrium dynamics from the standard SIS model. When k is in the middle range, agents initially do not wear masks and the

epidemic develops identically to the previous case, up until reaching the indifference curve. At this point, a measure S^{m*} of the susceptible agents start wearing a mask and the system is at steady state, represented by the black dot. When k is low, there is still no mask wearing at the onset of the epidemic, which initially develops no differently that when k is higher. However, agents eventually become indifferent between wearing a mask or not and at this point, the system reaches the saddle path that leads to the mask-wearing steady state (represented by the red dot). On this saddle path, all agents wear masks. In all three cases, dynamics are monotone in I, S and S^m.

4.5.2 SIR model

We now switch to the SIR specification, where infectious agents who recover become permanently resistant. Very few modifications are required. The HJB for susceptible agents remains identical to (4.14). The HJB for infectious agents also remains similar to that from the SIS model, but the term $(V_S - V_I)$ becomes $(V_R - V_I)$. Finally, the HJB for recovered agents is identical to the one derived for the SIR participation model, (4.10), with $N = P$ since resistant agents have no incentives to wear masks. We can then obtain a differential equation in ω,

$$\dot{\omega} = r\omega - \min\left\{k + \alpha(P)\tau^m \frac{I}{P}\omega, \alpha(P)\tau^n \frac{I}{P}\omega\right\} - r\frac{\psi}{r+\gamma}. \tag{4.18}$$

The law of motion is for I is given by (4.16), the low of motion for R by (4.12), and the aggregate measure of mask wearers by (4.17).

Definition 4.4 (Equilibrium definition)**.** *An equilibrium consists in a list of time paths for the three states variables* $\{I(t), R(t), S(t)\}$ *and the two control variables* $\{S^m(t), \omega(t)\}$ *such that* (4.18)*,* (4.16) *and* (4.12) *are satisfied, where* S^m *is given by* (4.17)*,* $S(t) = P - I(t) - R(t)$*,* $R(0)$ *and* $I(0)$ *are given.*

Steady-states Again, no different than standard SIR models, steady state requires the economy to be virus-free, $I = 0$, and any $S^* \in [0, P]$ and $R^* = P - S^*$ constitute a steady state of the system. We showed earlier that there is no mask-wearing when I is low, so $S^{m*} = 0$.

We now calibrate the model in order to study its dynamics.

4.5.3 Calibration

Most parameters were already present in the model with participation, calibrated in Section 4.4.3, and are kept identical. There are two additional parameters to calibrate: the transmissibility of the virus in a contact where the susceptible agent wears a mask while the infectious agent does not, τ^{mn}, and the cost of wearing a mask, k.

There is no general consensus regarding the efficacy of masks, which significantly depends on the type of mask as well as the fit of the mask to the wearer. While it is generally understood that masks are most efficient when they are worn by the source of the virus (the infectious agents), several studies do suggest that masks also confer a benefit to the wearer. For example, Li et al. (2020) claim that masks reduce the risk of infection for the wearer between 40% to 70%. We pick an efficacy of 50%, so that $\tau^{mn} = 0.5\tau^{nn} = 0.025$.

The cost of mask-wearing is calibrated using a heuristic approach driven by the idea that mask-wearing is costly because of the physical inconvenience it imposes when going out and engaging in social interactions—it may be harder to breathe or to communicate, for example. We then assume that k is proportional to the utility received from meetings. More specifically, $k(N) = x\alpha N\tilde{y}/100$, i.e., agents suffer a disutility cost equal to $x\%$ of their utility from having to wear a mask. We then pick $x \in \{1, 5, 10\}$, which corresponds to daily costs of around \$1.23, \$6.16 and \$12.33 when participation is full (averaging ten meetings per person

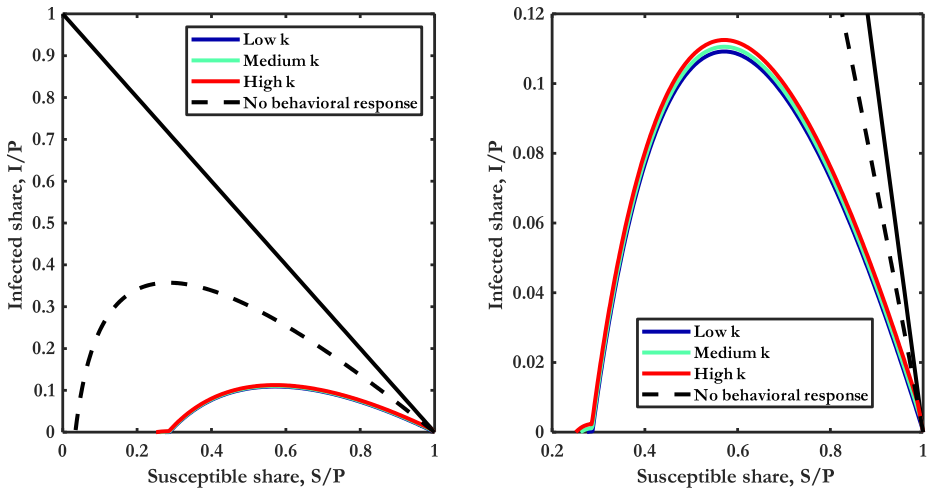

Figure 4.11: Equilibrium paths of the SIR model with mask-wearing decision, represented in a phase plane. Left panel - Paths for the three levels of the mask-wearing cost: low k in blue, medium k in green, and high k in red. The path with no mask-wearing decision, in dashed black, is provided for comparison. Right panel - Magnified view of the bottom-right of the graph displayed in the left panel.

per day). Those three values will be referred to as low, medium, and high k thereafter.

4.5.4 Results and discussion

I now describe some key equilibrium results. More detailed results are available in Table D.2 in Appendix D.1. Contrary to the model with participation, there are no complementarities between the mask-wearing decisions of agents, and the equilibrium is unique.

The relationship between I and S is plotted in the phase planes displayed in Figure 4.11, for the three levels of k considered. The left panel shows the full graph, so that the equilibrium paths can easily be compared with the equilibrium path obtained absent the mask-wearing decision. We first notice that the paths do not differ much across the three levels of k, and are difficult to distinguish one from another. While the shape of the curve is roughly similar to that followed by the benchmark curve—it first goes up then goes down with a slightly flatter slope, it is much flatter, from the very beginning. As a result, the number of infected

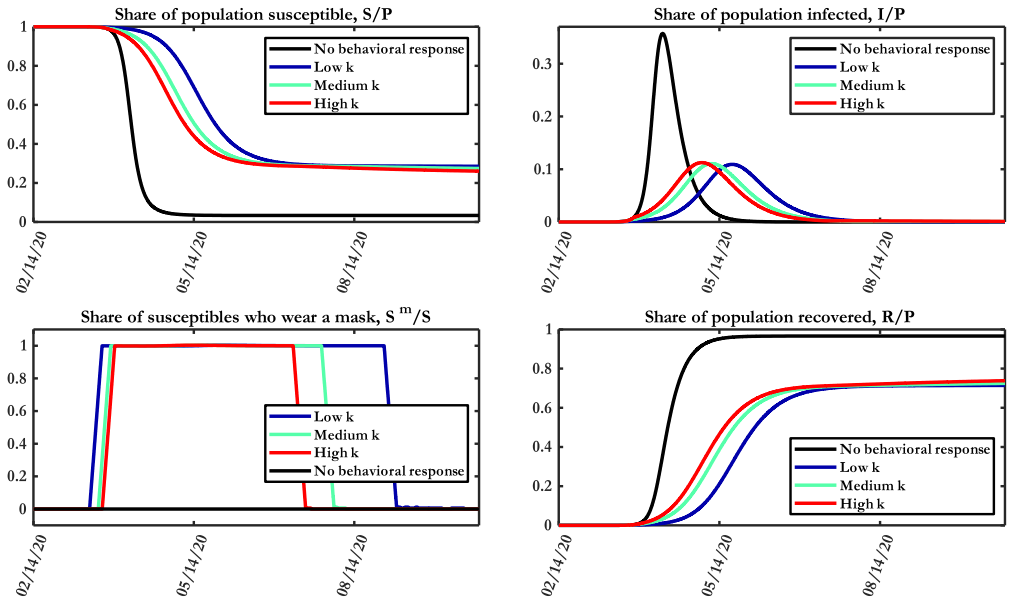

Figure 4.12: Time paths of epidemiological measures for the SIR model with mask-wearing, for the three levels of the mask-wearing cost: low k in blue, medium k in green, and high k in red. Note that 3 months lapse between each tick and label on the x-axis.

individuals is always lower (never breaking past 12%), the peak of the epidemic is reached for a lower S, and in the long-run, a considerably smaller number of individuals are infected.

One interesting feature to notice is that as I gets close to 0, the curve becomes significantly flatter until it reaches the steady state. This certainly corresponds to a shift back to no mask-wearing, which can be confirmed with the time plots. The right panel is a magnified view of the left panel, zooming in on the bottom-right corner. It is now easier to distinguish the three equilibrium paths. As expected, the lower the cost, the lower the peak of infections, and the lower the number of cumulative infections in the long run. More precisely, once at steady state, 72.51% of the population has been infected when k is low, 73.87% when k is in the middle, and 74.90% when k is high. Recall that absent any behavioral response, 96.63% of the population gets infected, while in the equilibrium with participation, between 78.15% and 79.30% were.

We can now look at the dynamics of the epidemic on the four panels of Figure 4.12. From

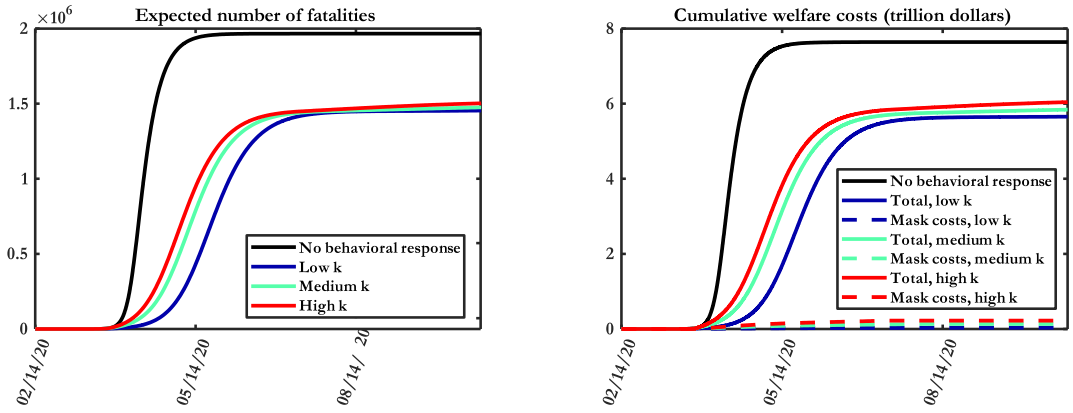

Figure 4.13: Time paths of expected number of fatalities and cumulative welfare losses for the SIR model with mask-wearing, for the three levels of the mask-wearing cost: low k in blue, medium k in green, and high k in red. Total welfare losses correspond to the discounted cumulative sum of sickness costs borne by infected agents and the mask-wearing costs borne by susceptible agents. Note that 3 months lapse between each tick and label.

the top-right panel, we can see that not only is the infection curve flattened when we allow for mask-wearing, but it is also delayed. The lower the cost of wearing a mask, the greater the delay. In the benchmark scenario, the epidemic peaks around mid-April. With high k, it peaks around the beginning of May, while with low k it peaks around the end of May.

All of those differences can be explained by the time paths for the share of susceptible individuals who wear a mask, displayed in the bottom left panel. Mask-wearing does not occur in equilibrium either at the very beginning or the very end of the epidemic, when I is low and the chance of infection is too low to warrant bearing the cost of mask-wearing. As I goes up, it eventually becomes rational to wear a mask and as expected, this occurs first when k is low. At this point, susceptible individuals sharply go from no mask-wearing to full mask-wearing, in the matter of less than two weeks. This occurs in the second half of March 2020. All susceptible individuals then keep wearing their mask until I is back to being low enough, again in a relatively quick fashion. This occurs at the end of July for the high k, during the first half of August for the medium k, and at the beginning of September for the low k.

We can now turn to the long-term impact for society and look at the expected number of fatalities as well as the welfare cost of the epidemic in the long-run. Those are represented in Figure 4.13. Because the number of total infections is relatively similar across all calibrations for k, the expected number of fatalities is relatively constant as well. We can see on the left panel that this number is around 1.5 million. Not only is it considerably smaller than the expected number of fatalities in the no-response scenario (approximately 2 million), it is also smaller than the expected number of fatalities in the model with the participation margin (approximately 1.6 million).

The right panel displays the discounted cumulative welfare losses. Those include the sickness costs borne by infected agents, and the mask-wearing costs borne by susceptible agents. While in the participation decision model, the welfare gains obtained thanks to the drop in infections and fatalities due to susceptible agents staying home were more than offset by the cost of foregone economic and social activity, the conclusion is much rosier here. Compared to the benchmark model, the economy goes from a cumulative loss of 7.6 trillion dollars in the benchmark case to between 5.7 and 6.1 trillion dollars when masks are introduced. Notably, the lower the cost of masks, the smaller the welfare loss, even though the losses due to the costs of wearing masks are negligible compared to the losses due to the costs of sickness (between 0.62% to 3.65%).

4.5.5 The relation between masks and participation

The possibility for agents to opt out of the market and stay home is now reintroduced in addition to the option to wear a mask, which revives the multiplicity of equilibria. As an illustration, I focus on equilibria where susceptible agents always coordinate on participating when multiple Nash equilibria exist ($x = 1$). The cost of masks is calibrated to the intermediate value.

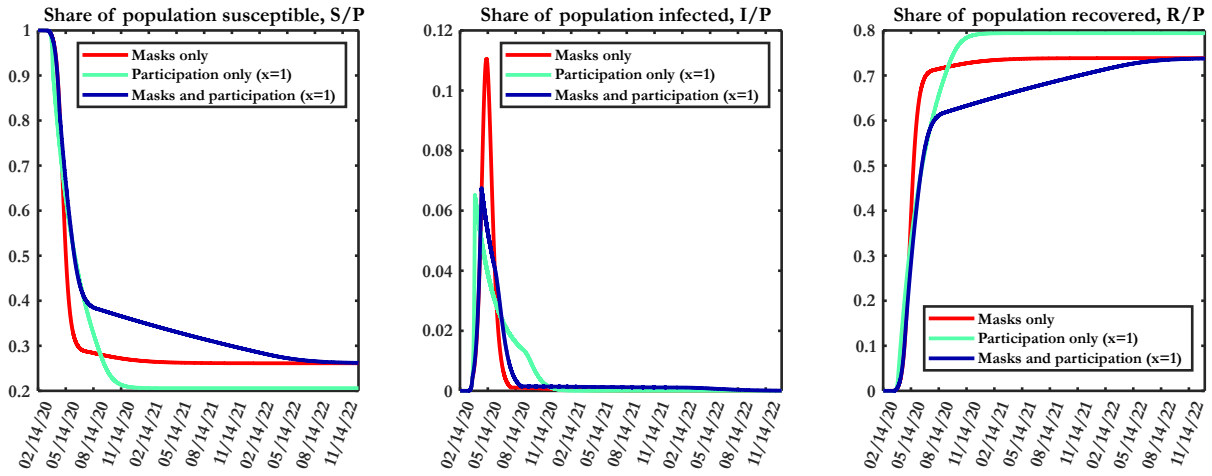

Figure 4.14: Time paths of epidemiological measures for the SIR model with mask-wearing and participation, assuming a medium cost for mask-wearing and coordination on participating (x=1). Note that 3 months lapse between each tick and label on the x-axis.

The time paths for epidemiological outcomes are represented in Figure 4.14. The middle panel displays the infection curve. The equilibrium path with both the mask and the participation margins is displayed in blue, and can be compared with the paths obtained when considering only masks (in red) or only participation (in green). We first notice that the infection curve is in between the two other curves. At its peak, active cases represent around 6.5% of the population, which is significantly lower than for the mask-only case, but a bit higher than the participation-only case. From the right panel, we can see that in the long run, the cumulative measure of agents that has been infected is very similar when considering the two margins as when considering masks only, even though in the short run, the curve is steeper when only considering masks.

Figure 4.15 provides some additional insights. The left panel represents the share of susceptible agents who participate. As expected, adding the mask margin dampens the reaction along the participation margin. Susceptible agents start to reduce their participation approximately two to three weeks later, and go back to full participation more than a month earlier. Additionally, the level of participation does not drop as much. At its lowest, it reaches around 67% of full participation, compared to 30% without the mask margin.

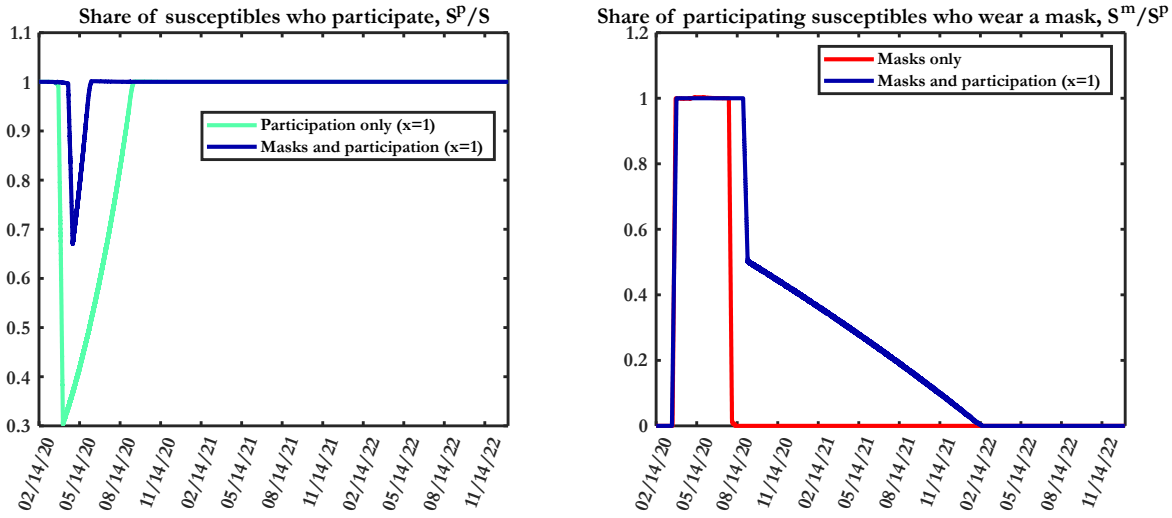

Figure 4.15: Time paths of participation and mask-wearing responses for the SIR model with mask-wearing and participation, assuming a medium cost for mask-wearing and coordination on participating (x=1). Note that 3 months lapse between each tick and label on the x-axis.

The right panel represents the share of participating susceptible agents who wear a mask, contrasting the equilibrium path when only the mask margin is active to that when the two margins are active. While the two paths originally overlap, they start to diverge just before August. At this point, susceptible agents entirely stop wearing masks when this is the only margin considered. On the other hand, when participation is also considered, mask-wearing continues at 100% for a few additional weeks, then declines steadily, but slowly, only reaching 0% around February of 2022. A key takeaway is that, maybe surprisingly, adding the participation margin reinforces the mask response. The intuition is that the mask margin is only effective for participating agents. If fewer agents participate, for a given share of mask wearers, infections are not as frequent, and immunity builds more slowly. In response, it becomes rational to keep wearing masks for a long time, as the number of active cases slowly vanishes.

4.6 Conclusion

This paper contrasted the impact of two response margins usually absent from epidemiological models of virus transmission—participation and mask-wearing—on infection dynamics and long-run outcomes in an economy where individuals gain utility from person-to-person social contacts. Rational agents and their decision making are embedded into a typical model of disease transmission in a micro-founded fashion, making use of search-and-matching methods to model interactions at a granular level.

When considering whether to participate in the matching process, individuals must weigh the risk of infection in each interaction with the benefits from that interaction. The more susceptible peers participate, the less risky the pool of participants, the smaller the risk of infection in a given interaction, and thus the higher the incentives for other susceptible agents to participate. These complementarities generate a continuum of equilibria and a rich set of dynamics, e.g. there can be multiple waves of infections, even absent any policy intervention. Another implication of the complementarities is that equilibria feature adverse selection, whereby susceptible agents leave the market while infectious agents stay, which comes at a very large welfare cost in the calibrated version of the model.

Predictions are markedly different when allowing agents to take precautions rather than to withdraw from the market. The equilibrium is unique, and infection dynamics are single-peaked. Mask-wearing being very efficient relative to its cost, both the human toll and welfare outcomes are considerably better under this specification than under the participation specification.

When both margins are considered simultaneously, a notable interplay arises. On one hand, the option to wear a mask reduces incentives to stay home, and dampens the drop in participation compared to a model with only a participation margin. On the other hand, the option to stay home makes the adoption of masks last longer, as a reduction in participation

means that fewer people get exposed to the virus, and herd immunity builds more slowly.

There are many interesting avenues for further research. One could explore the impact of different policies on behaviors, in particular when taking the two response margins into consideration. For example, could a mask mandate lead to undesirable outcomes by increasing the participation of agents? Another path would be to expand the model to multiple regions, e.g., with different densities, so as to investigate cross-regional contagion dynamics and policies such as border closures.

Bibliography

Afonso, G. and R. Lagos (2015): "Trade dynamics in the market for federal funds," *Econometrica*, 83, 263–313.

Anbarci, N., R. Dutu, and C.-J. Sun (2019): "On the timing of production decisions in monetary economies," *International Economic Review*, 60, 447–472.

Arora, N., P. Gandhi, and F. A. Longstaff (2012): "Counterparty credit risk and the credit default swap market," *Journal of Financial Economics*, 103, 280–293.

Aruoba, S. B., G. Rocheteau, and C. Waller (2007): "Bargaining and the Value of Money," *Journal of Monetary Economics*, 54, 2636–2655.

Atkeson, A. G., A. L. Eisfeldt, and P.-O. Weill (2015): "Entry and exit in OTC derivatives markets," *Econometrica*, 83, 2231–2292.

Aumann, R. J. and B. Peleg (1974): "A note on Gale's example," *Journal of Mathematical Economics*, 1, 209–211.

Barnichon, R. and A. Figura (2015): "Labor market heterogeneity and the aggregate matching function," *American Economic Journal: Macroeconomics*, 7, 222–49.

Baughman, G. and S. Rabinovich (2021): "Capacity choice, monetary trade, and the cost of inflation," *European Economic Review*, 134, 103698.

Bech, M. L. and E. Atalay (2010): "The topology of the federal funds market," *Physica A: Statistical Mechanics and its Applications*, 389, 5223–5246.

Berentsen, A., G. Menzio, and R. Wright (2011): "Inflation and unemployment in the long run," *American Economic Review*, 101, 371–98.

Berentsen, A. and G. Rocheteau (2003): "On the Friedman Rule in search models with divisible money," *Contributions in Macroeconomics*, 3.

Bethune, Z., B. Sultanum, and N. Trachter (2019): "Asset issuance in over-the-counter markets," *Review of Economic Dynamics*, 33, 4–29.

Bethune, Z. A. and A. Korinek (2020): "Covid-19 infection externalities: Trading off lives vs. livelihoods," working paper w27009, National Bureau of Economic Research.

Bi, Q., Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, et al. (2020): "Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study," *The Lancet Infectious Diseases*, 20, 911–919.

Blatter, M., S. Muehlemann, and S. Schenker (2012): "The costs of hiring skilled workers," *European Economic Review*, 56, 20–35.

Boeri, T., P. Garibaldi, and E. R. Moen (2018): "Financial constraints in search equilibrium: Mortensen Pissarides meet Holmstrom and Tirole," *Labour Economics*, 50, 144–155.

