Essays in International Economics and Macroeconomics

by

Jaeeun Seo

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ECONOMICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by: Jaeeun Seo
Department of Economics
May 15, 2024

Certified by: Arnaud Costinot
Professor of Economics, Thesis Supervisor

Accepted by: Isaiah Andrews
Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays in International Economics and Macroeconomics

by

Jaeeun Seo

Submitted to the Department of Economics
on May 15, 2024 in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN ECONOMICS

ABSTRACT

This dissertation consists of three independent essays in international economics and macroeconomics. The first chapter emphasizes the importance of persistent heterogeneity across workers in the economy's adjustment to sectoral shocks. The second chapter explains the spatial concentration of non-tradable consumption services based on consumers' trip-chaining behavior, which results in geographically correlated consumption location choices. The final chapter provides a novel mechanism by which persistent noise in signals endogenously generates optimism through dynamic learning.

The first chapter develops a sufficient statistics approach to evaluate the impact of sectoral shocks on labor market dynamics and welfare. Within a broad class of dynamic discrete choice models that allows for arbitrary time-invariant heterogeneity across workers, I show that knowledge of steady-state intersectoral worker flows over various time horizons is sufficient to evaluate labor supply responses to shocks as well as their aggregate welfare consequences. I also establish analytically that assuming away persistent worker heterogeneity—a common practice in the literature—necessarily leads to overestimation of steady-state worker flows, resulting in systematic biases in counterfactual predictions. As an illustration of our sufficient statistics approach, I revisit the consequences of the rise of import competition from China. Using US panel data to measure steady-state worker flows, I conclude that labor reallocation away from manufacturing is significantly slower, and the negative welfare effects on manufacturing workers are more severe than those predicted by models without persistent worker heterogeneity.

The second chapter shifts the focus to non-tradable consumption services market, such as restaurants and retail, in Seoul. To understand the spatial concentration of services, we first causally identify positive spillovers across services stores. We microfound these spillovers by incorporating the trip-chaining mechanism—whereby consumers make multiple purchases during their services travel—into a quantitative spatial model that endogenizes the spatial distribution of services. When calibrated to an original survey on trip chaining, this mechanism explains about one-third of the observed concentration. However, unlike standard agglomeration mechanisms, it does not lead to inefficiency nor does it exacerbate welfare inequality. Finally, we show that spatial linkages of services consumption play a crucial role in shaping the impact of the rise of work from home and of delivery services on the distribution of services.

In the third chapter, I propose a noisy rational expectations model with persistent noise. Firms learn about economic conditions from signals, and the noise in the signals is persistent rather than *i.i.d.* over time. Firms rationally account for the persistence of noise and update their interpretations of signals based on ex post observations of true economic conditions. I show that this process gives rise to a novel mechanism by which optimism arises endogenously, which in turn amplifies or dampens the effects of underlying shocks. In particular, this model can generate the delayed overreaction in firms' expectations documented in the literature. Moreover, strategic complementarity between firms and the resulting higher-order optimism further strengthen my mechanism. Finally, I distinguish empirically my rational theory of optimism from behavioral theories by exploiting the difference in the degree of overextrapolation between consensus and individual forecasts.

Thesis supervisor: Arnaud Costinot

Title: Professor of Economics

Acknowledgements

I am deeply indebted to my advisors, Arnaud Costinot, George-Marios Angeletos, and Dave Donaldson, for generously giving their time, guidance, and support throughout my years at MIT. Their insights, feedback, and encouragement were invaluable in completing this thesis.

I am also grateful to all the faculty and friends at MIT for their inspiration and helpful comments. In particular, I thank Martin Beraja for kindly agreeing to be the third reader of this thesis. The collaborative and stimulating environment at MIT—through seminars, research meetings, and casual conversations—has made my years at MIT an incredibly rewarding experience.

Last but not least, I would like to thank my family for their unconditional love and support, especially my mother, Soyoung Lee, and my wife, Ryungha Oh. Their constant encouragement has been essential throughout my life.

Table of Contents

*These two chapters are joint work with Ryungha Oh.

List of Figures

List of Tables

Chapter 1

Sectoral Shocks and Labor Market Dynamics: A Sufficient Statistics Approach

Joint with Ryungha Oh

1.1 Introduction

Labor markets in the United States, as well as in many other countries, have been subject to a variety of shocks, from globalization to the rise of automation, oil price shocks, and the Covid-19 pandemic. Although these shocks differ in many ways, they all have one thing in common: their effects tend to be highly asymmetric across sectors, potentially creating both winners and losers among workers. How much do winners gain and losers lose? Can workers exposed to a negative shock in one sector avoid, or at least mitigate, its adverse consequences by moving to another sector? And if so, what determines the extent of this reallocation and the time it takes?

The goal of this paper is to shed light on these questions. The premise of our analysis is that both workers' exposure to shocks and their subsequent sectoral mobility depend on their comparative advantage across sectors. If comparative advantage is weak or highly transient, we expect frequent sector changes of workers, resulting in small welfare losses or even gains for negatively exposed workers. If, instead, comparative advantage is strong and persistent, we expect many workers to remain stuck in the negatively affected sector for a long time, leading to more severe negative welfare consequences. Intuitively, whether comparative advantage is weak or strong can be revealed by workers' propensity to switch sectors *prior to shocks*—i.e., steady-state gross worker flows. If so, one might be able to use such information as sufficient statistics to evaluate the impact of sectoral shocks on labor market dynamics

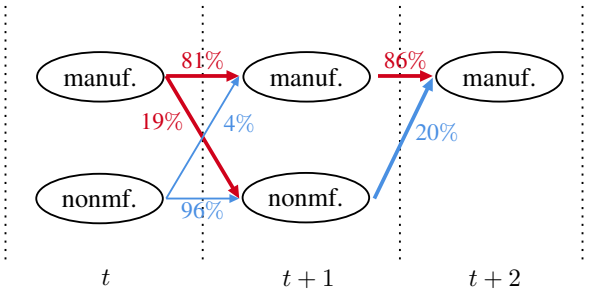


Figure 1. 1-Year and 2-Year Worker Flows in the Data

Notes: Arrows between years t and $t+1$ represent steady-state 1-year worker flows between manufacturing and non-manufacturing sectors. They represent $\Pr(s_{t+1} = \text{manuf.}|s_t = \text{manuf.}) = 1 - \Pr(s_{t+1} = \text{nonmf.}|s_t = \text{manuf.}) = 0.81$ and $\Pr(s_{t+1} = \text{nonmf.}|s_t = \text{nonmf.}) = 1 - \Pr(s_{t+1} = \text{manuf.}|s_t = \text{nonmf.}) = 0.96$. Arrows between years $t+1$ and $t+2$ provide additional information needed to compute the 2-year staying probability for the manufacturing sector. They represent $\Pr(s_{t+2} = \text{manuf.}|s_{t+1} = s_t = \text{manuf.}) = 0.86$ and $\Pr(s_{t+2} = \text{manuf.}|s_{t+1} = \text{nonmf.}, s_t = \text{manuf.}) = 0.20$. Assuming that the economy was in a steady state between years 1980 and 2000, these worker flows are computed by pooling all observations of the NLSY79 data over this period.

and welfare. In this paper, we formalize this general intuition and demonstrate its importance both theoretically and empirically.

We focus on a broad class of dynamic discrete choice models that allows for arbitrary time-invariant heterogeneity across workers. At each point in time, workers decide in which sector to work. They are subject to sector-switching costs and idiosyncratic shocks that are independently drawn each period from an extreme value distribution, as in the canonical dynamic discrete choice framework (e.g., Artuç, Chaudhuri, and McLaren, 2010). However, workers may have time-invariant differences in sector-specific productivity (or, more generally, in the utility they derive from being in a particular sector) and in sector-switching costs. We impose no restriction on these differences, in line with the general Roy model (e.g., Heckman and Honore, 1990). Within this general environment, we establish two theoretical results.

First, we develop a novel sufficient statistics approach that yields valid counterfactual predictions without estimating the sources or extent of worker heterogeneity. Aside from the discount factor and a single parameter that governs the dispersion of the idiosyncratic shocks, this approach requires only one input: steady-state gross intersectoral worker flows over different time horizons.[1] We show that this information is sufficient to construct sectoral welfare changes and dynamic labor reallocation in response to sectoral shocks, up to a first-order approximation. Steady-state worker flows summarize the effect of heterogeneity of workers on their mobility, and this is precisely what we need in order to predict the dynamic effect of sectoral shocks. We start by focusing on the labor supply side, examining the effect of exogenous changes in sectoral wages. However, we also demonstrate that when we augment the model with the

[1] Formally, we need probabilities that workers switch from sectors i to j after n periods, for all sectors i and j and all values of n.

labor demand side and endogenize wage changes, the same set of sufficient statistics, combined with knowledge of the labor demand side, can be used to perform counterfactual exercises.

Second, we show analytically how persistent worker heterogeneity shapes the consequences of sectoral shocks. We begin by characterizing the systematic bias in steady-state worker flows implied by the canonical dynamic discrete choice model without persistent worker heterogeneity (hereafter, the canonical model). Ignoring heterogeneity and the resulting self-selection would lead to underestimation of the long-run probabilities of workers remaining in the same sector. Intuitively, workers who have self-selected into a sector are more likely to choose the same sector in subsequent periods. More importantly, this bias, combined with the sufficient statistics result, implies systematic biases in the counterfactual predictions of the canonical model. In particular, the canonical model always underestimates the welfare losses of adversely affected workers and overestimates the speed of labor reallocation for given exogenous changes in sectoral wages. Moreover, we show that when wages are endogenously determined, underestimation of welfare losses is likely to be compounded by overestimation of labor reallocation. These findings suggest the importance of incorporating persistent worker heterogeneity in evaluating the consequences of sectoral shocks, which is quantified in the remainder of the paper.

The next part of our paper provides empirical estimates of our sufficient statistics for the United States. We first use the panel information in the National Longitudinal Survey of Youth 1979 (NLSY79) dataset to compute worker flows over various time horizons.[2] In line with our theoretical finding, we find that the steady-state worker flows observed in the data are inconsistent with those predicted by the canonical model without persistent worker heterogeneity. For example, the canonical model underestimates the probabilities of workers choosing the same sector after ten years by more than a factor of two. Simply controlling for workers' demographic and socioeconomic characteristics—such as gender, education, race, and age—explains little of this underestimation. This suggests that the persistent heterogeneity that drives this underestimation is mostly within these characteristics and unobserved to the econometrician.[3]

To illustrate the inconsistency, we plot the 1-year and 2-year worker flows observed in the NLSY data in Figure 1. A detailed description of the data and calculations will be provided in Section 1.4, along with additional analysis. The figure shows, for example, that 81% of typical manufacturing workers remain in manufacturing after 1 year. The canonical model, which assumes away persistent heterogeneity, necessarily implies that workers who choose to stay in the manufacturing sector in one year are as likely to stay in the following year as the typical manufacturing worker. However, the data reveal that these workers exhibit a higher probability of staying again in the following year

[2] Given data constraints, we only compute worker flows over horizons of less than or equal to 18 years. In Section 1.4, we discuss various ways of extrapolating longer-run worker flows from the available data.

[3] In addition, we provide suggestive evidence pointing to the importance of time-invariant worker heterogeneity, rather than duration-dependence mechanisms, in generating this underestimation. Specifically, we show that the underestimation is substantially reduced when the effect of worker heterogeneity is controlled for by their past sector choice history, while maintaining the contribution of duration dependence.

($86\% > 81\%$).[4] Similarly, workers who have self-selected into manufacturing in year t are more likely to choose it again in year $t+2$ even when they choose non-manufacturing sectors in year $t+1$ ($20\% > 4\%$). As a result of these discrepancies, the canonical model underestimates the probability of workers choosing the manufacturing sector again after 2 years.[5]

We then turn to estimation of the dispersion of the idiosyncratic shock. This estimation is based on the observation that the response of sectoral employment to wage shocks depends solely on the dispersion parameter and the sufficient statistics, independent of the specific details of worker heterogeneity. Thus, this parameter can be estimated by measuring the response of sectoral employment, conditional on our sufficient statistics. We put this idea into practice by extending the standard Euler equation approach in the literature to allow for arbitrary worker heterogeneity.

In the final part of our paper, we combine our empirical estimates with the sufficient statistics result to revisit two applications in the literature. First, we apply our findings to a hypothetical trade liberalization exercise of Artuç, Chaudhuri, and McLaren (2010), in which the economy experiences an unexpected permanent drop in manufacturing prices. This stylized exercise clearly illustrates how the failure to match worker flows across sectors at different horizons can lead to biases in counterfactual predictions. Second, as a more realistic application, we revisit an extensively studied topic: the impact of the rise in China's import competition on US labor markets. Following Caliendo, Dvorkin, and Parro (2019), we introduce a richer labor demand side that features international and intranational trade, input-output linkages, and multiple production inputs.[6] The results demonstrate that labor reallocation following the China shock is significantly smaller, by up to 50%, and the negative welfare effects on manufacturing workers are more severe than those predicted by the canonical model. In the absence of persistent worker heterogeneity, workers initially employed in the manufacturing sector in all states appear to benefit, on average, from the China shock because the model predicts that they can easily move to positively affected sectors. However, this is no longer the case when we do account for persistent worker heterogeneity. The welfare gains of manufacturing workers are close to zero even when they are positive, and manufacturing workers in five states experience welfare losses.

Related Literature. This paper is related to several strands of the literature. A large body of empirical literature studies the labor market impact of shocks that exhibit asymmetric effects across sectors, such as globalization (Goldberg and Pavcnik, 2007); automation (Acemoglu and Restrepo, 2020); the Covid-19 pandemic (Chetty et al., 2020); and oil price shocks (Keane and Prasad, 1996). Of particular relevance to our application is the literature that

[4] The data show considerable variation in the frequency of sector switching among workers. Some workers exhibit a high degree of sector mobility and transition frequently between different sectors. On the other hand, other workers remain employed in a particular sector for most of their careers and rarely change sectors. This fact alone is difficult to reconcile with the canonical model.

[5] These two gaps contribute almost equally to this underestimation. The canonical model implies that the 2-year staying probability is $81\% \times 81\% + 19\% \times 4\% = 66\%$, while the actual probability is $81\% \times 5\% + 19\% \times 16\% = 4\% + 3\%$ higher than this value.

[6] In contrast to their paper, we make the simplifying assumption that workers can switch sectors only within each state, and thus abstract from interstate migration.

examines the impact of the rise in China import competition on US labor markets (e.g., Autor, Dorn, and Hanson, 2013a; Autor et al., 2014; Acemoglu et al., 2016; Pierce and Schott, 2016). In this paper, we characterize welfare changes and labor reallocation in response to such sectoral shocks, taking into account the full general-equilibrium effect in a dynamic environment.

On the more structural side, recent papers have emphasized the importance of transitional dynamics in studying the effect of sectoral shocks and built models based on the dynamic discrete choice framework initially introduced in the IO literature (e.g., Rust, 1987a). An important early contribution is Artuç, Chaudhuri, and McLaren (2010), who adopt this framework to analyze the impact of a trade shock on labor market dynamics. Subsequent papers enrich this framework by incorporating more realistic elements to investigate the effect of the China shock—including trade and input-output linkages (Caliendo, Dvorkin, and Parro, 2019); involuntary unemployment due to downward nominal wage rigidities or search frictions (Rodriguez-Clare, Ulate, and Vasquez, 2022); and endogenous trade imbalances (Dix-Carneiro et al., 2023). We extend the literature by introducing arbitrary time-invariant worker heterogeneity to the framework and show how this heterogeneity significantly affects the results of counterfactual exercises.

In allowing for persistent worker heterogeneity, we relate to a large, mostly static literature that emphasizes self-selection based on comparative advantage (see Borjas, 1987; Heckman and Sedlacek, 1985; and Ricardo's theory of comparative advantage for early contributions and Lagakos and Waugh, 2013; Burstein, Morales, and Vogel, 2019; Hsieh et al., 2019; Porzio, Rossi, and Santangelo, 2022; Grigsby, 2022; and Adao, Beraja, and Pandalai-Nayar, 2023 for recent applications and developments). The work most closely related to ours includes Costinot and Vogel (2010); Adão (2016); Lee (2020), and Galle, Rodríguez-Clare, and Yi (2023), who also study the distributional effects of trade shocks. We contribute to this literature by embedding this mechanism within a dynamic discrete choice framework, which allows us to take transitional dynamics into account and highlight the importance of self-selection in a dynamic context.

In terms of methodology, we follow the sufficient statistics tradition (e.g., Chetty, 2009; Hulten, 1978; Arkolakis, Costinot, and Rodríguez-Clare, 2012; Baqaee and Farhi, 2020; McKay and Wolf, 2022; Beraja, 2023). In the present context, the use of sufficient statistics offers multiple advantages. First, by eliminating the need to estimate numerous primitives, it reduces computational costs and identification requirements while ensuring the transparency of the analysis. Second, our approach allows us to accommodate arbitrary worker heterogeneity without having to deal directly with the well-known identification challenges associated with unobserved heterogeneity (see, e.g., Heckman and Honore, 1990; French and Taber, 2011; Arellano and Bonhomme, 2017; Bonhomme, Lamadon, and Manresa, 2022).

Finally, our paper is also related to a large structural literature that incorporates rich heterogeneity—such as age, gender, education, nonpecuniary benefit, tenure, unobserved comparative advantage—in dynamic models. Prominent

examples include Keane and Wolpin (1997) and Lee and Wolpin (2006) in the labor literature and Dix-Carneiro (2014) and Traiberman (2019) in the trade literature. Our contribution to this body of literature is twofold. First, our results may serve to guide future structural analysis by suggesting that one can include worker flows over different time horizons as targeted moments in structural estimations to properly capture labor market dynamics. Second, our sufficient statistics provide a transparent way to assess which types of heterogeneity in the literature matter more for counterfactual predictions of the model.

Outline. The remainder of the paper is organized as follows. Section 1.2 presents a model of labor market dynamics with arbitrary time-invariant worker heterogeneity and derives our main sufficient statistics results. Section 1.3 analytically shows how ignoring persistent worker heterogeneity systematically affects the sufficient statistics and in turn biases counterfactual predictions regarding welfare and labor market dynamics. Section 1.4 measures our sufficient statistics using US panel data and estimates the bias in sufficient statistics due to ignoring persistent heterogeneity. Section 1.5 applies our sufficient statistics approach to study the impact of a hypothetical trade liberalization and the China shock, and Section 1.6 concludes. The Appendix contains proofs omitted in the main text.

1.2 Dynamic Discrete Choice Model with Persistent Worker Heterogeneity

In this section, we present a model that extends the dynamic discrete choice model of Artuç, Chaudhuri, and McLaren (2010) (hereafter, ACM) by allowing arbitrary time-invariant worker heterogeneity. We focus on workers' dynamic choice over sectors, though the same framework can be applied to analyze their geographic location or occupation choices by a simple relabeling.[7] We first describe the individual worker's dynamic discrete choice problem. The solution to this problem characterizes the dynamics of welfare and sectoral labor supply at individual level. We then show how we can aggregate the dynamics to macro level to derive equations that can be used to compute counterfactual changes in aggregate welfare and sectoral labor supply in response to sectoral shocks. We conclude by demonstrating how to combine this result with the labor demand side of the model to study the general equilibrium effects of sectoral shocks.

[7] Dynamic location or occupation choices have also been widely studied in the literature: location choices (Kennan and Walker, 2011; Amior and Manning, 2018; Allen and Donaldson, 2020; Bilal and Rossi-Hansberg, 2021; Howard and Shao, 2023; Kleinman, Liu, and Redding, 2023) and occupation choices (Lee, 2005; Artuç and McLaren, 2015; Traiberman, 2019), just to name a few.

1.2.1 Workers' Dynamic Discrete Choice Problem

Time is discrete and indexed by t. There are S sectors indexed by $i, j \in \mathcal{S} = \{1, \ldots, S\}$. There is a continuum of infinitely lived heterogeneous workers. We allow for arbitrary time-invariant heterogeneity of workers by assigning each worker a type $\omega \in \Omega$, which is drawn from an unknown distribution W over Ω. Importantly, the type ω may capture not only observable demographic and socioeconomic characteristics, such as gender and education, but also unobserved differences across workers. At each point in time, workers decide in which sector to work.[8] A worker of type ω who is employed in sector i at period t chooses in which sector to work in period $t+1$ in order to maximize her continuation value. The value of this worker in period t can be recursively written as[9]

$$V_{it}^{\omega} = w_{it}^{\omega} + \max_{j \in \mathcal{S}} \{\beta \, \mathbb{E}_t \, V_{jt+1}^{\omega} - C_{ij}^{\omega} + \rho^{\omega} \cdot \varepsilon_{jt}\}, \tag{1}$$

where the type-specific instantaneous utility, w_{it}^{ω}, can capture both the wage and nonpecuniary benefits from sector i in period t. For expositional purposes, we will refer to w_{it}^{ω} as sectoral wages (though in some of our applications, they will be the logarithm of the real wages). The term $C_{ij}^{\omega} \geq 0$ captures the type-specific cost of switching from sector i to j. The idiosyncratic shock ε_{jt} is worker-specific and reflects nonpecuniary motives for workers to switch sectors.[10] The expectation operator, $\mathbb{E}_t$, is taken over realizations of future wages and future idiosyncratic shocks, conditional on the information available in period t. The parameter ρ^{ω} governs the relative importance of idiosyncratic shocks in sector choice decisions and hence determines the elasticity of sectoral employment with respect to sectoral wages. We allow the sectoral wages and switching costs to vary arbitrarily across different types of workers, as can be seen from the superscripts ω in equation (1).

When $|\Omega| = 1$—which means that there are no persistent differences across workers—our model reduces to the canonical homogeneous-worker sector choice model of ACM or, more broadly, to the standard dynamic discrete choice model (e.g., Rust, 1987a). In these models, workers are ex ante homogeneous, and any ex post heterogeneity arising from different realizations of idiosyncratic shocks persists for only a single period. In the rest of the paper, we use the term *worker heterogeneity* exclusively to refer to persistent worker heterogeneity across different types of workers. When $C_{ij}^{\omega} = 0$ and $\rho^{\omega} = 0$, on the other hand, the worker problem boils down to choosing the sector that

[8] Following the timeline of ACM, workers make their sector choice decision one period in advance. This means that the sector choice decision for period t is made in period $t-1$. Once period t arrives, the instantaneous utility w_{it} from their chosen sector is realized, and workers enjoy the realized utility. They then observe the realized value of period t idiosyncratic shocks, $\{\varepsilon_{jt}\}_j$, and subsequently make a sector choice decision for period $t+1$. The expectation operator $\mathbb{E}_t$ is defined with respect to workers' information set at the time of their sector choice decision for period $t+1$.

[9] Given the time series of instantaneous utility, $\{w_{it}^{\omega}\}_{i,t,\omega}$, there are no interactions between workers of different types, so each type of worker solves problem (1) independently.

[10] Idiosyncratic shocks result in gross flows that are an order of magnitude larger than net flows, a pattern that is consistent with the observed data.

offers the highest wage, so this model reduces to the general Roy model of self-selection (e.g., Heckman and Honore, 1990).

As is standard in dynamic discrete choice models in trade, IO, and labor (e.g., Rust, 1987a; Aguirregabiria and Mira, 2010a), we assume that the idiosyncratic shocks, ε_{jt}, are drawn from a type I extreme-value distribution independently across workers, sectors, and time periods (see, e.g., McFadden, 1973). Let v_{it}^{ω} be the expected value derived from choosing sector i in period t for workers of type ω, taking the average over the realizations of idiosyncratic shocks $\{\varepsilon_{jt}\}_j$, and let F_{ijt}^{ω} denote the probability that workers of type ω in sector i in period t choose sector j in period $t+1$. Standard extreme-value algebra gives an analytical characterization of the ex ante value and sector choice probabilities:

$$v_{it}^{\omega} \equiv \mathbb{E}_{\varepsilon} V_{it}^{\omega} = w_{it}^{\omega} + \rho^{\omega} \ln \sum_{j \in \mathcal{S}} \left(\exp(\beta \, \mathbb{E}_t \, v_{jt+1}^{\omega}) / \exp(C_{ij}^{\omega}) \right)^{1/\rho}, \tag{2}$$

$$F_{ijt}^{\omega} \equiv \Pr_t(s_{t+1} = j | s_t = i, \omega) = \frac{\left(\exp(\beta \, \mathbb{E}_t \, v_{jt+1}^{\omega}) / \exp(C_{ij}^{\omega}) \right)^{1/\rho^{\omega}}}{\sum_{k \in \mathcal{S}} \left(\exp(\beta \, \mathbb{E}_t \, v_{kt+1}^{\omega}) / \exp(C_{ik}^{\omega}) \right)^{1/\rho^{\omega}}}, \tag{3}$$

where the expectation operator $\mathbb{E}_{\varepsilon}$ is taken over the realizations of $\{\varepsilon_{jt}\}_j$. Equation (2) expresses the value of being in sector i as the sum of the current period's instantaneous utility and a nonlinear aggregation of next-period values, net of switching costs. Equation (3) suggests that, all else being equal, workers are more likely to choose sectors with higher values net of switching costs. Because there is a continuum of workers for each type ω, we can characterize the law of motion of their sectoral employment share from their sector choice probabilities,

$$\ell_{jt+1}^{\omega} = \sum_{i \in \mathcal{S}} F_{ijt}^{\omega} \ell_{it}^{\omega}. \tag{4}$$

We also define the backward transition probability,

$$B_{jit}^{\omega} \equiv \Pr_t(s_t = i | s_{t+1} = j, \omega) = \frac{\ell_{it}^{\omega} F_{ijt}^{\omega}}{\ell_{jt+1}^{\omega}},$$

which is the probability that a type ω worker in sector j in period $t+1$ came from sector i in period t. We define $S \times S$ matrices F_t^{ω} and B_t^{ω}, whose (m,n)-element is F_{mnt}^{ω} and B_{mnt}^{ω}, respectively. We refer to them as the (forward) transition matrix and backward transition matrix, respectively. Note that the rows of these matrices sum to one.

1.2.2 Welfare and Labor Dynamics at the Micro Level

The system of equations (2)–(4) fully characterizes the labor supply side of the model. That is, given the series of sectoral wages, $\{w_{it}^{\omega}\}$, we can solve for the series of sectoral employment, $\{\ell_{it}^{\omega}\}$, and sectoral values, $\{v_{it}^{\omega}\}$, from

this system of equations. These are the two variables of interest in our counterfactual analysis. As a first step toward deriving a sufficient statistics result, we consider infinitesimal sectoral shocks $\{\mathrm{d}w_{it}^{\omega}\}$ and take first-order approximations of equations (2) and (4) around a steady state.[11]

A *steady state* is associated with time-invariant sectoral wages ($w_{it}^{\omega} = w_i^{\omega}$ for all t), where the type-specific value, sector choice probabilities, and sectoral labor supply remain constant over time. In line with our previous notation, we denote the steady-state forward and backward transition matrices as F^{ω} and B^{ω}, respectively. The following equations summarize the responses of the endogenous variables in terms of deviations from a steady state:

$$\mathrm{d}v_t^{\omega} = \mathrm{d}w_t^{\omega} + \beta F^{\omega}\, \mathbb{E}_t\, \mathrm{d}v_{t+1}^{\omega}, \tag{5}$$

$$\mathrm{d}\ln \ell_{t+1}^{\omega} = B^{\omega}\, \mathrm{d}\ln \ell_t^{\omega} + \tfrac{\beta}{\rho^{\omega}}(I - B^{\omega}F^{\omega})\, \mathbb{E}_t\, \mathrm{d}v_{t+1}^{\omega}, \tag{6}$$

where we use the vector notation,

$$\mathrm{d}v_t^{\omega} = \begin{pmatrix} \mathrm{d}v_{1t}^{\omega} & \cdots & \mathrm{d}v_{St}^{\omega} \end{pmatrix}^{\mathsf{T}},\ \mathrm{d}w_t^{\omega} = \begin{pmatrix} \mathrm{d}w_{1t}^{\omega} & \cdots & \mathrm{d}w_{St}^{\omega} \end{pmatrix}^{\mathsf{T}},\ \text{and}\ \mathrm{d}\ln \ell_t^{\omega} = \begin{pmatrix} \mathrm{d}\ln \ell_{1t}^{\omega} & \cdots & \mathrm{d}\ln \ell_{St}^{\omega} \end{pmatrix}^{\mathsf{T}}.$$

The algebra follows that of Kleinman, Liu, and Redding (2023), as described in Appendix A.4. It is worth highlighting the assumptions underlying these equations. Equation (5) is a direct application of the envelope theorem, often referred to the Williams-Daly-Zachary theorem in the discrete choice literature. Thus, it remains valid regardless of the distribution assumption on the idiosyncratic shock.[12] On the other hand, equation (6) is valid only under the assumption that idiosyncratic shocks follow a type I extreme-value distribution. Whereas Taylor's theorem implies that it is always possible to express $\mathrm{d}\ln \ell_{t+1}^{\omega}$ as a linear function of $\mathrm{d}\ln \ell_t^{\omega}$ and $\mathbb{E}_t\, \mathrm{d}v_{t+1}^{\omega}$ up to the first order, the sufficient statistics results we derive below rely on the coefficients of this linear function's being polynomials of the transition matrices B^{ω} and F^{ω}, which extreme-value distribution assumption guarantees.[13] A more detailed intuition is provided in Appendix A.1.4.

It is important to note that equations (5) and (6) cannot be confronted directly with data if worker type ω is unobservable. The key idea of this paper is that despite this unobservability, these type-specific equations can be aggregated into equations that solely involve a few observable statistics, which can be used to construct counterfactuals. This idea allows us to bypass the well-known challenges associated with explicitly specifying and estimating the distribution of underlying unobserved heterogeneity (e.g., Heckman and Honore, 1990).

[11] We come back to the quality of this first order approximation in Section 1.5. For the shocks that we consider, the approximation is good.

[12] In fact, the same envelope-type result extends to a much wider class of models, including models with duration dependence mechanism.

[13] This assumption is sufficient but not necessary. In Appendix A.1.4, we demonstrate that this result is not specific to the extreme-value distribution. Specifically, we show that a version of equation (6) holds for any distribution of the idiosyncratic shock in a limit case of a perturbed economy, in which dispersion of the idiosyncratic shock converges to zero. Thus, all results in this paper apply to this limit case.

In order to prepare the aggregation at the macro level, we solve these equations forward and backward to write the responses of sectoral values and sectoral employment as a function of the expected past and future wage changes $(\cdots, \mathrm{d}w^{\omega}_{t-1}, \mathrm{d}w^{\omega}_{t}, \mathrm{d}w^{\omega}_{t+1}, \cdots)$. Lemma 1 summarizes the result.

Lemma 1. *For a given sequence of changes in sectoral wages $\{\mathrm{d}w^{\omega}_{t}\}$, the changes in type-specific sectoral values and sectoral employment $\{\mathrm{d}v^{\omega}_{t}, \mathrm{d}\ln \ell^{\omega}_{t}\}$ are given by:*

$$\mathrm{d}v^{\omega}_{t} = \sum_{k\geq 0} (\beta F^{\omega})^{k}\, \mathbb{E}_{t}\, \mathrm{d}w^{\omega}_{t+k}, \tag{7}$$

$$\mathrm{d}\ln \ell^{\omega}_{t} = \tfrac{\beta}{\rho^{\omega}} \sum_{s\geq 0} (B^{\omega})^{s} (I - B^{\omega}F^{\omega}) \left(\sum_{k\geq 0} (\beta F^{\omega})^{k}\, \mathbb{E}_{t-s-1}\, \mathrm{d}w^{\omega}_{t-s+k} \right). \tag{8}$$

Workers are forward-looking, so all future shocks affect the value of workers and sectoral employment, as can be seen from equations (7) and (8). Due to the presence of switching costs and idiosyncratic shocks, labor reallocation is sluggish, so past shocks also affect sectoral employment, as can be seen from equation (8).

1.2.3 Welfare and Labor Dynamics at the Macro Level

We focus on the effect of sectoral shocks on variables aggregated across workers of different types ω. In particular, we define total sectoral employment and average sectoral value as follows:

$$\ell_{it} = \int_{\Omega} \ell^{\omega}_{it}\, \mathrm{d}W(\omega) \quad \text{and} \quad v_{it} = \int_{\Omega} v^{\omega}_{it}\, \mathrm{d}W(\omega | s = i) \tag{9}$$

where $W(\cdot | s = i)$ is the steady-state type distribution of workers in sector i. The total employment of sector i is obtained by summing the employment of different types of workers. Likewise, the average value of workers in sector i is given by taking the weighted average across different types of workers, using the steady-state type distribution of that sector as weights.[14] In so doing, we implicitly assume a utilitarian social welfare function with equal weights across all workers. Next, we define worker flow matrices. As we will demonstrate, these matrices are sufficient statistics for characterizing the welfare and labor market consequences of sectoral shocks.

Definition 1. *For each $k \in \mathbb{N}_0$, the k-*period worker flow matrix *$\mathcal{F}_k$ is an $S \times S$ matrix whose (i,j)-element is given by the steady-state share of workers in sector i who switch to sector j after k periods:*

$$(\mathcal{F}_k)_{i,j} = \Pr(s_{t+k} = j | s_t = i).$$

[14] We therefore abstract from distributional consequences of sectoral shocks across unobservable types ω. Our interest here is in comparing the welfare of workers initially employed in different sectors.

Unlike the type-specific transition matrices B^ω and F^ω in Lemma 1, these worker flow matrices can be computed directly from longitudinal information on workers' sector choices, as we will do with the NLSY data in Section 1.4.[15]

To derive an aggregation result, we make the following two assumptions.

Assumption 1. *Workers of different types share common sectoral shocks and common dispersion of idiosyncratic shocks:*

$$\mathrm{d}w_t^\omega = \mathrm{d}w_t \text{ and } \rho^\omega = \rho, \textit{ for all } t \textit{ and } \omega \in \Omega.$$

Assumption 2. *The bilateral switching costs between sectors satisfy either one of the following conditions:*

$$C_{ij}^\omega = C_{ji}^\omega \text{ or } C_{ij}^\omega = \begin{cases} C_i^\omega + \tilde{C}_j^\omega & \textit{if } i \neq j \\ 0 & \textit{if } i = j \end{cases},$$

for all $i, j \in \mathcal{S}$*, and* $\omega \in \Omega$*.*

It is worth emphasizing that the first part of Assumption 1 does not require that all workers have the same *level* of instantaneous utilities. For example, suppose workers have log utility and shocks are multiplicative to the wages of all workers. In this case, even if workers have different wages, the shocks manifest themselves as common additive shocks to instantaneous utilities for all workers.[16] Similarly, the second part of Assumption 1 does not necessarily imply that all workers have the same labor supply elasticity. Although the dispersion of idiosyncratic shocks governs the elasticity of sectoral labor supply with respect to sectoral shocks, the elasticity also depends on type-specific transition matrices, B^ω and F^ω, as can be seen in equation (8).[17] Although restrictive, this assumption is standard in the dynamic discrete choice literature, even in papers that incorporate rich heterogeneity of workers. Our approach can be applied to the models studied in those papers.

The conditions in Assumption 2 are often imposed in the literature for various purposes. For example, ACM (in Section IV.D) and Dix-Carneiro (2014) assume that the switching costs can be decomposed as in the second condition to reduce the number of parameters to be estimated. In a different context, Allen and Arkolakis (2014) and Desmet, Nagy, and Rossi-Hansberg (2018) assume the first condition for bilateral trade costs and bilateral switching costs, respectively, in order to simplify the equilibrium system into a single integral equation.

We are ready to state our main result.

[15] In Appendix A.1.1, we formally show that if we have access to infinite-length longitudinal information on workers' sector choices, we can directly observe the full series of worker flow matrices. However, since panel data have finite time dimension in practice, we can only calculate worker flow matrices $\mathcal{F}_k$ for low enough k's. In Section 1.4, we discuss various ways of extrapolating worker flow matrices.

[16] In many applications, however, the shocks of interest are known to have heterogeneous effects across *observed* types—for example, between high-skilled and low-skilled workers or across geographic regions. In such cases, we can account for this possibility by conducting the same analysis separately for each observed type of workers. This is the approach we take when we study the effect of the China shock on US labor markets in Section 1.5. However, this approach is infeasible for *unobserved* types. Therefore, we cannot dispense with the assumption that shocks are common across unobserved types.

[17] What this implies is that all heterogeneity in the elasticity of sectoral labor supply with respect to sectoral shocks is revealed by shares. This observation will be useful in deriving our sufficient statistics result.

Proposition 1. *Suppose that Assumptions 1 and 2 hold. For a given sequence of (common) changes in sectoral wages* $\{\mathrm{d}w_t\}$*, the changes in sectoral value and sectoral employment are given by:*

$$\mathrm{d}v_t = \sum_{k\geq 0} \beta^k \mathcal{F}_k \, \mathbb{E}_t \, \mathrm{d}w_{t+k}, \tag{10}$$

$$\mathrm{d}\ln \ell_t = \sum_{s\geq 0, k\geq 0} \frac{\beta^{k+1}}{\rho} (\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2}) \, \mathbb{E}_{t-s-1} \, \mathrm{d}w_{t-s+k}. \tag{11}$$

The logic behind this proposition is as follows. Starting from Lemma 1, we want to aggregate type-specific variables to the macro level. We first invoke Assumption 2, which simplifies the aggregation by giving the equality between the forward and backward transition matrices, $B^\omega = F^\omega$.[18] Although this equality is not strictly necessary for our purpose, it allows us to derive analytical results in Section 1.3 and reduces the data requirements needed to implement our sufficient statistics approach.[19] Figure A.2 shows that these two transition matrices are indeed very similar at the level of granularity at which both matrices can be computed from the data. After imposing this equality, we aggregate equations (7) and (8) to derive equations (10) and (11), respectively. In particular, the kth powers of the type-specific transition matrix $(F^\omega)^k$ are aggregated into the k-period worker flow matrix $\mathcal{F}_k$. Intuitively, since the (i,j)-element of the former is given by $\Pr(s_{t+k} = j | s_t = i, \omega)$, we can obtain the (i,j)-element of the latter, $\Pr(s_{t+k} = j | s_t = i)$, by taking an average over the type distribution of workers in sector i, $\Pr(\omega | s_t = i)$. In Appendix A.1.1, we prove Proposition 1 by introducing the population-average operator, which formalizes the idea of aggregation.[20]

The role of Lemma 1 and Assumption 1 should also be clear at this point. We have just seen that products of transition matrices can be aggregated to worker flow matrices, but when they are multiplied by another type-specific variable, such as $\mathrm{d}v^\omega_{t+1}$ in (5) and $\mathrm{d}w^\omega_{t+k}$ in (7), a complication arises because the aggregation then involves a covariance term that captures the extent to which two multiplicands comove across different worker types. Since the type index ω may include unobserved heterogeneity, it is not possible to characterize the covariance term without specifying the precise form of worker heterogeneity. Lemma 1 and Assumption 1 allow us to bypass this problem.

Proposition 1 establishes that in order to calculate the counterfactual changes in aggregate welfare and sectoral employment for a known sequence of exogenous wage changes, $\{\mathrm{d}w_t\}$, we only require knowledge of the worker flow matrices, $\{\mathcal{F}_k\}$, and the shape parameter ρ. In particular, we do not need full knowledge of the detailed worker heterogeneity (i.e., the distribution of types, W) and resulting self-selection that generates these worker flow matrices.

[18] See Appendix A.1.2 for a proof. Two matrices are equal if and only if the steady-state flow of type ω workers from sector i to j is equal to the flow from sector j to i for all sector pairs. However, the definition of the steady state does not necessarily imply this condition, since it only requires that the *total* outflow of type ω workers from a sector be equal to the *total* inflow into that sector. We further need Assumption 2 to guarantee that bilateral worker flows are balanced for all sector pairs.

[19] See Appendix A.1.2 for a general version of Proposition 1 without this assumption.

[20] In static settings, a similar aggregation idea for the first-order welfare effect is advanced in Kim and Vogel (2020) and Sprung-Keyser, Hendren, and Porter (2022), under conditions similar to the first part of Assumption 1.

What matters is how frequently workers switch sectors over time, not the specific structural determinants of these patterns.[21] In Section 1.4, we estimate worker flow matrices and the value of the parameter ρ using panel data. The following corollary summarizes the discussion.[22]

Corollary 1. *Consider a sequence of exogenous changes in sectoral wages,* $\{\mathrm{d}w_t\}$*. Together with* ρ*, the worker flow matrices,* $\{\mathcal{F}_k\}$*, constitute sufficient statistics for changes in sectoral values,* $\{\mathrm{d}v_t\}$*, and sectoral employment,* $\{\mathrm{d}\ln \ell_t\}$*.*

We have so far focused on the labor supply side and considered exogenously given wage changes. In Section 1.2.4, we show that even when we endogenize the wage by augmenting the model with the labor demand side, the same set of sufficient statistics, combined with knowledge of the labor demand side, constitutes sufficient statistics for counterfactual changes in sectoral values and sectoral employment.

At this point, it is worth discussing how our sufficient statistics approach relates to structural work in this area. Our approach stands in stark contrast to the standard structural approach to accounting for worker heterogeneity. First, our approach eliminates the need to estimate many primitives. This reduces computational costs and data requirements, while ensuring the transparency of the analysis. Like other studies of sufficient statistics (see, e.g., Chetty, 2009), our method yields counterfactual predictions that are immune to the Lucas critique, without requiring knowledge of the full structure of the model. Second, this approach effectively accommodates arbitrary worker heterogeneity without encountering the well-known challenges associated with estimating the distribution of unobserved heterogeneity from the data.[23] However, these advantages do not come without costs. First, we need to make restrictions on the set of shocks that can be studied and on the heterogeneity in the dispersion of idiosyncratic shocks (Assumption 1). Second, we rely on the first-order approximation around a steady state, the validity of which depends on the sectoral shock of interest. We will revisit this issue in Section 1.5.

[21] This result is surprising because, in principle, constructing the counterfactual in a dynamic context without a precisely specified model requires estimating all dynamic elasticities of sectoral values and sectoral employment with respect to shocks at all time horizons—that is, how past as well as future shocks affect these variables. McKay and Wolf (2022) propose a method to operationalize this approach in practice, but in general estimating all elasticities is challenging due to high information requirements and limited availability of data. Proposition 1 reveals that the dynamic discrete choice framework imposes a tight connection among these dynamic elasticities. This relationship enables us to parameterize these elasticities with a single parameter to be estimated, ρ, while the worker flow matrices contain all the remaining information needed to calculate the dynamic elasticities.

[22] In Arkolakis, Costinot, and Rodríguez-Clare (2012), a distinction is made between the ex ante sufficient statistics result and the ex post result. Proposition 1 provides the ex post sufficient statistics in the sense that this result is only useful if we can estimate or directly observe the change in wages resulting from the shock of interest. This is often feasible when examining the effects of shocks that occurred in the past. However, it becomes impossible when attempting to forecast the impact of hypothetical shocks.

[23] Both Dix-Carneiro (2014) and Traiberman (2019) incorporate unobserved heterogeneity in their analyses. However, due to the identification challenge, they are constrained to use a limited number of unobserved types. While this could allow them to capture the absolute advantage of workers, it is difficult to capture the comparative advantage of workers and hence their self-selection into sectors.

1.2.4 Closing the Model: Labor Demand and Equilibrium Wages

We conclude this section by extending the sufficient statistics result to the case in which the wage is endogenously determined by the labor market equilibrium, as in Heckman and Sedlacek (1985). For this purpose, we need to specify the labor demand side of the model. Although the main contribution of this paper centers on the labor supply side, our heterogeneous worker labor supply model can be integrated with any labor demand system. The specific nature of the labor demand system depends on preferences, technology, and good market structure. For expositional purposes, we specify it in a reduced-form manner in this section and return to its structural determinants in our applications in Section 1.5. Specifically, we assume that the wage of sector i is endogenously determined by the sectoral labor allocation $\{\ell_{jt}\}_j$ and exogenous shocks $\{\varepsilon_{jt}\}_j$ that affect the marginal productivity of labor. The variable ε_{jt} encompasses sector-specific factors, such as capital stock, technology shocks, policy variables, and the like. This relationship can be expressed as $w_{it}^{\omega} = f_i^{\omega}(\{\ell_{jt}\}_j, \{\varepsilon_{jt}\}_j)$.[24] In Appendix A.1.3, we show that under Assumption 1 we can write this relationship in terms of a first-order approximation as

$$\mathrm{d}w_t = D \cdot \mathrm{d}\ln \ell_t + E \cdot \mathrm{d}\varepsilon_t, \tag{12}$$

where there is no ω-index on matrices D and E.

Combining the labor demand curve represented by this equation with the labor supply curve characterized in the previous proposition, we can define the labor market equilibrium. It consists of paths of type-specific sectoral value, v_t^{ω}; type-specific labor allocation across sectors, ℓ_{t+1}^{ω}; type-specific sector choice probabilities, F_t^{ω}; aggregate sectoral value, v_t; and aggregate labor allocation, ℓ_{t+1}, that are measurable with respect to the period-t information set, and a path of sectoral wages, w_t, that are measurable with respect to the period-$(t-1)$ information set and the period-t shock such that: *(a)* type-specific variables $\{v_t^{\omega}, \ell_{t+1}^{\omega}, F_t^{\omega}\}$ solve problem (1) given the path of wages; *(b)* aggregate variables $\{v_t, \ell_t\}$ are consistent with the type-specific variables through equation (9); *(c)* wages are determined by the marginal productivity of labor, (12); and *(d)* the labor market clears.

The following proposition shows that the same set of worker flow matrices, combined with knowledge of the labor demand side, still constitutes sufficient statistics when wages are endogenously determined by the labor market equilibrium.

[24] For simplicity, we assume that labor demand, unlike labor supply, is determined in a static manner. However, the results below could be extended to models with dynamic labor demand decisions, without affecting any of the main insights.

Proposition 2. *Suppose that Assumptions 1 and 2 hold. For a given sequence of labor demand shocks* $\{\mathrm{d}\varepsilon_t\}$*, the equilibrium values of* $\{\mathrm{d}v_t, \mathrm{d}\ln\ell_t, \mathrm{d}w_t\}_t$ *are given by solution of the following system of equations:*

$$\begin{aligned}
\mathrm{d}v_t &= \sum_{k\geq 0} \beta^k \mathcal{F}_k \,\mathbb{E}_t\, \mathrm{d}w_{t+k}, \\
\mathrm{d}\ln\ell_t &= \sum_{s\geq 0, k\geq 0} \tfrac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathbb{E}_{t-s-1}\,\mathrm{d}w_{t-s+k}, \\
\mathrm{d}w_t &= D\cdot \mathrm{d}\ln\ell_t + E\cdot \mathrm{d}\varepsilon_t.
\end{aligned}$$

The intuition is simple. Conditional on a path of wage changes across time and across sectors, we can characterize the dynamic response of sectoral employment using Proposition 1. Conditional on the dynamics of sectoral employment, we can solve for prices and wages from the labor demand side to characterize the path of (real) wage changes. The equilibrium is determined as a fixed point of these relations.

Unlike Proposition 2, which requires knowledge of the path of wages, Proposition 1 requires the path of labor demand shocks $\{\varepsilon_i\}$. One can think of it as extending Proposition 2 to the case where labor demand are not perfectly elastic.

Given the shock path, $\{\mathrm{d}\varepsilon_t\}$, we can solve the system of equations to compute changes in values and sectoral employment. Conditional on the observed series of worker flow matrices, the value of ρ, and the reduced-form specification of the labor demand side, D and E, the responses of welfare, employment, and wages to labor demand shocks do not depend on the specific details of worker heterogeneity.

1.3 Employment and Welfare Implications of Persistent Heterogeneity

The canonical dynamic discrete choice model commonly used in the trade, labor, and IO literature abstracts from persistent worker heterogeneity. The sufficient statistics results in the previous section provide a way to account for arbitrary time-invariant heterogeneity. In this section, we use these results to demonstrate why incorporating this consideration matters in evaluating the consequences of sectoral shocks on welfare and labor reallocation.

Our sufficient statistics result highlights that worker heterogeneity affects the results of counterfactual exercises only through its effect on the model's predictions for a particular set of moments of the data: the worker flow matrices $\{\mathcal{F}_k\}$.[25] In this section, we first theoretically characterize a systematic bias in worker flow matrices implied by the canonical model, which, due to the lack of persistent worker heterogeneity, imposes that the k-period worker flow

[25] Another implication of the sufficient statistics result is that, conditional on the observed worker flow matrix series, the results of counterfactual exercises remain unchanged regardless of how we specify the underlying heterogeneity. However, in practice, not all worker flow matrices of the model match those observed in the data, and, as will be seen in this section, models with different worker heterogeneity yield different worker flow matrices.

matrix is equal to the one-period worker flow matrix to the kth power. In turn, the bias in worker flows leads to systematic biases in counterfactual predictions of welfare changes and labor reallocation.

1.3.1 Steady-state Worker Flow with and without Persistent Heterogeneity

Lemma 2 characterizes the restrictions that the absence of persistent worker heterogeneity imposes on the worker flow matrices $\{\mathcal{F}_k\}$.

Lemma 2. *Without persistent worker heterogeneity ($|\Omega| = 1$), we have $\mathcal{F}_k = (\mathcal{F}_1)^k$. With (non-degenerate) persistent worker heterogeneity, we have $(\mathcal{F}_k)_{ii} > \left((\mathcal{F}_1)^k\right)_{ii}$ for all $i \in \mathcal{S}$ and $k > 1$.*

The first part of Lemma 2 shows that without persistent heterogeneity, the Markovian structure of the model implies that the same transition probabilities are applied to all workers, which enables us to compute the k-period worker flow matrix by multiplying the one-period matrix k times.[26] The second part of Lemma 2 illustrates how accommodating worker heterogeneity relaxes this restriction. With worker heterogeneity, the diagonal elements of the k-period worker flow matrices could be larger than they would be in the absence of worker heterogeneity. Accordingly, if we wrongly ignore persistent worker heterogeneity, we systematically *underestimate* the probability of workers choosing the same sector after k periods, concluding that moving across sectors is more frequent than it actually is. To understand this underestimation, consider the case of $k = 2$, where we have

$$(\mathcal{F}_2)_{ii} = \sum_{j \in \mathcal{S}} \Pr(s_{t+1} = j | s_t = i) \Pr(s_{t+2} = i | \boldsymbol{s_t = i}, s_{t+1} = j), \tag{13}$$

$$\left((\mathcal{F}_1)^2\right)_{ii} = \sum_{j \in \mathcal{S}} \Pr(s_{t+1} = j | s_t = i) \Pr(s_{t+2} = i | s_{t+1} = j).$$

When workers are heterogeneous, the additional conditioning of $s_t = i$ in equation (13) increases the likelihood of choosing sector i again in period $t + 2$, since workers who have self-selected into sector i in period t are more likely to do so in subsequent periods. This result is reminiscent of the findings in Heckman (1981) and Heckman and Singer (1984) that the average hazard rate is biased toward negative duration dependence relative to each type-specific hazard rate.

Many widely used datasets only provide information on the short-run worker flows because they do not track individual workers, nor do they provide information on workers' past sector choice history or tenure.[27] In such situations, a common approach in the literature is to assume homogeneous workers and calibrate models by matching the one-period worker flow matrix. Lemma 2 shows how this calibration practice effectively extrapolates longer-run

[26] This is known as the *Chapman-Kolmogorov equation* in the theory of Markovian processes.

[27] Even when researchers use panel data that contain the necessary information, it is unclear whether the model correctly matches longer-run worker flows unless they are directly targeted.

worker flows and why this extrapolation is necessarily biased. In Section 1.4, we indeed show that the canonical model performs poorly in matching the longer-run worker flow patterns observed in the data.

1.3.2 Counterfactual Predictions with and without Persistent Heterogeneity

Combining Lemma 2 with our sufficient statistics result, we can theoretically characterize the systematic biases in counterfactual predictions that arise from assuming away persistent heterogeneity. For the moment, we consider shocks to exogenously given wages, as in Proposition 1. For simplicity, we focus on a uniform permanent shock, either positive or negative, to a sector $s \in \mathcal{S}$ that is known to workers in period 1:

$$\mathrm{d}w_{st} = \Delta \in \mathbb{R}, \ \forall t \geq 1. \tag{14}$$

For more general shocks, Appendix A.1.6 characterizes the effect of one-time shocks, from which we can calculate the effect of any sequence of shocks.

Counterfactual Welfare Changes. We begin with welfare changes. Compared to the predictions of the canonical model, workers who are initially employed in sector s are more likely to remain in the sector and to be affected by the wage change for a longer period of time. As a result, ignoring worker heterogeneity leads to underestimation of the welfare changes of these workers. This observation is formalized in Proposition 3.

Proposition 3. *Consider a uniform permanent shock of the form* (14) *known to workers in period* 1. *The canonical model, calibrated by matching the one-period worker flow matrix, underestimates the welfare effect on workers initially employed in sector* s, $|\,\mathrm{d}v_{s1}|$.

Counterfactual Employment Changes. We now turn to labor reallocation. A shock to sector s changes the employment share of that sector over time. This labor reallocation is characterized by equation (11) of Proposition 1, which involves terms of the form $\mathcal{F}_k - \mathcal{F}_{k+2}$. For ease of notation, we define b_k as the diagonal element of $\mathcal{F}_k - \mathcal{F}_{k+2}$ that corresponds to sector s:

$$b_k \equiv (\mathcal{F}_k - \mathcal{F}_{k+2})_{s,s}.$$

Roughly speaking, b_k measures the rate at which the probability of remaining in sector s decreases over time. To characterize the bias of the canonical model, we assume a single-crossing condition on b_k.

Assumption 3. *There exists* $\bar{k} \in \mathbb{N}$ *such that* b_k *is higher in the canonical model if and only if* $k \leq \bar{k}$.

This assumption requires that the probability of remaining in sector s initially decreases faster in the canonical model, but eventually decreases faster in the heterogeneous-worker model. Note that both models give the same value of the staying probabilities $(\mathcal{F}_k)_{s,s}$ for $k = 0, 1$, and the heterogeneous-worker model yields higher staying probabilities for all $k \geq 2$. Thus, staying probabilities must decrease faster in the canonical model for early periods. On the other hand, if staying probabilities converge to similar levels in both models as $k \to \infty$, then the decline should eventually become faster in the heterogeneous-worker model in order to compensate for the initial faster decline. Assumption 3 further requires the existence of a cutoff $\bar{k}$ at which the order of the speed of decline is reversed. In Appendix A.1.5, we show that this assumption indeed holds with $\bar{k} = 9$ (years) for the worker flow matrix series we observe in the data. Under this assumption, the following proposition shows that the canonical model initially overestimates the change in employment in sector s while underestimating the long-run labor reallocation.

Proposition 4. *Consider a uniform permanent shock of the form* (14) *known to agents in period* 1. *Under Assumption 3, there exists* $\bar{t} \in \mathbb{N} \cup \{\infty\} \backslash \{1\}$ *such that the canonical model, calibrated by matching the one-period worker flow matrix, overestimates the change in employment of sector* s *in period* t *if and only if* $1 < t \leq \bar{t}$.

The result implies that whether assuming away persistent heterogeneity leads to overestimation or underestimation of the labor reallocation depends on the time horizon. On the one hand, as discussed in Lemma 2, the canonical model overestimates the mobility of workers across sectors, leading to an overestimation of the change in employment of sector s. This intuition is what Proposition 4 describes when t is small. On the other hand, in the heterogeneous-worker model, once workers choose sector s, they have relatively higher probabilities of being stuck in that sector. Thus, in the face of a permanent negative (positive, respectively) shock, workers will dislike (like, respectively) sector s more relative to the canonical model. This aspect works in the opposite direction to our previous intuition and may become dominant when t is large enough.[28] In Appendix A.1, we show that the worker flow matrix series we observe in the data implies $\bar{t} = 11$ years. Thus, we can conclude that the canonical model overestimates the impact of shocks on sectoral employment within an 11-year horizon but underestimates their longer-run effects.

Until now, we have considered exogenous changes in sectoral wages. This scenario corresponds, for example, to a small open economy with linear technology affected by changes in world prices induced by trade liberalization. However, if wage changes are endogenously determined by labor market equilibrium, different models may also generate different predictions for wage changes in response to given exogenous shocks to the labor market. Interestingly, with endogenously determined wages, the underestimation of the welfare effect characterized in Proposition 3 is likely to be compounded by the overestimation of the speed of labor reallocation shown in Proposition 4. To see this, without loss of generality, consider a negative shock to a sector. Proposition 4 implies that in response to given negative wage changes, workers leave the sector more rapidly in the canonical model. The resulting decrease in labor

[28] This is not always the case. In particular, there exists $\bar{\beta} \in (0, 1)$ such that when $\beta > \bar{\beta}$, the canonical model overestimates the change in employment in sector s in all periods t.

supply raises the marginal productivity of labor in that sector, which partially offsets the initial decline in wages. Thus, the canonical model predicts a smaller decline in wages, at least in the short term. This, combined with discounting of the future, further contributes to underestimation of the welfare effect.

In sum, we show that the canonical model, without persistent worker heterogeneity, always underestimates the welfare losses of adversely affected workers and overestimates the speed of labor reallocation. In our counterfactual exercises in Section 1.5, we indeed document sizable differences in welfare effects and labor reallocation with and without persistent worker heterogeneity.

1.4 Sufficient Statistics in the Data

The sufficient statistics result in Proposition 2 requires three inputs to construct counterfactuals for a given shock of interest: the worker flow matrix series, values of the parameters ρ and β, and knowledge of the labor demand side. In this section, we first use longitudinal information in the NLSY data to compute the aggregate worker flow matrices and compare them with those implied by the canonical model without persistent worker heterogeneity. We then present a method for estimating the value of ρ without specifying worker heterogeneity, which extends the standard Euler-equation approach used in the literature. Finally, we impose $\beta = 0.96$ for the subsequent analysis. In Section 1.5, we close the model by specifying details on the labor demand side.

1.4.1 Observed Worker Flow Matrices

We compute worker flow matrices from the National Longitudinal Survey of Youth 1979, a rich dataset compiled by the US Bureau of Labor Statistics.[29] This survey follows a nationally representative sample of workers from 1979 onward annually through 1994 and biennially thereafter. The sample consists of workers who were between 14 and 21 years old as of December 31, 1978, and entered the labor market in the 1980s. The NLSY79 provides detailed information on education, race, gender, age, and, importantly, the sector of employment. Specifically, we identify a worker's sector of employment in a year as the sector in which the worker was employed in the first week of that year.[30] We mainly follow the data-cleaning procedure of Lise and Postel-Vinay (2020).[31] We consider worker flows across four broad sectors: (i) Agriculture and Construction; (ii) Manufacturing; (iii) Communication and Trade; and (iv) Services and Others.

[29] We also obtain quantitatively and qualitatively similar findings using the monthly Current Population Survey dataset; see Section 1.5.

[30] This ensures that we consistently measure mobility at a 1-year window.

[31] The survey comprises a cross-sectional subsample representative of young people living in the US and other subsamples that target ethnic minorities, people in the military, and the poor. We only use the representative subsample for our analysis. We also drop people seen in the military.

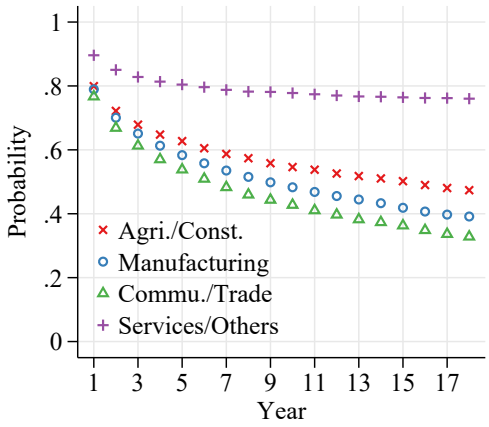


Figure 2. Worker Flow Matrix Series

Notes: Each marker in the figure represents the probability that workers choose the same sector after $k \in \{1, 2, \ldots, 18\}$ years, $\Pr(s_{t+k} = s | s_t = s)$. There are four sectors: Agriculture and Construction, Manufacturing, Communication and Trade, and Services and Others. Data source: NLSY79.

Assuming that the economy was in a steady state between 1980 and 2000, we calculate the series of worker flow matrices by pooling all observations in this period. Given the data constraints, we only calculate the k-year worker flow matrices, $\mathcal{F}_k$, up to $k = 18$. Figure 2 plots the diagonal elements of the obtained worker flow matrices. Each point represents the probability of workers choosing the same sector after k years (i.e., $\Pr(s_{t+k} = s | s_t = s)$; hereafter, *$k$-year staying probability*). 1-year staying probabilities are close to 80%, except in the services sector, in which it is just below 90%. k-period staying probabilities decrease in k, which reflects the diminishing impact of being in a particular sector in the past over time.

To apply sufficient statistics results, however, we need a full sequence of worker flow matrices from k equals one to infinity. Thus, we need a method to extrapolate longer-run worker flow matrices from the available finite-length data. Leveraging the structural model provides a natural method for extrapolation.[32] For this purpose, we estimate the structural model by matching the worker flow matrices we computed directly from the NLSY data, $\{\mathcal{F}_k\}_{k=1}^{18}$. See Appendix A.2.1 for details. The estimated structural model generates the full set of worker flow matrices, which are used in Section 1.5 to perform counterfactual exercises.[33] Since the structural model reduces to the canonical model when there is only one worker type, this extrapolation is a strict generalization of the canonical model's extrapolation.

[32] This approach has the advantage that extrapolation is disciplined by the model. However, the sufficient statistics approach does not, in principle, require that we estimate the details of the model. In Appendix A.2.3, we explore alternative extrapolation methods that do not rely on a structural model. In addition to extrapolation, there are two other benefits of the structural estimation. First, it serves as a proof of concept: By seeing whether our structural model can match the observed worker flow matrix series, we can test whether our framework is consistent with the data. Second, we can use the estimated structural model to evaluate the quality of the first-order approximation around a steady state. In Section 1.5, we compare the results of counterfactual exercises computed from the exact solution of the estimated model and those computed using the sufficient statistics result.

[33] In fact, we can simply treat the estimated model as if it were the true model and use it to perform counterfactual exercises directly. The sufficient statistics result guarantees that we will obtain correct counterfactual results regardless of whether this model is true.

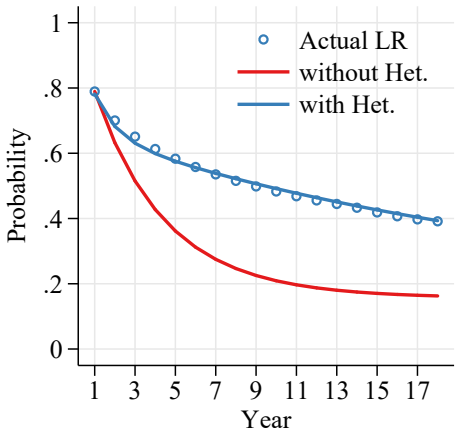


Figure 3. Actual and Model-implied Worker Flow Matrices: Manufacturing Staying Prob.

Notes: Blue dots in the figure represent k-year manufacturing staying probabilities. The blue solid line represents the fit of the estimated two-type worker model. The red solid line represents staying probabilities implied by the canonical model. Data source: NLSY79.

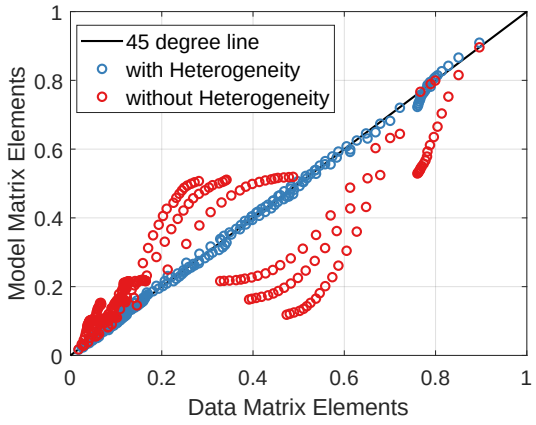


Figure 4. Actual and Model-implied Worker Flow Matrices: All 4×4 Elements

Notes: Blue dots in the figure plot elements of the worker flow matrix series implied by the estimated two-type worker model against those in the data. Red dots correspond to the canonical model. Data source: NLSY79.

Surprisingly, the model with only two worker types (i.e., $|\Omega| = 2$) closely matches the observed worker flow matrices. Figures 3 and 4 document the fit of the estimated model. The blue solid line in Figure 3 represents the model-implied k-year staying probabilities for the manufacturing sector, which is almost indistinguishable from the actual staying probabilities in the data (blue dots). In Figure 4, we use blue dots to plot all 4×4 elements of the model-implied worker flow matrix series against the actual values in the data. Most of the dots lie roughly on the 45-degree line.

1.4.2 Bias of the Canonical Model

To quantify the bias of the canonical model characterized in Lemma 2, we calculate the worker flow matrix series implied by the canonical model. Following Lemma 2, we compute the implied k-year worker flow matrix by multiplying the 1-year matrix k times. We first compare a specific diagonal element of the actual and the implied worker flow matrices: the k-year staying probability for the manufacturing sector. This is the element of primary interest because our main counterfactual exercise in Section 1.5 examines the impact of the China shock on US manufacturing sectors. Figure 3 plots both the actual staying probabilities (blue dots) and those implied by the canonical model (red line). It clearly shows that the canonical model significantly underestimates longer-run staying probabilities, which is in line with the prediction of Lemma 2 and becomes particularly pronounced at longer

horizons.[34,35] Figure A.5 further shows that this underestimation is not driven by the nonstationary nature of the data. The result remains qualitatively and quantitatively similar even when nonstationarity is taken into account.[36]

As we have seen in Section 1.3, this discrepancy arises because the likelihood of choosing the manufacturing sector is higher for workers who have previously chosen the manufacturing sector. Workers who have self-selected to stay in manufacturing exhibit a higher probability of staying again in the following year, perhaps due to particularly high switching costs. Similarly, workers who self-selected into manufacturing in the past are more likely to choose it again, possibly owing to their comparative advantage.

The canonical model also underestimates the diagonal elements of the worker flow matrices that correspond to the non-manufacturing sectors. In Figure 4, we plot all 4×4 elements of the worker flow matrix series implied by the canonical model against the actual values in the data (red dots). The points clustered below the 45-degree line correspond to the diagonal elements of the worker flow matrices, which are underestimated by the canonical model. To compensate for this underestimation, the off-diagonal elements are overestimated, as seen in other points clustered above the 45-degree line. Section 1.5 quantifies how this inconsistency translates into systematic biases in counterfactual predictions of the effects of sectoral shocks.

1.4.3 Understanding the Bias of the Canonical Model

Demographic and Socioeconomic Characteristics. Where does the bias of the canonical model come from? One possibility is that worker heterogeneity in terms of observable characteristics could explain most of the bias. If so, we can easily correct for the bias by simply conditioning on these characteristics, obviating the need for our sufficient statistics approach. However, we will demonstrate that this is not at all the case. The literature has discussed various types of demographic and socioeconomic characteristics of workers. Here, we focus on three dimensions—gender, race, and education—that have been found to be important determinants of sectoral choices and welfare outcomes (e.g., Dix-Carneiro, 2014; Lee and Wolpin, 2006). We divided people into male and female, Hispanic/Black and

[34] Howard and Shao (2023) document a similar pattern for dynamic location choices. They explain this finding with a model in which dynamics arise from spatially and persistently auto-correlated preferences rather than from moving costs.

[35] One concern is that this bias may mainly, or at least partly, reflect misclassification errors in the sector choice data. Dvorkin (2023) studies the bias resulting from misclassification errors in the industry (and occupation) information in the PSID and March CPS datasets. Such misclassification errors lead to overestimation of mobility, as in this paper. However, it is unlikely that our results are mainly due to these errors. First, the NLSY data and the monthly CPS (observations within four consecutive months) used in this paper are known to be less prone to such errors (Moscarini and Thomsson, 2007). Second, a similar bias has also been documented for dynamic location choices (Howard and Shao, 2023), which is difficult to attribute to misclassification errors.

[36] In fact, the stationarity assumption is likely to lead to an *underestimation* of the gap between the actual and implied staying probabilities. To see this, suppose that worker flow matrices for two periods $t = 1, 2$ are F^1 and F^2, respectively. Under the stationarity assumption, we calculate the worker flow matrix $\bar{F}$, which is applied to both periods, by taking an average of F^1 and F^2. Suppose that we put equal weights on F_1 and F_2. Then, we have

$$(F^1F^2)_{ii} \approx (F^1)_{ii}(F^2)_{ii} \leq \left(\frac{(F^1)_{ii} + (F^2)_{ii}}{2}\right)^2 = (\bar{F})_{ii}^2 \approx (\bar{F}^2)_{ii}.$$

Thus, the implied two-period staying probability is overestimated under the stationarity assumption, leading to a *smaller* gap between the actual and implied staying probabilities.

non-Hispanic/Black, and low-skilled (less than high school and high school) and high-skilled (some college and college or more). Unique combinations of these three dimensions of heterogeneity define eight worker types. In Figure A.6, we plot the actual manufacturing staying probabilities separately for the eight types. Indeed, workers who differ along these characteristics exhibit highly distinct sectoral movement patterns. For example, low-skilled non-Hispanic/Black males are more than twice as likely to stay in the manufacturing sector in the longer run than high-skilled Hispanic/Black females. However, Figure 5 shows that these characteristics do not explain the gap observed in Figure 3. We plot the manufacturing staying probabilities implied by the model that incorporates these three dimensions of observed characteristics. Specifically, we consider a model with the eight observed worker types. For each worker type ω^{obs}, we can calculate the 1-year transition matrix $(F^{\omega^{\text{obs}}})$ directly from the data. Since workers are assumed to be homogeneous within each of the eight types, their k-year transition matrix can then be computed by $(F^{\omega^{\text{obs}}})^k$. Thus, the model-implied aggregate k-year worker flow matrix is obtained by taking averages of these type-specific k-year transition matrices using the steady-state type composition as the weight. The green solid line shows the result and is almost indistinguishable from the red line, which plots the staying probabilities implied by the canonical model. These types of observed characteristics provide only a minor improvement in the model's ability to explain the actual pattern of worker flows. If we also incorporate age in the model, there is slight improvement in the fit, but the implied worker flow matrix series still significantly differs from the actual one (purple solid line).[37] In sum, workers with different demographic and socioeconomic characteristics do indeed behave differently, but this fact barely changes *aggregate* labor market dynamics.[38] In turn, the sufficient statistics result implies that adding these worker characteristics to a model does not cause substantial changes in the results of the counterfactual analysis.[39]

Pure Duration Dependence. The limited role of demographic and socioeconomic characteristics implies two possibilities: Either these characteristics are simply poor proxies for underlying worker heterogeneity or the gap in Figure 3 arises from mechanisms other than selection on worker heterogeneity, such as pure duration dependence.[40] A notable example is the accumulation of sector-specific human capital, whereby otherwise identical workers who have spent more time in manufacturing may have accumulated more manufacturing-specific human capital, which renders them more likely to choose the sector again. Other examples include fixed adjustment costs (e.g., Stokey, 2008); learning about match productivity (e.g., Jovanovic, 1979); and psychological choice models (e.g., Cain, 1976). This raises concern, given that only the first equation of Proposition 1 extends to the case with a duration dependence mechanism.