Brügemann, B., P. Gautier, and G. Menzio (2019): "Intra firm bargaining and Shapley values," *The Review of Economic Studies*, 86, 564–592.

Brunnermeier, M. K. and L. H. Pedersen (2009): "Market liquidity and funding liquidity," *The review of financial studies*, 22, 2201–2238.

Campello, M., E. Giambona, J. R. Graham, and C. R. Harvey (2011): "Liquidity management and corporate investment during a financial crisis," *The Review of Financial Studies*, 24, 1944–1979.

Chodorow-Reich, G. (2014): "The employment effects of credit market disruptions: Firm-level evidence from the 2008–9 financial crisis," *The Quarterly Journal of Economics*, 129, 1–59.

Chu, D. K., E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-harakeh, A. Bognanni, T. Lotfi, M. Loeb, et al. (2020): "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis," *The Lancet*, 395, 1973–1987.

Copeland, A., A. Martin, and M. Walker (2014): "Repo runs: Evidence from the tri-party repo market," *The Journal of Finance*, 69, 2343–2380.

Davis, S. J., R. J. Faberman, and J. C. Haltiwanger (2013): "The establishment-level behavior of vacancies and hiring," *The Quarterly Journal of Economics*, 128, 581–622.

Debreu, G. (1952): "A social equilibrium existence theorem," *Proceedings of the National Academy of Sciences*, 38, 886–893.

DeStefano, F., M. Haber, D. Currivan, T. Farris, B. Burrus, B. Stone-Wiggins, A. McCalla, H. Guled, H. Shih, P. Edelson, et al. (2011): "Factors associated with social contacts in four communities during the 2007–2008 influenza season," *Epidemiology & Infection*, 139, 1181–1190.

Di Maggio, M., A. Kermani, and Z. Song (2017): "The value of trading relations in turbulent times," *Journal of Financial Economics*, 124, 266–284.

Diamond, P. A. (1982): "Aggregate demand management in search equilibrium," *Journal of political Economy*, 90, 881–894.

Dudley, W. (2006): "Market and funding liquidity: an overview," Federal Reserve Bank of Atlanta 2016 Financial Markets Conference.

Duffie, D. (2011): *Dark markets: Asset pricing and information transmission in over-the-counter markets*, Princeton University Press.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005): "Over-the-counter markets," *Econometrica*, 73, 1815–1847.

Dugast, J., S. Üslü, and P.-O. Weill (2019): "A theory of participation in OTC and centralized markets," working paper w25887, National Bureau of Economic Research.

Dutu, R. and B. Julien (2008): "Ex-ante production, directed search and indivisible money," *Economics Bulletin*, 5.

Eichenbaum, M. S., S. Rebelo, and M. Trabandt (2020): "The macroeconomics of epidemics," working paper w26882, National Bureau of Economic Research.

Eisfeldt, A. L., B. Herskovic, S. Rajan, and E. Siriwardane (2018): "OTC intermediaries," working paper 3245966, SSRN.

Elsby, M. W., R. Michaels, and D. Ratner (2015): "The Beveridge curve: A survey," *Journal of Economic Literature*, 53, 571–630.

Engle, S., J. Keppo, E. K. M. Quercioli, L. Smith, and A. Wilson (2020): "The Behavioral SIR Model, with Applications to the Swine Flu and COVID-19 Pandemics," Tech. rep., University of Wisconsin.

Fan, K. (1952): "Fixed-point and minimax theorems in locally convex topological linear spaces," *Proceedings of the National Academy of Sciences of the United States of America*, 38, 121.

Farboodi, M., G. Jarosch, and R. Shimer (2020): "Internal and external effects of social distancing in a pandemic," working paper w27059, National Bureau of Economic Research.

Feehan, D. M. and C. Cobb (2019): "Using an online sample to estimate the size of an offline population," *Demography*, 56, 2377–2392.

Fershtman, C. (1990): "The importance of the agenda in bargaining," *Games and Economic Behavior*, 2, 224–238.

Fortune, P. et al. (2001): "Margin lending and stock market volatility," *New England Economic Review*, 3–26.

Friewald, N. and F. Nagler (2019): "Over-the-counter market frictions and yield spread changes," *The Journal of Finance*, 74, 3217–3257.

Gale, D. (1974): "Exchange equilibrium and coalitions: an example," *Journal of Mathematical Economics*, 1, 63–66.

Garibaldi, P., E. R. Moen, and C. A. Pissarides (2020): "Modelling contacts and transitions in the SIR epidemics model," *Covid Economics*, 5.

Gavazza, A. (2011): "The role of trading frictions in real asset markets," *American Economic Review*, 101, 1106–43.

Gavazza, A., S. Mongey, and G. L. Violante (2018): "Aggregate recruiting intensity," *American Economic Review*, 108, 2088–2127.

Geromichalos, A. and L. Herrenbrueck (2016a): "Monetary policy, asset prices, and liquidity in over-the-counter markets," *Journal of Money, Credit and Banking*, 48, 35–79.

——— (2016b): "The strategic determination of the supply of liquid assets," working paper 16-1, UC Davis.

Geromichalos, A., J. M. Licari, and J. Suárez-Lledó (2007): "Monetary policy and asset prices," *Review of Economic Dynamics*, 10, 761–779.

Glicksberg, I. L. (1952): "A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points," *Proceedings of the American Mathematical Society*, 3, 170–174.

Gorton, G. and A. Metrick (2012): "Securitized banking and the run on repo," *Journal of Financial economics*, 104, 425–451.

Green, R. C., B. Hollifield, and N. Schürhoff (2007): "Financial intermediation and the costs of trading in an opaque market," *The Review of Financial Studies*, 20, 275–314.

Greenstone, M., A. Mas, and H.-L. Nguyen (2020): "Do credit market shocks affect the real economy? Quasi-experimental evidence from the great recession and" normal" economic times," *American Economic Journal: Economic Policy*, 12, 200–225.

Gromb, D. and D. Vayanos (2002): "Equilibrium and welfare in markets with financially constrained arbitrageurs," *Journal of financial Economics*, 66, 361–407.

Gu, C., F. Mattesini, and R. Wright (2016): "Money and credit redux," *Econometrica*, 84, 1–32.

Hall, R. E., C. I. Jones, and P. J. Klenow (2020): "Trading off consumption and covid-19 deaths," working paper w27340, National Bureau of Economic Research.

Hall, R. E. and S. Schulhofer-Wohl (2018): "Measuring job-finding rates and matching efficiency with heterogeneous job-seekers," *American Economic Journal: Macroeconomics*, 10, 1–32.

Hendershott, T., D. Li, D. Livdan, and N. Schürhoff (2017): "Relationship trading in OTC markets," *Swiss Finance Institute Research Paper*.

Hollifield, B., A. Neklyudov, and C. Spatt (2017): "Bid-ask spreads, trading networks, and the pricing of securitizations," *The Review of Financial Studies*, 30, 3048–3085.

Holmstrom, B. and J. Tirole (1997): "Financial intermediation, loanable funds, and the real sector," *the Quarterly Journal of economics*, 112, 663–691.

Hornstein, A. and M. Kudlyak (2016): "Estimating matching efficiency with variable search effort," working paper, FRB Richmond.

Hu, T.-W. and G. Rocheteau (2013): "On the coexistence of money and higher-return assets and its social role," *Journal of Economic Theory*, 148, 2520–2560.

——— (2020): "Bargaining under liquidity constraints: Unified strategic foundations of the Nash and Kalai solutions," *Journal of Economic Theory*, 189, 105098.

Hugonnier, J., B. Lester, and P.-O. Weill (2020): "Frictional intermediation in over-the-counter markets," *The Review of Economic Studies*, 87, 1432–1469.

Jermann, U. and V. Quadrini (2012): "Macroeconomic effects of financial shocks," *American Economic Review*, 102, 238–71.

Kaas, L. and P. Kircher (2015): "Efficient firm dynamics in a frictional labor market," *American Economic Review*, 105, 3030–60.

Kalai, E. (1977): "Proportional solutions to bargaining situations: interpersonal utility comparisons," *Econometrica: Journal of the Econometric Society*, 1623–1630.

Kalai, E. and M. Smorodinsky (1975): "Other solutions to Nash's bargaining problem," *Econometrica: Journal of the Econometric Society*, 513–518.

Kermack, W. O. and A. G. McKendrick (1927): "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115, 700–721.

Kremer, M. (1996): "Integrating behavioral choice into epidemiological models of AIDS," *The Quarterly Journal of Economics*, 111, 549–573.

Kreps, D. M. and J. A. Scheinkman (1983): "Quantity precommitment and Bertrand competition yield Cournot outcomes," *The Bell Journal of Economics*, 326–337.

Krueger, A. B. (2017): "Where have all the workers gone? An inquiry into the decline of the US labor force participation rate," *Brookings papers on economic activity*, 2017, 1.

Krueger, D., H. Uhlig, and T. Xie (2020): "Macroeconomic dynamics and reallocation in an epidemic," working paper w27047, National Bureau of Economic Research.

Lagos, R. (2010): "Asset prices and liquidity in an exchange economy," *Journal of Monetary Economics*, 57, 913–930.

——— (2013): "Moneyspots: extraneous attributes and the coexistence of money and interest-bearing nominal bonds," *Journal of Political Economy*, 121, 127–185.

Lagos, R. and G. Rocheteau (2005): "Inflation, output, and welfare," *International Economic Review*, 46, 495–522.

——— (2007): "Search in asset markets: Market structure, liquidity, and welfare," *American Economic Review*, 97, 198–202.

——— (2008): "Money and capital as competing media of exchange," *Journal of Economic theory*, 142, 247–258.

——— (2009): "Liquidity in asset markets with search frictions," *Econometrica*, 77, 403–426.

Lagos, R., G. Rocheteau, and P.-O. Weill (2011): "Crises and liquidity in over-the-counter markets," *Journal of Economic Theory*, 146, 2169–2205.

Lagos, R., G. Rocheteau, and R. Wright (2017): "Liquidity: A new monetarist perspective," *Journal of Economic Literature*, 55, 371–440.

Lagos, R. and R. Wright (2005): "A unified framework for monetary theory and policy analysis," *Journal of political Economy*, 113, 463–484.

Lagos, R. and S. Zhang (2019a): "A monetary model of bilateral over-the-counter markets," *Review of Economic Dynamics*, 33, 205–227.

——— (2019b): "On money as a medium of exchange in near-cashless credit economies," working paper w25803, National Bureau of Economic Research.

——— (2020): "Turnover liquidity and the transmission of monetary policy," *American Economic Review*, 110, 1635–72.

Lebeau, L. (2020): "Credit frictions and participation in over-the-counter markets," *Journal of Economic Theory*, 189, 105100.

Leduc, S. and Z. Liu (2016): "Uncertainty shocks are aggregate demand shocks," *Journal of Monetary Economics*, 82, 20–35.

Leung, K., M. Jit, E. H. Lau, and J. T. Wu (2017): "Social contact patterns relevant to the spread of respiratory infectious diseases in Hong Kong," *Scientific reports*, 7, 1–12.

Li, D. and N. Schürhoff (2019): "Dealer networks," *The Journal of Finance*, 74, 91–144.

Li, T., Y. Liu, M. Li, X. Qian, and S. Y. Dai (2020): "Mask or no mask for COVID-19: A public health and market study," *PloS one*, 15, e0237691.

Li, Y., G. Rocheteau, and P.-O. Weill (2012): "Liquidity and the threat of fraudulent assets," *Journal of Political Economy*, 120, 815–846.

Liu, Y., A. A. Gayle, A. Wilder-Smith, and J. Rocklöv (2020): "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *Journal of travel medicine*.

Luo, L., D. Liu, X. Liao, X. Wu, Q. Jing, J. Zheng, F. Liu, S. Yang, H. Bi, Z. Li, et al. (2020): "Contact settings and risk for transmission in 3410 close contacts of patients with COVID-19 in Guangzhou, China: a prospective cohort study," *Annals of internal medicine*, 173, 879–887.

Macchiavelli, M. and X. A. Zhou (2019): "unding liquidity and market liquidity: the broker-dealer perspective," working paper 3311786, SSRN.

Mann, R. J. (1996): "Searching for Negotiability in Payment and Credit Systems," *Ucla L. Rev.*, 44, 951.

Masters, A. (2013): "Inflation and welfare in retail markets: prior production and imperfectly directed search," *Journal of Money, Credit and Banking*, 45, 821–844.

McAdams, D. (2020a): "Economic epidemiology in the wake of Covid-19," *economics*, 82120, 122900.

——— (2020b): "Nash sir: An economic-epidemiological model of strategic behavior during a viral epidemic," *Covid Economics*.

Mills, K. and B. McCarthy (2014): "The state of small business lending," *Harvard Business School*.

Monacelli, T., V. Quadrini, and A. Trigari (2011): "Financial markets and unemployment," working paper w17389, National Bureau of Economic Research.

Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. (2008): "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med*, 5, e74.

Nash, J. F. (1950): "The bargaining problem," *Econometrica: Journal of the Econometric Society*, 155–162.

O'Keeffe, D. (2018): "Understanding cryptocurrency transaction speeds," `https://medium.com/coinmonks/`, accessed: 2018-06-05.

O'Neill, B., D. Samet, Z. Wiener, and E. Winter (2004): "Bargaining with an agenda," *Games and Economic Behavior*, 48, 139–153.

Oran, D. P. and E. J. Topol (2020): "Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review," *Annals of internal medicine*, 173, 362–367.

Osborne, M. J. and A. Rubinstein (1990): *Bargaining and markets*, Academic Press Limited.

Pagnotta, E. S. and T. Philippon (2018): "Competing on speed," *Econometrica*, 86, 1067–1115.

Petrongolo, B. and C. A. Pissarides (2001): "Looking into the black box: A survey of the matching function," *Journal of Economic literature*, 39, 390–431.

Petrosky-Nadeau, N. (2014): "Credit, vacancies and unemployment fluctuations," *Review of Economic Dynamics*, 17, 191–205.

Petrosky-Nadeau, N. and E. Wasmer (2013): "The cyclical volatility of labor markets under frictional financial markets," *American Economic Journal: Macroeconomics*, 5, 193–221.

Pissarides, C. A. (2000): *Equilibrium unemployment theory*, MIT press.

——— (2009): "The unemployment volatility puzzle: Is wage stickiness the answer?" *Econometrica*, 77, 1339–1369.

Pruitt, S. W. and K. Tse (1996): "The price, volatility, volume, and liquidity effects of changes in Federal Reserve margin requirements on both marginable and nonmarginable OTC stocks," in *The Industrial Organization and Regulation of the Securities Industry*, University of Chicago Press, 317–358.

Randall, O. (2015): "How Do Inventory Costs Affect Dealer Behavior in the US Corporate Bond Market?" working paper 2590331, SSRN.

Rapp, A. C. (2018): "Middlemen matter: Corporate bond market liquidity and dealer inventory funding," working paper 2867531, SSRN.

Rocheteau, G. (2011): "Payments and liquidity under adverse selection," *Journal of Monetary Economics*, 58, 191–205.

Rocheteau, G. and E. Nosal (2017): *Money, payments, and liquidity*, MIT Press.

Rocheteau, G. and R. Wright (2005): "Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium," *Econometrica*, 73, 175–202.

——— (2013): "Liquidity and asset-market dynamics," *Journal of Monetary Economics*, 60, 275–294.

Rubinstein, A. (1982): "Perfect equilibrium in a bargaining model," *Econometrica: Journal of the Econometric Society*, 97–109.

Rubinstein, A. and A. Wolinsky (1985): "Equilibrium in a market with sequential bargaining," *Econometrica: Journal of the Econometric Society*, 1133–1150.

Şahin, A., J. Song, G. Topa, and G. L. Violante (2014): "Mismatch unemployment," *American Economic Review*, 104, 3529–64.

Salanié, F. and N. Treich (2020): "Public and private incentives for self-protection," *The Geneva Risk and Insurance Review*, 45, 104–113.

Sedláček, P. (2014): "Match efficiency and firms' hiring standards," *Journal of Monetary Economics*, 62, 123–133.

Serrano, R. (2008): "Nash program," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf and L. E. Blume, London: McMillan.

Shapley, L. S. (1969): "Utility comparison and the theory of games," *La decision*, 307–319.

Shi, S. (1995): "Money and prices: a model of search and bargaining," *Journal of Economic Theory*, 67, 467–496.

Ståhl, I. (1972): *Bargaining theory*, Stockholm: The Economic Research Institute at the Stockholm School of Economics.

Stole, L. A. and J. Zwiebel (1996): "Intra-firm bargaining under non-binding contracts," *The Review of Economic Studies*, 63, 375–410.

Toxvaerd, F. M. (2020): "Equilibrium social distancing," working paper, University of Cambridge.

Trejos, A. and R. Wright (1995): "Search, bargaining, money, and prices," *Journal of political Economy*, 103, 118–141.

——— (2016): "Search-based models of money and finance: An integrated approach," *Journal of Economic Theory*, 164, 10–31.

Üslü, S. (2019): "Pricing and liquidity in decentralized asset markets," *Econometrica*, 87, 2079–2140.

Vayanos, D. and P.-O. Weill (2008): "A search-based theory of the on-the-run phenomenon," *The Journal of Finance*, 63, 1361–1398.

Wallace, N. et al. (1998): "A dictum for monetary theory," *Federal Reserve Bank of Minneapolis Quarterly Review*, 22, 20–26.

Wasmer, E. and P. Weil (2004): "The macroeconomics of labor and credit market imperfections," *American Economic Review*, 94, 944–963.

Weill, P.-O. (2007): "Leaning against the wind," *The Review of Economic Studies*, 74, 1329–1354.

Wiener, Z. and E. Winter (1998): "Gradual bargaining," working paper OLIN-98-02, Washington University.

Worobey, M., J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, and P. Lemey (2020): "The emergence of SARS-CoV-2 in Europe and the US," *BioRxiv*.

Wright, R., S. X. Xiao, and Y. Zhu (2020): "Frictional capital reallocation with ex post heterogeneity," *Review of Economic Dynamics*, 37, S227–S253.

Zarutskie, R. and T. Yang (2016): "How did young firms fare during the great recession? evidence from the kauffman firm survey," in *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, University of Chicago Press, 253–290.

Zhu, T. and N. Wallace (2007): "Pairwise trade and coexistence of money and higher-return assets," *Journal of Economic Theory*, 133, 524–535.

Appendix A

Supplementary material for Chapter 1

A.1 Proofs

Proof of Lemma 1.1. First, note that the functions $p(y)$ and $k(y)$ are strictly increasing in y (thus invertible) and coincide when evaluated at y^* if $y^* < \infty$. In order to solve for (1.2), we can set up a Lagrangian,

$$\mathcal{L} = \theta \log [u(y) - p] + (1 - \theta) \log [p - c(y)] - \lambda_p(p - z) - \lambda_y(y - w), \tag{A.1}$$

where λ_p is the Lagrange multiplier on the payment capacity constraint and λ_y is the Lagrange multiplier on the inventory constraint. The Kuhn-Tucker conditions are

$$\frac{\theta u'(y)}{u(y) - p} - \frac{(1 - \theta)c'(y)}{p - c(y)} - \lambda_y = 0, \lambda_y \geq 0, y \leq w, \lambda_y(y - w) = 0 \tag{A.2}$$

$$\frac{1 - \theta}{p - c(y)} - \frac{\theta}{u(y) - p} - \lambda_p = 0, \lambda_p \geq 0, p \leq z, \lambda_p(p - z) = 0. \tag{A.3}$$

First, consider $\lambda_y > 0$ and $\lambda_p = 0$. Then $y = w < y^*$, and from (A.3) we get $p = (1-\theta)u(w) + \theta c(w) = k(w)$. The payment constraint $p \leq z$ implies $z \geq k(w)$. Second, consider $\lambda_y > 0$ and $\lambda_p > 0$. Then $y = w < y^*$ while $p = z$, so that $\lambda_y > 0$ and $\lambda_p > 0$ requires $p(w) \leq z \leq k(w)$. Third, consider $\lambda_y = 0$ and $\lambda_p > 0$. Then $p = z$, and from (A.2) we get $y = p^{-1}(z)$. The inventory constraint requires $y \leq w$, which implies $p^{-1}(z) \leq w$, or equivalently, $z \leq p(w)$. Finally, when $\lambda_y = \lambda_p = 0$, it must be that $u'(y) = c'(y)$, implying $y = y^*$ from (A.2) and $p = k(y^*) = p(y^*)$ from (A.3). The capacity constraints require $y^* \leq w$ and $k(y^*) \leq z$. When $z = 0$ or $w = 0$, it is trivial to show that the solution to the Nash bargaining problem is $p = y = 0$. □

Derivations for Section 1.2.3. We prove the claim according to which the trade breakdown described in Proposition 1.1 is robust to using the gradual bargaining mechanism proposed by Hu and Rocheteau (2020) for any $N < \infty$. It is assumed that when an offer is rejected by player $j \in \{h, \ell\}$, the round is over and players move on to the next bundle with probability $1 - \xi^j$, and the game ends with the complement probability. Assuming that $\xi^h = e^{-(1-\theta)\varepsilon}$ and $\xi^\ell = e^{-\theta\varepsilon}$, and taking the limit as $\varepsilon \to 0$, the unique subgame-perfect Nash equilibrium of the game is given by the last term of the sequence $\{(y_n, p_n)\}_{n=1}^N$ where $(y_0, p_0) = (0, 0)$ and

$$\begin{aligned}(y_n, p_n) \in \arg\max_{y,p} \left[u(y) - u(y_{n-1}) - (p - p_{n-1})\right]^\theta \left[(p - p_{n-1}) - c(y) + c(y_{n-1})\right]^{1-\theta} \\ \text{s.t. } y - y_{n-1} \leq \min(y^*, w)/N \text{ and } p \leq z.\end{aligned} \tag{A.4}$$

Let $N \in \mathbb{N}^+$ and $\bar{y}_n \equiv n \min(y^*, w)/N$. The solution to (A.4) is given by

$$\begin{cases} p_n = z = g(y_n, \bar{y}_{n-1}) & \text{if } z \leq g(\bar{y}_n, \bar{y}_{n-1}) \\ p_n = z \text{ and } y_n = \bar{y}_n & \text{if } g(\bar{y}_n, \bar{y}_{n-1}) < z < k(\bar{y}_n) \\ p_n = k(\bar{y}_n) \text{ and } y_n = \bar{y}_n & \text{if } z \geq k(\bar{y}_n), \end{cases} \tag{A.5}$$

where

$$g(y,\bar{y}) \equiv p(y) + [\Theta(y) - \theta]\,[u(\bar{y}) - c(\bar{y})]\,. \tag{A.6}$$

The payment function $g(y,\bar{y})$ is increasing in y and $\bar{y}$, with $g(y,0) = p(y)$ and $g(y,y) = k(y)$. From (A.5) we can derive the players' surpluses from trade,

$$\begin{aligned} S_m^h(z,w) = &\sum_{n=1}^{N} \mathbb{I}_{\{k(\bar{y}_{n-1}) \le z < g(\bar{y}_n,\bar{y}_{n-1})\}} \{u\,[y_n(z)] - z\} + \mathbb{I}_{\{g(\bar{y}_n,\bar{y}_{n-1}) < z < k(\bar{y}_n)\}}\,[u(\bar{y}_n) - z] \\ &+ \mathbb{I}_{\{z \ge k[\min(w,y^*)]\}} \theta\,\{u\,[\min(w,y^*)] - c\,[\min(w,y^*)]\}\,, \end{aligned} \tag{A.7}$$

and

$$\begin{aligned} S_m^\ell(z,w) = &\sum_{n=1}^{N} \mathbb{I}_{\{k(\bar{y}_{n-1}) \le z < g(\bar{y}_n,\bar{y}_{n-1})\}} \{z - c\,[y_n(z)]\} + \mathbb{I}_{\{g(\bar{y}_n,\bar{y}_{n-1}) < z < k(\bar{y}_n)\}}\,[z - c(\bar{y}_n)] \\ &+ \mathbb{I}_{\{z \ge k[\min(w,y^*)]\}} (1-\theta)\,\{u\,[\min(w,y^*)] - c\,[\min(w,y^*)]\}\,. \end{aligned} \tag{A.8}$$

Note that player h's surplus is strictly decreasing when $g(\bar{y}_n,\bar{y}_{n-1}) < z < k(\bar{y}_n)$ for all $n = 1,2,...,N$, first increasing then potentially decreasing when $k(\bar{y}_{n-1}) < z < g(\bar{y}_n,\bar{y}_{n-1})$, and constant when $z \ge k(\bar{y}_n) = k\,[\min(w,y^*)]$ (which is preceded by a decreasing piece, on $g(\bar{y}_N,\bar{y}_{N_1}) < z < k(\bar{y}_N)$). As a result, a solution to (1.11) can only belong to $U_{n=1}^N(k(\bar{y}_{n-1}), g(\bar{y}_n,\bar{y}_{n-1})]$. But for any $z \in U_{n=1}^N(k(\bar{y}_{n-1}), g(\bar{y}_n,\bar{y}_{n-1})]$, player ℓ's surplus could be increased by reducing w, implying that a solution to (1.12) would never lie in this interval. This proves that the two players' best responses cannot intersect for $(w,z) > 0$. Because $S^h(z,w) = S^\ell(z,w) = 0$ when $z = 0$ or $w = 0$, $z = w = 0$ is a subgame perfect Nash equilibrium, and it is the unique one when $N < \infty$.

Now, when $N \to \infty$, Proposition 3 of Hu and Rocheteau (2020) proves that the final allocation (y_N, p_N) converges to (y,p) such that $p = k(y) = \min\,[z, k[\min(y^*,w)]]$, which is the allocation we would obtain under Kalai bargaining, and the surpluses converge to $S_k^h(z,w)$

and $S_k^\ell(z,w)$, derived in the following paragraph. Those surpluses are monotone, and from 1.2.4 we know that a monetary equilibrium exists for i and ψ low enough. □

Derivations for Section 1.2.4. Under Kalai bargaining, the players' surpluses are given by

$$S_k^h(z,w) = \begin{cases} \theta\left\{u\left[\min(w,y^*)\right] - c\left[\min(w,y^*)\right]\right\} & \text{if } z \geq k\left[\min(w,y^*)\right] \\ \theta\left\{u\left[k^{-1}(z)\right] - c\left[k^{-1}(z)\right]\right\} & \text{otherwise,} \end{cases} \tag{A.9}$$

and $S_k^\ell(z,w) = (1-\theta)S_k^h(z,w)/\theta$. Best-responses are obtained by solving for player h's problem, $\max_{z\in\mathcal{R}^+}\left\{-iz + S_k^h(z,w)\right\}$ and player ℓ's problem, $\max_{w\in[0,\Omega]}\left\{-\psi w + S_k^\ell(z,w)\right\}$. □

Proof of Lemma 1.2. For (i), note that in bilateral meetings where credit is available, the Nash problem can be written as

$$\max_{y_c,p_c}(\varepsilon y_c - p_c)^\theta (p_c - y_c)^{1-\theta} \text{ s.t } y_c \leq w. \tag{A.10}$$

Because the Nash solution is Pareto efficient, we must have $y_c = w$. Taking the first order condition with respect to p_c yields $p_c = \left[(1-\theta)\varepsilon + \theta\right] y_c = \delta_2 y_c$. Part (ii) directly follows from Lemma 1.1, making use of $u(y) = \varepsilon y$, $c(y) = y$, and $y^* = \infty$. □

Proof of 1.3. In order to solve for the optimal amount of real balances carried from a period to another by the h-investor, one needs to solve for the maximization problem in (1.5). To do so, we first need to write the equation for $V^h(z)$,

$$\begin{aligned} V^h(z) = \max_{0\leq z^p\leq z} \mathbb{E}_{\tilde{w}}\Big\{ & \gamma\alpha\left[\varepsilon y_c(\tilde{w}) + W^h(z - p_c(\tilde{w}))\right] \\ & + \gamma(1-\alpha)\left[\varepsilon y_m(z,\tilde{w}) + W^h(z - p_m(z,\tilde{w}))\right] + (1-\gamma)W^h(z)\Big\}. \end{aligned} \tag{A.11}$$

Making use of the linearity of W^h, this simplifies to (1.7). Plugging (1.7) into (1.5), and

making use of $i \equiv (1+\pi)/\beta$, the maximization problem can then be reduced to

$$\max_{z\geq 0}\left\{-iz + \max_{0\leq z^p\leq z} \gamma \mathbb{E}_{\tilde{w}}\left[\alpha S_c^h(\tilde{w}) + (1-\alpha)S_m^h(z^p,\tilde{w})\right]\right\}, \tag{A.12}$$

from which we directly obtain $z^p = z$ for h-investors as long as $i > 0$. We can follow similar steps to solve the problem of an ℓ-investor. We start with $V^\ell(z)$,

$$\begin{aligned} V^\ell(z) = \max_{0\leq z^p\leq z, 0\leq w\leq \Omega} \mathbb{E}_{\tilde{z}}\Big\{ & \gamma\alpha\left[\Omega - y_c(w) + W^\ell(z + p_c(w))\right] \\ & + \gamma(1-\alpha)\left[\Omega - y_m(\tilde{z},w) + W^\ell(z+p_m(\tilde{z},w))\right] + (1-\gamma)\left[\Omega + W^\ell(z)\right]\Big\}. \end{aligned} \tag{A.13}$$

Making use of the linearity of W^ℓ, this simplifies to (1.8). Plugging (1.8) into (1.5), we obtain the following maximization problem,

$$\max_{z\geq 0}\left\{-iz + \max_{0\leq z^p\leq z, 0\leq w\leq \Omega} \gamma \mathbb{E}_{\tilde{z}}\left[\alpha S_c^\ell(w) + (1-\alpha)S_m^\ell(\tilde{z},w)\right]\right\}, \tag{A.14}$$

from which we get $z = z^p = 0$ for ℓ-investors. Part (ii) follows. □

Proof of Proposition 1.2. Parts (i.i), (i.ii) and (ii.ii) are proven in the text. A proof of the existence result in (ii.i) can be found in the proof of Proposition 1.3. We provide an additional proof here, making use of Glicksberg (1952) fixed-point theorem. To do so, we first define the game formally. Denote $\mathcal{I} = \{h, \ell\}$ the set of players, indexed by i. Player ℓ's strategy set is $S^\ell = [0, \Omega]$, while player h's strategy set is $S^h = [0, \delta_1\Omega]$. The upper bound $\delta_1\Omega$ is required in order for this strategy set to be compact. We know that player h would never carry more than $\delta_1\Omega$ as long as $i > 0$ (if $i = 0$, we can impose an arbitrarily large upper bound, which we know is without loss of generality since the value of money must be bounded in equilibrium). The players' payoffs can be written as

$$u_h(s_h, s_\ell) = -is_h + \gamma\left[\alpha\theta(\varepsilon - 1)s_\ell + (1-\alpha)S_m^h(s_h, s_\ell)\right], \tag{A.15}$$

and

$$u_\ell(s_\ell, s_h) = \gamma \left[\alpha(1-\theta)(\varepsilon-1)s_\ell + (1-\alpha)S_m^\ell(s_h, s_\ell)\right]. \tag{A.16}$$

In this set up, Debreu (1952), Glicksberg (1952) and Fan (1952) prove the existence of a pure-strategy Nash equilibrium so long as for all $i \in \mathcal{I}$, (i) S_i is non-empty, compact and convex, (ii) $u_i(s_i, s_{-i})$ is continuous in s_{-i}, and (iii) $u_i(s_i, s_{-i})$ is continuous and quasiconcave in s_i. While those three conditions hold when $\alpha \geq 1/\delta_2$, where Proposition 1.3 indeed predicts a symmetric equilibrium (N=1), condition (iii) is not satisfied when $\alpha < 1/\delta_2$. Indeed, in this parameter region, $u_\ell(s_\ell, s_h)$ is not quasiconcave in s_ℓ for $s_h < \delta_1 s_\ell$ (it is increasing on $[0, s_h/\delta_2]$, decreasing on $[s_h/\delta_2, s_h/\delta_1]$, and increasing on $[s_h/\delta_1, \Omega]$).

We then turn to another existence theorem that does not require the quasiconcavity of payoffs. Glicksberg (1952) proves the existence of a mixed-strategy Nash equilibrium so long as for all $i \in \mathcal{I}$, (i) S_i is non-empty and compact and (ii) $u_i(s_i, s_{-i})$ is continuous in s_i and s_{-i}. These two conditions are satisfied for any $\alpha \in (0, 1)$. In order to prove, more specifically, the existence of a monetary equilibrium when $\alpha < 1 - i/\left[\gamma\theta(\varepsilon-1)\right]$, we restrict the strategy sets to $S_h = [\epsilon_h, \delta_1\Omega]$ and $S_\ell = [\epsilon_\ell, \Omega]$, where ϵ_h and ϵ_ℓ are strictly positive but arbitrarily close to 0. Conditions (i) and (ii) still hold. We need to check that were the strategies $s_h = 0$ and $s_\ell = 0$ part of S_h and S_ℓ, they would not constitute a profitable deviation for either player.

We know from Lemma 1.4 that in any monetary equilibrium, $\bar{s}_\ell = \Omega$. Thus, in any monetary equilibrium, player ℓ's payoff is

$$v^\ell(\Omega) = \gamma\left[\alpha(1-\theta)(\varepsilon-1)\Omega + (1-\alpha)(1-\theta)(\varepsilon-1)\mathbb{E}(s_h)/\varepsilon\right] > 0. \tag{A.17}$$

Were player ℓ to deviate and pick $s_\ell = 0$, there would be no trade, leading to a payoff of 0.

Thus, player ℓ would not deviate.

Similarly, player h's payoff is

$$v^h(s_h) = -is_h + \gamma\left[\alpha\theta(\varepsilon-1)\mathbb{E}(s_\ell) + (1-\alpha)\mathbb{E}(S_m^h(s_h, s_\ell))\right]. \tag{A.18}$$

If she were to deviate and pick $s_h = 0$, she would not be able to trade in money-only meetings, and her payoff would be $v^h(0) = \gamma\alpha\theta(\varepsilon-1)\mathbb{E}(s_\ell)$. In any equilibrium obtained when $S_h = [\epsilon_h, \delta_1\Omega]$, $v^h(s_h) \geq v^h(\delta_1\underline{s}_\ell) = -i\delta_1\underline{s}_\ell + v^h(0) + \gamma(1-\alpha)\theta(\varepsilon-1)\delta_1\underline{s}_\ell$. Now, $v^h(\delta_1\underline{s}_\ell) > v^h(0)$ if and only if $i < \gamma\theta(1-\alpha)(\varepsilon-1)$, which is satisfied when $\alpha < 1 - i/\left[\gamma\theta(\varepsilon-1)\right]$. Hence, in any monetary equilibrium, $v^h(s_h) \geq v^h(\delta_1\underline{s}_\ell) > v^h(0)$, so that enlarging the strategy set to $S_h = [0, \delta_1\Omega]$ would not lead to any deviation to $s_h = 0$ from player h. As a result, there must exist a monetary equilibrium whenever $0 < \alpha < 1 - i/\left[\gamma\theta(\varepsilon-1)\right]$. □

Proof of Lemma 1.4. We first prove that $\bar{z} \leq \delta_1\bar{w}$. Consider $\bar{z}' > \delta_1\bar{w}$. The h-investor's net payoff from carrying $z = \delta_1\bar{w}$ is

$$v^h(\delta_1\bar{w}) = -i\delta_1\bar{w} + \gamma(1-\alpha)\int S_m^h(\delta_1\bar{w}, w)dF^w(w), \tag{A.19}$$

while her net payoff from carrying $z = \bar{z}'$ is

$$v^h(\bar{z}') = -i\bar{z}' + \gamma(1-\alpha)\int S_m^h(\bar{z}', w)dF^w(w). \tag{A.20}$$

For any $w > \delta 1\bar{w}/\delta_2$, $\delta_1 w \leq \delta_1\bar{w} < \delta_2 w$, so that $S_m^h(\bar{z}', w) < S_m^h(\delta_1\bar{w}, w)$. For any $w \leq \delta_1\bar{w}/\delta_2$, $\delta_1\bar{w} \geq \delta_2 w$, so that $S_m^h(\bar{z}', w) = S_m^h(\delta_1\bar{w}, w)$. As a result, $\int S_m^h(\delta_1\bar{w}, w)dF^w(w) > \int S_m^h(\bar{z}, w)dF^w(w)$, and $v^h(\delta_1\bar{w}) > v^h(\bar{z}')$ as long as $i \geq 0$, so that $\bar{z} \leq \delta_1\bar{w}$.

We now show that $\underline{z} \geq \delta_1\underline{w}$. Consider $\underline{z}' < \delta_1\underline{w}$. The h-investor's net payoff from carrying

$z = \delta_1 \underline{w}$ is

$$v^h(\delta_1 \underline{w}) = -i\delta_1 \underline{w} + \gamma(1-\alpha) \int S_m^h(\delta_1 \underline{w}, w) dF^w(w), \tag{A.21}$$

while her net payoff from carrying $z = \underline{z}'$ is

$$v^h(\underline{z}') = -i\underline{z}' + \gamma(1-\alpha) \int S_m^h(\underline{z}', w) dF^w(w). \tag{A.22}$$

Now, note that $z \leq \delta_1 \underline{w}$ implies $z \leq \delta_1 w$ for any $w \in \mathbb{W}$, so that $S_m^h(z, w) = \theta(\varepsilon - 1)z$. Then, $v^h(\underline{z}') > v^h(\delta_1 \underline{w})$ requires $\underline{z}' [\gamma(1-\alpha)\theta(\varepsilon-1) - i] > \delta_1 \underline{w} [\gamma(1-\alpha)\theta(\varepsilon-1) - i]$. Because $\underline{z}' < \delta_1 \underline{w}$, this implies $\gamma(1-\alpha)\theta(\varepsilon-1) - i < 0$. Only when $\gamma(1-\alpha)\theta(\varepsilon-1) - i \geq 0$ can a monetary equilibrium exist, so that in any monetary equilibrium, $v^h(\underline{z}') < v^h(\delta_1 \underline{w})$, and $\underline{z} \geq \delta_1 \underline{w}$.

Third, we prove that $\underline{w} \geq \min(\Omega, \underline{z}/\delta_2)$. Note that $w \leq \underline{z}/\delta_2$ implies $w \leq z/\delta_2$ for any $z \in \mathbb{Z}$, so that $S_m^\ell(z, w) = (1-\theta)(\varepsilon-1)w$. As a result, for any $w \leq \underline{z}/\delta_2$, $v^\ell(w) = (1-\theta)(\varepsilon-1)w$, strictly increasing in w. It follows that $\underline{w} \geq \min(\Omega, \underline{z}/\delta_2)$.

Finally, we prove that $\bar{w} = \Omega$. Recall that $\bar{z} \geq \delta_1 \bar{w}$, so that $\bar{w} \geq \bar{z}/\delta_1$, and thus $\bar{w} \geq z/\delta_1$ for any $z \in \mathbb{Z}$. As a result, $S_m^h(z, \bar{w}) = (1-\theta)(\varepsilon-1)z/\varepsilon$, and the net payoff of the ℓ-investor carrying $\bar{w}$ is equal to

$$v^\ell(\bar{w}) = \gamma \left[\alpha(1-\theta)(\varepsilon-1)\bar{w} + (1-\alpha)(1-\theta)\frac{\varepsilon-1}{\varepsilon}\mathbb{E}(z) \right], \tag{A.23}$$

strictly increasing in $\bar{w}$. Therefore, we must have $\bar{w} = \Omega$. □

Proof of Proposition 1.3. Condition 1 is always satisfied in region (ii.i) of Proposition 1.2,

which Proposition 1.3 is concerned with. Condition 2 amounts to

$$\begin{aligned} &\sum_{j=1}^{N-1} \frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\delta_2^{j-1} < 1 \\ \Leftrightarrow &\frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\frac{\delta_2^{N-1}-1}{\delta_2-1} < 1 \end{aligned} \tag{A.24}$$

By definition, $\frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\frac{\delta_2^{\tilde{N}-1}-1}{\delta_2-1} < 1$, where $\tilde{N}$ satisfies (1.16). Therefore, condition 2 holds for the equilibrium proposed. Finally, we look into the third condition. For $N = 1$, it amounts to $\alpha(1-\theta)(\varepsilon-1) - (1-\alpha) \geq 0$, that is, $\alpha \geq 1/\delta_2$. For $N > 1$, condition 3 amounts to $\alpha(1-\theta)(\varepsilon-1) + (1-\alpha)\left[(1-\Pr(z=z_1))(1-\theta)(\varepsilon-1) - \Pr(z=z_1)\right] \geq 0$, that is $\alpha \leq 1/\delta_2$. From (1.16), we know that $\frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\frac{\delta_2^{\tilde{N}-1}-1}{\delta_2-1} < 1$ for $N > 1$ and $\frac{\alpha}{1-\alpha}(1-\theta)(\varepsilon-1)\frac{\delta_2^{\tilde{N}}-1}{\delta_2-1} > 1$ for all N. Algebra manipulations yield $\alpha \geq 1/\delta_2$ when $N = 1$ and $1/\delta_2^N \leq \alpha \leq 1/\delta_2^{N-1}$ for $N > 1$, ensuring that the third condition also holds. □

Derivations of equilibrium results with a competitive asset market. (i) When $\alpha \in (0, 1]$, there is a unique equilibrium with $q = \varepsilon$, $w = y = \Omega$, $z = 0$, and $\mathcal{W} = \Omega + \gamma(\varepsilon-1)\Omega$. (ii) When $\alpha = 0$, there is a unique monetary equilibrium with $q = \gamma\varepsilon/(i+\gamma)$, $w = y = \Omega$, $z = q\Omega$, and $\mathcal{W} = \Omega + \gamma(\varepsilon-1)\Omega$ if $i \leq \gamma(\varepsilon-1)$. Otherwise, there exists a continuum of non-monetary equilibria with $q < 1$, $z = y = 0$, $w \in [0, \Omega]$, and $\mathcal{W} = \Omega$. □

Derivations of equilibrium results with a competitive asset market and a satiation point. i) When $\alpha > \Omega/\bar{y}$, there exists a unique equilibrium, with $q = \varepsilon$, $w = y = \Omega$, $z = 0$ and $\mathcal{W} = \Omega + \gamma(\varepsilon-1)\Omega$. (ii) When $\alpha = \Omega/\bar{y}$, there exists a continuum of equilibria indexed by $q \in [\min\{1, \gamma(1-\alpha)\varepsilon/[i+\gamma(1-\alpha)]\}, \varepsilon]$, with $w = y = \Omega$, $z = 0$, and $\mathcal{W} = \Omega + \gamma(\varepsilon-1)\Omega$. (iii) If $\alpha < \Omega/\bar{y}$, there exists a unique monetary equilibrium with $q = \gamma(1-\alpha)\varepsilon/[i+\gamma(1-\alpha)]$, $w = y = \Omega$, $z = q(\Omega - \alpha\bar{y})/(1-\alpha)$, and $\mathcal{W} = \Omega + \gamma(\varepsilon-1)\Omega$ if $\alpha < 1 - i/[\gamma(\varepsilon-1)]$. Otherwise, there is a continuum of non-monetary equilibria with $q = 1$, $y = \bar{y}$, $w \in [\alpha\bar{y}, \Omega]$, $z = 0$, and $\mathcal{W} = \Omega + \gamma\alpha(\varepsilon-1)\bar{y}$. □

Proof of Lemma 1.5. In order to show that F^z_α first-order stochastically dominates $F^z_{\alpha'}$, we need to show that $\Pr_\alpha[z \geq \zeta] \geq \Pr_{\alpha'}[z \geq \zeta]$ for all ζ, with a strict inequality for some ζ. We have

$$\Pr[z > \zeta] = \sum_{i=1}^{j(\zeta)} \Pr[z = z_i], \tag{A.25}$$

where $j(\zeta) = \max_{j \in \mathbb{N}^+} j$ such that $z_j > \zeta$ and $z_j = \delta_1(\delta_1/\delta_2)^{j-1}\Omega$. Plugging in for (1.17), we obtain

$$\begin{aligned} \Pr[z > \zeta] &= \begin{cases} \sum_{i=1}^{j(\zeta)} \frac{\alpha}{1-\alpha}\theta(\varepsilon-1)\delta_2^{i-1} & \text{if } j(\zeta) < N \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{\alpha}{1-\alpha}\theta(\varepsilon-1)\frac{\delta_2^{j(\zeta)}-1}{\delta_2-1} & \text{if } j(\zeta) < N \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \tag{A.26}$$

We can directly see that $\Pr[z > \zeta]$ increases in α, strictly so when $z_N < \zeta < z_1$, concluding the proof. We proceed similarly to show that F^w_α first-order stochastically dominates $F^w_{\alpha'}$. We have

$$\Pr[w > \omega] = 1 - \sum_{i=h(\omega)}^{N(\alpha)} \Pr[w = w_i] \tag{A.27}$$

where $h(\omega) = \in_{h \in \mathbb{N}^+} h$ such that $w_h \leq \omega$ and $w_h = (\delta_1/\delta_2)^{h-1}\Omega$. Plugging in for (1.18), we obtain

$$\Pr[w > \omega] = \begin{cases} 1 - \sum_{i=h(\omega)}^{N(\alpha)} \frac{(1-\alpha)\theta(\varepsilon-1)-i}{(1-\alpha)[\theta(\varepsilon-1)+1]} \left(\frac{\varepsilon}{\delta_1}\right)^{i-N(\alpha)} & \text{if } h(\omega) > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{A.28}$$

Focusing on the case when $h(\omega) > 1$, we obtain

$$\Pr[w > \omega] = 1 - \frac{(1-\alpha)\theta(\varepsilon-1) - i}{(1-\alpha)\left[\theta(\varepsilon-1)+1\right]} \left(\frac{\varepsilon}{\delta_1}\right)^{h(\omega)} \left[\frac{(\varepsilon/\delta_1) - (\varepsilon/\delta_1)^{-N(\alpha)}}{(\varepsilon/\delta_1) - 1}\right]. \tag{A.29}$$

Remark that $\frac{(1-\alpha)\theta(\varepsilon-1)-i}{(1-\alpha)[\theta(\varepsilon-1)+1]}$ is positive and strictly decreasing in α. Making use of Result 1, according to which $N(\alpha)$ decreases in α, we also get that $\frac{(\varepsilon/\delta_1)-(\varepsilon/\delta_1)^{-N(\alpha)}}{(\varepsilon/\delta_1)-1}$, which is positive, decreases in α. As a result, $Pr[w > \omega]$ is increasing, strictly so α when $w_n < \omega < w_1$, concluding the proof. □

Proof of Result 1.2. Let $0 < \alpha' < \alpha$. Because F^w_α first-order stochastically dominates $F^w_{\alpha'}$, it directly follows that $\mathbb{E}_\alpha(w) \geq \mathbb{E}_{\alpha'}(w)$.