[37] Age is a form of time-varying heterogeneity. To analyze this within the context of our persistent heterogeneity framework, we categorize age into two groups, "old" and "young," so that each represents about 50% of the total sample.

[38] Similar results can be found in various fields of economics. For example, Card, Rothstein, and Yi (2023) find that only a modest fraction of the variation in average wages across commuting zones is explained by differences in the observed characteristics of workers.

[39] This means that the discrepancy in Figure 3 is mostly driven by within-type heterogeneity instead of across-type heterogeneity. We can observe similar gaps between actual and predicted longer-run staying probabilities for each of the eight types. In particular, the gaps are more pronounced for female Hispanic/Black workers.

[40] Distinguishing dynamic selection based on heterogeneity from duration dependence mechanisms is a recurring theme in various fields of economics; see Heckman (1981) for an important early contribution along these lines. Alvarez, Borovičková, and Shimer (2016) also study how to distinguish between the two in the context of unemployment duration.

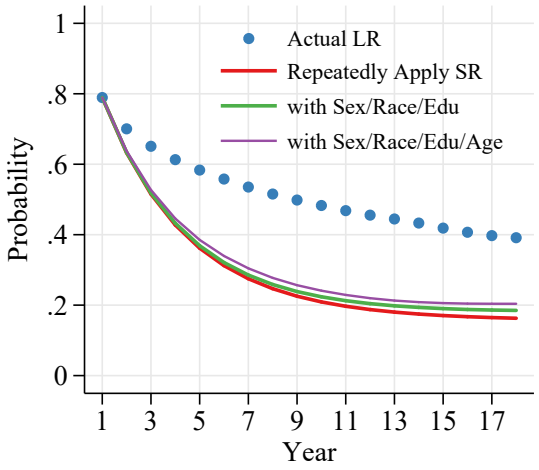


Figure 5. Actual and Model-implied Staying Probabilities, with and without Socio. Char.

Notes: Blue dots in the figure represent k-period manufacturing staying probabilities. The red solid line represents the fit of the estimated canonical model. The green line incorporates three observed characteristics. The purple line also incorporates age. Data source: NLSY79.

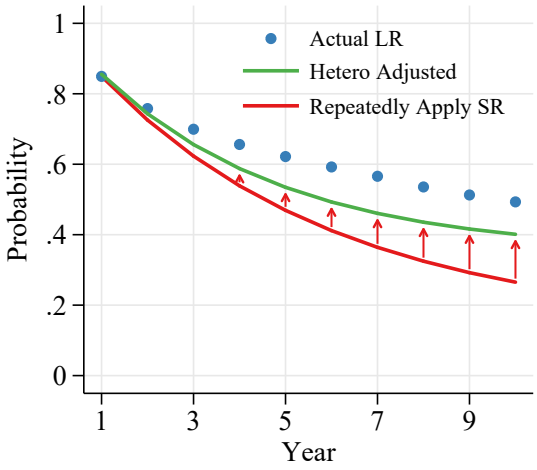


Figure 6. Actual and Model-implied Staying Probabilities, Heterogeneity Partially Controlled

Notes: Blue dots in the figure represent k-period manufacturing staying probabilities calculated using post-1990 data. Green and red lines represent the staying probabilities implied by the canonical model and the model with five worker types, respectively. Data source: NLSY79.

In response to this issue, we provide suggestive evidence that points to the importance of worker heterogeneity in generating the gap in Figure 3. Our strategy involves constructing an alternative proxy for worker heterogeneity. Instead of relying on demographic and socioeconomic characteristics, we leverage workers' sector choice histories prior to 1990 as a means to capture their heterogeneity. Differences among workers materialize as differences in their sector choice patterns, so their histories allow us to effectively control for their heterogeneity. Specifically, we use pre-1990 data to compute the 1-year manufacturing staying probability for each worker, then categorize workers into five types based on the quintiles of this probability. We then assess the extent to which accounting for these worker types narrows the gap between actual and model-implied staying probabilities. In Figure 6, the actual k-year staying probabilities calculated using post-1990 data are denoted by blue dots, while red and green lines represent the staying probabilities implied by the canonical model and the model with the five worker types, respectively. Note that the green line is computed under the assumption that workers are homogeneous within each of the five worker types and that there is no duration dependence mechanism in play. The green line is close to the blue dots, and the gap between the actual and implied staying probabilities is reduced by more than half. Given that worker heterogeneity is only partially controlled for by sector choice history, this result suggests that at least half of the gap is due to worker heterogeneity—the mechanism emphasized in this paper.

1.4.4 Estimation of ρ

In this section, we present a novel strategy for estimating the parameter ρ, which, in conjunction with the extrapolated worker flow matrix series, provides a complete description of the labor supply side of the model. Our goal is to propose an estimation method that does not require explicitly specifying the underlying heterogeneity.[41] The possibility of such an estimation is already suggested in the second equation of Proposition 1, which we restate here for convenience:

$$\mathrm{d}\ln \ell_t = \sum_{s\geq 0, k\geq 0} \frac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathbb{E}_{t-s-1}\,\mathrm{d}w_{t-s+k}. \tag{15}$$

This equation describes the response of sectoral employment to wage shocks, with the coefficients depending only on the value of ρ and the worker flow matrix series, independent of the specific details of worker heterogeneity. Thus, we can estimate ρ by measuring the responsiveness of sectoral employment to sectoral wage shocks, conditional on the observed worker flow matrix series.[42]

Implementation. In principle, equation (15) could be directly confronted with data to estimate ρ, but this requires that we fully specify how much information workers have about the future. The literature circumvents this demanding requirement by applying the Euler equation approach first used in ACM (e.g., Artuç and McLaren, 2015; Caliendo, Dvorkin, and Parro, 2019; Traiberman, 2019). We extend the Euler equation approach by allowing for arbitrary worker heterogeneity.[43]

The key idea of the Euler equation approach is to transform (15) into a recursive representation. For example, in the absence of persistent worker heterogeneity, we can use the restriction $\mathcal{F}_k = (\mathcal{F}_1)^k$ to rewrite equation (15) recursively as follows:[44]

$$\mathrm{d}\ln \ell_t = (\mathcal{F}_1^{-1} + \beta\mathcal{F}_1)\,\mathrm{d}\ln \ell_{t+1} - \beta\,\mathbb{E}_t\,\mathrm{d}\ln \ell_{t+2} - \frac{\beta}{\rho}(\mathcal{F}_1^{-1} - \mathcal{F}_1)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}. \tag{16}$$

[41] In Propositions 1 and 2, we showed that the worker flow matrix series serves as sufficient statistics for counterfactual exercises *when we know the value of the parameter* ρ. However, if the estimation procedure for ρ depends on how we specify worker heterogeneity, the distribution of worker heterogeneity can affect the results of counterfactual exercises through its effect on estimation of ρ. In this sense, our estimation method is comparable to estimation of the trade elasticity using the gravity equation of Arkolakis, Costinot, and Rodríguez-Clare (2012). Just as their estimation based on the gravity equation provides a valid estimate within a class of "quantitative trade models," our method yields a valid estimate within a specific class of dynamic discrete choice models.

[42] The responsiveness is inversely related to the value of ρ. A higher value of ρ indicates that sector choice decisions are primarily driven by idiosyncratic shocks. As a consequence, the impact of wage shocks on sectoral employment is relatively diminished when ρ is high.

[43] ACM (p.1040) write that "Perhaps the biggest weakness in the Euler-equation approach ... is that it assumes away workers with unobserved [heterogeneity] .... A full exploration of such effects probably requires a structural micro approach." However, the method described in this section demonstrates that up to the first-order approximation, this problem can be circumvented.

[44] Unlike equation (15), equation (16) contains only the term $\mathbb{E}_t\,\mathrm{d}w_{t+1}$ since the (expected) values of $\mathrm{d}\ln \ell_{t+1}$ and $\mathrm{d}\ln \ell_{t+2}$ summarize the effect of all other beliefs. Because of this advantage, the Euler equation approach has been widely used in the literature, although previous studies have used equations that involve migration probabilities rather than labor supply.

See Appendix A.2.2 for proofs of the results in this section. However, with arbitrary worker heterogeneity, this is impossible because the worker flow matrix series that determines the coefficients of equation (15) is based on empirical data and does not have a recursive structure. Nevertheless, we will demonstrate in two steps that it is still possible to derive an *approximate* recursive representation of equation (15). First, in Appendix A.2.2, we show that when there is a finite number of worker types, N, equation (15) always possesses an *exact* recursive representation of the form[45]

$$\mathrm{d}\ln \ell_t = \sum_{k=1}^{4N-2} \Gamma_k \, \mathbb{E}_t \, \mathrm{d}\ln \ell_{t+k} + \frac{\beta}{\rho} \sum_{k=1}^{4N-3} \Lambda_k \, \mathbb{E}_t \, \mathrm{d}w_{t+k} \tag{17}$$

where Γ_k and Λ_k are functions of worker flow matrix series, $\{\mathcal{F}_k\}$. Second, recall that a model with two worker types provides a close approximation to the observed worker flow matrix series $\{\mathcal{F}_k\}$ in the data. Combining these two findings, we can conclude that equation (15), with $\{\mathcal{F}_k\}$ from the data, can be approximately represented in the recursive form (17) with $N = 2$. In Appendix A.2.2, we provide formulas for computing $\{\Gamma_k\}_{k=1,\ldots,6}$ and $\{\Lambda_k\}_{k=1,\ldots,5}$. In Figure A.9, we use simulation to demonstrate that the obtained approximate recursive representation provides a close fit to the actual dynamics of sectoral employment.

We further modify the obtained recursive representation in two ways:

$$\ln \ell_t - \sum_{k=1}^{6} \Gamma_k \ln \ell_{t+k} = \frac{\beta}{\rho} \sum_{k=1}^{5} \Lambda_k w_{t+k} + \mathsf{ExpectationError}_{t+1,t+6} \, . \tag{18}$$

First, instead of deviations from steady-state values, $\mathrm{d}\ln \ell$ and $\mathrm{d}w$, we use the actual values, $\ln \ell$ and w. This is possible because the recursive representation always satisfies $\sum_{k=1}^{6} \Gamma_k = I$ and $\sum_{k=1}^{5} \Lambda_k = O$, where I is the identity matrix and O is the zero matrix. Second, the expected values are substituted with the realized values plus an expectation error term that depends on the news revealed between time $t+1$ and $t+6$.

Equation (18) is our regression specification, where we regress the left-hand-side variable on the explanatory variable on the right-hand side, $\sum_{k=1}^{5} \Lambda_k \, \mathrm{d}w_{t+k}$, to estimate $\frac{\beta}{\rho}$. However, since both the explanatory variable and the expectation error term are affected by newly revealed information between period $t+1$ and $t+6$, they are likely to be correlated. To address this concern, we follow ACM and use past values of sectoral labor allocation and wages as instruments. Any variables included in the period t information set are theoretically valid instruments for the explanatory variable, providing a consistent estimate of ρ. In particular, we use the 1-year lag of sectoral wages and sectoral employment shares. For this approach to be valid, it is necessary to assume that workers have rational expectations and that the error term in equation (18) only reflects errors in workers' forecasts. If this term also incorporates shocks to the labor supply curve (e.g., unexpected aggregate shifts in preferences for particular sectors), we must assume an exclusion restriction whereby such shocks are uncorrelated with our instruments. For a discussion

[45] For further insights into how this result relates to the findings of Granger and Morris (1976), see Appendix A.2.2.

Table 1: Estimation of ρ

	(1)	(2)	(3)
β/ρ	-0.286	1.164**	0.877***
	(0.311)	(0.550)	(0.296)
Implied ρ	-3.358	0.825**	1.095***
	(3.652)	(0.390)	(0.369)
Method	OLS	IV	IV
Persistent Heterogeneity	✓	✓	
No Persistent Heterogeneity			✓
Observations	136	136	136
First-stage F	–	38.7	11.1

Notes: OLS and IV estimation results for specification (18) (heterogeneous workers: Columns (1) and (2)) and (16) (homogeneous workers: Column (3)). Standard errors robust to heteroskedasticity are reported in parentheses, with *** : $p < 0.01$. Data source is NLSY79 and BLS.

of the strengths and weaknesses of this approach in the case of homogeneous workers, refer to ACM and Traiberman (2019). Also, we include sector fixed effects in the regression to isolate within-sector variation from across-sector variation.

Our estimation method requires data on sectoral employment and sectoral wages. We use the Bureau of Labor Statistics' Current Employment Surveys (CES) to compute the time series of these variables. We use the share of workers in the dataset employed in each sector as our measure of sectoral labor allocation and average wages in each sector as our measure of sectoral wages. We again consider four broad sectors.[46] In Section 1.5, we will compare the counterfactual predictions of our model with those of Caliendo, Dvorkin, and Parro (2019), who analyze a homogeneous counterpart of our model. To facilitate clear comparison, we assume that the variable w represents the logarithm of the sectoral wage.

Table 1 presents the estimation results. Column (1) estimates equation (18) by OLS. The estimated coefficient and the implied value of ρ are negative and insignificant. In Column (2), we estimate the same specification using IV. The implied value of ρ is 0.825. This means that a one standard deviation higher realization of the idiosyncratic shock is associated with 4.06% higher lifetime consumption (see Appendix A.2.4). Comparing the results in Columns (1) and (2), it appears that the estimated coefficient $\widehat{\beta/\rho}$ from OLS is biased downward due to the presence of expectation errors. This is consistent with the fact that expectation errors are likely to cause sectoral labor supply and sectoral wages to be negatively correlated.[47] The estimate in Column (2) implies that a temporary 1% decline

[46] The assumption of a four-sector model raises an econometric issue similar to the one discussed by ACM. If the true model consists of more than four sectors with a dispersion parameter ρ, is it valid to approximate the model with a four-sector model and estimate the dispersion parameter based on this approximation? More importantly, can we use it to conduct counterfactual exercises? We can demonstrate that under certain assumptions on switching costs, the validity of using the approximated model is supported by the fact that the maximum of i.i.d. type I extreme-value distributed random variables follows another type I extreme-value distribution with the same scale parameter.

[47] If workers wrongly expect an increase in wages in a sector, they would supply more labor to that sector. This increased labor supply would lead to a decrease in wages in that sector.

in manufacturing wages leads to a 0.35% decrease in the manufacturing share, while a permanent 1% decline in manufacturing wages is associated with a 1.15% decrease in the manufacturing share. In column (3), we assume that workers are homogeneous and estimate equation (16) by IV with the same set of instrument variables. The implied value of ρ is slightly higher than our preferred estimate, but we cannot reject the null of equality between the two. Estimates in Columns (2) and (3) are broadly consistent with the estimates of ρ in the literature, which range from 0.5 to 2. The original ACM and subsequent papers estimate ρ to be around two. A more recent paper, Artuç and McLaren (2015), suggests a value of $\rho = 0.62$. Also, Rodriguez-Clare, Ulate, and Vasquez (2022) obtain a similar value of $\rho = 0.56$.[48] This estimate in Column (2) is our preferred specification, and we will use it in the subsequent analysis.

We now move on to applications of our model, where we use our empirical estimates to quantify the effect of sectoral shocks.

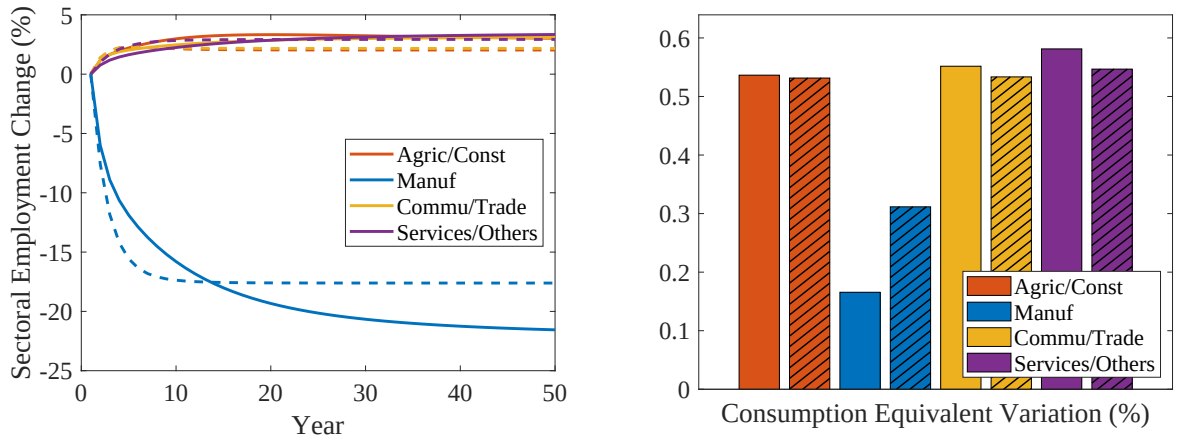
1.5 Applications

In this section, we consider two sectoral shocks and quantify the implication of persistent worker heterogeneity for welfare and labor reallocation. We first apply our results to a stylized trade liberalization exercise of ACM, which illustrates how the framework of the literature can easily be extended to allow for worker heterogeneity. For a more realistic application, we then examine the dynamic effects of the rise in China's import competition on US labor markets. We revisit this extensively studied topic using the sufficient statistics approach. Although we focus on these two exercises, our result can be applied more generally to other papers in the literature.

1.5.1 Permanent Decline in Manufacturing Prices

The first counterfactual exercise closely follows ACM and considers an unanticipated permanent 10% drop in manufacturing prices for a small open economy—for example, due to trade liberalization. Following Section 1.2, we incorporate persistent worker heterogeneity in the labor supply side of the model, which is calibrated in Section 1.4. We specify the labor demand side of the model as in ACM. We assume log utility with Cobb-Douglas consumption aggregate and a sector-specific CES production function with fixed capital stock. All goods are traded and their prices are exogenously given at world price level. Sectoral wages are competitively determined by the marginal productivity of labor. The labor demand side of the model is calibrated exactly as in ACM; see Appendix A.3.1 for

[48] Note that ACM and other papers in the literature assume instantaneous utility linear in wage. This implies that the value of ρ governs semi-elasticity $\frac{\partial \ln \ell_{t+k}}{\partial \text{wage}_t}$ instead of elasticity $\frac{\partial \ln \ell_{t+k}}{\partial \ln \text{wage}_t}$. However, because they normalize sectoral wages so that the average annualized wage equals unity, both the semi-elasticity and elasticity can be interpreted as the percentage change in sectoral employment in response to a percentage change in sectoral wages.



(a) Changes in Sectoral Employment Share

(b) Changes in Sectoral Values

Figure 7. Counterfactual Changes in Sectoral Employment and Welfare: Trade Liberalization

Notes: This figure plots the transitional dynamics following an unexpected permanent drop in manufacturing prices. Solid lines correspond to the prediction from the sufficient statistics in the data, and dashed lines correspond to the prediction of the canonical model without persistent worker heterogeneity.

details. Initially, the economy is in a steady state. Since wages are endogenously determined by the labor market equilibrium, we apply Proposition 2 to compute the perfect-foresight transition path following announcement of the shock in year 1 until the economy reaches a new steady state.

Figure 7 shows results of the counterfactual exercise. In Figure 7a, we plot the dynamics of sectoral employment predicted by the models with and without worker heterogeneity. The manufacturing employment share drops sharply, by around 20%. Importantly, the canonical model overestimates the short-term labor reallocation, resulting in a much faster transition to the new steady state. The transition is completed within 10 years in the absence of worker heterogeneity, while it takes more than 50 years with worker heterogeneity. At the same time, the canonical model underestimates the magnitude of long-term reallocation. All of these results are consistent with the predictions of Proposition 4.[49]

In Figure 7b, we plot changes in welfare, measured in terms of consumption-equivalent variation; that is, the proportional change in the lifetime consumption sequence that would have the same effect on household welfare as would the welfare effect of the China shock. In particular, solid bars represent the welfare changes of workers based on their initial sector of employment predicted by the model. Hatched bars represent welfare changes implied by the canonical model. The result shows that even workers who were initially employed in manufacturing—the import-competing sector—benefit from trade liberalization. The increase in option value, driven by the increase in

[49] In Appendix A.3.1, we compute the impulse responses of sectoral employment from the observed worker flow matrix series. As predicted by Proposition A.4, the canonical model tends to overestimate the short-term impact of shocks on sectoral employment but underestimates their long-term effects. Notably, within a 7-year time horizon, the canonical model consistently overestimates the effects of shocks on sectoral employment. This explains the difference between changes in sectoral employment in the models with and without worker heterogeneity documented in Figure 7a.

real wages in other sectors, more than compensates for the decline in manufacturing wages.[50] However, consistent with the prediction of Proposition 3, the canonical model overestimates the gains of manufacturing workers by a factor of almost two. At the same time, it slightly underestimates the gains of non-manufacturing sector workers, resulting in substantial underestimation of the distributional consequences of trade liberalization.

As discussed in Section 1.3, the welfare gains of manufacturing workers are overestimated in the canonical model for two reasons. Not only does it overestimate welfare gains for given wage changes, but it also predicts a smaller decline in manufacturing wages in the short term. To illustrate this, we show in Figure A.11 that if we used the changes in sectoral wages computed from the canonical model—instead of endogenizing them—and combined them with the worker flow matrix series as in Proposition 1, we would get a narrower gap between the welfare changes predicted by models with and without worker heterogeneity.

To gain deeper understanding of the disparities between models with and without heterogeneity, we plot changes in sectoral values and employment shares separately for each of the two worker types in Figure A.12.[51] When the shock hits the economy, type 1 workers, who have lower switching costs and comparative advantage in non-manufacturing sectors, can easily move out of manufacturing and enjoy a higher welfare gain from the shock. Over time, as more type 2 workers leave the manufacturing sector, manufacturing wages begin to recover. Given that type 2 workers are more likely to be stuck in the manufacturing sector once they choose it, they dislike manufacturing more than type 1 workers, resulting in a higher proportion of type 1 workers in the manufacturing sector in the long run.

Finally, we assess the quality of the first-order approximation around a steady state. In Figure A.13, we compare the results of the counterfactual exercises obtained using sufficient statistics formulas with those calculated from the exact solution of the estimated structural model.[52] Despite the relatively large magnitude of the 10% shock considered in this section, the sufficient statistics results deliver a close approximation. In particular, the approximation error is an order of magnitude smaller than the difference between the counterfactual predictions of the models with and without worker heterogeneity. The approximation error becomes almost negligible for a shock size of 1%, but becomes more pronounced as the shock size increases to 30%.

[50] Figure A.10 plots the time path of changes in sectoral wages. The real wage of the manufacturing sector initially overshoots because it takes time for labor to adjust. As the number of workers in manufacturing gradually declines, manufacturing wages rise and eventually exceed the preshock steady-state level. However, they increase less than those in other sectors. Thus, the manufacturing employment share declines over time and in the new steady state.

[51] An important caveat to this interpretation pertains to the identification of the structural model. In estimating the structural model, many configurations of the model primitives are almost equally successful in matching the worker flow matrix series. This means that even small changes in the observed matrices can lead to significantly different estimation results. Thus, statements made at the unobserved type level should be viewed with caution, since they may be far from the real world.

[52] Figure A.13 also demonstrates that the exact value changes always exceed those calculated using a first-order approximation. We can show analytically that the second-order term is always positive due to the option value.

1.5.2 The China Shock

As a second application, we apply our sufficient statistics result to a more realistic counterfactual exercise: the dynamic impact of the growth of China's manufacturing productivity and the resulting import competition on the welfare of US workers and labor reallocation. We closely follow the dynamic quantitative trade model of Caliendo, Dvorkin, and Parro (2019) (hereafter, CDP) and extend the labor supply side by allowing for arbitrary time-invariant worker heterogeneity.[53] We make two simplifying assumptions relative to the original specification of CDP. First, we consider four broad sectors—Manufacturing, Wholesale/Retail, Construction, and Services—and another auxiliary sector representing nonemployment.[54] Second, for reasons discussed shortly, we abstract from interstate migration and assume that workers can switch sectors only within each state.[55,56] We first describe the labor supply side, then specify the labor demand side and the shock of interest to close the model.

Labor Supply Side. Two observations motivate us to conduct a separate analysis for each of the 50 US states. First, it is well known that there is considerable variation in exposure to the China shock across different geographic regions in the US (see, for example, Autor, Dorn, and Hanson, 2013a; Acemoglu et al., 2016). This suggests that Assumption 1 is better imposed within each state. Second, as shown in Figure A.16, worker flow matrices differ significantly across states; workers in states with a higher manufacturing employment share are more likely to remain in the manufacturing sector over time. By applying the sufficient statistics approach at state level, we can account for such state-level heterogeneity. Under our simplifying assumptions, we can focus on 5-by-5 intersectoral worker flow matrices for each state.

State-level analysis requires computation of a worker flow matrix series for each state. However, the limited sample size of the NLSY data makes it difficult to estimate them accurately. Thus, we instead use the monthly Current Population Survey (CPS) dataset, which contains a substantial sample size for each state. This dataset tracks workers for 4 consecutive months, which allows us to compute worker flow matrices $\mathcal{F}_k^{\text{state}}$ for $k = 1, 2, 3$ months for each state. Specifically, we assume that the US was in a steady state before the China shock and compute worker

[53] A large body of subsequent literature studies additional elements that are missing in this framework: involuntary unemployment from downward nominal wage rigidities or search frictions (Kim and Vogel, 2021; Rodriguez-Clare, Ulate, and Vasquez, 2022); endogenous trade imbalances (Dix-Carneiro et al., 2023); occupation adjustment (Traiberman, 2019); learning and expectations (Fan, Hong, and Parro, 2023); and currency pegs (Kim, de la Barrera, and Fukui, 2023). Incorporating worker heterogeneity in models with these additional features is an important direction for future research.

[54] In CDP, there are 23 sectors: 12 from Manufacturing, 8 from Services, and 1 each for Wholesale/Retail, Construction, and nonemployment.

[55] CDP uses an 1150-by-1150 quarterly worker flow matrix between all US state-sector pairs (50 states, excluding the District of Columbia and the territories, and 22 sectors plus 1 additional sector representing non-employment). However, we need to estimate the longer-run worker flow matrix as well as the short-run one, and estimating them at this level of granularity is practically impossible.

[56] In principle, allowing for regional migration would dampen the welfare effects because it provides an additional margin of adjustment. However, US workers change sectors almost 10 to 100 times more often than they change states, which suggests that the majority of the labor adjustment occurs at the sector change margin. In the same context of the China shock, Rodriguez-Clare, Ulate, and Vasquez (2022) also show that ignoring migration makes little difference for their model's prediction. Autor, Dorn, and Hanson (2013a) and Autor et al. (2014) also demonstrate that regional migration is not an important mechanism through which the economy adjusts to the China shock.

flows by pooling transition observations between January 1995 and December 1999.[57,58] To compute worker flow matrices for other values of k, we follow Section 1.4 and estimate the structural model with two types of workers by matching the observed worker flow matrices. We then use the estimated model to extrapolate longer-run worker flow matrix series.[59] In Figure A.18, we compare the fits of the models with and without worker heterogeneity to the observed worker flow matrix series. While the fits of the models vary across states, the canonical model consistently underestimates the staying probabilities. Finally, we use the value of the parameter ρ estimated in Section 1.4.[60]

Labor Demand Side and the China Shock. The labor demand side of the model is more complex than in Section 1.5.1: It features a large number of labor markets distinguished by sector and region, international trade, interregional trade within the US, input-output linkages, and multiple production inputs. Appendix A.3.2 describes the primitive assumptions of the model regarding households' consumption and sector choices; intermediate goods and final goods producing firms' profit maximization; and the definition of a sequential competitive equilibrium. We follow CDP in calibrating the structural parameters of the labor demand side; see Appendix A.3.2 for details of the calibration. The shock of interest is the growth of China's manufacturing productivity. Following CDP, we consider the China shock to be a sequence of shocks to the growth rate of total factor productivity (TFP) of the Chinese manufacturing sector from 2000 to 2007, assuming a constant fundamental thereafter. We also assume that US agents anticipated the China shock in 2000 exactly as it occurred. We calibrate manufacturing productivity growth such that the model's predicted increase in US imports from China exactly matches the predicted increase in imports, using the increase in imports from China of the other eight advanced economies as an instrument. See Appendix A.3.1 for detailed calibration of the China shock.

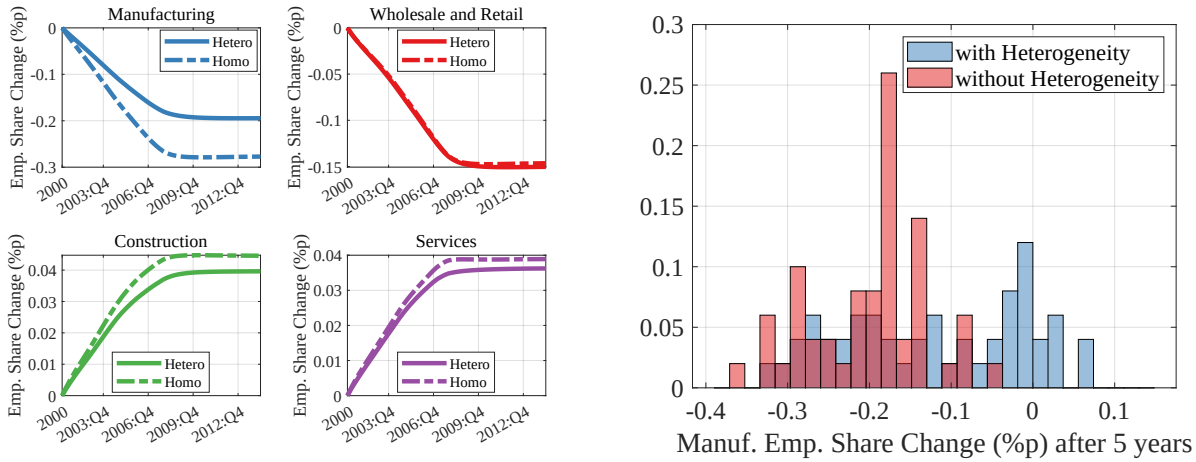
Counterfactual Results. For exogenous changes in the manufacturing productivity of China, US sectoral wages are endogenously determined by the labor market equilibrium. Thus, we again use Proposition 2 to calculate counterfactual changes in sectoral welfare and sectoral employment. Given the rich structure of the model, it is computationally demanding to estimate all exogenous state variables of the model—including productivities,

[57] In contrast to the literature, which often uses only one worker flow matrix—either monthly ($k = 1$) or quarterly ($k = 3$)—we exploit the information contained in all worker flow matrices to identify underlying persistent worker heterogeneity. For any k, if we use the k-period worker flow matrix $\mathcal{F}_k^{\text{state}}$ and rely on the homogeneous-worker assumption to calculate the remaining worker flow matrices, we would underestimate the diagonal elements of k'-period worker flow matrices for k' greater than k.

[58] The monthly CPS dataset also tracks workers again for 4 additional consecutive months after 8 months after the first 4 consecutive months. Thus, in principle, we can observe $\mathcal{F}_k^{\text{state}}$ for $k = 1, 2, 3, 12, 13, 14, 15$. However, it is well known that the monthly CPS dataset suffers from underestimation of staying probabilities when comparing the first 4 months with the second 4 months because sectors are coded independently between these months (e.g., Kambourov and Manovskii, 2013). This can be clearly seen in Figure A.17, where we plot the manufacturing staying probabilities. This underestimation is critical to our analysis, so we focus only on the first three worker flow matrices. This also minimizes concerns about sample attrition and the resulting selection (e.g., Moscarini and Thomsson, 2007).

[59] One concern is that this requires too much extrapolation. In Figure A.20, we compare the extrapolated state-specific worker flow matrix series with the aggregate worker flow matrix series we computed in Section 1.4. Reassuringly, the average state-specific worker flow matrix series behaves very similarly to the aggregate series.

[60] The estimated ρ is at the opposite extreme of the CDP's estimate within the range of estimates in the literature. Thus, we also report results under the CDP's estimate in Section 1.4.4. Another issue is that we estimate the parameter ρ at annual frequency. In Appendix A.3.2, we present a way to transform this to quarterly frequency. The resulting value is $\rho = 1.0011$.



(a) Manufacturing Employment: Aggregate Effect

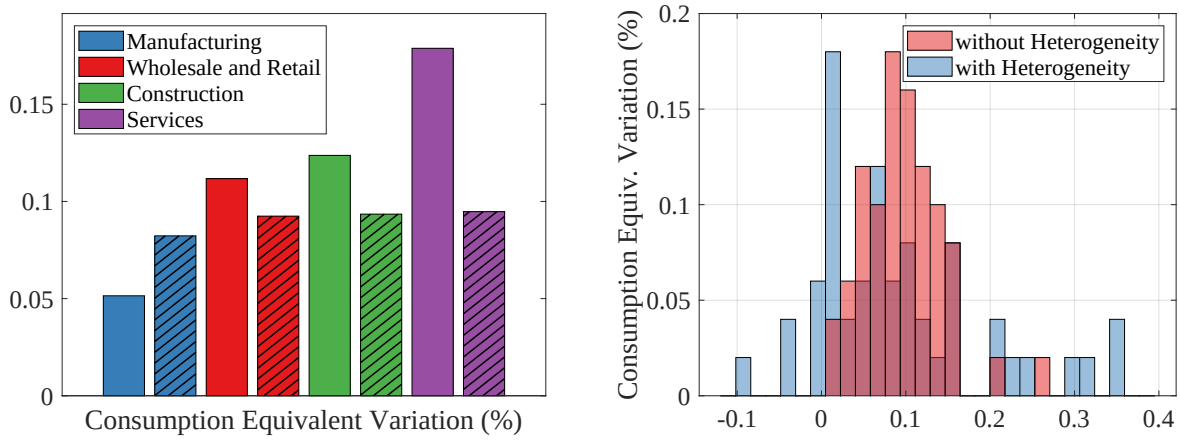
(b) Manufacturing Employment: Individual States

Figure 8. Effect of the China Shock on Welfare

labor mobility costs, and trade costs—for every period. We reduce the computational burden by extending the CDP's dynamic hat algebra to models with arbitrary worker heterogeneity using our sufficient statistics result; see Appendix A.3.2 for details.[61] Figures 8 and 9 plot the results of counterfactual exercises computed using the sufficient statistics approach.

Sectoral Employment Changes. Figure 8a plots the dynamic response of US sectoral employment to the China shock for models with and without worker heterogeneity. In both models, the increase in the manufacturing productivity of China shifts manufacturing production from the US to China, resulting in a decline in the share of US manufacturing employment (top left panel). With worker heterogeneity, the China shock reduces the share of manufacturing employment by around 0.2 percentage points (or, equivalently, 0.45 million manufacturing jobs) after 10 years. At the same time, the China shock increases employment in the construction and services sectors, as observed in the data. These sectors benefit from access to cheaper intermediate goods made available by the China shock. Dashed lines plot changes in the sectoral employment share predicted by the canonical model. As expected from Proposition 4, the canonical model consistently overestimates the extent of labor reallocation by up to 50%. Figure 8b presents a histogram of state-level changes in manufacturing employment after 5 years. The impact of the China shock varies across states in both models, but in line with Figure 8a, the contraction of manufacturing is faster in the canonical model.

[61] Dynamic hat algebra solves the equilibrium of the model in terms of time differences and differences between the actual and counterfactual economies. This method allows one to perform counterfactual exercises without the need to estimate the level of exogenous state variables of the model. For this reason, it is widely used in the literature (e.g., Rodriguez-Clare, Ulate, and Vasquez, 2022; Caliendo et al., 2021; Balboni, 2019; Kleinman, Liu, and Redding, 2023).



(a) Sectoral Welfare Changes: Aggregate Effect

(b) Manufacturing Welfare Changes: Individual States

Figure 9. Effect of the China Shock on Sectoral Employment

Welfare Changes. In terms of welfare, the heterogeneous-worker model predicts a 0.09% increase (in terms of consumption-equivalent variation), which is similar to the 0.11% welfare change predicted by the canonical model. Despite this similarity, the two models differ in their predictions regarding the distributional conflicts—how much the winners win and the losers lose. In Figure 9a, we plot sectoral welfare changes measured at the time the China shock was known to the US labor market. Again, solid bars represent sectoral welfare changes predicted using the sufficient statistics in the data, and hatched bars represent those implied by the canonical model. As in Section 1.5.1, both models predict that even workers who were initially employed in the manufacturing sector benefit from the China shock. However, as expected from the results of Proposition 3, the canonical model significantly overestimates the welfare gain of manufacturing workers, but underestimates the welfare gain of workers in non-manufacturing sectors. Thus, the model significantly underestimates the distributional impact of the China shock. In particular, in the absence of worker heterogeneity, workers who were initially employed in the manufacturing sector enjoy about the same welfare gains as non-manufacturing workers. However, once we account for worker heterogeneity and correctly match longer-run worker flow patterns, their average gain becomes less than half the gains of workers in the other sectors.

Similar to employment effects, the welfare impact of China's import competition varies substantially across regions. In Figure 9b, we present a histogram of state-level changes in the welfare of manufacturing workers. In the canonical model, manufacturing workers in all states benefit from the China shock. In contrast, the heterogeneous-worker model predicts that the welfare gain of manufacturing workers is close to zero in most states, and manufacturing workers from five states—Alabama, Illinois, Louisiana, Michigan, and Ohio—experience a decline in welfare.[62] These

[62] Workers from these five states continue to experience negative welfare effects even in the long run, which is in stark contrast to findings in the literature.

states have the highest manufacturing employment shares and experience higher reallocation from manufacturing to nonemployment. The figure also shows that worker heterogeneity not only amplifies the negative welfare effects of the China shock but also increases regional disparities in the welfare effects on manufacturing workers.

Figure A.22 presents the percentage changes in sectoral wages averaged across states induced by the China shock. The average effect is positive for all sectors, although some states experience declines in the manufacturing real wage. In contrast, non-manufacturing sectors experience wage increases in all states. More importantly, the heterogeneous-worker model predicts slower labor reallocation from manufacturing to non-manufacturing, resulting in a larger decline in manufacturing wages. Motivated by this observation, in Figure A.23 we plot sectoral welfare changes calculated by combining the worker flow matrix series implied by the heterogeneous-worker model and sectoral real wage changes implied by the model without worker heterogeneity. The result implies that around a quarter of the welfare change gap between the two models arises from this difference in real wage changes, and the remaining three-quarters result from differences in the worker flow matrix series. This highlights the importance of endogenizing wage changes when studying the role of worker heterogeneity.

In sum, results of the counterfactual exercises demonstrate the quantitative importance of accounting for worker heterogeneity.

1.6 Concluding Remarks

Large economic shocks often have asymmetric effects across different sectors. Such shocks can necessitate substantial labor reallocation across sectors and may have significant distributional consequences for workers employed in different sectors. The key determinant of the dynamic effects of sectoral shocks is the ease with which workers can switch sectors over time. In this paper, we develop a dynamic sector choice model that incorporates a self-selection mechanism based on persistent worker heterogeneity in an otherwise standard dynamic discrete choice framework. Our sufficient statistics approach, which relies on the information contained in steady-state worker flows over various horizons, highlights in a transparent way the critical role of this self-selection mechanism in shaping the dynamic effects of sectoral shocks. Assuming away persistent worker heterogeneity results in overestimation of steady-state worker flows, which in turn leads to underestimation of the distributional consequences and overestimation of the speed of labor reallocation.

By revisiting the two applications in the literature using our empirical estimates of the sufficient statistics, we illustrate the applicability of our approach and quantify the importance of the added flexibility from worker heterogeneity. Our results present a more pessimistic view of the consequences of sectoral shocks: The reallocation of workers is significantly slower, and the welfare losses of adversely affected workers are more severe than

previously suggested. Although we have focused on specific applications in this paper, the general insights we have developed could be applied more broadly—not only to other sectoral shocks but also to models that incorporate richer mechanisms. For example, if involuntary unemployment is considered, our finding that workers are more likely to be stuck in a negatively affected sector will materialize as a higher unemployment rate in that sector, leading to even greater welfare losses. Applying our approach in the context of different shocks and to models with additional structures is an important direction for future research.

Chapter 2

What Causes Agglomeration of Services? Theory and Evidence from Seoul

Joint with Ryungha Oh

2.1 Introduction

It is well documented that economic activities are highly concentrated in space. An extensive empirical literature documents the agglomeration of various economic activities, and a corresponding theoretical literature analyzes the mechanism that leads to such agglomeration.[1] However, much of this literature focuses primarily on manufacturing industries. Services, which are of comparable importance in terms of expenditure and employment shares, show an even higher degree of concentration. However, the agglomeration of services has been less studied, both theoretically and empirically.

To fill this gap, this paper studies the concentration of non-tradable consumption services, such as restaurants and retail stores. Unlike tradable goods, non-tradable services require that consumers travel to the location where the services are provided.[2] Consumers mainly travel to nearby regions due to spatial frictions, and often make multiple purchases per travel—that is, they exhibit trip-chaining behavior. For example, a consumer may first visit a retail store to shop and then go to a restaurant, or visit a nearby grocery store while waiting on a car repair.

We thank support from the Yale Economic Growth Center and Ryoichi Sasakawa Young Leaders Fellowship Fund. IRB Exemption Determination obtained through Yale University, ID 2000032896.

[1] See Combes and Gobillon (2015), Rosenthal and Strange (2004), and Duranton and Puga (2004) for reviews.

[2] Unlike our terminology, this is commonly referred to as *trade in services* (e.g., Lipsey, 2009; Eaton and Kortum, 2018; Agarwal, Jensen, and Monte, 2020). Another type of trade in services, examined by Muñoz (2023), involves the migration of workers employed by services firms.

Our paper demonstrates the importance of trip-chaining behavior in the concentration of the non-tradable services sector and its implications for efficiency and welfare. Trip chaining suggests that stores in a given location benefit from the presence of other stores, since a purchase at one store increases the likelihood of purchases at a nearby store.[3] We develop a theory of the non-tradable services market that incorporates trip chaining. To quantify the model, we use an original survey of trip chaining and micro datasets from Seoul. We find that spillovers generated by the trip-chaining mechanism account for about one-third of the observed concentration of non-tradable services. Furthermore, we show that despite its importance in concentration, trip-chaining behavior does not lead to inefficiency or exacerbate welfare inequality, which distinguishes it from standard agglomeration mechanisms.

We begin our analysis by documenting the presence of spillovers in the services sector, which is expected given the spatial concentration of services stores. Using a shift-share instrument approach, we causally identify positive spillovers across sectors. Specifically, we find that a 10% exogenous increase in the number of stores in one sector leads to a 3.6% increase in the number of services stores in other sectors in the same region, which indicates substantial positive spillovers. We then discuss the plausibility of the exclusion restrictions and relevance of the instruments in our setting.

Next, we develop a quantitative spatial model that endogenously determines the distribution of services stores. Central to our model is a novel microfounded demand spillover mechanism that arises from trip-chaining behavior in services travel. To incorporate this mechanism, we adopt a dynamic discrete choice framework and exploit its recursive structure to maintain tractability and a gravity equation. We also incorporate scale economies—a standard reduced-form approach for modeling spillovers—which capture the idea that the productivity of services stores in a region increases with the size of the services market. Both trip chaining and scale economies provide potential explanations for the observed spillovers in the services market. An exogenous increase in the number of stores in a particular sector benefits nearby stores, either by stimulating demand through increased foot traffic or by raising productivity levels.

However, we show that the trip-chaining mechanism and scale economies possess distinct efficiency properties. Trip chaining itself is not a source of inefficiency and does not exacerbate or mitigate underlying monopolistic distortions. This result holds regardless of the specific modeling approach used for trip chaining, as long as the model features a constant elasticity of substitution between individual stores. In contrast, when spillovers arise from external economies of scales, the decentralized economy is generically inefficient in terms of both interregional and intraregional resource allocation, and scale economies exacerbate monopolistic distortions. This inefficiency calls for further policy intervention. These findings highlight the importance of distinguishing between the specific mechanisms behind spillovers. To this end, we turn to estimation of the quantitative model, which proceeds in several steps.

[3] We use the term *services stores* or simply *stores* to refer to firms providing non-tradable consumption services.

First, we estimate the spatial friction parameters by fitting the model-implied gravity equation to the observed patterns of services travel. In the second step, we calibrate a subset of the parameters directly using our data. Importantly, we calibrate the degree of trip chaining based on the results of an online survey we conducted, which is specifically designed to collect information on the number of stores visited per travel. Third, we estimate the remaining structural parameters associated with non-tradable services, including the degree of economies of scale, using Bartik-motivated generalized method of moments estimation. This estimation strategy exploits exogenous variation from the shift-share instrument constructed from structural residuals and accounts for spatial linkages. Our results indicate that the trip-chaining mechanism explains a large fraction of the observed spillovers, which suggests that the non-tradable services market operates close to efficiency.

Our estimated model suggests that the spillovers generated by the trip-chaining mechanism explain a significant portion of the concentration of services stores, comparable to the role of location fundamentals or access to consumers. When we turn off the possibility of trip chaining, the dispersion of services, as measured by the standard deviation of the log number of stores, decreases by about 35%. However, trip chaining does not exacerbate inequality in services market access (SMA), which represents the value consumers in each region derive from services travel. Although trip chaining leads to an uneven distribution of services stores, which increases SMA inequality, the trip-chaining behavior itself reduces SMA inequality when the distribution of services is held fixed, and thus offsets the first channel. This occurs because trip chaining effectively reduces the travel disutility per purchase.

Finally, we conduct counterfactual exercises to examine the impact of the rise of work from home and delivery technology for non-tradable services on urban structure. Results indicate that the effect of work from home on the concentration of services depends heavily on the spatial linkages of services consumption between residential and business areas. Even after the rise of work from home, many business areas in Seoul remain highly concentrated due to their strong spatial linkages with residential areas, which attract a significant number of consumers who work from home. In contrast, the emergence of delivery technology has a significant effect on reducing the concentration of services stores. As spatial frictions decrease, stores in concentrated areas that are close to consumers lose their advantage. Interestingly, trip chaining also reduces concentration in industries that do not use delivery services. When fewer consumers visit concentrated areas, all stores are negatively affected by a decrease in potential customers who would otherwise make subsequent purchases.

Related Literature. This paper is related to several strands of the literature. First, the paper contributes to the literature on the agglomeration of economic activities, particularly in the context of the services sector. Studies have provided suggestive evidence of spillovers, such as the colocation of services stores and the increase in rental prices in areas with higher services store densities (e.g., Koster, Pasidis, and van Ommeren, 2019; Leonardi and Moretti, 2023). In addition, several studies have provided a microfoundation for these spillovers, such as consumer benefits from the

agglomeration of stores due to imperfect information or the desire to compare goods (e.g., Konishi, 2005; Takahashi, 2013; Eaton and Lipsey, 1979). However, these mechanisms operate primarily within individual sectors. In this paper, we provide direct evidence of spillovers across sectors and introduce a microfounded spillover mechanism that is consistent with both our empirical findings and the travel behavior of consumers in the data.

The literature on trade in non-tradable services extensively documents customers' travel patterns when purchasing non-tradable goods and services, and shows that spatial frictions are a first-order concern in consumption choices (e.g., Couture, 2016; Davis et al., 2019; Monte, Jensen, and Agarwal, 2020). Some studies specifically examine the determinants of customers' trip chaining patterns, and focus on how they are affected by agglomeration of services or transportation costs (e.g., Anas, 2007; Primerano et al., 2008; Bernardin Jr, Koppelman, and Boyce, 2009; Arentze, Oppewal, and Timmermans, 2005; Relihan, 2022). In contrast, our analysis focuses on how consumer behavior shapes the spatial distribution of services. Moreover, it goes beyond the typical sector-specific or localized analyses found in the literature by examining the entire spatial distribution of general non-tradable services goods in a city.

Finally, we build on the literature that has developed quantitative urban models (e.g., Ahlfeldt et al., 2015). This literature provides a framework for studying the internal structure of cities, including the population distribution of residences and workplaces, the impact of transportation infrastructure, and the effects of agglomeration economies. In this paper, we focus on the distribution of non-tradable consumption services within a city, which has been less studied despite its importance as a main advantage of cities, as shown by Glaeser, Kolko, and Saiz (2001), Couture and Handbury (2020), and Handbury and Weinstein (2015). Recent studies have used quantitative urban models to examine how the distribution of services is shaped, focusing on the effects of the spatial distribution of consumers (e.g., Couture et al., 2021; Almagro and Domínguez-Iino, 2020). Instead, this paper focuses on the importance of non-tradable services travel with trip-chaining behavior for the distribution of services. Miyauchi, Nakajima, and Redding (2022) also model trip chaining in a quantitative urban model and show the importance of the travel itinerary between home and work, which translates the concentration of tradable sectors into the concentration of services. We focus instead on how trip chaining creates spillover forces that operate across nearby services stores, especially within regions, and their impact on the concentration of services.

The rest of the paper is organized as follows. Section 2.2 discusses the background and data, provides reduced-form evidence on the spillover mechanism, and presents stylized facts on services travel. Section 2.3 develops a structural model that features the spillover mechanism, which is estimated in Section 2.4. In Section 2.5, we use the estimated model to investigate the importance of spillovers that arise from the trip-chaining mechanism. Finally, in Section 2.6, we perform counterfactual exercises on the urban structure in the future.

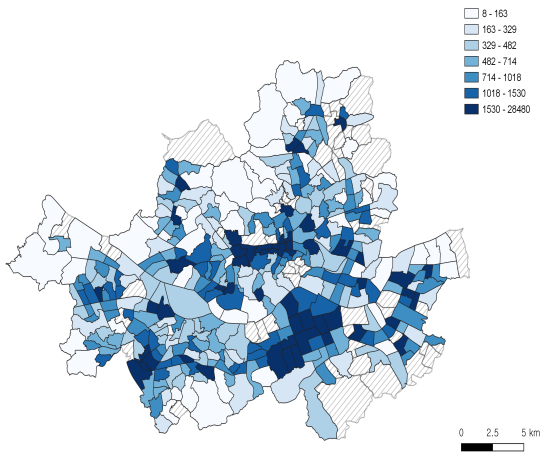


Figure 1. Number of Services Stores Per Area

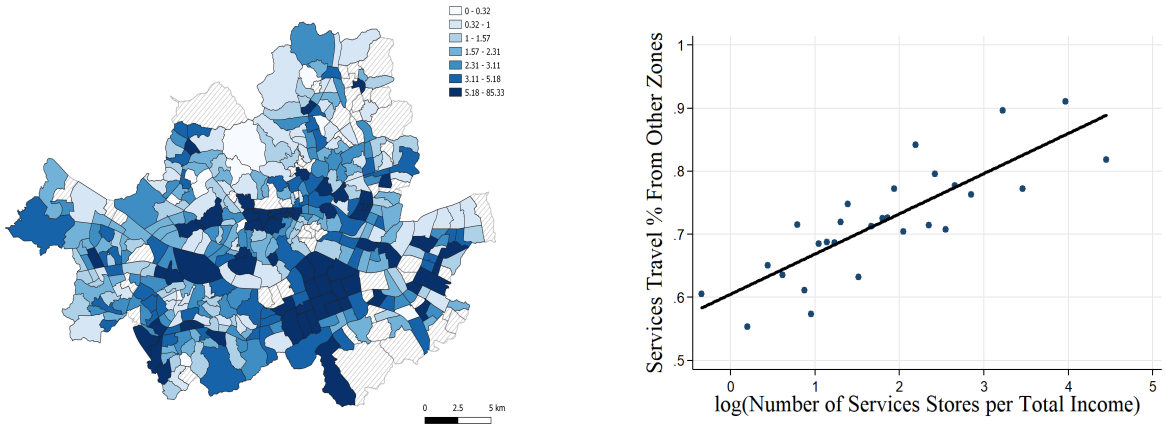
2.2 Motivating Evidence

We analyze the non-tradable services market in Seoul Special Metropolitan City, the capital of South Korea. With a population of approximately 10 million, Seoul accounts for 18.7% of the country's total population and contributes to 22% of its GDP. The geographic unit of this paper is the zone (*dong*) which is contained within a larger spatial unit called the district (*gu*). We use *zone* or *region* interchangeably in this paper. The zone is a granular geographic unit. Seoul consists of 425 zones distributed across 25 districts, and covers an area of 605.21 km^2 (or 233.67 mi^2). This results in an average zone size of about 1.4 km^2 (or 0.55 mi^2).

The supply of services is spatially concentrated and shows strong correlation among sectors. In Figure 1, we plot the number of services stores per area of each zone on the map of Seoul. We can see that the distribution of services across zones is highly uneven. Moreover, zones with high services supply are clustered together in a few areas. In addition, we observe a high correlation in the spatial distribution of services stores in three sectors: Food, Retail, and Other. The correlation coefficients of the log number of stores between each pair of sectors are all above 0.79. This indicates that if a region has many stores in one sector, it is likely to have many stores in the other sectors as well.

This concentration suggests the presence of spillovers, which we define as any forces that increase economic outcomes (here, services stores) as the size of the local economy increases (Combes and Gobillon, 2015). Two observations lend support to the idea that the concentration is not easily explained by differences in local demand or location productivity alone, which suggests the existence of spillover forces.[4] First, even after normalizing the number

[4]Zoning laws in Korea have limited impact on the distribution of services stores because of mild and narrow restrictions that are mostly confined to a few designated areas. In addition, the process of opening and closing services stores is quite efficient, with average quarterly entry and exit rates of 2.5% to 3.5%.



(a) Number of Services Stores Per Total Income

(b) Share of Consumers from Other Zones

Figure 2. Spatial Distribution of Services in Seoul

of stores by the total income of the population in each zone, the distribution of services remains highly concentrated, as shown in Figure 2a. In addition, Figure 2b shows that regions with a higher number of stores per total income tend to have a higher share of consumers from other zones. In the most concentrated areas, more than 80% of consumers travel from other zones for services consumption. Second, since we are analyzing the internal structure of a city, it is unlikely that local productivity differences are large enough to account for most of the spatial disparity.

In this section, we begin by describing the datasets we use in this paper. We then provide reduced-form evidence on spillovers in the non-tradable services sector. Using a shift-share instrument approach, we causally identify positive spillovers across sectors. The goal of the next section is to develop a structural model of the services sector that can explain these spillovers. To guide the modeling choice, we conclude this section by establishing two stylized facts about the demand for services.

2.2.1 Data Description

We rely on three main datasets from Seoul: the Korean Household Travel Survey and Online Household Services Travel Survey on the services demand side, and Seoul commercial area data on the services supply side. Before explaining the details of the data, we will clarify the concepts of travel and trip chaining, which we will refer to throughout the paper: Consumers' services *travel* exhibits trip-chaining behavior in the sense that it consists of a sequence of *trips* (or purchases) that begin and end at locations unrelated to services consumption, such as home or workplace.

Data 1: Korean Household Travel Survey. The first dataset is the Korean Household Travel Survey, a representative travel survey conducted by the Korea Transport Database. The survey asks individuals to report all of their travel on

a given day. The sample includes approximately 200,000 travel instances from 43,000 individuals within Seoul. We use weekday and weekend surveys, both conducted in 2010 and 2016. The dataset provides detailed information on travel, including origin and destination zones, mode of transportation, and demographic information. Importantly, the survey asks in detail about the purpose of the travel, which allows us to focus only on services travel and not on other travel such as commuting. We divide services travel into three categories: *Food*, *Retail*, and *Other*. The last category, Other, consists of recreational activities, exercise, touring, leisure, and private education.

Data 2: Seoul Commercial Area Data. Our second main dataset comes from the Seoul Commercial Area Analysis Service, which is a publicly available big data hub operated by the Seoul Metropolitan Government. It contains a rich set of variables, such as the number of services stores in each region and estimates of their sales and rents. These variables are constructed from confidential cell phone data, credit card transactions, and more. Most of the variables are provided at quarterly frequency starting in 2014, and we aggregate them to an annual frequency. The geographic unit of this dataset is a commercial area which is a smaller unit than a zone. There are 1,496 commercial areas in Seoul, so there are about 3.5 commercial areas in each zone.[5] We map each commercial area to a zone using the mapping table provided. A key advantage of this dataset is that variables are provided at subsector level: We have three sectors (Food, Retail, Other) and within each sector there are 9, 8, and 14 subsectors, respectively.[6] This allows us to use subsector composition to construct shift-share instruments.

Data 3: Online Household Services Travel Survey. To complement the Korean Household Travel Survey, which lacks certain details necessary to accurately quantify the strength of the trip-chaining mechanism,[7] we conduct a supplementary online survey that closely follows the structure of the Korean Household Travel Survey but includes additional questions on trip-chaining patterns. The survey specifically asks respondents about their trip-chaining experiences, including information on the total number of stores visited with at least one purchase, the location, sector, and subsector of each purchase made, and the origin and destination of the travel. We mainly ask about services travel experiences on 2 specific days within the last 7 days—1 during the week and 1 on the weekend.

The complete survey questions can be found in Appendix B.7. We use a survey provider called Embrain—the largest online survey company in Korea—to recruit respondents and conduct the survey online. We collected a

[5] One limitation of this dataset is that commercial areas do not completely cover all of Seoul: Only 375 of the 425 zones appear in the dataset. However, this is not a serious concern because it covers most of the services stores in Seoul. For example, it contains more than 96% of all restaurants in Seoul.

[6] In the raw data, there are 10, 47, and 43 subsectors, respectively, which amounts to a total of 100 subsectors. However, the effective number of subsectors in our dataset is 63 because we do not observe either the number of stores or sales estimates for the other subsectors. We aggregate some of the subsectors if they have no sales estimates or if they appear in only a few zones. After redefining the subsectors, we have a total of 31 subsectors.

[7] Respondents are instructed not to report trips that take less than 5 minutes on foot. In addition, the survey does not provide explicit guidelines on how to report consecutive trips, which may lead to an underreporting of such trips, especially if they are made for the same purpose. This is evidenced by the fact that 90% of the trips in the data are followed by return trips.

total of 2,000 responses in August 2022.[8] To ensure that the sample is representative of Seoul residents, we use a proportional stratified sampling approach based on gender and age group (6 groups for ages 14 to 64) to match the population in the census.[9] We carefully designed the survey to ensure that respondents understand concepts such as trip chaining and provide valid responses. Before the survey begins, we define important variables and provide examples, and respondents are required to read this information for at least 1 minute. To further improve response quality, we check for consistency across different questions. If we detect inconsistent answers, we display a message that encourages respondents to answer carefully and ensure consistency in their responses.

Others. In addition to the three main datasets, we use distance data between each pair of zones. We use the travel time between two zones to best approximate the effective distance. The geographic coordinate of each zone is identified by the location of the zone's community service center, which is a widely used reference point. By using Seoul Bus Open-API, we obtain the expected travel time for the optimal combination of public transportation options. These data can be considered to be similar to the Directions API available on Google Maps. We also use Seoul Business Survey and Seoul Population and Income data for estimation.

2.2.2 Reduced-form Evidence on Spillovers

To provide suggestive evidence of spillovers in the services sector, we estimate the effect of a plausibly exogenous increase in the number of stores in one sector on the number of stores in other sectors within the same zone.[10] For example, suppose there is an exogenous increase in the number of Retail stores in a zone. If this increase can somehow benefit the Food and Other sectors through spillovers, those two sectors would also experience an increase in the number of stores. To quantify cross-sector spillovers, we start with the following specification:

$$\Delta \log N_{jsd} = \alpha_1 + \beta_1 \Delta \log N_{js'} + \boldsymbol{X}'_{jsd}\gamma_1 + \varepsilon_{jsd}, \tag{1}$$

where $\Delta \log N_{jsd}$ is the growth rate of the number of stores in zone j, sector s, and subsector d between years $t = 2015$ and $t' = 2019$. The explanatory variable $\Delta \log N_{js'}$ is the growth rate of the number of stores in zone j and sector

[8] We believe that any potential bias in the results due to the Covid-19 pandemic is minimal. The Korean government lifted strict regulations in response to the pandemic in April 2022, and by the time of the survey, daily foot traffic had returned to pre-pandemic levels.

[9] Our sample may not be fully representative due to the use of an online survey platform. But given the high internet penetration rate in Korea (over 95%), we believe that the online survey method remains a reasonable approach for data collection in this context. However, we cannot rule out the possibility that the recruitment methods used by the survey company may cause a problem. Individuals with higher incomes may be less inclined to participate in online surveys due to the opportunity cost associated with their time.

[10] In this section we do not attempt to distinguish between the various mechanisms underlying spillover. For this purpose, we will combine a structural model with the survey data on trip chaining in the following sections.

$s' \neq s$, defined by the weighted sum of subsector-level growth rates,

$$\Delta \log N_{js'} = \sum_{d'} s_{js'd'} \Delta \log N_{js'd'},$$

where weight $s_{js'd'}$ is the revenue share of subsector d' in sector s' in year $t = 2015$. The covariate $\boldsymbol{X}_{jsd}$, which will be explained later, contains a group of controls.

One threat to identification is the possibility of common regional shocks that simultaneously affect the number of stores in all sectors in the same direction. In such cases, the OLS estimate of specification (1) would be biased upward and falsely indicate strong positive spillovers across sectors. To address this endogeneity concern, we use the shift-share instrument approach of Bartik (1991) to isolate exogenous variation in $\Delta \log N_{js'}$.[11] As an instrument, we use the predicted local growth in the number of stores in a sector, which is computed by interacting the initial subsector composition with city-level subsector growth rates:

$$\Delta \log N_{js'}^{\text{Bartik}} = \sum_{d'} s_{js'd',0} \Delta \log N_{\text{Seoul},s'd'},$$

where $s_{js'd',0}$ is the revenue share in the initial year $t_0 = 2014$, and $\Delta \log N_{\text{Seoul},s'd'}$ is the growth rate in the number of stores in subsector d' in all of Seoul.[12] Our instrument exploits how differential exposure to common city-level preference shifts affects the growth of the number of stores in a sector. For example, suppose that Japanese restaurants became popular throughout the city, and bars became unpopular. If a region initially had a higher share of Japanese restaurants and a lower share of bars, then it likely has a comparative advantage in the former. As a result, a larger share of the gains from the citywide change in preferences would tend to accrue to that region and lead to a higher growth in the *total* number of restaurants, including both Japanese restaurants and bars.[13] In Figure B.2, we show that our Bartik instruments exhibit sufficient spatial variation.

In this regard, our research design is closely related to that of Goldsmith-Pinkham, Sorkin, and Swift (2020) (hereafter, GSS).[14] For our instruments to be valid, the growth trend of sector s, the term ε_{jsd} in specification (1), should be uncorrelated with the initial subsector composition of sector $s' \neq s$, $\{s_{j,s',d',0}\}_{d'}$, conditional on controls $\boldsymbol{X}_{jsd}$. In many papers that use shift-share instruments (e.g., Autor, Dorn, and Hanson, 2013b), the exclusion restriction requires orthogonality between a sector's trend and *its* initial composition (in our context, this can be written as

[11] This approach is widely used in the trade and urban literature; examples include Topalova (2010), Autor, Dorn, and Hanson (2013b), and Dix-Carneiro and Kovak (2017). However, it is rarely used in studies on agglomeration economies. One notable example is Diamond (2016), who uses a shift-share labor demand shock

[12] In practice, we exclude the number of stores in j when calculating city-level growth, although this does not change the results because we have a sufficiently large number of zones.

[13] It is worth noting that a higher initial share may result in more intense competition, leading to smaller growth in the number of *Japanese* restaurants in this region. Nevertheless, as long as this cannibalization effect is not excessively strong, the higher initial share of Japanese restaurants still has a positive impact on the *total* number of restaurants. The results of first-stage regressions in Table 1 confirm that this is the case in our data.

[14] For a different approach to shift-share instruments, see Borusyak, Hull, and Jaravel (2020) and Adão, Kolesár, and Morales (2019).

$\varepsilon_{jsd} \perp \{s_{j,s,d,0}\}_d$). Our requirement is much less demanding than such exclusion restrictions. Moreover, we can even control for the subsector composition of sector s, in which case our exclusion restriction becomes

$$\{s_{j,s',d',0}\}_{d'} \perp \varepsilon_{jsd} \mid \{s_{j,s,d,0}\}_d, \boldsymbol{X}_{jsd}, \quad \forall s' \neq s, \tag{2}$$

which is even easier to hold. A possible concern is that zones with high shares of fast-growing subsectors may be affected by other positive growth shocks, which could confound our estimates of spillovers with these zone trends. In Appendix B.1.2, we conduct the diagnostic tests GSS recommend to further ensure the validity of our instruments.

In Table 1, we report results of the estimation of spillovers. In all specifications, we control for sector s and s' fixed effects, the subsector composition of sector s and of the third sector s''. In practice, instead of including the full vector of subsector shares $\{s_{jsd,0}\}_d \cup \{s_{js''d'',0}\}_{d''}$ as controls, we control for $\sum_d s_{jsd,0} \Delta \log N_{\text{Seoul},sd}$ and $\sum_{d''} s_{js''d'',0} \Delta \log N_{\text{Seoul},s''d''}$. In addition, to control for subsector-specific trends, we include as controls either the city-level growth rate of each subsector, $\Delta \log N_{\text{Seoul},sd}$, or the subsector fixed effect.

In Column (1), we report the result of ordinary least squares estimation. We find significant positive effects, but this result may be biased, as discussed. Therefore, in Column (2), we report the result from the instrumental variable estimation with the same set of controls as in Column (1). In the bottom panel, we report the result of the first-stage regression, where we find a statistically significant positive coefficient despite the presence of cannibalization effects indicated by a coefficient much smaller than 1. The IV regression coefficient is statistically and economically significant, which suggests that an exogenous 10% increase in the number of stores in one sector leads to an average 3.4% increase in the number of stores in other sectors.[15]

In Column (3), we show that the result is robust to the inclusion of additional controls. We include subsector fixed effects and district fixed effects, as well as controls for the growth rates of income, rents, and density. In addition, we control for the levels of these variables in 2015 to address concerns about the exclusion restrictions. We control for these as a precautionary measure, even though the correlations with the instruments are not statistically significant. See Appendix B.1.2 for a more detailed discussion. We find that the spillovers remain significantly large. From our preferred specification in Column (3), we find that a 10% exogenous increase in the number of stores increases the number of stores in other sectors by 3.6%.[16]

In Column (4), we use the theory-consistent specification implied by our model in Section 2.3 and include the same controls as in Column (3). See Appendix B.3.2 for how we can derive the specification using a first-order approximation of the model in Section 2.3. The theory-consistent specification requires that we regress $\Delta \log N_{jsd}$ not only on $\Delta \log N_{js'}$ but also on $\Delta \log \tilde{N}^{\text{Bartik}}_{js}$ and $\Delta \log \tilde{N}^{\text{Bartik}}_{js''}$, where we define quasi-Bartik instruments $\{\Delta \log \tilde{N}^{\text{Bartik}}_{js}\}_s$ by

[15] We find that an OLS coefficient is smaller than IV coefficients, which could be due to measurement errors.

[16] We also compute Conley HAC standard errors, which allow for arbitrary spatial correlation of errors within 3 km, and find that estimates remain statistically significant at the 5% level.

Table 1: Estimation Results

dependent variable: $\Delta \log N_{jsd}$

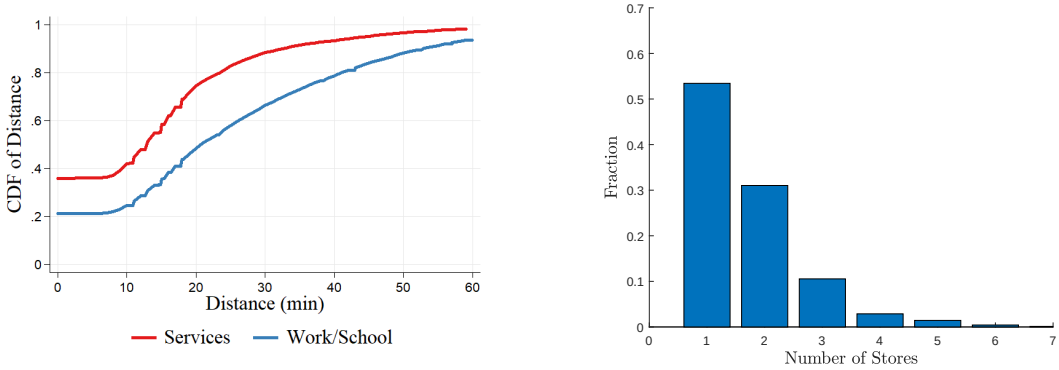
	(1) OLS	(2) IV	(3) IV	(4) IV
$\Delta \log N_{j,s'}$	0.181^{***}	0.341^{**}	0.356^{**}	0.337^{**}
	(0.027)	(0.145)	(0.161)	(0.149)
Sector FE, subsector trend	✓	✓		
Subsector, district FE			✓	✓
Additional controls			✓	✓
	First Stage Estimates			
$\Delta \log N_{j,s'}^{\text{Bartik}}$		0.627^{***}	0.660^{***}	.
		(0.112)	(0.115)	.
First-stage F stat		31.32	32.80	.
Observations	18,773	18,773	17,665	17,665

Notes: Equation estimates based on Seoul Commercial Area data for 2014, 2015, and 2019. We use the longest time period before emergence of the first Covid-19 case in January 2020. Observations are growth rates at zone-sector-subsector level. Standard errors are clustered at zone-sector level. We drop observations with $\Delta \log N_{jsd}$ more than 5 standard deviations from the mean for each subsector. Results remain largely unchanged when we do not implement this trimming, although the coefficients tend to be slightly larger in absolute terms.

interacting $t = 2015$ (instead of initial) subsector composition with city-level subsector growth rates. It also requires that we jointly instrument these three regressors using $(\Delta \log N_{js'}^{\text{Bartik}}, \Delta \log N_{js}^{\text{Bartik}}, \Delta \log N_{js''}^{\text{Bartik}})$. We find that the result is quantitatively similar to that in Column (3).

In Appendix B.1.3, we examine the *within*-sector effect of an exogenous increase in the number of stores. In particular, we estimate the effect on the number of stores in a subsector when there is an exogenous change in the number of stores in other subsectors in the *same* sector. As in the across-sector specification, we exploit the differential exposure to city-level preference shifts to identify exogenous variation in the number of stores. Our results suggest that the number of stores in a subsector decreases by 4.6% on average in response to a 10% exogenous increase in the number of stores in other subsectors within the same sector. This negative effect is likely due to competition between stores within the same sector (e.g., between Korean restaurants and Japanese restaurants) because consumers can more easily substitute between subsectors within a sector. There could be also spillovers that operate within sectors, but our results suggest that the competition is strong enough to produce the negative net effect within sectors. In Section 2.4.1, when we estimate the model, both of the patterns across and within sectors become crucial moments.

The results in Table 1 consistently suggest that an exogenous increase in the number of services stores has significant positive spillovers to other sectors. So far, we have remained agnostic about the specific mechanism that generates these results. In Section 2.3, we propose a novel mechanism that can explain them.



(a) CDF of Travel Distance

(b) The Number of Stores Per Travel

Figure 3. Stylized Facts

2.2.3 Stylized Facts on Services Travel

Before developing a model that microfounds the spillover mechanism, we conclude this section by documenting two stylized facts about consumers' services travel using the Korean Household Travel Survey and the Online Household Services Travel Survey. These facts will discipline our structural analysis in the next section.