The average trade volume in money-only meetings is

$$\begin{aligned}
\mathcal{Y}_m &= \sum_{i=1}^{N}\sum_{j=1}^{N} y_m(z_i, w_j)\Pr[z = z_i]\Pr[w = w_j] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \Pr[z = z_i]\Pr[w = w_j]\min(w_j, z_i/\delta_1) \\
&= \mathbb{E}^w\left[\mathbb{E}^z\left[\min(w, z/\delta_1)\right]\right],
\end{aligned} \tag{A.30}$$

where $\mathbb{E}^z\left[\min(w, z/\delta_1)\right] = \sum_{i=1}^{N} \Pr[z = z_i]\min(w, z_i/\delta_1) \equiv f(w)$ and $\mathbb{E}^w\left[\mathbb{E}^z\left[\min(w, z/\delta_1)\right]\right] = \sum_{j=1}^{N} \Pr[w = w_j] f(w_j)$. Note that the function $\min(w, z/\delta_1)$ is increasing and concave in z, so that $\mathbb{E}^z_\alpha\left[\min(w, z/\delta_1)\right] \geq \mathbb{E}^z_{\alpha'}\left[\min(w, z/\delta_1)\right]$, or $f_\alpha(w) \geq f_{\alpha'}(w)$, due to the first-order stochastic dominance of F^z_α over $F^z_{\alpha'}$. Similarly, $f(w)$ is increasing and concave in w, so that $\mathbb{E}^w_\alpha\left[f_\alpha(w)\right] \geq \mathbb{E}^w_{\alpha'}\left[f_\alpha(w)\right]$, and therefore $\mathbb{E}^w_\alpha\left[f_\alpha(w)\right] \geq \mathbb{E}^w_{\alpha'}\left[f_{\alpha'}(w)\right]$ It follows that $\mathcal{Y}_m$ increases in α.

The aggregate trade volume in the OTC market is $\mathcal{Y} = \gamma\left[\alpha\mathbb{E}(w) + (1-\alpha)\mathcal{Y}_m\right]$, from which

$$\frac{\partial \mathcal{Y}}{\partial \alpha} = \gamma\left[\mathbb{E}(w) - \mathcal{Y}_m + \alpha\frac{\partial \mathbb{E}(w)}{\alpha} + (1-\alpha)\frac{\partial \mathcal{Y}_m}{\partial \alpha}\right] > 0. \tag{A.31}$$

Indeed, $\mathbb{E}(w) > \mathcal{Y}_m$ since, by construction, asset buyers are sometimes constrained by their money holdings and cannot purchase all of the seller's inventory, and we know that the two partial derivatives are positive from results derived above.

Aggregate welfare is given by

$$\mathcal{W} = (1-\gamma)\Omega + \gamma\left\{\alpha\left[\varepsilon\mathcal{Y}_c + (\Omega - \mathcal{Y}_c)\right] + (1-\alpha)\left[\varepsilon\mathcal{Y}_m + (\Omega - \mathcal{Y}_m)\right]\right\}, \tag{A.32}$$

where $\mathcal{Y}_c$ is the average trade size in credit meetings, and thus equal to $\mathbb{E}(w)$. This expression simplifies to $\mathcal{W} = \Omega + \gamma(\varepsilon - 1)\mathcal{Y}$, from which we directly obtain that welfare increases in α. □

Proof of Result 1.3. Because F^z_α first-order stochastically dominates $F^z_{\alpha'}$ for $0 < \alpha' < \alpha$, it directly follows that $\mathbb{E}_\alpha(z) \geq \mathbb{E}_{\alpha'}(z)$. □

Proof of Result 1.4. The average price in money meetings is given by

$$\mathcal{P}_m = \sum_{i=1}^{N}\sum_{j=1}^{N} \Pr(w = w_j)\Pr(z = z_i)p_m(z_i, w_j), \tag{A.33}$$

where $p_m(z_i, w_j) = \delta_1$ if $i \geq j$ and δ_2 if $i < j$ and $g(w) \equiv \mathbb{E}^z_\alpha[p_m(z,w)] = \sum_{i=1}^{N}\Pr(w)\Pr(z = z_i)p_m(z_i, w)$. The function $p_m(z,w)$ is increasing and concave in z, so that $\mathbb{E}^z_{\alpha'}[p_m(z,w)] \leq \mathbb{E}^z_\alpha[p_m(z,w)]$, or $g_\alpha(w) > g_{\alpha'}(w)$, due to the first-order stochastic dominance of F^z_α over $F^z_{\alpha'}$. Now, the function $-g(w)$ is increasing and concave in w, so that $\mathbb{E}_\alpha[g_\alpha(w)] \leq \mathbb{E}_{\alpha'}[g_\alpha(w)]$, due to the first-order stochastic dominance of F^w_α over $F^w_{\alpha'}$. Combining those results, it is ambiguous whether $\mathbb{E}_\alpha[g_\alpha(w)]$ is greater or smaller than $\mathbb{E}_{\alpha'}[g_{\alpha'}(w)]$. Were real balances not complement to credit but either independent or substitute, we would have $g_\alpha(w) \leq g_{\alpha'}(w)$, and $\mathbb{E}_\alpha[g_\alpha(w)] \leq \mathbb{E}_{\alpha'}[g_{\alpha'}(w)]$, so that the average price in money meetings would unambiguously increase following a tightening of credit.

The average price in the OTC market is $\mathcal{P} = \alpha\delta_2 + (1-\alpha)\mathcal{P}_m$. The impact of a marginal increase in credit is

$$\frac{\partial \mathcal{P}}{\partial \alpha} = (\delta_2 - \mathcal{P}_m) + (1-\alpha)\frac{\partial \mathcal{P}_m}{\partial \alpha}. \tag{A.34}$$

The first term is positive. The second term may be positive or negative. If positive, then the aggregate price increases in access to credit. If negative, the aggregate price may increase or decrease when credit is more available. □

A.2 Long-lived assets and asset pricing

We investigate the asset pricing implications of payment and inventory frictions in the OTC market. While we come back to the model presented in 1.3.1, a few modifications are introduced. First, the one-period-lived asset is replaced by an infinitely-lived asset with supply Ω that can be traded both in the first-stage and in the second-stage markets. Second, the asset is not endowed to investors anymore. Instead, at the end of each period, investors must choose how much of the asset to accumulate for the upcoming period, w_{t+1}. The second-stage price of the asset in terms of the numéraire is ϕ_t. This is the price we will focus on. Investors still receive utility $\varepsilon_\chi y$ from holding y units of the asset at the end of the first stage.[1] To ensure the existence of gains from trade in the first stage (i.e., ensure that the asset is at least partially misallocated before the first stage), we now assume that investors are ex-ante homogeneous, and receive an idiosyncratic shock that determines their preferences, $\chi \in \{\varepsilon, 1\}$ at the beginning of each period. For simplicity, $\Pr(\chi = \varepsilon) = \Pr(\chi = 1) = 0.5$. We still let agents make a participation decision, i.e., they can choose the amount of holdings they wish to trade in the first stage, $z^p_{t+1} \leq z_{t+1}$ and $w^p_{t+1} \leq w_{t+1}$. To abstract from search

[1] This can be interpreted as the asset being traded cum-dividend in the first stage and ex-dividend in the second stage.

frictions, we assume $\gamma = 1$.

The maximum expected discounted utility of an investor that enters the second stage with a portfolio (w_t, z_t) is

$$\begin{aligned} W_t(w_t, z_t) = &\max_{z_{t+1}\geq 0, w_{t+1}} c_t + \frac{\beta}{2}\left[V_{t+1}^h + V_{t+1}^\ell\right] \\ &\text{s.t. } 0 \leq w_{t+1} \text{ and } c_t + \phi_t w_{t+1} + (1+\pi) z_{t+1} \leq \phi_t w_t + z_t + T_t, \end{aligned} \tag{A.35}$$

The maximum expected discounted utility of an investor that enters the first stage with portfolio (w_t, z_t) and receives the high valuation shock is

$$\begin{aligned} V_t^h &= \max_{z_t^p \leq z_t} \left\{\varepsilon[w_t + y_t(z_t^p, \hat{w}_t^p)] + W_t[w_t + y_t(z_t^p, \hat{w}_t), z_t - p_t(z_t^p, \hat{w}_t)]\right\} \\ &= \max_{z_t^p \leq z_t} \left\{[(\varepsilon + \phi) y_t(z_t^p, \hat{w}_t) - p_t(z_t^p, \hat{w}_t)] + \varepsilon w_t + W_t(w_t, z_t)\right\}, \end{aligned} \tag{A.36}$$

Similarly, the maximum expected discounted utility of an investor that enters the first stage with portfolio (w_t, z_t) and receives the low valuation shock is

$$\begin{aligned} V_t^\ell &= \max_{w_t^p \leq w_t} \left\{[w_t - y_t(\hat{z}_t^p, w_t^p)] + W_t[w_t - y_t(\hat{z}_t^p, w_t^p), z_t + p_t(\hat{z}_t^p, w_t^p)]\right\} \\ &= \max_{w_t^p \leq w_t} \left\{[p_t(\hat{z}_t^p, w_t^p) - (1+\phi) y_t(\hat{z}_t^p, w_t^p)] + w_t + W_t(w_t, z_t)\right\}. \end{aligned} \tag{A.37}$$

In the right-hand sides of the second lines of (A.36) and (A.37), the terms in square brackets correspond to the surplus from trade in the first stage, while the second part corresponds to the investors' baseline utility in autarky.

Plugging (A.36) and (A.37) into (A.35), restricting our attention to equilibria where the rate-of-return on the asset is constant, $\phi_{t+1} = \phi_t = \phi$, and using the fact derived earlier that $z_t = z_t^p$ in equilibrium, we can write the portfolio and participation decision problem faced

by investors in the second stage as two independent problems,

$$\begin{aligned} z_t &= \arg\max \left\{ -iz_t + \frac{1}{2}\left[(\varepsilon+\phi)y_t - p_t\right] \right\} \\ (w_t, w_t^p) &= \arg\max \left\{ (\bar{\varepsilon} - r\phi)w_t + \frac{1}{2}\left[p_t - (1+\phi)y_t\right] \right\} \text{ s.t. } w_t^p \leq w_t, \end{aligned} \tag{A.38}$$

where $\bar{\varepsilon} \equiv (\varepsilon+1)/2$ is the average valuation of the asset.

To understand better the asset pricing implications of trading frictions in the OTC market, it is helpful to start by deriving a benchmark price—the equilibrium second-stage price were the asset traded in a competitive markets in both stages, with no credit frictions nor search frictions.

Lemma A.1 (Benchmark pricing)**.** *When the asset market is competitive and* $\alpha = 1$*,* $\phi = \varepsilon/r \equiv \phi^*$*.*

In the absence of any frictions, the asset can be perfectly reallocated towards the investors who value it the most. As a result, it is priced at their marginal valuation, which we call the fundamental value of the asset, ϕ^*.

We can now investigate how ϕ evolves as we introduce frictions, either by requiring the asset market to be an OTC market with bilateral trade and bargaining, by removing access to credit, or both.

Proposition A.1 (Illiquidity premia)**.** *In equilibrium, when*
(i) the asset market is competitive and $\alpha = 0$ *(payment friction),*

$$\phi = \begin{cases} \frac{1+i}{1+i(2+1/r)}\phi^* = \phi^* - \frac{(1+r)i}{i+r+2ri}\frac{\varepsilon}{r} & \text{for } i \leq \frac{(\varepsilon-1)r}{\varepsilon+1+2r} \\ \frac{\bar{\varepsilon}}{r} = \phi^* - \frac{(\varepsilon-1)}{2r} & \text{otherwise,} \end{cases} \tag{A.39}$$

(ii) the asset market is an OTC market and $\alpha = 1$ (market structure friction),

$$\phi = \phi^* - \frac{\theta(\varepsilon - 1)}{2r}, \tag{A.40}$$

(iii) the asset market is an OTC market and $\alpha = 0$ (both frictions),

$$\phi = \frac{\bar{\varepsilon}}{r} = \phi^* - \frac{(\varepsilon - 1)}{2r}. \tag{A.41}$$

First, note that when the first-stage terms of trade are determined via pairwise bargaining and credit is not accessible (case (iii)), the asset is priced at its average value, $\bar{\varepsilon}/r$. This is consistent with the fact that under this scenario, the asset is completely illiquid in the first stage. Indeed, strategic interactions identical to those presented in 1.3.2 shut the OTC market down. The illiquidity premium is equal to $-(\varepsilon - 1)/2r$. The higher the valuation difference between h- and ℓ- investors, the larger the premium.[2] One may wonder how to reconcile the former results with Lagos and Zhang (2019a), who have a similar setup—trade is bilateral and credit is not accessible, yet show that asset price is higher than the illiquid price $\bar{\varepsilon}/r$, which they interpret as a speculative premium. Indeed, in their setup, the bilateral terms of trade are determined by take-it-or-leave-it-offers. These correspond to special cases of our bargaining protocol, with $\theta = 0$ or $\theta = 1$. Giving all of the bargaining power to the h-investor leaves the ℓ-investor indifferent between bringing any amount of assets, thereby eliminating the strategic interactions that shut the OTC market down. For this reason, the asset is liquid OTC, pushing its second-stage price up.

In between the illiquid case and the perfectly-liquid case, there are two cases in which the asset is partially liquid. On one hand, the introduction of bilateral bargaining frictions (that is, frictions due to the market structure) commands an illiquidity premium equal to $-\theta(\varepsilon -$

[2]The same illiquidity premium exists when the market is competitive and credit is not accessible and the cost of holding real balances, i, is too high to sustain a monetary equilibrium.

$1)/2r$. This premium disappears as the h-investor's bargaining power goes to 0, but tends to the full illiquidity premium as her bargaining power goes to 1. On the other hand, credit frictions generate an illiquidity premium that becomes larger as i goes up and disappears as i goes to 0—by making money holdings costless, the Friedman rule eliminates the payment constraint faced by h-investors and generates outcomes analogous to an environment where credit is available. The premium term increases in i, thereby decreasing the second-stage resale price. This is consistent with the "turnover liquidity" mechanism highlighted by Lagos and Zhang (2020). When the nominal interest rate i increases, money holdings are more costly, reducing the amount of money held by potential buyers. This creates a downwards price impact in the first-stage market, diminishing the surplus earned by ℓ-investors. Because investors expect a lower gain from turning the asset over, the second-stage price goes down. When credit is available, h-investors are not bound by real balances, so that the turnover mechanism disappears.

We can easily show that the ranking between the two types of illiquidity premia depends on the relative magnitude of i and θ. The payment premium is larger in absolute value than the market structure premium as long as

$$i > \frac{(\varepsilon - 1)r\theta}{2\varepsilon(1+r) - (\varepsilon - 1)(1+2r)\theta}. \tag{A.42}$$

The right-hand-side is stricly increasing in θ, it is equal to 0 when $\theta = 0$ and it reaches $(\varepsilon-1)r/(\varepsilon+1+2r)$ when $\theta = 1$. Thus, there exists a threshold for θ under which the impact of payments frictions on the asset price is more severe than that of the market structure, and over which the opposite is true.

Finally, while allowing investors to make a participation decision at the intensive margin has no impact on pricing when trade is competitive or when credit is available, it is required in order for the asset price to be determinate when the environment features credit frictions

and pairwise trading.

Appendix B

Supplementary material for Chapter 2

B.1 Proofs of Lemmas and Propositions

The proofs of Proposition 2.1 and Proposition 2.3 can respectively be found in Appendix B.2 and Appendix B.3.

Proof of Lemma 2.1. The Pareto frontier is derived from the program

$$u^b = \max_{y,p\geq 0} \left\{ u(y) - p + u_0^b \right\} \quad \text{s.t. } p - v(y) + u_0^s \geq u^s,\ p \leq \delta\tau.$$

The consumer chooses the terms of trade, (y, p), to maximize his utility subject the constraint that he must guarantee some utility level u^s to the producer. If $\delta\tau \geq u^s - u_0^s + v(y^*)$, then $y = y^*$ and $p = u^s - u_0^s + v(y^*)$. Moreover, $u^b + u^s = u(y^*) - v(y^*) + u_0^b + u_0^s$. If $\delta\tau < u^s - u_0^s + v(y^*)$, then $p = \delta\tau = u^s - u_0^s + v(y)$, i.e., $y = v^{-1}(\delta\tau - u^s + u_0^s)$. □

Proof of Proposition 2.2. Since $p(\tau) = \delta\tau$, equation (2.1) implies that $u^b(\tau) = u_0^b + u\left[y(\tau)\right] -$

$\delta\tau$, and hence

$$u^{b\prime}(\tau) = u'(y)\, y'(\tau) - \delta. \tag{B.1}$$

The change in the consumer's utility along the gradual bargaining path the change in DM consumption as the consumer adds assets to the negotiating table, net of the asset transfer (the second term on the right side). From (2.15) and (B.1), we obtain (2.17). The total transfer of assets is $p(y) = \int_0^y \delta \frac{\partial \tau}{\partial x} dx$ where from (2.17) $\partial\tau/\partial x$ coincides with $1/y'(\tau)$ evaluated at x. □

Proof of Proposition 2.4. We assume that, with no loss of generality, $z \leq p_\infty(y^*)$. This also allows us to assume that (2.27) has interior solutions, and, summing (2.27) from $n = 1$ to N:

$$\sum_{n=1}^{N} \left[\int_{y_{n-1}}^{y_n} \frac{\upsilon'(y_n)}{u'(y_n) + \upsilon'(y_n)} u'(x) dx + \int_{y_{n-1}}^{y_n} \frac{u'(y_n)}{u'(y_n) + \upsilon'(y_n)} \upsilon'(x) dx \right] = z.$$

It can be expressed more compactly as

$$\int_0^{y_N} \left[1 - \Theta\left(x; \frac{z}{N}\right) \right] u'(x) + \Theta\left(x; \frac{z}{N}\right) \upsilon'(x) dx = z,$$

where

$$\Theta\left(x; \frac{z}{N}\right) = \sum_{n=1}^{N} \frac{u'(y_n)}{u'(y_n) + \upsilon'(y_n)} 1_{(y_{n-1}, y_n]}(x)$$

and $1_{(y_{n-1}, y_n]}(x)$ is the indicator function for the interval $(y_{n-1}, y_n]$. Note that for all $N < +\infty$ and for all $x \notin \{y_n\}$,

$$\Theta\left(x; \frac{z}{N}\right) < \frac{u'(x)}{u'(x) + \upsilon'(x)}.$$

Hence,

$$\int_0^{y_N} \left[1 - \Theta\left(x; \frac{z}{N}\right)\right] u'(x) + \Theta\left(x; \frac{z}{N}\right) \upsilon'(x) dx > \int_0^{y_N} \frac{2\upsilon'(x)u'(x)}{u'(x) + \upsilon'(x)} dx.$$

So for all $N < +\infty$, the payment to finance y_N units of consumption, the left side of the inequality, is larger than the one when $N = +\infty$, the right side of the inequality. Hence, the consumer extracts the largest surplus when $N = +\infty$. □

Proof of Proposition 2.5. For each $y \in (0, y^*]$, equation (2.38) gives a negative relationship between s and y, denoted by $s = s(y)$, with $\lim_{y \to 0} s(y) = +\infty$, and $s(y)$ is strictly decreasing. Given this function, equilibrium is given by y that satisfies (2.39), i.e.,

$$\left[\rho - \alpha\theta \left(\frac{u'(y) - \upsilon'(y)}{\upsilon'(y)}\right)\right] p(y) \leq (1 + \rho) Ad, \quad \text{" = " if } \; y < y^* \tag{B.2}$$

Since the left side of (B.2) is strictly increasing in y from 0 when $y = \underline{y} < y^*$ solution to $\rho = \alpha\theta \left[u'(\underline{y}) - \upsilon'(\underline{y})\right] / \upsilon'(\underline{y})$ to $\rho p(y^*)$ when $y = y^*$ and the right side is constant and strictly positive, there is a unique y that satisfies (B.2).

Part 1. If $p(y^*) \leq (1 + \rho) Ad/\rho$, then the solution is $y = y^*$ and $s(y^*) = 0$. If $p(y^*) > (1 + \rho) Ad/\rho$, the solution is such that $y \in \left(\underline{y}, y^*\right)$ and $s > 0$.

Part 2. Suppose the terms of trade are determined according to the generalized Nash solution:

$$(y, p) \in \arg\max \left[u(y) - p\right]^{\theta} \left[p - \upsilon(y)\right]^{1-\theta} \quad \text{s.t.} \quad p \leq z.$$

The first-order condition with respect to y gives

$$p = p(y) \equiv \frac{(1 - \theta)\upsilon'(y)u(y) + \theta u'(y)\upsilon(y)}{\theta u'(y) + (1 - \theta)\upsilon'(y)}.$$

For all $\theta \in (0,1)$, it is easy to check that $u(y) - p(y)$ reaches a maximum for some $\tilde{y}_\theta < y^*$. The buyer's problem corresponds to a choice of y solution to

$$\max_{y \geq 0} \{-sp(y) + \alpha [u(y) - p(y)]\}.$$

The solution is such that $y \leq \tilde{y}_\theta < y^*$ for all $s \geq 0$ and all $\theta < 1$.

Part 3. If the asset is fiat money, then $R = (1+\pi)^{-1}$ where π is the money growth rate and the spread is

$$s = (1+\rho)(1+\pi) - 1,$$

which can be interpreted as a nominal interest rate on an illiquid asset. From (2.38), y is the unique solution to

$$s = \alpha\theta \left(\frac{u'(y)}{v'(y)} - 1 \right).$$

If $s > 0$, $u'(y)/v'(y) > 1$ implies $y < y^*$. It is easy to check that y decreases with s because the right side is decreasing in y and as s tends to 0, y approaches y^*. □

Proof of Proposition 2.6. Under Nash bargaining the seller's surplus from a trade is:

$$u^s(\omega, z) = \begin{cases} \frac{\varepsilon_h - \varepsilon_\ell}{2}\omega & \geq \frac{\varepsilon_h + \varepsilon_\ell}{2} \\ z - \varepsilon_\ell \omega & \text{if } \frac{z}{\omega} \in \left[\frac{2\varepsilon_h \varepsilon_\ell}{\varepsilon_h + \varepsilon_\ell}, \frac{\varepsilon_h + \varepsilon_\ell}{2} \right) \\ \frac{(\varepsilon_h - \varepsilon_\ell) z}{2\varepsilon_h} & < \frac{2\varepsilon_h \varepsilon_\ell}{\varepsilon_h + \varepsilon_\ell}. \end{cases} \tag{B.3}$$

If z/ω is sufficiently high, then all DM goods are purchased by the buyer who only spends a fraction of his real balances. In that case, the seller's surplus increases with ω. If z/ω is in some intermediate range, then the buyer can still purchase all the DM goods of the seller but he has to spend all his real balances. In this case, the seller's surplus decreases with ω.

Finally, if z/ω is low, then the buyer can only purchase a fraction of the seller's DM goods, and the seller's surplus is constant. As a result, the seller's surplus reaches a maximum when $p \leq z$ starts to bind, i.e., $\omega = 2z/(\varepsilon_h + \varepsilon_\ell)$. The surplus of a buyer in a bilateral match is

$$u^b(z,\omega) = \begin{cases} \frac{\varepsilon_h - \varepsilon_\ell}{2}\omega & \geq \frac{\varepsilon_h+\varepsilon_\ell}{2} \\ \varepsilon_h\omega - z \quad \text{if} \quad \frac{z}{\omega} & \in \left[\frac{2\varepsilon_h\varepsilon_\ell}{\varepsilon_h+\varepsilon_\ell}, \frac{\varepsilon_h+\varepsilon_\ell}{2}\right) \\ \frac{(\varepsilon_h-\varepsilon_\ell)}{2\varepsilon_\ell}z & < \frac{2\varepsilon_h\varepsilon_\ell}{\varepsilon_h+\varepsilon_\ell}. \end{cases} \tag{B.4}$$

Let $\bar{z}$ denote the highest value on the support of $F^b(z)$. Then,

$$\omega \leq \min\left\{\frac{2\bar{z}}{\varepsilon_h + \varepsilon_\ell}, \Omega\right\}. \tag{B.5}$$

Let $\bar{\omega}$ denote the highest value in the support of $F^s(\omega)$. The solution is such that

$$z \leq \frac{2\varepsilon_h\varepsilon_\ell\bar{\omega}}{\varepsilon_h + \varepsilon_\ell}. \tag{B.6}$$

It can be checked that $(\varepsilon_h + \varepsilon_\ell)/2 > 2\varepsilon_h\varepsilon_\ell/(\varepsilon_h + \varepsilon_\ell)$, i.e., the intersection of the two best-response functions, (B.5) and (B.6), is such that the only Nash equilibrium is $\bar{z} = \bar{\omega} = 0$.

Under gradual bargaining, the Pareto frontier of the bargaining set, $u^b = \max(\varepsilon_h y - p)$ s.t. $p - \varepsilon_\ell y \geq u^s$, $p \leq z$, and $y \leq \omega$, is given by:

$$\begin{aligned} H\left(u^b, u^s, z, \omega\right) &= (\varepsilon_h - \varepsilon_\ell)\omega - u^b - u^s \quad \text{if} \quad u^s \leq z - \varepsilon_\ell\omega \\ &= \frac{(\varepsilon_h - \varepsilon_\ell)z}{\varepsilon_\ell} - \frac{\varepsilon_h}{\varepsilon_\ell}u^s - u^b \quad \text{otherwise.} \end{aligned}$$

Hence, the gradual bargaining solution requires

$$u^{b\prime}(z) = -\frac{1}{2}\frac{\partial H/\partial z}{\partial H/\partial u^b} = \frac{1}{2}\frac{(\varepsilon_h - \varepsilon_\ell)}{\varepsilon_\ell}.$$

From the definition $u^{b\prime}(z) = \varepsilon_h \partial y/\partial z - 1$, it follows that $\partial z/\partial y = 2\varepsilon_h\varepsilon_\ell/(\varepsilon_h + \varepsilon_\ell)$. Integrating this expression, the payment function is $p(y) = \frac{2\varepsilon_h\varepsilon_\ell}{\varepsilon_h+\varepsilon_\ell} y$. The buyer's choice of y is given by:

$$\max_{y\in[0,\omega]} \left\{ -s\frac{2\varepsilon_h\varepsilon_\ell}{\varepsilon_h+\varepsilon_\ell} y + \alpha \left[\varepsilon_h y - \frac{2\varepsilon_h\varepsilon_\ell}{\varepsilon_h+\varepsilon_\ell} y \right] \right\}.$$

It can be re-expressed as:

$$\max_{y\in[0,\omega]} \left[-s2\varepsilon_\ell + \alpha\left(\varepsilon_h - \varepsilon_\ell\right) \right] y.$$

Provided that $s \leq \alpha\left(\varepsilon_h - \varepsilon_\ell\right)/(2\varepsilon_\ell)$, it is optimal to choose $y = \omega$ and to hold $z = p(\omega)$. The surplus of the seller is:

$$\begin{aligned} u^s(\omega, z) &= \min\left\{ p(\omega) - \varepsilon_\ell\omega, z - \frac{(\varepsilon_h+\varepsilon_\ell)}{2\varepsilon_h} z \right\} \\ &= \min\left\{ \varepsilon_\ell \left(\frac{\varepsilon_h - \varepsilon_\ell}{\varepsilon_h + \varepsilon_\ell} \right) \omega, \frac{\varepsilon_h - \varepsilon_\ell}{2\varepsilon_h} z \right\}. \end{aligned}$$

The seller's surplus is monotone (weakly) increasing in ω. Hence, $\omega = \Omega$ is a weakly dominant strategy. □

Proof of Lemma 2.2. By (2.46), an optimal $[\sigma_j]_{j=0}^J$ maximizes

$$\omega(\overline{\tau}) = \sum_{j=0}^{J} \int_0^{\overline{\tau}} \delta_j \sigma_j(x) dx = \sum_{j=0}^{J} \delta_j \Delta_j,$$

where $\Delta_j \equiv \int_0^{\overline{\tau}} \sigma_j(x)dx$, subject to feasibility. We can then rewrite this problem as

$$\max_{\Delta_j, j=0,...,J} \sum_{j=0}^{J} \delta_j \Delta_j, \text{ subject to } \sum_{j=0}^{J} \Delta_j = \overline{\tau} \text{ and } 0 \leq \Delta_j \leq a_j/\delta_j \text{ for all } j = 0, ..., J,$$

where the constraints follow from feasibility requirement on $[\sigma_j]_{j=0}^J$. Now, let $\widetilde{j} \geq 0$ satisfy

$$\sum_{j=0}^{\widetilde{j}-1} a_j/\delta_j < \overline{\tau} \leq \sum_{j=0}^{\widetilde{j}} a_j/\delta_j.$$

Since $\delta_0 \geq \delta_1 \geq ... \geq \delta_J$, it is optimal to choose $\Delta_j = a_j/\delta_j$ for all $j = 0, ..., \widetilde{j} - 1$, $\Delta_{\widetilde{j}} = \overline{\tau} - \sum_{j=0}^{\widetilde{j}-1} a_j/\delta_j$, and $\Delta_j = 0$ for all $j > \widetilde{j}$. Hence, $[\sigma_j^*]_{j=0}^J$ restricted to $[0, \overline{\tau}]$ is optimal. It is also uniquely optimal if $\delta_0 > \delta_1 > ... > \delta_J$. □

Proof of Lemma 2.3. Define $\Omega_j(\mathbf{a}) = \sum_{k=0}^{j-1} a_k$ for all $j = 1, ..., J+1$ with $\Omega_0(\mathbf{a}) = 0$. We can then rewrite (2.48) as

$$S(\mathbf{a}) = \theta \sum_{j=0}^{J} \int_{\Omega_j}^{\Omega_{j+1}} e^{-\lambda\left[\frac{(\omega-\Omega_j)}{\delta_j}+T_j\right]} \ell[p^{-1}(\omega)] d\omega,$$

where $p^{-1}(\omega) = y^*$ whenever $\omega > p(y^*)$, and we have changed the variable from τ to $\omega = \omega^*(\tau)$; note that for all $\omega \in (\Omega_j, \Omega_{j+1})$,

$$\begin{aligned} (\omega^*)^{-1}(\omega) &= \frac{(\omega - \Omega_j)}{\delta_j} + T_j, \\ \frac{d}{d\omega}(\omega^*)^{-1}(\omega) &= \frac{1}{\delta_j}. \end{aligned}$$

Now, let $k \geq 0$ be given. We shall compute the derivative of $S(\mathbf{a})$ w.r.t. a_k. We will compute it by grouping the terms inside the summation into three groups: terms with $j < k$, the term with $j = k$, and terms with $j > k$. Note that $S(\mathbf{a})$ depends on a_k through terms Ω_j and T_j with $j > k$ and hence, for $j < k$,

$$\frac{\partial}{\partial a_k} \int_{\Omega_j}^{\Omega_{j+1}} e^{-\lambda\left[\frac{(\omega-\Omega_j)}{\delta_j}+T_j\right]} \ell[p^{-1}(\omega)] d\omega = 0,$$

for $j = k$,

$$\frac{\partial}{\partial a_k}\int_{\Omega_k}^{\Omega_{k+1}} e^{-\lambda\left[\frac{(\omega-\Omega_k)}{\delta_k}+T_k\right]}\ell[p^{-1}(\omega)]d\omega = -e^{-\lambda T_k}\ell[p^{-1}(\Omega_k)],$$

and for $j > k$,

$$\begin{aligned}
&\frac{\partial}{\partial a_k}\int_{\Omega_j}^{\Omega_{j+1}} e^{-\lambda\left[\frac{(\omega-\Omega_j)}{\delta_j}+T_j\right]}\ell[p^{-1}(\omega)]d\omega \\
= \;& -e^{-\lambda T_j}\ell\left[p^{-1}(\Omega_j)\right] + e^{-\lambda\left[\frac{(\Omega_{j+1}-\Omega_j)}{\delta_j}+T_j\right]}\ell[p^{-1}(\Omega_{j+1})] + \int_{\Omega_j}^{\Omega_{j+1}}\lambda\left[\frac{1}{\delta_j}-\frac{1}{\delta_k}\right]e^{-\lambda\left[\frac{(\omega-\Omega_j)}{\delta_j}+T_j\right]}\ell[p^{-1}(\omega)]d\omega \\
= \;& -e^{-\lambda T_j}\ell\left[p^{-1}(\Omega_j)\right] + e^{-\lambda T_{j+1}}\ell[p^{-1}(\Omega_{j+1})] + \int_{\Omega_j}^{\Omega_{j+1}}\lambda\left[\frac{1}{\delta_j}-\frac{1}{\delta_k}\right]e^{-\lambda\left[\frac{(\omega-\Omega_j)}{\delta_j}+T_j\right]}\ell[p^{-1}(\omega)]d\omega.
\end{aligned}$$

Thus, adding the terms up across j, we obtain

$$\frac{\partial}{\partial a_k}S(\mathbf{a}) = \theta\sum_{j=k+1}^{J}\int_{\Omega_j}^{\Omega_{j+1}}\lambda\left[\frac{1}{\delta_j}-\frac{1}{\delta_k}\right]e^{-\lambda\left[\frac{(\omega-\Omega_j)}{\delta_j}+T_j\right]}\ell[y(\omega)]d\omega + \theta e^{-\lambda T_{J+1}}\ell[y(\Omega_{J+1})],$$

where the terms $e^{-\lambda T_j}\ell[p^{-1}(\Omega_j)]$ cancel one another except for the very last one. Equation (2.51) is obtained by another change of variable back to τ. □

Proof of Proposition 2.7. (1) The equilibrium is solved recursively. The FOC (2.50) when $j = 0$ determines a_0. Note that, in the expression (2.51) for $\frac{\partial}{\partial a_0}S(\mathbf{a})$, it depends on a_0 only through $T_1, ..., T_{J+1}$, and it is strictly decreasing in a_0. Indeed, this follows directly from the fact that T_j is strictly increasing in a_0 and the difference $T_{j+1} - T_j$ is not, and that $e^{-\lambda\tau}\ell[y(\tau)]$ strictly decreases with τ. Now, the right side of (2.50) is also strictly positive at $a_0 = 0$ provided that $\delta_0 > \delta_1$ and equal to 0 as a_0 goes to ∞. The threshold for the nominal interest rate below which a monetary equilibrium exists is

$$\bar{\iota} = \alpha\theta\lambda\sum_{k=1}^{J}\frac{(\delta_0-\delta_k)}{\delta_0}\int_{T_k}^{T_{k+1}} e^{-\lambda\tau}\ell\left[y(\tau)\right]d\tau + \alpha\theta e^{-\lambda T_{J+1}}\ell\left[y(T_{J+1})\right],$$

where $T_1 = 0$, and $T_j = \sum_{k=1}^{j-1} A_k/\delta_k$ for all $j \in \{2, ..., J+1\}$.

Given a_0, the spreads $\{s_j\}_{j=1}^J$ are determined by (2.50), with $A_0 = a_0$ and $T_j = \sum_{k=0}^{j-1} A_k/\delta_k$ for all $j \in \{1, ..., J+1\}$. From (2.50) we can compute the difference between two consecutive spreads:

$$s_j - s_{j+1} = \alpha\theta\lambda \frac{(\delta_j - \delta_{j+1})}{\delta_j} \int_{T_{j+1}}^{T_{j+2}} e^{-\lambda\tau} \ell\left[y(\tau)\right] d\tau.$$

Hence, $s_j - s_{j+1} > 0$ requires $\delta_j - \delta_{j+1} > 0$ and $y(T_{j+1}) < y^*$, i.e., $\Omega_{j+1} = \sum_{k=0}^{j} A_k < p(y^*)$.

(2) We can simplify the expression of the velocity of asset j given by (2.52) as

$$\begin{aligned} \mathcal{V}_j &\equiv \frac{\alpha \int_0^{+\infty} \lambda e^{-\lambda x} \int_0^x \omega_j^{*\prime}(\tau) 1_{\{\omega^*(\tau) < p(y^*)\}} d\tau dx}{A_j} = \frac{\alpha \int_0^{+\infty} e^{-\lambda\tau} \omega_j^{*\prime}(\tau) 1_{\{\omega^*(\tau) < p(y^*)\}} d\tau}{A_j} \\ &= \frac{\alpha \int_{T_j}^{T_{j+1}} e^{-\lambda\tau} \delta_j 1_{\{\omega^*(\tau) < p(y^*)\}} d\tau}{A_j}, \end{aligned}$$

where the first equality changes the order of integration and the second uses the fact that $\omega_j^{*\prime}(\tau) = \delta_j 1_{\{T_j \leq \tau < T_{j+1}\}}$. Using the expressions for T_j and T_{j+1} we distinguish three cases:

$$\mathcal{V}_j = \begin{cases} A_j^{-1}\lambda^{-1}\alpha\delta_j e^{-\lambda T_j}\left(1 - e^{-\frac{\lambda}{\delta_j}A_j}\right) \\ A_j^{-1}\lambda^{-1}\alpha\delta_j e^{-\lambda T_j}\left[1 - e^{-\frac{\lambda}{\delta_j}[p(y^*) - \Omega_j]}\right] \\ 0 \end{cases} \quad \text{if } p(y^*) \begin{cases} \geq \Omega_{j+1} \\ \in (\Omega_j, \Omega_{j+1}) \\ \leq \Omega_j \end{cases}.$$

Thus, $\mathcal{V}_j > 0$ if and only if $p(y^*) > \Omega_j$. Moreover, for any j with $p(y^*) > \Omega_j$,

$$\begin{aligned} \mathcal{V}_j - \mathcal{V}_{j+1} &\geq A_j^{-1}\lambda^{-1}\alpha\delta_j e^{-\lambda T_j}\left(1 - e^{-\frac{\lambda}{\delta_j}A_j}\right) - A_{j+1}^{-1}\lambda^{-1}\alpha\delta_{j+1} e^{-\lambda T_{j+1}}\left(1 - e^{-\frac{\lambda}{\delta_{j+1}}A_{j+1}}\right) \\ &= \alpha e^{-\lambda T_{j+1}}\left[\delta_j A_j^{-1}\lambda^{-1}\left(e^{\frac{\lambda}{\delta_j}A_j} - 1\right) - \delta_{j+1}A_{j+1}^{-1}\lambda^{-1}\left(1 - e^{-\frac{\lambda}{\delta_{j+1}}A_{j+1}}\right)\right] > 0, \end{aligned}$$

where the inequality follows from the fact that

$$\frac{e^{\frac{\lambda}{\delta_j}A_j}-1}{\frac{\lambda}{\delta_j}A_j} > 1 > \frac{1-e^{-\frac{\lambda}{\delta_{j+1}}A_{j+1}}}{\frac{\lambda}{\delta_{j+1}}A_{j+1}}.$$

(3) It follows directly from (2.50) and the fact that:

$$\begin{aligned} |s_j - s_{j+1}| &= \alpha\theta\lambda\frac{(\delta_j-\delta_{j+1})}{\delta_j}\int_{T_{j+1}}^{T_{j+2}} e^{-\lambda\tau}\left[\frac{u'[y(\tau)]-v'[y(\tau)]}{v'[y(\tau)]}\right]d\tau \\ &\leq \alpha\theta\lambda\frac{(\delta_j-\delta_{j+1})}{\delta_j}e^{-\lambda T_{j+1}}\left[\frac{u'[y(T_{j+1})]-v'[y(T_{j+1})]}{v'[y(T_{j+1})]}\right]d\tau, \end{aligned} \tag{B.7}$$

which converges to zero as $\lambda \to \infty$. □

B.2 Proof of Proposition 2.1 and extension to asymmetric bargaining powers

As assumed in the main text, the number of bargaining rounds, N, is even, and the producer is the first to make an offer while the consumer is the last. We obtain essentially the same results for the other cases (either N is odd or the producer is making the last offer), as will be discussed in the proof. Here we also normalize $u_0^b = u_0^s = 0$. Also, with no loss of generality, we normalize δ to be one.

We define intermediate payoffs as the utilities that the players would enjoy based on the agreements reached up to some round $n \in \{1, ..., N\}$. Let (y_n, p_n) denote the cumulative offers that are agreed upon up to round n. Feasibility requires $0 \leq p_n - p_{n-1} \leq z/N$ and $0 \leq y_n - y_{n-1}$ for all $n = 1, ..., N$ and $p_0 = y_0 = 0$. From (2.1) and (2.2), we have $u_n^b = u(y_n) - p_n$ and $u_n^s = -v(y_n) + p_n$. The payoffs over terminal histories are simply u_N^b

and u_N^s. If we restrict $y \in [0, y^*]$, then there is a one-to-one correspondence between the intermediate allocation (y, p) and the intermediate payoff (u^b, u^s) such that $H(u^b, u^s, p) = 0$.

The rest of the section consists in proving Proposition 2.1 followed by the extension to asymmetric bargaining powers. The proof contains four parts: the first gives a full characterization of the equilibrium payoffs of any subgame; the second gives equilibrium intermediate payoffs; the third proves uniqueness; the fourth characterize the solution as N goes to infinity.

Final equilibrium payoffs

To solve the game, we need to solve all possible subgames. A subgame is characterized by the intermediate payoffs, denoted by (u_0^b, u_0^s) with the corresponding allocation denoted by (y_0, p_0), and the number of rounds remaining for bargaining, denoted by J. That is, the subgame begins at round $N - J + 1$, with the intermediate payoff (u_0^b, u_0^s) that results from the bargaining in the first $N - J$ rounds. (The entire game has $(u_0^b, u_0^s) = (0, 0)$ and $J = N$.) Feasibility requires $p_0 \leq (N - J)z/N$, and we only consider $y_0 < y^*$ so that there are still gains from trade to be exploited. Our first lemma describes the final payoffs of such a game. Let $S(y) = u(y) - v(y)$ and $S^* = S(y^*)$.

Lemma B.1. *Consider a game* $[(u_0^b, u_0^s), J]$ *with* $0 \leq u_0^b + u_0^s < S^*$, *and* $p_0 = u[S^{-1}(u_0^b + u_0^s)] - u_0^b = u_0^s + v[S^{-1}(u_0^b + u_0^s)]$. *Equilibrium final payoffs,* $(\widetilde{u}_J^b, \widetilde{u}_J^s)$, *correspond to the last term of the sequence,* $\{(\widetilde{u}_j^b, \widetilde{u}_j^s)\}_{j=0}^J$, *defined as* $(\widetilde{u}_0^b, \widetilde{u}_0^s) = (u_0^b, u_0^s)$, *and*

$$H(\widetilde{u}_j^b, \widetilde{u}_{j-1}^s, p_0 + jz/N) = 0 \text{ and } \widetilde{u}_j^s = \widetilde{u}_{j-1}^s, \text{ for } j \geq 1 \text{ odd}, \tag{B.8}$$

$$H(\widetilde{u}_{j-1}^b, \widetilde{u}_j^s, p_0 + jz/N) = 0 \text{ and } \widetilde{u}_j^b = \widetilde{u}_{j-1}^b, \text{ for } j \geq 2 \text{ even}. \tag{B.9}$$

The proof of Lemma B.1 uses backward induction. When $J = 1$, the game $[(u_0^b, u_0^s), 1]$ is

a standard take-it-or-leave-it offer game (with the consumer making the offer). In equilibrium, the consumer makes an offer that leaves the producer indifferent between rejecting or accepting, with the final payoff to the producer $\widetilde{u}_1^s = u_0^s$. Taking this as given, the consumer spends up to z/N units of assets so that his final payoff $\widetilde{u}_1^b$ satisfies $H(\widetilde{u}_1^b, u_0^s, p_0 + z/N) = 0$. (Note that the buyer will spend exactly z/N unless y^* is achieved with a slack liquidity constraint.) This proves (B.8) with $J = 1$.

Now consider $J = 2$, and the producer makes the first offer. If the consumer rejects the offer, the subgame becomes $[(u_0^b, u_0^s), 1]$, and the consumer can guarantee himself a final payoff of $\widetilde{u}_1^b$, which we call the consumer's *reservation payoff*. Take this as given, the producer's offer is acceptable as long as the offer leads to a consumer final payoff no less than $\widetilde{u}_1^b$. Thus, the producer's offer maximizes his final payoff, u_2^s, subject to $u_2^b \geq \widetilde{u}_1^b$. Equivalently, the producer final payoff $\widetilde{u}_2^s$ solves $H(\widetilde{u}_1^b, \widetilde{u}_2^s, p_0 + 2z/N) = 0$. This proves (B.9) with $J = 2$. We illustrate this logic in Figure B.1.

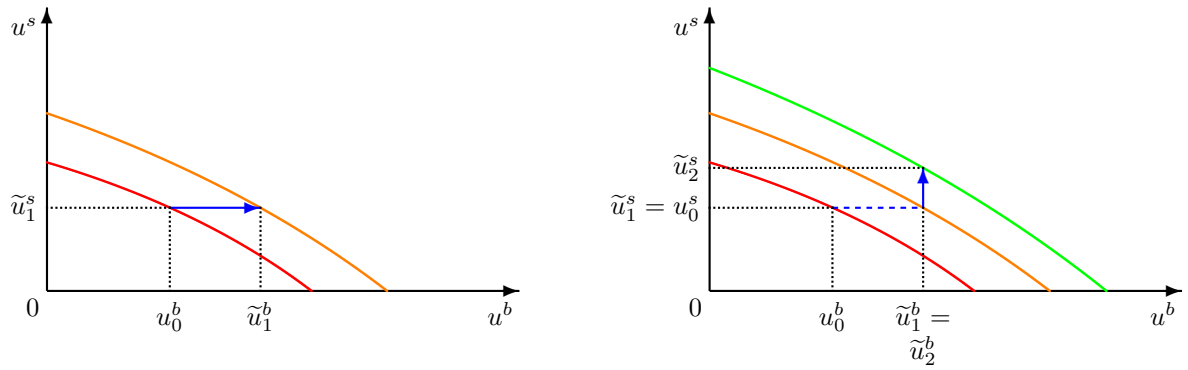

Figure B.1: Construction of $\widetilde{u}_1^b$ and $\widetilde{u}_2^s$

We continue this argument by induction. Suppose that the final payoffs are given by (B.8) and (B.9) for any game $[(u_0^b, u_0^s), J-1]$ with $J \geq 3$ and consider a game $[(u_0^b, u_0^s), J]$ with J odd and the consumer is making the first offer. If the producer rejects the offer, his reservation payoff would be $\widetilde{u}_{J-1}^s$. Following the same logic, the consumer's offer maximizes his final payoff u_b^J subject to the constraint that the producer's final payoff is no less than

his reservation payoff, $\widetilde{u}^s_{J-1}$. Thus, the final payoffs in the game $[(u^b_0, u^s_0), J]$, denoted by $(\widetilde{u}^b_J, \widetilde{u}^s_J)$, solve $H(\widetilde{u}^b_J, \widetilde{u}^s_{J-1}, p_0 + Jz/N) = 0$ and $\widetilde{u}^s_J = \widetilde{u}^s_{J-1}$. The case for J even is similar. This proves (B.8) and (B.9) for J.