Fact 1: Consumers travel to other zones for services—and mainly to nearby zones—which accounts for a significant portion of travel.

Of all types of travel, services travel accounts for 26.1%. This makes it the second most common type of travel after work- and school-related travel, which accounts for 59.8% of all travel. Consumers travel to other zones with a probability of 60%, which indicates that a significant part of the demand for services comes from other regions. However, services travel is often limited to nearby regions. In Figure 3a, we plot the cumulative distribution of distance (in minutes) for both services travel and work and school related travel. The cumulative distribution function for services travel increases in distance much faster than for work- and school-related travel. Nearly 90% of services travel is concentrated within 30 minutes of distance. The average distance of services travel is about 12.5 minutes, which is only one-half the distance of work- and school-related travel (24.8 minutes). Our findings suggest that people are more sensitive to distance when traveling for services than when commuting for work or school.

Fact 2: Consumers make purchases at multiple stores during their services travel.

To assess the extent of trip chaining, we ask the following question in the online survey: "In total, how many purchases did you make per travel? Please write down the total number of stores that you visited with at least one purchase". Figure 3b shows the distribution of the number of stores. About half of consumers reported visiting multiple stores with at least one purchase, and about 20% visited three or more stores per travel. On average, consumers made purchases at 1.72 stores per travel. Additionally, our survey reveals that the majority of travel—approximately

70%—originates from the consumer's home. Furthermore, in about 70% of travel, consumers return to their initial location after completing their purchases.[17] Finally, we find that the sector of the initial purchase does not significantly influence the choice of the sector for the subsequent purchase. About 53% of consumers purchase services goods in the same sector for their first and second purchases, which is similar to the expected probability of 51% if they decide what to buy independently across purchases. These results guide our modeling choices for trip chaining in Section 2.3. More details on the results of the online survey can be found in Appendix B.1.1.

2.3 Theoretical Framework

In this section, we develop a structural model of non-tradable services that explains the positive spillovers across services stores. The key features of the model are services travel and trip-chaining behavior, and the stylized facts discussed earlier guide us in how to model these features. To capture consumers' frequent travel to nearby regions, we assume that consumers can travel to other regions for services consumption, but they are subject to disutility from distance. We also model consumers' trip-chaining behavior explicitly, while maintaining tractability by using a dynamic discrete choice framework as in Rust (1987b) and Aguirregabiria and Mira (2010b). In the model, the spatial distribution of non-tradable services supply is endogenously determined by local characteristics, the spatial distribution of consumers, and their services travel. We also assume that the non-tradable services sector is subject to external economies of scale, which is the most common reduced-form way to model spillovers in the literature.

In our model, both trip chaining and external economies of scale can generate positive spillovers. On the one hand, when consumers make more than one purchase, they are likely to visit nearby stores for successive purchases because they face the disutility from distance. Thus, an exogenous increase in the number of stores in one region attracts potential customers and increases demand for stores in the same and nearby regions, which generates positive spillovers.[18] On the other hand, this exogenous increase also generates positive spillovers by raising the effective productivity of nearby stores through external economies of scale. In Appendix B.3.1, we formalize this intuition by showing analytically that these two mechanisms contribute to positive spillovers. However, distinguishing between different spillover mechanisms is crucial because they can have different implications for the efficiency of decentralized services markets and for regional inequality in access to services markets. In Section 2.3.3, we show that consumers' trip-chaining behavior is not a source of inefficiency. In addition, we show in Section 2.5 that trip chaining does not lead to greater regional inequality in access to services markets, despite its contribution to higher concentration. These findings are in stark contrast to the implications of external economies of scale. External economies of scale

[17] We define the end of travel as when a consumer visits any locations that are not intended for the purchase. Services travel during commuting is not frequent, at least in Seoul, and accounts for only 16% of total travel.

[18] Economists have long recognized that spatially clustered firms can potentially increase profits from larger aggregate demand—a market size effect that encourages clustering.

are generically inefficient and always lead to greater regional inequality. In Section 2.4, we use an original survey on trip chaining to disentangle the role of the trip-chaining mechanism. The result suggests that the trip-chaining mechanism alone explains a large fraction of the spillovers observed in the data.

For expositional purposes, we introduce a model of non-tradable services in the main text in which we take input costs and the spatial distribution of consumers as given. In Appendix B.4.2, we describe how we endogenize input prices and the spatial distribution of consumers—who optimally decide where to live and where to work within a city—in our baseline model in a way that does not change our estimation procedure. In our counterfactual exercises, we will allow these general equilibrium forces to operate.

2.3.1 A Model of Non-Tradable Services

Consider a city that consists of a set of discrete zones $j \in \mathcal{J} \equiv \{1, 2, \cdots, J\}$, with many consumers and stores in each zone. Zones differ in their distance to other zones, rent, wages, the location fundamentals of services sectors, and the number of consumers nearby. A measure $M_{i,i'}$ of workers with an average income of $I_{i'}$ reside in zone $i \in \mathcal{J}$ and work in zone $i' \in \mathcal{J}$. These workers serve as consumers in the services market.

The services market consists of three sectors indexed by $s \in \mathcal{S} \equiv \{1, 2, 3\}$: Food ($s = 1$), Retail ($s = 2$), and Other ($s = 3$). Each sector s consists of a finite number of subsectors indexed by $d \in \mathcal{D}_s$. For each zone-sector-subsector pair (j, s, d), there is a continuum of monopolistically competitive firms with measure N_{jsd} that supply corresponding services goods. For the services market, we will use the terms *firm* and *store* interchangeably, but we need to distinguish them from firms that produce tradable goods.

We first characterize the utility maximization problem of consumers, which determines the demand for services stores. In doing so, we present a tractable model of services travel with trip chaining. We then characterize the profit maximization problem of stores. The free-entry condition endogenously determines the spatial distribution of services stores. Derivations and proofs in this section can be found in Appendix B.2.

Consumers. Consider a worker who lives in zone $i \in \mathcal{J}$ and works in zone $i' \in \mathcal{J}$ with total income I. The *consumption* utility of this worker is given by[19]

$$\mathcal{U}^C_{ii'}(I) = \max_{\tilde{C}, C_r, C_w, C_\ell} \left(\frac{\tilde{C}}{\mu_{\tilde{c}}}\right)^{\mu_{\tilde{c}}} \left(\frac{C_r}{\mu_r}\right)^{\mu_r} \left(\frac{C_w}{\mu_w}\right)^{\mu_w} \left(\frac{C_\ell}{\mu_\ell}\right)^{\mu_\ell} \quad \text{s.t.} \quad p^{tradable}\tilde{C} + P_i C_r + P_{i'} C_w + r_i C_\ell \leq I \tag{3}$$

where $\tilde{C}$ and C_ℓ are the consumption of tradable goods and floor space, which we discuss in more detail when we explain the residence and workplace choices of workers. Workers consume services goods through their services

[19] The final utility depends on both consumption utility and residential amenity.

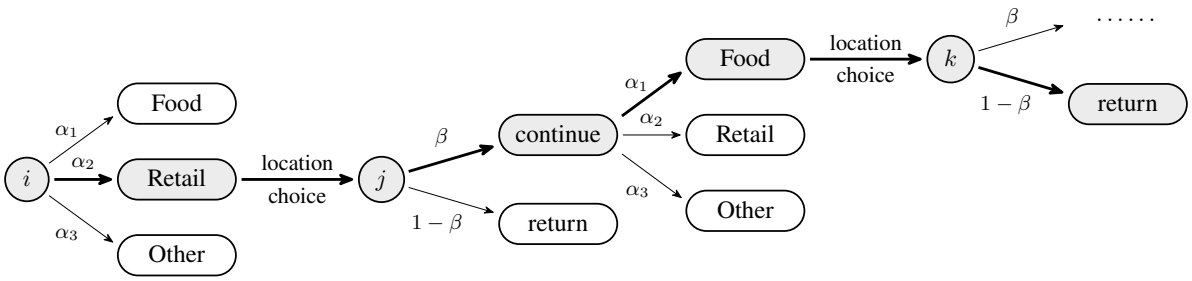


Figure 4. Timeline of Services Travel

travel, which can start from either their residence or workplace zone. Consumption amounts are denoted by C_r and C_w, respectively, with the corresponding price indices P_i and $P_{i'}$, which will be specified shortly when we model services travel. We assume that Cobb-Douglas shares sum to 1, $\mu_{\tilde{c}} + \mu_r + \mu_w + \mu_\ell = 1$.

Services Travel. We model services travel with trip-chaining behavior in a way that can easily be embedded in quantitative urban models. We draw on the data patterns of services travel and trip chaining we document in Section 2.2 to inform our modeling choices. Our priority is to maintain tractability and gravity equations, which we achieve by the extreme value assumption and the recursive structure of the dynamic discrete choice framework.[20]

We consider a consumer who starts her services travel from zone i, which can be either her residence zone or workplace zone, with a slight abuse of notation. The services travel consists of multiple purchases, and the timeline is as follows. The consumer's first decision is to determine the sector and location of the initial purchase. We assume the consumer randomly draws a sector $s \in \mathcal{S}$ with probability α_s, where $\sum_{s\in\mathcal{S}} \alpha_s = 1$, and then chooses a region $j \in \mathcal{J}$ for her initial purchase. After the first purchase, the consumer decides whether to continue her services travel with probability $\beta \in [0, 1)$ or return to region i with probability $1 - \beta$. If she chooses to continue her services travel, she randomly draws a new services sector $s' \in \mathcal{S}$, independent of the previous sector choice s, then chooses a new region $k \in \mathcal{J}$ for her second purchase.[21] In this case, she travels from region j to region k. Importantly, her disutility from distance is measured with respect to the previous region j, not the initial region i. This continues until the consumer decides to end her services travel.[22] The timeline and a specific instance of services travel are illustrated in Figure 4. We first formulate the consumer's problem for each purchase. Consider a consumer who visits zone j to

[20] An alternative modeling approach, following Gentzkow (2007), involves modeling consumers' discrete choice while allowing them to choose any subset from the choice set, which can create complementarities between goods. However, this approach requires solving a combinatorial problem, which can be challenging to achieve tractability and gravity equations. Moreover, when applied at region level, this approach does not permit consumers to visit the same region multiple times, which limits its ability to explain spillovers within a location.

[21] The assumption that the new sector choice is independent of the previous sector is consistent with the services travel pattern in Section 2.2.3. In Appendix B.1.1, we demonstrate that other assumptions on services travel we make are also supported by the data.

[22] Although consumers do not plan their entire travel in advance, they are forward-looking and consider the possibility of continuing their travel when deciding where to go for their first purchase.

make a purchase in sector s. Given spending e for this purchase, she maximizes the effective consumption q_{js}:

$$q_{js}^* = \max_{\{q_{jsd}(\omega)\}_{d,\omega}} \underbrace{\left(\sum_{d\in\mathcal{D}_s} \phi_{jsd}^{\frac{1}{\sigma}} \cdot q_{jsd}^{1-\frac{1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}}_{\equiv q_{js}} \text{ where } q_{jsd} = \left(\int_0^{N_{jsd}} q_{jsd}(\omega)^{1-\frac{1}{\rho}}\,\mathrm{d}\omega\right)^{\frac{\rho}{\rho-1}}$$
$$\text{s.t.} \quad \sum_{d\in\mathcal{D}_s} \int_0^{N_{jsd}} p_{jsd}(\omega) q_{jsd}(\omega)\,\mathrm{d}\omega \le e,$$

where $q_{jsd}(\omega)$ is the quantity purchased in a store ω in (j,s,d) pair; $p_{jsd}(w)$ is the corresponding price; and ϕ_{jsd} is an exogenous preference shifter. The utility function has a nested CES structure in which the upper tier aggregates quantities across subsectors and the lower tier aggregates quantities across individual stores within a subsector.[23] We assume that stores in different subsectors are substitutable, and stores within a subsector are even more substitutable, $1 < \sigma \le \rho$. As is standard with nested CES utility, the maximized effective consumption is given by $q_{js}^* = \frac{e}{p_{js}}$, where p_{js} is the corresponding CES price index,

$$p_{js} = \left(\sum_{d\in\mathcal{D}_s} \phi_{jsd} \cdot p_{jsd}^{1-\sigma}\right)^{\frac{1}{1-\sigma}} \quad \text{where } p_{jsd} = \left(\int_0^{N_{jsd}} p_{jsd}(\omega)^{1-\rho}\,\mathrm{d}\omega\right)^{\frac{1}{1-\rho}}.$$

We now turn our attention from individual purchases to services travel. Let sequential purchases be indexed by $t = 0, 1, 2, \ldots$. For each purchase t, a realization of idiosyncratic shocks, $\varepsilon_t = (\varepsilon_t^j)^j$, and a realization of the sector, s_t, together define the state variable $\sigma_t = (\varepsilon_t, s_t)$. We use σ^t to denote the history up to time t, which has unconditional probability $\pi(\sigma^t)$. The expected value from services travel starting from i with expected total spending equal to E is given by

$$\tilde{V}(i,E) = \max_{\{q_t(\cdot), j_t(\cdot)\}_t} \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U(q_t(\sigma^t)) - \tilde{d}(j_{t-1}(\sigma^{t-1}), j_t(\sigma^t)) + \nu\varepsilon_t^{j_t(\sigma^t)}\Big)\pi(\sigma^t) \tag{4}$$
$$\text{s.t.} \quad \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} q_t(\sigma^t) p_{j_t(\sigma^t)s(\sigma^t)} \pi(\sigma^t) \le E \tag{5}$$

where $j_{-1}(\sigma^{-1}) = i$. We assume $U(\cdot) \equiv \log(\cdot)$, which is a common choice in the dynamic discrete choice literature.[24] After drawing a sector for purchase t, the consumer observes the idiosyncratic component of utility ε_t, which follows a type I extreme value distribution. The consumer then chooses where to visit, $j_t(\sigma^t)$, and how much to consume, $q_t(\sigma^t)$, for purchase t. Traveling between regions j and j' entails disutility $\tilde{d}(j,j')$, which represents the spatial frictions in services consumption. We assume that $\tilde{d}(j,j')$ is given by $\tau d(j,j') + \varphi \mathbb{1}_{j\neq j'}$, where d is the

[23] An alternative interpretation of the nested CES utility structure is that once a consumer arrives in zone j to make a purchase in sector s, they observe preference shocks correlated within sectors and choose the individual store (d,ω) that provides the highest utility. See Verboven (1996) for the equivalence between nested logit models and nested CES models.

[24] Note that log utility with a type I extreme value distributed additive error term is equivalent to linear utility with a Fréchet distributed multiplicative error term.

distance between two regions, and the second term captures the border effect. In Appendix B.3.1, we show that we need to assume $1 + \frac{1}{\nu} < \rho$ to guarantee the stability of the equilibrium. We maintain this parameter restriction throughout the paper.

In Appendix B.2, we prove that it is optimal for consumers to equalize the expenditure across purchases, independent of the regions and sectors they visit. Using this fact, we show that the maximization problem (4) can be recursively expressed as

$$\tilde{V}(i, E) \equiv V(i, e) = \mathbb{E}\left[\sum_s \alpha_s \left(\max_j \left\{U(e/p_{js}) - \tilde{d}(i, j) + \beta V(j, e) + \nu\varepsilon^j\right\}\right)\right] \tag{6}$$

where $V(i, e)$ is the expected value from services travel starting from i with an equalized per-purchase expenditure e, and e and E are related by $E = e+\beta e+\beta^2 e+\cdots = \frac{e}{1-\beta}$. The expectation is taken over realizations of the idiosyncratic shocks. We borrow this recursive formulation from the dynamic discrete choice literature, but with β representing the probability of continuing services travel, rather than the discount factor. Using standard extreme-value algebra, we can further simplify this value to

$$V(i, e) = \sum_s \alpha_s \nu \log\left(\sum_j \exp\left(U(e/p_{js}) - \tilde{d}(i, j) + \beta V(j, e)\right)^{1/\nu}\right). \tag{7}$$

Finally, we define the consumption index for services travel from zone i with per-purchase spending e as[25]

$$C(i, e) \equiv \exp\big((1 - \beta) \cdot V(i, e)\big). \tag{8}$$

We can show that this consumption index is linear in spending e. This allows us to define the price index of services travel from zone i, denoted by P_i, as the amount of per-purchase spending e needed to buy one unit of consumption index. In other words, in the consumption utility maximization problem (3), a worker who lives in zone i and works in zone i' has to pay $P_i C_r + P_{i'} C_w$ in order to consume C_r and C_w units of consumption goods from services travel.

Services Stores. Within each zone-sector-subsector pair (j, s, d), there is a measure N_{jsd} of homogeneous stores indexed by ω. Under monopolistic competition, the stores choose how much to produce given the inverse demand function they face. In particular, they maximize the following profits net of operating costs:

$$\pi_{jsd} \equiv \max_{\substack{p_{jsd}(\omega),\, q_{jsd}(\omega),\\ H_{jsd}(\omega),\, L_{jsd}(\omega)}} p_{jsd}(\omega) q_{jsd}(\omega) - \sum_j \big(r_j H_{jsd}(\omega) + w_j L_{jsd}(\omega)\big) - \tfrac{(\rho-1)^{\sigma-1}}{\rho^\sigma} \cdot C_{jsd}$$

[25] By applying the exponentiation operation to counteract the logarithmic operation in U, we achieve the linearity necessary to define the price index for services travel. We also multiply the expected value by $(1 - \beta)$ to offset the impact of continuation probability β on the expected number of purchases, $\frac{1}{(1-\beta)}$, This isolates the effect of combining purchases through trip chaining, holding fixed the expected number of purchases.

subject to the inverse demand function and production function. Here, $q_{jsd}(\omega)$ is the quantity produced; $p_{jsd}(\omega)$ is the price they set; $H_{jsd}(\omega)$ and $L_{jsd}(\omega)$ are the land and labor they use in production; r_j and w_j are corresponding input prices; and C_{jsd} is a fixed operating cost with a convenient normalization $\frac{(\rho-1)^{\sigma-1}}{\rho^\sigma}$. The inverse demand function comes from the consumer's utility maximization and will soon be characterized. The technology is given by

$$q_{jsd}(\omega) = A_{jsd} \cdot H_{jsd}(\omega)^\gamma L_{jsd}(\omega)^{1-\gamma},$$

where $A_{jsd} \geq 0$ is the common productivity of stores in (j, s, d). The measure of stores N_{jsd} is determined by the free entry condition, $\pi_{jsd} = 0$. Note that we allow for the possibility of $A_{jsd} = 0$, in which case we have $N_{jsd} = 0$.

External Economies of Scale. We introduce external economies of scale as an additional force of agglomeration beyond the trip-chaining mechanism. We allow the productivity A_{jsd} and fixed operating cost C_{jsd} of a region to depend on the size of its services sector. In the literature on agglomeration, size is often measured in terms of total employment. However, in our model, the production of services goods requires not only labor but also land and fixed costs for store creation. Therefore, we assume that productivity and fixed operating costs depend on the total resources spent on production and store creation in region j:

$$A_{jsd} = A_{jsd}(\Upsilon_{1j}, \Upsilon_{2j})$$
$$C_{jsd} = C_{jsd}(\Upsilon_{1j}, \Upsilon_{2j})$$

where $\Upsilon_{1j} = \sum_{s,d} \frac{c_{jsd}}{A_{jsd}} \left(\int_0^{N_{jsd}} q_{jsd}(\omega)\, \mathrm{d}\omega \right)$ and $\Upsilon_{2j} = \sum_{s,d} C_{jsd} N_{jsd}$ represent the total resources expended on production and variety creation for region j, respectively.

2.3.2 Equilibrium

We start with the definition of equilibrium. Given input prices $\{r_j, w_j\}$, the consumer distribution $\{M_{ii'}\}$, and their average income $\{I_{i'}\}$, the *equilibrium* of the non-tradable services market consists of a set of allocations $\{q_{jsd}(\omega)\}$, prices $\{p_{jsd}(\omega)\}$, the distribution of stores $\{N_{jsd}\}$, and the values of productivities and fixed operating costs $\{A_{jsd}, C_{jsd}\}$ such that *(i)* consumers optimally choose their consumption plans given prices and the distribution of stores; *(ii)* stores optimally choose their production plans and prices given the demand they face; *(iii)* productivities and fixed operating costs are endogenously determined, *(iv)* all markets clear; and *(v)* the free entry condition is satisfied. We first characterize the solution of the consumer's utility maximization problem and then use the result to solve the store's profit maximization problem.

Consumer Problem. Consumers solve a standard discrete choice problem (6), and it is well known (e.g., see Train, 2003) that the probability of choosing region j for sector s consumption from region i is given by

$$\pi_{ij}^{s} \equiv \Pr(i \to j|s) = \frac{\exp\left(-\log p_{js} - \tilde{d}(i,j) + \beta V(j)\right)^{1/\nu}}{\sum_{j' \in \mathcal{J}} \exp\left(-\log p_{j's} - \tilde{d}(i,j') + \beta V(j')\right)^{1/\nu}} \tag{9}$$

where $V(j) \equiv V(j,1)$. This yields a structural gravity equation for services travel flows for each sector. In Sections 2.4 and 2.5, we assess the level of regional inequality, with respect to access to the services market and examine how it is affected by the economic environment. To this end, we define services market access (SMA) for zone i as the inverse of the price index $1/P_i$. Services market access summarizes the impact of the spatial distribution of services stores on the attractiveness of zone i as a starting point for services travel (Donaldson and Hornbeck, 2016). For instance, a zone with high SMA indicates that there are many stores located in or around the zone, which makes it a more desirable starting point for services travel.

Definition 1 (Services Market Access). *The* services market access SMA_i *for zone* i *is recursively defined as*

$$SMA_i^{1/(1-\beta)} = \prod_s \left(\sum_j \left(p_{js}^{-1/\nu} \cdot \exp(-\tilde{d}(i,j))^{1/\nu} \cdot SMA_j^{\beta/(\nu(1-\beta))} \right) \right)^{\alpha_s \nu}.$$

Store Problem. With this characterization of consumer choices, we turn to the store's problem. We start with two observations. First, consumers spend a share μ_c^r (a share μ_c^w, respectively) of their income on services travel that starts from their residence zone (workplace zone, respectively). Second, consumer demand is homogeneous of degree one with respect to expenditure. These two observations imply that what is important for firms is the *effective distribution of consumers*, $\{E_i\}$, which is defined as

$$E_i \equiv \mu_r \sum_{i' \in \mathcal{J}} M_{ii'} \cdot I_{i'} + \mu_w \sum_{i' \in \mathcal{J}} M_{i'i} \cdot I_i.$$

Then, it is as if there were a single representative consumer in each zone i who spends E_i on services travel and always starts the services travel from zone i.

We show that the total consumer spending in region j and sector s is given by (see Appendix B.2)

$$R_{js} \equiv \sum_{t=0}^{\infty} (1-\beta)\beta^t \alpha_s \boldsymbol{E}^{\mathsf{T}} \Pi^t \boldsymbol{\pi}_j^s = (1-\beta)\alpha_s \boldsymbol{E}^{\mathsf{T}} (I - \beta\Pi)^{-1} \boldsymbol{\pi}_j^s, \tag{10}$$

where $\boldsymbol{E} = (E_1, \cdots, E_J)^{\mathsf{T}}$ is the vector of the effective distribution of consumers; $\boldsymbol{\pi}_j^s = (\pi_{1j}^s, \cdots, \pi_{Jj}^s)^{\mathsf{T}}$ denotes the vector of location choice probabilities; and Π is a $J \times J$ matrix with (i,i')-element $\pi_{ii'} = \sum_s \alpha_s \pi_{ii'}^s$. Given the

aggregate demand on (j, s), the demand for an individual store, $q_{jsd}(\omega)$, is isoelastic, which results in constant-markup pricing, $p_{jsd}(\omega) = \frac{\rho}{\rho-1}\frac{c_{jsd}}{A_{jsd}}$, where the unit cost c_{jsd} is given by $c_{jsd} = \left(\frac{r_j}{\gamma}\right)^{\gamma}\left(\frac{w_j}{1-\gamma}\right)^{1-\gamma}$.

The number of stores, N_{jsd}, is determined by the free-entry condition, which equates profits with operating costs. We summarize the result in the following proposition.

Proposition 1 (Number of Stores). *The number of stores in (j, s, d) is determined by firm optimization and the free-entry condition:*

$$N_{jsd}^{1-\frac{\sigma-1}{\rho-1}} = C_{jsd}^{-1}\tilde{A}_{jsd}^{-(1-\sigma)}c_{jsd}^{1-\sigma}p_{js}^{-(1-\sigma)}(1-\beta)\alpha_s \boldsymbol{E}^{\intercal}(I-\beta\Pi)^{-1}\boldsymbol{\pi}_j^s$$

where $\tilde{A}_{jsd} = A_{jsd} \cdot \phi_{jsd}^{\frac{1}{\sigma-1}}$ is the composite productivity of stores in (j, s, d), which combines productivity and consumer preferences.

In Section 2.4, we will confront our structural model with the data by nonlinearly estimating the structural parameters. Before moving to the estimation, however, a first-order approximation analysis similar to those in Costinot et al. (2019) and Fajgelbaum et al. (2021) would be helpful to understand the role of trip chaining and external economies of scale. Following their approach, Appendix B.3.1 presents analytical results that characterize the effects of these mechanisms on the spatial distribution of services stores. In particular, we find that a favorable shock in a sector $s' \neq s$ has a positive effect on the number of stores in sector s in the same region. This effect vanishes as the degree of trip chaining and the external economies of scale approach zero. In Appendix B.3.2, we use these results to establish a theory-consistent specification for the reduced-form estimation of spillovers, which corresponds to column (4) of Table 1. This specification yields an IV coefficient of interest that converges in probability to a positive value, which again vanishes as the degree of trip chaining and the external economies of scale go to zero. In sum, through the mechanisms of trip chaining and external economies of scale, our structural model can *qualitatively* match the reduced-form evidence of spillovers presented in Section 2.2.2. In Section 2.4, we further show that this model can *quantitatively* match these patterns.

2.3.3 Efficiency Properties of Equilibrium

In this section, we demonstrate the different efficiency properties of trip chaining and external economies of scale, although both are potential explanations for spillovers in the services market. On the one hand, trip chaining itself is not a source of inefficiency. Trip-chaining behavior only affects the mapping from underlying quantities to utility, as formalized in Appendix B.4.1, so in a hypothetical world with no underlying inefficiencies, we could invoke the first welfare theorem to conclude that trip chaining does not introduce inefficiency. However, the presence of monopolistic distortions complicates the efficiency implications of trip chaining. Nevertheless, we can show that

trip-chaining behavior neither exacerbates nor mitigates monopolistic distortions, which implies that trip chaining does not introduce any additional inefficiency. On the other hand, when spillovers arise from external economies of scale, the decentralized economy is generically inefficient. This observation suggests that the presence of spillovers in the data does not necessarily indicate inefficiency in non-tradable services markets. Indeed, in Section 2.4, we find that the trip-chaining mechanism largely explains the observed spillovers, which implies that the non-tradable services market is close to efficient.

Efficiency Properties of Trip Chaining. Our argument proceeds in two steps to show that trip-chaining behavior does not lead to additional inefficiency. For the purpose of this discussion, we assume for the moment the absence of external economies of scale. We start with a constrained social planner problem that focuses on resource allocation *within* the services market. The social planner maximizes social welfare under the constraint that the resource allocation between tradable goods consumption and the services market cannot be changed. We assume that the social planner maximizes the Pareto weighted sum of the log utilities of consumers, where the Pareto weight is proportional to their income. This particular choice of Pareto weights guarantees that the decentralized allocation is constrained efficient when trip chaining is not allowed ($\beta = 0$). By considering this benchmark social planner problem, we can exclusively examine the potential inefficiency that arises from the trip-chaining mechanism.

In this economy, there are two potential sources of inefficiency in resource allocation. The first is the inefficient allocation of resources across consumption regions, and the second is the inefficient allocation of resources between production and store creation within a consumption region, which reflects the quantity-diversity trade-off discussed by Dixit and Stiglitz (1977). The first part of Proposition 2 shows that the decentralized resource allocation within the services market—across regions and between production and variety creation—remains efficient regardless of the presence of trip chaining. Further details and formal proofs of the results presented in this section can be found in Appendix B.4.1.[26] Importantly, our proof demonstrates that the result does not rely on the specific modeling of trip chaining, as long as the model features a constant elasticity of substitution between individual stores. For example, the result holds when consumers plan their entire itinerary, including the number of trips and where to visit, before starting their services travel.

In addition, we consider an unconstrained social planner problem that involves resource allocation between tradable goods consumption and the services market. Due to the underlying monopolistic distortions in the services sector, the unconstrained social planner would reallocate resources from tradable goods consumption to nontradable services consumption and store creation, consistent with the finding of Dixit and Stiglitz (1977). However, this inefficiency does not interact with trip chaining, which means that the amount of resource reallocation is independent of the value of β. In particular, the second part of the proposition shows that the social planner increases the number

[26]These results extend the CES efficiency results established by Dixit and Stiglitz (1977) and Dhingra and Morrow (2019) by incorporating nested aggregation, heterogeneous consumers, and external economies of scale.

of non-tradable services stores proportionally more than the decentralized number of stores, but this proportionality does not depend on the presence of trip chaining. Based on these observations, we conclude that the trip-chaining mechanism does not represent an additional source of inefficiency in this economy.

Proposition 2 (Efficiency Properties of Trip Chaining).

(1) When trip chaining is not allowed ($\beta = 0$), the decentralized equilibrium coincides with the solution to the constrained social planner problem, where the Pareto weights on the log utilities of consumers are proportional to their income. Even when trip chaining is allowed ($\beta > 0$), the decentralized equilibrium still aligns with the solution to the same constrained social planner problem with exactly the same Pareto weights.

(2) The unconstrained social planner chooses the number of non-tradable services stores $\{N^*_{jsd}\}$ given by

$$N^*_{jsd} = \chi \cdot N^{de}_{jsd},$$

where N^{de}_{jsd} represents the number of stores in the decentralized equilibrium and $\chi = \frac{\rho}{\rho-(1-\mu)} > 1$ is a constant that is unaffected by the presence of trip chaining. The optimal allocation can be implemented by the combination of a tax on tradable goods and a subsidy on non-tradable services, which are again independent of the presence of trip chaining.

Efficiency Properties of External Economies of Scale. Again, we consider both constrained and unconstrained social planner problems to illustrate the inefficiency of the services market with external economies of scale. The first part of Proposition 3 shows that the non-tradable services market with external economies of scale is generically constrained inefficient, except for the special case of isoelastic external economies of scale. We define external economies of scale as *isoelastic* when they take the form of either

$$A_{jsd} = \bar{A}_{jsd} \cdot \Upsilon_{1j}^{\varepsilon} \text{ and } C_{jsd} = \bar{C}_{jsd} \cdot \Upsilon_{2j}^{-\varepsilon} \tag{11}$$

or

$$A_{jsd} = \bar{A}_{jsd} \cdot \Upsilon_{j}^{\varepsilon} \text{ and } C_{jsd} = \bar{C}_{jsd} \cdot \Upsilon_{j}^{-\varepsilon}, \tag{12}$$

where $\Upsilon_j = \Upsilon_{1j} + \Upsilon_{2j}$. In Appendix B.4.1, we characterize the inefficiency in terms of both interregional and intraregional resource allocation. In addition, the presence of external economies of scale exacerbates the inefficient allocation of resources between tradable goods and non-tradable services. In particular, the second part of the proposition shows that the extent of resource reallocation increases with the degree of external economies of scale.

Taken together, these findings highlight the inherent inefficiency that arises in the non-tradable services market with external economies of scale.

Proposition 3 (Efficiency Properties of External Economies of Scale).

(1) With isoelastic external economies of scale, the decentralized equilibrium solves the social planner problem. However, beyond this special case, the non-tradable services market is constrained inefficient in terms of both interregional and intraregional allocation.

(2) With isoelastic external economies of scale, the unconstrained social planner chooses the number of non-tradable services stores $\{N^*_{jsd}\}$ *as*

$$N^*_{jsd} = \chi(\varepsilon) \cdot N^{de}_{jsd},$$

where N^{de}_{jsd} *represents the number of stores in the decentralized equilibrium, and* $\chi(\varepsilon) > 1$ *is an increasing function of* ε*. The optimal allocation can be implemented through a combination of a tax on tradable goods and a subsidy for non-tradable services.*

Nonparametrically estimating the specific form of external economies of scale is beyond the scope of this paper. Therefore, in the next section where we estimate the structural model, we make the assumption of isoelastic external economies of scale and focus on estimating a single parameter ε.[27] It is worth noting that both forms of isoelastic external economies of scale, (11) and (12), are isomorphic in terms of the changes in endogenous variables, because the ratio $\Upsilon_{1j} : \Upsilon_{2j} : \Upsilon_j = 1 : (\rho - 1) : \rho$ remains constant in the decentralized equilibrium.

2.4 Estimation

In this section, we discuss the quantification of the model with particular emphasis on estimating the degree of trip chaining and external economies of scale. We estimate the former using the results of our original survey, and the latter is estimated by matching residual spillovers. In particular, we introduce a novel Bartik-motivated generalized method of moments estimation approach that allows us to exploit exogenous variation from a shift-share design, while accounting for spatial linkages.

[27] Thus, in the estimated model, external economies of scale do not lead to constrained inefficiency. However, as emphasized in Proposition 3, this is just a knife-edge case.

2.4.1 Parameter Estimation

We estimate the parameters of the model in four steps. First, the model-implied gravity equation allows us to estimate the parameters $\tilde{\tau} \equiv \tau/\nu$ and $\tilde{\varphi} \equiv \varphi/\nu$ without solving the full model. Second, we calibrate a subset of parameters using our data, which includes the degree of trip chaining. Third, we estimate the remaining parameters for our model of non-tradable services in Section 2.3.1. As in Ahlfeldt et al. (2015), we invert the model to back out the values of exogenous variables that rationalize the observed services market data. We then use these inverted data as input for the GMM estimation, with moments motivated by the reduced-form evidence. Finally, we calibrate the parameters for the general equilibrium component of our model. In this section, we focus on the first three steps of the estimation procedure. See Appendix B.5.2 for estimation of the general equilibrium parameters.

Gravity Equation. From the location choice probability of consumers, we derive a gravity equation for services travel, which is similar to the conventional gravity equation for trade or commuting flows:

$$\ln \pi_{ij}^{s} = -\tilde{\tau} d(i,j) - \tilde{\varphi} \mathbb{1}\{i \neq j\} + \mathsf{FE}_{js} + \mathsf{FE}^{is} + \epsilon_{ij},$$

where π_{ij}^{s} is the probability that a consumer from region i chooses region j when purchasing a good in sector s, and $d(i,j)$ is the travel time distance between two zones, measured in minutes.[28] The destination–sector fixed effect FE_{js} captures the price index and the expected continuation value, while the origin–sector fixed effect FE^{is} measures market access for consumers. The normalized coefficients $\tilde{\tau} = \tau/\nu$ and $\tilde{\varphi} = \varphi/\nu$ represent the semi-elasticity of services travel and the border effect, respectively. As described in Fact 1, a significant fraction of consumers stay in the same zone when purchasing services goods. This observation motivates us to include the border effect to improve model fit. Finally, the error term ϵ_{ij} captures the measurement error that is independent of the other variables on the right-hand side.

Table 2 reports estimation results. Column (1) reports the results of OLS estimation without the border effect. This estimate implies that an additional 10-minute increase in distance reduces services travel flows by about 30%. When the border effect is included in Column (2), the estimate drops to 0.016, since the border effect accounts for the significant decline in services travel around zero distance. In addition, our gravity fit improves with inclusion of the border effect, increasing R-squared from 0.533 to 0.592.

Despite the large number of observations in our data, due to the granularity of our geographic unit, we observe that a substantial fraction of pairs of regions have zero travel between them.[29] To incorporate these zero observations

[28] Travel time varies depending on the transportation mode chosen. But, for simplicity, we assume that consumers choose the optimal public transportation combination, as explained in Section 2.2.1. We use the same distance measure for commuting decisions in the general equilibrium model.

[29] Out of approximately 540K pairs of zones and sectors (=$424^2 \times 3$), we only observe flows for about 10K pairs, which account for less than 2% of total pairs. We do not use a larger geographic unit because services travel is highly sensitive to distance.

Table 2: Estimation Results: Gravity Equation

	OLS		PPML	
	(1)	(2)	(3)	(4)
Distance ($\tilde{\tau}$)	0.033***	0.016***	0.161***	0.152***
	(0.0011)	(0.0010)	(0.0014)	(0.0025)
Border effect ($\tilde{\varphi}$)		1.084***		0.357***
		(0.0380)		(0.0644)
Fixed Effects	✓	✓	✓	✓
Observations	8,409	8,409	522,512	522,512
(pseudo) R^2	0.533	0.592	0.539	0.539

Notes: Data source: Household Travel Survey (2016) for both weekdays and weekends. Distance is measured in minutes. Fixed effects represent destination-sector (j, s) and origin-sector (i, s) fixed effects. Robust standard errors are shown in parentheses, with *** $p < 0.001$. For PPML, we report pseudo R-squared in the last row.

into estimation, in Column (3) we report the results of Poisson pseudo maximum likelihood (PPML) estimation with the same specification (see, e.g., Silva and Tenreyro, 2006). The semi-elasticity is 16.1% in Column (3), which is five times higher than that in Column (1) due to the inclusion of pairs with zero travel in the estimation. In Column (4), which is our preferred specification, we include the border effect. The result shows that an additional 10-minute increase in travel-time distance reduces services travel flows by about 80%. The (pseudo) R-squared is reported in the last row of the table.

Parameter Calibration. We calibrate five types of parameters: $\{\alpha_s\}_s, \{\mu_{\bar{c}}, \mu_r, \mu_w, \mu_\ell\}, \gamma, \rho$, and β. The most important parameter is travel continuation probability β, which governs the magnitude of spillovers from trip chaining. We calibrate this parameter directly from the Online Household Services Travel Survey. According to the survey, the number of services stores visited for purchases is on average 1.72 per services travel, which implies $\beta = 0.419$. We then estimate the Cobb-Douglas share of each sector, α_s, directly from the revenue shares of each services sector. To calibrate Cobb-Douglas expenditure shares of services travel, μ_r and μ_w, we calculate the ratio between the share of services travel from home and workplace, $\frac{\mu_r}{\mu_w}$, using information on the origin locations from the Household Travel Survey. Next, we divide the total revenue of services sectors obtained from our commercial data by the total income of workers in Seoul to compute the total expenditure share of services, $\mu_r + \mu_w$. These two moments allow us to calibrate the values of μ_r and μ_w. The remaining parameters are difficult to calibrate from our dataset, so we rely on aggregate moments or central values from the literature. For the share of household spending on housing, we use $\mu_\ell = 25.3\%$ from the Seoul Household Consumption Spending Survey from 2006. This estimate is in line with those in the literature, such as Ahlfeldt et al. (2015). For the Cobb-Douglas share of firm spending on commercial floor space γ, we use 20%, which is the value commonly used in the literature (e.g., Valentinyi and Herrendorf, 2008;

Ahlfeldt et al., 2015). Finally, we set ρ equal to 9 based on Couture (2016), whose estimates range from 8.4 to 9.2. He uses detailed information on restaurants and household travel to estimate the elasticity of substitution across stores.

Bartik GMM. We estimate the remaining parameters of the model of non-tradable services. In particular, we estimate ε, σ, and ν using the generalized method of moments, which proceeds in two steps. Given the parameterized model of non-tradable services, we first back out local composite productivity $\log \tilde{A}_{jsd}$ by inverting the model (e.g., Berry, 1994; Ahlfeldt et al., 2015). See Appendix B.5.1 for details on model inversion. Note that composite productivity is a structural residual of our model, which captures productivity and preferences. We then construct three types of moment conditions with the composite productivity, which can be used to estimate the three parameters. These moment conditions are based on the same identification idea as the reduced-form evidence in Section 2.2.[30]

To isolate exogenous changes in the composite productivity of each sector in each region, we compute the predicted change in composite productivity by interacting the initial subsector composition with the city-level growth in composite productivity across subsectors:[31,32]

$$\Delta \log \tilde{A}_{js}^{Bartik} = \sum_{d'} s_{jsd',0} \cdot \Delta \log \tilde{A}_{\text{Seoul},sd'}.$$

This variable is defined analogously to the Bartik instruments in Section 2.2, but using composite productivity instead of the number of stores. Our first set of moment conditions is given by

$$\Delta \log \tilde{A}_{jsd} \perp_j \Delta \log \tilde{A}_{js'}^{Bartik} \quad \text{for all } (s, s', d) \text{ with } s \neq s'. \tag{13}$$

This condition requires that the change in the composite productivity of a sector is orthogonal to the exogenous change in the composite productivity of another sector.

For each subsector d in a region, the following instrumental variable captures the exogenous change in the composite productivity of subsectors other than d in the region:

$$\Delta \log \tilde{A}_{js,-d}^{Bartik} = \sum_{d' \neq d} s_{jsd',0} \cdot \Delta \log \tilde{A}_{\text{Seoul},sd'}.$$

[30] A comparison with the approach in Section 2.2 is in order. In Section 2.2, we construct Bartik instruments based on the number of stores, which is an endogenous variable. In this section, we instead construct Bartik instruments based on composite productivity $\tilde{A}$, which provides two advantages. First, considering shocks to exogenous variables, we can get a clearer economic interpretation of estimation results. Second, our structural model allows us to easily incorporate spatial linkages in a theory-consistent manner. Incorporating spatial linkages in reduced-form shift-share research designs is inherently challenging, and this limitation is often acknowledged in the literature on shift-share instruments (see, e.g., GSS). Adão, Arkolakis, and Esposito (2020) also emphasize this point and extend the shift-share design to incorporate spatial linkages and general equilibrium effects.

[31] We estimate the city-level change using leave-one-out averages excluding region j.

[32] Our estimation method requires either $\Delta \log \tilde{A}_{jsd} \perp \{s_{js'd',0}\}_{s'd'}$, as in GSS, or $\Delta \log \tilde{A}_{\text{Seoul},sd}$ being as-good-as-randomly assigned, as in Borusyak, Hull, and Jaravel (2020).

The second set of moment conditions is

$$\Delta \log \tilde{A}_{jsd} \perp_j \Delta \log \tilde{A}^{Bartik}_{js,-d} \quad \text{for all } (s,d) \tag{14}$$

This condition imposes orthogonality similarly to the first condition, but between subsectors within a sector instead of between sectors. Likewise, the third set of moment conditions requires orthogonality across nearby regions. For each region j, we calculate the weighted average of the changes in the composite productivity of other regions, in which the weights $\varrho(j,j')$ are j-specific:

$$\Delta \log \tilde{A}^{Bartik}_{-js} = \sum_{j' \neq j} \varrho(j,j') \cdot \Delta \log \tilde{A}^{Bartik}_{j's'}.$$

In particular, we put a higher weight on region j' if this region is a closer substitute for region j. Specifically, we use the share of consumers in j who choose j' for services travel as our weight. Then, our final moment conditions can be written as

$$\Delta \log \tilde{A}_{jsd} \perp_j \Delta \log \tilde{A}^{Bartik}_{-js} \quad \text{for all } (s,d) \tag{15}$$

We estimate three parameters using these three types of moment conditions, (13)–(15), which are stacked in vector form in the following moment condition:

$$\mathbb{E}_j[\boldsymbol{m}(\boldsymbol{X}_j, \{\varepsilon, \sigma, \nu\})] = 0.$$

GMM estimates solve

$$\{\hat{\varepsilon}, \hat{\sigma}, \hat{\nu}\} \in \underset{\{\varepsilon,\sigma,\nu\}}{\operatorname{argmin}} \left(\frac{1}{J} \sum_{j \in \mathcal{J}} \boldsymbol{m}\big(\boldsymbol{X}_j, \{\varepsilon, \sigma, \nu\}\big) \right)' \mathscr{W} \left(\frac{1}{J} \sum_{j \in \mathcal{J}} \boldsymbol{m}\big(\boldsymbol{X}_j, \{\varepsilon, \sigma, \nu\}\big) \right),$$

where $\mathscr{W}$ is the efficient GMM weighting matrix. We numerically minimize the objective function to obtain GMM estimates.

Identification. Whereas the model's complexity makes it difficult to deliver a formal argument of identification, we can provide an intuitive explanation of how each type of moment condition separately identifies each of the remaining parameters. First, an exogenous increase in the composite productivity in (j,s') has a positive spillover effect on (j,s,d) for $s \neq s'$. If we postulate a weaker spillover than it actually is, the spillover alone cannot fully explain the change in the number of stores in (j,s,d). The remaining part should be explained by an increase in the composite productivity of (j,s,d), which results in a spurious positive correlation between $\Delta \log \tilde{A}_{jsd}$ and $\Delta \log \tilde{A}^{\text{Bartik}}_{js'}$. Holding fixed the value of the calibrated parameter β, the parameter ε mainly controls the magnitude of this spillover effect. Thus, the first set of moment conditions requires selecting ε in such a way that these terms are uncorrelated across

sectors. Similarly, an exogenous increase in the composite productivity in $(j, s, -d)$ has both a spillover effect and a negative competition effect on (j, s, d). Holding fixed the spillover effect controlled by β and ε, if we assume too small competition effects, a spurious negative correlation arises between $\Delta \log \tilde{A}_{jsd}$ and $\Delta \log \tilde{A}^{\text{Bartik}}_{js,-d}$. The parameter σ mainly controls the magnitude of this competition effect. Therefore, the second set of moment conditions requires that, with β and ε held fixed, the parameter σ is chosen to render these terms uncorrelated across subsectors. Finally, an exogenous increase in the composite productivity in a given zone j has both a positive spillover effect and a negative competition effect on nearby zones, $-j$. Holding fixed the spillover effect again, the parameter ν mainly controls the magnitude of the spatial competition. Thus, the third set of moment conditions requires that, with β, ε, and σ held fixed, the parameter ν needs to be selected to ensure that these terms are uncorrelated across zones.

Estimation Results. Table 3 summarizes estimation results.[33] First, we find that the estimated degree of external economies of scale ε is not significantly different from zero. This finding suggests that the trip-chaining mechanism dominantly accounts for spillovers in the services market, and leaves limited room for other mechanisms to contribute significantly. Although it is difficult to find a directly comparable estimate, this estimate stands in stark contrast to the tradable goods sector, for which the literature extensively documents evidence of the presence of strong scale economies, which contribute to the agglomeration of industrial production. This literature emphasizes several mechanisms at play, such as sharing, matching, and learning, but our result indicates that these mechanisms play a limited role in the services market.[34]

We find that the dispersion of taste shocks ν is about 0.35. Our estimation results suggest that consumers' idiosyncratic preferences are more dispersed than their preferences for residence or workplace choices, which are estimated by Ahlfeldt et al. (2015). Finally, our estimate of substitution across subsectors σ is about 4.8. This estimate is comparable to that of Couture (2016), who estimate the elasticity of substitution across different types of restaurant cuisine.

[33] In this section, we use data from the years 2014, 2015, and 2018, which is different from the data used in Section 2.2, where we used data from the years 2014, 2015, and 2019. In 2019, the data sources for constructing sales estimates were changed, and we find that this caused noise that differs across subsectors. The number of stores does not have the same issue, since it has been consistently collected. The estimation results in Section 2.2 would remain qualitatively similar if we had used the year 2018 instead.

[34] This result is perhaps not surprising, given the distinctive features of the services market. First, mechanisms based on the relationship between firms, such as input-output linkages, are not relevant for the services sector because services firms mostly cater to households rather than other services firms. Moreover, the geographic unit of analysis in our study is much smaller than those used in the literature, so the agglomeration mechanism for services should have a higher rate of spatial decay. In addition, the mechanisms discussed in the literature, such as comparison shopping, tend to operate within a sector rather than across sectors, which is inconsistent with our findings of across-sector spillovers in Section 2.2. In contrast, our trip-chaining mechanism provides a natural explanation for across-sector spillovers with a high rate of spatial decay. This arises from consumers' disutility from travel, which is highly sensitive to distance, and from trip chaining, which combines purchases from different sectors.

Table 3: Estimation Results

	Description	Value	Source
$\frac{\tau}{\nu}$	Services travel elasticity	0.152 (0.002)	Gravity
$\frac{\varphi}{\nu}$	Services travel border effects	0.357 (0.064)	Gravity
β	Travel continuation probability	0.419 (0.006)	Online Survey
ε	External economies of scale	0.006 (0.045)	GMM
σ	Substitution across subsectors within a sector	4.851 (0.131)	GMM
ν	Dispersion of taste shocks	0.351 (0.000)	GMM
ρ	Substitution across stores within a subsector	9	Couture (2016)
γ	Share of firm expenditure on floor space	0.2	Valentinyi and Herrendorf (2008)
α_s	Expenditure share on Food, Retail, Other	0.31, 0.51, 0.18	Revenue shares
μ_r, μ_w	Share on services from home and workplace	0.209, 0.062	Spending share
μ_ℓ	Share on housing	0.25	Literature

Notes: Standard errors from the gravity equation estimation and efficient GMM estimation are in parentheses.

2.4.2 Estimation Results: SMA Inequality

Equipped with the estimated model, we can compute each region's services market access (SMA), which represents the value that consumers in each region derive from services travel. We can compute SMA only after estimating the model, because SMA depends not only on the spatial distribution of services stores but also on a number of key parameters—travel cost parameters, elasticities of substitution, and, most importantly, the trip-chaining parameter. Together, they map the spatial distribution of services stores to the spatial distribution of SMA.

In the left panel of Figure 5, we plot the histogram of (log) SMA. The standard deviation of (log) SMA across zones is 0.22, which indicates limited but nonnegligible variation across regions. For example, a consumer who begins their services travel in the top 25% zone can enjoy 33% higher welfare per spending compared with those who begin their services travel in the bottom 25% zone.[35] As an alternative measure of inequality, we plot the Lorentz curve for SMA in the right panel of Figure 5. The corresponding Gini coefficient is 0.12.

However, the dispersion of the number of services stores is much larger than that of the SMA. In the middle panel, we plot the histogram of the (log) number of services stores, which has a much thicker right tail. The standard deviation of the (log) number of stores is 1.03, which is 10 times larger than that of the SMA. In addition, the interquartile ratio and the Gini coefficient are 3.29 and 0.56, respectively, both of which are substantially larger than

[35] For all statistics, we use the effective population distribution—i.e., the weighted sum of residence and population distribution—and assign 77% weight to the former based on the share of travel starting from home.

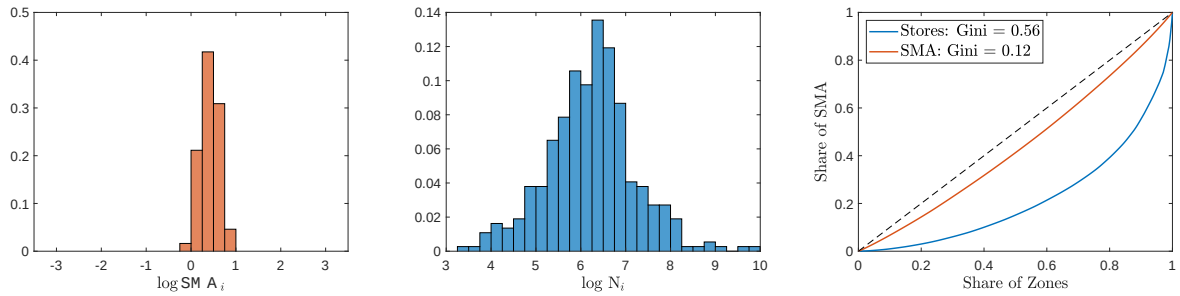


Figure 5. Spatial Disparity in the Number of Stores and SMA

those of the SMA.[36] Thus, we conclude that inequality in access to the services market is significant, but less than expected from the uneven distribution of stores. The gap between these inequalities arises from the possibilities of services travel and trip chaining: consumers can travel to other regions and make multiple purchases in regions with many stores without incurring additional travel costs.

In Appendix B.6, we show that SMA inequality exacerbates real income inequality between high-skilled and low-skilled workers. This is because high-skilled workers tend to live and work in areas with better access to the services market. Since consumers allocate a significant portion of their income (27%) to services goods, SMA—the inverse of the price index of services goods—has a significant impact on the price index they face. Indeed, the impact of SMA inequality on real income inequality is substantial and comparable to that of housing rents.

2.5 Importance of Spillovers from Trip Chaining

In this section, we examine the importance of spillovers from trip chaining. We first explore their importance in agglomeration of non-tradable services stores. We then explore the welfare implications, focusing on the distinctive features of trip chaining.

2.5.1 Spillovers and Agglomeration of Services

To investigate the importance of spillovers from trip chaining in the agglomeration of non-tradable services, we set β to 0 to turn off trip chaining and calculate the concentration of the counterfactual distribution of services stores. The result indicates that trip chaining plays a substantial role in agglomeration of services. In the left panel of Figure 6, we present a scatter plot that compares the log number of services stores with and without trip chaining, and in the right panel, we depict the Lorenz curves for the distributions of services stores. Without trip chaining, the number of stores in concentrated areas decreases significantly. The dispersion of services, as measured by the standard deviation

[36] The Gini coefficient of household income in Korea is 0.314 (World Bank, 2016).

of the log number of stores, decreases by 35% (from 1.03 to 0.67).[37] Similarly, the right panel shows that the Lorenz curve of services stores shifts significantly, and the Gini coefficient decreases by 40%. This suggests that more than one-third of the concentration of services is attributable to spillovers arising from the trip-chaining mechanism.

In contrast, we find that external economies of scale have limited impact on the concentration of services, as expected from the estimate of ε being close to zero. To examine the role of external economies of scale, we set parameter ε to zero, along with trip-chaining parameter β. This eliminates all spillover forces in the services sector. As shown in Figure 6, this causes little change to the counterfactual distribution of services stores (represented by the red cross markers) and the Lorenz curve (red line). The standard deviation of the log number of stores decreases by only an additional 2%. Therefore, we conclude that spillovers explain more than one-third of the concentration in services—and of the two spillover mechanisms, the trip-chaining mechanism accounts for about 95% of the total contribution of spillovers.

What explains the remaining 63% of the concentration? We find that location fundamentals explain about one-half of the rest, or about one-third of the total concentration. If we also turn off the regional differences in location fundamentals—composite productivity and costs—the standard deviation of the log number of stores decreases by an additional 29%. In the left panel of Figure 6, the counterfactual distribution represented by the yellow squares is much less dispersed. Finally, the remaining 34% of the concentration arises from differential access to consumers, which stems from the combination of the distribution of effective consumers E_i and spatial frictions. If we further assume that there are no spatial frictions in this economy, all services stores have the same advantage in terms of proximity to consumers, regardless of their locations. Without differences in fundamentals or access to consumers, the dispersion of services stores disappears completely, as shown by the purple dots in Figure 6.

The Lorenz curves in the right panel of Figure 6 confirm our results. As we turn off each channel one by one, the curves approach the 45-degree line. The Gini coefficients decrease to 60%, 59%, 35%, and 0% of the baseline, respectively.

2.5.2 Welfare Implications of the Trip-chaining Mechanism

In Section 2.3.3, we discuss the efficiency implications of the trip-chaining mechanism. In particular, we show that the spillovers that arise from trip chaining do not lead to any inefficiency in the decentralized economy, whereas external economies of scale are generically inefficient.

In this section, we turn our attention to the effect of trip chaining on SMA inequality. As discussed in Section 2.5.1, trip chaining increases the concentration of services stores and thus contributes to higher SMA inequality. However, its

[37] In Sections 2.5 and 2.6, we report the result fixing the distribution of consumers, and focus on reallocation of the services sector. Results are both qualitatively and quantitatively similar when we use the general equilibrium model to perform counterfactual exercises. For example, turning off the trip-chaining mechanism leads to a 34% decrease in the standard deviation of the log number of stores when we further endogenize decisions on residential areas and workplaces, which is only 1 percentage point smaller.

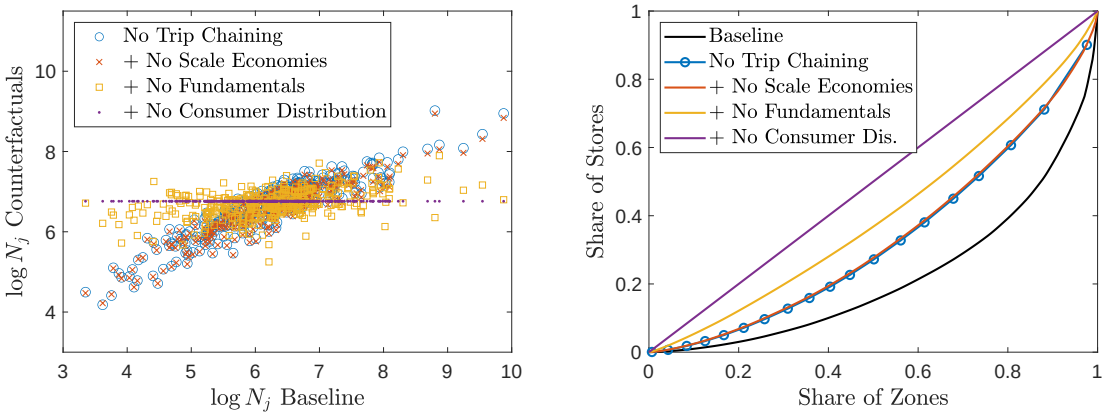


Figure 6. Importance of Trip Chaining in Agglomeration

Notes: We turn off each mechanism sequentially on top of the previous one. For example, scale economies represent an economy in which both the trip-chaining mechanism and external economies of scales are turned off.

total effect on SMA inequality is ambiguous due to a countervailing force. Trip-chaining behavior allows consumers to make multiple purchases per travel, leading to a lower disutility cost per purchase. This makes it less costly to live or work in a region with fewer services stores nearby. Thus, holding the spatial distribution of services stores fixed, trip-chaining behavior itself reduces SMA inequality.

Despite the significant changes in the concentration of stores we documented above, we find that the net effect of the trip chaining on SMA inequality is close to zero due to the countervailing force. In the left panel of Figure 7, we plot for each zone the counterfactual value of the (log) SMA when there is no spillover from trip chaining against the actual value (with blue circles). The figure clearly shows that trip chaining has a limited impact on dispersion of the SMA. SMA inequality, as measured by the standard deviation of the (log) SMA, slightly increases by 4.5%. The right panel also confirms this finding, by showing that the Lorenz curve and the Gini coefficient barely change.

To decompose the effects of the two opposing forces, we first isolate the effect of changes in the distribution of stores. We compute SMA inequality using the counterfactual distribution of services stores without trip chaining, but still allowing consumers to make multiple purchases per travel. These results are represented by the red squares in the left panel of Figure 7, which shows that a lower concentration of services stores leads to a decrease in inequality. The standard deviation of (log) SMA becomes 12.5% smaller than the actual value.

Our analysis suggests that the trip-chaining mechanism does not exacerbate SMA inequality, despite its substantial contribution to the concentration of services stores. This finding again highlights the importance of identifying the mechanisms that derive agglomeration, since their welfare implications may differ substantially. For example, if agglomeration arises from external economies of scale rather than the trip-chaining mechanism, the countervailing force related to changes in travel patterns does not operate. In such cases, spillovers always lead to greater inequality

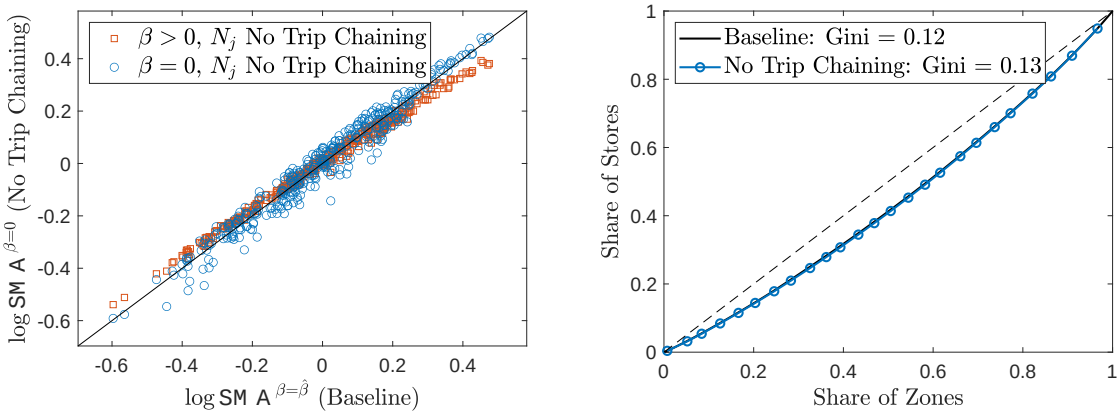


Figure 7. Importance of Trip Chaining in SMA

in access to the services market, and an increase in concentration always goes hand in hand with an increase in SMA inequality.

2.6 Urban Structure in the Future

Consumption services can play a more important role in the success of cities than production (Glaeser, Kolko, and Saiz, 2001). Couture and Handbury (2020) demonstrate that non-tradable services have been a driving force behind recent urbanization in the United States. Our model provides insights into how the urban structure, particularly the distribution of services, may change in the future due to various factors such as policy changes, transportation networks, or technological advances. We examine the impact of the rise of work from home and delivery technology, which have unexpectedly been accelerated by the COVID-19 pandemic.

2.6.1 Work from Home

The COVID-19 pandemic has changed the way people work: Between April and December 2020, about one-half of paid work hours in the US were supplied from home. This shift is not temporary, and research suggests that work from home will remain at around 20% (Barrero, Bloom, and Davis, 2021). Similarly, in Seoul, the proportion of remote or hybrid work doubled from 4.5% to 9% between 2016 and 2018, and the pandemic has further accelerated this trend, with around 18% of workers experiencing remote work in 2022.

Studies provide evidence that work from home may have significant impacts on the distribution of services. However, its impacts to date have been uneven across cities, and its long-term consequences remain uncertain. While services stores in large US cities become less spatially concentrated, shifting from dense city centers to suburban areas (e.g., Althoff et al., 2022; Duranton and Handbury, 2023; Duguid et al., 2023), Seoul did not experience a decrease in the concentration of services stores during the pandemic. According to Seoul Commercial Area Data, from 2019 to 2022, about one-half of the top 10% zones with the largest share of working population experienced growth above the citywide median level. In addition, the standard deviation of the log number of stores across zones did not decrease, but instead increased by 8%. Which characteristics of cities determine the impact of work from home? And furthermore, how will it reshape the distribution of services in the long run?

The spatial linkages of services consumption between residential and business areas are critical in understanding the impact of the rise of work from home. As workers shift to working remotely, the distribution of consumers also shifts from business areas to residential areas, which are typically less concentrated in the data. However, it remains unclear whether this shift will actually lead to a reduction in the concentration of services, since a significant proportion of purchases involve services travel. If spatial linkages between residential and business areas are strong, consumers may still travel for services from their homes to business areas while working from home.

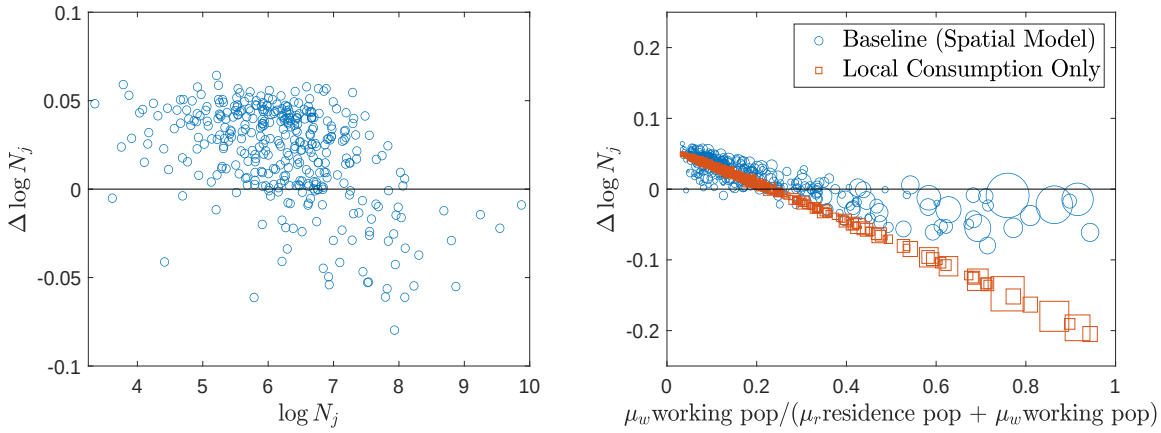
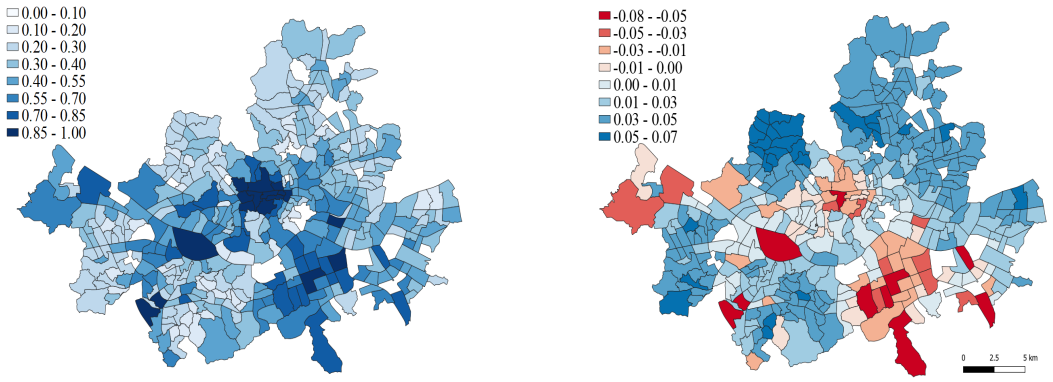


Figure 8. Changes in the Number of Stores after Work from Home

To analyze the long-run effects of work from home, we assume that 20% of workers will work remotely in the future, as predicted by Barrero, Bloom, and Davis (2021), and compute the counterfactual distribution of services stores. When working from home, consumers always start their services travel from home rather than from their workplace. In the left panel of Figure 8, we plot the percentage change in the number of stores after the rise of work from home against the current number of stores. Although concentrated areas tend to experience a decline in the number of stores, the magnitude of the change in concentration is limited, which is qualitatively consistent with the empirical pattern we document above. The standard deviation of the log number of stores decreases by only 1.7%.

We find that the limited impact on services concentration is due to the strong spatial linkages between business and residential areas. Specifically, the demand for services in certain concentrated business areas remains high because they continue to attract a significant number of consumers even when they work from home. This is confirmed in the right panel of Figure 8, in which we plot (with blue circles) the change in the number of stores against the share of the workplace population. The figure shows that regions with a high share of the workplace population do not necessarily experience significant declines in the number of stores. Although some zones with a workplace population share above 70% experience a significant decrease in the number of stores (about 10%), many of these zones experience only a minor decline (less than 1%). To further illustrate this point, Figure 8 also plots (with red circles) the effect of work from home on the number of stores in a model in which consumers purchase services only in the zone in which they begin their services travel. In this case, we find a strong relationship between the share of the workplace population and the change in the number of stores.

The impact of spatial linkages can clearly be seen in Figure 9. We plot the share of the workplace population and the change in the number of stores on the map of Seoul. We can compare the two largest business areas in Seoul, *Jong-ro* in the north and *Gangnam* in the southeast (see Figure 9a). Although Jong-ro has a higher share of the workplace population than Gangnam, it experiences a smaller decline in the number of services stores after



(a) Share of the Workplace Population

(b) Change in the Number of Stores

Figure 9. Work from Home: Map

the rise of work from home (see Figure 9b). Jong-ro has stronger spatial linkages with its surrounding residential areas. As a result, it continues to attract consumers from nearby regions, which offsets the decline in demand from local consumers who work in the area. In contrast, Gangnam, which is surrounded by several business areas, faces challenges in attracting consumers when people work from home.

2.6.2 Delivery Services

As transportation and internet technology continue to improve, the importance of delivery for non-tradable services has grown rapidly. Consumers in the U.S. can now buy retail goods online (e.g., from Amazon) and order food for home delivery (e.g., Uber Eats). South Korea has also experienced significant growth in delivery services in the Food and Retail sectors. Online retail sales in South Korea almost doubled from $94 billion to $150 billion between 2017 and 2021. The share of restaurant sales that are delivered to consumers has also grown rapidly in recent years, and accounts for 15% of total sales (Statistics Korea, 2022). Delivery services eliminate the disutility from distance, which renders non-tradable services effectively tradable. Thus, delivery technology can have a significant impact on both the spatial distribution of services firms and the welfare of consumers.

To examine the impact of delivery services, we consider a counterfactual scenario in which $\theta \in [0, 0.5]$ fraction of the total demand for the Food and Retail sectors is fulfilled by delivery. We assume that spatial frictions are completely eliminated in the Retail sector and reduced by 50% in the Food sector when services are delivered. In particular, we reduce the distance disutility parameters (φ, τ) to either zero or to half of the baseline values, and consumers purchase a single services good at a time.[38]

[38] For the restaurant sector, spatial frictions decrease but still exist even when delivery is available because delivery platforms typically charge fees based on distance. We choose 50% as an approximation, but the results are qualitatively the same independent of a specific number we choose.

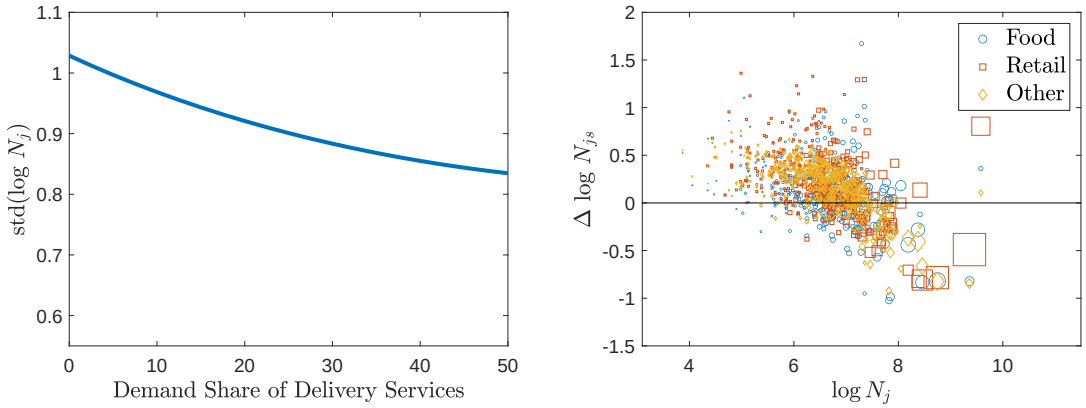


Figure 10. Delivery Services: Concentration of Services Stores

Improvements in delivery technology lead to a substantial reduction in concentration. In the left panel of Figure 10, we plot how the concentration decreases as we increase the share of delivery services. For example, when one-half of total demand is fulfilled by delivery ($\theta = 0.5$), the standard deviation of the log number of stores decreases by 19%. Advances in delivery technology help services stores located in unfavorable geographic locations by reducing spatial frictions. For example, stores located in remote areas that were previously connected to a relatively small consumer base can now access the entire market through better spatial linkages.