Before we proceed, we give some comments on how to handle the case when the first best is reached at some point of the game. Once we reach y^*, that is, once $\widetilde{u}^b_j + \widetilde{u}^s_j = u(y^*) - v(y^*)$, the sequence $\{(\widetilde{u}^b_j, \widetilde{u}^s_j)\}_{j=0}^J$ is constant afterwards and in equilibrium there is no trade in rounds after j. Note that this is consistent with our definition of simple SPE. Thus, we may only consider the case where

$$\widetilde{u}^b_{J-1} + \widetilde{u}^s_{J-1} < S^*. \tag{B.10}$$

Equilibrium Intermediate Payoffs

We now construct the sequence of intermediate payoffs (and the corresponding allocations and offers) that will lead to final payoffs. We emphasize that the sequence $\{(\widetilde{u}^b_j, \widetilde{u}^s_j)\}_{j=0}^J$ used to construct the final payoffs is distinct from the sequence of intermediate payoffs, as we will illustrate shortly. To do so, we expand the notation slightly to explicate the recursive nature of the sequence $\{(\widetilde{u}^b_j, \widetilde{u}^s_j)\}_{j=0}^J$. As mentioned, at each step according to (B.8)-(B.9), the next payoff is computed by either a rightward or upward shift to the next Pareto frontier. Formally, we define two operators, $F_r(u^b, u^s)$ and $F_u(u^b, u^s)$ given by

$$F_r(u^b, u^s) = (u^{b\prime}, u^{s\prime}) \text{ such that } u^{s\prime} = u^s \text{ and } H(u^{b\prime}, u^s, p + z/N) = 0, \tag{B.11}$$

$$F_u(u^b, u^s) = (u^{b\prime}, u^{s\prime}) \text{ such that } u^{b\prime} = u^b \text{ and } H(u^b, u^{s\prime}, p + z/N) = 0, \tag{B.12}$$

where $p = u[S^{-1}(u^b + u^s)] - u^b$. The operator $F_r(u^b, u^s)$ moves from (u^b, u^s) to the next Pareto frontier by a rightward shift, and $F_u(u^b, u^s)$ moves upward. It then follows directly from (B.8) and (B.9) that, for all j even,

$$(\widetilde{u}^b_{j+1}, \widetilde{u}^s_{j+1}) = F_r(\widetilde{u}^b_j, \widetilde{u}^s_j), \tag{B.13}$$

$$(\widetilde{u}^b_{j+2}, \widetilde{u}^s_{j+2}) = F_u(\widetilde{u}^b_{j+1}, \widetilde{u}^s_{j+1}) = (F_u \circ F_r)(\widetilde{u}^b_j, \widetilde{u}^s_j). \tag{B.14}$$

Our construction of equilibrium intermediate payoffs follows backward induction from the final payoffs constructed in Lemma B.1. Consider a game $[(u^b_0, u^s_0), J]$ with J even. Lemma B.1 shows that the final payoffs to the agents are given by $(\widetilde{u}^b_J, \widetilde{u}^s_J)$. Let $(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$ denote the equilibrium intermediate payoff for the agents at the end of round-$(J-1)$ bargaining. Applying Lemma B.1 to the game $[(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1}), 1]$, the equilibrium payoff to that game is given by $F_r(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$. Thus, subgame perfection requires

$$F_r(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1}) = (\widetilde{u}^b_J, \widetilde{u}^s_J). \tag{B.15}$$

The solution to (B.15) is to move from $(\widetilde{u}^b_J, \widetilde{u}^s_J)$ leftward to the previous Pareto frontier: formally, it is given by

$$H[\widehat{u}^b_{J-1}, \widetilde{u}^s_J, p_0 + (J-1)z/N] = 0, \ \widehat{u}^s_{J-1} = \widetilde{u}^s_J. \tag{B.16}$$

In general, the same argument shows that the equilibrium intermediate payoff at the end of

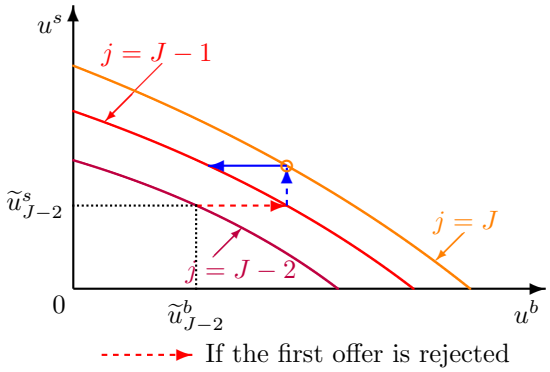

Figure B.2: Backward induction

round-$(J-j)$ bargaining, denoted by $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})$, must satisfy

$$\begin{aligned} (F_u \circ F_r)^{j/2}(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j}) &= (\widetilde{u}^b_J, \widetilde{u}^s_J) \text{ for } j \text{ even}, \\ F_r[(F_u \circ F_r)^{(j-1)/2}(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})] &= (\widetilde{u}^b_J, \widetilde{u}^s_J) \text{ for } j \text{ odd}. \end{aligned} \tag{B.17}$$

According to (B.17), for j even, if we start with $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})$, it should reach the final payoffs, $(\widetilde{u}^b_J, \widetilde{u}^s_J)$, by $j/2$ rightward and upward shifts to next Pareto frontiers, one rightward shift followed by an upward one. Now, by repeated use of (B.14), we have that $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j}) = (\widetilde{u}^b_{J-j}, \widetilde{u}^s_{J-j})$ for all j even. For j odd, if we start with $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})$, it should reach the final payoffs, $(\widetilde{u}^b_J, \widetilde{u}^s_J)$, by $(j-1)/2$ rightward and upward shifts to next Pareto frontiers, plus one more rightward shift. Hence, $(\widetilde{u}^b_{J-j}, \widetilde{u}^s_{J-j})$ can be obtained from $(\widetilde{u}^b_J, \widetilde{u}^s_J)$ by first a leftward shift to the previous Pareto frontier, followed by $(j-1)/2$ downward and leftward shifts to previous frontiers. Figure B.2 illustrates this process for $j = 2$. We have the following lemma.

Lemma B.2. *Consider a game $[(u^b_0, u^s_0), J]$ be given with J even that satisfies (B.10). There is a unique sequence, $\{(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})\}_{j=0}^{J-1}$, with corresponding sequence of allocation denoted by $\{\widehat{y}_{J-j}\}_{j=0}^{J-1}$, possibly except for $(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$, that satisfies (B.17), which also enjoys the*

following properties:

$$\widehat{u}^b_{J-j} > \widehat{u}^b_{J-j-1} \text{ for all } j = 0, ..., J-2;\ \widehat{u}^b_1 > u^b_0; \tag{B.18}$$

$$\widehat{u}^s_{J-j} > \widehat{u}^s_{J-j-1} \text{ for all } j = 1, ..., J-2;\ \widehat{u}^s_1 > u^s_0; \tag{B.19}$$

$$\widehat{y}_j > \widehat{y}_{j-1} \text{ for all } j = 2, ..., J;\ \widehat{y}_1 > y_0. \tag{B.20}$$

The proof of Lemma B.2 is based on induction on j and uses the fact that $u(y) - v(y)$ is strictly concave. The proof is rather straightforward but tedious and the detailed proof is available upon request. Moreover, since $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j}) = (\widetilde{u}^b_{J-j}, \widetilde{u}^s_{J-j})$ for all j even, (B.18) and (B.19) imply that the two sequences, $\{(\widetilde{u}^b_j, \widetilde{u}^s_j)\}_{j=1}^{J-1}$ and $\{(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})\}_{j=1}^{J-1}$ in fact nests one another, and hence, if one sequence converges to some limit, the other also converges to the same limit. We also remark that while we have assumed J to be even, an analogous lemma for J odd holds as well. In that case, $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j}) = (\widetilde{u}^b_{J-j}, \widetilde{u}^s_{J-j})$ for all j odd, but we need to compute $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})$ for j even with an alternative sequence analogous to the one we constructed for the case with J even and j odd.

Uniqueness of SPE

Here we prove our uniqueness claim. First we show that, for any subgame, $[(u^b_0, u^s_0), J]$, the equilibrium final payoffs in any SPE is given by (B.8)-(B.9), denoted by $(\widetilde{u}^b_J, \widetilde{u}^s_J)$. For $J = 1$ this is the standard ultimatum game and the uniqueness is standard. Suppose that we have uniqueness for $J-1$, $J \geq 2$. Then, fix a SPE and consider the game at the first bargaining round, and, without loss of generality, assume that producer is making an offer and J is even. We show that the consumer can guarantee a final payoff of $\widetilde{u}^b_J$ and the producer can guarantee $\widetilde{u}^s_J$ at the first round. First, by rejecting the producer offer, by the induction hypothesis, the

unique equilibrium payoff to the consumer is $\widetilde{u}^b_{J-1} = \widetilde{u}^b_J$, and hence any offer that leads to a final payoff lower than $\widetilde{u}^b_J$ will be rejected. For the consumer, Lemma B.2 shows that there exists a unique intermediate payoff, (u^b_1, u^s_1), such that $F_r \circ (F_u \circ F_r)^{(J-2)/2}(u^b_1, u^s_1) = (\widetilde{u}^b_J, \widetilde{u}^s_J)$, and such intermediate payoff is achievable with some offer (y_1, d_1). By offering $(y_1 + \varepsilon, d_1)$ for ε small the producer can guarantee consumer acceptance and hence, taking ε to zero, the producer can guarantee a final payoff of $\widetilde{u}^s_J$. Since the payoffs, $(\widetilde{u}^b_J, \widetilde{u}^s_J)$, lie on the Pareto frontier achievable by the two agents with total assets the consumer has, and each can guarantee the respective payoff, this final payoff is unique.

Now we show that the intermediate payoffs we constructed are unique in a simple SPE. Note first that in a simple SPE, the game effectively ends when active rounds end. Let J be the number of active rounds and the final payoffs are given by $(\widetilde{u}^b_J, \widetilde{u}^s_J)$, and, by backward induction, (B.17) must hold. Lemma B.2 implies that there is a unique solution to that except for $(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$. However, that payoff can be pinned down by the fact that buyer has to spend z/N in a simple SPE in round $J-1$. Finally, when the output corresponding to $(\widetilde{u}^b_J, \widetilde{u}^s_J)$ is less than y^*, then $J = N$, and the solution to (B.17) is unique for all j. Since y^* is not achievable in any subgame, it follows that the SPE is unique.

Convergence to Gradual Nash Solution

We consider convergence of games with N even. The limit will be the same for N odd and hence we have convergence. Here we show that the limit intermediate payoffs converge as N approaches infinity in simple SPE in the following sense. Now, for each N and each $n \in \{1, 2, \ldots, N\}$, define

$$[u^b_N(\tau), u^s_N(\tau)] = (u^b_n, u^s_n) \text{ if } \tau \in [(n-1)z/N, nz/N),$$

where (u_n^b, u_n^s) is an equilibrium intermediate payoff in the game with N rounds. We then show that $[u_N^b(\tau), u_N^s(\tau)]$ converges (pointwise) to $[u^b(\tau), u^s(\tau)]$, the solution to (2.6) and (2.7).

As we have seen, the sequence of intermediate equilibrium payoffs, $\{(\widehat{u}_n^b, \widehat{u}_n^s)\}_{n=1}^N$, satisfies $(\widehat{u}_n^b, \widehat{u}_n^s) = (\widetilde{u}_n^b, \widetilde{u}_n^s)$ for n even. Consider two bargaining rounds, $n-1$ and $n+1$, where n is an odd number. So, $(\widetilde{u}_{n-1}^b, \widetilde{u}_{n-1}^s)$ and $(\widetilde{u}_{n+1}^b, \widetilde{u}_{n+1}^s)$ are corresponding equilibrium intermediate payoffs.

Fix some τ and let $nz/N \to \tau$ as N goes to infinity. Let $\Delta u^b = \widetilde{u}_{n+1}^b - \widetilde{u}_{n-1}^b$ (note that $\widetilde{u}_{n+1}^b = \widetilde{u}_n^b$) denote the buyer's incremental payoffs (on the equilibrium path) in rounds $n-1$ and $n+1$, and $\Delta u^s = \widetilde{u}_{n+1}^s - \widetilde{u}_{n-1}^s$ (note that $\widetilde{u}_n^s = \widetilde{u}_{n-1}^s$) denote the producer's incremental payoff (on the equilibrium path) in rounds $n-1$ and $n+1$. Similarly, let $\Delta z = 2z/N$. Then we have

$$H(\widetilde{u}_{n-1}^b, \widetilde{u}_{n-1}^s;\ \frac{n-1}{N}z) = 0 \tag{B.21}$$

$$H(\widetilde{u}_{n-1}^b + \Delta u^b, \widetilde{u}_{n-1}^s;\ \frac{n-1}{N}z + \frac{\Delta z}{2}) = 0 \tag{B.22}$$

$$H(\widetilde{u}_{n-1}^b + \Delta u^b, \widetilde{u}_{n-1}^s + \Delta u^s;\ \frac{n-1}{N}z + \Delta z) = 0. \tag{B.23}$$

According to (B.21) and (B.22), the producer's intermediate payoff is unchanged at $\widetilde{u}_{n-1}^s$ while the consumer's intermediate payoff increases by Δu^b. The amount of assets up for negotiation on the n^{th} frontier are nz/N. According to (B.23), at the end of round $n+1$ the intermediate payoffs are obtained by moving vertically from the n^{th} frontier to the $(n+1)^{\text{th}}$ frontier (since $n+1$ is even).

A first-order Taylor series expansion of (B.22) in the neighborhood of $(u^b, u^s, \tau) = \left(\widetilde{u}_{n-1}^b, \widetilde{u}_{n-1}^s, \frac{n-1}{N}z\right)$

yields:

$$H(\widetilde{u}^b_{n-1} + \Delta u^b, \widetilde{u}^s_{n-1};\ \frac{n}{N}z) = H_1 \Delta u^b + H_3 \frac{\Delta z}{2} + o(\Delta u^b) + o(\frac{1}{N}),$$

where $\lim_{N\to\infty, nz/N\to\tau} \frac{o(\Delta u^b)}{\Delta u^b} = \lim_{N\to\infty} N o(\frac{1}{N}) = 0$, we used that $H(\widetilde{u}^b_{n-1}, \widetilde{u}^s_{n-1};\ \frac{n-1}{N}z) = 0$ from (B.21), and the partial derivatives H_1, H_2, and H_3 are evaluated at $\left(\widetilde{u}^b_{n-1}, \widetilde{u}^s_{n-1}, \frac{n-1}{N}z\right)$. Similarly, a first-order Taylor series expansion of (B.23) yields

$$H(\widetilde{u}^b_{n-1}+\Delta u^b, \widetilde{u}^s_{n-1}+\Delta u^s;\ \frac{n+1}{N}z) = H_1 \Delta u^b + H_2 \Delta u^s + H_3 \Delta z + o(\Delta u^b) + o(\Delta u^s) + o(\frac{1}{N}),$$

where $\lim_{N\to\infty, nz/N\to\tau} \frac{o(\Delta u^b)}{\Delta u^b} = \lim_{N\to\infty, nz/N\to\tau} \frac{o(\Delta u^s)}{\Delta u^s} = \lim_{N\to\infty} N o(\frac{1}{N}) = 0$. Using that $H = 0$ for payoffs on the Pareto frontiers, we obtain that

$$\begin{aligned}
H_1 \Delta u^b + o(\Delta u^b) &= -H_3 \frac{\Delta z}{2} + o(\frac{1}{N}), \\
H_1 \Delta u^b + o(\Delta u^b) + H_2 \Delta u^s + o(\Delta u^s) &= -H_3 \Delta z + o(\frac{1}{N}), \\
o(\Delta u^b) + H_2 \Delta u^s + o(\Delta u^s) &= -H_3 \frac{\Delta z}{2} + o(\frac{1}{N}).
\end{aligned}$$

From the first one and rearranging terms, we obtain

$$\frac{\Delta u^b}{\Delta z} = -\frac{H_3}{2H_1} + \frac{o(\Delta u^b)}{H_1 \Delta z} + \frac{o(\frac{1}{N})}{H_1 \Delta z}.$$

Note that

$$\lim_{N\to\infty} \frac{o(\frac{1}{N})}{H_1 \Delta z} = \frac{o(\frac{1}{N})N}{H_1 z} = 0 \text{ and } \lim_{N\to\infty} \frac{o(\Delta u^b)}{H_1 \Delta z} = \lim_{N\to\infty} \frac{o(\Delta u^b)}{H_1 z \Delta u^b}(\Delta u^b N) = 0,$$

where

$$\Delta u^b N = (\widetilde{u}^b_{n+1} - \widetilde{u}^b_{n-1})N \in [[1 - v'(\widetilde{y}_{n+1})/u'(\widetilde{y}_{n+1})]z, [1 - v'(\widetilde{y}_{n-1})/u'(\widetilde{y}_{n-1})]z]$$

and hence its limit exists and is bounded away from zero by the concavity of the function $S(\bullet)$. Thus,

$$\frac{\partial u^b}{\partial \tau} = \lim_{N\to\infty} \frac{\Delta u^b}{\Delta z} = -\frac{1}{2}\frac{H_3}{H_1} = -\frac{1}{2}\frac{\partial H/\partial \tau}{\partial H/\partial u^b}.$$

Similarly, combining these two equations and rearranging, we obtain

$$\frac{\Delta u^s}{\Delta z} = -\frac{H_3}{2H_2} + \frac{o(\Delta u^b)}{H_2 \Delta z} + \frac{o(\Delta u^s)}{H_2 \Delta z} + \frac{o(\frac{1}{N})}{H_2 \Delta z}.$$

By the same arguments, we have

$$\frac{\partial u^s}{\partial \tau} = \lim_{N\to\infty} \frac{\Delta u^s}{\Delta z} = -\frac{1}{2}\frac{H_3}{H_2} = -\frac{1}{2}\frac{\partial H/\partial \tau}{\partial H/\partial u^s}.$$

These correspond to (2.6) and (2.7).

Asymmetric bargaining powers

Here we revise our game to support gradual Nash solution with asymmetric bargaining power, denoted by θ. The parameter θ affects the game as follows. We assume that the number of rounds is an even number N, and the producer is the one making the first offer and the consumer is making the last offer.

1. In each round $n \in \{1, 3, \ldots, N-1\}$, it is the producer's turn to make an offer, with asset transfer within the range $[0, 2(1-\theta)z/N]$; the consumer then decides to accept

or reject the offer.

2. In each round $n \in \{2, 4, \ldots, N\}$, it is the consumer's turn to make an offer, with asset transfer within the range $[0, 2\theta z/N]$; the producer then decides to accept or reject the offer.

Note that at the end of an odd round n, the maximum cumulative asset transfer is $[2(n-1)+2(1-\theta)]z/N$, and at the end of an even round n, the maximum cumulative asset transfer is nz/N.

As before, to solve the game, we need to solve all possible subgames. Also, such subgame can still be characterized by $[(u_0^b, u_0^s), J]$, where (u_0^b, u_0^s) is the intermediate payoff at the beginning of the subgame and J is the number of remaining bargaining rounds.

Proposition B.1. *Fix some $\theta \in [0, 1]$. There exists a SPE in each alternating-ultimatum offer game, and all SPE share the same final payoffs. When the output level corresponding to the final payoffs is less than y^*, the SPE is unique and is simple; otherwise, there is a unique simple SPE. Moreover, in any simple SPE, the intermediate payoffs, $\{(u_n^b, u_n^s)\}_{n=1,2,\ldots,N}$, converge to the solution $[u^b(\tau), u^s(\tau)]$ to the differential equations (2.19) and (2.20) as N approaches ∞ and $[(n-1)+2(1-\theta)]z/N$ or nz/N approaches τ.*

Note that Proposition 2.1 is a special case of Proposition B.1 with $\theta = 1/2$. The proof of Proposition B.1 follows exactly the same outline as that of Proposition 2.1. In particular, we will use the same technique to compute the final payoffs for any subgame, but with necessary modification to accommodate the fact that the consumer has control over θ fraction of assets to be negotiated every two rounds. As before, we can denote an arbitrary subgame by $[(u_0^b, u_0^s), J]$ with $0 \leq u_0^b + u_0^s < u(y^*) - v(y^*)$.

The final payoff is computed as follows. Define $\{(\widetilde{u}_j^b, \widetilde{u}_j^s)\}_{j=0}^J$ as $(\widetilde{u}_0^b, \widetilde{u}_0^s) = (u_0^b, u_0^s)$, and

$$H(\widetilde{u}_j^b, \widetilde{u}_{j-1}^s, p_0 + 2\theta z/N + (j-1)z/N) = 0, \text{ and } \widetilde{u}_j^s = \widetilde{u}_{j-1}^s, \text{ for } j \geq 1 \text{ odd}, \quad \text{(B.24)}$$

$$H(\widetilde{u}_{j-1}^b, \widetilde{u}_j^s, p_0 + jz/N) = 0, \text{ and } \widetilde{u}_j^b = \widetilde{u}_{j-1}^b, \text{ for } j \geq 2 \text{ even}, \quad \text{(B.25)}$$

where $p_0 = u[S^{-1}(u_0^b + u_0^s)] - u_0^b$. Below we show that the final equilibrium payoffs for the agents are given by $(\widetilde{u}_J^b, \widetilde{u}_J^s)$.

The logic behind this construction is exactly the same as the symmetric case, except for the fact that the consumer and the producer controls different shares of assets up for negotiation. In particular, when $J = 1$, the game $[(u_0^b, u_0^s), 1]$ is a standard take-it-or-leave-it offer game (with the consumer making the offer). Since the consumer can offer up to additional $2\theta z/N$ units of assets, the final payoff is computed by a rightward shift to next Pareto frontier with intermediate payments $p_0 + 2\theta z/N$, as in (B.24) with $j = 0$. When $J = 2$, the producer makes the first offer and take the final payoff for consumer in case he rejects the offer as given. Note that with $J = 2$ the final Pareto frontier has intermediate payment of $p_0 + 2z/N$, as in (B.25) with $j = 0$.

To compute the intermediate payoffs, we first define the functions F_r and F_u analogous to (B.11) and (B.12):

$$F_r(u^b, u^s) = (u^{b\prime}, u^{s\prime}) \text{ such that } u^{s\prime} = u^s \text{ and } H(u^{b\prime}, u^s, p + 2\theta z/N), \quad \text{(B.26)}$$

$$F_u(u^b, u^s) = (u^{b\prime}, u^{s\prime}) \text{ such that } u^{b\prime} = u^b \text{ and } H(u^b, u^{s\prime}, p + 2(1-\theta)z/N), \quad \text{(B.27)}$$

where $p = u[S^{-1}(u^b + u^s)] - u^b$. Now we are ready to explain how to compute intermediate payoffs. Consider a game $[(u_0^b, u_0^s), J]$ with J even. Using the same backward induction

argument as in the symmetric case, if $(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$ is the equilibrium intermediate payoff for the agents at the end of round-$(J-1)$ bargaining, then

$$F_r(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1}) = (\widetilde{u}^b_J, \widetilde{u}^s_J). \tag{B.28}$$

As before, the solution would be obtained by a leftward shift, but, under θ, to the lower Pareto frontier with intermediate payment lowered by $2\theta z/N$; that is,

$$H(\widehat{u}^b_{J-1}, \widetilde{u}^s_J, p_0 + Jz/N - 2\theta z/N) = 0,\ \widehat{u}^s_{J-1} = \widetilde{u}^s_J. \tag{B.29}$$

Note that in this case, $(\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$ and $(\widetilde{u}^b_{J-1}, \widetilde{u}^s_{J-1})$ do not lie on the same Pareto frontier unless $\theta = 1/2$.