In the right panel of Figure 10, we plot the change in the number of stores in each region for each sector. With the increasing use of delivery services, some of the demand for the Food and Retail sectors is now being fulfilled by these services, which reduces the importance of the geographic location of stores for revenues. As a result, stores in previously concentrated areas lose their comparative advantage, which leads to a decline in the number of restaurants and retail stores in such regions. Interestingly, the decline in concentration is also observed in the Other sector, which is not directly affected by delivery services. This unique phenomenon is a consequence of trip chaining, which generates positive spillovers across sectors within regions. As consumers visit concentrated areas less often to eat out or shop for retail goods due to the rise of delivery services, subsequent purchases from trip chaining decrease. This reduces demand for the Other sector in these areas, which leads to a decrease in the number of stores.

Finally, the rise of delivery services also leads to a significant decrease in welfare inequality. Welfare inequality, as measured by the standard deviation of log SMA, decreases by 36%. Access to delivery technology eliminates the disutility of travel, which disproportionately benefits consumers in remote areas with previously lower SMA. This suggests that policymakers should consider investments in delivery services or transportation technologies for nontradable goods as an effective means of reducing inequality in access to the services market. This is particularly important given the significant impact of SMA inequality on real income inequality, as discussed in Section 2.4.2.

Chapter 3

Persistent Noise, Feedback, and Endogenous Optimism: A Rational Theory of Overextrapolation

3.1 Introduction

How do agents form expectations based on the information they have? Can waves of optimism and pessimism play a role in driving the economy? Several strands of macroeconomic literature provide insights into incorporating the role of optimism in business cycle models, focusing on either behavioral changes in beliefs or exogenous shifts in rational expectations, such as sunspots, noise shocks in public signals, and sentiment shocks.

In this paper, we capture optimism in an entirely different yet rational way. Agents learn about economic conditions from signals, but they are uncertain about how to interpret these signals. They are said to be optimistic if they interpret their signals in an optimistic way. Agents try to correct their optimism rationally, and in the presence of strategic interaction between agents, they try to learn the optimism of others as well. In this process, optimism is endogenously determined by the dynamic path of fundamentals and acts as an amplification or dampening mechanism.[1]

The key to our mechanism is the assumption of persistent noise. Agents observe noisy signals about economic conditions, but in contrast to the literature, we assume that noise terms are persistent rather than *i.i.d.* over time. Agents rationally account for the persistence of noise and update their beliefs about noise based on ex post observations of true economic conditions. I show that this process gives rise to a novel mechanism by which optimism arises endogenously. In particular, when agents observe better-than-expected economic conditions, they update their expectation of the noise term downward, implying that they interpret their future signals more optimistically.

[1] Similarly, Angeletos and Lian (2019) study a confidence multiplier that varies endogenously and amplifies demand shocks.

To illustrate the main idea, consider firms that need to forecast market demand for their products in order to set prices or make production plans. Firms observe noisy signals about market demand at the beginning of each year, for example, from consumer surveys. Firms also receive feedback on their prior forecast by observing the realized market demand ex post through the sale of their products. Suppose that there is a positive shock to market demand, but firms were not well informed about it from their consumer surveys. Then, realized market demand is higher than firms expected, leading them to believe that they were too pessimistic in interpreting the survey results. This, in turn, makes them overly optimistic when they make new forecast based on a new survey result. In this way, the effect of the positive shock propagates into the next year.

We begin by calling into question the validity of the *i.i.d.* noise assumption commonly made in the literature and provide two arguments in favor of persistent noise. We model this persistence by assuming that the noise terms follow an AR(1) process. Rational agents then take this persistence into account when forming their expectations and try to learn the noise terms in order to best interpret their signals. Their beliefs about the noise terms affect their forecast and hence their actions. We then formally define the notion of optimism. If agents underestimate the noise terms in their signals, they will overestimate the fundamentals for given values of the signals, so we call them optimistic.

We then introduce a macroeconomic noisy rational expectations model in which firms' output choices are made under dispersed information about their productivity. We characterize how firms dynamically learn persistent noise terms in their signals and how this novel channel of learning endogenously generates optimism, which either amplifies or dampens the underlying productivity shock. We assume that firms receive feedback on their previous forecast by observing the actual productivity realization. We assume that the productivity shock consists of two components. Firms receive noisy signals about the first *partly-observed* component and choose their output levels. However, true productivity is also affected by the second *unobserved* component. This distinction formalizes the idea that there are many shocks in the real world that differ in the extent to which economic agents are informed about them.[2] Our main finding is that, with persistent noise, the effect of partly-observed shocks on the next period's output is dampened because they make firms pessimistic, while the effect of unobserved shocks is amplified because they make firms optimistic. Following the literature, we further assume that firms are relatively well informed about idiosyncratic shocks, so that partly-observed shocks correspond to idiosyncratic shocks while unobserved shocks correspond to aggregate shocks. In this case, aggregate optimism fluctuates procyclically with the underlying aggregate shocks, and aggregate output exhibits delayed overreaction to aggregate productivity shocks, as in Angeletos, Huo, and Sastry (2020).

[2] For example, firms make decisions after receiving noisy information about some shocks that affect market demand, but realized market demand is also affected by some other shocks. Firms may be relatively well informed about shocks that are realized before they make forecasts, or about shocks that are idiosyncratic to firms, which may then be well reflected in the consumer survey. In contrast, firms are likely to be less well informed about shocks that are realized after the forecasts have been made, or about aggregate shocks, such as shocks to aggregate demand.

Next, we introduce strategic complementarity in our baseline model. This gives firms an incentive to forecast others' optimism, others' beliefs about others' optimism, and so on (i.e., higher-order optimism). We characterize how firms update their higher-order optimism and how this higher-order optimism in turn affects firms' output choices through its effect on higher-order beliefs about aggregate productivity. We first show that the main result continues to holds in the extended model with strategic complementarity: the effect of partly-observed shocks is dampened while the effect of unobserved shocks is amplified. Moreover, we show that the presence of strategic complementarity and the resulting higher-order optimism always strengthen our mechanism relative to the baseline model. In other words, when firms observe better than expected economic conditions, they become optimistic not only about their signals (first-order optimism) but also about others' optimism (higher-order optimism). This result is in stark contrast to the literature, which instead documents that the higher the degree of strategic complementarity is, the less responsive agents are to the underlying shocks.

Finally, we turn to the empirical content of our theory: how can we interpret forecast data through the lens of our framework? There is a large body of empirical work that uses survey forecast data to measure agents' expectations directly. This literature often assumes that forecasters do not observe past realizations, even ex post. This is partly because observing past realizations makes the problem essentially static in their setting and makes it difficult to explain what we can observe in the dynamic forecast data. Our theory provides a completely different way of interpreting the forecast data. We view these forecast data as the results of agents' dynamic learning, in which they can observe past realizations but are trying to learn how to interpret their own information, i.e., noise terms. Our model explains prominent empirical findings in the literature, including Coibion and Gorodnichenko (2015); Kohlhas and Walther (2020); and Angeletos, Huo, and Sastry (2020). However, many other standard models can also explain these findings, especially behavioral theories of overextrapolation combined with information frictions. To distinguish our model from these theories, we exploit the difference between the degree of overextrapolation in consensus and individual forecasts. In the IBES dataset, analysts' expectations for earnings growth exhibit overextrapolation from past realizations only when we aggregate them into consensus forecasts. This is consistent only with our rational theory of overextrapolation.

The rest of the paper is organized as follows. Section 3.2 justifies our assumption of persistent noise and formalizes the notion of optimism. Section 3.3 develops a macroeconomic model without strategic complementarity. We characterize the learning of firms and see how our mechanism amplifies or dampens the underlying productivity shocks. In Section 3.4, we use an extended model to explore further implications when strategic complementarity is present. Section 3.5 discusses the new interpretation of dynamic forecast data and provides empirical evidence that is suggestive of our theory. Section 3.6 concludes.

3.2 Persistent Noise Terms and Optimism

In this section, we introduce the key element of our theory—persistent noise terms in signals—and illustrate how it naturally leads to a definition of optimism. The literature that investigates the role of expectations and information often postulates that agents receive noisy signals about the true state of nature (hereafter, fundamental) and that they know the stochastic relationships between the signals and the fundamental. For example, signals are often modeled as the fundamental plus a random noise term with a known distribution:

$$s_t = a_t + \xi_t$$

where s_t is the signal about the fundamental a_t at time t, and ξ_t is the noise term. This assumption is just a modeling device that captures the idea that we are partly informed about the true state of the world while preserving the tractability of the models. Most of the papers in the literature, however, assume that noise is a random variable independent over time.[3] This time-independence assumption greatly simplifies the analysis, and, combined with normality assumptions, often leads to closed-form solutions.

Whether this *i.i.d.* assumption is a good or bad description of the real world depends on how we interpret the noisy signal; i.e., what the real-world counterpart of the noise is. There are two prominent interpretations in the literature, and we will argue in this section that whichever interpretation we adopt, *persistent* noise is a more realistic assumption. We then discuss the implications of this persistence in the following sections.

First, we can literally interpret the signal as noise-ridden information about the fundamental, and agents directly observe this signal. In the real world, information sources are always biased, and the bias is likely to be persistent over time. Agents, however, do not know the exact value of this bias. This introduces an additional component of noise, which makes the perceived noise also likely to be persistent.[4]

[3] One crucial reason for this assumption is that with persistent variables that cannot be observed perfectly, we have to tackle the infinite regress problem as in Townsend (1983). A large number of works have explored how to solve the infinite regress problem using either guess-and-verify, approximation, or the frequency domain technique. A partial list of these works includes Sargent (1991), Kasa (2000), Nimark (2017), Rondina and Walker (2018), and Huo and Takayama (2018). Even in those works, it is often assumed that only fundamentals are persistent, while noise follows *i.i.d.* Notable exceptions are Huo and Takayama (2015, 2018), but their main focuses are on the methodological contribution rather than on economic implications of persistent noise terms.

[4] This makes no difference in static settings as it only adds another layer of uncertainty about the stochastic relationship. To see this point, consider a signal $s = a + \xi$ subject to bias, $\xi = bias + e$. An agent believes that the bias is distributed as $bias \sim F(\cdot)$ and that the (unbiased) error term is distributed as $e \sim G(\cdot)$. In static settings, it is isomorphic to the case in which the agent believes that the noise term ξ follows a *known distribution*

$$P(\xi \leq \hat{\xi}) = \int F(\hat{\xi} - e) dG(e).$$

In contrast, this additional layer of uncertainty is important in dynamic settings. If the bias in an information source is persistent, agents then have incentives to correct this bias over time. To see this clearly, consider a dynamic extension in which the agent uses the signal $s_t = a_t + \xi_t$ to form expectations about a_t for two periods $t = 0, 1$. Suppose we assume that $\xi_t = bias + e_t$ so that the bias is time-invariant. Assume further that the agent can observe the true realization of a_0 at the beginning of period 1. The agent then makes another prediction about a_1 after observing s_1. Then, the agent tries to correct the bias by comparing the previous signal s_0 with the realization a_0.

The second interpretation comes from the literature on rational inattention. Consider a slightly generalized version of the attention problem studied by Sims (2003) and Mackowiak and Wiederholt (2009):

$$\begin{aligned} \min_{b_0, b(\cdot), c_t(\cdot)} \quad & \mathbb{E}\Big[\big(\mathbb{E}[a_t|s^t] - a_t\big)^2\Big] \\ \text{s.t.} \quad & \mathcal{I}(\{s_t\};\{a_t\}) \leq \kappa \\ & a_t = \rho_a a_{t-1} + \varepsilon_t \\ & s_t = b_0 + b(L)\varepsilon_t + c_t(L)\tilde{\xi}_t \end{aligned}$$

where ε_t and $\tilde{\xi}_t$ follow independent Gaussian white noise processes.[5] The decision maker chooses a signal process s_t to forecast a_t, subject to a constraint on the information flow between $\{s_t\}$ and $\{a_t\}$, which sets an upper bound for

$$\mathcal{I}(\{s_t\};\{a_t\}) \equiv \lim_{T\to\infty} \frac{1}{T} I(s_1, \cdots, s_T; a_1, \cdots, a_T)$$

where $I(\cdot;\cdot)$ denotes the mutual information. Sims (2003) and Mackowiak and Wiederholt (2009) show that it is without loss to assume that the decision maker makes a forecast after observing a signal of the form "true state plus a time-independent noise term:"

$$s_t = a_t + \sigma \cdot \tilde{\xi}_t$$

where σ is a constant. In other words, such signals are optimal when agents can choose an information structure under the above constraint.

This result provides an elegant justification for the *i.i.d.* assumption. It is, however, not robust in the sense that it depends crucially on the precise form of the information constraint. For example, when we consider other forms of information constraints such as

$$I(s_t; a_t) \leq \kappa, \quad \forall t, \tag{1}$$

then we can easily show that agents can always be better off by inducing correlations in their signals.

Lemma 1. *The signals with i.i.d. noise terms, $s_t = a_t + \sigma \cdot \tilde{\xi}_t$, cannot attain the minimum of the following attention problem*

$$\begin{aligned} \min_{b_0, b(\cdot), c_t(\cdot)} \quad & \mathbb{E}\Big[\big(\mathbb{E}[a_t|s^t] - a_t\big)^2\Big] \\ \text{s.t.} \quad & I(s_t; a_t) \leq \kappa, \quad \forall t \\ & a_t = \rho_a a_{t-1} + \varepsilon_t \\ & s_t = b_0 + b(L)\varepsilon_t + c_t(L)\tilde{\xi}_t. \end{aligned}$$

[5] This is a generalized version as we consider $c_t(\cdot)$ instead of a time invariant function $c(\cdot)$.

Instead, signals with persistent noise terms, $s_t = a_t + \sigma \cdot \tilde{\xi}$ where $\tilde{\xi} \sim \mathcal{N}(0,1)$ attain the minimum.

Proof. All proofs are in Appendix C.1. □

The intuition is simple: under this information constraint, agents can make noise terms correlated across time periods *for free*.[6] By doing so, however, agents can dynamically learn and correct for the persistent component of the noise.[7] Of course, there is no a priori reason why the information constraint in the real world should be given by (1). However, there is also no reason to prefer the original information constraint of Sims (2003) and Mackowiak and Wiederholt (2009). Thus, Lemma 1 gives us a takeaway that unless the information constraint in the real world is exactly the same as that postulated by Sims (2003) and Mackowiak and Wiederholt (2009), decision makers would optimally choose correlated noise terms in order to exploit their ability to correct the persistent component of the noise terms over time.

Illustrative Example. We have argued that regardless of how we interpret the noise term, its counterpart in the real world is likely to be correlated across time periods. To further illustrate this argument, consider a forecaster who makes predictions about, say, the annual US inflation rate. There are various information sources she can use to make her prediction, but suppose that she only gets information from newspapers. There are N newspapers available, each of which is informative about the inflation rate. The forecaster chooses to subscribe to a subset of the newspapers while being subject to an attention cost. Thus, these newspapers can be thought of as information sources in Myatt and Wallace (2004) and Pavan (2016). First, note that if a particular newspaper gives a positively biased view of the inflation rate this year, then it is likely to give a view biased in the same direction next year. Consider a constraint on the attention cost which requires the forecaster to read at most, say, three newspapers each year. Under this constraint, which is analogous to constraint (1), the forecaster would optimally choose to read the same set of newspapers each year because she can at least partly correct the bias in the newspapers by herself. Since the bias in each newspaper is persistent, and the forecaster would choose to read the same newspapers, it is as if she were receiving a noisy signal whose noise term is correlated across time periods. In contrast, under a constraint analogous to the original constraint in Sims (2003) and Mackowiak and Wiederholt (2009), if she reads three newspapers this year, then it would be strictly more costly (in terms of cognitive cost) to read the same set of newspapers next year because she can obtain strictly more information from those newspapers. Thus, we reach the counterintuitive conclusion that it can be optimal to read three *new* newspapers each year.

[6] This is costly in the original formulation of Sims (2003) and Mackowiak and Wiederholt (2009).

[7] This is obvious when agents receive feedback later on so the objective function changes to $\mathbb{E}\left[\left(\mathbb{E}\left[a_t|s^t, a^{t-1}\right] - a_t\right)^2\right]$. However, this also holds without such feedback because agents can ultimately learn the precise value of $\tilde{\xi}$ after observing an infinite number of signal realizations.

Remark. To fix ideas, we talked about biased information sources and agents who try to learn the bias. In the real world, however, information sources are not only biased, but also lack a pre-specified way of interpreting them. For example, if you observe that the current unemployment rate is 5%, how can you interpret that number as a signal about the current inflation rate? It is indeed informative about the inflation rate to some extent, but it is not like observing a random variable distributed around the inflation rate, as we assumed. Forecasters need to interpret the information sources they have, but they are uncertain about how to interpret them. Thus, the forecaster in the previous example can also be viewed as the one who tries to correct the way she interprets newspapers. In this sense, "learning noise terms" in this paper also means "learning how to interpret information."

We have argued so far that it is natural to assume persistent noise terms. From now on, we model this persistence by simply assuming that agents observe a noisy signal about the fundamental a_t:

$$s_t = a_t + \xi_t$$

whose noise term ξ_t is autocorrelated[8]

$$\xi_t = \rho \xi_{t-1} + \eta_t.$$

This is only a small departure from the literature and enables us to maintain tractability.

The persistence of the noise terms naturally leads to a formal definition of optimism. A crucial difference from the model with *i.i.d.* noise terms is that agents try to learn the noise terms ξ_t in their signals. Agents' expectations about the noise terms affect how optimistically they interpret the signals, which in turn affects their forecasts and hence their decisions. We consider two information sets $\tilde{\Omega}_t$ and Ω_t, which are the information sets of an agent right before and right after observing a signal s_t, respectively. From the perspective of an outside observer, if an agent underestimates her noise term ξ_t, then for a given realization of her signal s_t, she would overestimate the fundamental a_t. Therefore, an agent is *optimistic* in interpreting her signal s_t if her belief about the noise term ξ_t is *lower* than its true value. This discussion leads to the following definition of (first-order) optimism.[9]

Definition 1. *An agent is said to be* ex-ante (ex-post, respectively) optimistic *if she underestimates the noise term in her signal:*

$$\mathbb{E}[\xi_t|\tilde{\Omega}_t] < \xi_t \quad (\mathbb{E}[\xi_t|\Omega_t] < \xi_t, \textit{ respectively})$$

[8] Throughout the paper, we will use the letter ξ to denote noise terms.

[9] In Section 3.4, we will define higher-order optimism when there are multiple agents who play a game with strategic complementarity.

The ex-ante (ex-post, respectively) optimism *of an agent is defined as the extent to which she underestimates the noise term:*

$$\tilde{\mathcal{O}}_t \equiv \xi_t - \mathbb{E}[\xi_t|\tilde{\Omega}_t] \quad (\mathcal{O}_t \equiv \xi_t - \mathbb{E}[\xi_t|\Omega_t], \textit{ respectively})$$

Thus, ex-ante optimism captures how optimistic an agent is before observing a signal, and ex-post optimism captures how optimistic she is in interpreting a realized signal. In the next section, we will see how agents' dynamic learning about the persistent noise terms endogenously generates optimism, and how the resulting optimism affects equilibrium outcomes.

3.3 The Baseline Model

We begin with a macroeconomic model in which firms' output choices are made under incomplete information about their productivity. The model structure is closely related to the island model of Angeletos and La'O (2009b) and Benhabib, Wang, and Wen (2015), but the timeline is more closely aligned with Kohlhas and Walther (2020). In this section, we consider specific parameter values under which there is no strategic complementarity between firms' decisions. This allows us to focus on how firms learn about persistent noise terms in their *own* signals. Compared to the benchmark case with *i.i.d.* noise terms, this novel channel of learning endogenously generates optimism and either amplifies or dampens underlying shocks, depending on how much firms are informed about the shocks. In the next section, we consider the case with strategic complementarity and show that the presence of strategic complementarity always strengthens our mechanism.

Timeline. There is an infinite number of periods $t = 0, 1, \ldots$ and a representative household consisting of a continuum of workers. We use the island analogy of Lucas (1972) to capture the incompleteness of information in the real world. There is a continuum of islands $i \in [0, 1]$, each of which has its own labor market and own information set. Island i is inhabited by a continuum of firms $j \in [0, 1]$, each of which specializes in the production of differentiated commodities. We will index these firms and their commodities by $(i, j) \in [0, 1] \times [0, 1]$. The timeline is as follows. First, at the beginning of each period, the household sends one worker to each island. Second, after observing noisy signals about island-specific productivity, firms commit to their output levels, and workers post wages at which they commit to supply any amount of labor. Third, the island-specific productivity is realized, and firms' labor demand is determined by the committed level of output and the productivity. Finally, workers return to their home and commodity markets open. Prices adjust to clear the markets.

Household. A representative household consists of a continuum of workers who solve a team problem of jointly maximizing the household utility, which is given by

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(\log C_t - \nu \cdot N_t)\right]$$

where $N_t = \int_0^1 N_{it}\,\mathrm{d}i$ is the total labor supply of its workers. For simplicity, we assume a unit intertemporal elasticity of substitution and a unit Frisch elasticity of labor supply, but none of our results qualitatively depend on this assumption. The consumption C_t has a nested structure. First, it is CES aggregation of the consumption bundle $\{C_{it}\}_{i\in[0,1]}$ from different islands,

$$C_t = \left(\int_0^1 C_{it}^{1-\frac{1}{\sigma}}\,\mathrm{d}i\right)^{\frac{\sigma}{\sigma-1}}$$

where $\sigma \leq 1$ is the elasticity of substitution across consumption from different islands. Second, the consumption C_{it} from island i is also CES aggregation of the consumption bundle $\{C_{ijt}\}_{j\in[0,1]}$ from monopolistic firms in island i

$$C_{it} = \left(\int_0^1 C_{ijt}^{1-\frac{1}{\eta}}\,\mathrm{d}j\right)^{\frac{\eta}{\eta-1}} \tag{2}$$

where $\eta > 1$ is the elasticity of substitution across firms. We normalize the price index to one,

$$1 = P_t \equiv \left(\int_0^1 P_{it}^{1-\sigma}\,\mathrm{d}i\right)^{\frac{1}{1-\sigma}} \quad \text{where } P_{it} = \left(\int_0^1 P_{ijt}^{1-\eta}\,\mathrm{d}j\right)^{\frac{1}{1-\eta}}$$

where P_{ijt} is the price of the good from firm j in island i, P_{it} is the price index for goods in island i.

The budget constraint dictates that the total purchase of consumption goods and bonds cannot exceed the total income, which consists of profits, wage, and payment from the bond:[10]

$$\int_0^1\int_0^1 P_{ijt}C_{ijt}\,\mathrm{d}i\,\mathrm{d}j + B_{t+1} \leq \int_0^1\int_0^1 \Pi_{ijt}\,\mathrm{d}i\,\mathrm{d}j + \int_0^1 W_{it}N_{it}\,\mathrm{d}i + (1+R_t)B_t$$

where B_t is the bond holding in period t, R_t is the gross interest rate between period t and $t+1$, Π_{ijt} is the profit of firm j in island i, W_{it} is the wage in island i, and N_{it} is the labor supply of its worker sent to island i. When workers jointly maximize the utility of the household they belong to, they are subject to informational constraints. The labor supply decisions of workers sent to different islands are based on different information sets. Once they return to their home, all the information is shared, and the household makes consumption and saving decisions.

[10] We do not need to distinguish nominal terms and real terms because we normalize $P_t = 1$.

Firms. Firm j in island i has a production function

$$Y_{ijt} = A_{it} \cdot N_{ijt}^{\theta}$$

where A_{it} is the common productivity of firms in island i, N_{ijt} is the firm's employment, and $\theta \in (0, 1]$ governs the decreasing return to scale in production. After commodity markets open and prices clear these markets, the firm's realized profit is

$$\Pi_{ijt} = P_{ijt} Y_{ijt} - W_{it} N_{ijt}.$$

Market Clearing. A key feature of the model is that firms and workers make decisions under imperfect information. After observing a signal for island-specific productivity, firms commit to an output level, Y_{ijt}, and workers post wages W_{it}. As will become evident in Lemma 2 below, this assumption makes firms choose higher output levels when they are more optimistic. Under this assumption, labor market clearing is trivial: given island-specific productivity, A_{it}, Firm j in island i demands

$$N_{ijt} = (Y_{ijt}/A_{it})^{1/\theta} \tag{3}$$

units of labor at the equilibrium wage W_{it}. After production takes place, goods markets open, and the price P_{ijt} adjusts to clear the market: $C_{ijt} = Y_{ijt}$ for all (i, j).

Shocks and Information. The only uncertainty is on the island-specific productivity, A_{it}, which follows an AR(1) process in log,[11]

$$a_{it} \equiv \log A_{it} = \rho_a a_{it-1} + \varepsilon_{it}.$$

We further assume that the innovation ε_{it} consists of two components,

$$\varepsilon_{it} = \varepsilon_{it}^{p} + \varepsilon_{it}^{u}$$

where $\varepsilon_{it}^{p} \sim \mathcal{N}\left(0, \sigma_p^2\right)$ and $\varepsilon_{it}^{u} \sim \mathcal{N}(0, \sigma_u^2)$ are independent across time periods and independent of one another.[12] We define corresponding aggregate components of these variables as

$$a_t \equiv \int_0^1 a_{it}\, \mathrm{d}i, \;\; \varepsilon_t^{p} \equiv \int_0^1 \varepsilon_{it}^{p}\, \mathrm{d}i, \;\; \text{and} \;\; \varepsilon_t^{u} \equiv \int_0^1 \varepsilon_{it}^{u}\, \mathrm{d}i.$$

[11] Another way to model uncertainty is to assume island-specific preference shocks and aggregate demand shocks. However, this approach is more difficult to implement because it involves endogenous signals.

[12] In this section, we do not impose restrictions on the correlation structure across islands because it makes no difference in the absence of strategic interaction between firms in different islands.

The difference between the two components is the extent to which firms are informed about them. The shock ε_{it}^{p} is called *partly-observed* because, before making their decisions, firms and workers in island i receive a noisy signal

$$s_{it} = \rho_a a_{it-1} + \varepsilon_{it}^{p} + \xi_{it}. \tag{4}$$

Thus, firms are at least partly informed about the component ε_{it}^{p} when they make their decisions in period t. In contrast, the shock ε_{it}^{u} is called *unobserved* because it is not contained in the information set of firms and workers when they make their decisions at period t. Because the distinction between partly-observed and unobserved shocks is important for our mechanism, we summarize it in Definition 2. This distinction formalizes the earlier observation that there are many shocks in the real world that differ in the extent to which economic agents are informed about them.

Definition 2. *Both partly-observed shocks ε_{it}^{p} and unobserved shocks ε_{it}^{u} drive island-specific productivity. But when firms and workers make their decisions, they only receive signals about the partly-observed shocks, while they are completely uninformed about the unobserved shocks.*

Given a period-t information set Ω_{it}, which we will soon specify, firms and workers in island i maximize their expected profits and expected household utility, respectively. Formally, firm j in island i chooses the level of Y_{ijt} by solving

$$\begin{aligned}
\max_{Y_{ijt}} \quad & \mathbb{E}\left[C_t^{-1} \cdot \Pi_{ijt} \middle| \Omega_{it}\right] \\
\text{s.t.} \quad & \Pi_{ijt} = P_{ijt} Y_{ijt} - W_{it} N_{ijt} \\
& Y_{ijt} = \left(\frac{P_{ijt}}{P_{it}}\right)^{-\eta} \left(\frac{P_{it}}{P_t}\right)^{-\sigma} Y_t \\
& Y_{ijt} = A_{it} \cdot N_{ijt}^{\theta}
\end{aligned}$$

where the second constraint comes from the isoelastic demand relation. Also, the representative worker in island i chooses the level of W_{it} in a competitive way by solving

$$\begin{aligned}
\max_{W_{it}} \quad & \mathbb{E}\left[C_t^{-1} \frac{W_{it}}{P_t} N_{it} - \nu N_{it} \middle| \Omega_{it}\right] \\
\text{s.t.} \quad & N_{it} = \begin{cases} 0 & \text{if } W_{it} > W_{it}' \\ \displaystyle\int_0^1 \left(\frac{Y_{ijt}}{A_{it}}\right)^{\frac{1}{\theta}} \mathrm{d}j & \text{if } W_{it} = W_{it}' \\ \infty & \text{if } W_{it} < W_{it}' \end{cases}
\end{aligned}$$

where W_{it}' is the wage level that other workers in island i choose. We should have $W_{it}' = W_{it}$ in the equilibrium.

Persistent Noise and Feedback. The modeling assumptions so far are standard in the literature, with the possible exception of the timing assumption. We now present our two main assumptions. First, we assume that the noise terms in the signals are persistent:

$$\xi_{it} = \rho\xi_{it-1} + \eta_{it} \quad \text{where } \eta_{it} \sim \mathcal{N}\big(0, (1-\rho^2)\sigma_\eta^2\big)$$

where the innovation η_{it} is *i.i.d.* across time periods and across islands.[13] After observing the signal, firms form their beliefs about the island-specific productivity a_{it}.

Second, we assume that firms in island i receive feedback on their previous forecast by observing the actual productivity realization a_{it} after making their decisions. Thus, the information set Ω_{it} on which firms' forecasts are based contains not only their signals up to time t but also the history of previous feedback[14]

$$\Omega_{it} = (\cdots, s_{it-2}, a_{it-2}, s_{it-1}, a_{it-1}, s_{it}).$$

This is a natural assumption in our setting. We assume that firms commit to the output level, Y_{ijt}. Thus, when labor market opens, they need to know the exact level of their productivity A_{it} in order to compute the amount of labor needed to fulfill their commitment, which is given by equation (3). Likewise, workers can compute island-specific productivity based on the labor demand of firms in their island. The main role of this assumption is to allow agents to receive feedback on their previous forecasts by observing the actual realization ex post. However, it also plays an important role in making the analysis tractable, and is exactly the same assumption that many papers adopt in order to simplify learning to an essentially static one. In general, firms learn the values of both a_{it} and ξ_{it} dynamically. The presence of feedback, however, essentially allows us to abstract from dynamic learning about a_{it} and focus on dynamic learning about ξ_{it}. The logic of this section, however, would still hold insofar as firms observe the true productivity ex post with sufficient precision. Lastly, we define an additional information set that agents possess right before observing the signal,

$$\tilde{\Omega}_{it} = \Omega_{it} \backslash (s_{it}) = (\cdots, s_{it-2}, a_{it-2}, s_{it-1}, a_{it-1}).$$

[13] Whether these noise terms are correlated across islands is irrelevant in this section as we will assume away strategic interaction between islands.

[14] One might argue that it is also natural to assume that firms and workers can learn from the commodity markets. Indeed firms and workers can fully learn the aggregate shock ε_t^y from observing the prices and quantities in the commodity markets. However, we consider agents suffering a form of internal schizophrenia as in the vast majority of the literature. We think of the firms having two personalities. One choosing Y_{ijt} is inattentive and do not learn from the commodity markets, and another, who does not communicate with the former, adjusts the price to clear the commodity market. See Angeletos and La'O (2009a) for trade-offs in this modeling choice.

Table 1: Timeline

		$\vdots$
		$\boxed{a_{it-1}} = \rho_a a_{it-1} + \varepsilon^p_{it-1} + \varepsilon^u_{it-1}$
period t	stage 1	$\boxed{s_{it}} = \rho_a a_{it-1} + \varepsilon^p_{it} + \xi_{it}$ where $\xi_{it} = \rho \xi_{it-1} + \eta_{it}$ Commit to Y_{ijt} and W_{it}
	stage 2	$\boxed{a_{it}} = \rho_a a_{it-1} + \varepsilon^p_{it} + \varepsilon^u_{it}$
		$\vdots$

Note: The variables in boxes are those observed by agents in island i. All the shocks (except for ξ_{it}) indexed by t are independent across time periods. All different types of shocks are independent of one another.

Recall that we define ex ante and ex post optimism as

$$\tilde{\mathcal{O}}_{it} = \xi_{it} - \tilde{\mathbb{E}}_{it}[\xi_{it}] \quad \text{where} \quad \tilde{\mathbb{E}}_{it}[\cdot] \equiv \mathbb{E}[\cdot|\tilde{\Omega}_{it}]$$

$$\mathcal{O}_{it} = \xi_{it} - \mathbb{E}_{it}[\xi_{it}] \quad \text{where} \quad \mathbb{E}_{it}[\cdot] \equiv \mathbb{E}[\cdot|\Omega_{it}].$$

We summarize the timeline of the model in Table 1. To highlight the difference in timing between receiving the signal s_{it} and receiving the feedback a_{it}, we consider each period as comprising two stages. Finally, we define an equilibrium as follows where we write $\Omega \equiv (\Omega_{it})_i$ and $A \equiv (A_{it})_i$.

Definition 3. *A* rational expectations equilibrium *is a sequence of allocations* $\{C_{ijt}(\Omega, A), Y_{ijt}(\Omega_{it}), N_{ijt}(\Omega_{it}, A_{it})\}$ *and prices* $\{W_{it}(\Omega_{it}), P_{ijt}(\Omega, A)\}$ *such that (i) In stage 1, workers and firms maximize their expected objective functions based on the information they have; (ii) In stage 2, the representative household maximizes its utility, taking the prices as given; and (iii) All markets clear.*

Illustrative Example (Continued). To illustrate that the main message of this section is not limited to our macroeconomic example, let us return to the previous forecaster example. A forecaster i makes a forecast, or nowcast, $y_{it} = \mathbb{E}_{it}[a_{it}]$ about the inflation rate a_{it} each year. The forecaster is indeed partly informed about some shocks from the newspapers. However, there are numerous other shocks that affect the inflation rate, which are not covered in the newspapers or whose effects on the inflation rate are not even conceived by the forecaster. These are captured in the unobserved shock ε^u_{it}. Subsequently, the forecaster can observe the realized value of the inflation rate.[15]

[15] Based on the literature on rational inattention, such as Sims (2003) and Mackowiak and Wiederholt (2009), one might argue that even if the realized inflation rates are publicly revealed, the forecasters might not pay attention to them. Nevertheless, it is unlikely that the forecasters who have made a prediction for the inflation rate do not pay close attention to the realized value of it.

Optimality Conditions. The optimal wage choice of the representative worker in island i is given by

$$W_{it} = \nu \cdot \left(\mathbb{E}\left[C_t^{-1} \middle| \Omega_{it}\right]\right)^{-1},$$

which is an intratemporal optimality condition equating the marginal disutility from labor with the marginal utility from consumption. Log-linearizing this condition yields[16]

$$w_{it} = \mathbb{E}[c_t|\Omega_{it}] \tag{5}$$

where we use small letters to denote log deviations from steady state values. The higher the aggregate consumption that workers expect, the higher the wage needed to compensate them. Next, consider the firm's optimization problem. Firm j in island i solves

$$\max_{Y_{ijt}} \mathbb{E}_{it}\left[Y_t^{-1}\left\{Y_{ijt}^{1-\frac{1}{\eta}} Y_t^{\frac{1}{\eta}} P_{it}^{1-\frac{\sigma}{\eta}} - W_{it}\left(\frac{Y_{ijt}}{A_{it}}\right)^{\frac{1}{\theta}}\right\}\right]$$

where we impose $C_t = Y_t$. The first order condition gives

$$\left(1-\tfrac{1}{\eta}\right) Y_{ijt}^{-\frac{1}{\eta}} \mathbb{E}_{it}\left[Y_t^{-1+\frac{1}{\eta}} P_{it}^{1-\frac{\sigma}{\eta}}\right] = \frac{1}{\theta} Y_{ijt}^{\frac{1}{\theta}-1} \mathbb{E}_{it}\left[Y_t^{-1} W_{it} A_{it}^{-\frac{1}{\theta}}\right].$$

Log-linearizing, we have

$$\left(1-\tfrac{1}{\eta}-\tfrac{1}{\theta}\right) y_{ijt} = \mathbb{E}_{it}\left[-\tfrac{1}{\eta} y_t + w_{it} - \tfrac{1}{\theta} a_{it} - \left(1-\tfrac{\sigma}{\eta}\right) p_{it}\right].$$

The symmetry across firms in island i implies

$$P_{it} = P_{ijt} = \left(\frac{Y_{it}}{Y_t}\right)^{-\frac{1}{\sigma}},$$

which in conjunction with equation (5) shows that the equilibrium of our microfounded model can be represented by the perfect Bayesian equilibrium of games with strategic complementarity, as in Angeletos and La'O (2009a).

Lemma 2. *Firms' equilibrium output choices are characterized, up to a log-linear approximation, by the solution of the following fixed-point problem:*

$$y_{ijt} = \mathbb{E}_{it}[(1-\alpha) a_{it} + \alpha y_t]$$

[16] Since Lucas (1972), the log linearization is frequently used in the literature as it allows a simple representation of the equilibrium with a signal extraction problem.

where the degree of strategic complementarity α is given by

$$\alpha = \frac{1/\sigma - 1}{1/\theta + 1/\sigma - 1} \in [0, 1).$$

In this section, we maintain the following assumption in order to assume away strategic complementarity, $\alpha = 0$, so that we can focus on how firms learn about their own noise terms.

Assumption 1 (No Strategic Complementarity). $\sigma = 1$.

Remark. Before proceeding, it is worth discussing some of the modeling choices we have made. First, it makes no difference whether we assume a general CRRA utility from consumption, $\frac{C_t^{1-\gamma}}{1-\gamma}$. We can follow the same steps to show that the equilibrium output choice is given by

$$y_{ijt} = \mathbb{E}_{it}[(1-\alpha)\tilde{a}_{it} + \alpha y_t],$$

where $\tilde{a}_{it} = \frac{1/\theta}{1/\theta+\gamma-1} a_{it}$ is the normalized productivity. With this normalization, we obtain the same result as in the case with $\gamma = 1$. Second, if we assume that the disutility from labor has a form $\nu \cdot \frac{N_t^{1+\varepsilon}}{1+\varepsilon}$ instead of $\nu \cdot N_t$, then we have an additional aggregate productivity term in the representation, $y_{ijt} = \mathbb{E}_{it}[(1-\alpha)a_{it} + \alpha y_t + \beta \cdot a_t]$ for some constant $\beta > 0$. This only complicates the learning of firms without providing further insight, so we assume constant marginal disutility from labor. Third, if we assume that firms commit to N_{ijt} instead of Y_{ijt}, then the optimal labor demand choice N_{ijt} is decreasing instead of increasing in $\mathbb{E}_{it}[A_{it}]$. This is because the CES aggregation (2) features diminishing marginal returns, which implies that if productivity doubles, firms would increase output less than twice as much. Thus, optimism leads to lower output, which is the opposite of what we are trying to capture.

Benchmark: *i.i.d.* noise terms. We conclude this section with a benchmark case with *i.i.d.* noise terms, as in the literature. This case is nested in the previous model with $\rho = 0$. Since we abstract from dynamic learning of productivity, there is nothing left to learn dynamically, and the problem essentially becomes a repetition of static learning problems. Thus, what happened in one period has no effect on firms' decisions in the next period. In particular, it does not change firms' optimism in the next period. Consequently, shocks can influence tomorrow's output solely through their impact on productivity, irrespective of whether they are partly-observed or unobserved. This is one of the reasons why Woodford (2003) does not adopt the assumption of Lucas (1972) that fundamentals—monetary disturbances—become public information within a period. Given that monetary statistics are reported promptly, Woodford (2003) follows Sims (2003) in assuming limited attention. In this paper, however, underlying shocks will have persistent effects even if firms observe productivity within a period.

We can guess and verify the coefficients of a linear equilibrium. Proposition 1 summarizes the result.

Proposition 1. *If noise terms in signals are i.i.d., $\rho = 0$, the optimal output choice of firms is given by*

$$y_{ijt+1} = \rho_a^2 a_{it-1} + \rho_a \varepsilon_{it}^p + \rho_a \varepsilon_{it}^u + \tilde{K} \varepsilon_{it+1}^p + \tilde{K} \eta_{it+1} \quad \text{where } \tilde{K} = \tfrac{\sigma_p^2}{\sigma_\eta^2 + \sigma_p^2} \in (0,1),$$

which can be aggregated into

$$y_{t+1} = \rho_a^2 a_{t-1} + \rho_a \varepsilon_t^p + \rho_a \varepsilon_t^u + \tilde{K} \varepsilon_{t+1}^p.$$

Thus, the effects of period-t shocks on period-$(t+1)$ outcomes are identical to their effects on the fundamental:

$$\frac{\partial y_{t+1}}{\partial \varepsilon_t^u} = \frac{\partial a_{t+1}}{\partial \varepsilon_t^u} \quad \textit{and} \quad \frac{\partial y_{t+1}}{\partial \varepsilon_t^p} = \frac{\partial a_{t+1}}{\partial \varepsilon_t^p}.$$

Note that a contemporaneous partly observed shock affects firms' decisions less than one-for-one, reflecting the fact that firms cannot fully identify this shock. On the other hand, a contemporaneous unobserved shock cannot affect their decisions as it is not in their information set.

Persistent noise terms. Let us go back to our main case with persistent noise terms, $\rho \in (0,1)$. We will characterize how firms update their beliefs about their noise terms and how this learning affects firms' forecasts of the productivity and hence their output choices. We write firms' belief about ξ_{it-1} right before observing s_{it} as $\xi_{it-1}|\tilde{\Omega}_{it} \sim \mathcal{N}(m_{it-1}, V_{t-1})$. After observing s_{it}, firms in island i make a forecast about a_{it} according to Bayes' rule:

Lemma 3. *Bayesian updating leads to the following forecast:*

$$\begin{aligned}
\mathbb{E}_{it}[a_{it}] &= \rho_a a_{it-1} + K_t\big(s_{it} - \rho_a a_{it-1} - \rho m_{it-1}\big) \quad \text{where } K_t = \tfrac{\sigma_p^2}{\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2 + \sigma_p^2} \in (0,1) \\
&= \rho_a a_{it-1} + K_t\big(\varepsilon_{it}^p + \tilde{\mathcal{O}}_{it}\big) \\
&= \rho_a a_{it-1} + \varepsilon_{it}^p + \mathcal{O}_{it}
\end{aligned}$$

Since firms in island i know the value of a_{it-1} at this point, it directly affects their forecasts. The contemporaneous partly-observed shock has less than a one-for-one effect (we will soon show that a positive realization of ε_{it}^p reduces $\mathcal{O}_{it}$), while the contemporaneous unobserved shock has no effect, as in the benchmark. The difference is that now firms' forecasts also depend on how optimistic they are. Optimistic firms interpret their signals more optimistically, which leads them to make optimistic forecasts.

After making a forecast, firms in island i receive feedback at stage 2 by observing the true productivity, a_{it}, and update their beliefs about the noise term by looking back on their previous forecasts. This learning can be characterized by the Kalman filter and the results are summarized in the next lemma.

Lemma 4. *The law of motions for m_{it} and V_t are given by*

$$m_{it} = (\gamma_1(V_{t-1}) + \gamma_2(V_{t-1})) \cdot s_{it} + \gamma_3(V_{t-1}) \cdot \rho m_{it-1} - \gamma_1(V_{t-1}) \cdot \rho_a a_{it-1} - \gamma_2(V_{t-1}) \cdot a_{it}$$
$$V_t = \frac{\sigma_p^2 \sigma_u^2 (\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2)}{(\sigma_p^2 + \sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2 \sigma_u^2}$$

where

$$\gamma_1(V_{t-1}) = \frac{\sigma_u^2(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2)}{(\sigma_p^2+\sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}$$
$$\gamma_2(V_{t-1}) = \frac{\sigma_p^2(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2)}{(\sigma_p^2+\sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}$$
$$\gamma_3(V_{t-1}) = \frac{\sigma_p^2\sigma_u^2}{(\sigma_p^2+\sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}.$$

Note that $\gamma_1(V_{t-1}), \gamma_2(V_{t-1}), \gamma_3(V_{t-1}) \in (0,1)$ and $\gamma_1(V_{t-1}) + \gamma_2(V_{t-1}) + \gamma_3(V_{t-1}) = 1$.

We can easily prove that there is a unique fixed point V such that $V_{t-1} = V$ implies $V_t = V$. We can also show that the sequence $(V_t)_t$ converges to this fixed point for any initial value of $V_0 \geq 0$. Thus, we will consider a stationary environment in which $V_t = V$ for all t. We can then write the law of motion for m_{it} in a time-invariant form:

$$m_{it} = (\gamma_1 + \gamma_2) \cdot s_{it} + \gamma_3 \cdot \rho m_{it-1} - \gamma_1 \cdot \rho_a a_{it-1} - \gamma_2 \cdot a_{it} \quad \text{where } \gamma_i \equiv \gamma_i(V).$$

Also, we define the stationary Kalman gain as $K \equiv \frac{\sigma_p^2}{\rho^2 V + (1-\rho^2)\sigma_\eta^2 + \sigma_p^2} \in (0,1)$. We are now able to characterize the dynamics of optimism. Recall that we defined the optimism as the extent to which firms in island i underestimate ξ_{it}.

Proposition 2. *The law of motions for ex ante and ex post optimism, $\tilde{\mathcal{O}}_{it} \equiv \xi_{it} - \tilde{\mathbb{E}}_{it}[\xi_{it}]$ and $\mathcal{O}_{it} \equiv \xi_{it} - \mathbb{E}_{it}[\xi_{it}]$, are given by*

$$\tilde{\mathcal{O}}_{it+1} = \gamma_3 \rho \tilde{\mathcal{O}}_{it} - \rho\gamma_1 \varepsilon_{it}^p + \rho\gamma_2 \varepsilon_{it}^u + \eta_{it+1}$$
$$\mathcal{O}_{it+1} = \gamma_3 \rho \mathcal{O}_{it} - (1-K)\varepsilon_{it}^p + K\rho\gamma_2 \varepsilon_{it}^u + K\eta_{it+1}.$$

Thus, a positive realization of partly-observed shocks (unobserved shocks, respectively) makes firms pessimistic (optimistic, respectively) next period.

First, we can see that there is inertia in optimism, as firms can only correct their optimism through noisy learning.[17] Thus, a positive shock in the noise term, η_{it+1}, increases the firm's optimism, which decays slowly over time. More interesting is the response of optimism to the underlying shocks. The partly observed shock and the

[17] In the discussion below, the optimism means both ex ante and ex post optimism.

unobserved shock have opposite effects on optimism. The intuition is simple. Suppose that realized productivity $a_{it} = \rho_a a_{it-1} + \varepsilon^p_{it} + \varepsilon^u_{it}$ is greater than its expected value $\rho_a a_{it-1}$, and firms in island i observe this increase at the end of period-t. If this increase were solely due to an increase in ε^u_{it}, it would not be reflected in s_{it} at all, so the realized value of a_{it} is *higher-than-expected* from the perspective of firms in island i. This makes them think that they were too pessimistic in interpreting s_{it}, which in turn induces them to interpret s_{it+1} more optimistically in the next period. Therefore, a positive innovation in the unobserved shock increases firms' optimism. In contrast, if the increase in the fundamental is due only to the high ε^p_{it}, agents rationally attribute this increase to both ε^p_{it} and ε^u_{it} when they observe a_{it}. However, the high realization of ε^p_{it} was fully reflected in s_{it}. Thus, the realized value of a_{it} is *lower-then-expected* for firms in island i. This makes them possess a more pessimistic belief in the next period.

Before turning to the analysis of firms' output choices, we discuss comparative statics results for γ_1, γ_2, and γ_3 with respect to variance parameters, σ^2_p, σ^2_u, and σ^2_η. Lemma 5 summarizes the results.

Lemma 5. *We have the following comparative statics.*

(1) γ_1 *is increasing in* σ^2_u *and* σ^2_η*, while decreasing in* σ^2_p

(2) γ_2 *is increasing in* σ^2_p *and* σ^2_η*, while decreasing in* σ^2_u

(3) γ_3 *is increasing in* σ^2_p *and* σ^2_u*, while decreasing in* σ^2_η

To understand this result, first consider Part (2), which states how the effect of the unobserved shock on optimism, γ_2, depends on the variance parameters. The main mechanism that changes optimism is the rational confusion between various shocks. If the partly-observed shock is relatively more volatile, then firms misattribute an increase in the unobserved shock more to the partly-observed shock, so they underestimate their noise terms more. Thus, we get a larger effect of the unobserved shock on optimism. Following the same logic, σ^2_u tends to reduce the effect of the unobserved shock on optimism. Moreover, since optimism arises as firms overestimate or underestimate their ξ_{it}, and firms are more likely to do so when σ^2_η is high, the effect of the unobserved shock on optimism tends to increase in σ^2_η. This explains Part (2), and we can apply the same argument to Part (1). For Part (3), note that γ_3 is the coefficient that determines the degree of inertia in optimism. This inertia comes from the rational confusion of firms between ξ_{it} and $(\varepsilon^p_{it}, \varepsilon^u_{it})$, which prevents them from fully correcting their optimism. This explains why γ_3 is decreasing in the relative size of σ^2_η compared to σ^2_p and σ^2_u.

Combining the results so far, we can characterizes the dynamics of the output choice as in the next theorem, which is our first main result.

Theorem 1. *The optimal output choice of firms is given by*

$$y_{ijt+1} = \rho_a^2 a_{it-1} + (\rho_a - \rho K \gamma_1)\varepsilon^p_{it} + (\rho_a + \rho K \gamma_2)\varepsilon^u_{it} + K\varepsilon^p_{it+1} + K\eta_{it+1} + \rho\gamma_3 K \tilde{\mathcal{O}}_{it}$$

hence the aggregate output is

$$y_{t+1} = \rho_a^2 a_{t-1} + (\rho_a - \rho K \gamma_1)\varepsilon_t^p + (\rho_a + \rho K \gamma_2)\varepsilon_t^u + K\varepsilon_{t+1}^p + \rho\gamma_3 K \int_0^1 \tilde{\mathcal{O}}_{it}\,\mathrm{d}i.$$

Thus, the effects of partly-observed shocks (unobserved shocks, respectively) on the next period outcomes are dampened (amplified, respectively) compared to their effects on the productivity:

$$\frac{\partial y_{t+1}}{\partial \varepsilon_t^u} > \frac{\partial a_{t+1}}{\partial \varepsilon_t^u} \quad \textit{and} \quad \frac{\partial y_{t+1}}{\partial \varepsilon_t^p} < \frac{\partial a_{t+1}}{\partial \varepsilon_t^p}.$$

Compared to the benchmark case in Proposition 1, this theorem establishes that the effect of the unobserved shock on the next period output is amplified by its effect on the agent's optimism.[18] At the same time, the effect of the partly-observed shock on the next period output is dampened because firms become pessimistic after a positive innovation in the partly-observed shock. A contemporaneous innovation η_{it+1} has a positive effect on output because firms cannot fully distinguish it from other shocks, and its effect decays slowly over time as firms correct their optimism. A special case of interest is that with $\rho_a = 0$ (*i.i.d.* productivity). In this case, we can observe that optimism propagates the effect of unobserved shocks to the next period, while the effect of partly-observed shocks is negative in the next period. Note that shocks in this case cannot affect future output if we assume *i.i.d.* noise, as in the literature. However, with persistent noise, these shocks can affect future output through their effects on firms' optimism.

Illustrative Example (Continued). Consider the inflation rate forecaster example again. Suppose the forecaster predicted an inflation rate of 2% based on her reading of the newspapers. Now suppose that the actual inflation rate turns out to be 1%. How does her interpretation of the newspaper change? The fact that the inflation rate turned out to be *lower-than-expected* leads her to believe that she was *too optimistic* in interpreting the contents of the newspapers. So, the forecaster would rationally take this into account the next time she makes the forecasts, and interpret the contents of the newspapers in a more pessimistic way.

Implication. The results so far may sound like "anything goes." Indeed, they imply that a shock can be either amplified or dampened, depending on how much firms are informed about the shock; i.e., where a given shock falls on the spectrum of *degree of observability*, from fully observed to completely unobserved.[19]

[18] Recall, however, that we have assumed away the sluggish response of expectations to the innovation in productivity. Thus, a correct interpretation of this result is that the presence of persistent noise terms amplifies (dampens, respectively) the effect of unobserved (partly-observed, respectively) shocks compared to the case with *i.i.d.* noise terms.

[19] Actually, with rational forecasts, the effect of one shock can be amplified precisely because the effect of another shock is dampened, and vice versa.

There are two ways to overcome this anything goes interpretation. First, we can use forecast data to measure the degree of observability of a shock of interest, and then our theory disciplines its dynamic effects. Another way is to assume that the partly-observed shocks are more likely to be idiosyncratic, while the unobserved shocks are more likely to be common across agents. The assumption that agents are relatively well informed about idiosyncratic shocks and less informed about aggregate shocks is often considered plausible in the literature.[20] In line with this, we will make an additional assumption.

Assumption 2. *Firms make decisions based on noisy information about purely idiosyncratic shocks, but productivity also depends on aggregate shocks. That is,*

- *Partly-observed shocks are purely island specific:* $\varepsilon_{it}^{p} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_p^2\right)$ *across islands*
- *Unobserved shocks are common:* $\varepsilon_{it}^{u} \equiv \varepsilon_{t}^{aggr} \sim \mathcal{N}(0, \sigma_u^2)$

Under this assumption, partly-observed shocks still lead to rational confusion, but they are averaged out among the continuum of islands, $\varepsilon_t^p \equiv \int_0^1 \varepsilon_{it}^p \, \mathrm{d}i = 0$. Thus, we can aggregate Proposition 2 and Theorem 1 as if the aggregate economy were driven only by unobserved shocks.

Corollary 1. *Under Assumption 2, the aggregate output and aggregate optimism,* $\tilde{\mathcal{O}}_t = \int_0^1 \tilde{\mathcal{O}}_{it} \, \mathrm{d}i$*, follow*

$$y_{t+1} = \rho_a^2 a_{t-1} + \beta_u \varepsilon_t^{aggr} + \beta_{\mathcal{O}} \tilde{\mathcal{O}}_t$$
$$\tilde{\mathcal{O}}_{t+1} = \gamma_3 \rho \tilde{\mathcal{O}}_t + \rho \gamma_2 \varepsilon_t^{aggr}$$

Thus, the aggregate shock has no contemporaneous effect on outcomes, while it has an amplified effect on the next period outcomes:

$$\frac{\partial y_{t+1}}{\partial \varepsilon_t^{aggr}} > \frac{\partial a_{t+1}}{\partial \varepsilon_t^{aggr}} \quad \textit{while} \quad \frac{\partial y_{t+1}}{\partial \varepsilon_{t+1}^{aggr}} = 0 < \frac{\partial a_{t+1}}{\partial \varepsilon_{t+1}^{aggr}}.$$

When firms make decisions based on noisy information about their idiosyncratic shocks while the economic condition also depends on an aggregate shock, aggregate optimism fluctuates procyclically with the aggregate shock and thus has an amplified effect on aggregate output after firms receive feedback on their previous forecasts. In contrast, the aggregate shock has no contemporaneous effect on aggregate action and forecast. In other words, the aggregate shock that has little effect on contemporaneous expectations would be amplified later when firms receive feedback. We find suggestive evidence of this result in Angeletos, Huo, and Sastry (2020). They show that, in response to aggregate shocks, agents' expectations underreact initially but overshoot later. They attribute this *delayed*

[20] For a prominent example, Mackowiak and Wiederholt (2009) calibrate their model by matching the price changes observed in data and conclude that firms pay more attention to idiosyncratic conditions than to aggregate conditions. This is because idiosyncratic conditions are more volatile than aggregate conditions. The theory of Kohlhas and Walther (2020) also relies on this asymmetry.

Table 2: Interaction Between Persistent Noise and Feedback

	Without feedback	With feedback
i.i.d. noise	No optimism	No optimism Static learning
Persistent noise	Higher-than-expected outcome $\Rightarrow$ pessimism	Higher-than-expected outcome $\Rightarrow$ optimism

overreaction to a combination of dispersed information and behavioral over-extrapolation. However, there finding can also be well understood using our result; expectations initially underreact due to the fact that agents are not informed about the aggregate shock—this part is identical to Angeletos, Huo, and Sastry (2020)—and overshoot later on when they receive feedback and adjust their optimism. It is worth noting that Corollary 1 does not depend on the exact form of Assumption 2. In Section 3.5 we obtain a qualitatively similar delayed overreaction as long as firms are relatively well informed about idiosyncratic shocks.

Remark. We conclude this section by discussing the importance of the interaction between the persistent noise terms and the presence of feedback in obtaining our mechanism. Table 2 summarizes the discussion. First, if the noise terms are *i.i.d.*, then the presence of feedback does not affect the qualitative results. Firms learn nothing ex post, so feedback does not affect firms' learning. Second, the presence of feedback is crucial for our mechanism to work. We essentially assume that firms observe two signals, s_{it} and a_{it}, and that the second signal provides feedback on the forecast made with the first signal. One might think that the second signal plays a redundant role in the sense that, even if firms only have the first signal, s_{it}, they can receive feedback from the future signal, s_{it+1}, and learn the persistent noise terms. However, to formalize our mechanism, it is crucial to incorporate the second signal in the model.[21] To see this, suppose that firms in island i observe only the first signal, $s_{it} = \rho_a a_{it-1} + \varepsilon^p_{it} + \xi_{it}$, in each period. Firms can indeed get feedback on s_{it} when they observe the next period signal, s_{it+1}, since it contains some information about ξ_{it}. However, this feedback gives a result that is exactly the opposite of our previous intuition; with this feedback, higher-than-expected outcomes make firms *pessimistic*.[22] The reason is that when firms observe a higher-than-expected signal in period $t+1$ due to a positive innovation in ε^u_{it}, they partly attribute this surprise to a higher realization of ξ_{it+1}, which means that they become pessimistic. We conclude that what underlies our mechanism is the interaction between persistent noise terms and the presence of feedback.

[21] In general, we need another noisy signal about the true realization whose noise term is not much correlated with the first signal.
[22] Acharya, Benhabib, and Huo (2019) document a similar result.

3.4 Strategic Complementarity

In this section, we illustrate how the introduction of strategic complementarity provides additional insights. With strategic complementarity, firms have incentives to predict the actions of other firms in different islands. To do so, they try to forecast the optimism of other firms.[23] Firms in this model are concerned not only with the optimism of others (second-order optimism), but also with higher-order optimism—how other firms think about others' optimism, how other firms think about others' beliefs about others' optimism, and so on. We first characterize how firms update their higher-order optimism, and how this higher-order optimism in turn affects firms' output choices through its effects on higher-order beliefs about productivity. In particular, the introduction of strategic complementarity and the resulting higher-order optimism always work in the direction of reinforcing the mechanism we document in the previous section.

The presence of higher-order optimism makes it difficult to solve the model due to the infinite regress problem of Townsend (1983), so we make some simplifying assumptions in order to obtain sharp analytical results. First, we consider a two-period version of the model where periods are indexed by $t = 0, 1$. Second, we assume that productivity is *i.i.d.* across periods (i.e., $\rho_a = 0$), thereby focusing on how agents learn the noise terms. Third, we assume that the noise terms are time-invariant, which we denote by ξ_i without t index, so that agents in island i observe a signal of the form

$$s_{it} = \varepsilon_{it}^p + \xi_i$$

where ξ_i is independent across islands[24]

$$\xi_i \overset{iid}{\sim} \mathcal{N}\big(0, \sigma_\xi^2\big).$$

These assumptions are not essential for our results but simplify our exposition. In a numerical exercise in Section 3.4.1, we will show that the main message of this section does not rely on these simplifying assumptions. Last but not least, we assume that islands share one common productivity, which we denote by a_t without i index,

$$a_t = \varepsilon_t^p + \varepsilon_t^u$$

Remark. The last assumption of common productivity requires further explanation. Our main goal in this section is to study the role of strategic complementarity on firms' learning and optimal output choices under incomplete

[23] This is analogous to the literature on higher-order beliefs, where agents try to forecast beliefs of other agents in order to forecast others' actions.

[24] We can alternatively allow the possibility that noise terms are positively correlated across agents. Then, agents try to learn this common component, which can generate additional channel through which underlying shocks affect the (higher-order) optimism.

information. However, if we assume that ε_{it}^p is a pure idiosyncratic shock, $\int_0^1 \varepsilon_{it}^p \, \mathrm{d}i = 0$, then the presence of strategic complementarity not only affects the learning of firms but also reduces the importance of productivity in choosing output. To see this clearly, consider the following two static examples, which use the notation of our model but are more similar to Woodford (2003) and Angeletos and La'O (2009a) in terms of the information structure.

Example 1. There is a continuum of agents $i \in [0,1]$ who share a common fundamental $a \sim \mathcal{N}(0, \sigma_a^2)$. Each agent i chooses an action y_i after observing a private signal $s_i = a + \xi_i$ where $\xi_i \sim \mathcal{N}\big(0, \sigma_\xi^2\big)$ is i.i.d. across agents. Agents' best response is assumed to be $y_i = (1-\alpha)\,\mathbb{E}_i\, a_i + \alpha\,\mathbb{E}_i\, y$ where $y = \int_0^1 y_j \, \mathrm{d}j$. We can show that the equilibrium action is given by

$$y_i = \frac{(1-\alpha)\sigma_a^2}{\sigma_\xi^2 + (1-\alpha)\sigma_a^2} s_i \quad \textit{and} \quad y = \frac{(1-\alpha)\sigma_a^2}{\sigma_\xi^2 + (1-\alpha)\sigma_a^2} a$$

Example 2. The only difference from Example 1 is that fundamental $a_i \sim \mathcal{N}(0, \sigma_a^2)$ is i.i.d. across agents, hence $\int_0^1 a_j \, \mathrm{d}j = 0$. In this case, we always have $y = 0$ and the equilibrium action is given by

$$y_i = (1-\alpha)\,\mathbb{E}_i\, a_i = \frac{(1-\alpha)\sigma_a^2}{\sigma_\xi^2 + \sigma_a^2} s_i$$

Thus, even under complete information ($\sigma_\xi^2 = 0$), when fundamental is purely idiosyncratic as in Example 2, a higher degree of strategic complementarity makes agents less responsive to the change in fundamental. This comparative static is neither our goal of this section nor the inertia documented by Woodford (2003) and Angeletos and La'O (2009a). Instead, what these papers document is that, in Example 1, a higher degree of strategic complementarity makes agents less responsive to the change in fundamental *only when* information is incomplete ($\sigma_\xi^2 > 0$). This is because the higher the degree of strategic complementarity is, the more weight agents put on the common prior. Thus, the role of our last assumption is to isolate the effect of strategic complementarity on firms' learning.[25]

As before, all firms observe the realized productivity a_t at the end of each period, which depends also on the unobserved shock, ε_t^u. Thus, firms in island i have three different information sets

$$\Omega_{i0} = (s_{i0}), \quad \tilde{\Omega}_{i1} = (s_{i0}, a_0), \quad \text{and} \quad \Omega_{i1} = (s_{i0}, a_0, s_{i1}).$$

To introduce strategic complementarity, we depart from Assumption 1 and assume the following.

Assumption 3 (Strategic Complementarity). *The trade linkage is strong enough to induce strategic complementarity in output choices across islands: $1/\sigma > 1$.*

[25] We can assume instead that $a_{it} = \varepsilon_{it}^p + \varepsilon_t^u$ where $\int_0^1 \varepsilon_{jt}^p \, \mathrm{d}j = 0$. In this case, however, a relevant comparative statics is changing α when agents' best response is given by $y_{it} = \mathbb{E}_{it}\, a_{it} + \alpha\,\mathbb{E}_{it}\, y_t$.

Table 3: Timeline

period 0	stage 1	$\boxed{s_{i0}} = \varepsilon_0^p + \xi_i$ $y_{i0} = (1-\alpha) \cdot \mathbb{E}_{i0}\, a_0 + \alpha \cdot \mathbb{E}_{i0}\, y_0$ Commit to Y_{ij0} and W_{i0}
	stage 2	$\boxed{a_0} = \varepsilon_0^p + \varepsilon_0^u$
period 1	stage 1	$\boxed{s_{i1}} = \varepsilon_1^p + \xi_i$ $y_{i1} = (1-\alpha) \cdot \mathbb{E}_{i1}\, a_1 + \alpha \cdot \mathbb{E}_{i1}\, y_1$ Commit to Y_{ij1} and W_{i1}
	stage 2	$\boxed{a_1} = \varepsilon_1^p + \varepsilon_1^u$

Note: The variables in boxes are those observed by agent i. The distribution of each shock is given by $\varepsilon_t^p \sim \mathcal{N}(0, \sigma_p^2), \varepsilon_t^u \sim \mathcal{N}(0, \sigma_u^2)$ and $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$. All the shocks indexed by i are independent across agents. All the shocks indexed by t are independent across time. All different types of shocks are independent of one another.

The inverse of the substitutability between goods from different islands, $1/\sigma$, governs the strength of a trade linkage. With a strong trade linkage, firms increase their output choices when they expect others to do so. At the same time, however, the log utility features diminishing marginal utility, which makes firms decrease their output choices when they expect other firms to increase their output levels. This is because household is expected to have low marginal utility from consumption, increasing the equilibrium wage. When Assumption 3 holds, the first effect dominates the second, and the optimal output choices feature strategic complementarity across islands:

$$y_{ijt} = (1-\alpha)\, \mathbb{E}_{it}[a_t] + \alpha\, \mathbb{E}_{it}[y_t] \quad \text{where } \; y_t = \int_0^1 y_{jt}\, \mathrm{d}j$$

Here, the weight $\alpha = \frac{1/\sigma - 1}{1/\theta + 1/\sigma - 1} \in (0,1)$ is the degree of strategic complementarity and y_t is the average action of other firms. We summarize the timeline of the model in Table 3. We assume that this structure of the model is common knowledge among all firms and workers.

Higher-Order Optimism. One might argue that there is no point for firms to learn the average optimism of others because the *i.i.d.* noise terms of a continuum of firms are averaged out. However, this is not the case. Firms try to correct their optimism by observing their signals, so their endogenous optimism tend to move in the same direction. Higher-order optimism is about how firms think about this comovement, how firms think about others' beliefs about this comovement, and so on. The literature often assumes that optimism arises exogenously, and that this optimism is correlated across economic agents in order for it to affect the economy. However, it is difficult to justify this

correlation if we are silent about the origin of optimism. Our paper provides a justification: optimism is correlated across agents because they observe the same economic outcomes.

Now we formally define the higher-order optimism. Recall that we define the optimism of a firm in island i at time t as the extent to which this firm underestimates its noise term, ξ_{it}. Likewise, from a firm's perspective, other firms in different islands are expected to be optimistic in interpreting their signals if they are expected to underestimate their noise terms. As we consider an environment with strategic complementarity, we can call it the firm's optimism about others' optimism or the *second-order optimism*. We proceed in a similar manner to define (ex-post) higher-order optimism.[26]

Definition 4. *The* ex-post hth-order optimism *of firms in island i is defined recursively by*

$$\mathcal{O}_{it}^h \equiv \mathbb{E}\left[\int_0^1 \mathcal{O}_{jt}^{h-1}\, dj \middle| \Omega_{it}\right], \;\; h = 2, 3, 4, \cdots \quad \text{where } \mathcal{O}_{it}^1 \equiv \mathcal{O}_{it}.$$

We also define the average ex-post higher-order optimism *by*

$$\mathcal{O}_t^h = \int_0^1 \mathcal{O}_{jt}^h\, dj.$$

Equipped with this definition, we will now solve for the equilibrium. It is well known that the period-0 equilibrium is unique.

Lemma 6. *In period 0, there is a unique equilibrium, in which firms choose*

$$y_{i0} = \theta \cdot s_{i0} \quad \text{where } \theta = \tfrac{(1-\alpha)\sigma_p^2}{(1-\alpha)\sigma_p^2 + \sigma_\xi^2} \in (0, 1)$$

and hence the aggregate output is $y_0 = \theta \cdot \varepsilon_0^p$.

For a given value of α, high σ_p^2 or low σ_ξ^2 implies that signals are more informative about productivity. Firms then respond more to their signals. All else equal, the degree of strategic complementarity α reduces θ because it makes agents put more weight on higher-order beliefs, which are more anchored in their mean-zero prior.

What we are interested in, however, is not the period-0 equilibrium.[27] Our focus is on how firms learn their own and others' noise terms and how this learning changes the effect of period-0 shocks on period-1 outcomes. Note that we can characterize the period-1 equilibrium without calculating firms' higher-order optimism or higher-order beliefs about the fundamental. Starting with a guess of a linear equilibrium, we can compute first-order beliefs

[26] We can use ex-ante higher-order optimism instead, which is defined analogously. It turns out that ex-post higher-order optimism, however, is much easier to keep track of under the presence of strategic complementarity, so we focus only on ex-post ones in this section.

[27] Actually, we do not even need to solve the period-0 equilibrium.

about the endogenous aggregate output y_1, which gives the updated linear best response. The fixed point of this guess-and-verify process gives a unique linear equilibrium for period 1. The result is summarized in Lemma 7, which corresponds to Theorem 1 for the case with strategic complementarity.

Lemma 7. *In period 1, there is a unique linear equilibrium in which the equilibrium output is given by*[28]

$$y_1 = \gamma_p \varepsilon_0^p + \gamma_u \varepsilon_0^u + \gamma_p' \varepsilon_1^p$$

where

$$\begin{aligned}
\gamma_p &= -\frac{\sigma_u^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} \\
\gamma_u &= \frac{\sigma_p^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} \\
\gamma_p' &= \frac{(1-\alpha)\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+\sigma_u^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2},
\end{aligned}$$

so $\gamma_u > 0 > \gamma_p$ *and* $\gamma_p' \in (0,1)$.

This lemma shows that the intuition of the previous section is extended to a model with strategic complementarity: Unobserved shocks are propagated to period 1 ($\gamma_u > 0$), while partly-observed shocks have negative effects on period-1 outcomes ($\gamma_p < 0$). A natural question that follows is whether the strategic complementarity and resulting higher-order optimism strengthen or weaken our mechanism. In order to answer this question and to fully understand the period-1 equilibrium, we need to keep track of higher-order optimism and its effects on higher-order beliefs about the productivity. We first characterize the (ex-post) higher-order optimism in Lemma 8.

Lemma 8. *After observing* $(s_{i0}, a_0, s_{i1})'$*, higher-order optimism of firms in island* i *is given by*

$$\begin{aligned}
\mathcal{O}_{i1} &\equiv \xi_i - \mathbb{E}_{i1}[\xi_i] = \underset{1\times 4}{Q}\begin{pmatrix}\varepsilon_0^p & \xi_i & \varepsilon_0^u & \varepsilon_1^p\end{pmatrix}' \\
\mathcal{O}_{i1}^h &\equiv \mathbb{E}_{i1}\left[\int_0^1 \mathcal{O}_{j1}^{h-1}\,dj\right] = QT^{h-1}\begin{pmatrix}\varepsilon_0^p & \xi_i & \varepsilon_0^u & \varepsilon_1^p\end{pmatrix}'
\end{aligned}$$

for some matrices $\underset{1\times 4}{Q}$ *and* $\underset{4\times 4}{T}$*, where the sign of each element is*

$$Q = \begin{pmatrix} - & + & + & - \end{pmatrix} \quad \text{and} \quad QT^{h-1} = \begin{pmatrix} - & - & + & - \end{pmatrix}.$$

For future reference, we also note that the second element of Q*,* $Q_{1,2}$*, is decreasing in* σ_ξ^2 *and increasing in* σ_p^2 *and* σ_u^2.

[28] The results below show that this is a unique equilibrium even if we allow for the possibility of a nonlinear equilibrium.

First of all, the first-order optimism is increasing in ξ_i (see the sign of the second element of Q) because firms are unable to fully identify an increase in ξ_i as they rationally confuse it with changes in ε_0^p and ε_0^u. It is then immediate that $Q_{1,2}$ is decreasing in σ_ξ^2 and increasing in σ_p^2 and σ_u^2 as in Part (3) of Lemma 5.

Second, the first-order optimism is decreasing in ε_0^p and increasing in ε_0^u as in the previous section (see the first and third elements of Q). More importantly, higher-order optimism always moves in the same direction as the first-order optimism in response to ε_0^p and ε_0^u (see the signs of the first and third elements of QT^{h-1}). To understand this, consider a firm in the *average* island i in the sense that $\xi_i = 0$. Suppose that there is a positive unit innovation in ε_0^u and that all other aggregate shocks remain zero. Then, a firm in the island i will observe higher-than-expected fundamental a_0 so her first-order optimism will be positive,

$$\mathcal{O}_{i1} = -\mathbb{E}_{i1}[\xi_i] > 0.$$

Since noise terms are symmetrically distributed around 0, this inequality means that she expects that the noise terms of other islands are on average $\mathcal{O}_{i1}$ units higher than her noise term. This in turn means that the first-order optimism of firms in other islands are on average $\mathcal{O}_{i1} \cdot Q_{1,2}$ units higher than her first-order optimism. At the same time, from her perspective, she is the one who is expected to have zero optimism (i.e., $\mathbb{E}_{i1}[\mathcal{O}_{i1}] = 0$). Thus, we can conclude that her second-order optimism is given by $\mathcal{O}_{i1}^2 = \mathcal{O}_{i1} \cdot Q_{1,2}$. This explains why the second-order optimism moves in the same direction as the first-order optimism in response to ε_0^u. In other words, firms who view a_0 as higher-than-expected will on average think that firms in other islands are likely to have higher noise terms than theirs, thereby being optimistic on average. Similar reasoning can be recursively applied to show that all the higher-order optimism is given by $\mathcal{O}_{i1}^h = \mathcal{O}_{i1} \cdot Q_{1,2}^{h-1}$, which also moves in the same direction. We can similarly see that higher-order optimism also moves in the same direction as the first-order optimism in response to ε_0^p.

A final observation is that, even though we start with the assumption that noise terms are *i.i.d.*, the fact that agents try to correct their optimism by observing their signals makes their optimism comove, which generates non-trivial average higher-order optimism, as we claimed before.

Next question is why this higher-order optimism is important. How does it affect the outcome in period 1? We characterize the role of higher-order optimism in Lemma 9 and Corollary 2.

Lemma 9. *Higher-order beliefs can be written as functions of ε_1^p and cumulative sums of higher-order optimism:*

$$\mathbb{E}_{i1}\,\bar{\mathbb{E}}_1^{h-1}[a_1] = \varepsilon_1^p + \mathcal{O}_{i1} + \mathcal{O}_{i1}^2 + \cdots + \mathcal{O}_{i1}^h \quad (\textit{with} \quad \mathbb{E}_{i1}\,a_1 = \varepsilon_1^p + \mathcal{O}_{i1})$$
$$\bar{\mathbb{E}}_1^h[a_1] = \varepsilon_1^p + \mathcal{O}_1 + \mathcal{O}_1^2 + \cdots + \mathcal{O}_1^h$$

where we write $\bar{\mathbb{E}}_1[\cdot] = \int_0^1 \mathbb{E}_{i1}[\cdot]\,\mathrm{d}i$ and $\bar{\mathbb{E}}_1^h[\cdot] = \int_0^1 \mathbb{E}_{i1}\,\bar{\mathbb{E}}_1^{h-1}[\cdot]\,\mathrm{d}i$.

Corollary 2. *The aggregate output in period 1 is a weighted average of higher-order beliefs, and hence is a weighted sum of higher-order optimism:*

$$\begin{aligned} y_1 &= \sum_{h=1}^{\infty}(1-\alpha)\alpha^{h-1}\bar{\mathbb{E}}_1^h[a_1] \\ &= \varepsilon_1^p + \sum_{h=1}^{\infty}\alpha^{h-1}\mathcal{O}_1^h. \end{aligned} \tag{6}$$

It is well known that aggregate output is determined by higher-order beliefs. Thus, Lemma 9 naturally leads to Corollary 2. In Lemma 8, we characterize how underlying shocks affect higher-order optimism, which in conjunction with Corollary 2 characterizes how underlying shocks affect aggregate output in period 1. This essentially gives the equivalent result of Lemma 7, but we have tracked higher-order optimism and higher-order beliefs to understand the mechanism behind it. We are now ready to answer the main question of this section: do strategic complementarity and the resulting higher-order optimism strengthen or weaken our mechanism?

With $\alpha = 0$, we return to the case without strategic complementarity where we have $y_1 = \varepsilon_1^p + \mathcal{O}_1$. With $\alpha > 0$, we have additional higher-order optimism terms in equation (6). In Lemma 8, we saw that these additional terms move in the same direction as first-order optimism. Thus, we can conclude that the presence of strategic complementarity and the resulting higher-order optimism always *strengthen* our mechanism relative to the case without strategic complementarity. In other words, when agents observe higher-than-expected outcomes, they become optimistic not only about their signals (first-order optimism) but also about others' optimism (higher-order optimism). Furthermore, higher α means higher coefficients on the first-order and higher-order optimism terms in equation (6). Therefore, the response of y_1 to the underlying shocks ε_0^p and ε_0^u is even higher when we have stronger strategic complementarity. This discussion is summarized in the following theorem, which is our second main result.

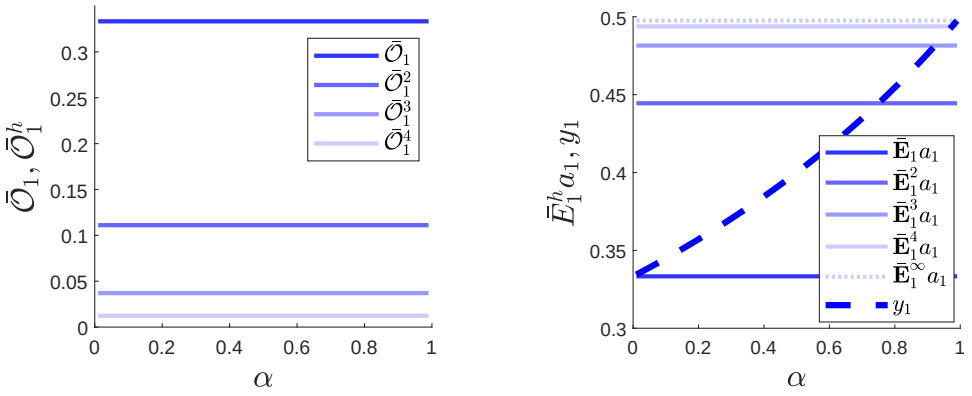
Theorem 2. *The effects of period-0 shocks on period-1 outcome are increasing in the degree of strategic complementarity:*[29]

$$\frac{\partial}{\partial\alpha}\left(\frac{\partial y_1}{\partial\varepsilon_0^u}\right) > 0 \quad \textit{and} \quad \frac{\partial}{\partial\alpha}\left|\frac{\partial y_1}{\partial\varepsilon_0^p}\right| > 0.$$

Thus, the strategic complementarity and the resulting higher-order optimism always strengthen the amplification and dampening we documented in Section 3.3.

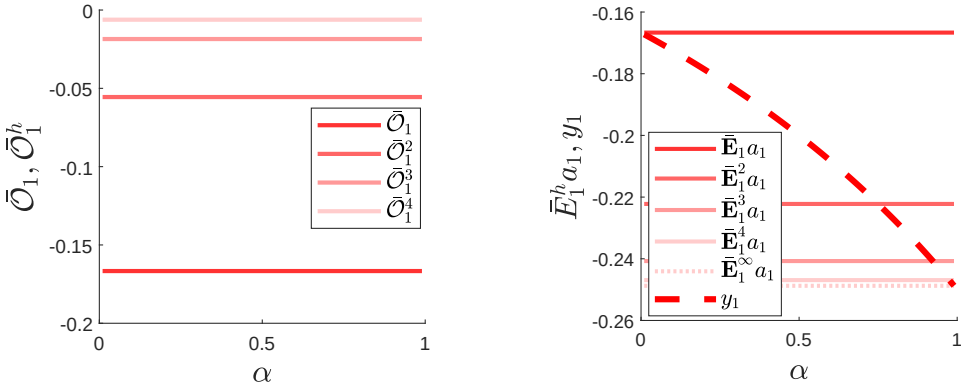
Remark. This theorem is in stark contrast to the results in Woodford (2003), Morris and Shin (2002), and Angeletos and Pavan (2007), which instead document that the higher the degree of strategic complementarity is, the less responsive the agents are to underlying shocks. This can be clearly seen in Example 1 where we have $\frac{\partial}{\partial\alpha}\left(\frac{\partial y}{\partial a}\right) < 0$.

[29] Recall that $\frac{\partial\mathcal{O}_1^h}{\partial\varepsilon_0^u}$ is positive while $\frac{\partial\mathcal{O}_1^h}{\partial\varepsilon_0^p}$ is negative for all orders h.



(a) Higher-Order Optimism (b) Higher-Order Beliefs and Outcome

Figure 1. Effects of a Unit Increase in Unobserved Shocks



(a) Higher-Order Optimism (b) Higher-Order Beliefs and Outcome

Figure 2. Effects of a Unit Increase in Partly-observed Shocks

This is because higher strategic complementarity implies that the equilibrium actions of agents are more anchored to the common prior, so agents are less responsive to contemporaneous shocks. In our model, however, optimistic agents are on average expect that others are more optimistic than they are, so higher strategic complementarity makes agents more responsive to period 0 shocks.

We can illustrate our findings using a parametrized example. We set $\sigma_u^2 = \sigma_\xi^2 = 1$ for the variance of the unobserved shock and noise terms, and $\sigma_p^2 = 2$ for the variance of the partly-observed shock.[30] We then change the degree of strategic complementarity α from 0 to 1. Figure 1 corresponds to unobserved shocks ε_0^u and Figure 2 to partly-observed shocks ε_0^p. These figures clearly illustrate our findings: (i) higher-order optimism moves in the same direction as the first-order optimism (Lemma 8), (ii) higher-order beliefs are cumulative sums of higher-order optimism (Lemma 9), and (iii) the effects of underlying shocks are increasing in α (Theorem 2).

[30] We set σ_p^2 higher than σ_u^2 according to the notion that agents are more concerned about volatile shocks; see footnote 20.

We conclude this section with comparative statics. We have seen so far that higher-order optimism determines higher-order beliefs, which in turn determine outcomes. Thus, we can focus on how the values of underlying parameters change the effects of shocks on higher-order optimism. The results are summarized in the following lemma.

Lemma 10. *The effect of the period-0 unobserved shock on higher-order optimism, $\partial \mathcal{O}_{i1}^h / \partial \varepsilon_0^u$, is*

(i) Increasing in σ_p^2

(ii) Decreasing in σ_u^2 if h is low (e.g., $h = 1$), while increasing in σ_u^2 if h is sufficiently high

(iii) Increasing in σ_ξ^2 if h is low (e.g., $h = 1$), while decreasing in σ_ξ^2 if h is sufficiently high.

Similarly, the effect of period-0 partly-observed shock on higher-order optimism, $\left|\partial \mathcal{O}_{i1}^h / \partial \varepsilon_0^p\right|$ is

(i) Increasing in σ_u^2

(ii) Decreasing in σ_p^2 if h is low (e.g., $h = 1$), while increasing in σ_p^2 if h is sufficiently high

(iii) Increasing in σ_ξ^2 if h is low (e.g., $h = 1$), while decreasing in σ_ξ^2 if h is sufficiently high.

Recall that we prove in Lemma 5 that the effect of the unobserved shock is increasing in the variance of the partly-observed shock and noise terms while decreasing in the variance of the unobserved shock. This explains the first part of Lemma 10 for the first-order optimism ($h = 1$). For comparative statics for higher-order optimism, we discussed in Lemma 8 that the effect of the unobserved shock on higher-order optimism can be decomposed into[31]

$$\frac{\partial \mathcal{O}_{i1}^h}{\partial \varepsilon_0^u} = \frac{\partial \mathcal{O}_{i1}}{\partial \varepsilon_0^u} \cdot Q_{1,2}^{h-1}.$$

The first term reflects the fact that higher-order optimism increases precisely because the first-order optimism increases. In addition, for a given increase in the first-order optimism, the second term determines the increase in higher-order optimism. Recall that this second term is increasing in both σ_p^2 and σ_u^2 and decreasing in σ_ξ^2. Why do we have different comparative statics for the first and second terms? Lemma 8 tells us that the first term originates from firm i's rational confusion between ε_0^u and (ε_0^p, ξ_i), while the second term is originated from other firms' rational confusion between their noise terms and $(\varepsilon_0^p, \varepsilon_0^u)$. Thus, the first term is decreasing in the relative variance of ε_0^u, and the second term is decreasing in the relative variance of ξ_i. If h is low, then the effect of variance parameters on the first term dominates that on the second term so that h^{th}-order optimism has the same comparative statics as the first-order optimism. On the other hand, for sufficiently high h, the effect of variance on the second term dominates

[31] Recall that we considered a unit innovation in ε_0^u, so we can interpret $\mathcal{O}_{i1}^h$ there as $\frac{\partial \mathcal{O}_{i1}^h}{\partial \varepsilon_0^u}$.

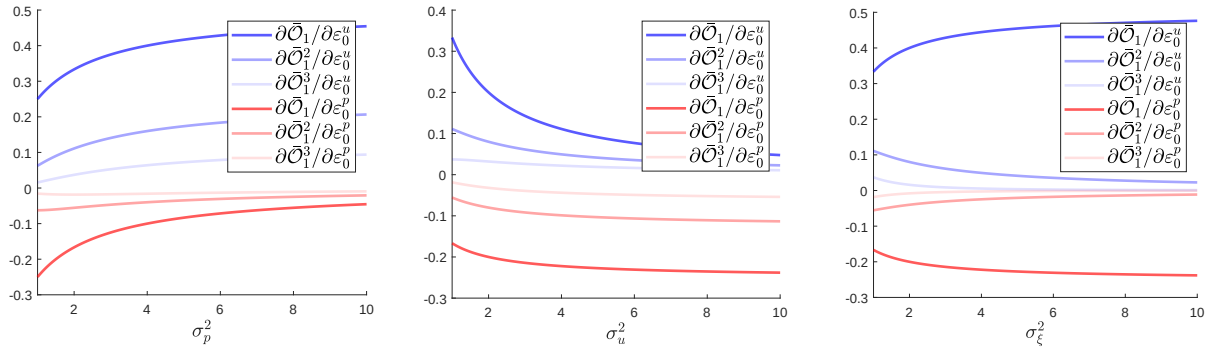


Figure 3. Comparative Statics with Respect to Variance

that on the first term and h^{th}-order optimism is increasing in σ_p^2 and σ_u^2 and decreasing in σ_ξ^2. This explains the first part of Lemma 10, and we can apply the same argument for the second part. Figure 3 illustrates the results. We use the same parameter values as in the previous numerical exercise and change the value of each variance one by one. We then calculate the effects of underlying shocks on higher-order optimism.

The aggregate output in period 1 is a function of all orders of optimism, and variance parameters, $\sigma_p^2, \sigma_u^2, \sigma_\xi^2$, can have different effects depending on the order of optimism. Lemma 11, however, shows that the effects on the first-term above always dominate the effects on the second term when it comes to the aggregate output.

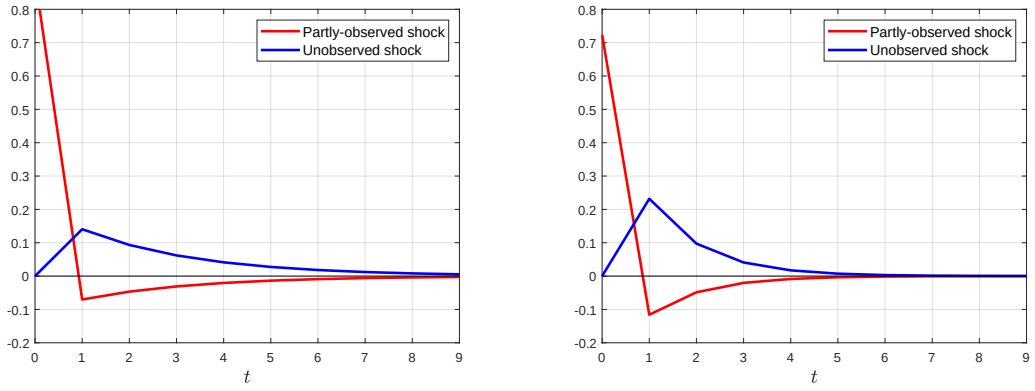
Lemma 11. *The effect of period-0 unobserved shocks on the period-1 aggregate outcome y_1 is (i) increasing in σ_p^2, (ii) decreasing in σ_u^2, and (iii) increasing in σ_ξ^2. Likewise, the effect of period-0 partly-observed shocks on the period-1 outcome is (i) increasing in σ_u^2, (ii) decreasing in σ_p^2, and (iii) increasing in σ_ξ^2.*

To sum up, the presence of strategic complementarity and the resulting higher-order optimism strengthen our mechanism as a result of two facts: Higher-order beliefs are cumulative sums of higher-order optimism, and higher-order optimism always move in the same direction as the first-order optimism in response to underlying shocks.

3.4.1 Numerical Exercise: Infinite Period with Strategic Complementarity

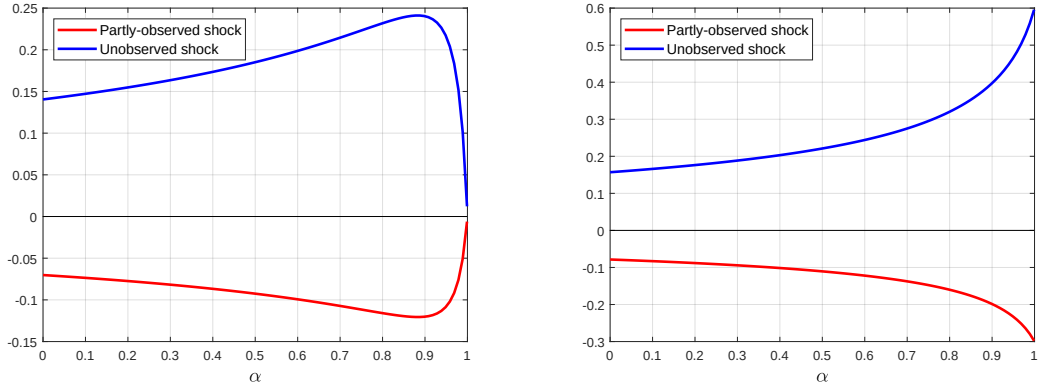
In this section, we discuss the robustness of the results in Section 3.4. We relax the restrictive two-period assumptions and instead assume infinite periods. In order to prevent firms from fully learning their noise terms, we assume as in Section 3.3 that noise terms follow an AR(1) process with $\rho \in (0, 1)$ and $\sigma_\eta^2 > 0$. Except for these two assumptions, the model is the same as in Section 3.4.

We utilize the method of Woodford (2003) to solve for the equilibrium dynamics of the aggregate output; see Appendix C.2 for details. We use the same parameters as in Section 3.4 with $\sigma_\eta^2 = 0.5$ and $\rho = 0.9$ and numerically calculate the trajectory of the economy after innovations in the underlying shocks. These parameters are arbitrary, but they are neither implausible nor qualitatively essential for the results below. Figure 4 plots the impulse responses



(a) $\alpha = 0$ (b) $\alpha = 0.8$

Figure 4. Impulse Response of Aggregate Output



(a) $\frac{\partial y_{t+1}}{\partial \varepsilon_t^p}$ and $\frac{\partial y_{t+1}}{\partial \varepsilon_t^u}$ (b) $\frac{\partial y_{t+1}}{\partial \varepsilon_t^p} \Big/ \frac{\partial y_t}{\partial \varepsilon_t^p}$ and $\frac{\partial y_{t+1}}{\partial \varepsilon_t^u} \Big/ \frac{\partial y_t}{\partial \varepsilon_t^p}$

Figure 5. Comparative Statics with Respect to α

of aggregate output to positive innovations in partly-observed and unobserved shocks. Figure 4a corresponds to the case without strategic complementarity, and Figure 4b corresponds to the case with strategic complementarity.

We can see that Lemma 7 continues to hold in this infinite horizon model: unobserved shocks are propagated to period 1, while partly-observed shocks have negative effects on the next period outcome. Also, comparing Figure 4a and Figure 4b, the effect of α is in line with Theorem 2: with higher degree of strategic complementarity, we seem to have stronger effects of period-t shocks on the period-$(t+1)$ outcome. However, if we plot $\frac{\partial y_{t+1}}{\partial \varepsilon_t^u}$ and $\frac{\partial y_{t+1}}{\partial \varepsilon_t^p}$ as a function of α in Figure 5a, then it turns out that this is not the case for very high values of α. In particular, both vanish as α converges to one. Does this mean that the intuition we obtain in the previous section is wrong? The answer is no. Note that the importance of underlying shocks goes to zero as α goes to 1, which can be clearly seen by $\lim_{\alpha \to 1} \frac{\partial y_t}{\partial \varepsilon_t^p} = 0$. For sufficiently high α, this force is dominant, so that the effects of period-t shocks on the period-$(t+1)$ output also go to zero. In Figure 5b, we plot the relative size of $\frac{\partial y_{t+1}}{\partial \varepsilon_t^u}$ and $\frac{\partial y_{t+1}}{\partial \varepsilon_t^p}$ compared to $\frac{\partial y_t}{\partial \varepsilon_t^p}$.

We can observe that these relative effects are indeed increasing in the degree of strategic complementarity, which is indeed in line with the result of Theorem 2. In this sense, we can conclude that the main message of Section 3.4 does not rely on its simplifying assumptions.

3.5 Implication on Forecast Survey Data

What is the empirical content of our model? In this section, we consider an abstract version of the model in Section 3.3, in which a_{it} can be interpreted as any fundamental of interest. In Section 3.5.1, we first illustrate how this model provides an alternative interpretation of the survey data, while explaining the prominent empirical findings in the literature in a unified way. In Section 3.5.2, we show that the gap between consensus-level and individual-level overextrapolation helps distinguish our rational theory of overextrapolation from other behavioral theories. This result is reminiscent of Angeletos, Huo, and Sastry's (2020) finding that the gap between consensus-level underreaction and individual-level overreaction speaks to the role of information frictions.

Model. As in Section 3.3, the fundamental follows an AR(1) process $a_{it} = \rho_a a_{it-1} + \varepsilon^p_{it} + \varepsilon^u_{it}$, and agents observe signals $s_{it} = \rho_a a_{it-1} + \varepsilon^p_{it} + \eta_{it}$. Thus, the optimism follows the law of motion $\tilde{\mathcal{O}}_{it+1} = \rho\gamma_3\tilde{\mathcal{O}}_{it} - \rho\gamma_1\varepsilon^p_{it} + \rho\gamma_2\varepsilon^u_{it} + \eta_{it+1}$. The forecast is then given by $\mathbb{E}_{it}[a_{it}] = \rho_a^2 a_{it-2} + (\rho_a - \rho K\gamma_1)\varepsilon^p_{it-1} + (\rho_a + \rho K\gamma_2)\varepsilon^u_{it-1} + K\varepsilon^p_{it} + K\eta_{it} + \rho\gamma_3 K\tilde{\mathcal{O}}_{it-1}$. Note that our timing convention implies that when agent i makes a forecast in period t, her information set $\Omega_{it} = (\cdot, s_{it-2}, a_{it-2}, s_{it-1}, a_{it-1}, s_{it})$ does not contain the realized fundamental a_{it}. Literature, however, often assumes that a_{it} is contained in the period-t information set, so we introduce a notation to make our results comparable to the literature: $\mathbb{F}_{it}[\cdot] \equiv \tilde{\mathbb{E}}_{it+1}[\cdot] = \mathbb{E}\left[\,\cdot \mid (\cdots, s_{it-2}, a_{it-2}, s_{it-1}, a_{it-1}, s_{it}, a_{it})\right]$. We denote aggregate variables by either omitting i-index or using a bar over variables: $x_t = \int x_{it}\,\mathrm{d}i$ for $x \in \{a, \varepsilon^p, \varepsilon^u, \tilde{\mathcal{O}}\}$, $\overline{\mathbb{E}_{it}\, a_{it+k}} = \int_0^1 \mathbb{E}_{jt}\, a_{jt+k}\,\mathrm{d}j$, and $\overline{\mathbb{F}_{it} a_{it+k}} = \int_0^1 \mathbb{F}_{jt} a_{jt+k}\,\mathrm{d}j$. We write $\sigma_p^2 = \mathrm{Var}(\varepsilon^p_{it})$, $\sigma_u^2 = \mathrm{Var}(\varepsilon^u_{it})$, $\bar{\sigma}_p^2 = \mathrm{Var}(\varepsilon^p_t)$ and $\bar{\sigma}_u^2 = \mathrm{Var}(\varepsilon^u_t)$. Note that we always have $\bar{\sigma}_p^2 \le \sigma_p^2$ and $\bar{\sigma}_u^2 \le \sigma_u^2$ since the shocks may have idiosyncratic components which are canceled out when we aggregate them. This means that the degrees of commonality defined below are always less than or equal to one.

Definition 5. *We define* the degrees of commonality *of the partly-observed and unobserved shocks as*

$$\mathcal{C}_p = \frac{\bar{\sigma}_p^2}{\sigma_p^2} \text{ and } \mathcal{C}_u = \frac{\bar{\sigma}_u^2}{\sigma_u^2}, \text{ respectively.}$$

On the one hand, if a shock is fully idiosyncratic and hence always take zero value when we integrate it across agents, the degree of commonality is zero. On the other hand, if a shock is common then the degree of commonality is one. The relative size of $\mathcal{C}_p$ and $\mathcal{C}_u$ plays a central role in Section 3.5.1. A special case is when agents share a common

fundamental a_t, such as the inflation rate of the economy or GDP growth. This necessarily implies $\mathcal{C}_p = \mathcal{C}_u = 1$. But, in the real world, even for such a common fundamental, forecast data can be better interpreted by a model with idiosyncratic fundamentals. To illustrate this, consider forecasters who form expectations about the US output growth. They have their own ways to view the US output growth, a_{it}, which is not necessarily the same as the true US output growth, a_t, even if it is unbiased, $\int_0^1 a_{it}\,\mathrm{d}i = a_t$. If this is the case, the feedback these forecasters receive is likely to be also in terms of their view of the US output growth, a_{it}, not in terms of the true US output growth, a_t. Thus, we assume hereafter that $\mathcal{C}_p$ and $\mathcal{C}_u$ are not necessarily equal to one even when we consider common fundamentals.

3.5.1 Empirical Findings in the Literature

Many empirical papers use panel survey data to measure agents' expectations directly. These papers often assume that forecasters do not observe past realizations, even ex post, and dynamically learn fundamentals from signals. This assumption is necessary because observing past realizations makes learning essentially static in their settings and makes it difficult to explain the dynamic pattern of forecast data. The literature often relies on rational inattention to justify this assumption. But, it is unlikely that forecasters who have made a prediction for a variable do not pay close attention to the realized value of it.

This paper gives a totally different way of interpreting dynamic forecast data. This paper views the same survey data as outputs of dynamic learning by forecasters, who can observe the past inflarion rates but are trying to learn how to interpret their own information; i.e., noise terms. Our model allows forecasters to observe past realizations ex post, while still being able to explain several empirical findings that have been explained using standard models. In this section, we first illustrate how our model explains the prominent empirical findings in the literature in a unified way. In particular, we consider the empirical findings of Coibion and Gorodnichenko (2015) (hereafter, CG), Kohlhas and Walther (2020) (hereafter, KW), and Angeletos, Huo, and Sastry (2020) (hereafter, AHS).

Coibion and Gorodnichenko (2015). We start with the finding of CG. They demonstrate the underreaction of the consensus forecast by showing that forecast errors are positively correlated with forecast revisions. They run the following regression

$$a_{it+k} - \mathbb{E}_{it}\, a_{it+k} = \alpha_i + \delta(\overline{\mathbb{E}_{it}\, a_{it+k}} - \overline{\mathbb{E}_{it-1}\, a_{it+k}}) + \text{error}_{it}$$

and obtain a positive coefficient estimate, $\hat{\delta} > 0$. They explain this by the gradual adjustment of average forecasts. The same result can be obtained in our model, but here it is based instead on the gradual adjustment of average optimism.

Proposition 3. *In our model, there exists a threshold $\lambda \in (0,1)$ such that $\frac{\mathcal{C}_p}{\mathcal{C}_u} > \lambda$ implies*

$$\mathrm{Cov}(a_{it+k} - \mathbb{E}_{it}\, a_{it+k}, \overline{\mathbb{E}_{it}\, a_{it+k}} - \overline{\mathbb{E}_{it-1}\, a_{it+k}}) > 0, \ \textit{for } k \geq 1,$$

This means that, unless partly observed shocks are mostly averaged out, we can obtain the underreaction of the consensus forecast as in CG.

Kohlhas and Walther (2020). KW show the coexistence of underreaction to new information and overextrapolation from recent realizations of the forecasted variable. The evidence for underreaction is the same as in CG, while for overextrapolation, they run the following regression for US output growth:[32]

$$a_{t+k} - \overline{\mathbb{E}_{it+1}\, a_{it+k}} = \alpha + \gamma a_t + \text{error}_t$$

and obtain a negative coefficient estimate, $\hat{\gamma} < 0$. This directly implies that consensus forecast features overextrapolation to the recent realization. They show that this can happen if rational agents pay more attention to procyclical components of the variable. A simpler explanation is based on a behavioral overextrapolation model in which agents' perceived persistence of the output growth is higher than the true persistence. In the next proposition, we will argue that we can obtain the same result using our model, in which agents overextrapolate to recent realizations because of the endogenous change in optimism.

Proposition 4. *Suppose that agents in our model are relatively well informed about idiosyncratic components of shocks in the sense that $\mathcal{C}_p < \mathcal{C}_u$. Then, we have*

$$\mathrm{Cov}(a_{t+k} - \overline{\mathbb{E}_{it+1}\, a_{it+k}}, a_t) < 0, \ \textit{for } k \geq 1.$$

Also note that the contemporaneous effect of a_t on the forecast error is positive:

$$\mathrm{Cov}(a_{t+k} - \overline{\mathbb{E}_{it}\, a_{it+k}}, a_t) > 0.$$

As discussed in Assumption 2, it is often assumed in the literature that agents are relatively well informed about idiosyncratic shocks. First, idiosyncratic shocks are more agent-specific, hence, it is easier to get information about them. Also, idiosyncratic shocks are likely to be more volatile than the aggregate shocks. Agents thus rationally pay more attention to the idiosyncratic shocks. This necessarily implies $\mathcal{C}_p < \mathcal{C}_u$. The first part of Proposition 4 says that our model predicts the finding of KW under this condition. The second part says that there is a reversal of covariance,

[32] Their original specification has $a_{t+k} - \bar{\mathbb{E}}_t a_{t+k}$ on the left hand side. Under their timing convention, however, a_t is in the agents' information set when they form the expectation about a_{t+k}; so it should be $\bar{\mathbb{E}}_{t+1} a_{t+k}$ under our timing convention.

$\text{Cov}(a_{t+k} - \overline{\mathbb{E}_{it+1}\, a_{it+k}}, a_t) < 0 < \text{Cov}(a_{t+k} - \overline{\mathbb{E}_{it}\, a_{it+k}}, a_t)$. This is reminiscent of Theorem 1, which states that a component that does not affect the period-t expectation has a larger effect on the period-$(t+1)$ expectation. This combination of overextrapolation and information friction is essential in many papers in the literature to explain the finding of KW. AHS explain it with the combination of behavioral overextrapolation and information friction. Our model and the model of KW essentially embed rational mechanisms of overextrapolation—persistent noise and feedback in our model and asymmetric attention in KW—into information friction models.

Angeletos, Huo, and Sastry (2020). AHS document delayed overreaction of consensus forecasts—consensus forecasts initially underreact and overshoot later on. This finding is consistent with their model, which combines incomplete information and behavioral over-extrapolation. Since my model provides a rational theory of the overextrapolation of consensus forecasts, we can obtain the same result. In our model, expectations initially underreact due to incomplete information and overshoot later on when agents receive feedback.[33]

3.5.2 Distinguish the Rational Theory from Behavioral Theories

Not only our model but also many behavioral theories of overextrapolation can obtain Proposition 4. Moreover, as AHS pointed out, these theories, combined with information friction, can potentially explain Proposition 3 as well. How can we test our model against other behavioral overextrpolation models? Smoking gun evidence comes from exploiting the difference between the degree of overextrapolation in consensus and individual forecasts. For example, consider the following two regressions.

$$a_{t+1} - \overline{\mathbb{E}_{it+1}\, a_{it+1}} = \beta_0 + \beta_{aggr} a_t + \text{error}_{t+1} \tag{7}$$

$$a_{it+1} - \mathbb{E}_{it+1} a_{it+1} = \beta_0 + \beta_{ind} a_{it} + \text{error}_{it+1} \tag{8}$$

Because a_{it} is contained in agent i's information set, our model can only generate overextrapolation at the consensus level, whereas in Proposition 5 we show that behavioral theories necessarily have the same coefficient for both regression specifications. We can compare the estimated coefficients of these regressions to distinguish our rational theory from behavioral theories.[34]

[33] They also show that behavioral over-extrapolation—misspecification in the stochastic process of *fundamental*—leads to the overreaction of individual forecasts as documented in Bordalo et al. (2020). Similarly, in our model, the misspecification in the stochastic process of *noise terms* leads to overreaction of individual forecasts. In particular, Proposition C.1 states that when the perceived persistence of noise is greater than the true persistence, individual forecast errors are negatively correlated with forecast revisions, implying the individual-level overreaction.

[34] Consider an extended version of KW model with individual-specific fundamental. Agent i has the fundamental $y_{it} = \sum_j x_{ijt}$, where j-th component is determined by $x_{ijt} = a_j \theta_{it} + u_{ijt}$ where θ_{it} denotes a latent factor that follows an AR(1) process, $\theta_{it} = \rho_a \theta_{it-1} + \eta_{it}$. Agent i observes noisy signals $s_{ijt} = x_{ijt} + \varepsilon_{ijt}$. The shocks u_{ijt}, η_{it}, and ε_{ijt} are normally distributed, serially uncorrelated, and mutually independent. In this model, we have $\widehat{\beta}_{aggr} < \widehat{\beta}_{ind}$ if and only if $\frac{\text{Var}(\int u_{ijt}\, \text{d}i)}{\text{Var}(u_{ijt})} > \frac{\text{Var}(\int \eta_{it}\, \text{d}i)}{\text{Var}(\eta_{it})}$. But there is no reason to expect $\widehat{\beta}_{ind} = 0$.

Proposition 5. *Suppose that the fundamental follows an AR(1) process*

$$a_{it} = \rho_a a_{it-1} + \varepsilon_{it}$$

and agents receive signals about the fundamental with normally distributed noise terms

$$s_{it} = a_{it} + \eta_{it}.$$

Consider the following three theories with behavioral elements

- *Extrapolation: Agents observe a_{it} (i.e., $\mathrm{Var}(\eta_{it}) = 0$) when they form expectations about a_{it+1}, but perceived AR(1) coefficient $\hat{\rho}_a$ is higher than the true one, ρ_a. We write $\mathbb{E}_{it+1}[\cdot] = \mathbb{E}[\cdot|\cdots, a_{it-1}, a_{it}]$.*
- *AHS: Perceived AR(1) coefficient, $\hat{\rho}_a$, is higher than the true one, ρ_a, and perceived precision of s_{it} is higher than the true one.*
- *Diagnostic expectation: $\mathbb{E}_{it}\, a_{i,t+k} = \mathbb{E}^{rational}_{it-1}\, a_{i,t+k} + g_k(s_{it} - \mathbb{E}^{rational}_{it-1}\, a_{it})$ with $g_k > K \cdot \rho^k$ where K is the Kalman gain.*

We always have

$$\widehat{\beta}_{aggr} = \widehat{\beta}_{ind}.$$

In our model, suppose again that agents are relatively well informed about idiosyncratic components of shocks in the sense that $\mathcal{C}_p < \mathcal{C}_u$. Then, we have

$$\widehat{\beta}_{aggr} < 0 \quad \textit{and} \quad \widehat{\beta}_{ind} = 0.$$

Gennaioli, Ma, and Shleifer (2016) report both $\widehat{\beta}_{aggr}$ and $\widehat{\beta}_{ind}$ for CFOs' and analysts' expectations on earnings growth, which is copied in Table 4, although their focus is not on comparing the coefficients. The coefficients in panel (A) correspond to $\widehat{\beta}_{aggr}$, and those in panel (B) correspond to $\widehat{\beta}_{ind}$. We can make two observations. First, analysts expectations feature a pattern consistent with Proposition 5; i.e., $\widehat{\beta}_{aggr} < 0$ and $\widehat{\beta}_{ind} \approx 0$. Second, CFOs expectations give higher extrapolation both at the consensus level and at the individual level. But the differences between two are approximately the same. These two observations are suggestive of the interpretation that analysts expectations are approximately rational but overextrapolate from past realizations once we aggregate them to consensus expectations, and that CFOs expectations are additionally subject to behavioral overextrapolation.[35] It is difficult, however, to formally map these estimates to the coefficients $\widehat{\beta}_{aggr}$ and $\widehat{\beta}_{ind}$ in Proposition 5 because Gennaioli, Ma, and Shleifer

[35] Another interesting observation is that the differences between coefficients, $\widehat{\beta}_{aggr} - \widehat{\beta}_{ind}$, which measure the extra overextrapolation in the consensus forecasts, are almost identical for analyst expectation and CFO expectation.

Table 4: Tables 8 and 9 of Gennaioli, Ma, and Shleifer (2016)

A. Aggregate Evidence

	Realized − Expected Next 12m Earnings Growth	
	(1) Analyst	(2) CFO
Past 12m earnings/asset (%)	−0.0456 (−3.68)	−0.0881 (−6.48)
Observations	106	57

B. Firm-Level Evidence

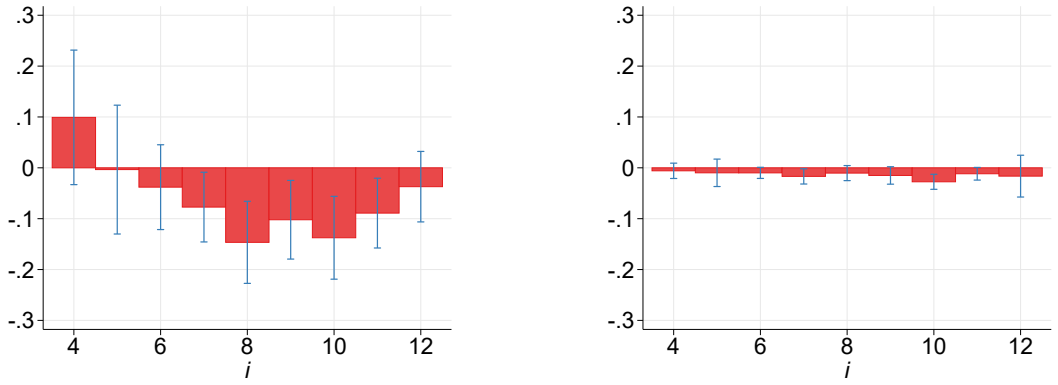
	Realized − Expected Next 12m Earnings Growth	
	(1) Analyst	(2) CFO
Past 12m earnings/asset (%)	−0.0080 (−7.43)	−0.0511 (−5.14)
Firm fixed effects	Y	Y
Observations	103,930	606

Notes: In panel (A), the dependent variable is aggregate earnings growth in the next 12 months minus aggregate expectations of earnings growth in the next 12 months. Independent variables include aggregate earnings/asset in the four quarters prior to quarter $t-1$. In panel (B), the dependent variable is firm-level earnings growth in the next 12 months minus expectations of earnings growth in the next 12 months. Independent variables include firm-level earnings/asset in the four quarters prior to quarter $t-1$. t-statistics in parentheses. See Gennaioli, Ma, and Shleifer (2016) for details.

(2016) regress forecast errors of earning growth on past earnings per asset, not on past earnings growth. Thus, we redo their estimation using past earnings growth as independent variables. One should be cautious when choosing the length of time periods because our theory essentially implies initial underextrapolation and overextrapolation later on (See Proposition 4). Suppose that a_{it} denotes firm-level earning growth over one year starting from time t. We experiment with various values of the length of time periods between t and $t+1$, from four quarters ($i=4$) to twelve quarters ($i=12$).[36] Figure 6 shows the estimated coefficients of the regression specifications (7) and (8). Reassuringly, this again features a pattern consistent with Proposition 5 for intermediate values of i, $7 \leq i \leq 11$.

Remark. The experience effects are studied by Malmendier and Nagel (2011, 2016) and subsequent papers. Recent evidence suggests longlasting effects of past personal experiences on expectations and behaviors. For example, personal lifetime experiences in the stock market affect future stock market investment behavior. This is inconsistent

[36] Because a_{it} denotes yearly earnings growth, we set $i \geq 4$ to ensure that there is no overlap between a_{it} and a_{it+1}.



(a) Aggregate Evidence: $\widehat{\beta}_{aggr}$ (b) Firm-Level Evidence: $\widehat{\beta}_{ind}$

Figure 6. Overextrapolation in Analyst Expectations

Notes: This figure plots the estimated coefficients of specifications (7) and (8). We vary the length of a unit time period, from four quarters ($i = 4$) to twelve quarters ($i = 12$).

with traditional economic models, in which there is no difference between personally experiencing an event and hearing about it. The literature on experience effects emphasizes the longlasting neuropsychology effects as a key mechanism. The following corollary provides a natural explanation of experience effects through the lens of our model.

Corollary 3. *In our model, suppose again that agents are relatively well informed about idiosyncratic components of shocks in the sense that $\mathcal{C}_p < \mathcal{C}_u$. Then, when we run the following regression*

$$a_{it+1} - \mathbb{E}_{it+1}\, a_{it+1} = \beta_0 + \tilde{\beta}_{aggr} a_t + \widetilde{\text{error}}_{it+1},$$

we have

$$\widehat{\tilde{\beta}}_{aggr} = \widehat{\beta}_{aggr} < 0.$$

Suppose you have experienced large negative stock market returns, $a_t < 0$. Then Corollary 3 implies that you become pessimistic about the stock market, $a_{it+1} - \mathbb{E}_{it+1}\, a_{it+1} > 0$, being less likely to participate in it. Moreover, expectations adjust only for those who experience this negative shock, because this overextrapolation arises from agents evaluating their previous forecasts based on the feedback they receive. Those who only hear about the negative shock have not had a chance to make a prediction, so they would not show overextrapolative behavior. In other words, our model provides a novel reason why the cognitive process of making a prediction and evaluating it affects the formation of future expectations.

3.6 Conclusion

We begin with two observations. First, noise in agents' signals is likely to be persistent regardless of its real-world counterpart. Second, in the real world, agents receive feedback on their past forecasts. With persistent noise and feedback, agents try to learn about the noise in their signals and the noise of others, and optimism arises endogenously. With this additional channel of learning, feedback on previous forecasts affects expectations about the noise, and shocks with different degrees of observability have different effects on the dynamics of aggregate outcomes through their different effects on optimism. We obtain a novel mechanism by which *rational agents become overoptimistic after observing higher-than-expected outcomes of the economy*, and this optimism amplifies/propagates the underlying shocks. Here, optimism is not only about one's own signals but also about others' optimism when there is strategic complementarity. Our model gives us a new way to interpret forecast dynamics in survey data—learning how to interpret information rather than learning fundamentals. This interpretation is consistent with many empirical findings in the literature.

References

Acemoglu, Daron, David Autor, David Dorn, Gordon H Hanson, and Brendan Price (2016) "Import Competition and the Great US Employment Sag of the 2000s," *Journal of Labor Economics*, 34 (S1), S141–S198.

Acemoglu, Daron, and Pascual Restrepo (2020) "Robots and Jobs: Evidence From US Labor Markets," *Journal of political economy*, 128 (6), 2188–2244.

Acharya, Sushant, Jess Benhabib, and Zhen Huo (2019) "The Anatomy of Sentiment-driven Fluctuations," (April).

Adão, Rodrigo (2016) "Worker Heterogeneity, Wage Inequality, and International Trade: Theory and Evidence from Brazil," *Working Paper* (November).

Adão, Rodrigo, Costas Arkolakis, and Federico Esposito (2020) "General Equilibrium Effects in Space: Theory and Measurement."

Adao, Rodrigo, Martin Beraja, and Nitya Pandalai-Nayar (2023) "Fast and Slow Technological Transitions,"Technical report, Tech. rep., MIT, Mimeo.

Adão, Rodrigo, Michal Kolesár, and Eduardo Morales (2019) "Shift-share designs: Theory and inference," *Quarterly Journal of Economics*, 134 (4), 1949–2010.

Agarwal, Sumit, J Bradford Jensen, and Ferdinando Monte (2020) "Consumer mobility and the local structure of consumption industries."

Aguirregabiria, Victor, and Pedro Mira (2010a) "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, 156 (1), 38–67.

——— (2010b) "Dynamic Discrete Choice Structural Models: A Survey," *Journal of Econometrics*, 156 (1), 38–67.

Ahlfeldt, Gabriel M, Stephen J Redding, Daniel M Sturm, and Nikolaus Wolf (2015) "The Economics of Density: Evidence From the Berlin Wall," *Econometrica*, 83 (6), 2127–2189.

Allen, Treb, and Costas Arkolakis (2014) "Trade and the Topography of the Spatial Economy," *The Quarterly Journal of Economics*, 129 (3), 1085–1140.

Allen, Treb, and Dave Donaldson (2020) "Persistence and Path Dependence in the Spatial Economy,"Technical report, National Bureau of Economic Research.

Almagro, Milena, and Tomás Domínguez-Iino (2020) "Location Sorting and Endogenous Amenities: Evidence from Amsterdam," (August).

Althoff, Lukas, Fabian Eckert, Sharat Ganapati, and Conor Walsh (2022) "The Geography of Remote Work," *Regional Science and Urban Economics*, 93, 103770.

Alvarez, Fernando E, Katarína Borovičková, and Robert Shimer (2016) "Decomposing Duration Dependence in a Stopping Time Model,"Technical report, National Bureau of Economic Research.

Amior, Michael, and Alan Manning (2018) "The Persistence of Local Joblessness," *American Economic Review*, 108 (7), 1942–1970.

Anas, Alex (2007) "A Unified Theory of Consumption, Travel and Trip Chaining," *Journal of Urban Economics*, 62 (2), 162–186.

Angeletos, George Marios, Zhen Huo, and Karthik A Sastry (2020) "Imperfect Macroeconomic Expectations : Evidence and Theory," (April 2).

Angeletos, George Marios, and Jennifer La'O (2009a) "Noisy Business Cycles," *NBER Macroeconomics Annual*, 24.

——— (2009b) "Incomplete information, higher-order beliefs and price inertia," *Journal of Monetary Economics*, 56, S19–S37.

Angeletos, George Marios, and Chen Lian (2019) "Confidence and the Propagation of Demand Shocks."

Angeletos, George Marios, and Alessandro Pavan (2007) "Efficient Use of Information and Social Value of Information," *Econometrica*, 75 (4), 1103–1142.

Arcidiacono, Peter, and John Bailey Jones (2003) "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71 (3), 933–946.

Arellano, Manuel, and Stéphane Bonhomme (2017) "Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models," *Annual Review of Economics*, 9, 471–496.

Arentze, Theo A, Harmen Oppewal, and Harry JP Timmermans (2005) "A Multipurpose Shopping Trip Model To Assess Retail Agglomeration Effects," *Journal of Marketing Research*, 42 (1), 109–115.

Arkolakis, Costas, Arnaud Costinot, and Andrés Rodríguez-Clare (2012) "New trade models, same old gains?," *American Economic Review*, 102 (1), 94–130.

Artuç, Erhan, Shubham Chaudhuri, and John McLaren (2010) "Trade shocks and labor adjustment: A structural empirical approach," *American Economic Review*, 100 (3), 1008–1045.

Artuç, Erhan, and John McLaren (2015) "Trade policy and wage inequality: A structural analysis with occupational and sectoral mobility," *Journal of International Economics*, 97 (2), 278–294.

Autor, David H, David Dorn, and Gordon H Hanson (2013a) "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American economic review*, 103 (6), 2121–2168.

——— (2013b) "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 103 (6), 2121–2168.

Autor, David H, David Dorn, Gordon H Hanson, and Jae Song (2014) "Trade Adjustment: Worker-level Evidence," *The Quarterly Journal of Economics*, 129 (4), 1799–1860.

Balboni, Clare Alexandra (2019) *In Harm's Way? Infrastructure Investments and the Persistence of Coastal Cities* Ph.D. dissertation, London School of Economics and Political Science.

Baqaee, David Rezza, and Emmanuel Farhi (2020) "Productivity and Misallocation in General Equilibrium," *The Quarterly Journal of Economics*, 135 (1), 105–163.

Barrero, Jose Maria, Nicholas Bloom, and Steven J Davis (2021) "Why Working From Home Will Stick,"Technical report, National Bureau of Economic Research.

Bartik, Timothy J (1991) "Who Benefits from State and Local Economic Development Policies?," *W.E. Upjohn Institute.*, 237–253.

Benhabib, Jess, Pengfei Wang, and Yi Wen (2015) "Sentiments and Aggregate Demand Fluctuations," *Econometrica*, 83 (2), 549–585.

Beraja, Martin (2023) "A Semistructural Methodology for Policy Counterfactuals," *Journal of Political Economy*, 131 (1), 190–201.

Bernardin Jr, Vincent L, Frank Koppelman, and David Boyce (2009) "Enhanced Destination Choice Models Incorporating Agglomeration Related To Trip Chaining While Controlling for Spatial Competition," *Transportation Research Record*, 2132 (1), 143–151.

Berry, Steven T (1994) "Estimating Discrete-choice Models of Product Differentiation," *The RAND Journal of Economics*, 242–262.

Bilal, Adrien, and Esteban Rossi-Hansberg (2021) "Location As an Asset," *Econometrica*, 89 (5), 2459–2495.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2022) "Discretizing Unobserved Heterogeneity," *Econometrica*, 90 (2), 625–643.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer (2020) "Overreaction in Macroeconomic Expectations," (Marcg), 1–72.

Borjas, George J (1987) "Self-Selection and the Earnings of Immigrants," *The American Economic Review*, 531–553.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2020) "Quasi-Experimental Shift-Share Research Designs."

Burstein, Ariel, Eduardo Morales, and Jonathan Vogel (2019) "Changes in Between-group Inequality: Computers, Occupations, and International Trade," *American Economic Journal: Macroeconomics*, 11 (2), 348–400.

Cain, Glen G (1976) "The Challenge of Segmented Labor Market Theories To Orthodox Theory: A Survey," *Journal of economic literature*, 14 (4), 1215–1257.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro (2019) "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock," *Econometrica*, 87 (3), 741–835.

Caliendo, Lorenzo, Luca David Opromolla, Fernando Parro, and Alessandro Sforza (2021) "Goods and Factor Market Integration: A Quantitative Assessment of the EU Enlargement," *Journal of Political Economy*, 129 (12), 3491–3545.

Card, David, Jesse Rothstein, and Moises Yi (2023) "Location, Location, Location," Working Paper 31587, National Bureau of Economic Research.

Chetty, Raj (2009) "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-form Methods," *Annu. Rev. Econ.*, 1 (1), 451–488.

Chetty, Raj, John N Friedman, Michael Stepner et al. (2020) "The Economic Impacts of COVID-19: Evidence From a New Public Database Built Using Private Sector Data,"Technical report, national Bureau of economic research.

Coibion, Olivier, and Yuriy Gorodnichenko (2015) "Information rigidity and the expectations formation process: A simple framework and new facts," *American Economic Review*, 105 (8), 2644–2678.

Combes, Pierre Philippe, Gilles Duranton, and Laurent Gobillon (2012) "The costs of agglomeration: House and land prices in French cities," 86 (4), 1556–1589.

Combes, Pierre-Philippe, and Laurent Gobillon (2015) "The Empirics of Agglomeration Economies," 5, 247–348.

Costinot, Arnaud, and Jonathan Vogel (2010) "Matching and Inequality in the World Economy," *Journal of Political Economy*, 118 (4), 747–786.

Couture, Victor (2016) "Valuing the Consumption Benefits of Urban Density," (September).

Couture, Victor, Cecile Gaubert, Jessie Handbury, and Erik Hurst (2021) "Income Growth and the Distributional Effects of Urban Spatial Sorting,"Technical report.

Couture, Victor, and Jessie Handbury (2020) "Urban Revival in America," *Journal of Urban Economics*, 119, 103267.

Davis, Donald R, Jonathan I Dingel, Joan Monras, and Eduardo Morales (2019) "How Segregated Is Urban Consumption?," *Journal of Political Economy*, 127 (4), 1684–1738.

Davis, Morris A, and François Ortalo-Magné (2011) "Household Expenditures, Wages, Rents," *Review of Economic Dynamics*, 14 (2), 248–261.

Desmet, Klaus, Dávid Krisztián Nagy, and Esteban Rossi-Hansberg (2018) "The Geography of Development," *Journal of Political Economy*, 126 (3), 903–983.

Dhingra, Swati, and John Morrow (2019) "Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity," *Journal of Political Economy*, 127 (1), 196–232.

Diamond, Rebecca (2016) "The Determinants and Welfare Implications of US Workers' Diverging Location Choices By Skill: 1980-2000," *American Economic Review*, 106 (3), 479–524.

Dix-Carneiro, Rafael (2014) "Trade Liberalization and Labor Market Dynamics," *Econometrica*, 82 (3), 825–885.

Dix-Carneiro, Rafael, João Paulo Pessoa, Ricardo Reyes-Heroles, and Sharon Traiberman (2023) "Globalization, Trade Imbalances, and Labor Market Adjustment," *The Quarterly Journal of Economics*, 138 (2), 1109–1171.

Dixit, Avinash K, and Joseph E Stiglitz (1977) "Monopolistic Competition and Optimum Product Diversity," *The American economic review*, 67 (3), 297–308.

Donaldson, Dave, and Richard Hornbeck (2016) "Railroads and American Economic Growth: A "Market Access" Approach," *Quarterly Journal of Economics*, 131 (2), 799–858.

Duguid, James, Bryan Kim, Lindsay Relihan, and Chris Wheat (2023) "The Impact of Work-from-Home On Brick-and-Mortar Retail Establishments: Evidence From Card Transactions," *Available at SSRN 4466607*.

Duranton, Gilles, and Jessie Handbury (2023) "Covid and Cities, Thus Far,"Technical report, National Bureau of Economic Research.

Duranton, Gilles, and Diego Puga (2004) "Micro-foundations of Urban Agglomeration Economies," in *Handbook of regional and urban economics*, 4, 2063–2117: Elsevier.

Dvorkin, Maximiliano A (2023) "International trade and labor reallocation: misclassification errors, mobility, and switching costs."

Eaton, B Curtis, and Richard G Lipsey (1979) "Comparison Shopping and the Clustering of Homogeneous Firms," *Journal of Regional Science*, 19 (4), 421–435.

Eaton, Jonathan, and Samuel Kortum (2018) "Trade in Goods and Trade in Services," in *World Trade Evolution*, 82–125: Routledge.

Fan, Jingting, Sungwan Hong, and Fernando Parro (2023) "Learning and Expectations in Dynamic Spatial Economies,"Technical report, National Bureau of Economic Research.

French, Eric, and Christopher Taber (2011) "Identification of Models of the Labor Market," in *Handbook of labor economics*, 4, 537–617: Elsevier.

Galle, Simon, Andrés Rodríguez-Clare, and Moises Yi (2023) "Slicing the Pie: Quantifying the Aggregate and Distributional Effects of Trade," *The Review of Economic Studies*, 90 (1), 331–375.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer (2016) "Expectations and Investment," *NBER Macroeconomics Annual*, 30 (1), 379–431.

Gentzkow, Matthew (2007) "Valuing new goods in a model with complementarity: Online newspapers," *American Economic Review*, 97 (3), 713–744.

Glaeser, Edward L, Jed Kolko, and Albert Saiz (2001) "Consumer City," *Journal of economic geography*, 1 (1), 27–50.

Goldberg, Pinelopi Koujianou, and Nina Pavcnik (2007) "Distributional Effects of Globalization in Developing Countries," *Journal of economic Literature*, 45 (1), 39–82.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020) "Bartik Instruments: What, When, Why, and How," *American Economic Review*, 110 (8), 2586–2624.

Granger, Clive WJ, and Michael J Morris (1976) "Time Series Modelling and Interpretation," *Journal of the Royal Statistical Society: Series A (General)*, 139 (2), 246–257.

Grigsby, John R (2022) "Skill Heterogeneity and Aggregate Labor Market Dynamics,"Technical report, National Bureau of Economic Research.

Handbury, J., and D. E. Weinstein (2015) "Goods Prices and Availability in Cities," *The Review of Economic Studies*, 82 (1), 258–296.

Heckman, James J (1981) "Heterogeneity and State Dependence," in *Studies in labor markets*, 91–140: University of Chicago Press.

Heckman, James J, and Bo E Honore (1990) "The Empirical Content of the Roy Model," *Econometrica: Journal of the Econometric Society*, 1121–1149.

Heckman, James J, and Guilherme Sedlacek (1985) "Heterogeneity, Aggregation, and Market Wage Functions: an Empirical Model of Self-selection in the Labor Market," *Journal of political Economy*, 93 (6), 1077–1125.

Heckman, James, and Burton Singer (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica: Journal of the Econometric Society*, 271–320.

Henry, Neil W (1971) "The Retention Model: a Markov Chain with Variable Transition Probabilities," *Journal of the American Statistical Association*, 66 (334), 264–267.

Howard, Greg, and Hansen Shao (2023) "The Dynamics of Internal Migration: A New Fact and Its Implications."

Hsieh, Chang-Tai, Erik Hurst, Charles I Jones, and Peter J Klenow (2019) "The Allocation of Talent and Us Economic Growth," *Econometrica*, 87 (5), 1439–1474.

Hulten, Charles R (1978) "Growth Accounting with Intermediate Inputs," *The Review of Economic Studies*, 45 (3), 511–518.

Huo, Zhen, and Naoki Takayama (2015) "Higher Order Beliefs, Confidence, and Business Cycles."

——— (2018) "Rational Expectations Models with Higher Order Beliefs," (February).

Jovanovic, Boyan (1979) "Job Matching and the Theory of Turnover," *Journal of political economy*, 87 (5, Part 1), 972–990.

Kambourov, Gueorgui, and Iourii Manovskii (2013) "A Cautionary Note On Using (March) Current Population Survey and Panel Study of Income Dynamics Data To Study Worker Mobility," *Macroeconomic Dynamics*, 17 (1), 172–194.

Kasa, Kenneth (2000) "Forecasting the Forecasts of Others in the Frequency Domain," *Review of Economic Dynamics*, 3, 726–756.

Keane, Michael P, and Eswar S Prasad (1996) "The Employment and Wage Effects of Oil Price Changes: A Sectoral Analysis," *The Review of Economics and Statistics*, 389–400.

Keane, Michael P, and Kenneth I Wolpin (1997) "The Career Decisions of Young Men," *Journal of political Economy*, 105 (3), 473–522.

Kennan, John, and James R Walker (2011) "The Effect of Expected Income On Individual Migration Decisions," *Econometrica*, 79 (1), 211–251.

Kim, Bumsoo, Marc de la Barrera, and Masao Fukui (2023) "Currency Pegs, Trade Imbalances and Unemployment: A Reevaluation of the China Shock."

Kim, Ryan, and Jonathan Vogel (2020) "Trade and Welfare (Across Local Labor Markets)," (April).

——— (2021) "Trade Shocks and Labor Market Adjustment," *American Economic Review: Insights*, 3 (1), 115–130.

Kleinman, Benny, Ernest Liu, and Stephen J Redding (2023) "Dynamic Spatial General Equilibrium," *Econometrica*, 91 (2), 385–424.

Kohlhas, Alexandre N, and Ansgar Walther (2020) "Asymmetric Attention," (September).

Konishi, Hideo (2005) "Concentration of Competing Retail Stores," *Journal of Urban economics*, 58 (3), 488–512.

Koster, Hans RA, Ilias Pasidis, and Jos van Ommeren (2019) "Shopping Externalities and Retail Concentration: Evidence From Dutch Shopping Streets," *Journal of Urban Economics*, 114, 103194.

Lagakos, David, and Michael E Waugh (2013) "Selection, Agriculture, and Cross-country Productivity Differences," *American Economic Review*, 103 (2), 948–980.

Lee, Donghoon (2005) "An Estimable Dynamic General Equilibrium Model of Work, Schooling, and Occupational Choice," *International Economic Review*, 46 (1), 1–34.

Lee, Donghoon, and Kenneth I Wolpin (2006) "Intersectoral Labor Mobility and the Growth of the Service Sector," *Econometrica*, 74 (1), 1–46.

Lee, Eunhee (2020) "Trade, Inequality, and the Endogenous Sorting Ofheterogeneous Workers," *Journal of International Economics*, 125, 103310.

Leonardi, Marco, and Enrico Moretti (2023) "The Agglomeration of Urban Amenities: Evidence From Milan Restaurants," *American Economic Review: Insights*, 5 (2), 141–157.

Lipsey, Robert E (2009) "1. Measuring International Trade in Services," in *International Trade in Services and Intangibles in the Era of Globalization*, 27–74: University of Chicago Press.

Lise, Jeremy, and Fabien Postel-Vinay (2020) "Multidimensional Skills, Sorting, and Human Capital Accumulation," *American Economic Review*.

Lucas, Robert E (1972) "Expectations and the neutrality of money," *Journal of Economic Theory*, 4 (2), 103–124.

Mackowiak, Bartosz, and Mirko Wiederholt (2009) "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, 99 (3), 769–803.

Malmendier, Ulrike, and Stefan Nagel (2011) "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?," *The quarterly journal of economics*, 126 (1), 373–416.

——— (2016) "Learning From Inflation Experiences," *The Quarterly Journal of Economics*, 131 (1), 53–87.

Manski, Charles F (1993) "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 60 (3), 531–542.

McFadden, Daniel (1973) "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontier in Econometrics*.

McKay, Alisdair, and Christian K Wolf (2022) "What Can Time-Series Regressions Tell Us About Policy Counterfactuals?,"Technical report, National Bureau of Economic Research.

Milgrom, Paul, and Ilya Segal (2002) "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70 (2), 583–601.

Miyauchi, Yuhei, Kentaro Nakajima, and Stephen J Redding (2022) "The Economics of Spatial Mobility: Theory and Evidence Using Smartphone Data,"Technical Report 1836, London School of Economics CEP Discussion Paper.

Monte, Ferdinando, J Bradford Jensen, and Sumit Agarwal (2020) "Consumer Mobility and the Local Structure of Consumption Industries."

Morris, Stephen, and Hyun Song Shin (2002) "Social Value of Public Information," *American Economic Review*, 92 (5).

Moscarini, Giuseppe, and Kaj Thomsson (2007) "Occupational and Job Mobility in the US," *The Scandinavian Journal of Economics*, 109 (4), 807–836.

Muñoz, Mathilde (2023) "Trading Non-Tradables: The Implications of Europe's Job Posting Policy."

Myatt, David P, and Chris C Wallace (2004) "Adaptive play by idiosyncratic agents," *Games and Economic Behavior*, 48, 124–138.

Nimark, Kristoffer P (2017) "Dynamic Higher Order Expectations," (March).

Oh, Ryungha, and Jaeeun Seo (2023) "What Causes Agglomeration of Services? Theory and Evidence From Seoul," *Working Paper*, https://ryunghaoh.github.io/files/Services.pdf.

Pavan, Alessandro (2016) "Attention, Coordination, and Bounded Recall," *Discussion Papers* (July).

Pierce, Justin R, and Peter K Schott (2016) "The Surprisingly Swift Decline of US Manufacturing Employment," *American Economic Review*, 106 (7), 1632–1662.

Porzio, Tommaso, Federico Rossi, and Gabriella Santangelo (2022) "The Human Side of Structural Transformation," *American Economic Review*, 112 (8), 2774–2814.

Primerano, Frank, Michael AP Taylor, Ladda Pitaksringkarn, and Peter Tisato (2008) "Defining and Understanding Trip Chaining Behaviour," *Transportation*, 35 (1), 55–72.

Relihan, Lindsay (2022) "Is Online Retail Killing Coffee Shops? Estimating the Winners and Losers of Online Retail Using Customer Transaction Microdata,"Technical Report 1836, London School of Economics CEP Discussion Paper.

Rodriguez-Clare, Andres, Mauricio Ulate, and Jose P Vasquez (2022) "Trade with Nominal Rigidities: Understanding the Unemployment and Welfare Effects of the China Shock."

Rondina, Giacomo, and Todd B Walker (2018) "Confounding Dynamics," (October).

Rosenthal, Stuart S, and William C Strange (2004) "Evidence On the Nature and Sources of Agglomeration Economies," in *Handbook of regional and urban economics*, 4, 2119–2171: Elsevier.

Rust, John (1987a) "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

——— (1987b) "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

Sargent, Thomas J (1991) "Equilibrium with signal extraction from endogenous variables," *Journal of Economic Dynamics and Control*, 15, 245–273.

Silva, JMC Santos, and Silvana Tenreyro (2006) "The Log of Gravity," *The Review of Economics and statistics*, 88 (4), 641–658.

Sims, Christopher A (2003) "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Sprung-Keyser, Ben, Nathaniel Hendren, and Sonya Porter (2022) *The Radius of Economic Opportunity: Evidence From Migration and Local Labor Markets*: US Census Bureau, Center for Economic Studies.

Stokey, Nancy L (2008) *The Economics of Inaction: Stochastic Control Models with Fixed Costs*: Princeton University Press.

Takahashi, Takaaki (2013) "Agglomeration in a City with Choosy Consumers Under Imperfect Information," *Journal of Urban Economics*, 76, 28–42.

Topalova, Petia (2010) "Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India," *American Economic Journal: Applied Economics*, 2 (4), 1–41.

Townsend, Robert M (1983) "Forecasting the Forecasts of Others," *Journal of Political Economy*, 91 (4), 546–588.

Traiberman, Sharon (2019) "Occupations and import competition: Evidence from Denmark," *American Economic Review*, 109 (12), 4260–4301.

Train, Kenneth E. (2003) *Discrete Choice Methods with Simulation* (October), Cambridge, UK, 1–334.

Tsivanidis, Nick (2019) "Evaluating the impact of urban transit infrastructure: Evidence from bogota's transmilenio."

Valentinyi, Ákos, and Berthold Herrendorf (2008) "Measuring factor income shares at the sectoral level," *Review of Economic Dynamics*, 11 (4), 820–835.

Verboven, Frank (1996) "The nested logit model and representative consumer theory," *Economics Letters*, 50 (1), 57–63.

Woodford, Michael (2003) "Imperfect Common Knowledge and the Effects of Monetary Policy," in *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*.

Table of Contents: Appendix

Appendix A

Sectoral Shocks and Labor Market Dynamics: A Sufficient Statistics Approach

A.1 Theoretical Appendix

A.1.1 Aggregation Result: Proposition 1

In this section, we first prove a general result that does not rely on Assumption 2. In the next section, we show how Assumption 2 simplifies this result. To state the general result, we define another type of worker flow matrices. As we will demonstrate, two types of worker flow matrices are sufficient statistics for characterizing the welfare and labor market consequences of sectoral shocks.

Definition A.1. *For each $m, k \in \mathbb{N}$, the* (m,k)-period worker flow matrix*, is an $S \times S$ matrix denoted as $\mathcal{F}_{m,k}$ whose (i,j)-element is given by*

$$(\mathcal{F}_{m,k})_{i,j} = \Pr(s_{\tau(t,m)+k} = j | s_t = i),$$

where the random variable $\tau(t,m) \equiv \min\{\tau \geq t : s_\tau = s_{t-m}\}$ denotes the first period in which a worker returns to the sector she chose m periods ago.

The (i,j)-element of the (m,k)-period worker flow matrix equals the steady-state probability that a randomly selected worker from sector i will move to sector j after k periods after returning to the sector she chose m periods ago.

To formalize the idea of aggregation, we define the population-average operator.

Definition A.2. *The* population-average operator $\bar{\mathbb{E}}_\omega$ *is an operator that can be applied to* $S \times N$ *matrices for any* $N \in \mathbb{N}$. *It maps a type-specific matrix to an aggregate matrix,*

$$\begin{pmatrix} a_{11}^\omega & \cdots & a_{1N}^\omega \\ \vdots & \ddots & \vdots \\ a_{S1}^\omega & \cdots & a_{SN}^\omega \end{pmatrix} \mapsto \sum_\omega \begin{pmatrix} \tilde{\ell}_1^\omega a_{11}^\omega & \cdots & \tilde{\ell}_1^\omega a_{1N}^\omega \\ \vdots & \ddots & \vdots \\ \tilde{\ell}_S^\omega a_{S1}^\omega & \cdots & \tilde{\ell}_S^\omega a_{SN}^\omega \end{pmatrix}$$

where $\tilde{\ell}_i^\omega = \frac{\ell_i^\omega}{\sum_{\omega'} \ell_i^{\omega'}} = \Pr(\omega | s_t = i)$ *is the steady-state proportion of type* ω *in sector* i.

If the i-th row of a matrix contains variables related to sector i, then the steady-state type distribution of sector i gives the appropriate weights for computing the average across different types. This is precisely how the population-average operator is defined. The following lemma shows that certain type-specific variables can be converted to their aggregate equivalents through application of the population-average operator.[1]

Lemma A.1. *If we apply the population-average operator to* $\mathrm{d} \ln \ell_{t+1}^\omega$, $\mathrm{d}v_t^\omega$, $(F^\omega)^k$, *or* $(B^\omega)^m (F^\omega)^k$, *we obtain aggregate variables:*

$$\begin{aligned} \bar{\mathbb{E}}_\omega \, \mathrm{d} \ln \ell_{t+1}^\omega &= \mathrm{d} \ln \ell_{t+1} \\ \bar{\mathbb{E}}_\omega \, \mathrm{d}v_t^\omega &= \mathrm{d}v_t \\ \bar{\mathbb{E}}_\omega \left[(F^\omega)^k \right] &= \mathcal{F}_k \\ \bar{\mathbb{E}}_\omega \left[(B^\omega)^m (F^\omega)^k \right] &= \mathcal{F}_{m,k}. \end{aligned}$$

In particular, with infinite-length longitudinal information on workers' sector choices, we can observe $\bar{\mathbb{E}}_\omega[(F^\omega)^k]$ and $\bar{\mathbb{E}}_\omega[(B^\omega)^m (F^\omega)^k]$ for all $k, m \in \mathbb{N}_0$.[2] We can now apply the population-average operator to the left- and right-hand sides of equations (7) and (8) to derive a general version of the sufficient statistics result. First, we have

$$\begin{aligned} \mathrm{d}v_t = \bar{\mathbb{E}}_\omega \, \mathrm{d}v_t^\omega &= \bar{\mathbb{E}}_\omega \Big[\sum_{k \geq 0} (\beta F^\omega)^k \, \mathbb{E}_t \, \mathrm{d}w_{t+k} \Big] \\ &= \sum_{k \geq 0} \beta^k \bar{\mathbb{E}}_\omega \Big[(F^\omega)^k \Big] \, \mathbb{E}_t \, \mathrm{d}w_{t+k} \\ &= \sum_{k \geq 0} \beta^k \mathcal{F}_k \, \mathbb{E}_t \, \mathrm{d}w_{t+k}. \end{aligned}$$

[1] The population-average operator does not transform all type-specific variables into their aggregate counterparts. For example, nonlinear functions of $\mathrm{d} \ln \ell_{t+1}^\omega$, $\mathrm{d}v_t^\omega$, or $(F^\omega)^k$ do not possess this property (e.g., $\bar{\mathbb{E}}_w \, \mathrm{d}\ell_{t+1}^\omega \neq \mathrm{d}\ell_{t+1}$).