In general, we can still use (B.17) to compute the equilibrium intermediate payoff at the end of round-$(J-j)$ bargaining, denoted by $(\widehat{u}^b_{J-j}, \widehat{u}^s_{J-j})$, with F_r and F_u defined by (B.26)-(B.27), and we have an analogous result to that of Lemma B.2 for the existence and uniqueness of such a sequence. For j even the terms are obtained as before. For j odd, we need a second sequence, $\{(\overline{u}^b_{J-j}, \overline{u}^s_{J-j})\}_{j=0}^{J-1}$ as follows: $(\overline{u}^b_J, \overline{u}^s_J) = (\widetilde{u}^b_J, \widetilde{u}^s_J)$, and

$$\begin{aligned} H(\overline{u}^b_{J-j}, \overline{u}^s_{J-j+1}, p_0 + (J-j-1)z/N + 2(1-\theta)z/N) &= 0, && \text{(B.30)} \\ \text{and } \overline{u}^s_{J-j} &= \overline{u}^s_{J-j+1} \text{ for } j \geq 1 \text{ odd}, \\ H(\overline{u}^b_{J-j+1}, \overline{u}^s_{J-j}, p_0 + (J-j)z/N) &= 0, && \text{(B.31)} \\ \text{and } \overline{u}^b_{J-j} &= \overline{u}^b_{J-j+1} \text{ for } j \geq 1 \text{ even}. \end{aligned}$$

Graphically, for j odd, $(\overline{u}^b_{J-j}, \overline{u}^s_{J-j})$ is obtained from $(\overline{u}^b_{J-j+1}, \overline{u}^s_{J-j+1})$ by moving toward left to the next lower Pareto frontier, with a decrease of incremental transfer of $2\theta z/N$; for j even, $(\overline{u}^b_{J-j}, \overline{u}^s_{J-j})$ is obtained from $(\overline{u}^b_{J-j+1}, \overline{u}^s_{J-j+1})$ by moving downward to the next

lower Pareto frontier, with a decrease of incremental transfer of $2(1-\theta)z/N$. Note that $(\overline{u}^b_{J-1}, \overline{u}^s_{J-1}) = (\widehat{u}^b_{J-1}, \widehat{u}^s_{J-1})$ given by (B.29). Note also that, in contrast to the symmetric case, $(\widetilde{u}^b_{J-j}, \widetilde{u}^s_{J-j})$ is situated in the same Pareto frontier as $(\overline{u}^b_{J-j}, \overline{u}^s_{J-j})$ if and only if j is even; for j odd, $(\overline{u}^b_{J-j}, \overline{u}^s_{J-j})$ lies on a different frontier.

Now we show that the intermediate payoffs converge to the same limit. As in the symmetric case, consider convergence of games with N even. The limit will be the same for N odd and hence we have convergence. By the above arguments we have that the sequence of intermediate equilibrium payoffs at the end of each round is given by $\{(\widehat{u}^b_n, \widehat{u}^s_n)\}_{n=1}^N$ with $(u^b_0, u^s_0) = (0,0)$, and that $(\widehat{u}^b_n, \widehat{u}^s_n) = (\widetilde{u}^b_n, \widetilde{u}^s_n)$ for n even. Consider two bargaining rounds, n and $n+2$ with n even. So, $(\widetilde{u}^b_n, \widetilde{u}^s_n)$ and $(\widetilde{u}^b_{n+2}, \widetilde{u}^s_{n+2})$ are corresponding equilibrium intermediate payoffs. Let $\Delta u^b = \widetilde{u}^b_{n+2} - \widetilde{u}^b_n$ denote the buyer's incremental payoffs (on the equilibrium path) in rounds n and $n+2$, and $\Delta u^s = \widetilde{u}^s_{n+2} - \widetilde{u}^s_n$ denote the producer's incremental payoff (on the equilibrium path) in rounds n and $n+2$. Let $\Delta z = 2z/N$ be the corresponding change in assets. Then we have

$$H(\widetilde{u}^b_n, \widetilde{u}^s_n;\ nz/N) = 0 \tag{B.32}$$

$$H(\widetilde{u}^b_n + \Delta u^b, \widetilde{u}^s_n; \theta\Delta z + \frac{n}{N}z) = 0 \tag{B.33}$$

$$H(\widetilde{u}^b_n + \Delta u^b, \widetilde{u}^s_n + \Delta u^s;\ nz/N + \Delta z) = 0. \tag{B.34}$$

A first-order Taylor series expansion of (B.33) in the neighborhood of $(u^b, u^s, \tau) = \left(\widetilde{u}^b_n, \widetilde{u}^s_n, \frac{n}{N}z\right)$ yields:

$$H(\widetilde{u}^b_n + \Delta u^b, \widetilde{u}^s_n; \theta\Delta z + \frac{n}{N}z) = H_1\Delta u^b + H_3\theta\Delta z + o(\Delta u^b) + o(\frac{1}{N}),$$

where $\lim_{N\to\infty, nz/N\to\tau} \frac{o(\Delta u^b)}{\Delta u^b} = \lim_{N\to\infty} No(\frac{1}{N}) = 0$, we used that $H(\widetilde{u}^b_n, \widetilde{u}^s_n; \frac{n}{N}z) = 0$ from (B.32), and the partial derivatives H_1, H_2, and H_3 are evaluated at $\left(\widetilde{u}^b_n, \widetilde{u}^s_n, \frac{n}{N}z\right)$. Similarly,

a first-order Taylor series expansion of (B.34) yields

$$H(\widetilde{u}_n^b + \Delta u^b, \widetilde{u}_n^s + \Delta u^s;\ \frac{n+2}{N}z) = H_1\Delta u^b + H_2\Delta u^s + H_3\Delta z + o(\Delta u^b) + o(\Delta u^s) + o(\frac{1}{N}),$$

where $\lim_{N\to\infty, nz/N\to\tau} \frac{o(\Delta u^b)}{\Delta u^b} = \lim_{N\to\infty, nz/N\to\tau} \frac{o(\Delta u^s)}{\Delta u^s} = \lim_{N\to\infty} No(\frac{1}{N}) = 0$. Using that $H = 0$ for payoffs on the Pareto frontiers, we obtain that

$$\begin{aligned}
H_1\Delta u^b + o(\Delta u^b) &= -H_3\theta\Delta z + o(\frac{1}{N}),\\
H_1\Delta u^b + o(\Delta u^b) + H_2\Delta u^s + o(\Delta u^s) &= -H_3\Delta z + o(\frac{1}{N}),\\
H_2\Delta u^s + o(\Delta u^s) + o(\Delta u^b) &= -(1-\theta)H_3\Delta z + o(\frac{1}{N})
\end{aligned}$$

From the first equation with rearranging, we obtain

$$\frac{\Delta u^b}{\Delta z} = -\theta\frac{H_3}{H_1} + \frac{o(\Delta u^b)}{H_1\Delta z} + \frac{o(\frac{1}{N})}{H_1\Delta z}.$$

Similarly, from the third equation with rearranging, we obtain

$$\frac{\Delta u^s}{\Delta z} = -(1-\theta)\frac{H_3}{H_2} + \frac{o(\Delta u^b)}{H_2\Delta z} + \frac{o(\Delta u^s)}{H_2\Delta z} + \frac{o(\frac{1}{N})}{H_2\Delta z}.$$

Thus, we have

$$\begin{aligned}
\frac{\partial u^b}{\partial\tau} &= \lim_{N\to\infty, 2n/N\to\tau} \frac{\Delta u^b}{\Delta z} = -\theta\frac{H_3}{H_1} = -\theta\frac{\partial H/\partial\tau}{\partial H/\partial u^b},\\
\frac{\partial u^s}{\partial\tau} &= \lim_{N\to\infty, 2n/N\to\tau} \frac{\Delta u^s}{\Delta z} = -(1-\theta)\frac{H_3}{H_2} = -(1-\theta)\frac{\partial H/\partial\tau}{\partial H/\partial u^s}.
\end{aligned}$$

B.3 Proof of Proposition 2.3

We use backward induction to prove Proposition 2.3.

Round N

Consider the alternating-offer game in the last round, N. The cumulative offer up to round N is (y_{N-1}, d_{N-1}) with associated payoff (u^b_{N-1}, u^s_{N-1}). So, if no agreement is reached in round N, the terminal payoffs are (u^b_{N-1}, u^s_{N-1}). The maximum wealth that can be negotiated at the end of round N is $z_N = d_{N-1} + z/N$. We will show that at the limit, when ξ_N goes to 1 (the risk of breakdown vanishes), the unique SPE payoffs of the subgame starting at the beginning of round N are determined according to the symmetric Nash solution:

$$\max_{u^b_N, u^s_N} \left(u^b_N - u^b_{N-1}\right)\left(u^s_N - u^s_{N-1}\right) \quad \text{s.t.} \quad H(u^b_N, u^s_N, z_N) = 0. \tag{B.35}$$

The terminal payoffs maximize the Nash product subject to the constraint that they belong to the Pareto frontier associated with z_N units of wealth. The ratio of the first-order conditions give

$$\frac{u^b_N - u^b_{N-1}}{u^s_N - u^s_{N-1}} = \frac{H_2(u^b_N, u^s_N, z_N)}{H_1(u^b_N, u^s_N, z_N)}. \tag{B.36}$$

At the optimum the slope of the Nash product is equal to the slope of the Pareto frontier.

We focus on the existence of the SPE and its construction. For simplicity we assume that y^* is never achieved. For the proof of uniqueness, see Rubinstein (1982). When it is his turn to make an offer the consumer proposes the (cumulative) offer (y^b, d^b) and the producer proposes (y^s, d^s). The consumer and the producer have a reservation surplus to accept offers, u^b and

u^s, respectively, which are determined endogenously below. The consumer's offer solves:

$$u_N^b = \max_{y^b, d^b} \left\{ u(y^b) - d^b \right\} \quad \text{s.t.} \quad -v(y^b) + d^b \geq u^s \text{ and } d^b \leq z_N. \tag{B.37}$$

The consumer maximizes his surplus subject to the constraint that his offer must generate a surplus for the producer that is at least equal to u^s and the offer must be feasible, $d^b \leq z_N$. Hence, u_N^b satisfies $H(u_N^b, u^s, z_N) = 0$. A solution to (B.37) exists provided that $u(y) - v(y) \geq u^s$ where $y = \min\{u^{-1}(z_N), y^*\}$. The reservation surplus of the producer solves

$$u^s = (1 - \xi_N) u_{N-1}^s + \xi_N \left[-v(y^s) + d^s \right]. \tag{B.38}$$

If the producer rejects the offer, his expected utility is equal to the weighted average of u_{N-1}^s, if the negotiation ends, and $-v(y^s) + d^s$ if the producer has the opportunity to make a counter-offer. Similarly, the producer's offer solves:

$$u_N^s = \max_{y^s, d^s} \left\{ -v(y^s) - d^s \right\} \quad \text{s.t. } u(y^s) - d^s = u^b \text{ and } d^s \leq z_N, \tag{B.39}$$

where the reservation surplus of the consumer solves:

$$u^b = (1 - \xi_N) u_{N-1}^b + \xi_N \left[u(y^b) - d^b \right]. \tag{B.40}$$

Hence, u_N^s satisfies $H(u^b, u_N^s, z_N) = 0$. A solution to (B.39) exists provided that $u(y) - v(y) \geq u^b$ where $y = \min\{v^{-1}(z_N), y^*\}$. Substituting u^b and u^s by their expressions given by (B.38) and (B.40), the equilibrium payoffs, (u_N^b, u_N^s), solve the following system of equations:

$$H \left[u_N^b, (1 - \xi_N) u_{N-1}^s + \xi_N u_N^s, z_N \right] = 0, \tag{B.41}$$

$$H \left[(1 - \xi_N) u_{N-1}^b + \xi_N u_N^b, u_N^s, z_N \right] = 0. \tag{B.42}$$

It is standard to check that for all $\xi_N < 1$ this system admits a unique solution. See Figure B.3. By virtue of the one-stage-deviation principle, the proposed strategies form a SPE.

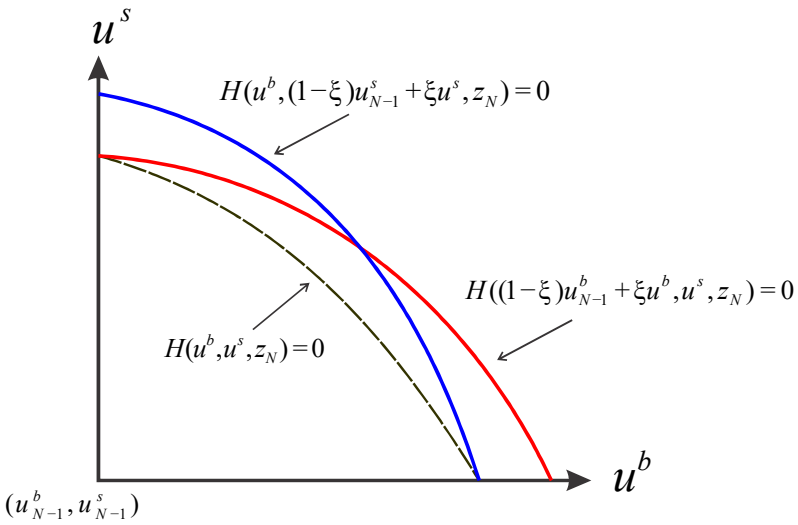

Figure B.3: Determination of equilibrium payoffs

Let us consider the limit as ξ_N approaches to 1. Using a first-order Taylor series expansion we can rewrite (B.41)-(B.42) as:

$$\begin{aligned} H\left(u_N^b, u_N^s, z_N\right) - H_2\left(u_N^b, u_N^s, z_N\right)(1-\xi_N)\left(u_N^s - u_{N-1}^s\right) &= o[(1-\xi_N)], \\ H\left(u_N^b, u_N^s, z_N\right) - H_1\left(u_N^b, u_N^s, z_N\right)(1-\xi_N)\left(u_N^b - u_{N-1}^b\right) &= o[(1-\xi_N)], \end{aligned}$$

where H_j is the partial derivative with respect to the j^{th} argument, and $o[(1-\xi_N)]/(1-\xi_N)$ converges to 0 as ξ_N converges to 1. Rearranging the terms and take limits, we obtain:

$$\lim_{\xi_N \to 1} H_2\left(u_N^b, u_N^s, z_N\right)\left(u_N^s - u_{N-1}^s\right) - H_1\left(u_N^b, u_N^s, z_N\right)\left(u_N^b - u_{N-1}^b\right) = \lim_{\xi_N \to 1} o[(1-\xi_N)]/(1-\xi_N) = 0. \tag{B.43}$$

This equation coincides with the FOC for (B.35). Hence, the solution to the alternating-offer round game corresponds to the Nash solution with disagreement points (u_{N-1}^b, u_{N-1}^s).

Terminal payoffs

We now make the following proposition for the determination of the terminal payoffs starting from any arbitrary round, and let N be the total number of rounds. When solving the game with N rounds, we take the limit on the probability of negotiation breakdown. We solve the game by taking ξ_N to one first and obtain the solution to the subgame beginning from round N. Then we solve round $N-1$, taking the limit of ξ_N at 1 as given. Then we take ξ_{N-1} to one, and so on.

We also need to expand the notation slightly. Let $\left(u_{n-1}^b, u_{n-1}^s\right)$ be a given intermediate payoff at the beginning of round n, and let d_{n-1} be the corresponding cumulative transfer of assets; i.e., $H\left(u_{n-1}^b, u_{n-1}^s, d_{n-1}\right) = 0$. Define $F\left(u_{n-1}^b, u_{n-1}^s\right) = \left(u_n^b, u_n^s\right)$ to be the solution of

$$\max_{u_n^b, u_n^s} \left(u_n^b - u_{n-1}^b\right)\left(u_n^s - u_{n-1}^s\right) \quad \text{s.t.} \quad H\left(u_n^b, u_n^s, d_{n-1} + z/N\right) = 0. \tag{B.44}$$

Proposition B.2. *Consider the subgame starting from the beginning of round $n \in \{1, ..., N\}$ with intermediate payoffs, $\left(u_{n-1}^b, u_{n-1}^s\right)$, where $H\left(u_{n-1}^b, u_{n-1}^s, d_{n-1}\right) = 0$. Take limits in the following order: $\xi_N \to 1$, $\xi_{N-1} \to 1$,..., $\xi_n \to 1$. The terminal payoffs, $\left(u_N^b, u_N^s\right)$, are obtained recursively from $\left(u_{n-1}^b, u_{n-1}^s\right)$ according to:*

$$\max_{u_{n+j}^b, u_{n+j}^s} \left(u_{n+j}^b - u_{n+j-1}^b\right)\left(u_{n+j}^s - u_{n+j-1}^s\right) \quad s.t. \quad H\left(u_{n+j}^b, u_{n+j}^s, z_{n+j}\right) = 0, \quad j = 0...N-n, \tag{B.45}$$

where $z_{n+j} = d_{n-1} + (1+j)z/N$.

The recursion (B.45) generates a sequence of payoffs, $\left\{\left(u_{n+j}^b, u_{n+j}^s\right)\right\}_{j=0}^{N-n}$, where each element, $\left(u_{n+j}^b, u_{n+j}^s\right)$, corresponds to the Nash solution of a bargaining problem with endogenous disagreement points, $\left(u_{n+j-1}^b, u_{n+j-1}^s\right)$, and Pareto frontier corresponding to the wealth z_{n+j}. We illustrate this construction in Figure 2.6 for the subgame starting in $N-2$.

We prove the proposition by induction. We have shown that the proposition holds for round N. We now show that if the proposition holds for some arbitrary round n, then it holds for round $n-1$. Consider the beginning of round $n-1$ with intermediate payoffs, $\left(u_{n-2}^b, u_{n-2}^s\right)$, where $H\left(u_{n-2}^b, u_{n-2}^s, d_{n-2}\right) = 0$. We also assume that at round $n-1$, it is the consumer to make the first offer.

In order to characterize the outcome of the alternating offer bargaining game in round $n-1$ we need to compute the payoffs in case the negotiation ends without an agreement. In the event of a breakdown in round $n-1$, then the players move to round n but keep the same intermediate payoffs, $\left(u_{n-2}^b, u_{n-2}^s\right)$. By inductive assumption, since the proposition holds for round n, the terminal payoffs in that subgame, denoted $\left(u_{N-1}^b, u_{N-1}^s\right)$, are given by $\left(u_{N-1}^b, u_{N-1}^s\right) = F^{N-n}\left(u_{n-1}^b, u_{n-1}^s\right)$, if we take the limits $\xi_N \to 1$, $\xi_{N-1} \to 1$,..., $\xi_n \to 1$, in that order.

Since our induction hypothesis allows us to compute the terminal payoffs from any intermediate payoffs in the beginning of round n, for any outcome from round $n-1$, we can compute the continuation value. First let

$$\mathcal{H}_N = \{(u_N^b, u_N^s) \geq 0 : H\left(u_N^b, u_N^s, z_N\right) \geq 0\}$$

be the set of all possible individually rational final payoffs given the initial disagreement point. We use $\mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s)$ to denote the set of all terminal payoffs, (u_N^b, u_N^s), attainable from (u_{n-2}^b, u_{n-2}^s), for which the corresponding allocation is given by (y_{n-2}, d_{n-2}), according to the induction hypothesis, if an offer at round $n-1$ is accepted:

$$\begin{aligned}\mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s) \;=\; & \{F^{N-n+1}(u_{n-1}^b, u_{n-1}^s) : \exists (y_{n-1}, d_{n-1}) \geq (y_{n-2}, d_{n-2}), d_{n-1} - d_{n-2} \leq z/N \\ & \text{such that } u_{n-1}^b = u(y_{n-1}) - d_{n-1}, u_{n-1}^s = -v(y_{n-1}) + d_{n-1}\}.\end{aligned}$$

Note that $\mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s) \subset \mathcal{H}_N$ is nonempty, as $(u_{N-1}^b, u_{N-1}^s) \equiv F^{N-n+1}(u_{n-2}^b, u_{n-2}^s) \in \mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s)$, which is attained if no trade is offered. Moreover, $(\hat{u}_N^b, \hat{u}_N^s) = F^{N-n+2}(u_{n-2}^b, u_{n-2}^s) \in \mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s)$ as well, which is attained if the offer corresponding to $(\hat{u}_{n-1}^b, \hat{u}_{n-1}^s) = F(u_{n-2}^b, u_{n-2}^s)$, denoted by $(\hat{y}_{n-1}, \hat{d}_{n-1})$, is offered and accepted. Moreover, since the cumulative offer, $(\hat{y}_{n-1}, \hat{d}_{n-1})$, is interior, i.e., $(\hat{y}_{n-1}, \hat{d}_{n-1}) > (y_{n-2}, d_{n-2})$, by continuity, there exists a neighborhood $\mathcal{O}$ around $(\hat{u}_N^b, \hat{u}_N^s)$ such that

$$\mathcal{O} \cap \mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s) \tag{B.46}$$

is open relative to $\mathcal{H}_N$.

Thus, using these terminal payoffs, the game in round $n-1$ can be reduced to the following game: the two players take turns to make an offer $(u_N^b, u_N^s) \in \mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s)$. If accepted, the game ends with the terminal payoff (u_N^b, u_N^s). Otherwise, with probability ξ_{n-1} the other player makes an offer; with probability $1 - \xi_{n-1}$ the game ends with payoff $\left(u_{N-1}^b, u_{N-1}^s\right)$. Note that only payoffs $(u_N^b, u_N^s) \geq \left(u_{N-1}^b, u_{N-1}^s\right)$ are relevant, for offers that lead to other payoffs are dominated by them. We claim that for ξ_{n-1} sufficiently large, the equilibrium payoffs, (u_N^b, u_N^s), solve the following system of equations:

$$H\left[u_N^b, (1-\xi_{n-1})u_{N-1}^s + \xi_{n-1}u_N^s, z_N\right] = 0, \tag{B.47}$$

$$H\left[(1-\xi_{n-1})u_{N-1}^b + \xi_{n-1}u_N^b, u_N^s, z_N\right] = 0. \tag{B.48}$$

First we note that if $\mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s) = \mathcal{H}_N^{IR} \equiv \{\left(u_N^b, u_N^s\right) \in \mathcal{H}_N : \left(u_N^b, u_N^s\right) \geq \left(u_{N-1}^b, u_{N-1}^s\right)\}$, then this follows from the same argument as that for round N. The set $\mathcal{H}_N^{IR}$ consists of all individually rational final payoffs relative to the disagreement point $\left(u_{N-1}^b, u_{N-1}^s\right)$. Now, since $\mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s) \subset \mathcal{H}_N$ and anything that is not individually rational is dominated by

$\left(u_{N-1}^b, u_{N-1}^s\right)$, the proof is still valid as long as the final payoffs correspond to the solutions, $[u_N^b, (1-\xi_{n-1})u_{N-1}^s + \xi_{n-1}u_N^s]$ and $\left[(1-\xi_{n-1})u_{N-1}^b + \xi_{n-1}u_N^b, u_N^s\right]$, belong to $\mathcal{U}_N^b(u_{n-2}^b, u_{n-2}^s)$. By earlier argument we know that those solutions converge to $(\hat{u}_N^b, \hat{u}_N^s)$. Thus, for ξ_{n-1} sufficiently large, such solutions also belong to $\mathcal{O}$ given by (B.46). Finally, the fact that the solution converges to the Nash solution as ξ_n approaches 1 follows exactly the same argument as round N. This proves that the proposition holds at $n-1$. Given that it holds at N, by induction it holds for all $n \geq 0$.

Intermediate payoffs

We determine the equilibrium terminal payoffs at the start of the whole game by using the initial condition $\left(u_0^b, u_0^s\right) = (0,0)$ and (B.45), i.e.,

$$\max_{u_n^b, u_n^s} \left(u_n^b - u_{n-1}^b\right)\left(u_n^s - u_{n-1}^s\right) \quad \text{s.t.} \quad H\left(u_n^b, u_n^s, \frac{n}{N}z\right) = 0.$$

We obtain a sequence $\{\left(u_n^b, u_n^s\right)\}_{n=0}^N$ where the last term corresponds to the terminal payoffs. Let's now denote $\{\left(\tilde{u}_n^b, \tilde{u}_n^s\right)\}_{n=0}^N$ the sequence of intermediate payoffs along the SPE. We determine this sequence by backward induction starting from $\left(\tilde{u}_N^b, \tilde{u}_N^s\right) = \left(u_N^b, u_N^s\right)$. Consider the alternating offer game in round N. Its solution is given by

$$\left(u_N^b, u_N^s\right) = \arg\max_{u_N^b, u_N^s} \left(u_N^b - \tilde{u}_{N-1}^b\right)\left(u_N^s - \tilde{u}_{N-1}^s\right) \quad \text{s.t.} \quad H(u_N^b, u_N^s, z) = 0.$$

By the definition of $\{\left(u_n^b, u_n^s\right)\}_{n=0}^N$ it follows that $\left(\tilde{u}_{N-1}^b, \tilde{u}_{N-1}^s\right) = \left(u_{N-1}^b, u_{N-1}^s\right)$.