[2] For the second type of worker flow matrices, this requires an additional assumption that $(\mathcal{F}_k)_{i,j}$ is strictly positive for all $i, j \in \mathcal{S}$, and $k \in \mathbb{N}$.

Second, we have

$$\begin{aligned}
\mathrm{d}\ln \ell_t &= \bar{\mathbb{E}}_\omega \, \mathrm{d}\ln \ell_t^\omega \\
&= \bar{\mathbb{E}}_\omega \left[\frac{\beta}{\rho} \sum_{s\geq 0} (B^\omega)^s (I - B^\omega F^\omega) \left(\sum_{k\geq 0} (\beta F^\omega)^k \, \mathbb{E}_{t-s-1} \, \mathrm{d}w_{t-s+k}^\omega \right) \right] \\
&= \bar{\mathbb{E}}_\omega \left[\sum_{s\geq 0, k\geq 0} \frac{\beta^{k+1}}{\rho} \left((B^\omega)^s (F^\omega)^k - (B^\omega)^{s+1} (F^\omega)^{k+1} \right) \mathbb{E}_{t-s-1} \, \mathrm{d}w_{t-s+k} \right] \\
&= \sum_{s\geq 0, k\geq 0} \frac{\beta^{k+1}}{\rho} (\mathcal{F}_{s,k} - \mathcal{F}_{s+1,k+1}) \, \mathbb{E}_{t-s-1} \, \mathrm{d}w_{t-s+k}.
\end{aligned}$$

The result is summarized in the following proposition.

Proposition A.1 (Sufficient Statistics Result without Assumption 2). *Suppose that Assumption 1 holds. For a given sequence of (common) changes in sectoral wages* $\{\mathrm{d}w_t\}$*, the changes in sectoral value and sectoral employment are given by*

$$\begin{aligned}
\mathrm{d}v_t &= \sum_{k\geq 0} \beta^k \mathcal{F}_k \, \mathbb{E}_t \, \mathrm{d}w_{t+k}, \\
\mathrm{d}\ln \ell_t &= \sum_{s\geq 0, k\geq 0} \frac{\beta^{k+1}}{\rho} (\mathcal{F}_{s,k} - \mathcal{F}_{s+1,k+1}) \, \mathbb{E}_{t-s-1} \, \mathrm{d}w_{t-s+k}.
\end{aligned}$$

Proposition A.1 establishes that in order to construct the counterfactual changes in welfare and sectoral employment for a given sequence of sectoral wage changes, we only require knowledge of the two types of worker flow matrices, $\{\mathcal{F}_k\}$ and $\{\mathcal{F}_{s,k}\}$, and the parameter ρ that governs the dispersion of the idiosyncratic shocks. Proposition 2 can be generalized in a similar manner.

A.1.2 Simplified Sufficient Statistics Result under Assumption 2

Lemma A.1 demonstrates that with infinite-length longitudinal information on workers' sector choices, we can observe two types of worker flow matrices, $\{\mathcal{F}_k\}$ and $\{\mathcal{F}_{m,k}\}$, for all k and m. In practice, however, we can only observe them for small enough k and m due to the finite nature of the real-world datasets. The following lemma characterizes a certain symmetry property satisfied by the second type of worker flow matrices that can be used to reduce the data requirements.

Lemma A.2. *We have*

$$\begin{pmatrix} \ell_1 & & 0 \\ & \ddots & \\ 0 & & \ell_N \end{pmatrix} \mathcal{F}_{m,k} = (\mathcal{F}_{k,m})^\mathsf{T} \begin{pmatrix} \ell_1 & & 0 \\ & \ddots & \\ 0 & & \ell_N \end{pmatrix}.$$

However, the data requirements for estimating the second type of worker flow matrices are still very demanding. In this section, we show how these data requirements can be significantly reduced under an additional assumption.

Lemma A.3. *Under Assumption 2, we have $B^\omega = F^\omega$ for all $\omega \in \Omega$. In this case, we have $\mathcal{F}_{m,k} = \mathcal{F}_{m+k}$.*

Assumption 2 imposes a certain structure on the sector-switching costs. Lemma A.3 shows that this assumption implies the equality between the backward transition matrix and the forward transition matrix, which in turn implies that the second type of worker flow matrices reduces to the first type of worker flow matrices.[3] Thus, we can focus on the first type of worker flow matrices, and Proposition A.1 is simplified to Proposition 1 in the main text.

In the left panel of Figure A.2, we compute the transition matrices between the four sectors considered in the main text and plot the 16 elements of the backward transition matrices against the corresponding elements of the forward transition matrices. All elements lie closely on the 45-degree line. It is not possible to test this implication empirically at the unobserved type level. Instead, we compute the backward and forward transition matrices for observed types. Specifically, we consider four dimensions of observed heterogeneity, as in Section 1.4: gender, race, education, and age, which results in 16 groups. In the right panel of Figure A.2, we compare the matrices for all 16 groups. Again, there is a striking similarity between the backward and forward transition matrices.

A.1.3 First-Order Approximation of Labor Demand Side

Suppose that type-specific sectoral wages are determined by the labor allocation and exogenous shocks:

$$w_{it}^\omega = f_i^\omega\big(\{\ell_{jt}\}_j, \{\varepsilon_{jt}\}_j\big)$$

for $i \in \mathcal{S}$ and $\omega \in \Omega$. Up to a first-order approximation, we can write

$$\mathrm{d}w_{it}^\omega = \sum_j \frac{\partial f_i^\omega}{\partial \ln \ell_{jt}} \cdot \mathrm{d}\ln \ell_{jt} + \sum_j \frac{\partial f_i^\omega}{\partial \varepsilon_{jt}} \mathrm{d}\varepsilon_{jt}.$$

Assumption 1 requires that $\mathrm{d}w_{it}^\omega = \mathrm{d}w_{it}$ for all $\omega \in \Omega$ for any realizations of $\{\ell_{jt}\}_j$ and $\{\varepsilon_{jt}\}_j$. This in turn requires that

$$\mathrm{d}w_{it}^\omega = \mathrm{d}w_{it} \equiv \sum_j \frac{\partial f_i^{\omega_1}}{\partial \ln \ell_{jt}} \cdot \mathrm{d}\ln \ell_{jt} + \sum_j \frac{\partial f_i^{\omega_1}}{\partial \varepsilon_{jt}} \mathrm{d}\varepsilon_{jt}$$

[3] Assumption 2 is almost necessary and sufficient in the sense that when the number of sectors is three, the necessary and sufficient condition is $C_{12} + C_{23} + C_{31} = C_{13} + C_{32} + C_{21}$. When the number of sectors is four, the necessary and sufficient condition is $C_{ij} + C_{jk} + C_{ki} = C_{ik} + C_{kj} + C_{ji}$ for all distinct $i, j, k \in \{1, 2, 3, 4\}$; and for all distinct i, j (and remaining k, ℓ), we have either $C_{ij} = C_{ji}$ or $C_{ik} + C_{j\ell} = C_{i\ell} + C_{jk}$.

for all $\omega \in \Omega$ for a given $\omega_1 \in \Omega$.[4] Thus, we can write

$$\begin{aligned} \mathrm{d}w_t &= \begin{pmatrix} \frac{\partial f_1^{\omega_1}}{\partial \ln \ell_{1t}} & \frac{\partial f_1^{\omega_1}}{\partial \ln \ell_{2t}} & \cdots & \frac{\partial f_1^{\omega_1}}{\partial \ln \ell_{St}} \\ \frac{\partial f_2^{\omega_1}}{\partial \ln \ell_{1t}} & \frac{\partial f_2^{\omega_1}}{\partial \ln \ell_{2t}} & \cdots & \frac{\partial f_2^{\omega_1}}{\partial \ln \ell_{St}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_S^{\omega_1}}{\partial \ln \ell_{1t}} & \frac{\partial f_S^{\omega_1}}{\partial \ln \ell_{2t}} & \cdots & \frac{\partial f_S^{\omega_1}}{\partial \ln \ell_{St}} \end{pmatrix} \mathrm{d} \ln \ell_t + \begin{pmatrix} \frac{\partial f_1^{\omega_1}}{\partial \varepsilon_{1t}} & \frac{\partial f_1^{\omega_1}}{\partial \varepsilon_{2t}} & \cdots & \frac{\partial f_1^{\omega_1}}{\partial \varepsilon_{St}} \\ \frac{\partial f_2^{\omega_1}}{\partial \varepsilon_{1t}} & \frac{\partial f_2^{\omega_1}}{\partial \varepsilon_{2t}} & \cdots & \frac{\partial f_2^{\omega_1}}{\partial \varepsilon_{St}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_S^{\omega_1}}{\partial \varepsilon_{1t}} & \frac{\partial f_S^{\omega_1}}{\partial \varepsilon_{2t}} & \cdots & \frac{\partial f_S^{\omega_1}}{\partial \varepsilon_{st}} \end{pmatrix} \mathrm{d}\varepsilon_t. \\ &= D \cdot \mathrm{d} \ln \ell_t + E \cdot \mathrm{d}\varepsilon_t. \end{aligned}$$

A.1.4 Intuition for the Sufficient Statistics Result

We begin by providing intuition at the micro (i.e., type) level, then show how this intuition is preserved at the macro level when type-specific equations are aggregated.

Intuition at the Micro Level. Key equations at the micro level are (5) and (6). As discussed in the main text, equation (5) is just an application of the envelope theorem (e.g., Milgrom and Segal, 2002). The envelope theorem implies that changes in future optimal sector choices do not contribute to the change in current welfare. Thus, we can evaluate the effect of changes in future sectoral wages using workers' sector-choice probabilities as weights. For example, equation (7) shows that the effect of a unit change in the period-$(t+k)$ wage of sector j on the period-t welfare of workers in sector i is given by the probability that these workers transition to sector j in period $t+k$ (after discounting the future):

$$\frac{\partial v_{it}^{\omega}}{\partial \mathbb{E}_t \, w_{j,t+k}^{\omega}} = \beta^k \Pr(s_{t+k} = j | s_t = i, \omega) \equiv \beta^k (F^{\omega})^k.$$

Thus, the matrix $(F^{\omega})^k$ is—by providing information about workers' k-period sector choices—informative about the effect of future sectoral shocks on current welfare.

Equation (6) summarizes the labor reallocation in response to sectoral shocks, which is rewritten here for convenience:

$$\mathrm{d} \ln \ell_{t+1}^{\omega} = B^{\omega} \, \mathrm{d} \ln \ell_t^{\omega} + \tfrac{\beta}{\rho^{\omega}} (I - B^{\omega} F^{\omega}) \, \mathbb{E}_t \, \mathrm{d} v_{t+1}^{\omega}. \tag{6}$$

The first term captures the mechanical effect of changes in labor allocation in the previous period, which holds under any assumption about the distribution of the idiosyncratic shocks. On the other hand, the second term captures the response of workers' sector choices to changes in sectoral values in period $t+1$. In particular, it implies

$$\left. \frac{\partial \ln \ell_{it+1}^{\omega}}{\partial \mathbb{E}_t \, v_{jt+1}^{\omega}} \right|_{\text{fix } \ell_t} = \tfrac{\beta}{\rho} \Big(\mathbb{1}_{i=j} - \big(B^{\omega} F^{\omega} \big)_{ij} \Big).$$

[4] Note that this does not necessarily imply that $w_{it}^{\omega} = w_{it}$ for all $\omega \in \Omega$.

To understand this equation, suppose first that $i \neq j$. Then, for a given value of $\frac{\beta}{\rho}$, this means that the semi-elasticity of employment in sector i with respect to the expected value of sector j is proportional to the (i,j) element of $B^\omega F^\omega$. Note that we have

$$\begin{aligned}(B^\omega F^\omega)_{ij} &= \sum_{k \in \mathcal{S}} B^\omega_{ik} F^\omega_{kj} \\ &= \sum_{k \in \mathcal{S}} \Pr(s_t = k | s_{t+1} = i, \omega) \Pr(s_{t+1} = j | s_t = k, \omega).\end{aligned}$$

Thus, this element can be interpreted as the probability that workers who choose $s_{t+1} = i$ will switch to $s_{t+1} = j$ if they are allowed to choose period-$(t+1)$ sector again after redrawing period-$(t+1)$ idiosyncratic shocks. Since their sector choice depends on the realization of idiosyncratic shocks, this probability is non-zero. If this probability is high for a given sector pair (i,j), it means that there are many workers who are at the margin between sectors i and j. In this case, a small change in the value of sector j leads to a relatively larger decline in employment in sector i.

Second, consider $i = j$. Then for a similar reason, an increase in the value of sector i leads to a relatively larger increase in sector i employment when many workers are at the margin between sector i and other sectors, which is measured by $1 - (B^\omega F^\omega)_{ii}$.

Moving to the longer run, consider changes in the values in period $t+1$ that are known to workers in period $t+1-\tau$, $\tau \in \mathbb{N}$. To make the intuition clear, we assume that $\beta = 1$, but the intuition remains the same for $\beta < 1$. Then, the effect of the shock on period t labor allocation can be written as follows:

$$\begin{aligned}\frac{\partial \ln \ell^\omega_{it+1}}{\partial \, \mathbb{E}_{t+1-\tau} \, v^\omega_{jt+1}} &= \tfrac{1}{\rho}\Big((I - B^\omega F^\omega) + (B^\omega F^\omega - (B^\omega)^2 (F^\omega)^2) + \cdots + ((B^\omega)^{k-1}(F^\omega)^{k-1} - (B^\omega)^k (F^\omega)^k) \Big)_{ij} \\ &= \tfrac{1}{\rho}\Big(\mathbb{1}_{i=j} - \big((B^\omega)^k (F^\omega)^k\big)_{ij} \Big).\end{aligned}$$

Again, the (i,j)-element of the matrix $(B^\omega)^k (F^\omega)^k$ can be interpreted as the probability that workers who choose sector i in period $t+1$ will switch to sector j if they are allowed to choose sectors again for periods $t+2-\tau, t+3-\tau, \cdots, t+1$ after redrawing idiosyncratic shocks for these periods. This again measures the k-period indifference of workers between sectors i and j, and thus provides information on the responsiveness of workers to shocks known k-periods ahead of time.

In this sense, the steady-state transition matrices and their kth powers are—by providing information about the indifference of workers between sectors–informative about the response of sectoral employment to sectoral shocks. The role of the extreme-value assumption is to make this qualitative relationship exact. As an alternative justification, we show in Appendix A.1.4 that we have the same result up to a multiplicative constant in the limit of idiosyncratic volatility converging to zero.

Intuition at the Macro Level. At the micro level, the intuition can be formulated in terms of products of transition matrices. Appendix A.1.1 demonstrates that we can aggregate type-specific equations to derive macro-level equations, establishing the sufficient statistics result. In particular, Lemma A.3 shows that the products of transition matrices can be aggregated to two types of worker flow matrices. This preserves the same intuition at the macro level: The matrix $\mathcal{F}_k$ is—by providing information about the *average* worker's k-period sector choices—informative about the effect of future sectoral shocks on current welfare; and the matrix $\mathcal{F}_{s,k}$ is—by providing information about the *average* workers' indifference between sectors—informative about the response of sectoral employment to sectoral shocks.

A.1.4.1 Small Idiosyncratic Shock Limit

Our goal is to show that a version of equation (6) is valid in the limit as ρ approaches zero, without any distribution assumption on the idiosyncratic shocks. Without loss of generality, consider a two-period version of the model in Section 1.2, where two periods are denoted by t and $t+1$ for notation consistent with equation (6). To make the limit nontrivial, we modify the canonical model in two ways. First, we perturb the sectoral values of workers in period $t+1$, which are written as $\Delta_s = \beta \mathbb{E}_t V_{st+1}$ for notational simplicity. Fix a positive number ε. For each pair of sectors $i \neq j \in \mathcal{S}$, consider a set of workers, $\Omega_{ij}(\varepsilon)$, who have $|\Delta_i - \Delta_j| < \varepsilon$ and $\Delta_s < \min\{\Delta_i, \Delta_j\} - \varepsilon$ for all $s \in \mathcal{S} \backslash \{i, j\}$. Since we will take a limit along which the importance of idiosyncratic shocks converges to zero, we *assume* that workers in the set $\Omega_{ij}(\varepsilon)$ never choose a sector other than i and j. For workers in $\Omega_{ij}(\varepsilon)$, the relative value of sector i to j is distributed as follows:[5]

$$\Delta_{ij} \equiv \Delta_i - \Delta_j \sim g_{ij}(\cdot), G_{ij}(\cdot).$$

Second, switching costs are given by $C_{ij} = \rho \cdot \tilde{C} \cdot \mathbb{1}_{i \neq j}$.[6] Finally, we dispense with the distributional assumption on the idiosyncratic shock: Idiosyncratic utility from sector i relative to sector j follows an unknown distribution:

$$\rho \cdot \varepsilon_{it} - \rho \cdot \varepsilon_{jt} \equiv \rho \cdot \varepsilon_{ij} \quad \text{where } \varepsilon_{ij} \sim f(\cdot), F(\cdot).$$

Within this environment, the following proposition establishes that equation (6) holds up to $o(\rho)$ and up to a multiplicative constant, which depends only on the shape of the distribution of idiosyncratic shocks and the switching costs.

[5] We consider workers with $|\Delta_{ij}| < \bar{\varepsilon}$, so this is a local distribution around 0.
[6] If the switching costs are fixed along the limit of $\rho \to 0$, no workers would change sectors over time.

Proposition A.2. *When $g_{ij}(\cdot)$ is continuous around 0 and ε_{ij} has a finite first moment for all $i, j \in \mathcal{S}$, we have*

$$\frac{\partial \ln \ell_{t+1}}{\partial \mathbb{E}_t v_{t+1}} = \textit{const.} \cdot \tfrac{\beta}{\rho} \cdot (I - BF) + o(\rho),$$

where the constant term only depends on the shape of $F(\cdot)$ and $\tilde{C}$.

A.1.5 Single Crossing Condition

We know from Lemma 2 that, compared with the canonical model calibrated by matching the one-period worker flow matrix, the model with worker heterogeneity implies lower values of b_k at least for $k = 0$ and $k = 1$:

$$b_0 = \left(\mathcal{F}_0 - \mathcal{F}_2\right)_{ss} = 1 - (\mathcal{F}_2)_{ss}$$
$$b_1 = \left(\mathcal{F}_1 - \mathcal{F}_3\right)_{ss} = (\mathcal{F}_1)_{ss} - (\mathcal{F}_3)_{ss}.$$

Thus, $\bar{k} \geq 1$. In Figure A.3, we plot the difference between $\{b_k\}$ implied by the canonical model and those observed in the data (extrapolated using the method described in the main text). Initially, the canonical model yields larger values of b_k, but eventually it leads to smaller values compared with those implied by the data (although not plotted, this is true for all values of k greater than 9; i.e., there is no more crossing).

A.1.6 Response to a One-time Shock

Consider a one-time negative shock to a sector $s \in \mathcal{S}$ that is known to agents in period 1:

$$\mathrm{d}w_{s\tau} = -\Delta < 0 \tag{A.1}$$

for given $\tau > 1$. The effect of any series of negative shocks to sector s can be calculated as the sum of the effects of such one-time negative shocks.

Effects on Sectoral Welfare. With worker heterogeneity, workers initially employed in sector s are more likely to stay in sector s when the shock hits the sector. Thus, they suffer more from the one-time shock.

Proposition A.3. *Consider a one-time negative shock to sector s of the form* (A.1) *known to agents in period* 1. *The canonical model, calibrated by matching the one-period worker flow matrix, underestimates the negative welfare effect on workers initially employed in sector s, $\mathrm{d}v_{s1}$.*

This result in turn implies that for any series of negative shocks to sector s, the canonical model underestimates the negative welfare effect on workers initially employed in sector s, proving Proposition 3.

Effects on Sectoral Employment. The following proposition characterizes the condition under which the canonical model overestimates the decline in employment in sector s in period $t > 1$.

Proposition A.4. *Consider a one-time negative shock to sector s of the form* (A.1) *known to agents in period* 1. *Under Assumption 3, there exists a decreasing function $B : \mathbb{N} \to \mathbb{N}$ such that the canonical model overestimates the decline in employment in sector s in period t in response to the shock if and only if $|t - \tau| \leq B(t \wedge \tau)$, where $t \wedge \tau$ denotes the minimum of t and τ.*

When $|t-\tau|$ and/or $t \wedge \tau$ are small, the canonical model calibrated by matching the one-period worker flow matrix overestimates the decline in employment in sector s in period t in response to the shock, $\frac{\partial \ln \ell_{s,t}}{\partial w_{s,\tau}}$.[7]

The result implies that whether the models without worker heterogeneity overestimate or underestimate labor reallocation depends on the time horizon. On the one hand, as discussed in Lemma 2, the canonical model overestimates the mobility of workers across sectors, leading to overestimation of the decline in employment in a negatively affected sector. This intuition is what Corollary 4 describes when $|t - \tau|$ and/or $t \wedge \tau$ are small; i.e., when the shock is recently known or when the period affected by the shock is close to the period of interest. On the other hand, in the canonical model, workers have relatively lower probabilities of remaining in a sector, which implies that their sector choices in a given period do not have long-lasting impacts on their future sector choices. This aspect works in the opposite direction to our previous intuition and can become dominant if $|t - \tau|$ and/or $t \wedge \tau$ are large enough. For example, suppose that t is much larger than τ. A negative shock to a sector's wage in period τ reduces employment in the sector around that period. However, this reduced employment has only a limited impact on employment in the sector in the distant period t in the canonical model. Figure A.4 plots the relative size of the decline in employment of sector s in period t in response to a shock to the period τ wage predicted by the canonical model and that implied by the data. As expected, the decline is overestimated in the canonical model for small t or τ or small $|t - \tau|$ (red cells). In short, the canonical model tends to overestimate the short-term impact of shocks on sectoral employment but underestimates their long-term effects. In particular, within a 7-year time horizon, the canonical model consistently overestimates the impact of shocks on sectoral employment.

A.1.7 Two-Sector Model

In this section we consider a special case of our model with two sectors, $\mathcal{S} = \{1, 2\}$. This special case serves two purposes. First, it highlights the restriction on the worker flow matrix series imposed by the class of dynamic discrete

[7] If w is log wage, this measures the elasticity of sectoral employment with respect to sectoral wages.

choice models studied in this paper. Second, the analytical results derived in this section will be used for proofs in Appendix A.4.

For simplicity, we assume that there are a finite number of worker types. The types are indexed by $i = 1, \cdots, I$, and the population share of type i is $\theta_i \in (0,1)$. We denote the transition matrix of type-i workers by

$$F_i = \begin{pmatrix} \bar{\alpha}_i & \alpha_i \\ \beta_i & \bar{\beta}_i \end{pmatrix}$$

where $\bar{\alpha}_i = 1 - \alpha_i$ and $\bar{\beta}_i = 1 - \beta_i$. We also write $F_i^k \equiv \begin{pmatrix} \bar{\alpha}_{i,k} & \alpha_{i,k} \\ \beta_{i,k} & \bar{\beta}_{i,k} \end{pmatrix}$. For example, we have

$$F_i^2 = \begin{pmatrix} \bar{\alpha}_i^2 + \alpha_i\beta_i & \alpha_i(\bar{\alpha}_i + \bar{\beta}_i) \\ \beta_i(\bar{\alpha}_i + \bar{\beta}_i) & \bar{\beta}_i^2 + \alpha_i\beta_i \end{pmatrix} \text{ and } F_i^3 = \begin{pmatrix} \cdot & \alpha_i(\bar{\alpha}_i^2 + \bar{\alpha}_i\bar{\beta}_i + \bar{\beta}_i^2 + \alpha_i\beta_i) \\ \beta_i(\bar{\alpha}_i^2 + \bar{\alpha}_i\bar{\beta}_i + \bar{\beta}_i^2 + \alpha_i\beta_i) & \cdot \end{pmatrix}.$$

By induction, we can obtain a general formula for the off-diagonal elements of the matrix F_i^k

Lemma A.4. *F_i^k has the following form:*

$$F_i^k = \begin{pmatrix} 1 - \alpha_i f^k(\bar{\alpha}_i + \bar{\beta}_i) & \alpha_i f^k(\bar{\alpha}_i + \bar{\beta}_i) \\ \beta_i f^k(\bar{\alpha}_i + \bar{\beta}_i) & 1 - \beta_i f^k(\bar{\alpha}_i + \bar{\beta}_i) \end{pmatrix}$$

where $f^k(x) = \frac{1-(x-1)^k}{2-x}$.

The steady-state sectoral employment share of type i workers are given by

$$\begin{pmatrix} \Pr(\text{sector 1}|\text{type } i) \\ \Pr(\text{sector 2}|\text{type } i) \end{pmatrix} = \begin{pmatrix} \frac{\beta_i}{\alpha_i+\beta_i} \\ \frac{\alpha_i}{\alpha_i+\beta_i} \end{pmatrix} \equiv \begin{pmatrix} \tilde{\beta}_i \\ \tilde{\alpha}_i \end{pmatrix},$$

which gives

$$\tilde{\ell}_1^i = \frac{\tilde{\beta}_i\theta_i}{\sum_j \tilde{\beta}_j\theta_j} \text{ and } \tilde{\ell}_2^i = \frac{\tilde{\alpha}_i\theta_i}{\sum_j \tilde{\alpha}_j\theta_j}.$$

Thus, the k-period worker flow matrix is given by

$$\mathcal{F}_k = \begin{pmatrix} 1 - \frac{\sum_i \tilde{\beta}_i\theta_i\alpha_{i,k}}{\sum_i \tilde{\beta}_i\theta_i} & \frac{\sum_i \tilde{\beta}_i\theta_i\alpha_{i,k}}{\sum_i \tilde{\beta}_i\theta_i} \\ \frac{\sum_i \tilde{\alpha}_i\theta_i\beta_{i,k}}{\sum_i \tilde{\alpha}_i\theta_i} & 1 - \frac{\sum_i \tilde{\alpha}_i\theta_i\beta_{i,k}}{\sum_i \tilde{\alpha}_i\theta_i} \end{pmatrix} \equiv \begin{pmatrix} 1 - \sum_i x_i\alpha_{i,k} & \sum_i x_i\alpha_{i,k} \\ \sum_i y_i\beta_{i,k} & 1 - \sum_i y_i\beta_{i,k} \end{pmatrix}$$

where $x_i = \frac{\tilde{\beta}_i\theta_i}{\sum_j \tilde{\beta}_j\theta_j}$ and $y_i = \frac{\tilde{\alpha}_i\theta_i}{\sum_j \tilde{\alpha}_j\theta_j}$.

Finally, we use the results to characterize a restriction imposed on the off-diagonal elements of the worker flow matrix series.

Proposition A.5. *We have*

$$\frac{(\mathcal{F}_k)_{1,2}}{(\mathcal{F}_k)_{2,1}} = \frac{\alpha_1 x_1}{\beta_1 y_1} \quad \textit{for all } k \geq 1.$$

This proposition shows that the ratio between the two off-diagonal elements is identical for all worker flow matrix series. Thus, if these rations are far from constant in the data, we cannot match the worker flow matrix series observed in the data with the class of models we consider in this paper. The result in Appendix A.2.1 demonstrates that this is not the case.

A.2 Structural Estimation Appendix

A.2.1 Extrapolation Using the Structural Model

We estimate the structural model by matching the observed worker flow matrices. Specifically, we estimate the number of worker types along with their respective steady-state instantaneous utility vectors w_i^ω and switching costs C_{ij}^ω by matching 18 worker flow matrices (i.e., 216 moments). Note from the worker's sector-choice problem (1) that only the ratio between these values and the parameter ρ can be identified from the observed worker flow matrices. Thus, we only estimate these ratios until we estimate the value of ρ in Section 1.4.4. Following Assumption 2, we impose symmetry on the switching costs. The estimation process involves two steps: We first maximize the likelihood of observing $\{\mathcal{F}_k\}_{k=1}^{18}$ to estimate $\{\frac{1}{\rho}w_i^\omega, \frac{1}{\rho}C_{ij}^\omega\}$ for a given number of worker types, then use the Bayesian information criterion to determine the number of worker types.

Table A.1 shows the estimation result. The Bayesian information criterion supports the model with two worker types. Figure A.1 plots the resulting transition matrix for each type of worker. The first type has a comparative advantage in non-manufacturing sectors and low switching costs. Thus, workers of this type switch sectors frequently, as indicated by the small diagonal elements of the transition matrix in Figure A.1. In contrast, the second type has much higher switching costs, so workers of this type rarely move to other sectors. In Appendix A.2.4, we show how to interpret the figures in Table A.1. In particular, paying one unit of switching costs means paying 3.25% of lifetime consumption. Thus, the switching costs in Table A.1 are at most less than 20% of lifetime consumption. This is smaller than the estimates of ACM, who find that the *average* switching cost is at least 20% of lifetime consumption. Our estimates are close to those of Artuç and McLaren (2015), in which the switching costs are distributed around 12% of lifetime consumption.[8]

We also estimate primitives of the canonical by matching the one-period worker flow matrix, $\mathcal{F}_1$ (see Figure A.7 for the results). All parameters, including elements of the transition matrix, lie between the corresponding parameters for the model with two worker types.

Model Fit. The fit of the model with two worker types, documented in Figures 3 and 4, is surprising for two reasons. First, the degrees of freedom (19 parameters) are much smaller than the number of moments we target (216 moments).[9] Second, the dynamic discrete choice framework imposes systematic restrictions on the model-implied

[8] As Artuç and McLaren (2015) argue, one reason for the smaller estimates is the inclusion of sector-specific nonpecuniary benefits in the model, which are absent in ACM's model.

[9] Suppose we want to match only the 1-year worker flow matrix. We can perfectly match this matrix with only one type of worker if we can choose an arbitrary transition matrix for this type. In terms of degrees of freedom, we match $N(N-1)$ values with $N(N-1)$ parameters, where $N = 4$ is the number of sectors. Suppose we also want to match one more worker flow matrix. This exercise can be seen as matching the level and slope of the dots in Figure 2. At least in terms of degrees of freedom, we can achieve this with only two types of workers: matching $2N(N-1)$ values with $2N(N-1)+1$ parameters ($N(N-1)$ parameters for each transition matrix, and 1 for the type share). However, we also want to

Table A.1: Estimation Results

$\lvert\Omega\rvert = 2$	Type ω_1 (31.1%)		Type ω_2 (68.9%)	
Sector	Wage	Switching Cost	Wage	Switching Cost
Agri/Const	0.58	0.00 1.53 1.69 1.55	1.02	0.00 4.62 5.62 5.46
Manufacturing	0.60	1.53 0.00 1.33 1.41	1.02	4.62 0.00 4.88 4.93
Commu/Trade	0.66	1.69 1.33 0.00 0.98	1.00	5.62 4.88 0.00 3.72
Services/Others	0.77	1.55 1.41 0.98 0.00	1.07	5.46 4.93 3.72 0.00

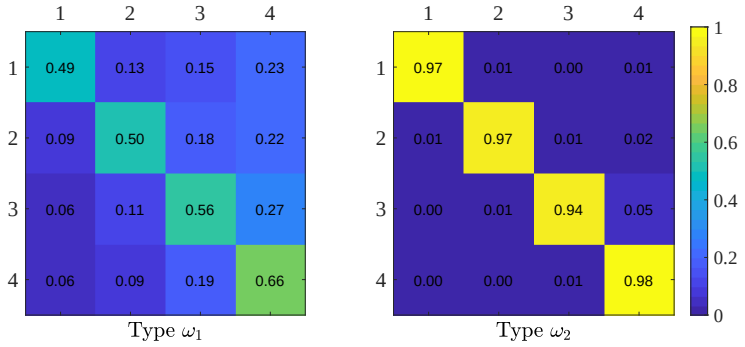


Figure A.1. Type-Specific Transition Matrix

worker flow matrices, so we would not be able to match every worker flow matrix series even with an arbitrary number of worker types.

The flexibility due to worker heterogeneity, characterized in Lemma 2, is necessary to match the observed worker flow matrices. As seen in Section 1.4.2, the canonical model with one type of workers fails to match the observed worker flow matrices. Table A.1 clearly reveals why the canonical model significantly underestimates longer-run staying probabilities. Workers of the second type rarely change sectors and have comparative advantage in manufacturing. Thus, conditioning on the fact that workers have previously self-selected into the manufacturing sector greatly increases the probability that they are the second type, and thus increases the probability that they will stay or choose again the manufacturing sector in subsequent periods. While the fit of the model improves substantially with two types of workers, the additional increase in fit from adding more types of workers is negligible, causing the Bayesian information criterion to choose the two-type worker model. See Figure A.8 for comparison of the fits of the two-type model and the five-type model.

Identification. Comparing the fits of the models does not necessarily identify the true number of worker types, let alone the fact that two worker types cannot capture the multifaceted nature of real-world worker differences. However, a key feature of our approach is that it does not require identification of all elements of the true model. As long as the estimated model closely approximates the observed worker flow matrix series, our sufficient statistics result

match the overall shape of the dots in Figure 2, and confine ourselves to the case in which the type-specific transition matrices are generated by the structural primitives. Thus, we end up with 19 parameters that can be used to match 216 moments.

ensures that the model always provides valid counterfactuals for the outcome of interest—namely, aggregate welfare and employment. This feature distinguishes our approach from the latent variable approach in the literature, such as the finite mixture model and k-means clustering (e.g., Arcidiacono and Jones, 2003; Heckman and Singer, 1984; Bonhomme, Lamadon, and Manresa, 2022), in which the validity of counterfactual predictions requires a higher level of confidence in identification.

A.2.2 Estimation of ρ

Section 1.4.4 proposes a method to estimate the parameter ρ, which is based on the second equation of Proposition 1:

$$\mathrm{d}\ln \ell_t = \sum_{s\geq 0, k\geq 0} \tfrac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathbb{E}_{t-s-1}\,\mathrm{d}w_{t-s+k}.$$

We refer to this relation as a forward-looking infinite-order MA process, because it resembles infinite-order MA processes, but involving forward-looking variables.

We first use equations (2) and (3) to prove equation (16) under the homogeneous worker assumption. Imposing Assumption 1 and assumptions in Lemma A.3, these equations become

$$\begin{aligned} \mathrm{d}v_t^\omega &= \mathrm{d}w_t + \beta F^\omega\,\mathbb{E}_t\,\mathrm{d}v_{t+1}^\omega \\ \mathrm{d}\ln \ell_{t+1}^\omega &= F^\omega\,\mathrm{d}\ln \ell_t^\omega + \tfrac{\beta}{\rho}(I - (F^\omega)^2)\,\mathbb{E}_t\,\mathrm{d}v_{t+1}^\omega \end{aligned} \tag{A.2}$$

Pre-multiplying one-period forwarded version of equation (A.2) with βF^ω and taking expectation operator $\mathbb{E}_t$, we have

$$\beta F^\omega\,\mathbb{E}_t\,\mathrm{d}\ln \ell_{t+2}^\omega = \beta(F^\omega)^2\,\mathrm{d}\ln \ell_{t+1}^\omega + \tfrac{\beta}{\rho}(I - (F^\omega)^2)\beta F^\omega\,\mathbb{E}_t\,\mathrm{d}v_{t+2}^\omega \tag{A.3}$$

Subtracting equation (A.3) from equation (A.2), we have

$$\mathrm{d}\ln \ell_{t+1}^\omega = F^\omega\,\mathrm{d}\ln \ell_t^\omega + \beta F^\omega(\mathbb{E}_t\,\mathrm{d}\ln \ell_{t+2}^\omega - F^\omega\,\mathrm{d}\ln \ell_{t+1}^\omega) + \tfrac{\beta}{\rho}(I - (F^\omega)^2)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}.$$

Rearranging it, we obtain equation (16):

$$\mathrm{d}\ln \ell_t = (\mathcal{F}_1^{-1} + \beta\mathcal{F}_1)\,\mathrm{d}\ln \ell_{t+1} - \beta\,\mathbb{E}_t\,\mathrm{d}\ln \ell_{t+2} - \tfrac{\beta}{\rho}(\mathcal{F}_1^{-1} - \mathcal{F}_1)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}.$$

We refer to this equation as a *forward-looking process with order 2* because we write a period t variable as a function of period $t+1$ and $t+2$ variables and a shock.

Now, suppose that there are N number of worker types, $\omega \in \{1, 2, \ldots, N\}$. We derive a recursive representation of the form

$$\mathrm{d}\ln \ell_t = \sum_{k=1}^{K} \mathbf{\Gamma}_k \, \mathbb{E}_t \, \mathrm{d}\ln \ell_{t+k} + \tfrac{\beta}{\rho} \sum_{k=1}^{K'+1} \mathbf{\Lambda}_k \, \mathbb{E}_t \, \mathrm{d}w_{t+k}.$$

We refer to this equation as a forward-looking process with order(K, K'). Note that the change in aggregate labor supply can be written as

$$\mathrm{d}\ln \ell_{t+1} = \bar{\mathbb{E}}_\omega[\mathrm{d}\ln \ell^\omega_{t+1}] = L_1 \, \mathrm{d}\ln \ell^1_{t+1} + L_2 \, \mathrm{d}\ln \ell^2_{t+1} + \cdots + L_N \, \mathrm{d}\ln \ell^N_{t+1}$$

where each $\mathrm{d}\ln \ell^\omega_{t+1}$ follows a forward-looking process with order w, and L_ω is a diagonal matrix, whose i-th diagonal element is given by the stationary proportion of type ω in sector i. Granger and Morris (1976) show that the scalar-weighted sum of N number of autoregressive processes of order 2 follows an autoregressive process of order at most $2N$. Here, we instead have diagonal-matrix-weighted sum of N number of forward-looking process of order 2, but we can apply a modified version of their proof to show proposition. Due to an invertibility issue, we need forward-looking process of order $(4N-2, 4N-4)$ instead of order $2N$. The next subsection is devoted to the proof of Proposition A.6.

Proposition A.6. *If the number of types is N, $\mathrm{d}\ln \ell_t$ has a recursive representation of the form*

$$\mathrm{d}\ln \ell_t = \sum_{k=1}^{4N-2} \mathbf{\Gamma}_k \, \mathbb{E}_t \, \mathrm{d}\ln \ell_{t+k} + \tfrac{\beta}{\rho} \sum_{k=1}^{4N-3} \mathbf{\Lambda}_k \, \mathbb{E}_t \, \mathrm{d}w_{t+k}.$$

A.2.2.1 Proof of Proposition A.6

We prove this proposition for the case with $N = 2$. The same proof can be inductively applied to show the case with $N > 2$. For the case with two worker types, aggregate labor supply is given by

$$\mathrm{d}\ln \ell_t = L_1 \, \mathrm{d}\ln \ell^1_t + L_2 \, \mathrm{d}\ln \ell^2_t \tag{A.4}$$

where

$$\mathrm{d}\ln \ell^1_{t+1} - \beta F^1 \, \mathbb{E}_t \, \mathrm{d}\ln \ell^1_{t+2} = F^1 \, \mathrm{d}\ln \ell^1_t - \beta (F^1)^2 \, \mathrm{d}\ln \ell^1_{t+1} + \tfrac{\beta}{\rho}(I - (F^1)^2) \, \mathbb{E}_t \, \mathrm{d}w_{t+1} \tag{A.5}$$

$$\mathrm{d}\ln \ell^2_{t+1} - \beta F^2 \, \mathbb{E}_t \, \mathrm{d}\ln \ell^2_{t+2} = F^2 \, \mathrm{d}\ln \ell^2_t - \beta (F^2)^2 \, \mathrm{d}\ln \ell^2_{t+1} + \tfrac{\beta}{\rho}(I - (F^2)^2) \, \mathbb{E}_t \, \mathrm{d}w_{t+1}. \tag{A.6}$$

Using equation (A.5) to cancel out $\mathrm{d}\ln\ell_t^1$ from equation (A.4), we have

$$
\begin{aligned}
&L_1((F^1)^{-1}+\beta F^1)L_1^{-1}\,\mathrm{d}\ln\ell_{t+1}-\beta\,\mathbb{E}_t\,\mathrm{d}\ln\ell_{t+2}-\mathrm{d}\ln\ell_t\\
&=L_1((F^1)^{-1}+\beta F^1)(\mathrm{d}\ln\ell_{t+1}^1+L_1^{-1}L_2\,\mathrm{d}\ln\ell_{t+1}^2)-\beta\,\mathbb{E}_t(L_1\,\mathrm{d}\ln\ell_{t+2}^1+L_2\,\mathrm{d}\ln\ell_{t+2}^2)-(L_1\,\mathrm{d}\ln\ell_t^1+L_2\,\mathrm{d}\ln\ell_t^2)\\
&=L_1(F^1)^{-1}\big((I+\beta(F^1)^2)\,\mathrm{d}\ln\ell_{t+1}^1-\beta F^1\,\mathbb{E}_t\,\mathrm{d}\ln\ell_{t+2}^1-F^1\,\mathrm{d}\ln\ell_t^1\big)+\mathbb{E}_t\,\Xi_t\\
&=L_1(F^1)^{-1}\tfrac{\beta}{\rho}(I-(F^1)^2)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}+\mathbb{E}_t\,\Xi_t
\end{aligned}
$$

where

$$
\begin{aligned}
\Xi_t&=L_1((F^1)^{-1}+\beta F^1)L_1^{-1}L_2\,\mathrm{d}\ln\ell_{t+1}^2-\beta L_2\,\mathrm{d}\ln\ell_{t+2}^2-L_2\,\mathrm{d}\ln\ell_t^2\\
&=L_1((F^1)^{-1}+\beta F^1)L_1^{-1}L_2\,\mathrm{d}\ln\ell_{t+1}^2-\beta L_2\,\mathrm{d}\ln\ell_{t+2}^2\\
&\quad-L_2(F^2)^{-1}\Big(((I+\beta(F^2)^2)\,\mathrm{d}\ln\ell_{t+1}^2-\beta F^2\,\mathrm{d}\ln\ell_{t+2}^2-\tfrac{\beta}{\rho}(I-(F^2)^2)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}\Big)\\
&=\Big(L_1((F^1)^{-1}+\beta F^1)L_1^{-1}L_2-L_2((F^2)^{-1}+\beta F^2)\Big)\,\mathrm{d}\ln\ell_{t+1}^2+\tfrac{\beta}{\rho}L_2((F^2)^{-1}-F^2)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}.
\end{aligned}
$$

This can be rearranged to

$$
\begin{aligned}
&L_1((F^1)^{-1}+\beta F^1)L_1^{-1}\,\mathrm{d}\ln\ell_{t+1}-\beta\,\mathbb{E}_t\,\mathrm{d}\ln\ell_{t+2}-\mathrm{d}\ln\ell_t\\
&=\tfrac{\beta}{\rho}\big(L_1((F^1)^{-1}-F^1)+L_2((F^2)^{-1}-F^2)\big)\,\mathbb{E}_t\,\mathrm{d}w_{t+1}+\big(L_1((F^1)^{-1}+\beta F^1)L_1^{-1}-L_2((F^2)^{-1}+\beta F^2)L_2^{-1}\big)L_2\,\mathrm{d}\ln\ell_{t+1}^2.
\end{aligned}
$$

or equivalently

$$y_t=\Psi x_t$$

where

$$
\begin{aligned}
y_t&=\boldsymbol{X}\,\mathrm{d}\ln\ell_{t+1}-\beta\,\mathbb{E}_t\,\mathrm{d}\ln\ell_{t+2}-\mathrm{d}\ln\ell_t-\tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\mathrm{d}w_{t+1}\\
x_t&=\mathrm{d}\ln\ell_{t+1}^2\\
\boldsymbol{X}&=L_1((F^1)^{-1}+\beta F^1)L_1^{-1}\\
\boldsymbol{Y}&=L_1((F^1)^{-1}-F^1)+L_2((F^2)^{-1}-F^2)\\
\Psi&=\big(L_1((F^1)^{-1}+\beta F^1)L_1^{-1}-L_2((F^2)^{-1}+\beta F^2)L_2^{-1}\big)L_2.
\end{aligned}
$$

From equation (A.6), the law of motion of x_t is given by

$$x_t = \boldsymbol{A} x_{t+1} + \boldsymbol{B}\,\mathbb{E}_{t+1}\, x_{t+2} + \varepsilon_{t+1}$$

where $\boldsymbol{A} = (F^2)^{-1} + \beta F^2$, $\boldsymbol{B} = -\beta I$, and $\varepsilon_t = \frac{\beta}{\rho}\boldsymbol{C}\,\mathbb{E}_t\,\mathrm{d}w_{t+1}$ where $\boldsymbol{C} = -((F^2)^{-1} - F^2)$.

Lemma A.5. *Suppose $x_t = \boldsymbol{A}x_{t+1} + \boldsymbol{B}x_{t+2} + \varepsilon_{t+1} \in \mathbb{R}^S$ and $y_t = \boldsymbol{Z}x_t$ where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{S\times S}$ are invertible and $\boldsymbol{Z} \in \mathbb{R}^{S\times S}$ is of rank $S-1$. Then, we can always write*

$$y_t = \Theta_1 y_{t+1} + \Theta_2 y_{t+2} + \Theta_3 y_{t+3} + \Theta_4 y_{t+4} + \Omega_1\varepsilon_{t+1} + \Omega_2\varepsilon_{t+2} + \Omega_3\varepsilon_{t+3}.$$

for some matrices $\Theta_i = \Theta_i(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{Z})$ and $\Omega_i = \Omega_i(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{Z})$.

Proof. Write $\boldsymbol{x} = \begin{pmatrix} x_{t+3} \\ x_{t+4} \end{pmatrix}$, then we have

$$\begin{aligned}
y_{t+4} &= \begin{pmatrix} \boldsymbol{O} & \boldsymbol{Z} \end{pmatrix}\boldsymbol{x} \equiv \boldsymbol{M}_1\boldsymbol{x} \\
y_{t+3} &= \begin{pmatrix} \boldsymbol{Z} & \boldsymbol{O} \end{pmatrix}\boldsymbol{x} \equiv \boldsymbol{M}_2\boldsymbol{x} \\
y_{t+2} &= \boldsymbol{Z}(\boldsymbol{A}x_{t+3} + \boldsymbol{B}x_{t+4} + \varepsilon_{t+3}) \\
&= \begin{pmatrix} \boldsymbol{ZA} & \boldsymbol{ZB} \end{pmatrix}\boldsymbol{x} + \boldsymbol{Z}\varepsilon_{t+3} \\
&\equiv \boldsymbol{M}_3\boldsymbol{x} + \boldsymbol{Z}\varepsilon_{t+3} \\
y_{t+1} &= \boldsymbol{Z}(\boldsymbol{A}x_{t+2} + \boldsymbol{B}x_{t+3} + \varepsilon_{t+2}) \\
&= \boldsymbol{Z}(\boldsymbol{A}(\boldsymbol{A}x_{t+3} + \boldsymbol{B}x_{t+4} + \varepsilon_{t+3}) + \boldsymbol{B}x_{t+3} + \varepsilon_{t+2}) \\
&= \begin{pmatrix} \boldsymbol{Z}(\boldsymbol{A}^2 + \boldsymbol{B}) & \boldsymbol{ZAB} \end{pmatrix}\boldsymbol{x} + \boldsymbol{ZA}\varepsilon_{t+3} + \boldsymbol{Z}\varepsilon_{t+2} \\
&\equiv \boldsymbol{M}_4\boldsymbol{x} + \boldsymbol{ZA}\varepsilon_{t+3} + \boldsymbol{Z}\varepsilon_{t+2} \\
y_t &= \boldsymbol{Z}(\boldsymbol{A}x_{t+1} + \boldsymbol{B}x_{t+2} + \varepsilon_{t+1}) \\
&= \boldsymbol{Z}(\boldsymbol{A}(\boldsymbol{A}(\boldsymbol{A}x_{t+3} + \boldsymbol{B}x_{t+4} + \varepsilon_{t+3}) + \boldsymbol{B}x_{t+3} + \varepsilon_{t+2}) + \boldsymbol{B}(\boldsymbol{A}x_{t+3} + \boldsymbol{B}x_{t+4} + \varepsilon_{t+3}) + \varepsilon_{t+1}) \\
&= \begin{pmatrix} \boldsymbol{Z}(\boldsymbol{A}^3 + \boldsymbol{AB} + \boldsymbol{BA}) & \boldsymbol{Z}(\boldsymbol{A}^2\boldsymbol{B} + \boldsymbol{B}^2) \end{pmatrix}\boldsymbol{x} + \boldsymbol{Z}(\boldsymbol{A}^2 + \boldsymbol{B})\varepsilon_{t+3} + \boldsymbol{ZA}\varepsilon_{t+2} + \boldsymbol{Z}\varepsilon_{t+1} \\
&\equiv \boldsymbol{M}_5\boldsymbol{x} + \boldsymbol{Z}(\boldsymbol{A}^2 + \boldsymbol{B})\varepsilon_{t+3} + \boldsymbol{ZA}\varepsilon_{t+2} + \boldsymbol{Z}\varepsilon_{t+1}.
\end{aligned}$$

Note that $\boldsymbol{N}_1 \equiv \boldsymbol{M}_1$ is of rank $S-1$, $\boldsymbol{N}_2 \equiv \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix}$ is of rank $2(S-1)$. If all rows of $\boldsymbol{M}_3$ can be written as a linear combination of rows of $\boldsymbol{N}_2$, then we can write $y_t = \Theta_1 y_{t+1} + \Theta_2 y_{t+2}$. If not, $\boldsymbol{N}_3 \equiv \begin{pmatrix} \boldsymbol{N}_2 \\ \boldsymbol{M}_3 \end{pmatrix}$ is of rank at least

$2(S-1)+1$.[10] By the same logic, we either prove the result, or $\boldsymbol{N}_4 \equiv \begin{pmatrix} \boldsymbol{N}_3 \\ \boldsymbol{M}_4 \end{pmatrix}$ is of rank (at least) $2S$. Thus, all rows of $\boldsymbol{M}_5$ can be written as a linear combination of rows of $\boldsymbol{N}_4$. This implies that we can find matrices $\Theta_1, \cdots, \Theta_4$ such that

$$\boldsymbol{M}_5 = \Theta_1 \boldsymbol{M}_4 + \Theta_2 \boldsymbol{M}_3 + \Theta_3 \boldsymbol{M}_2 + \Theta_4 \boldsymbol{M}_1.$$

Thus, we only need

$$\boldsymbol{Z}(\boldsymbol{A}^2 + \boldsymbol{B})\varepsilon_{t+3} + \boldsymbol{Z}\boldsymbol{A}\varepsilon_{t+2} + \boldsymbol{Z}\varepsilon_{t+1} = \Theta_1(\boldsymbol{Z}\boldsymbol{A}\varepsilon_{t+3} + \boldsymbol{Z}\varepsilon_{t+2}) + \Theta_2(\boldsymbol{Z}\varepsilon_{t+3}) + \Omega_1\varepsilon_{t+1} + \Omega_2\varepsilon_{t+2} + \Omega_3\varepsilon_{t+3}$$

or

$$\Omega_1 = \boldsymbol{Z}, \;\; \Omega_2 = \boldsymbol{Z}\boldsymbol{A} - \Theta_1\boldsymbol{Z}, \;\; \text{and} \;\; \Omega_3 = \boldsymbol{Z}(\boldsymbol{A}^2 + \boldsymbol{B}) - \Theta_1\boldsymbol{Z}\boldsymbol{A} - \Theta_2\boldsymbol{Z}. \qquad \square$$

Lemma A.6. Ψ *is non-invertible.*

Proof. Denote population share of type ω as θ^ω, transition matrix as F^ω, and stationary distribution over sectors as π^ω. Define $\pi \equiv \sum_\omega \theta^\omega \pi^\omega$, then we have

$$\pi^\intercal L_\omega = \theta^\omega (\pi^\omega)^\intercal \;\; \text{and} \;\; (\pi^\omega)^\intercal F^\omega = (\pi^\omega)^\intercal$$

Thus,

$$\pi^\intercal \Psi = \pi^\intercal \Big(L_1((F^1)^{-1} + \beta F^1)L_1^{-1} - L_2((F^2)^{-1} + \beta F^2)L_2^{-1} \Big) L_2 = (\pi^\intercal + \beta\pi^\intercal - \pi^\intercal - \beta\pi^\intercal)L_2 = 0. \qquad \square$$

By combining the previous two lemmas, we can write

$$y_t = \Theta_1 y_{t+1} + \Theta_2 \, \mathbb{E}_{t+1} \, y_{t+2} + \Theta_3 \, \mathbb{E}_{t+1} \, y_{t+3} + \Theta_4 \, \mathbb{E}_{t+1} \, y_{t+4} + \Omega_1 \varepsilon_{t+1} + \Omega_2 \, \mathbb{E}_{t+1} \, \varepsilon_{t+2} + \Omega_3 \, \mathbb{E}_{t+1} \, \varepsilon_{t+3}.$$

[10] With numerical simulation, we can see that generically $\boldsymbol{N}_3$ is of rank $2(S-1)+1$.

where $\Theta_i = \Theta_i(\boldsymbol{A}, \boldsymbol{B}, \Psi)$ and $\Omega_i = \Omega_i(\boldsymbol{A}, \boldsymbol{B}, \Psi)$. Plugging in the definitions of y_t and ε_t, we have

$$\begin{aligned}
-\operatorname{d}\ln \ell_t + \boldsymbol{X}\operatorname{d}\ln \ell_{t+1} - \beta\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+2} - \tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+1} &= \Theta_1(-\operatorname{d}\ln \ell_{t+1} + \boldsymbol{X}\mathbb{E}_t \operatorname{d}\ln \ell_{t+2} - \beta\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+3} - \tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+2}) \\
&\quad + \Theta_2(-\mathbb{E}_t \operatorname{d}\ln \ell_{t+2} + \boldsymbol{X}\mathbb{E}_t \operatorname{d}\ln \ell_{t+3} - \beta\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+4} - \tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+3}) \\
&\quad + \Theta_3(-\mathbb{E}_t \operatorname{d}\ln \ell_{t+3} + \boldsymbol{X}\mathbb{E}_t \operatorname{d}\ln \ell_{t+4} - \beta\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+5} - \tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+4}) \\
&\quad + \Theta_4(-\mathbb{E}_t \operatorname{d}\ln \ell_{t+4} + \boldsymbol{X}\mathbb{E}_t \operatorname{d}\ln \ell_{t+5} - \beta\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+6} - \tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+5}) \\
&\quad + \tfrac{\beta}{\rho}\Omega_1\boldsymbol{C}\mathbb{E}_t\,\operatorname{d}w_{t+2} + \tfrac{\beta}{\rho}\Omega_2\boldsymbol{C}\mathbb{E}_t\,\operatorname{d}w_{t+3} + \tfrac{\beta}{\rho}\Omega_3\boldsymbol{C}\mathbb{E}_t\,\operatorname{d}w_{t+4}
\end{aligned}$$

or equivalently

$$\begin{aligned}
\operatorname{d}\ln \ell_t &= (\Theta_1 + \boldsymbol{X})\operatorname{d}\ln \ell_{t+1} + (-\Theta_1\boldsymbol{X} + \Theta_2 - \beta I)\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+2} + (\beta\Theta_1 - \Theta_2\boldsymbol{X} + \Theta_3)\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+3} \\
&\quad + (\beta\Theta_2 - \Theta_3\boldsymbol{X} + \Theta_4)\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+4} + (\beta\Theta_3 - \Theta_4\boldsymbol{X})\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+5} + (\beta\Theta_4)\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+6} \\
&\quad - \tfrac{\beta}{\rho}\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+1} + (\tfrac{\beta}{\rho}\Theta_1\boldsymbol{Y} - \tfrac{\beta}{\rho}\Omega_1\boldsymbol{C})\,\mathbb{E}_t\,\operatorname{d}w_{t+2} + (\tfrac{\beta}{\rho}\Theta_2\boldsymbol{Y} - \tfrac{\beta}{\rho}\Omega_2\boldsymbol{C})\,\mathbb{E}_t\,\operatorname{d}w_{t+3} + (\tfrac{\beta}{\rho}\Theta_3\boldsymbol{Y} - \tfrac{\beta}{\rho}\Omega_3\boldsymbol{C})\,\mathbb{E}_t\,\operatorname{d}w_{t+4} + \tfrac{\beta}{\rho}\Theta_4\boldsymbol{Y}\mathbb{E}_t\,\operatorname{d}w_{t+5}.
\end{aligned}$$

This can be rearranged to obatin Proposition A.6.[11]

$$\begin{aligned}
\operatorname{d}\ln \ell_t &= \boldsymbol{\Gamma}_1 \operatorname{d}\ln \ell_{t+1} + \boldsymbol{\Gamma}_2\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+2} + \boldsymbol{\Gamma}_3\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+3} + \boldsymbol{\Gamma}_4\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+4} + \boldsymbol{\Gamma}_5\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+5} + \boldsymbol{\Gamma}_6\,\mathbb{E}_t \operatorname{d}\ln \ell_{t+6} \\
&\quad + \tfrac{\beta}{\rho}(\boldsymbol{\Lambda}_1\,\mathbb{E}_t\,\operatorname{d}w_{t+1} + \boldsymbol{\Lambda}_2\,\mathbb{E}_t\,\operatorname{d}w_{t+2} + \boldsymbol{\Lambda}_3\,\mathbb{E}_t\,\operatorname{d}w_{t+3} + \boldsymbol{\Lambda}_4\,\mathbb{E}_t\,\operatorname{d}w_{t+4} + \boldsymbol{\Lambda}_5\,\mathbb{E}_t\,\operatorname{d}w_{t+5}).
\end{aligned}$$

□

A.2.3 Alternative Methods of Extrapolation

Pure Extrapolation. Suppose we have panel data of length $K < \infty$. From the data, we can observe the conditional probability

$$\Pr\left(s_{t+k} = s_k | s_t = s_0, s_{t+1} = s_1, \cdots, s_{t+k-1} = s_{k-1}\right),$$

[11] If the number of types is 3, then we need $(\operatorname{d}\ln \ell_{t+1}, \cdots, \operatorname{d}\ln \ell_{t+10})$, and with the number of types W, we need $(\operatorname{d}\ln \ell_{t+1}, \cdots, \operatorname{d}\ln \ell_{t+4W-2})$.

for all $s_0, s_1, \ldots, s_k \in \mathcal{S}$ and k less than K. To extrapolate the probabilities with k greater than or equal to K, we truncate the history and assume the following:

$$\begin{aligned}&\Pr\left(s_{t+k} = s_k \middle| s_t = s_0, s_{t+1} = s_1, \cdots, s_{t+k-1} = s_{k-1}\right)\\ &= \Pr\left(s_{t+k} = s_k | s_{t+k-K+1} = s_{k-K+1}, \cdots, s_{t+k-1} = s_{k-1}\right).\end{aligned}$$

In short, we assume a $(K-1)$-th order Markov process and calculate the probabilities accordingly. In this sense, this extrapolation is a strict generalization of the canonical model's extrapolation, which is based on the assumption that the sector choice follow a first-order Markov process.

Extrapolation Using Retention Model. The retention model developed in Henry (1971) is based on the idea that Markov chains can be viewed as the conjunction of two processes: One determines whether workers change sectors or not, and the other governs which sector they choose, conditional on sector switching. Any transition matrix F can be decomposed as follows:

$$F = F^{\text{diag}} + (I - F^{\text{diag}})F^{\text{off-diag}}$$

where

$$\begin{aligned}(F^{\text{diag}})_{i,j} &= F_{i,j} \cdot \mathbb{1}_{i=j}\\ (F^{\text{off-diag}})_{i,j} &= \Pr(s_{t+1} = j | s_t = i, s_{t+1} \neq i) = \frac{F_{i,j}}{1 - F_{i,i}}.\end{aligned}$$

The idea of extrapolation is to treat the two parts of the process separately. For example, we can extrapolate each element of the first part of the worker flow matrices to compute the full series of $\{(\mathcal{F}_k)^{\text{diag}}\}$. Then, we can assume that the second part remains constant for all k: $(\mathcal{F}_k)^{\text{off-diag}} = (\mathcal{F}_1)^{\text{off-diag}}$ for all k.

A.2.4 Interpretation of the Estimation Results

Denote the log wages and switching costs estimated under the normalization $\rho = 1$ by $\ln w^0$ and C^0. Then, the true log wages and switching costs are given by

$$\ln w = \rho \cdot \ln w^0 \text{ and } C = \rho \cdot C^0.$$

Thus, when $C^0 = 1$, the consumption equivalent variation of paying switching cost is implicitly given by

$$\frac{\ln w(1 - \texttt{CEV})}{1 - \beta} = \frac{\ln w}{1 - \beta} - \rho,$$

which gives

$$\texttt{CEV} = 1 - \exp(-\rho(1 - \beta)) = 3.25\%.$$

Likewise, our estimate of Fréchet parameter is 0.825, which means that the consumption equivalent variation corresponding to one standard deviation lower realization of the idiosyncratic shock is given by

$$\frac{\ln w(1 - \texttt{CEV})}{1 - \beta} = \frac{\ln w}{1 - \beta} - 2 \cdot \frac{\pi\rho}{\sqrt{6}}$$

where we multiply two because it is the difference between two realizations of idiosyncratic shocks. This gives

$$\texttt{CEV} = 1 - \exp(-\frac{2\pi\rho(1 - \beta)}{\sqrt{6}}) = 8.12\%.$$

In contrast, ACM assume linear utility function, so the consumption equivalent variation can be computed as

$$\frac{w(1 - \texttt{CEV}_{\text{ACM}})}{1 - \beta} = \frac{w}{1 - \beta} - C_{\text{ACM}}$$

Thus, we have (given their normalization of average wages to one)

$$\texttt{CEV}_{\text{ACM}} = \frac{(1 - \beta)C_{\text{ACM}}}{w} = 19.7\%$$

where we use the number in Panel IV of Table 3, which is used in their counterfactual exercises. Likewise, their estimate of Fréchet parameter is 1.884, which means that the consumption equivalent variation corresponding to one standard deviation lower realization of the idiosyncratic shock is given by

$$\frac{w(1 - \texttt{CEV}_{\text{ACM}})}{1 - \beta} = \frac{w}{1 - \beta} - 2 \cdot \frac{\pi\rho_{\text{ACM}}}{\sqrt{6}}$$

which gives

$$\texttt{CEV}_{\text{ACM}} = \frac{2\pi\rho_{\text{ACM}}(1 - \beta)}{w\sqrt{6}} = 14.5\%.$$

A.3 Empirical Applications Appendix

A.3.1 Application 1: Hypothetical Trade Liberalization

Closing the Model. We consider the first version of the model in ACM, in which all sectors produce tradable output, and world prices are exogenously determined. Each sector s has a constant elasticity of substitution (CES) production function

$$y_t^s = \psi^s(\alpha^s(L_t^s)^{\rho^s} + (1-\alpha^s)(K_t^s)^{\rho^s})^{\frac{1}{\rho^s}}$$

with fixed sector-specific capital $K_t^s = 1$ normalized to one. The parameters should satisfy $\alpha^s \in (0,1), \rho^s < 1$, and $\psi^s > 0$. Then, the sectoral wage is given by the marginal productivity of labor

$$w_t^s = p_t^s \alpha^s \psi^s (L_t^s)^{\rho^s - 1}(\alpha^s (L_t^s)^{\rho^s} + (1-\alpha^s))^{\frac{1-\rho^s}{\rho^s}}$$

where p_t^s is the domestic price of the sector-s good. Without loss of generality, we normalize the domestic prices to one in the initial steady state prior to the shock. Finally, workers have identical Cobb-Douglas with shares μ^s for sector s.

We follow the calibration strategy of ACM (except for the number of sectors). We set the values of α^s, ρ^s, and ψ^s to minimize the Euclidean distance between the model-implied values of sectoral wages, sectoral labor shares, and sector share of GDP; and the values computed from the data. The values of μ^s are calibrated from consumption shares from the Bureau of Labor Statistics (BLS).

A.3.2 Application 2: The China Shock

A.3.2.1 CDP's Model Extended with Worker Heterogeneity

In this section and the next, we closely follow the modeling decisions and notation of CDP. See CDP for more details and equilibrium characterization. CDP consider a world with N locations (indexed by n or i) and J sectors (indexed by j or k), where sector $j = 0$ represents non-employment. Time is discrete and indexed by $t \in \mathbb{N}_0$. In each location-sector combination, (n, j), there is a competitive local labor market.

Heterogeneous Workers. In each location n, there is a continuum of forward-looking workers who optimally decide which sector to supply their labor for each period. The heterogeneity of workers are indexed by ω. Similar to equation

(1), the value of a worker of type ω who is employed in sector j at period t is given by

$$V_{jt}^{n,\omega} = U(c_{jt}^{n,\omega}) + \max_k \{\beta \, \mathbb{E}_t \, V_{kt+1}^{n,\omega} - C_{jk}^{n,\omega} + \rho \cdot \varepsilon_{kt}\}$$

where $c_{jt}^{n,\omega} = \prod_k (c_{jt}^{k,n,\omega})^{\alpha^k}$ is a Cobb-Douglas aggregator across local sectoral goods, with the corresponding price index $P_t^n = \prod_j (P_t^{nk}/\alpha^k)^{\alpha^k}$. Following CDP, we assume $U(c) = \log c$.

Households employed in a local labor market (n, j) earn a nominal wage of w_t^{nj}, consuming $c_{jt}^{n,\omega} = \tilde{c}^{nj,\omega} \cdot w_t^{nj}/P_t^n$ units of consumption aggregate where $\tilde{c}^{nj,\omega}$ is a type-specific shifter representing non-pecuniary sectoral preferences. Non-employed household (who chooses sector $j = 0$) consume $c_{0t}^{n,\omega} = b^{n,\omega}$ units of consumption aggregate in terms of home production. The ex-ante value and sector choice probabilities are characterized as in equations (2) and (3):

$$v_{jt}^{n,\omega} = U(c_{jt}^{n,\omega}) + \rho \ln \sum_k (\exp(\beta \, \mathbb{E}_t \, v_{kt+1}^{n,\omega}) / \exp(C_{jk}^{n,\omega}))^{1/\rho} \tag{A.7}$$

$$F_{jkt}^{n,\omega} = \frac{(\exp(\beta \, \mathbb{E}_t \, v_{kt+1}^{n,\omega}) / \exp(C_{jk}^{n,\omega}))^{1/\rho}}{\sum_{k'} (\exp(\beta \, \mathbb{E}_t \, v_{k't+1}^{n,\omega}) / \exp(C_{jk'}^{n,\omega}))^{1/\rho}}. \tag{A.8}$$

Thus, the law of motion of sectoral labor supply of type ω workers in region n is

$$\ell_{kt+1}^{n,\omega} = \sum_j F_{jkt}^{n,\omega} \ell_{jt}^{n,\omega}. \tag{A.9}$$

Finally, let $L_t^{nj} = \sum_\omega \ell_{jt+1}^{n,\omega}$ be the total labor supply to local labor market (n, j). In this environment, Assumption 1 holds, and we implicitly maintain Assumption 2 (with n superscript) as well.

Production. For each sector j, there is a continuum of different varieties of intermediate goods. Each region-sector combination draws a variety-specific productivity z^{nj}, which follows a Fréchet distribution with the dispersion parameter θ^j. Without loss, each variety is indexed by $z^j = (z^{1j}, z^{2j}, \ldots, z^{Nj})$. In each local labor market, (n, j), there is a continuum of perfectly competitive firms producing variety z^j. They have a Cobb-Douglas technology combining labor (l), structures (h), and local sectoral goods from all sectors (M):

$$q_t^{nj} = z^{nj} (A_t^{nj} (h_t^{nj})^{\xi^n} (l_t^{nj})^{1-\xi^n})^{\gamma^{nj}} \prod_k (M_t^{nj,nk})^{\gamma^{nj,nk}}$$

where A_t^{nj} is a sector-region specific productivity. Thus, the unit cost of producing this intermediate good is

$$\frac{x_t^{nj}}{z^{nj} (A_t^{nj})^{\gamma^{nj}}} \quad \text{where} \quad x_t^{nj} = B^{nj} ((r_t^{nj})^{\xi^n} (w_t^{nj})^{1-\xi^n})^{\gamma^{nj}} \prod_k (P_t^{nk})^{\gamma^{nj,nk}} \tag{A.10}$$

where B^{nj} is a constant, r_t^{nj} is the rental price of structures, and P_t^{nk} is the price of the local sector-k goods.

Local sectoral goods are produced from intermediate goods in a competitive way, which are then used as final consumption and as materials for the production of intermediate varieties. The technology is given by:

$$Q_t^{nj} = \left(\int (\tilde{q}_t^{nj}(z^j))^{1-1/\eta^{nj}} \, \mathrm{d}\phi^j(z^j) \right)^{\eta^{nj}/(\eta^{nj}-1)}$$

where $\tilde{q}_t^{nj}(z^j)$ is the quantity of variety z^j used in the production, and $\phi^j(\cdot)$ is the joint distribution of the vector z^j. The intermediate good of variety z^j is sourced from a country with the minimum price, taking into account bilateral iceberg-type trade costs (κ). The minimized price is then given by

$$p_t^{nj}(z^j) = \min_i \left\{ \frac{\kappa_t^{nj,ij} x_t^{ij}}{z^{ij} (A_t^{ij})^{\gamma^{ij}}} \right\}.$$

Thus, the price of the local sectoral good is

$$P_t^{nj} = \Gamma\left(\frac{1+\theta^j-\eta^{nj}}{\theta^j} \right) \cdot \left(\sum_i (x_t^{ij} \kappa_t^{nj,ij})^{-\theta^j} (A_t^{ij})^{\theta^j \gamma^{ij}} \right)^{-1/\theta^j} \tag{A.11}$$

Finally, the share of total expenditure in local market (n,j) on goods from market (i,j) is given by

$$\pi_t^{nj,ij} = \frac{(x_t^{ij} \kappa_t^{nj,ij})^{-\theta^j} (A_t^{ij})^{\theta^j \gamma^{ij}}}{\sum_{i'} (x_t^{i'j} \kappa_t^{nj,i'j})^{-\theta^j} (A_t^{i'j})^{\theta^j \gamma^{i'j}}} \tag{A.12}$$

Structure Rentier. There is a continuum of structure rentiers in each region n. The own the local structures of fixed amount $\{H^{nj}\}_j$ and rent them to local firms. The received rents are aggregated at the global-level, and rentiers in each region n receive a constant share ι^n of the total global revenue:

$$\iota^n \chi_t \quad \text{where } \chi_t = \sum_i \sum_k r_t^{ik} H^{ik}.$$

Market Clearing. Market clearing for goods market, labor market, and structure market is given by

$$X_t^{nj} = \sum_k \gamma^{nk,nj} \sum_i \pi_t^{ik,nk} X_t^{ik} + \alpha^j \left(\sum_k w_t^{nk} L_t^{nk} + \iota^n \chi_t \right) \tag{A.13}$$

$$w_t^{nj} L_t^{nj} = \gamma^{nj} (1-\xi^n) \sum_i \pi_t^{ij,nj} X_t^{ij} \tag{A.14}$$

$$r_t^{nj} H^{nj} = \gamma^{nj} \xi^n \sum_i \pi_t^{ij,nj} X_t^{ij} \tag{A.15}$$

where X_t^{nj} is the total expenditure on sector j good in region n.

Equilibrium. Following CDP, we group exogenous state variables of the economy into time-varying ones and time-invariant ones:

$$\Theta_t \equiv (\{A_t^{nj}\}_{n,j}, \{\kappa_t^{nj,ij}\}_{n,i,j}) \text{ and } \bar{\Theta} \equiv (\{C_{jk}^{n,\omega}\}_{j,k,n,\omega}, \{H^{nj}\}_{n,j}, \{\tilde{c}^{nj,\omega}\}_{n,j,\omega}, \{b^{n,\omega}\}_{n,\omega}).$$

Given the initial distribution of labor and the path of exogenous state variables $(\{\ell_{j0}^{n,\omega}\}_{j,n,\omega}, \{\Theta_t\}_{t=0}^{\infty}, \bar{\Theta})$, a *sequential competitive equilibrium* corresponds to a sequence of $\{L_t, F_t, v_t, x_t, P_t, \pi_t, X_t, w_t, r_t\}_{t=0}^{\infty}$, where $L_t = \{\ell_{jt}^{n,\omega}\}_{j,n,\omega}$, $F_t = \{F_{jkt}^{n,\omega}\}_{j,k,n,\omega}$, $v_t = \{v_{jt}^{n,\omega}\}_{j,n,\omega}$, $x_t = \{x_t^{nj}\}_{n,j}$, $P_t = \{P_t^{n,j}\}_{n,j}$, $\pi_t = \{\pi_t^{ij,nj}\}_{i,j,n}$, $X_t = \{X_t^{nj}\}_{n,j}$, $w_t = \{w_t^{nj}\}_{n,j}$, and $r_t = \{r_t^{nj}\}_{n,j}$, such that households optimally make sector choice decisions, as described in (A.7)–(A.9); firms maximize their profits, as described in (A.10)–(A.12); all markets clear, as described in (A.13)–(A.15). A *stationary equilibrium* is a sequential competitive equilibrium such that $\{L_t, F_t, v_t, x_t, P_t, \pi_t, X_t, w_t, r_t\}$ is time-invariant.

A.3.2.2 Dynamic Hat Algebra with Worker Heterogeneity

Following CDP, we solve for the equilibrium in time differences. We denote by $\dot{y}_{t+1} \equiv (y_{t+1}^1/y_t^1, y_{t+1}^2/y_t^2, \cdots)$ the proportional change in any scalar or vector. The following proposition corresponds to Propositions 1 and 2 of CDP, but allowing for worker heterogeneity.

Proposition A.7 (Solving the model). *Suppose that the economy is initially starting from a stationary equilibrium at period $t = 0$. Up to the first-order approximation around a stationary equilibrium, given a sequence of changes in exogenous state variables, $\{\dot{\Theta}_t\}_{t=1}^{\infty}$ satisfying $\lim_{t\to\infty} \dot{\Theta}_t = 1$, known to agents in period $t = 1$, the solution to the sequential equilibrium in time differences does not require information on the level of the exogenous state variables*

$\{\Theta_t\}_{t=0}^{\infty}$ *or* $\bar{\Theta}$*, and solves the following system of equations:*

$$
\begin{aligned}
v_t^{nj} &= v_0^{nj} + \sum_k \beta^k \mathcal{F}_k \ln\left(\frac{\dot{w}_{t+k}^n}{\dot{P}_{t+k}^n} \cdot \frac{\dot{w}_{t+k-1}^n}{\dot{P}_{t+k-1}^n} \cdot \cdots \cdot \frac{\dot{w}_1^n}{\dot{P}_1^n}\right) \\
L_t^{nj} &= L_0^{nj} \cdot \exp\left(\left(\sum_{s=0}^{t-2}\sum_{k=0}^{\infty} \frac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k}^n - \mathcal{F}_{s+k+2}^n) \ln\left(\frac{\dot{w}_{t-s+k}^n}{\dot{P}_{t-s+k}^n} \cdot \frac{\dot{w}_{t-s+k-1}^n}{\dot{P}_{t-s+k-1}^n} \cdot \cdots \cdot \frac{\dot{w}_1^n}{\dot{P}_1^n}\right)\right)_j\right) \\
\dot{x}_{t+1}^{nj} &= (\dot{L}_{t+1}^{nj})^{\gamma^{nj}\xi^n} (\dot{w}_{t+1}^{nj})^{\gamma^{nj}} \prod_k (\dot{P}_{t+1}^{nk})^{\gamma^{nj,nk}} \\
\dot{P}_{t+1}^{nj} &= \left(\sum_i \pi_t^{nj,ij} (\dot{x}_{t+1}^{ij} \dot{\kappa}_{t+1}^{nj,ij})^{-\theta^j} (\dot{A}_{t+1}^{ij})^{\theta^j \gamma^{ij}}\right)^{-1/\theta^j} \\
\pi_{t+1}^{nj,ij} &= \pi_t^{nj,ij} \left(\frac{\dot{x}_{t+1}^{ij} \dot{\kappa}_{t+1}^{nj,ij}}{\dot{P}_{t+1}^{nj}}\right)^{-\theta^j} (\dot{A}_{t+1}^{ij})^{\theta^j \gamma^{ij}} \\
X_{t+1}^{nj} &= \sum_k \gamma^{nk,nj} \sum_i \pi_{t+1}^{ik,nk} X_{t+1}^{ik} + \alpha^j \left(\sum_k \dot{w}_{t+1}^{nk} \dot{L}_{t+1}^{nk} w_t^{nk} L_t^{nk} + \iota^n \chi_{t+1}\right) \\
\dot{w}_{t+1}^{nj} \dot{L}_{t+1}^{nj} w_t^{nj} L_t^{nj} &= \gamma^{nj}(1-\xi^n) \sum_i \pi_{t+1}^{ij,nj} X_{t+1}^{ij}
\end{aligned}
$$

where $\chi_{t+1} = \sum_i \sum_k \frac{\xi^i}{1-\xi^i} \dot{w}_{t+1}^{ik} \dot{L}_{t+1}^{ik} w_t^{ik} L_t^{ik}$ *and* $\dot{w}_t^n$ *is a vector whose* j*th element is* $\dot{w}_t^{nj}$*.*

Proof. The last five equations write the equilibrium conditions for the static multicountry interregional trade model in time differences. See CDP for a proof of this representation. Note that the real wage in period t can be written as

$$
\frac{w_t^{nj}}{P_t^n} = \frac{\dot{w}_t^{nj}}{\dot{P}_t^n} \cdot \frac{\dot{w}_{t-1}^{nj}}{\dot{P}_{t-1}^n} \cdot \cdots \cdot \frac{\dot{w}_1^{nj}}{\dot{P}_1^n} \frac{w_0^{nj}}{P_0^n}.
$$

Since the economy is initially starting from a stationary equilibrium at period $t = 0$, we have

$$
\mathrm{d}\ln\left(\frac{w_t^{nj}}{P_t^n}\right) = \ln\left(\frac{\dot{w}_t^{nj}}{\dot{P}_t^n} \cdot \frac{\dot{w}_{t-1}^{nj}}{\dot{P}_{t-1}^n} \cdot \cdots \cdot \frac{\dot{w}_1^{nj}}{\dot{P}_1^n}\right). \tag{A.16}
$$

Note that with shocks known to agent at period $t = 1$, the sufficient statistics results of Proposition 1 can be simplified to

$$
\begin{aligned}
\mathrm{d}v_t &= \sum_{k=0}^{\infty} \beta^k \mathcal{F}_k \,\mathrm{d}w_{t+k} \\
\mathrm{d}\ln \ell_t &= \sum_{s=0}^{t-2}\sum_{k=0}^{\infty} \frac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathrm{d}w_{t-s+k}.
\end{aligned}
$$

Plugging expression (A.16) into these equations gives the first two equations. □

In the baseline economy, the path of exogenous state variables are given by $\{\Theta_t\}_{t=0}^{\infty}$ and $\bar{\Theta}$. In the counterfactual economy, we consider changes in exogenous state variables. We denote the new path by $\{\Theta'_t\}_{t=0}^{\infty}$. We assume that agents learn about these changes at period $t = 1$. This proposition corresponds to Proposition 3 of CDP, but allowing for worker heterogeneity. It shows how to solve for the counterfactual changes in endogenous variables in time differences and relative to a baseline economy without the need to estimate the level of the exogenous state variables. We denote by $\hat{y}_{t+1} \equiv \dot{y}'_{t+1}/\dot{y}_{t+1}$ the ratio of time differences between the counterfactual equilibrium and the baseline equilibrium.

Proposition A.8 (Solving for Counterfactuals). *Suppose that the economy is initially starting from a stationary equilibrium at period $t = 0$. Up to the first-order approximation around a stationary equilibrium, given a baseline equilibrium, $\{L_t, \pi_t, X_t\}_{t=0}^{\infty}$, and a counterfactual sequence of changes in exogenous state variables, $\{\hat{\Theta}_t\}_{t=1}^{\infty}$ satisfying $\lim_{t\to\infty} \hat{\Theta}_t = 1$, known to agents in period $t = 1$, the solution to the counterfactual sequential equilibrium in time differences does not require information on the level of the exogenous state variables $\{\Theta_t\}_{t=0}^{\infty}$ or $\bar{\Theta}$, and solves the following system of equations:*

$$
\begin{aligned}
v_t'^{nj} &= v_t^{nj} + \sum_k \beta^k \mathcal{F}_k \ln\left(\frac{\hat{w}^n_{t+k}}{\hat{P}^n_{t+k}} \cdot \frac{\hat{w}^n_{t+k-1}}{\hat{P}^n_{t+k-1}} \cdot \cdots \cdot \frac{\hat{w}^n_1}{\hat{P}^n_1}\right) \\
L_t'^{nj} &= L_t^{nj} \cdot \exp\left(\left(\sum_{s=0}^{t-2}\sum_{k=0}^{\infty} \frac{\beta^{k+1}}{\rho}(\mathcal{F}^n_{s+k} - \mathcal{F}^n_{s+k+2}) \ln\left(\frac{\hat{w}^n_{t-s+k}}{\hat{P}^n_{t-s+k}} \cdot \frac{\hat{w}^n_{t-s+k-1}}{\hat{P}^n_{t-s+k-1}} \cdot \cdots \cdot \frac{\hat{w}^n_1}{\hat{P}^n_1}\right)\right)_j\right) \\
\hat{x}^{nj}_{t+1} &= (\hat{L}^{nj}_{t+1})^{\gamma^{nj}\xi^n} (\hat{w}^{nj}_{t+1})^{\gamma^{nj}} \prod_k (\hat{P}^{nk}_{t+1})^{\gamma^{nj,nk}} \\
\hat{P}^{nj}_{t+1} &= \left(\sum_i \pi_t'^{nj,ij} \dot{\pi}^{nj,ij}_{t+1} (\hat{x}^{ij}_{t+1} \hat{\kappa}^{nj,ij}_{t+1})^{-\theta^j} (\hat{A}^{ij}_{t+1})^{\theta^j \gamma^{ij}}\right)^{-1/\theta^j} \\
\pi'^{nj,ij}_{t+1} &= \pi_t'^{nj,ij} \dot{\pi}'^{nj,ij}_{t+1} \left(\frac{\hat{x}^{ij}_{t+1}\hat{\kappa}^{nj,ij}_{t+1}}{\hat{P}^{nj}_{t+1}}\right)^{-\theta^j} (\hat{A}^{ij}_{t+1})^{\theta^j\gamma^{ij}} \\
X'^{nj}_{t+1} &= \sum_k \gamma^{nk,nj} \sum_i \pi'^{ik,nk}_{t+1} X'^{ik}_{t+1} + \alpha^j \left(\sum_k \hat{w}^{nk}_{t+1} \hat{L}^{nk}_{t+1} w_t'^{nk} L_t'^{nk} \dot{w}^{nk}_{t+1} \dot{L}^{nk}_{t+1} + \iota^n \chi'_{t+1}\right) \\
\hat{w}^{nj}_{t+1} \hat{L}^{nj}_{t+1} &= \frac{\gamma^{nj}(1-\xi^n)}{w_t'^{nk} L_t'^{nk} \dot{w}^{nk}_{t+1} \dot{L}^{nk}_{t+1}} \sum_i \pi'^{ij,nj}_{t+1} X'^{ij}_{t+1}
\end{aligned}
$$

where $\chi'_{t+1} = \sum_i \sum_k \frac{\xi^i}{1-\xi^i} \hat{w}^{ik}_{t+1} \hat{L}^{ik}_{t+1} w_t'^{ik} L_t'^{ik} \dot{w}^{ik}_{t+1} \dot{L}^{ik}_{t+1}$ and $\hat{w}^n_t$ is a vector whose jth element is $\hat{w}^{nj}_t$.

A.3.2.3 Calibration of the Model

There are 87 regions, 50 US states and 37 other countries, and 4 sectors. The model has the following parameters: value added shares $(\{\gamma^{nj}\}_{n,j})$, the share of structures in value added $(\{\xi^n\}_n)$, the input-output coefficients $(\{\gamma^{nk,nj}\}_{n,k,j})$, rentier shares $(\{\iota^n\}_n)$, consumption Cobb-Douglas shares $(\{\alpha^j\}_j)$, the discount factor (β), the sectoral trade

elasticities ($\{\theta^j\}_j$), and the inverse sector-choice elasticity (ρ).[12] The year 2000 corresponds to the period $t = 0$ of the model. To apply dynamic hat algebra, we use data on bilateral trade flows π_t and value added $\{w_t^{nj} L_t^{nj} + r_t^{nj} H^{nj}\}_{n,j}$ from year 2000 to 2007. The data comes from the World Input-Output Database (WIOD), the 2002 Commodity Flow Survey (CFS), and regional employment data from the Bureau of Economic Analysis (BEA). See CDP for more details. Finally, we need to identify the magnitude of the China shock.

Parameters. Following CDP, value added shares ($\{\gamma^{nj}\}_{n,j}$), the share of structures in value added ($\{\xi^n\}_n$), and the input-output coefficients ($\{\gamma^{nk,nj}\}_{n,k,j}$) are constructed from the BEA and the WIOD data. Rentier shares ($\{\iota^n\}_n$), consumption Cobb-Douglas shares ($\{\alpha^j\}_j$) are calculated from the constructed trade and production data. We set the quarterly discount factor to $\beta = 0.99$. We use the sectoral trade elasticities from Dix-Carneiro et al. (2023), $\theta^j = 4$. Finally, we obtain the inverse migration elasticity at a quarterly frequency from the estimate in Section 1.4.4. In particular, the value of rho at a quarterly frequency is calibrated such that both the yearly and quarterly analysis deliver the same elasticity of labor with respect to the wage changes. Up to a first-order approximation, the response of labor to a permanent change in wage $w_t = w$ known to households at $t = 1$ is given by

$$\mathrm{d} \ln \ell_t^{\text{quarterly}} = \sum_{s=0}^{t-2} \sum_{k=0}^{\infty} \frac{(\beta^{\text{quarterly}})^{k+1}}{\rho^{\text{quarterly}}} (\mathcal{F}_{s+k}^{\text{quarterly}} - \mathcal{F}_{s+k+2}^{\text{quarterly}}) \, \mathrm{d}w$$
$$\mathrm{d} \ln \ell_t^{\text{yearly}} = \sum_{s=0}^{t-2} \sum_{k=0}^{\infty} \frac{(\beta^{\text{yearly}})^{k+1}}{\rho^{\text{yearly}}} (\mathcal{F}_{s+k}^{\text{yearly}} - \mathcal{F}_{s+k+2}^{\text{yearly}}) \, \mathrm{d}w.$$

We calculate the value of $\rho^{\text{quarterly}}$ that minimizes the difference between $\dfrac{\mathrm{d} \ln \ell_2^{\text{yearly}}}{\mathrm{d}w}$ and $\dfrac{\mathrm{d} \ln \ell_2^{\text{quarterly}}}{\mathrm{d}w}$:

$$\left\| \sum_{k=0}^{\infty} \frac{(\beta^{\text{quarterly}})^{k+1}}{\rho^{\text{quarterly}}} (\mathcal{F}_k^{\text{quarterly}} - \mathcal{F}_{k+2}^{\text{quarterly}}) - \sum_{k=0}^{\infty} \frac{(\beta^{\text{yearly}})^{k+1}}{\rho^{\text{yearly}}} (\mathcal{F}_k^{\text{yearly}} - \mathcal{F}_{k+2}^{\text{yearly}}) \right\|_2.$$

The resulting value is $\rho = 1.0011$. Figure A.15a plots the elements of $\dfrac{\mathrm{d} \ln \ell_2^{\text{yearly}}}{\mathrm{d}w}$ against the correponding elements of $\dfrac{\mathrm{d} \ln \ell_2^{\text{quarterly}}}{\mathrm{d}w}$. In Figure A.15b, we also compare the (normalized) sectoral value changes at the yearly and quarterly frequency:

$$(1 - \beta^{\text{quarterly}}) \, \mathrm{d}v_0 = (1 - \beta^{\text{quarterly}}) \sum_{k=0}^{\infty} (\beta^{\text{quarterly}})^k \mathcal{F}_k^{\text{quarterly}} \, \mathrm{d}w$$
$$(1 - \beta^{\text{yearly}}) \, \mathrm{d}v_0 = (1 - \beta^{\text{yearly}}) \sum_{k=0}^{\infty} (\beta^{\text{yearly}})^k \mathcal{F}_k^{\text{yearly}} \, \mathrm{d}w.$$

[12] Without loss of generality we can ignore $\{\eta^{nj}\}_{n,j}$ as they only appear in the constant term of the price index.

China Shock. Following CDP, we first compute the predicted increases in US imports from China between 2000 and 2007 using the increases in imports from China of other eight advanced economies during the same period as an instrument. Given this plausibly China-driven increase in imports, we calibrate the increase in China's manufacturing TFP, $\hat{A}_t^{\text{China,manufacturing}}$, from 2000 to 2007 such that the structural model the increase in imports from China that exactly matches the predicted imports increase.

A.4 Omitted Proofs

Proof of Equations (5) and (6). Taking the first-order approximations to equations (2) and (4) and the definition of F_{ijt}, we have

$$\begin{aligned}
\mathrm{d}v_{it} &= \mathrm{d}w_{it} + \rho \sum_j \left. \frac{\Big(\exp(\beta\, \mathbb{E}_t\, v_{jt+1})/\exp(C_{ij})\Big)^{1/\rho}}{\sum_k \Big(\exp(\beta\, \mathbb{E}_t\, v_{kt+1})/\exp(C_{ik})\Big)^{1/\rho}} \right|_{\text{steady state}} \cdot \frac{\beta}{\rho}\, \mathbb{E}_t\, \mathrm{d}v_{jt+1} \\
&= \mathrm{d}w_{it} + \beta \sum_j F_{ij}\, \mathbb{E}_t\, \mathrm{d}v_{jt+1} \qquad \text{(A.17)} \\
\mathrm{d}\ln \ell_{jt+1} &= \sum_i \left. \frac{F_{ijt}\ell_{it}}{\ell_{jt+1}} \right|_{\text{steady state}} \cdot (\mathrm{d}\ln \ell_{it} + \mathrm{d}\ln F_{ijt}) \\
&= \sum_i B_{ji} \cdot (\mathrm{d}\ln \ell_{it} + \mathrm{d}\ln F_{ijt}) \qquad \text{(A.18)} \\
\mathrm{d}\ln F_{ijt} &= \frac{\beta}{\rho}\left(\mathbb{E}_t\, \mathrm{d}v_{jt+1} - \sum_k F_{ik}\, \mathbb{E}_t\, \mathrm{d}v_{kt+1}\right). \qquad \text{(A.19)}
\end{aligned}$$

Plugging equation (A.19) into equation (A.18), we have

$$\begin{aligned}
\mathrm{d}\ln \ell_{jt+1} &= \sum_i B_{ji}\, \mathrm{d}\ln \ell_{it} + \frac{\beta}{\rho} \sum_i B_{ji} \cdot \left(\mathbb{E}_t\, \mathrm{d}v_{jt+1} - \sum_k F_{ik}\, \mathbb{E}_t\, \mathrm{d}v_{kt+1}\right) \\
&= \sum_i B_{ji}\, \mathrm{d}\ln \ell_{it} + \frac{\beta}{\rho} \cdot \left(\mathbb{E}_t\, \mathrm{d}v_{jt+1} - \sum_i B_{ji} \sum_k F_{ik}\, \mathbb{E}_t\, \mathrm{d}v_{kt+1}\right). \qquad \text{(A.20)}
\end{aligned}$$

With vector notation, equations (A.17) and (A.20) become equations (5) and (6). □

Proof of Lemma 1. Equation (5) relates $\mathrm{d}v_t^\omega$ with $\mathrm{d}v_{t+1}^\omega$. Solving this equation forward, we can write $\mathrm{d}v_t^\omega$ as a linear combination of the expected value of a sequence $(\mathrm{d}w_t^\omega, \mathrm{d}w_{t+1}^\omega, \mathrm{d}w_{t+2}^\omega, \cdots)$,

$$\mathrm{d}v_t^\omega = \sum_{k\geq 0} (\beta F^\omega)^k\, \mathbb{E}_t\, \mathrm{d}w_{t+k}^\omega.$$

Plugging this result into equation (6), we obtain a formula that relates $\mathrm{d}\ln \ell_{t+1}^\omega$ with $\mathrm{d}\ln \ell_t^\omega$,

$$\mathrm{d}\ln \ell_{t+1}^\omega = B^\omega\, \mathrm{d}\ln \ell_t^\omega + \frac{\beta}{\rho}(I - B^\omega F^\omega) \sum_{k\geq 0} (\beta F^\omega)^k\, \mathbb{E}_t\, \mathrm{d}w_{t+k+1}^\omega.$$

Solving this equation backward, we can write $\mathrm{d}\ln \ell_t^\omega$ as a linear combination of the expected value of a two-sided sequence $(\cdots, \mathrm{d}w_{t-1}^\omega, \mathrm{d}w_t^\omega, \mathrm{d}w_{t+1}^\omega, \cdots)$,

$$\mathrm{d}\ln \ell_t^\omega = \tfrac{\beta}{\rho} \sum_{s\geq 0} (B^\omega)^s (I - B^\omega F^\omega) \left(\sum_{k\geq 0} (\beta F^\omega)^k \, \mathbb{E}_{t-s-1} \, \mathrm{d}w_{t-s+k}^\omega \right). \qquad \square$$

Proof of Lemma A.3. Write $\theta_j = \exp(\frac{\beta}{\rho} \mathbb{E}\, v_j')$ and $k_{ij} = \exp(\frac{1}{\rho} C_{ij})$, then

$$F_{ij} = \frac{\theta_j / k_{ij}}{\varphi_i} \quad \text{where } \varphi_i = \sum_k \theta_k / k_{ik}.$$

First, assume

$$k_{ij} = \begin{cases} k_i \cdot \tilde{k}_j & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}.$$

Then, we have

$$\varphi_i = \sum_k \theta_j / k_{ik} = \theta_i - \frac{\theta_i}{k_i \tilde{k}_i} + \frac{1}{k_i} \sum_k \frac{\theta_k}{\tilde{k}_k} \tag{A.21}$$

and

$$\begin{aligned} \ell_i &= \sum_k \ell_k F_{ki} = \sum_k \ell_k \frac{\theta_i / k_{ki}}{\varphi_k} \\ &= \ell_i \frac{\theta_i}{\varphi_i} + \ell_i \frac{\theta_i}{\varphi_i} \frac{1}{k_i \tilde{k}_i} + \frac{\theta_i}{\tilde{k}_i} \sum_k \frac{\ell_k}{\varphi_k k_k} \end{aligned}$$

hence, rearranging,

$$\begin{aligned} \ell_i &= \left(1 - \frac{\theta_i}{\varphi_i} \left(1 - \frac{1}{k_i \tilde{k}_i}\right)\right)^{-1} \frac{\theta_i}{\tilde{k}_i} \sum_k \frac{\ell_k}{\varphi_k k_k} \\ &\overset{(A.21)}{=} \frac{\varphi_i \theta_i k_i}{\tilde{k}_i} \frac{\sum_k \frac{\ell_k}{\varphi_k k_k}}{\sum_k \frac{\theta_k}{\tilde{k}_k}}. \end{aligned}$$

Thus, we can show that $\ell_i F_{ij} = \ell_j F_{ji}$ from

$$\frac{\ell_i F_{ij}}{\ell_j F_{ji}} = \frac{\frac{\varphi_i \theta_i k_i}{\tilde{k}_i} \cdot \frac{\theta_j / k_{ij}}{\varphi_i}}{\frac{\varphi_j \theta_j k_j}{\tilde{k}_j} \cdot \frac{\theta_i / k_{ji}}{\varphi_j}} = 1.$$

second, assume instead $k_{ij} = k_{ji}$ for all j, i. Under this assumption, we can easily check that

$$\ell_i = \frac{\theta_i \varphi_i}{\sum_k \theta_k \varphi_k}$$

solves the equation

$$\ell_i = \sum_k \ell_k F_{ki}, \ \forall i.$$

Thus, we again have

$$\frac{\ell_i F_{ij}}{\ell_j F_{ji}} = \frac{\theta_i \varphi_i \frac{\theta_j / k_{ij}}{\varphi_i}}{\theta_j \varphi_j \frac{\theta_i / k_{ji}}{\varphi_j}} = 1. \qquad \square$$

Proof of Lemma 2. It is without loss to assume that there are two sectors. From Lemma A.4, we have

$$\begin{aligned}
(\mathcal{F}_k)_{11} &= 1 - \frac{\sum_\omega \tilde{\beta}_\omega \theta_\omega \alpha_\omega f^k(2 - \alpha_\omega - \beta_\omega)}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} \\
\left((\mathcal{F}_1)^k\right)_{11} &= 1 - \frac{\sum_\omega \tilde{\beta}_\omega \theta_\omega \alpha_\omega}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} f^k\left(2 - \frac{\sum_\omega \tilde{\beta}_\omega \theta_\omega \alpha_\omega}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} - \frac{\sum_\omega \tilde{\alpha}_\omega \theta_\omega \beta_\omega}{\sum_\omega \tilde{\alpha}_\omega \theta_\omega}\right).
\end{aligned}$$

Thus, we need to show that

$$\frac{\sum_\omega \tilde{\beta}_\omega \theta_\omega \alpha_\omega \tilde{f}^k(\alpha_\omega + \beta_\omega)}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} < \frac{\sum_\omega \tilde{\beta}_\omega \theta_\omega \alpha_\omega}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} \tilde{f}^k\left(\frac{\sum_\omega \tilde{\beta}_\omega \theta_\omega \alpha_\omega}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} + \frac{\sum_\omega \tilde{\alpha}_\omega \theta_\omega \beta_\omega}{\sum_\omega \tilde{\alpha}_\omega \theta_\omega}\right)$$

where $\tilde{f}^k(x) = f^k(2 - x)$, or equivalently

$$\frac{\sum_\omega \frac{\alpha_\omega \beta_\omega}{\alpha_\omega + \beta_\omega} \theta_\omega \tilde{f}^k(\alpha_\omega + \beta_\omega)}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} < \frac{\sum_\omega \frac{\alpha_\omega \beta_\omega}{\alpha_\omega + \beta_\omega} \theta_\omega}{\sum_\omega \tilde{\beta}_\omega \theta_\omega} \tilde{f}^k\left(\frac{\sum_\omega \frac{\alpha_\omega \beta_\omega}{\alpha_\omega + \beta_\omega} \theta_\omega}{\left(\sum_\omega \tilde{\beta}_\omega \theta_\omega\right) \cdot \left(\sum_\omega \tilde{\alpha}_\omega \theta_\omega\right)}\right)$$

Define $g^k(x) = x \cdot \tilde{f}^k(x)$, then this becomes

$$\frac{\sum_\omega \tilde{\alpha}_w \tilde{\beta}_w \theta_w g^k(\alpha_w + \beta_w)}{\left(\sum_\omega \tilde{\beta}_\omega \theta_\omega\right) \cdot \left(\sum_\omega \tilde{\alpha}_\omega \theta_\omega\right)} < g^k\left(\frac{\sum_\omega \tilde{\alpha}_w \tilde{\beta}_w \theta_\omega (\alpha_w + \beta_w)}{\left(\sum_\omega \tilde{\beta}_\omega \theta_\omega\right) \cdot \left(\sum_\omega \tilde{\alpha}_\omega \theta_\omega\right)}\right).$$

Define, $\tau_\omega = \frac{\tilde{\alpha}_\omega \tilde{\beta}_\omega \theta_\omega}{(\sum_\omega \tilde{\beta}_\omega \theta_\omega)\cdot(\sum_\omega \tilde{\alpha}_\omega \theta_\omega)}$, then $\kappa \equiv \sum_\omega \tau_w \in (0,1)$. We can inductively show that $g^k(x) \in [0,1]$ is weakly increasing and weakly concave for $x \in [0,1]$. Thus,

$$\begin{aligned} g^k\left(\frac{\sum_\omega \tilde{\alpha}_w \tilde{\beta}_w \theta_\omega (\alpha_w + \beta_w)}{(\sum_\omega \tilde{\beta}_\omega \theta_\omega)\cdot(\sum_\omega \tilde{\alpha}_\omega \theta_\omega)}\right) &= g^k\left(\sum_\omega \tau_w(\alpha_\omega + \beta_\omega)\right) \\ &= g^k\left(\kappa \sum_\omega \frac{\tau_\omega}{\kappa}(\alpha_w + \beta_w)\right) \\ &\geq \kappa g^k\left(\sum_\omega \frac{\tau_\omega}{\kappa}(\alpha_w + \beta_w)\right) \\ &> \kappa \sum_\omega \frac{\tau_\omega}{\kappa} g^k(\alpha_w + \beta_w) \\ &= \frac{\sum_\omega \tilde{\alpha}_w \tilde{\beta}_w \theta_w g^k(\alpha_w + \beta_w)}{(\sum_\omega \tilde{\beta}_\omega \theta_\omega)\cdot(\sum_\omega \tilde{\alpha}_\omega \theta_\omega)}. \end{aligned}$$

□

Proof of Proposition 3. We prove a more general result in Proposition A.3. □

Proof of Proposition 4. For $t \geq 2$,

$$\begin{aligned} \mathrm{d}\ln \ell_t &= \sum_{s\geq 0, k\geq 0} \frac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathbb{E}_{t-s-1}\,\mathrm{d}w_{t-s+k} \\ &= \sum_{s=0}^{t-2}\sum_{k\geq 0} \frac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathrm{d}w_{t-s+k} \\ &= \sum_{k\geq 0} \frac{\beta^{k+1}}{\rho}(\mathcal{F}_k + \mathcal{F}_{k+1} - \mathcal{F}_{t+k-1} - \mathcal{F}_{t+k})\,\mathrm{d}w \end{aligned}$$

where $\mathrm{d}w = (0, \cdots, 0, \overbrace{-\Delta}^{s\text{th}}, 0, \cdots, 0)$. Thus, we can write

$$|\,\mathrm{d}\ln \ell_{st}| = \Delta \cdot \sum_{k\geq 0} \frac{\beta^{k+1}}{\rho}\left(\sum_{s=0}^{t-2} b_{s+k}\right).$$

This gives

$$|\,\mathrm{d}\ln \ell_{st+1}| - |\,\mathrm{d}\ln \ell_{st}| = \Delta \cdot \sum_{k\geq 0} \frac{\beta^{k+1}}{\rho} b_{k+t-1}$$

Under the single-crossing condition of Assumption 3, if this term is higher in the heterogeneous-worker model, then $|\,\mathrm{d}\ln \ell_{st+2}| - |\,\mathrm{d}\ln \ell_{st+1}|$ is also higher in the heterogeneous-worker model. This means that there exists a cutoff $\bar{t} \in \mathbb{N} \cup \{\infty\}$ such that the canonical model calibrated by matching the one-period worker flow matrix overestimates

the decline in employment of sector s in period t if and only if $1 < t \leq \bar{t}$. Note that

$$b_0 + \cdots + b_{t-2} = (I + \mathcal{F}_1 - \mathcal{F}_{t-1} - \mathcal{F}_t)_{ss}$$

is always higher in the canonical model. Thus, $|\,\mathrm{d}\ln \ell_{st}|$ is always higher in the canonical model if β is sufficiently close to zero. In that case, $\bar{t} = \infty$ is possible. On the other hand, for $t = 2$, we have

$$\begin{aligned}
|\,\mathrm{d}\ln \ell_{s2}| &= \Delta \cdot \sum_{k\geq 0} \tfrac{\beta^{k+1}}{\rho}(\mathcal{F}_k - \mathcal{F}_{k+2})_{s,s} \\
&= \Delta \cdot \tfrac{\beta}{\rho}\big((I - \mathcal{F}_2) + \beta(\mathcal{F}_1 - \mathcal{F}_3) + \beta^2(\mathcal{F}_2 - \mathcal{F}_4) + \cdots\big)_{s,s} \\
&= \Delta \cdot \tfrac{\beta}{\rho}\big(I + \beta\mathcal{F}_1 - (1-\beta^2)\mathcal{F}_2 - \beta(1-\beta^2)\mathcal{F}_3 - \cdots\big)_{s,s},
\end{aligned}$$

which is always higher in the canonical model. Thus, $\bar{t} \neq 1$. □

Proof of Proposition A.3. We can write the absolute changes in the value as

$$|\,\mathrm{d}v_{s1}| = \beta^{\tau-1}(\mathcal{F}_{\tau-1})_{s,s}\Delta \geq \beta^{\tau-1}\big((\mathcal{F}_1)^{\tau-1}\big)_{s,s}\Delta = \Big|\,\mathrm{d}v_{s1}\big|_{\text{canonical}}\Big|. \qquad \square$$

Proof of Proposition A.4. For $t \geq 2$,

$$\begin{aligned}
\mathrm{d}\ln \ell_t &= \sum_{s\geq 0, k\geq 0} \tfrac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathbb{E}_{t-s-1}\,\mathrm{d}w_{t-s+k} \\
&= \sum_{s=0}^{t-2}\sum_{k\geq 0} \tfrac{\beta^{k+1}}{\rho}(\mathcal{F}_{s+k} - \mathcal{F}_{s+k+2})\,\mathrm{d}w_{t-s+k} \\
&= \tfrac{\beta}{\rho}(\mathcal{F}_{t-2} - \mathcal{F}_t)\,\mathrm{d}w_2 \\
&+ \Big(\tfrac{\beta}{\rho}(\mathcal{F}_{t-3} - \mathcal{F}_{t-1}) + \tfrac{\beta^2}{\rho}(\mathcal{F}_{t-1} - \mathcal{F}_{t+1})\Big)\,\mathrm{d}w_3 \\
&+ \Big(\tfrac{\beta}{\rho}(\mathcal{F}_{t-4} - \mathcal{F}_{t-2}) + \tfrac{\beta^2}{\rho}(\mathcal{F}_{t-2} - \mathcal{F}_t) + \tfrac{\beta^3}{\rho}(\mathcal{F}_t - \mathcal{F}_{t+2})\Big)\,\mathrm{d}w_4 \\
&+ \cdots \\
&+ \Big(\tfrac{\beta}{\rho}(\mathcal{F}_0 - \mathcal{F}_2) + \cdots + \tfrac{\beta^{t-1}}{\rho}(\mathcal{F}_{2t-4} - \mathcal{F}_{2t-2})\Big)\,\mathrm{d}w_t \\
&+ \Big(\tfrac{\beta^2}{\rho}(\mathcal{F}_1 - \mathcal{F}_3) + \cdots + \tfrac{\beta^t}{\rho}(\mathcal{F}_{2t-3} - \mathcal{F}_{2t-1})\Big)\,\mathrm{d}w_{t+1} \\
&+ \cdots \\
&\equiv \sum_{\tau=2}^{\infty} \boldsymbol{A}_{\tau,t}\,\mathrm{d}w_\tau
\end{aligned}$$

where the impulse response functions are given by

$$\boldsymbol{A}_{\tau,t} = \begin{cases} \sum_{s=0}^{t-2} \frac{\beta^{s+\tau-t+1}}{\rho}(\mathcal{F}_{2s+\tau-t} - \mathcal{F}_{2s+\tau-t+2}) & \text{if } \tau \geq t \\ \sum_{s=0}^{\tau-2} \frac{\beta^{s+1}}{\rho}(\mathcal{F}_{2s+t-\tau} - \mathcal{F}_{2s+t-\tau+2}) & \text{if } t > \tau. \end{cases}$$

Thus, the s-th diagonal element of $\boldsymbol{A}_{\tau,t}$ is a weighted sum of $\{b_{2s+|t-\tau|}\}_{s=0,\ldots,t\wedge\tau-2}$ where more weights are given to those with small s. Thus, under Assumption 3, the s-th diagonal element of $\boldsymbol{A}_{\tau,t}$ is higher in the canonical model when $|t-\tau|$ and/or $t \wedge \tau$ are small. In particular, for given $t \wedge \tau$, we can find $B \in \mathbb{N}$ such that the s-th diagonal element of $\boldsymbol{A}_{\tau,t}$ is higher in the canonical model when $|t-\tau| \leq B$. We can show that $B \geq 1$:

$$\begin{aligned}(\boldsymbol{A}_{t,t})_{ss} &= \sum_{s=0}^{t-2} \frac{\beta^{s+1}}{\rho} b_{2s} \\ &= \tfrac{\beta}{\rho} - \tfrac{\beta(1-\beta)}{\rho}(\mathcal{F}_2)_{ss} - \tfrac{\beta^2(1-\beta)}{\rho}(\mathcal{F}_4)_{ss} - \cdots - \tfrac{\beta^{t-2}(1-\beta)}{\rho}(\mathcal{F}_{2t-4})_{ss} - \tfrac{\beta^m}{\rho}(\mathcal{F}_{2t-2})_{ss},\end{aligned}$$

$$\begin{aligned}(\boldsymbol{A}_{t,t+1})_{ss} &= \sum_{s=0}^{t-2} \frac{\beta^{s+2}}{\rho} b_{2s+1} \\ &= \tfrac{\beta^2}{\rho}(\mathcal{F}_1)_{ss} - \tfrac{\beta^2(1-\beta)}{\rho}(\mathcal{F}_3)_{ss} - \tfrac{\beta^3(1-\beta)}{\rho}(\mathcal{F}_5)_{ss} - \cdots - \tfrac{\beta^{t-1}(1-\beta)}{\rho}(\mathcal{F}_{2t-3})_{ss} - \tfrac{\beta^t}{\rho}(\mathcal{F}_{2t-1})_{ss},\end{aligned}$$

and

$$\begin{aligned}(\boldsymbol{A}_{t+1,t})_{ss} &= \sum_{s=0}^{t-2} \frac{\beta^{s+1}}{\rho} b_{2s+1} \\ &= \tfrac{\beta}{\rho}(\mathcal{F}_1)_{ss} - \tfrac{\beta(1-\beta)}{\rho}(\mathcal{F}_3)_{ss} - \tfrac{\beta^2(1-\beta)}{\rho}(\mathcal{F}_5)_{ss} - \cdots - \tfrac{\beta^{t-2}(1-\beta)}{\rho}(\mathcal{F}_{2t-3})_{ss} - \tfrac{\beta^m}{\rho}(\mathcal{F}_{2t-1})_{ss}.\end{aligned}$$

Thus, by Lemma 2, we can see that $(\boldsymbol{A}_{\tau,t})_{ss}$ is higher in the canonical model if $|t-\tau| \leq 1$. This proves that B maps $\mathbb{N}$ into itself. □

Proof of Lemma A.1. The second equation holds by construction. For the first equation, we have

$$[\bar{\mathbb{E}}_\omega \, \mathrm{d}\ln \ell^\omega_{t+1}]_i = \sum_\omega \tilde{\ell}^\omega_i \, \mathrm{d}\ln \ell^\omega_{it+1} = \frac{\sum_\omega \ell^\omega_i \, \mathrm{d}\ln \ell^\omega_{it+1}}{\ell_i} = \frac{\sum_\omega \mathrm{d}\ell^\omega_{it+1}}{\ell_i} = \frac{\mathrm{d}\ell_{it+1}}{\ell_i} = \mathrm{d}\ln \ell_{it+1}.$$ □

For the remaining equations, fix time t, we then have

$$\begin{aligned}
\left[\bar{\mathbb{E}}_\omega[(F^\omega)^k]\right]_{ij} &= \sum_\omega \tilde{\ell}_i^\omega \left[(F^\omega)^k\right]_{ij} \\
&= \sum_\omega \Pr(\omega|s_t = i) \cdot \Pr(s_t = i, s_{t+k} = j|s_t = i, \omega) \\
&= \sum_\omega \Pr(s_t = i, s_{t+k} = j, \omega|s_t = i) \\
&= \Pr(s_t = i, s_{t+k} = j|s_t = i) \\
&= (\mathcal{F}_k)_{ij}.
\end{aligned}$$

Again fix time t, and define a random variable $\tau(t, m)$ as follows:

$$\tau(t, m) \equiv \min\{\tau > t : s_\tau = s_{t-m}\}$$

which is well-defined based on two assumptions we begin with.

$$\begin{aligned}
\left[\bar{\mathbb{E}}_\omega[(B^\omega)^m(F^\omega)^k]\right]_{ij} &= \sum_\omega \tilde{\ell}_i^\omega \sum_h [(B^\omega)^m]_{ih} \cdot \left[(F^\omega)^k\right]_{hj} \\
&= \sum_\omega \Pr(\omega|s_t = i) \sum_h \Pr(s_{t-m} = h, s_t = i|s_t = i, \omega) \cdot \Pr(s_{\tau(t,m)+k} = j|s_{t-m} = h, s_t = i, \omega) \\
&= \sum_\omega \sum_h \Pr(s_{t-m} = h, s_t = i, s_{\tau(t,m)+k} = j, \omega|s_t = i) \\
&= \sum_\omega \Pr(s_t = i, s_{\tau(t,m)+k} = j, \omega|s_t = i) \\
&= \Pr(s_t = i, s_{\tau(t,m)+k} = j|s_t = i)
\end{aligned}$$

Proof of Lemma A.2. We have

$$\begin{aligned}
\ell_i \cdot \left[\bar{\mathbb{E}}_\omega\left[(B^\omega)^m(F^\omega)^k\right]\right]_{ij} &= \sum_\omega \ell_i^\omega \sum_h [(B^\omega)^m]_{ih} \cdot \left[(F^\omega)^k\right]_{hj} \\
&= \sum_\omega \sum_h \ell_h^\omega [(F^\omega)^m]_{hi} \cdot \left[(F^\omega)^k\right]_{hj} \\
&= \sum_\omega \sum_h \ell_j^\omega [(F^\omega)^m]_{hi} \cdot \left[(B^\omega)^k\right]_{jh} \\
&= \ell_j \left[\bar{\mathbb{E}}_\omega (B^\omega)^k (F^\omega)^m\right]\right]_{ji}.
\end{aligned}$$

Thus,

$$(LHS)_{ij} = \ell_i \cdot \left[\bar{\mathbb{E}}_\omega\left[(B^\omega)^m (F^\omega)^k\right]\right]_{ij} = \ell_j \left[\bar{\mathbb{E}}_\omega\left(B^\omega\right)^k (F^\omega)^m\right]\Big]_{ji} = (RHS)_{ij} \qquad \square$$

Proof of Proposition A.2. Consider shifts in the distributions of Δ_i and Δ_j by $\rho \cdot \mathrm{d}\Delta_i$ and $\rho \cdot \mathrm{d}\Delta_j$ units, respectively. Denote the share of workers working in sector s as L_s. The following lemma represents levels and changes of sector choice probabilities in terms of $\mathrm{d}\Delta_i$ and $\mathrm{d}\Delta_j$.

Lemma A.7. *When $g_{ij}(\cdot)$ is continuous around 0 and ε_{ij} has a finite first moment for all $i, j \in \mathcal{S}$, we have*

$$\begin{aligned}
F_{ij} &= \frac{\rho g_{ij}(0)\mathcal{F}_1}{L_i} + o(\rho) \\
\mathrm{d}F_{ijt} &= \left(\frac{\rho g_{ij}(0)\mathcal{F}_2}{L_i} + o(\rho)\right) \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) \\
\mathrm{d}\ln F_{ijt} &= (\mathcal{F} + o(1)) \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) \\
F_{ii} &= 1 + o(1) \\
\mathrm{d}F_{iit} &= \mathcal{F} \cdot \left(\mathrm{d}\Delta_i - \sum_j F_{ij}\,\mathrm{d}\Delta_j\right) + o(\rho) \\
\mathrm{d}\ln F_{iit} &= \mathcal{F} \cdot \left(\mathrm{d}\Delta_i - \sum_j F_{ij}\,\mathrm{d}\Delta_j\right) + o(1)
\end{aligned}$$

where the constants $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}$ only depend on $F(\cdot)$ and $\tilde{C}$.

Plugging the final equation into equation (A.18), we have

$$\begin{aligned}
\mathrm{d}\ln \ell_i &= \sum_j B_{ij}\,\mathrm{d}\ln F_{jit} \\
&= F_{ii}\,\mathrm{d}\ln F_{iit} + \sum_{j\neq i} B_{ij}\,\mathrm{d}\ln F_{jit} \\
&= \mathrm{d}F_{iit} + \sum_{j\neq i} B_{ij}\,\mathrm{d}\ln F_{jit} \\
&= \mathcal{F} \cdot \left(\mathrm{d}\Delta_i - \sum_j F_{ij}\,\mathrm{d}\Delta_j\right) + \sum_{j\neq i} T_{ij} \cdot \mathcal{F} \cdot (\mathrm{d}\Delta_i - \mathrm{d}\Delta_j) + o(\rho) \\
&= \mathcal{F} \cdot \left(2\,\mathrm{d}\Delta_i - \sum_j F_{ij}\,\mathrm{d}\Delta_j - \sum_j T_{ij}\,\mathrm{d}\Delta_j\right) + o(\rho)
\end{aligned}$$

Lemma A.8. *We have $2I - F - B = I - BF + o(\rho)$.*

Proof. This directly follows from $(I-B)(I-F)=o(\rho)$, which in turn comes from the fact that all the elements of both matrices $I-F$ and $I-B$ are proportional to ρ. □

Thus, we obtain

$$\begin{aligned}\mathrm{d}\ln\ell &= \mathcal{F}\cdot(2I-F-B)\,\mathrm{d}\Delta + o(\rho)\\ &= \mathcal{F}\cdot\tfrac{\beta}{\rho}(I-BF)\,\mathbb{E}_t\,\mathrm{d}v_{t+1} + o(\rho)\end{aligned}$$ □

Proof of Lemma A.7. We start with the following lemma.

Lemma A.9 (Conditional Distribution). *Suppose that in the steady state, shares ℓ_i and ℓ_j of workers in $\Omega_{ij}(\varepsilon)$ are working in sectors i and j, respectively, and that they have*

$$\Delta_{ij}|_{\text{in sector } i,\,\Omega_{ij}(\varepsilon)} \sim g^i_{ij}(\cdot), G^i_{ij}(\cdot) \text{ and } \Delta_{ij}|_{\text{in sector } j,\Omega_{ij}(\varepsilon)} \sim g^j_{ij}(\cdot), G^j_{ij}(\cdot).$$

Then, these conditional distributions satisfy

$$g^i_{ij}(x)\ell_i = \frac{F\left(\frac{x}{\rho}-\tilde{C}_{ji}\right)}{F\left(\frac{x}{\rho}-\tilde{C}_{ji}\right)+F\left(-\frac{x}{\rho}-\tilde{C}_{ij}\right)}g_{ij}(x).$$

Without a shock, the share of workers in $\Omega_{ij}(\varepsilon)$ who move from i to j is given by

$$\begin{aligned}\ell_{ij} = \ell_i\Pr(\Delta_{ij}+\rho\varepsilon_{ij}<-\rho\tilde{C}_{ij}|\text{in } i) &= \begin{cases}\ell_i\cdot\int_{-\infty}^{\infty} G^i_{ij}(-\rho\varepsilon_{ij}-\rho\tilde{C}_{ij})\,\mathrm{d}F(\varepsilon_{ij})\\ \ell_i\cdot\int_{-\infty}^{\infty} F\left(\frac{-\Delta_{ij}-\rho\tilde{C}_{ij}}{\rho}\right)\mathrm{d}G^i_{ij}(\Delta_{ij})\end{cases}\\ &= \int_{-\infty}^{\infty}\frac{F\left(\frac{x}{\rho}-\tilde{C}_{ji}\right)\cdot F\left(-\frac{x}{\rho}-\tilde{C}_{ij}\right)}{F\left(\frac{x}{\rho}-\tilde{C}_{ji}\right)+F\left(-\frac{x}{\rho}-\tilde{C}_{ij}\right)}g_{ij}(x)\,\mathrm{d}x\\ &= \rho\int_{-\infty}^{\infty}\frac{F(t-\tilde{C}_{ji})\cdot F(-t-\tilde{C}_{ij})}{F(t-\tilde{C}_{ji})+F(-t-\tilde{C}_{ij})}g_{ij}(\rho t)\,\mathrm{d}t\end{aligned} \tag{A.22}$$

Suppose that the distributions of Δ_i and Δ_j shift by $\rho\cdot\mathrm{d}\Delta_i$ and $\rho\cdot\mathrm{d}\Delta_j$, respectively. Then, new share is given by

$$\begin{aligned}\ell_{ij}+\mathrm{d}\ell_{ij} &= \ell_i\cdot\Pr(\Delta_{ij}+\rho(\mathrm{d}\Delta_i-\mathrm{d}\Delta_j)+\rho\varepsilon_{ij}<-\rho\tilde{C}_{ij}|\text{in } i)\\ &= \ell_i\cdot\int_{-\infty}^{\infty}F\big(-\tfrac{\Delta_{ij}}{\rho}-\tilde{C}_{ij}-(\mathrm{d}\Delta_i-\mathrm{d}\Delta_j)\big)\,\mathrm{d}G^i_{ij}(\Delta_{ij})\\ &= \int_{-\infty}^{\infty}\frac{F\left(\frac{x}{\rho}-\tilde{C}_{ji}\right)\cdot F\left(-\frac{x}{\rho}-\tilde{C}_{ij}-(\mathrm{d}\Delta_i-\mathrm{d}\Delta_j)\right)}{F\left(\frac{x}{\rho}-\tilde{C}_{ji}\right)+F\left(-\frac{x}{\rho}-\tilde{C}_{ij}\right)}g_{ij}(x)\,\mathrm{d}x\end{aligned}$$

$$= \int_{-\infty}^{\infty} \frac{F\Big(\frac{x}{\rho} - \tilde{C}_{ji}\Big) \cdot \Big(F\Big(-\frac{x}{\rho} - \tilde{C}_{ij}\Big) - f\Big(-\frac{x}{\rho} - \tilde{C}_{ij}\Big) \cdot (\mathrm{d}\Delta_i - \mathrm{d}\Delta_j)\Big)}{F\Big(\frac{x}{\rho} - \tilde{C}_{ji}\Big) + F\Big(-\frac{x}{\rho} - \tilde{C}_{ij}\Big)} g_{ij}(x)\,\mathrm{d}x$$

Thus,

$$\begin{aligned}
\mathrm{d}\ell_{ij} &= \int_{-\infty}^{\infty} \frac{F\Big(\frac{x}{\rho} - \tilde{C}_{ji}\Big) \cdot f\Big(-\frac{x}{\rho} - \tilde{C}_{ij}\Big)}{F\Big(\frac{x}{\rho} - \tilde{C}_{ji}\Big) + F\Big(-\frac{x}{\rho} - \tilde{C}_{ij}\Big)} g_{ij}(x)\,\mathrm{d}x \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) \\
&= \rho \int_{-\infty}^{\infty} \frac{F\big(t - \tilde{C}_{ji}\big) \cdot f\big(-t - \tilde{C}_{ij}\big)}{F\big(t - \tilde{C}_{ji}\big) + F\big(-t - \tilde{C}_{ij}\big)} g_{ij}(\rho t)\,\mathrm{d}t \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) \qquad \text{(A.23)}
\end{aligned}$$

We will take a limit $\rho \to 0$ to equations (A.22) and (A.23). Note that we have

$$\begin{aligned}
\left| \frac{F(t - \tilde{C}_{ji}) \cdot F(-t - \tilde{C}_{ij})}{F(t - \tilde{C}_{ji}) + F(-t - \tilde{C}_{ij})} g_{ij}(\rho t) \right| &\le \big(F(-t - \tilde{C}_{ij})\mathbb{1}_{t \ge 0} + F(t - \tilde{C}_{ji})\mathbb{1}_{t<0}\big) \cdot \sup_x \{g_{ij}(x)\} \\
\left| \frac{F\big(t - \tilde{C}_{ji}\big) \cdot f\big(-t - \tilde{C}_{ij}\big)}{F\big(t - \tilde{C}_{ji}\big) + F\big(-t - \tilde{C}_{ij}\big)} g_{ij}(\rho t) \right| &\le f(-t - \tilde{C}_{ij}) \cdot \sup_x \{g_{ij}(x)\}
\end{aligned}$$

with

$$\begin{aligned}
\int_0^{\infty} F(-t - \tilde{C}_{ij})\,\mathrm{d}t + \int_{-\infty}^{0} F(t - \tilde{C}_{ji})\,\mathrm{d}t &= \int_0^{\infty} \int_{-\infty}^{-t-\tilde{C}_{ij}} f(x)\,\mathrm{d}x\,\mathrm{d}t + \int_{-\infty}^{0} \int_{-\infty}^{t-\tilde{C}_{ji}} f(x)\,\mathrm{d}x\,\mathrm{d}t \\
&= \int_{-\infty}^{-\tilde{C}_{ij}} \int_0^{-x-\tilde{C}_{ij}} f(x)\,\mathrm{d}t\,\mathrm{d}x + \int_{-\infty}^{-\tilde{C}_{ji}} \int_{x+\tilde{C}_{ji}}^{0} f(x)\,\mathrm{d}t\,\mathrm{d}x \\
&= \int_{-\infty}^{-\tilde{C}_{ij}} (-x - \tilde{C}_{ij}) f(x)\,\mathrm{d}x + \int_{-\infty}^{-\tilde{C}_{ji}} (-x - \tilde{C}_{ji}) f(x)\,\mathrm{d}x,
\end{aligned}$$

which is finite because $\mathbb{E}[\varepsilon_{ij}]$ is well-defined, and

$$\int_{-\infty}^{\infty} f(-t - \tilde{C}_{ij})\,\mathrm{d}t = 1,$$

which is also finite. Thus, we can apply dominated convergence theorem to conclude

$$\begin{aligned}
\ell_{ij} &= \rho \cdot \underbrace{\int_{-\infty}^{\infty} \frac{F(t - \tilde{C}_{ji}) \cdot F(-t - \tilde{C}_{ij})}{F(t - \tilde{C}_{ji}) + F(-t - \tilde{C}_{ij})}\,\mathrm{d}t}_{\mathcal{F}_1^{ij}} \cdot g_{ij}(0) + o(\rho) \\
\mathrm{d}\ell_{ij} &= \left(\rho \cdot \underbrace{\int_{-\infty}^{\infty} \frac{F(t - \tilde{C}_{ji}) \cdot f(-t - \tilde{C}_{ij})}{F(t - \tilde{C}_{ji}) + F(-t - \tilde{C}_{ij})}\,\mathrm{d}t}_{\mathcal{F}_2^{ij}} \cdot g_{ij}(0) + o(\rho) \right) \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i)
\end{aligned}$$

Assume $\tilde{C}_{ji} = \tilde{C}_{ij} = \tilde{C}$ for all $i \neq j$ and write $\mathcal{F}_1 = \mathcal{F}_1^{ij}$ and $\mathcal{F}_2 = \mathcal{F}_2^{ij}$. Then, we have

$$\begin{aligned}
F_{ij} = \frac{\ell_{ij}}{L_i} &= \frac{\rho g_{ij}(0)\mathcal{F}_1}{L_i} + o(\rho) \\
\mathrm{d}F_{ij} = \frac{\mathrm{d}\ell_{ij}}{L_i} &= \left(\frac{\rho g_{ij}(0)\mathcal{F}_2}{L_i} + o(\rho)\right) \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) \\
\implies \mathrm{d}\ln F_{ij} &= \left(\frac{\mathcal{F}_2}{\mathcal{F}_1} + o(1)\right) \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) \equiv (\mathcal{F} + o(1)) \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i)
\end{aligned}$$

for all $i \neq j$, Thus,

$$\begin{aligned}
F_{ii} &= 1 - \sum_{j \neq i} F_{ij} = 1 + o(1) \\
\mathrm{d}F_{ii} &= -\sum_{j \neq i} \mathrm{d}F_{ij} = -\sum_{j \neq i} F_{ij}\,\mathrm{d}\ln F_{ij} = -\sum_{j \neq i} F_{ij}\mathcal{F} \cdot (\mathrm{d}\Delta_j - \mathrm{d}\Delta_i) + o(\rho) \\
&= \mathcal{F} \cdot \left(\mathrm{d}\Delta_i - \sum_j F_{ij}\,\mathrm{d}\Delta_j\right) + o(\rho) \\
\implies \mathrm{d}\ln F_{ii} &= \mathcal{F} \cdot \left(\mathrm{d}\Delta_i - \sum_j F_{ij}\,\mathrm{d}\Delta_j\right) + o(1).
\end{aligned}$$

□

Proof of Lemma A.9. In the steady state, we have

$$\begin{aligned}
\ell_i G_{ij}^i(x) &= \ell_i \Pr(\Delta_{ij} \leq x | \text{in } i) \\
&= \ell_i \Pr(\Delta_{ij} \leq x | \text{in } i) \Pr(\text{stay in } i | \Delta_{ij} \leq x, \text{in } i) + \ell_j \Pr(\Delta_{ij} \leq x | \text{in } j) \Pr(\text{leave } j | \Delta_{ij} \leq x, \text{in } j) \\
&= \ell_i \int_{-\infty}^{x} \underbrace{\Pr(\Delta_{ij} + \rho\varepsilon_{ij} \geq -\rho\tilde{C}_{ij})}_{=\bar{F}(-\frac{\Delta_{ij}}{\rho} - \tilde{C}_{ij})} \mathrm{d}G_{ij}^i(\Delta_{ij}) + \ell_j \int_{-x}^{\infty} \underbrace{\Pr(\Delta_{ji} + \rho\varepsilon_{ji} < -\rho\tilde{C}_{ji})}_{=F(-\frac{\Delta_{ji}}{\rho} - \tilde{C}_{ji})} \mathrm{d}G_{ji}^j(\Delta_{ji})
\end{aligned}$$

Differentiating with respect to x, we have

$$\ell_i g_{ij}^i(x) = \ell_i \bar{F}\left(-\frac{x}{\rho} - \tilde{C}_{ij}\right) g_{ij}^i(x) + \ell_j F\left(\frac{x}{\rho} - \tilde{C}_{ji}\right) g_{ji}^j(-x)$$

or

$$F\left(-\frac{x}{\rho} - \tilde{C}_{ij}\right) g_{ij}^i(x)\ell_i = F\left(\frac{x}{\rho} - \tilde{C}_{ji}\right) g_{ij}^j(x)\ell_j.$$

We also have

$$g_{ij}^i(x)\ell_i + g_{ij}^j(x)\ell_j = g_{ij}(x),$$

so we can conclude that

$$g_{ij}^i(x)\ell_i = \frac{F\left(\frac{x}{\rho} - \tilde{C}_{ji}\right)}{F\left(\frac{x}{\rho} - \tilde{C}_{ji}\right) + F\left(-\frac{x}{\rho} - \tilde{C}_{ij}\right)} g_{ij}(x). \qquad \square$$

Proof of Lemma A.4.

$$\begin{aligned} F_i^{k+1} &= \begin{pmatrix} 1-\alpha_i f^k & \alpha_i f^k \\ \beta_i f^k & 1-\beta_i f^k \end{pmatrix} \begin{pmatrix} \bar{\alpha}_i & \alpha_i \\ \beta_i & \bar{\beta}_i \end{pmatrix} \\ &= \begin{pmatrix} \bar{\alpha}_i - \alpha_i\bar{\alpha}_i f^k + \alpha_i\beta_i f^k & \alpha_i(1-\alpha_i f^k + \bar{\beta}_i f^k) \\ \beta_i(\bar{\alpha}_i f^k + 1 - \beta_i f^k) & \alpha_i\beta_i f^k + \bar{\beta}_i - \beta_i\bar{\beta}_i f^k \end{pmatrix} \\ &= \begin{pmatrix} 1-\alpha_i f^{k+1} & \alpha_i f^{k+1} \\ \beta_i f^{k+1} & 1-\beta_i f^{k+1} \end{pmatrix}. \end{aligned}$$

Thus, we have $f^{k+1}(x) = (x-1)\cdot f^k(x) + 1$. Define $g^k(x) = x \cdot f^k(2-x) - 1$, then we have

$$\begin{aligned} g^{k+1}(x) &= x \cdot f^{k+1}(2-x) - 1 \\ &= x \cdot \big((1-x)\cdot f^k(2-x) + 1\big) - 1 \\ &= (1-x)g^k(x). \end{aligned}$$

and $g^1(x) = -(1-x)$. Thus, $g^k(x) = -(1-x)^k$ and hence $f^k(x) = \frac{1+g^k(2-x)}{2-x} = \frac{1-(x-1)^k}{2-x}$. $\square$

Proof of Proposition A.5. Note from the definitions of x_i and y_i that we have

$$\frac{\alpha_i x_i}{\alpha_1 x_1} = \frac{\beta_i y_i}{\beta_1 y_1}.$$

Let λ_i denote this ratio. We then have

$$\frac{(\mathcal{F}_k)_{1,2}}{(\mathcal{F}_k)_{2,1}} = \frac{\sum_i x_i\alpha_{i,k}}{\sum_i y_i\beta_{i,k}} = \frac{\sum_i \alpha_i x_i\left(\frac{\alpha_{i,k}}{\alpha_i}\right)}{\sum_i \beta_i y_i\left(\frac{\beta_{i,k}}{\beta_i}\right)} = \frac{\alpha_1 x_1 \sum_i \lambda_i\left(\frac{\alpha_{i,k}}{\alpha_i}\right)}{\beta_1 y_1 \sum_i \lambda_i\left(\frac{\beta_{i,k}}{\beta_i}\right)} = \frac{\alpha_1 x_1}{\beta_1 y_1}$$

where the last equality uses $\frac{\alpha_{i,k}}{\alpha_i} = \frac{\beta_{i,k}}{\beta_i} = f^k(\bar{\alpha}_i + \bar{\beta}_i)$. $\square$

A.5 Additional Figures

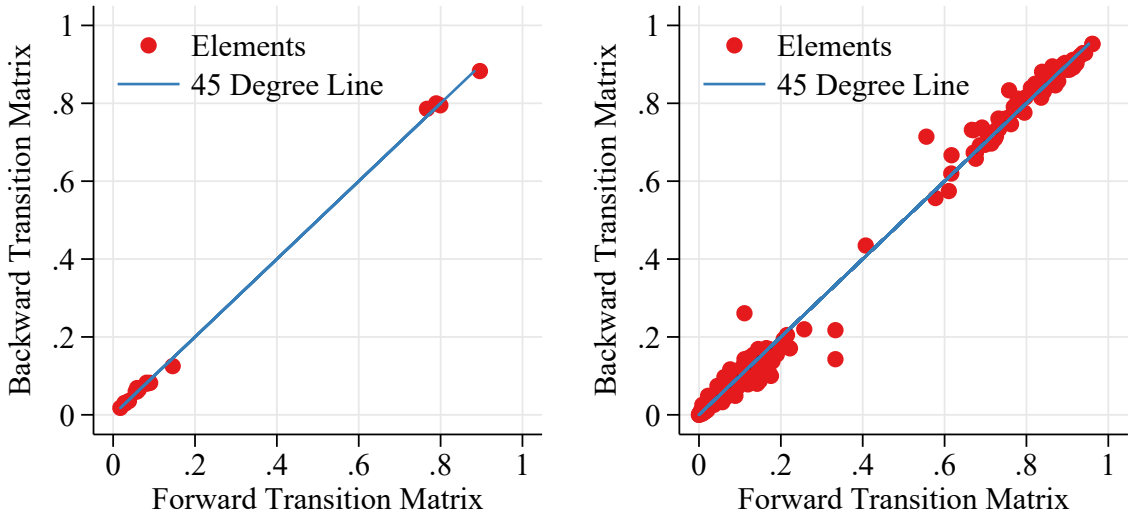


Figure A.2. The Backward and Forward Transition Matrices: NLSY

Notes: Assuming that the economy was in a steady state between years 1980 and 2000, the backward and forward transition matrices are computed by pooling all observations of the NLSY79 data over this period. In the left panel, we plot the elements of the aggregate forward transition matrix against those of the aggregate backward transition matrix. With four sectors, the backward and forward transition matrices are four-by-four matrices with sixteen elements. In the right panel, we consider four dimensions of observed heterogeneity—sex, race, education, and age, leading to sixteen groups—and compare the matrices for all sixteen groups.

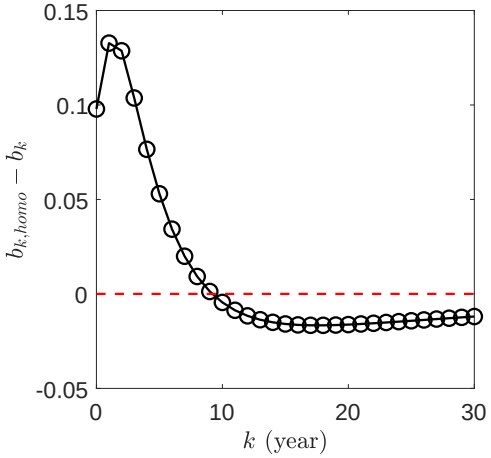


Figure A.3. Differences in b_k Series: Manufacturing Sector

Notes: For each k, this figure plots the difference between the diagonal elements of $(\mathcal{F}_k - \mathcal{F}_{k+2})$ corresponding to the manufacturing sector implied by the canonical model and those observed in the data. For the canonical model, we compute $\mathcal{F}_k$ by multiplying $\mathcal{F}_1$ k times. Since we only observe a finite number of worker flow matrices in the data, we extrapolate it using the estimated structural model. The same pattern is observed for the other sectors. Data source: NLSY79.

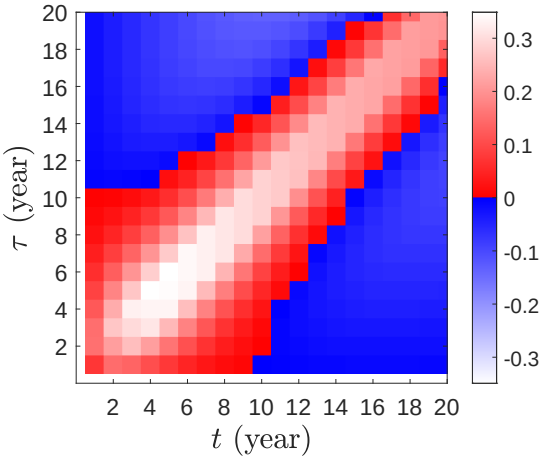


Figure A.4. Differences in the Response of Sectoral Employment

Notes: This figure plots the values of $\frac{\partial \ln \ell_{s,t}}{\partial w_{s,\tau}}\Big|_{\text{canonical}} - \frac{\partial \ln \ell_{s,t}}{\partial w_{s,\tau}}\Big|_{\text{data}}$ for $1 \leq t, \tau \leq 20$. These derivatives are calculated from equation (11) using the worker flow matrices $\mathcal{F}_k$ from the NLSY79 data for $\frac{\partial \ln \ell_{s,t}}{\partial w_{s,\tau}}\Big|_{\text{data}}$ and using the worker flow matrices implied by the canonical model for $\frac{\partial \ln \ell_{s,t}}{\partial w_{s,\tau}}\Big|_{\text{canonical}}$.

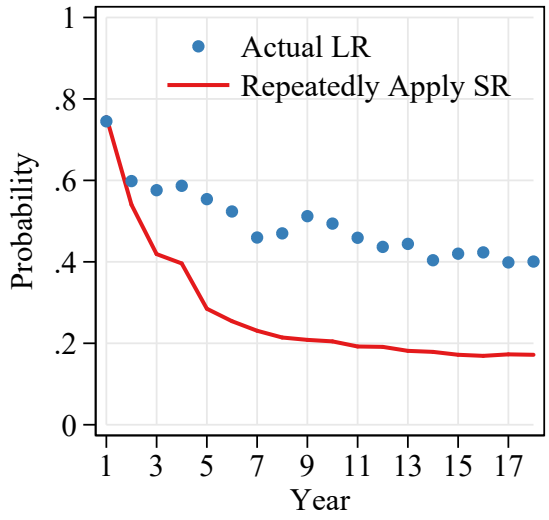


Figure A.5. Actual and Model-Implied Staying Probabilities: Non-Stationarity

Notes: For each $k = 1, \ldots, 18$ (year), this figure plots the actual probability of workers who choose manufacturing in 1980 choosing manufacturing again in year $1980+k$ (blue dots) and the probability computed by repeatedly multiplying time-varying one-year worker flow matrices (red line); i.e., the diagonal element of the matrix $\prod_{\kappa=1}^{k-1} \mathcal{F}_1^{\kappa}$ corresponding to the manufacturing sector. $\mathcal{F}_1^{\kappa}$ is the aggregate worker flow matrix computed using transition observations between years $1980 + \kappa - 1$ and $1980 + \kappa$ from the NLSY data.