Let's now move to round $N-1$. The disagreement point is $\left(\hat{u}_{N-1}^b, \hat{u}_{N-1}^s\right)$ solution to

$$\left(\hat{u}_{N-1}^b, \hat{u}_{N-1}^s\right) = \arg\max_{u_{N-1}^b, u_{N-1}^s} \left(u_{N-1}^b - \tilde{u}_{N-2}^b\right)\left(u_{N-1}^s - \tilde{u}_{N-2}^s\right) \quad \text{s.t.} \quad H\left(u_{N-1}^b, u_{N-1}^s, \frac{N-1}{N}z\right) = 0.$$

Given this disagreement point the terminal payoffs solve:

$$\max_{u_N^b, u_N^s} \left(u_N^b - \hat{u}_{N-1}^b\right)\left(u_N^s - \hat{u}_{N-1}^s\right) \quad \text{s.t.} \quad H(u_N^b, u_N^s, z) = 0.$$

It follows that $\left(\hat{u}_{N-1}^b, \hat{u}_{N-1}^s\right) = \left(u_{N-1}^b, u_{N-1}^s\right)$ and hence $\left(\tilde{u}_{N-2}^b, \tilde{u}_{N-2}^s\right) = \left(u_{N-2}^b, u_{N-2}^s\right)$. We can iterate this procedure and obtain that $\left(\tilde{u}_n^b, \tilde{u}_n^s\right) = \left(u_n^b, u_n^s\right)$ for all n. This then proves (2.23).

Gradual bargaining: limit as $N \to \infty$

The FOCs of the Nash problems above give

$$\frac{u_n^s - u_{n-1}^s}{u_n^b - u_{n-1}^b} = \frac{H_1\left(u_n^b, u_n^s, \frac{n}{N}z\right)}{H_2\left(u_n^b, u_n^s, \frac{n}{N}z\right)}.$$

Denote $\tau = nz/\delta N$. Divide both the numerator and the denominator of the left side by $z/\delta N$ and take the limit as N tends to infinity to obtain $u^{s\prime}(\tau)/u^{b\prime}(\tau)$. This gives:

$$\frac{u^{s\prime}(\tau)}{u^{b\prime}(\tau)} = \frac{H_1\left(u_\tau^b, u_\tau^s, \delta\tau\right)}{H_2\left(u_\tau^b, u_\tau^s, \delta\tau\right)}.$$

This differential equation coincides with (2.8).

B.4 Repeated Rubinstein game: the asymmetric case

We now generalize the game of Section 2.3 and study succinctly the case where consumers and producers are asymmetric by assuming that they bargain according to the generalized Nash solution in each of the $N \in \mathbb{N}$ stages of the game. The consumer's bargaining power is θ and the producer's bargaining power is $1 - \theta$. As before, one could provide strategic foundations

for the use of the generalized Nash solution in each stage by considering a Rubinstein (1982) alternating-offer game where the risk of breakdown after an offer has been rejected depends on the identity of the responder.

The N-round game is solved by backward induction. Consider the last stage and suppose the interim agreement is $\tilde{o} \equiv (\tilde{y}, \tilde{p})$ with $\tilde{y} < y^*$ (so that there are gains from trade). The solution to the subgame with a single remaining stage, $o_1(\tilde{o}) \equiv (y_1, p_1)$, is

$$o_1(\tilde{o}) \in \arg\max_{y_1, p_1} \left[u(y_1) - p_1 - u(\tilde{y}) + \tilde{p}\right]^{\theta} \left[-\upsilon(y_1) + p_1 + \upsilon(\tilde{y}) - \tilde{p}\right]^{1-\theta}$$
$$\text{s.t.} \quad p_1 - \tilde{p} \leq \frac{z}{N}. \quad \text{(B.49)}$$

The payoffs in case of disagreement correspond to $\tilde{o}$. The feasibility constraint requires that the consumer does not spend more than the last z/N units of assets on the bargaining table. We now move to stage $n = N - 1$ where we keep the same notation for the interim agreement, $\tilde{o} = (\tilde{y}, \tilde{p})$. The disagreement point is then $o_1(\tilde{o})$. The solution is the final outcome, $o_2(\tilde{o}) \equiv (y_2, p_2)$, given by:

$$o_2(\tilde{o}) \in \arg\max_{y_2, p_2} \left[u(y_2) - p_2 - u(y_1) + p_1\right]^{\theta} \left[-\upsilon(y_2) + p_2 + \upsilon(y_1) - p_1\right]^{1-\theta}$$
$$\text{s.t.} \quad p_2 - \tilde{p} \leq \frac{2z}{N}. \quad \text{(B.50)}$$

The players who have perfect foresight negotiate the final outcome, o^2, by taking into account that the agreement of stage $N - 1$ affects the outcome of the last stage as given by (B.49). The solution is obtained by applying the generalized Nash solution recursively. Given $\tilde{o}$, the disagreement point in stage $N-1$, (y_1, p_1), is obtained from (B.49). Given (y_1, p_1), the offer (y_2, p_2) is obtained from (B.50). We need to show that there is an interim agreement in $N-1$ that makes (y_2, p_2) feasible in round N. From the comparison of (B.49) and (B.50) it follows immediately that this interim agreement is $o_1(\tilde{o})$. So, ultimately, it is as if the solution in

each stage is the naive general Nash solution with a backward-looking disagreement point.

We can iterate this reasoning to obtain a sequence of offers, $\{(y^n, p^n)\}_{n=0}^N$ with $(y^0, p^0) = (0, 0)$, that satisfies:

$$(y_n, p_n) \in \arg\max_{y,p} \left[u(y) - p - u(y_{n-1}) + p_{n-1}\right]^{\theta} \left[-\upsilon(y) + p + \upsilon(y_{n-1}) - p_{n-1}\right]^{1-\theta}$$
$$\text{s.t.} \quad p_n \leq \frac{nz}{N}. \quad \text{(B.51)}$$

As long as $p_n \leq \frac{nz}{N}$ binds, the solution takes the form:

$$\frac{z}{N} = \frac{(1-\theta)\upsilon'(y_n)\left[u(y_n) - u(y_{n-1})\right] + \theta u'(y_n)\left[\upsilon(y_n) - \upsilon(y_{n-1})\right]}{\theta u'(y_n) + (1-\theta)\upsilon'(y_n)}.$$

Summing across all stages, and assuming that $y_N < y^*$, the total output solves

$$z = \sum_{n=1}^{N} \int_{y_{n-1}}^{y_n} \frac{(1-\theta)\upsilon'(y_n)u'(x) + \theta u'(y_n)\upsilon'(x)}{\theta u'(y_n) + (1-\theta)\upsilon'(y_n)} dx.$$

The integrand is a weighted average of $u'(x)$ and $\upsilon'(x)$ where the weights depend on bargaining powers. As N goes to infinity, the right side converges to the asymmetric gradual solution described in Section 2.2.3.

Appendix C

Supplementary material for Chapter 3

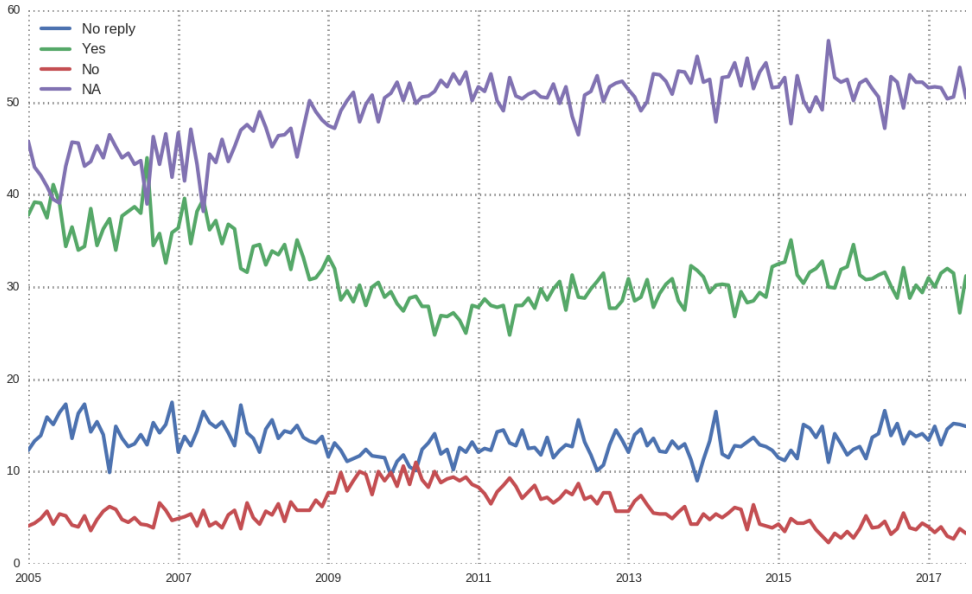

Figure C.1: "During the last three months, was your firm able to satisfy its borrowing needs?", NFIB, n=1031.

Appendix D

Supplementary material for Chapter 4

D.1 Results tables

See Tables D.1 and D.2.

D.2 Alternative calibrations

Farboodi et al. (2020) and Bethune and Korinek (2020) Farboodi et al. (2020) and Bethune and Korinek (2020), and the present paper use different approaches to calibrate the timeline of the model as well as $\alpha\tau$.

To match the timeline of the model to the real world, Farboodi et al. (2020) set time $t = 0$ to March 13th, 2020. This corresponds to the date when the authors first observe social distancing in the US, from which they infer that it corresponds to the date when the population became aware of the virus. They then make use of the fact that by then, 51 fatalities had been reported, in order to derive the implied distribution of S, I and R at time 0.

	R^*/P (%)	Deaths (M)	Max I/P (%)	Peak date	Min S^p/S (%)	End date
			Benchmark			
	96.63	1.9665	35.73	12-Apr-20	-	11-Sep-20
x			**Probability-based rule**			
0	78.15	1.5903	1.66	21-Feb-25	0.00	23-Jan-27
0.1	78.15	1.5904	1.65	8-Dec-24	9.83	13-Nov-26
0.2	78.15	1.5904	1.64	18-Aug-24	19.24	24-Jul-26
0.3	78.15	1.5903	1.57	7-Jul-23	29.34	19-Jun-25
0.4	78.15	1.5903	3.70	27-Dec-20	38.77	18-Nov-22
0.5	78.14	1.5903	5.44	31-Jul-20	34.87	2-Jun-22
0.6	78.14	1.5903	5.70	7-Jun-20	32.91	30-Mar-22
0.7	78.14	1.5903	5.88	9-May-20	31.77	21-Feb-22
0.8	78.15	1.5903	5.97	22-Apr-20	31.37	29-Jan-22
0.9	78.37	1.5948	6.14	10-Apr-20	30.86	28-Dec-21
1	79.39	1.6157	6.53	2-Apr-20	30.40	22-Oct-21
T			**Fatigue rule (months)**			
1	79.39	1.6157	6.53	16-Apr-20	0	5-Nov-21
5	79.39	1.6157	6.50	11-Aug-20	0	2-Mar-22
10	79.39	1.6157	6.51	4-Jan-21	0	26-Jul-22
15	79.39	1.6157	6.52	30-May-21	0	19-Dec-22
20	79.4	1.6158	6.50	24-Oct-21	0	15-May-23
25	79.39	1.6156	6.50	20-Mar-22	0	9-Oct-23
29	79.39	1.6158	6.44	16-Jul-22	0	3-Feb-24
30	79.40	1.6158	1.73	16-Feb-22	0	22-Oct-23
31	79.04	1.6085	1.72	22-Feb-22	0	26-Nov-23
32	78.49	1.5974	1.68	11-Mar-22	0	17-Jan-24
33	78.15	1.5904	1.66	19-Mar-22	0	20-Feb-24
34	78.15	1.5904	1.66	19-Mar-22	0	20-Feb-24
35	78.15	1.5904	1.66	19-Mar-22	0	20-Feb-24
36	78.15	1.5904	1.66	19-Mar-22	0	20-Feb-24
A			**Active-cases rule (% of P)**			
0.1	78.33	1.5942	1.67	-	17.10	30-Jun-22
0.2	79.39	1.6156	2.03	-	13.97	2-Jan-22
0.3	79.39	1.6157	3.04	-	9.22	3-Nov-21
0.4	79.39	1.6157	4.05	-	6.88	12-Oct-21
0.5	79.39	1.6156	5.06	-	5.53	3-Oct-21
0.6	79.39	1.6157	6.08	-	4.60	30-Sep-21

Table D.1: Full set of results for the calibrated SIR model with participation. Peak date corresponds to the date when the highest I/P is reached. End date corresponds to the date when $I < 1$.

	R^*/P (%)	Deaths (M)	Max I/P (%)	Peak date	Max S^m/S (%)	End date
Low k	72.51	1.4757	10.92	21-May	100	9-Feb-27
Mid k	73.87	1.5033	11.06	9-May	100	5-Oct-23
High k	74.90	1.5243	11.25	4-May	100	15-Nov-22

Table D.2: Full set of results for the calibrated SIR model with mask-wearing. Peak date corresponds to the date when the highest I/P is reached. End date corresponds to the date when $I < 1$.

In comparison, the present paper's timeline would imply 51 fatalities by March 10th absent any reaction along the participation margin. Therefore, under the assumption that the population did not react until March 13th, the two calibrations are very close.

However, because being aware of the virus before reacting through social distancing turns out to be consistent with equilibrium behavior in my model, the approach followed by Farboodi et al. (2020) may not be the most appropriate. Additionally, there is some evidence supporting that the US population may have gained awareness of the virus earlier than March 13th. On January 20th, the Center for Disease Control and Prevention (CDC) announced that three airports would start screening for COVID-19. The next day, it confirmed the first case in Washington state. On January 31st, the World Health Organization (WHO) declared a global public health emergency, and the US declared a public health emergency on February 3rd. Correspondingly, the Google Trend tools show that searches for the term "coronavirus" experienced a first significant uptick during the last week of January.

As for Bethune and Korinek (2020), they calibrate their timeline by fitting time 0 to mid-May, assuming that by that time, 0.3% of the population was infected. This is considerably different from the calibrations in this paper and in Farboodi et al. (2020). For comparison, with my calibration and when agents coordinate on participating ($x = 1$), 0.3% of the population would be infected by March 24, implying a discrepancy of more than one month at the minimum.

Table D.3 shows the results obtained when using the timing from Farboodi et al. (2020), de-

noted FSJ, compared to the timing used in this paper, denoted L. By construction, outcomes with the benchmark model are exactly identical. Surprisingly, outcomes for the participation model with $x = 1$ and for the mask model are also identical (up to numerical errors). This is because in those two cases, there is actually no reaction from individuals, on either margin, before March 13th. Thus, even if my calibration allows for agents to react, while Farboodi et al. (2020) do not, the equilibrium paths remain identical. Results are a bit different for the case with $x = 0$, where agents react from the very beginning (February 15th), if allowed to do so. Then, the timeline of the epidemic differs noticeably. In particular, it is much shorter in the Farboodi et al. (2020) setup, which is intuitive since a higher number of people is infected earlier on.

	Benchmark		**Participation**				**Masks**	
			x = 1		**x = 0**		**Mid k**	
	L	**FJS**	**L**	**FJS**	**L**	**FJS**	**L**	**FJS**
R^*/P **(%)**	96.63	96.63	79.39	79.39	78.15	78.15	73.87	73.87
Deaths (M)	1.9665	1.9665	1.6157	1.6157	1.5903	1.5904	1.5033	1.5033
Max I/P **(%)**	35.73	35.73	6.53	6.52	1.66	1.66	11.06	11.05
Peak date	12-Apr-20	12-Apr-20	2-Apr-20	2-Apr-20	21-Feb-25	22-Sep-21	9-May-20	9-May-20
Min S^p/S **(%)**	-	-	30.40	30.40	0	0	100	100
End date	11-Sep-20	11-Sep-20	22-Oct-21	22-Oct-21	23-Jan-27	28-Aug-23	5-Oct-23	5-Oct-23

Table D.3: Robustness of results to alternative timeline, used in Farboodi et al. (2020). Peak date corresponds to the date when the highest I/P is reached. End date corresponds to the date when $I < 1$.

Second, the present paper is, to the best of my knowledge, the first paper in the economic literature related to COVID-19 to use a micro-founded approach to calibrating the matching rate of individuals (αP) and the transmissibility of the virus (τ). Most of the literature estimates those two parameters jointly to match the contagion dynamics observed at the beginning of the epidemic (under the assumption that at that time, behavioral responses had not yet kicked in). Farboodi et al. (2020) target the growth rate of infections by the beginning of the epidemic, which they estimate to 30%. Noting that $\dot{I}(0)/I(0) = \alpha\tau S(0) - \gamma$, we can then easily solve for $\alpha\tau = [\dot{I}(0)/I(0) + \gamma]/S(0)$. In comparison, this paper's calibration yields a growth rate of the measure of infected agents of 35.7% by March 13th, assuming

no behavioral response until then. Bethune and Korinek (2020) target a basic reproduction number, σ, of 2.5. Since $\sigma = \alpha P \tau / \gamma$, this again allows to easily solve for $\alpha\tau = \sigma\gamma/P$. The calibration used in the present paper implies a basic reproduction number of 3.5, while that used in Farboodi et al. (2020) implies a basic reproduction number of 3.1. All of those numbers are consistent with the range of reproduction number estimated by epidemiological studies, from 1.5 to 7 (see Liu et al. (2020) for example).

Baseline utility In the calibration presented in 4.4.3, it is assumed that social contacts are the only source of income/utility, since $\alpha P \tilde{y}$ is calibrated to match the median yearly consumption in the US. This assumption may result in overstating the cost of staying home and forgoing social activities if some amount of income/utility can be gained without requiring to go out or to come into contact with other people. We now relax this assumption.

Denote y^h the baseline flow utility from consumption enjoyed by individuals regardless or whether they are at home or engaging in social interactions, and $Y = \alpha P \tilde{y} + y^h$ the total flow utility of an individual in a world with no virus (ensuring full participation from all individuals). We can now normalize Y to 1 and vary the share of consumption coming from social engagement, $\alpha P \tilde{y}/Y$, by varying $\tilde{y}$.

The table below presents the results obtained when solving for the equilibrium path with the participation margin active, imposing $x = 1$ (susceptible agents coordinate on participating whenever possible), for $\tilde{y} \in \{0.7, 0.8, 0.9, 1\}$. The specification with $\tilde{y} = 1$ corresponds to the specification assumed in the main text.

Varying the share of total consumption that requires interpersonal has little impact on the cumulative measure of agents infected by the virus in the long run.

In the short run, it does impact participation: a lower $\tilde{y}$ makes the benefit of going out smaller relative to the risk of infection. As a result, the lower $\tilde{y}$, the bigger the reaction of

	$\tilde{y}$			
	0.7	**0.8**	**0.9**	**1**
R^*/P **(%)**	78.04	78.51	78.97	79.39
Deaths (M)	1.5882	1.5978	1.6070	1.6157
Max I/P **(%)**	4.28	5.01	5.75	6.52
Peak date	31-Mar-20	01-Apr-20	01-Apr-20	02-Apr-20
Min S^p/S **(%)**	29.99	29.99	30.40	30.40
End date	25-Mar-22	21-Jan-22	02-Dec-21	22-Oct-21
Welfare cost ($T)	8.3471	8.3404	8.3325	8.3241

Table D.4: Robustness of results to varying the share of total utility requiring social contacts. Peak date corresponds to the date when the highest I/P is reached. End date corresponds to the date when $I < 1$.

susceptible agents, the lower the peak of the infection curve, and the longer the epidemic lasts.

The impact on welfare of varying the magnitude of $\tilde{y}$ results from two different forces. A lower $\tilde{y}$ directly implies that a drop in social activities is not as costly. But we saw that a lower $\tilde{y}$ encourages a comparatively larger reduction in activity, which indirectly could increase the welfare cost. The last line of table D.4 shows that the indirect effect dominates—the welfare cost is slightly lower when $\tilde{y}$ is higher.

D.3 Numerical algorithm for the SIR model with participation

In this section, I describe the algorithm used to solve the SIR model with participation. First, the model was discretized. The laws of motion for the measure of infected and recovered agents are given by

$$I_{t+1} = [1 - \gamma + \alpha\tau(P - R_t - I_t)]\, I_t \tag{D.1}$$

and

$$R_{t+1} = R_t + \gamma I_t. \tag{D.2}$$

The difference between the lifetime discounted utility of a susceptible agent and that of an infectious agent is given by

$$\omega_t = -\alpha(S_t^p + I_t + R_t) \min\left\{\tilde{y}, \tau \frac{I_t}{S_t^p + I_t + R_t} \beta \omega_{t+1}\right\} + (1-\beta)\omega^* + \beta\omega_{t+1}, \tag{D.3}$$

where $\beta \equiv 1/(1+r)$ and $\omega^* \equiv \psi/(1-\beta+\beta\gamma)$. Finally, the decision for a susceptible agent to participate or to stay home is governed by

$$a_{j,t} \begin{cases} = 0 & < \\ \in [0,1] \text{ if } \tilde{y} & = \tau \frac{I_t}{S_t^p + I_t + R_t} \beta \omega_{t+1}. \\ = 1 & > \end{cases} \tag{D.4}$$

The model is then solved forward, following these steps:

(1) Set I_0, R_0 and S_0 to their calibrated values, and pick a guess for ω_0.

(2) Making use of (D.3), compute $\omega_1(S_0^p)$ for $S_0^p = S_0$ and $S_0^p = 0$, so as to determine whether we are in the multiplicity region or one of the two dominance regions. If $\tilde{y} < \tau(I_0/P)\beta\omega_1(S_0^p = S_0)$, the unique Nash equilibrium is such that $S_0^p = 0$. If $\tilde{y} > \tau[I_0/(I_0 + R_0)]\beta\omega_1(S_0^p = 0)$, the unique Nash equilibrium is such that $S_0^p = S_0$. Otherwise, we are in the multiplicity region, and S_0^p is determined by one of the following coordination rules, chosen ex-ante and used for the whole algorithm:

- *Always participate:* $S_0^p = S_0$
- *Never participate:* $S_0^p = 0$

- *Participation with probability x:* draw d from a uniform distribution bounded by 0 and 1. If $d \leq x$, $S_0^p = S_0$, otherwise $S_0^p = 0$
- *Fatigue rule:* for a given T, if time 0 is less than T, $S_0^p = 0$, otherwise $S_0^p = S_0$
- *Active cases rule:* for a given A, if $I_0 < A$, then $S_0^p = S_0$, otherwise $S_0^p = 0$

(3) Record the corresponding I_1, R_1, S_1 and ω_1.

(4) Iterate over steps (2) and (3) to obtain ω_2, ω_3, ... up to ω_M, where M is set to a very large number.

(5) Let ϵ be an arbitrarily small number. If $|\omega_M - \omega^*| < \epsilon$, the algorithm has converged. If $\omega_M - \omega^* > \epsilon > 0$, go back to step (1), with a lower initial guess for ω_0. Otherwise, go back to step (1), with a higher initial guess for ω_0.