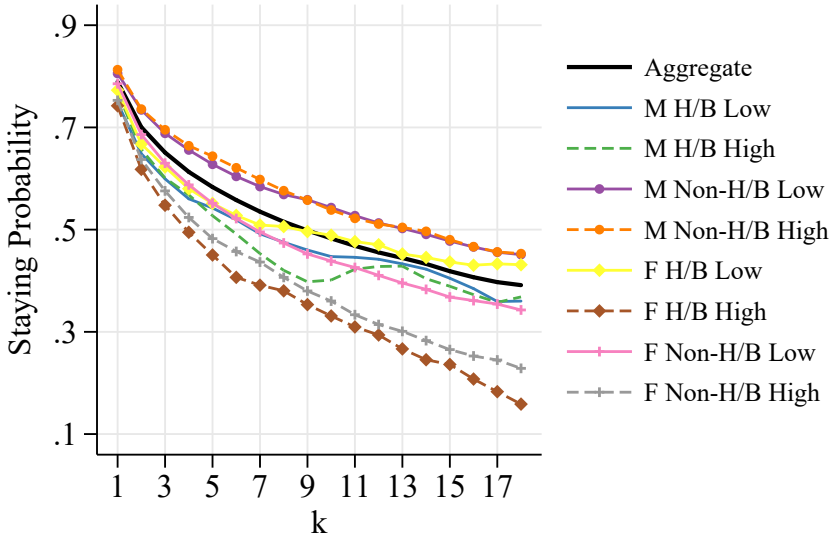
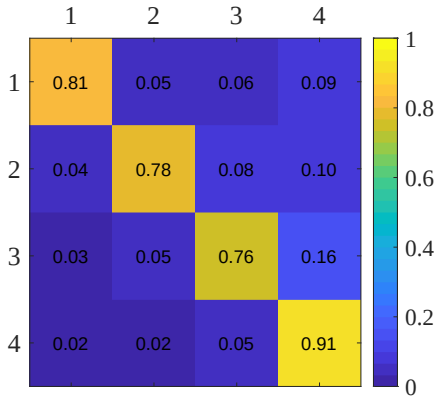


Figure A.6. Actual Staying Probabilities for Different Worker Groups

Notes: For each unique combination of (male, female), (Hispanic/Black, non-Hispanic/Black), and (low-skilled, high-skilled), this figure plots the steady-state k-year manufacturing staying probabilities, $\Pr(s_{t+k} = \text{manufacturing} | s_t = \text{manufacturing})$. Data source: NLSY79.

Wage	Switching Cost			
$\begin{pmatrix} 0.87 \\ 0.85 \\ 0.84 \\ 1.00 \end{pmatrix}$	$\begin{pmatrix} 0.00 & 2.86 & 2.95 & 3.15 \\ 2.86 & 0.00 & 2.46 & 2.88 \\ 2.95 & 2.46 & 0.00 & 2.22 \\ 3.15 & 2.88 & 2.22 & 0.00 \end{pmatrix}$			

(a) Primitives



(b) Transition Matrix

Figure A.7. Estimation Result: Canonical Model

Notes: Panel (a) shows the estimated values of the primitives of the canonical model. The four sectors are Agriculture and Construction; Manufacturing; Communications and Trade; and Services and Others. Panel (b) shows the resulting transition matrix (or, equivalently, one-year worker flow matrix).

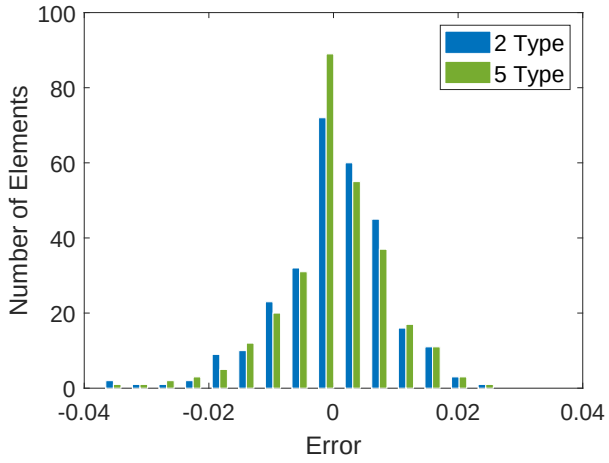
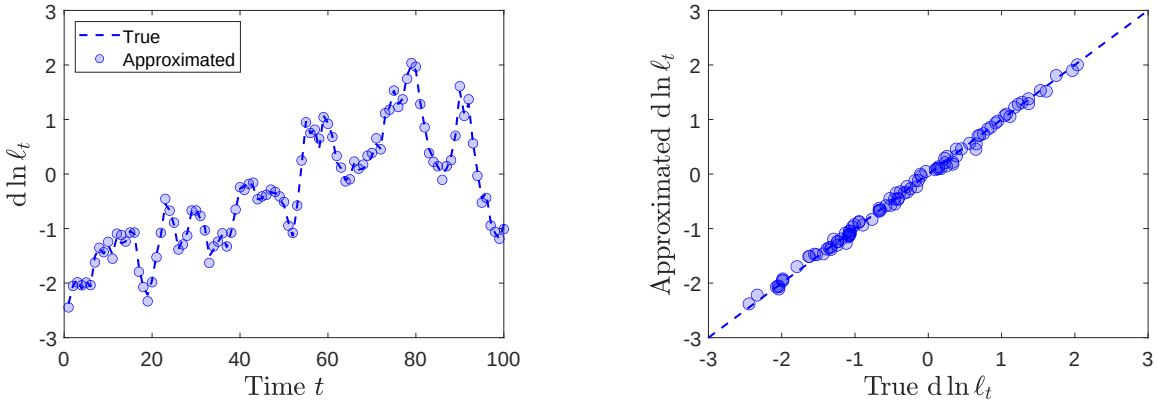


Figure A.8. Fits of Two-type Model and Five-type Model

Notes: For each of the two-type and five-type models, we compute the model-implied worker-flow matrix series, $\{\mathcal{F}_k\}_{k=1,\ldots,18}$. There are a total of 288 elements. This figure plots the differences between these elements and the actual values observed in the data for each of the two models. The five-type model provides a slightly better fit, but the difference is not significant.



(a) (b)

Figure A.9. Fit of Recursive Representation

Notes: Plugging randomly generated values of $\{\mathrm{d}w_t\}$ (from standard normal distribution) and worker flow matrices computed from the NLSY data (extrapolated using the structural model) into equation (15), we can generate a sequence of changes in sectoral employment $\{\mathrm{d}\ln \ell_t\}$. Using the computed values of $\{\Gamma_k\}_{k=1,\ldots 6}$ and $\{\Lambda_k\}_{k=1,\ldots 5}$ and equation (17) instead, we can also compute an approximated sequence of changes in sectoral employment. These figures compare the actual sequence with the approximated sequence.

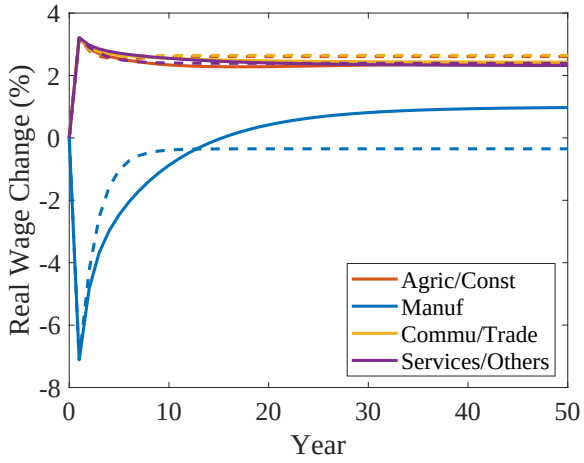


Figure A.10. Changes in Sectoral Real Wages

Notes: This figure plots changes in sectoral real wages over time following an unexpected permanent drop in manufacturing prices. Solid lines correspond to the prediction from the sufficient statistics in the data, and dashed lines correspond to the prediction of the canonical model, without persistent worker heterogeneity.

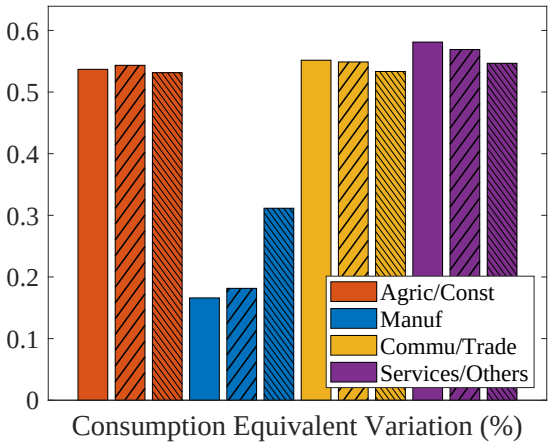
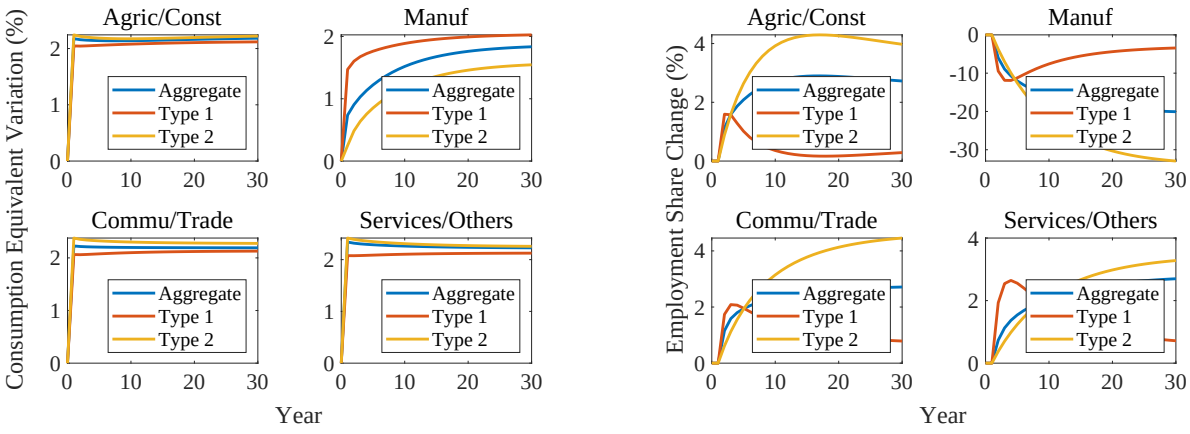


Figure A.11. Changes in Sectoral Values: Exogenous and Endogenous Wage Changes

Notes: This figure plots changes in sectoral values in terms of consumption-equivalent variation for workers initially employed in different sectors. The plain bars correspond to the prediction from sufficient statistics in the data. The rightmost hatched bars (Northwest to Southeast) correspond to the prediction of the canonical model. The hatched bars in the middle (Northeast to Southwest) correspond to the prediction made by combining the wage changes of the canonical model and the sufficient statistics.



(a) Changes in Sectoral Values

(b) Change in Sectoral Employment Share

Figure A.12. Counterfactual Changes in Employment Share and Welfare: Type-specific Changes

Notes: This figure plots the transitional dynamics following an unexpected permanent drop in manufacturing prices for each type of worker. The orange line corresponds to type-1 workers (frequent movers), and the yellow line corresponds to type-2 workers (infrequent movers).

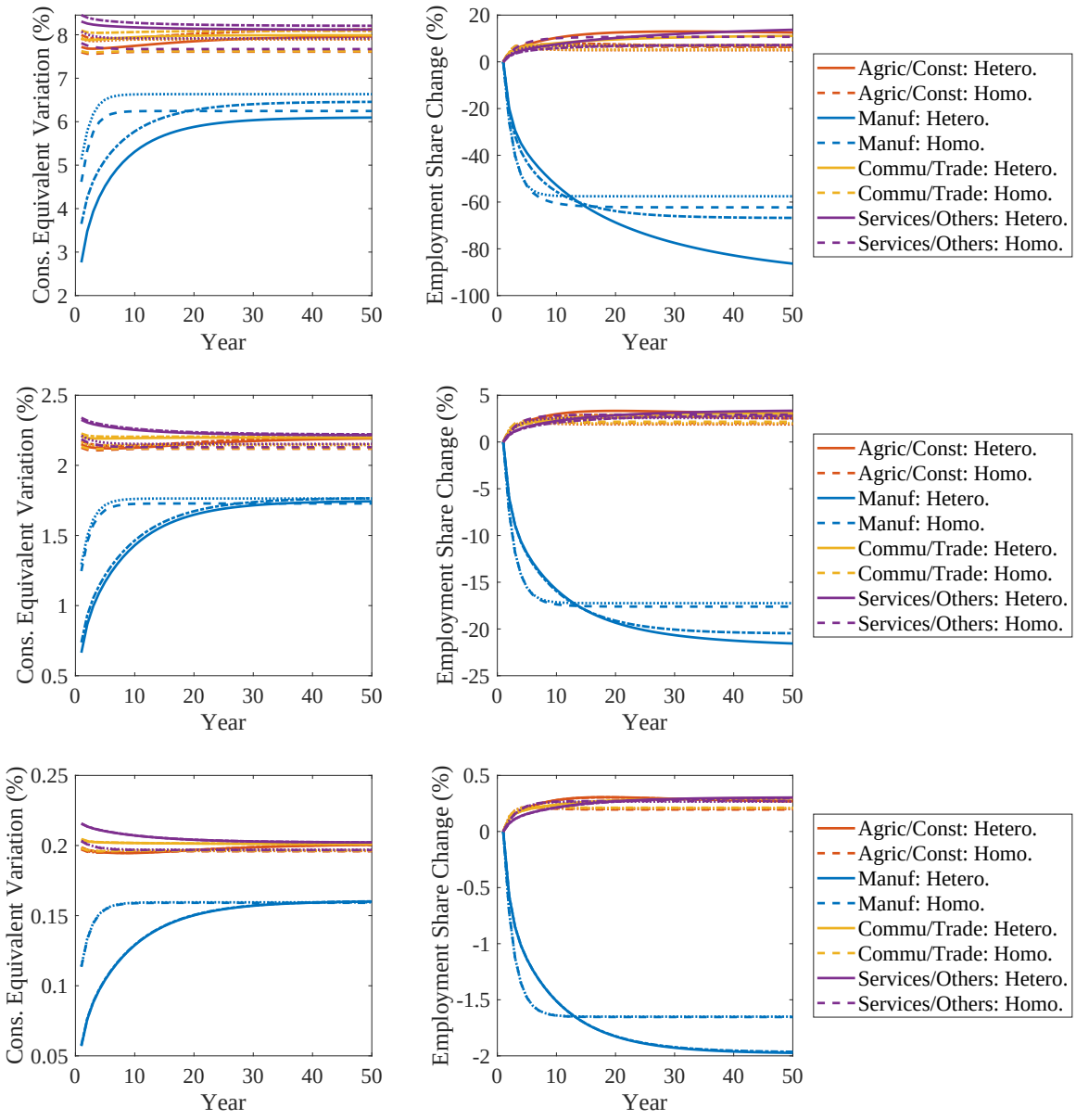


Figure A.13. The Quality of First-Order Approximation

Notes: This figure compares the transitional dynamics of welfare (left column) and employment share (right column) obtained using sufficient statistics formula with those calculated from the exact solution of the estimated structural model. The top row corresponds to 30% drop in manufacturing prices, the middle row to 10%, and the bottom row to 1%.

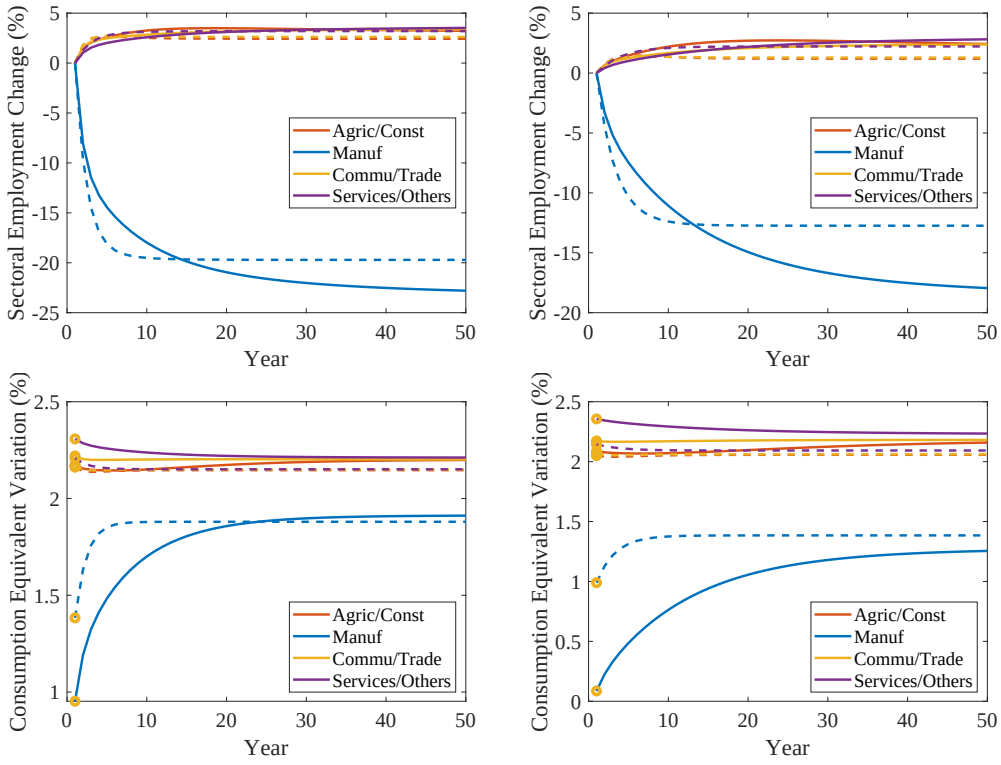
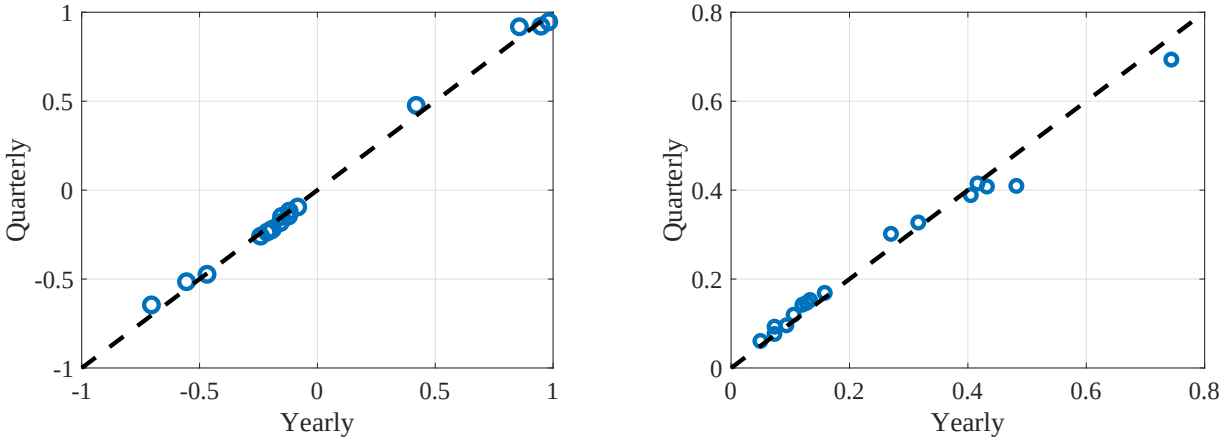


Figure A.14. Counterfactual Exercises with Different Values of ρ

Notes:



(a) Coefficients on Sectoral Employment Changes

(b) Coefficients on Sectoral Welfare Changes

Figure A.15. Yearly to Quarterly

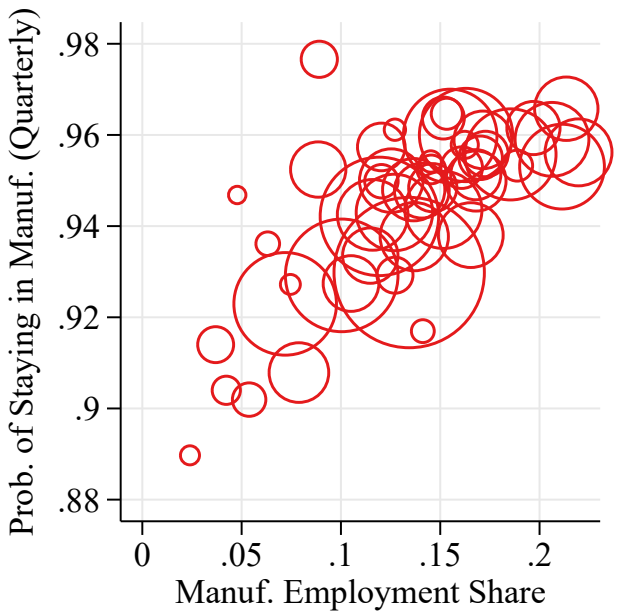


Figure A.16. Manufacturing Shares and Staying Probabilities Across States

Notes: This figure plots the probability of working in manufacturing sector again after one year against the manufacturing employment share. The size of the circle represents the relative size of the manufacturing sector. It is the largest in California, followed by Texas, Ohio, and Michigan.

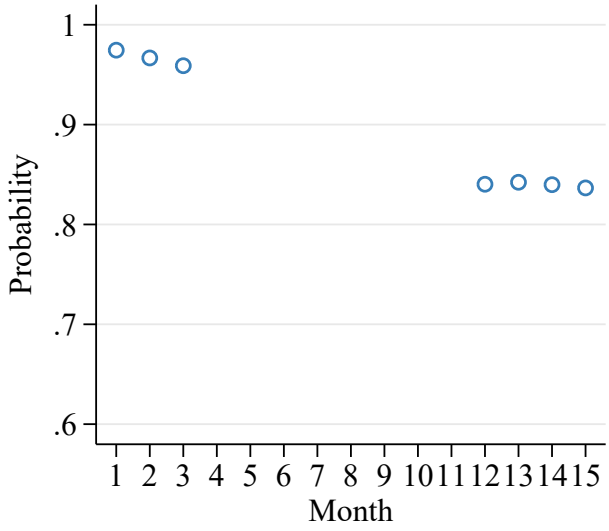


Figure A.17. Aggregate Worker Flow Matrix Series: Monthly CPS

Notes: This figure clearly shows that the initial three circles are not in line with the remaining four circles. In particular, the last four circles should be shifted upwards. This is a well-known problem of the CPS dataset.

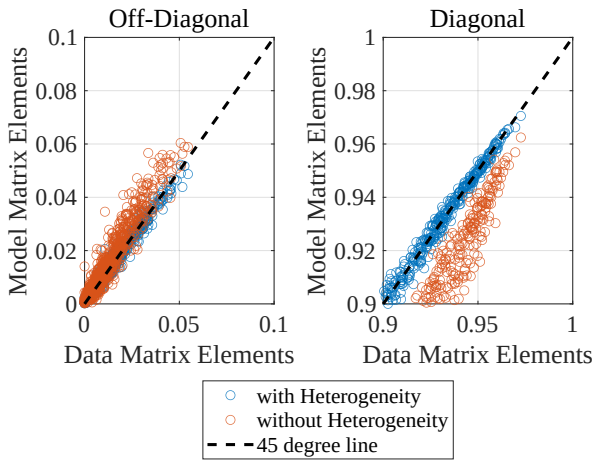


Figure A.18. Fit of the Models with and without Heterogeneity: State-Level Worker Flow Matrices

Notes: In this figure, we plot the model-implied state-level worker flow matrix series for $k = 2, 3$ against that in the data. Blue circles represent the results of the heterogeneous-worker model and orange circles represent those of the canonical model, which exactly matches the 1-month worker flow matrix. Due to the short time horizon, worker flow matrices have diagonal elements close to one and off-diagonal elements close to zero. We separately plot them in the left and right panels.

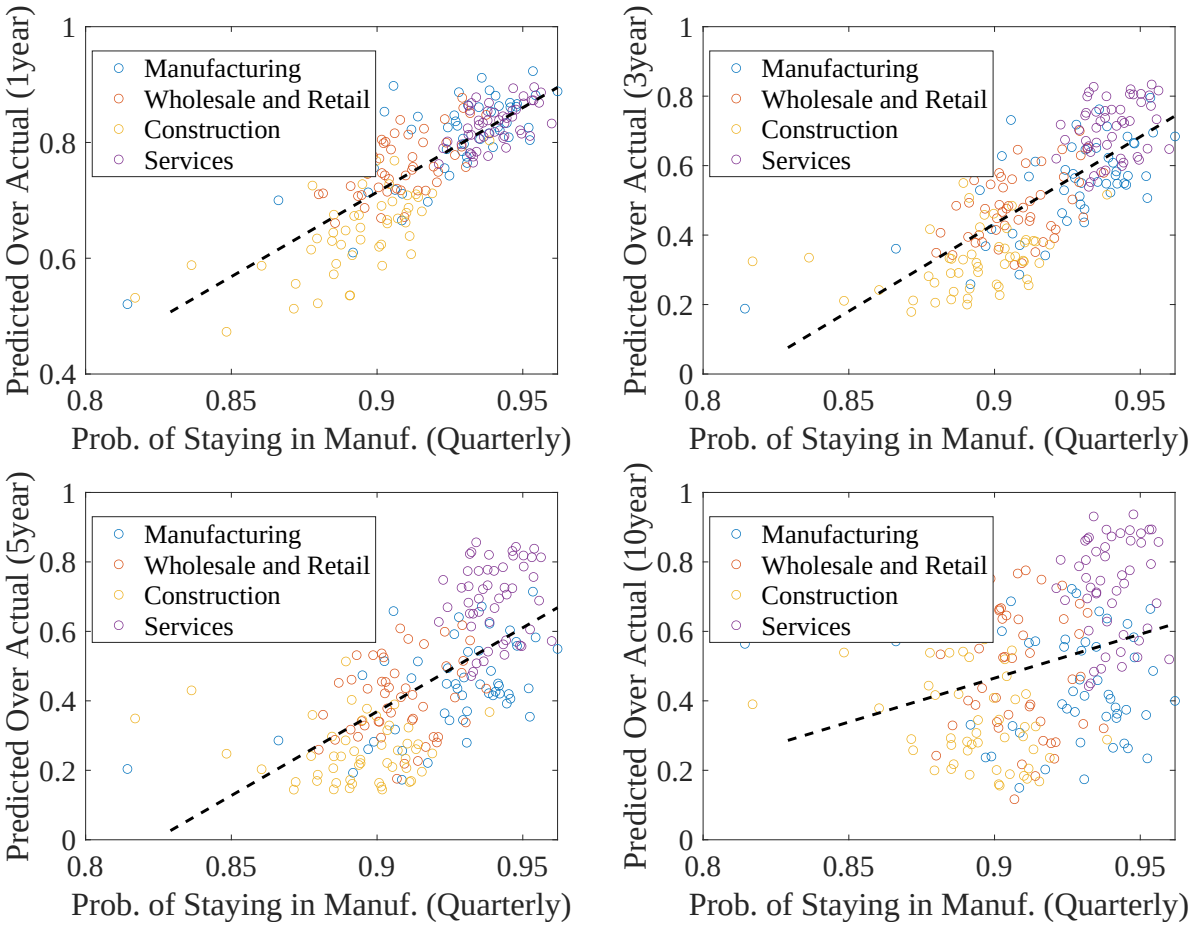
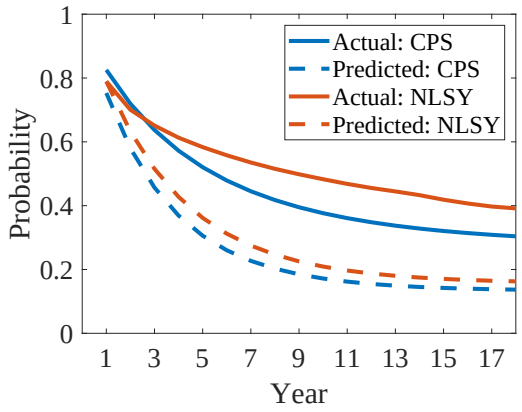
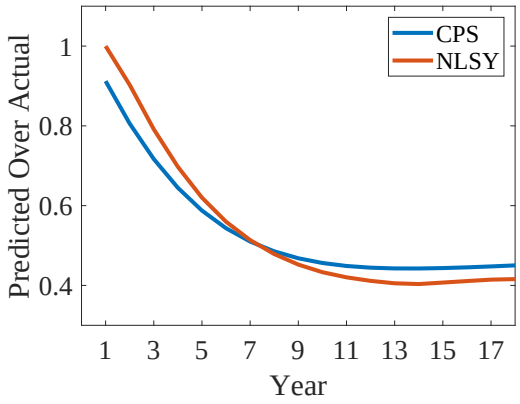


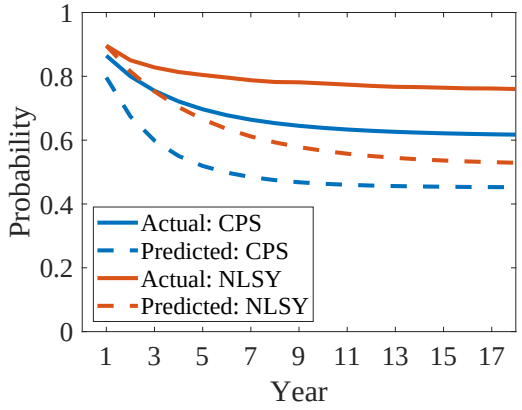
Figure A.19. Predicted and Actual Staying Probabilities: State-Level



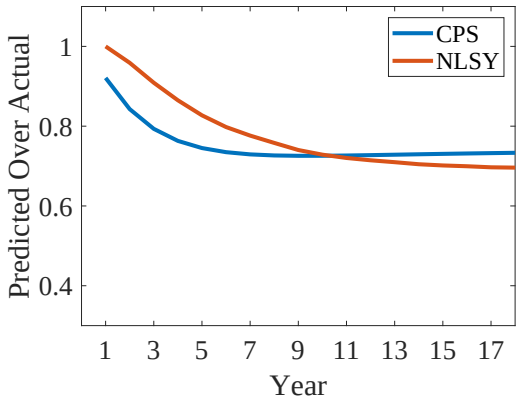
(a) Manufacturing: Actual and Predicted Probabilities



(b) Manufacturing: Ratio between Actual and Predicted



(c) Services: Actual and Predicted Probabilities



(d) Services: Ratio between Actual and Predicted

Figure A.20. Comparison: CPS and NLSY

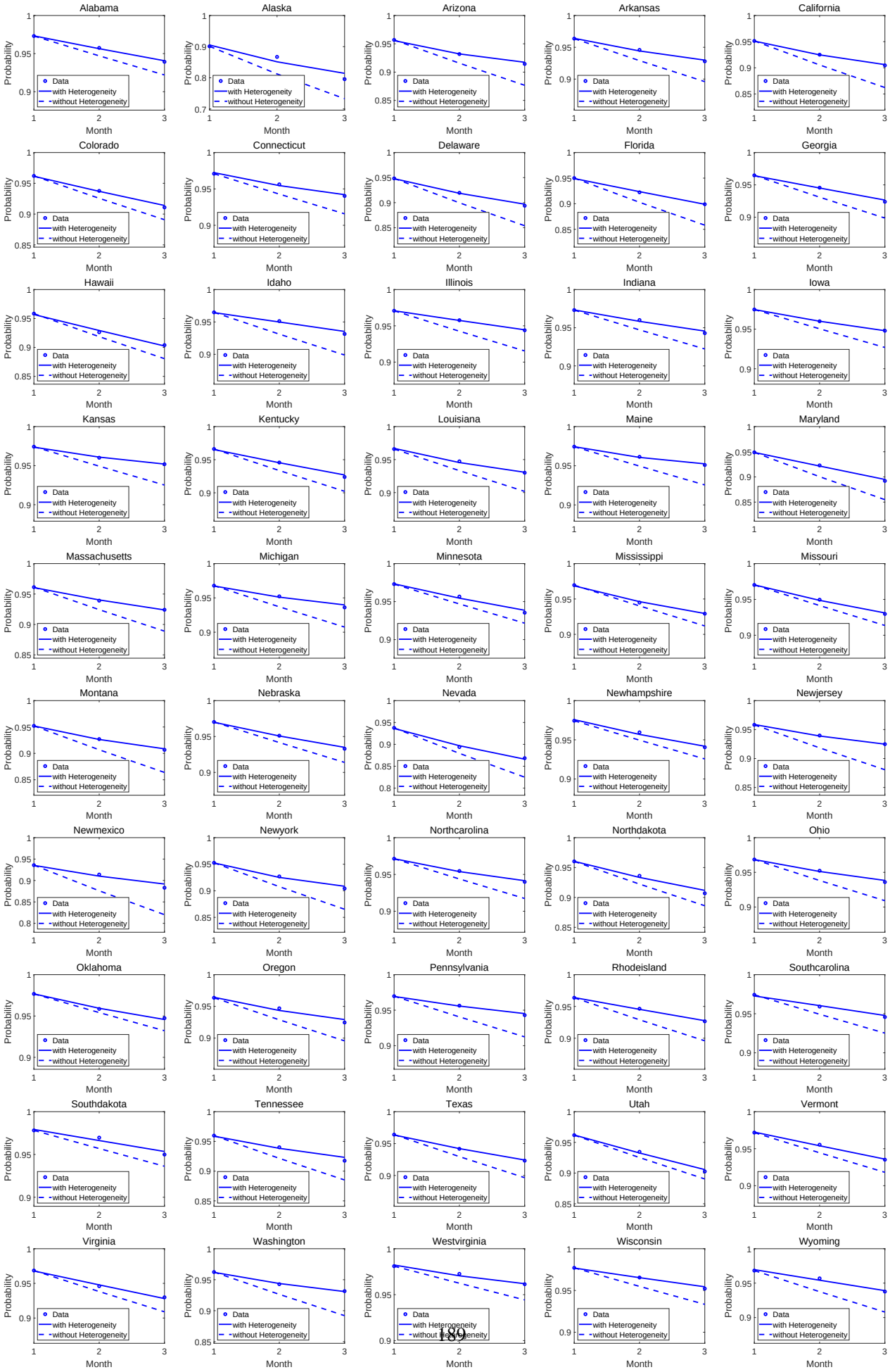
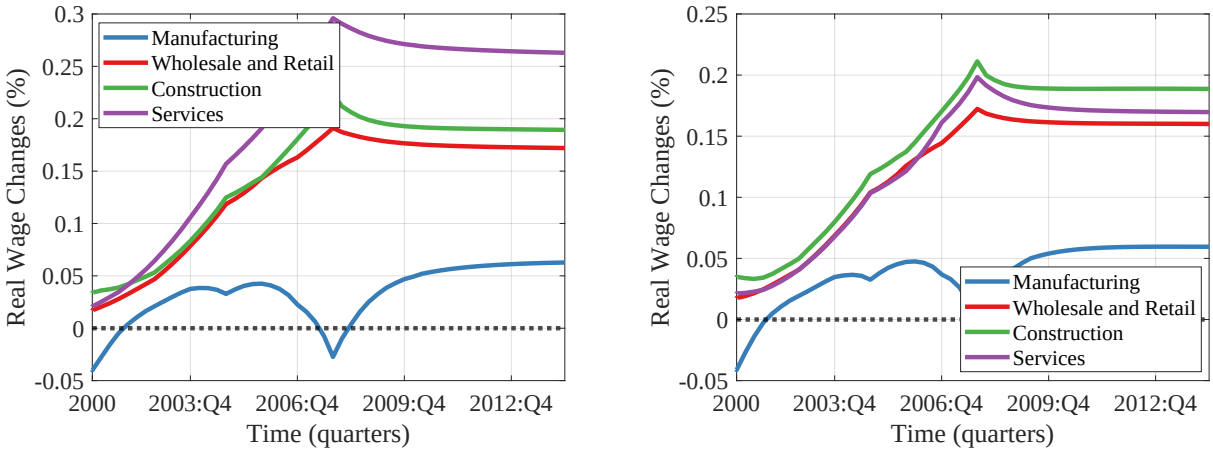


Figure A.21. Model Fit to State-Level Worker Flow Matrix Series: Manufacturing Sector



(a) With Worker Heterogeneity

(b) Without Worker Heterogeneity

Figure A.22. Changes in Real Wages

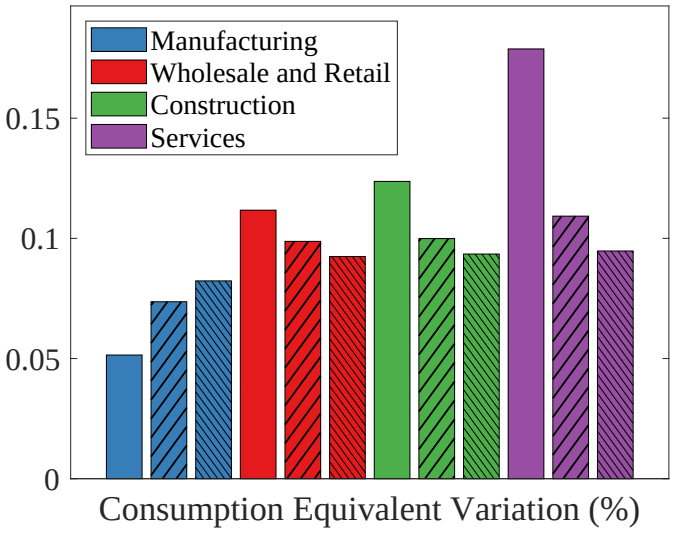


Figure A.23. Exogenous Wage

Appendix B

What Causes Agglomeration of Services? Theory and Evidence from Seoul

B.1 Empirical Analysis

B.1.1 Online Travel Survey

We present additional results from the online survey. Although a few recent studies document travel patterns using credit card transaction data or smartphone data, the online survey provides a more in-depth understanding of services travel, particularly how consumers combine multiple purchases into a single trip when faced with spatial frictions. These findings allow us to assess whether our assumptions on services travel in Section 2.3.1 are realistic.

Travel Distances. In Figure B.1, we plot the distribution of travel distances for commuting, services travel for the first purchase, and services travel for trip chaining. First, distances are comparable to Figure 3a, which reassures the quality of the online survey. Any slight differences between the two datasets might be attributed to variation in the survey year or sampling. Second, consumers tend to visit nearby regions for the second purchase. This observation implies that positive spillovers from trip chaining would spatially decay fast. It is worth noting that this observed pattern aligns with the recursive structure of our structural model. If consumers tend to travel to concentrated areas for their first purchase, they are more likely to stay in those areas for subsequent purchases, resulting in shorter travel distances.

Trip Chaining. The magnitude of trip chaining, represented by the average number of purchases made during a single instance of travel, is found to be 1.72 in the online survey. We find that the average number of purchases is comparable among travel with the different types of origin and destination (home, school/workplace, others), ranging

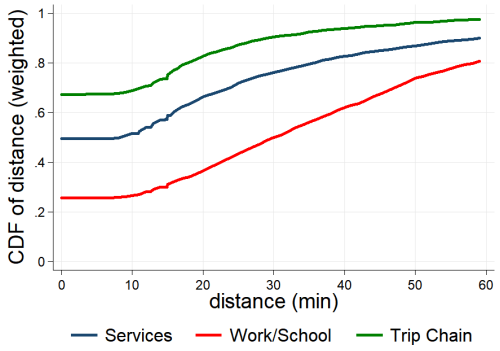


Figure B.1. CDF of Travel Distance (Online Survey)

Table B.1: Sector Choices of the First and Second Purchases

1st visit	2nd visit		
	Food	Retail	Other
Food	52.5% (52.9%)	42.2% (41.4%)	5.3% (5.7%)
Retail	34.7% (52.9%)	60.7% (41.4%)	4.6% (5.7%)
Other	48.4% (52.9%)	25.8% (41.4%)	25.8% (5.7%)

Notes: We compute an average percentage of visits to each sector in parentheses, which has to be a probability if choices of sectors are completely *i.i.d.*.

between 1.83 to 2.73. We also find that the number of purchases by the first sectors are comparable: 2, 1.8, and 1.9 for Food, Retail, and Other, respectively.[1] These results suggest that it is reasonable to model the trip chaining parameter as common for all types of travel.

Sector Choices of Purchases. We find that sectors of subsequent purchases do not critically depend on the sector of the previous purchase. In Table B.1, we calculate the conditional probability of choosing each sector for the second purchase, given the sector of the first purchase among travel with at least two purchases. We aim to examine whether there are systematic differences in these conditional probabilities across the three different sectors of the first purchase. To this end, we also compute the average (unconditional) probability of visits of each sector in parentheses. We find that the conditional probabilities for the second visit do not differ significantly from the unconditional probabilities in the parentheses. These findings provide reassurance that assuming random sector choices across purchases would not create bias in understanding the impacts of positive demand spillovers from trip chaining.

[1] In our online survey, we ask respondents to report on the travel that resulted in the most purchases in a single day in order to maximize the number of responses with multiple purchases. As a result, the average number of purchases with travel details is higher than the overall average.

Recursive Structure and Travel Distances. We model services travel in a recursive manner, which has several advantages as outlined in Section 2.3.1. An alternative approach would be to assume that consumers plan their entire itinerary and optimize their travel routes accordingly. This approach may result in different location choices for services stores, as consumers may prefer to visit stores that are located along their routes in order to minimize travel disutility.

We test whether our model-implied travel distance is significantly different from the distance that is minimized over the entire route. For the latter, we calculate the sum of distances between each pair of zones, an origin j_o, 1st purchase j_1, 2nd purchase j_2 if applicable, and a destination j_d.[2] In contrast, the model implied distances include only distances between j_o, j_1, and j_2, excluding j_d.

For two distances to be identical, two conditions must be satisfied. First, consumers must return to their origin zone ($j_o = j_d$). In our data, we find that 70.5% of consumers return to the same location. Second, they must make all purchases in a single zone. A total of 62.4% of trips satisfy both of these conditions. It is worth noting that this number represents a lower bound, as we only asked about travel with the maximum number of purchases for a given day. Additionally, even if consumers made purchases in two different zones, if the second location is not on the route back to their home, the model-implied distance may not be significantly different from the optimal distance.

B.1.2 Robustness of Shift-Share Design

In this section, we check the credibility of our shift-share design. Our instruments exploit differential city-wide trends across subsectors. For example, the top 3 growing subsectors are Japanese restaurantes, Cafe, and stree foods. The bottom 3 subsectors are computers, clothing, and bars. We first confirm that our instruments have enough variation in the Figure B.2. Below, we perform a few diagnostic tests that are suggested by Goldsmith-Pinkham, Sorkin, and Swift (2020).

Correlates of the Instruments. We first examine the correlation between our instruments and the characteristics of each zone, specifically rents, population density, and average income in 2015. Table B.2 confirms that the instruments do not exhibit significant explanatory power for these variables across regions.

Although insignificant, the instruments show a positive correlation with income. This correlation might arise if regions with higher average incomes initially have a higher proportion of subsectors that become popular in the following years. For instance, this scenario is plausible if wealthier individuals lead the popularity trend. Our identification strategy may be threatened if income in 2015 is correlated with regional shocks during the sample period. This could occur if wealthy individuals are more likely to live in regions that are likely to grow, or if

[2] Because we only ask about locations up to the second purchase, we ignore any additional distances that may have resulted from additional purchases.

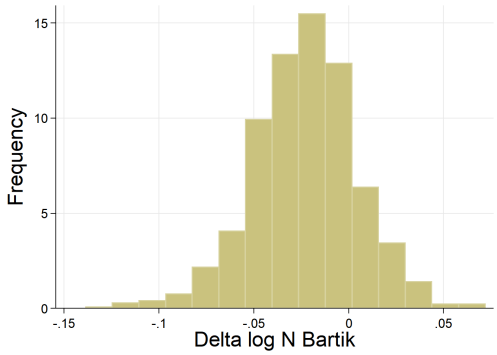


Figure B.2. Histogram of $\log N_{js}^{\text{Bartik}}$

Table B.2: Correlations with the Instruments

	$\log(\text{rent})_j$	$\log(\text{pop density})_j$	$\log(\text{income})_j$
$\Delta \log N_{js}^{\text{Bartik}}$	0.309	-1.084	0.274
	(0.316)	(0.775)	(0.172)
Observations	1,122	1,122	1,046

Notes: We always control for sector and district fixed effects. Robust standard errors are shown in parentheses, with * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

they have a role in driving regional growth. Although we do not find a significant correlation, such a correlation could potentially introduce an upward bias in our estimates, and thus we still control for them in our analysis as a precautionary measure.

Pre-trends. In order to investigate whether there are any pre-trends in the change in the number of stores, we cannot use our main dataset, the Seoul Commercial dataset, as it only covers years from 2014 onwards. Instead, we turn to the Seoul Business Survey, a publicly available administrative dataset spanning from 2006. This annual dataset includes information on the number of businesses in various sectors, including agriculture, manufacturing, and others, in each zone. However, the classification in this dataset is relatively broad, comprising only 19 sectors for the entire economy. Thus, we narrow our focus to four sectors that are relevant to consumption services: wholesale and retail trade, accommodation and food services activities, real estate activities and renting and leasing, and arts, sports and recreation related services.

Using this dataset, we calculate the pre-trends in the growth rates of the number of stores between 2009 and 2013. Since the classification in this dataset does not match our main analysis, preventing us from grouping them into the three sectors we use in our main analysis, we instead concentrate on variation at the zone level rather than at the zone-sector level. We compute the changes in the number of stores in the four sectors listed above and construct

Table B.3: Pre-Trends: Number of Stores

	pre-trend (2009–2013)		trend (2015–2019)	
Instruments constructed from Seoul Commercial Dataset				
$\Delta \log N_j^{\text{Bartik}}$	-0.084	-0.145	0.192	0.296
	(0.230)	(0.283)	(0.177)	(0.208)
District FE		✓		✓
Instruments constructed from Seoul Business Survey				
$\Delta \log N_j^{\text{Bartik}}$	-0.264	-0.105	0.296**	0.285**
	(0.251)	(0.262)	(0.123)	(0.132)
District FE		✓		✓

Notes: $N = 365$. Data sourc: Seoul Commercial Dataset and Seoul Business Survey. Robust standard errors are shown in parentheses, with * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

zone-level instruments as follows: $\Delta \log N_j^{\text{IV}} = \sum_{s,d} s_{jsd,0} \Delta \log N_{\text{Seoul},sd}$, where the formula remains the same as in our main analysis, but we aggregate across different sectors.

In the upper panel of Table B.3, we examine the relationship between our instruments, constructed from the Seoul Commercial dataset, and the pre-trends (trends, respectively) of the number of stores computed from the Seoul Business Survey. The left two columns show no significant association between the pre-period regional changes in the number of stores and our instruments. Coefficients are almost zero and they are not significant. In contrast, the trends during our sample period computed using the Seoul Business Survey are positively correlated with our instruments, as shown in third and forth columns. This reassures that instruments can successfully predict the growth of services stores. However, they are less statistically significant compared to the first stage in Table 1 due to the difference in data sources.

To further confirm the credibility of our instruments, we repeat the analysis using the instruments constructed from Seoul Business Survey. We use the same definition of instruments, but we interact share of the number of business in 2014 instead of revenues shares, and sum over the four different sectors mentioned earlier. The results, shown in the bottom panel of Table B.3, indicate that these new instruments are not correlated with the pre-trends, but are positively correlated and statistically significant when considering the trends.

In conclusion, we find that our results are not influenced by trends in the changes in the number of stores across regions.

B.1.3 Spillovers Within-Sector

In addition to cross-sector spillovers, we investigate spillovers *within* sector which can differ from across-sector spillovers due to competition forces. For example, we are interested in whether an exogenous increase in the number of Korean restaurants increases or decreases the number of other food stores. In this example, demand will substitute toward Korean restaurants, leading to smaller or even negative spillovers on net. We run the following regression:[3]

$$\Delta \log N_{jsd} = \alpha_2 + \beta_2 \Delta \log N_{js,-d} + \boldsymbol{X}'_{jsd}\gamma_2 + u_{jsd}$$

where $\Delta \log N_{jsd}$ is the same as before, and $\Delta \log N_{js,-d}$ is the growth rate of the number of stores of zone j and sector s excluding subsector d, which is defined by

$$\Delta \log N_{js,-d} = \sum_{d' \neq d} s_{jsd'} \Delta \log N_{jsd'},$$

where $s_{jsd'}$ is the revenue share of subsector d' in sector s, excluding subsector d for year $t = 2015$, and $\boldsymbol{X}'_{jsd}$ is the covariate. This specification has the same endogeneity concerns as before, so we instrument $\Delta \log N_{js,-d}$ with a Bartik instrument

$$\Delta \log N^{\text{Bartik}}_{js,-d} = \sum_{d' \neq d} s_{jsd',0} \Delta \log N_{\text{Seoul},sd'}$$

where $s_{jsd',0}$ is the revenue share in year $t_0 = 2014$.

Columns (1)–(4) report the results for the within-sector specifications. For all four columns, we control for sector s fixed effects and the subsector composition of the other two sectors, s' and s''. As in cross-sector specifications, we control for subsector-specific trends. Column (1) is the result of the ordinary least squares. We find that the number of stores in a given subsector increases when there is an exogenous increase in the number of stores in other subsectors within the same sector. However, this result may be biased upward. Moving to Column (2), where we use Bartik instruments to address endogeneity concerns, we obtain the opposite result. In Columns (3) and (4), we include subsector fixed effects, district fixed effects, and the same set of controls as in Columns (3) and (4) in Table 1. Focusing on our main specification in Column (3), the number of stores in a given subsector decreases by 4.6% when there is a 10% exogenous increase in the number of stores in other subsectors within the same sector. Again, in the last column, we use the specification guided by the theory as derived in Appendix B.3.2, and it yields similar results.

[3] One might view this as an example of peer effect regression of Manski (1993). However, we can interpret the result as a causal one for two reasons. First, we use leave-one-out weighted average $\Delta \log N_{js,-d}$ and instrument it with shift-share instruments. Second, we can get similar estimates when we make arbitrary division between subjects and peers (e.g., estimate the effect of Korean restaurants on Japanese restaurants, not *vice versa*).

Table B.4: Number of Stores Results: Within-Sector

dependent variable: $\Delta \log N_{jsd}$

	(1)	(2)	(3)	(4)
	OLS	IV	IV	IV
$\Delta \log N_{js,-d}$	0.188***	-0.434*	-0.456*	-0.416*
	(0.042)	(0.252)	(0.257)	(0.234)
Sector FE, subsector trend	✓	✓		
Subsector, district FE			✓	✓
Additional controls			✓	✓
	FIRST STAGE ESTIMATES			
$\Delta \log N^{\text{Bartik}}_{js,-d}$		0.657***	0.672***	.
		(0.130)	(0.140)	.
First-stage F stat		25.62	23.06	.
Observations	9380	9379	8827	8827

Notes: Equation estimates based on Seoul Commercial area data for 2014, 2015 and 2019. More details are explained in the notes of Table 1.

B.2 Derivations in Section 2.3.1

Expenditure Equalization and Recursive Formulation. Taking the first-order condition of the maximization problem (4) with respect to $q_t(\sigma^t)$, we have $\beta^t \frac{1}{q_t(\sigma^t)} \pi(\sigma^t) = \lambda \beta^t p_{j_t(\sigma^t)s(\sigma^t)} \pi(\sigma^t)$ where λ is the Lagrange multiplier associated with the budget constraint (5). This immediately shows that the optimal expenditure is equalized across all regions and purchases, $e_t(\sigma^t) \equiv q_t(\sigma^t) \cdot p_{j_t(\sigma^t)s(\sigma^t)} = \lambda^{-1}$. This expenditure equalization implies that changing the region for purchase t, $j_t(\sigma^t)$, does not affect the budget constraint (5). This means that the consumer always chooses the region j that maximizes the sum of instantaneous and continuation utilities. Thus, the maximization problem (4) can be recursively expressed as equation (6). Standard extreme-value algebra simplifies this further to equation (7).

Linear Consumption Index. Combining equations (7) and (8), we have

$$C(i,e)^{1/(1-\beta)} = e \cdot \prod_s \left(\sum_j \left(p_{js}^{-1/\nu} \cdot \exp(-\tilde{d}(i,j))^{1/\nu} \cdot C(j,e)^{\beta/(\nu(1-\beta))} \right) \right)^{\alpha_s \nu}.$$

Thus, the consumption index is linear in her spending as claimed, and we can define the price index of services travel from zone i as $P_i \equiv \frac{1}{C(i,1)}$.

Total Consumer Spending. Note first that if consumers in zone i spend E_i on services travel, their per-purchase expenditure is given by $(1-\beta)E_i$. In their first purchases, they choose sector s with probability α_s and region j with probability π^s_{ij}. Thus, the spending on (j,s) is

$$\sum_{i\in\mathcal{J}}(1-\beta)\alpha_s\pi^s_{ij}E_i \equiv (1-\beta)\alpha_s \boldsymbol{E}^{\mathsf{T}}\boldsymbol{\pi}^s_j$$

where $\boldsymbol{E} = (E_1,\cdots,E_J)^{\mathsf{T}}$ and $\boldsymbol{\pi}^s_j = (\pi^s_{1j},\cdots,\pi^s_{Jj})^{\mathsf{T}}$. Likewise, in their second purchases, the spending on (j,s) is given by

$$\sum_{i\in\mathcal{J}}\sum_{i'\in\mathcal{J}}(1-\beta)\beta\alpha_s E_i\pi_{ii'}\pi^s_{i'j} \equiv (1-\beta)\beta\alpha_s \boldsymbol{E}^{\mathsf{T}}\Pi\boldsymbol{\pi}^s_j$$

where $\pi_{ii'} = \sum_s \alpha_s\pi^s_{ii'}$ and Π is a $J\times J$ matrix with (i,i')-element $\pi_{ii'}$. In a similar manner, the total spending on (j,s), which is the sum of the spending from all purchases can be computed as in equation (10).

Demand Function. Given the aggregate demand on (j,s), R_{js}, demand for an individual store ω in (j,s,d) is given by

$$q_{jsd}(\omega) = \left(\frac{p_{jsd}(\omega)}{p_{jsd}}\right)^{-\rho}\cdot\phi_{jsd}\cdot\left(\frac{p_{jsd}}{p_{js}}\right)^{-\sigma}\cdot\frac{R_{js}}{p_{js}}.$$

This isoelastic demand function implies constant markups, $p_{jsd}(\omega) = \frac{\rho}{\rho-1}\frac{c_{jsd}}{A_{jsd}}$, where the unit cost c_{jsd} is given by the solution of the following cost minimization problem subject to

$$c_{jsd} = \min_{H_{jsd}(\omega),L_{jsd}(\omega)} r_jH_{jsd}(\omega) + w_jL_{jsd}(\omega) \text{ s.t. } H_{jsd}(\omega)^\gamma L_{jsd}(\omega)^{1-\gamma} \le 1 = \left(\frac{r_j}{\gamma}\right)^\gamma\left(\frac{w_j}{1-\gamma}\right)^{1-\gamma}.$$

B.3 Justification of IV Specification: Proofs

B.3.1 First-Order Approximation (Section 2.3.2)

In this section, we take the first-order approximations to the equilibrium conditions in Section 2.3.2, which we summarize in the following lemma for convenience. For simplicity, we first start with the case without external economies of scale, assuming that A_{jsd} and C_{jsd} are exogenous variables. At the end of the section, we return to the case with external economies of scale.[4]

Lemma B.1. *Below equations summarize the equilibrium conditions in Section 2.3.2.*

$$N_{jsd}^{1-\frac{1-\sigma}{1-\rho}} = C_{jsd}^{-1}\tilde{A}_{jsd}^{-(1-\sigma)}c_{jsd}^{1-\sigma}p_{js}^{-(1-\sigma)}R_{js} \tag{B.1}$$

[4] Detailed derivations and proofs are omitted and are available upon request.

$$p_{js} = \left(\sum_d N_{jsd}^{\frac{1-\sigma}{1-\rho}} \tilde{A}_{jsd}^{-(1-\sigma)} \left(\frac{\rho}{\rho - 1} \right)^{1-\sigma} c_{jsd}^{1-\sigma} \right)^{1/(1-\sigma)} \tag{B.2}$$

$$V(i) \equiv \exp(\mathbb{E}\, v(i,1)) = \prod_s \left(\sum_j p_{js}^{-1/\nu} \exp(-\tilde{\tau} d(i,j) - \tilde{\varphi} \mathbb{1}_{i \neq j}) \cdot V(j)^{\tilde{\beta}} \right)^{\alpha_s \nu} \tag{B.3}$$

$$R_{js} = (1-\beta)\alpha_s \boldsymbol{E}^{\mathsf{T}} (I - \beta \Pi)^{-1} \boldsymbol{\pi}_j^s = \alpha_s \left((1-\beta) \sum_i E_i \pi_{ij}^s + \beta \sum_i \sum_k E_i \pi_{ik} \pi_{kj}^s \right) + o(\beta) \tag{B.4}$$

$$\pi_{ij}^s = \frac{p_{js}^{-1/\nu} \exp(-\tilde{\tau} d(i,j) - \tilde{\varphi} \mathbb{1}_{i \neq j}) \cdot V(j)^{\tilde{\beta}}}{\sum_{j'} p_{j's}^{-1/\nu} \exp(-\tilde{\tau} d(i,j') - \tilde{\varphi} \mathbb{1}_{i \neq j'}) \cdot V(j')^{\tilde{\beta}}}. \tag{B.5}$$

where $\pi_{ij} = \sum_s \alpha_s \pi_{ij}^s$, $\boldsymbol{\Pi} = (\pi_{ij})$, *and* $\boldsymbol{\pi}_j^s = (\pi_{1j}^s, \cdots, \pi_{Jj}^s)^{\mathsf{T}}$.

Assumption B.1 (Parameter Restriction). *We assume* $1 + \frac{1}{\nu}, \sigma \in (1, \rho)$.

We first log linearize the equilibrium conditions, (B.1)–(B.5). Then, we view these linearized equations as exact equilibrium conditions and take the first-order approximation up to o($\boldsymbol{x}$) where $\boldsymbol{x} = (\beta, (\pi_{ik}^s)_{i \neq k, s})'$, i.e., we assume that β and π_{ij}^s are small for $i \neq j$ and ignore second-order terms.[5] To simplify notation, we write log deviations as $n_{jsd} = \mathrm{d} \log N_{jsd}$, $\hat{C}_{jsd} = \mathrm{d} \log C_{jsd}$, $\tilde{a}_{jsd} = \mathrm{d} \log \tilde{A}_{jsd}$, $\hat{c}_{jsd} = \mathrm{d} \log c_{jsd}$, $\hat{p}_{js} = \mathrm{d} \log p_{js}$, $r_{js} = \mathrm{d} \log R_{js}$, $\hat{\pi}_{ij}^s = \mathrm{d} \log \pi_{ij}^s$, and $e_i = \mathrm{d} \log E_i$. We also define sector-level variables $x_{js} \equiv \sum_d \theta_{jsd} x_{jsd}$ for $x \in \{n, \hat{C}, \tilde{a}, \hat{c}\}$ where the weight θ_{jsd} is the revenue share of d in (j,s) given by

$$\theta_{jsd} \equiv \frac{N_{jsd}^{\frac{1-\sigma}{1-\rho}} \tilde{A}_{jsd}^{-(1-\sigma)} c_{jsd}^{1-\sigma}}{\sum_{d'} N_{jsd'}^{\frac{1-\sigma}{1-\rho}} \tilde{A}_{jsd'}^{-(1-\sigma)} c_{jsd'}^{1-\sigma}} = \frac{R_{jsd}}{R_{js}}.$$

This choice of weights is consistent with our specifications in Section 2.2.2. Finally, we define two more variables $\check{a}_{js} \equiv \tilde{a}_{js} - \hat{c}_{js}$, which captures the combined effect of changes in productivity and input costs, and $\check{a}_{js}^* = \check{a}_{js} - \frac{1}{\rho - 1} \hat{C}_{js}$, which additionally captures the effect of changes in operating costs. Equilibrium in terms of log deviations is characterized in the following lemma.

Lemma B.2 (Log linearization). *When* $e_i = 0$ *for all* i, *we have*

$$\begin{aligned}
\hat{p}_{js} &= -\check{a}_{js}^* - \tfrac{1}{\rho - 1} r_{js} \\
r_{js} &= \sum_i (1-\beta) \lambda_{ij}^s \hat{\pi}_{ij}^s + \sum_i \sum_k \sum_{s'} \beta \lambda_{ikj}^{s's} (\hat{\pi}_{ik}^{s'} + \hat{\pi}_{kj}^s) \\
\hat{\pi}_{ij}^s &= \sum_{j'} (\mathbb{1}_{j'=j} - \pi_{ij'}^s) \left(-\tfrac{1}{\nu} \hat{p}_{j's} - \tilde{\beta} \sum_{s'} \alpha_{s'} \sum_{j''} \pi_{j'j''}^{s'} \hat{p}_{j''s'} \right) + o(\boldsymbol{x}) \\
n_{jsd} &= \tfrac{\rho - 1}{\rho - \sigma} \left(-\hat{C}_{jsd} + (\sigma - 1) \check{a}_{jsd} + (\sigma - 1) \hat{p}_{js} + r_{js} \right)
\end{aligned}$$

[5] To be precise, we consider a sequence of models indexed by J (e.g., the number of regions), where, as $J \to \infty$, β and π_{ij}^s converges to zero. We write $A = B + o(\boldsymbol{x}^k)$ hereafter when $\lim_{J \to \infty} \frac{A_J - B_J}{\|\boldsymbol{x}_J^k\|} = 0$ for $k \in \mathbb{N}_0$ and write $A = B + O(\boldsymbol{x}^k)$ when $\limsup_{J \to \infty} \frac{A_J - B_J}{\|\boldsymbol{x}_J^k\|} < \infty$.

where $\lambda_{ij}^{s} = \dfrac{E_i \pi_{ij}^{s}}{(1-\beta)\sum_i E_i \pi_{ij}^{s} + \beta \sum_{iks'} E_i \alpha_{s'} \pi_{ik}^{s'} \pi_{kj}^{s}}$ *and* $\lambda_{ikj}^{s's} = \dfrac{E_i \alpha_{s'} \pi_{ik}^{s'} \pi_{kj}^{s}}{(1-\beta)\sum_i E_i \pi_{ij}^{s} + \beta \sum_{iks'} E_i \alpha_{s'} \pi_{ik}^{s'} \pi_{kj}^{s}}$.

The resulting equilibrium conditions are linear but still not tractable because of general equilibrium feedback between zones. To proceed, we focus on shocks to a given zone j_0 that satisfies the following small open zone assumption:

Assumption B.2 (Small Open Zone Assumption). *Write $R_{i\to js}$ to denote the revenue of (j,s) coming from consumers who start their services travel from region i:*

$$R_{i\to js} \equiv (1-\beta)\alpha_s \Big(0, \cdots, 0, \overbrace{E_i}^{i\text{th}}, 0, \cdots, 0\Big)(I - \beta\Pi)^{-1}\boldsymbol{\pi}_j^{s}.$$

We assume that zone j_0 is small *in the sense that*

$$\pi_{jj_0}^{s}, \frac{R_{j_0\to js}}{R_{js}} = o(\boldsymbol{x}) \text{ for } j \neq j_0 \text{ and } \sum_{j\neq j_0} \pi_{jj_0}^{s} = O(1).$$

Roughly speaking, this assumption requires that the share of zone j_0 as both a travel destination and a revenue source is small enough to ignore any complicated general equilibrium feedback from other zones to zone j_0. As a result, we can solve for the changes in the number of stores in terms of exogenous shocks. Under this assumption, we can simplify the equilibrium conditions in Lemma B.2.

Proposition B.1. *Under Assumption B.2, equilibrium is characterized by (up to $o(\boldsymbol{x})$)*

$$\begin{aligned}
\hat{p}_{js}, r_{js} &= 0 \quad \text{for } j \neq j_0 \\
\hat{p}_{j_0 s} &= -\check{a}_{j_0 s}^{*} - \tfrac{1}{\rho - 1} r_{j_0 s} \\
r_{j_0 s} &= \Psi_{j_0 s} \cdot \hat{q}_{j_0, s} + \sum_{s'} \Phi_{j_0 s}^{s'} \cdot \hat{q}_{j_0 s'} \\
\hat{\pi}_{ij}^{s} &= \begin{cases} 0 & \text{if } i, j \neq j_0 \\ -\pi_{j_0 j_0}^{s} \hat{q}_{j_0 s} & \text{if } i = j_0, j \neq j_0 \\ \hat{q}_{j_0 s} & \text{if } i \neq j_0, j = j_0 \\ (1 - \pi_{j_0 j_0}^{s}) \hat{q}_{j_0 s} & \text{if } i = j = j_0 \end{cases}
\end{aligned}$$

where $\hat{q}_{j_0 s} = -\frac{1}{\nu}\hat{p}_{j_0 s} - \tilde{\beta} \sum_{s'} \alpha_{s'} \pi_{j_0 j_0}^{s'} \hat{p}_{j_0 s'}$ *and*

$$\begin{aligned}
\Psi_{j_0 s} &= \left(1 - \left((1-\beta)\lambda_{j_0 j_0}^{s} + \sum_{is'} \beta \lambda_{ij_0 j_0}^{s's}\right)\pi_{j_0 j_0}^{s}\right) \\
\Phi_{j_0 s}^{s'} &= \left(\sum_i \beta \lambda_{ij_0 j_0}^{s's}(1 - \pi_{j_0 j_0}^{s} \cdot \mathbb{1}_{i=j_0})\right).
\end{aligned}$$

For future reference, note that $\Psi_{j_0 s}, \Phi_{j_0 s}^{s'} \in (0,1)$ *and that* $\Phi_{j_0 s}^{s'} = o(1)$.

The following corollary characterizes the responses of other endogenous variables. When there are favorable shocks to zone j_0, it experiences an increase in the number of stores, an increase in revenue, a decrease in the price index, and increases in travel inflows. In contrast, the effects on other regions are ambiguous.[6]

Corollary B.1. *When there are favorable shocks in j_0 ($\tilde{a}_{j_0 s} > 0, c_{j_0 s} < 0$, and $\hat{C}_{j_0 s} < 0$ for all s), we have*

$$\hat{p}_{j_0 s} < 0, r_{j_0 s} > 0, \hat{\pi}^s_{j_0 j_0} > 0, \hat{\pi}^s_{i j_0} > 0, \hat{\pi}^s_{j_0 j} < 0, \text{ and } n_{j_0 s} > 0 \tag{B.6}$$

for $i, j \neq j_0$.

Proposition B.1 gives us a simultaneous equation system that determines the equilibrium values of endogenous variables. We can further solve for endogenous variables in terms of exogenous shocks.

Proposition B.2. *Under Assumption B.2 and up to $o(\boldsymbol{x})$, we have*

$$n_{j_0 s} = \gamma^1_{j_0 s} \check{a}^*_{j_0 s} - \hat{C}_{j_0 s} + \sum_{s' \neq s} \tau^{2s'}_{j_0 s} \check{a}^*_{j_0 s'} \tag{B.7}$$

$$n_{j_0 s d} = \gamma^1_{j_0 s} \check{a}^*_{j_0 s} - \hat{C}_{j_0 s d} + \kappa(\check{a}^*_{j_0 s d} - \check{a}^*_{j_0 s}) + \sum_{s' \neq s} \tau^{2s'}_{j_0 s} \check{a}^*_{j_0 s'} \tag{B.8}$$

where
$$\kappa = \frac{(\sigma-1)(\rho-1)}{(\rho-\sigma)}, \quad \gamma^1_{j_0 s} = B_{j_0 s} \frac{1}{1 - \frac{\Theta_{j_0 s}}{\rho - 1}}, \quad \text{and} \quad \tau^{2s'}_{j_0 s} = \Lambda^{s'}_{j_0 s} \frac{1}{1 - \frac{\Theta_{j_0 s}}{\rho-1}} \frac{1}{1 - \frac{\Theta_{j_0 s'}}{\rho-1}}.$$

The shocks $\{\check{a}^*_{j_0 s d}\}$ and $\{\hat{C}_{j_0 s d}\}$ are the composite effect of preference, technology, input costs, and the operating cost. Moreover, all coefficients are positive, and $\tau^{2s'}_{j_0 s}$ vanishes as $\beta \to 0$.

Note that the effect of $(j, s, -d)$ on (j, s, d) is determined by the sign of $\gamma^1_{j_0 s} - \kappa$, which is in principle ambiguous and can be characterized by the following proposition. This shows that a shock to a subsector within the same sector can have negative spillovers when the competition forces dominate the positive effect.

Corollary B.2. *When there are favorable shocks on subsectors $d' \neq d$ in a region-sector pair (j_0, s), subsector d experiences a decline in the number of stores (i.e., the competition effect is dominant) if and only if*

$$\nu(\sigma - 1) > (1 - \lambda^s_{j_0 j_0} \pi^s_{j_0 j_0}). \tag{B.9}$$

Finally, we return to the case with external economies of scale. Note that the constant markup pricing implies that Υ_{1j} is always proportional to Υ_{2j}, which in turn gives $\mathrm{d} \log \Upsilon_{1j} = \mathrm{d} \log \Upsilon_{2j}$. Thus, up to the first-order approximation,

[6] These effects are $o(\boldsymbol{x})$, so we need to take second-order approximations to determine their signs. To see why these effects are ambiguous, consider four regions, A, B, C, and D. Services travel is possible only from B to A; from C to B; and from C to D. In this case, when there are favorable shocks in region A. The number of stores in B decreases, so consumers in region C substitute toward region D. Thus, favorable shocks in region A decrease the number of stores in region B, while increasing the number of stores in region D.

we can write

$$a_{jsd} = \bar{a}_{jsd} + \varepsilon_a \cdot \mathrm{d}\log \Upsilon_{2j}$$
$$\hat{C}_{jsd} = \bar{\hat{C}}_{jsd} - \varepsilon_c \cdot \mathrm{d}\log \Upsilon_{2j}$$

for some $\varepsilon_a, \varepsilon_c \in \mathbb{R}$. We assume positive external economies of scale, assuming $\varepsilon_a, \varepsilon_c \geq 0$. The shifters $\bar{a}_{jsd}$ and $\bar{C}_{jsd}$ are assumed to be exogenous. The following proposition extends the results of Proposition B.2. In particular, an exogenous increase in the effective productivity of sector $s' \neq s$ has a positive effect on sector s through trip chaining and external economies of scale.

Proposition B.3. *Under Assumption B.2 and up to $o(\boldsymbol{x})$, we have*

$$n_{j_0 s} = \tilde{\gamma}^1_{j_0 s} \bar{\tilde{a}}^*_{j_0 s} + \check{\gamma}^1_{j_0 s} \bar{\hat{C}}_{j_0 s} + \sum_{s' \neq s} \tilde{\tau}^{2s'}_{j_0 s} \bar{\tilde{a}}^*_{j_0 s'} + \sum_{s' \neq s} \check{\tau}^{2s'}_{j_0 s} \bar{\hat{C}}_{j_0 s'} \tag{B.10}$$

$$n_{j_0 sd} = n_{j_0 s} - (\bar{\hat{C}}_{j_0 sd} - \bar{\hat{C}}_{j_0 s}) + \kappa(\bar{\tilde{a}}^*_{j_0 sd} - \bar{\tilde{a}}^*_{j_0 s}). \tag{B.11}$$

where all $\tilde{\gamma}$'s and $\tilde{\tau}$'s are positive, with the latter vanishing as $(\beta, \varepsilon_a, \varepsilon_c) \to (0, 0, 0)$.

B.3.2 Justification of IV Specification

In this section, we interpret the reduced-form estimates in Section 2.2.2 through the lens of our structural model and its first-order approximated equilibrium conditions. We again start with the case without external economies of scale and show how the results extend. We relax the assumption that there are shocks only to region j_0 and assume instead that

$$n_{js} = \gamma^1_{js} \check{a}^*_{js} - \hat{C}_{js} + \sum_{s' \neq s} \tau^{2s'}_{js} \check{a}^*_{js'}$$
$$n_{jsd} = n_{js} + \kappa(\check{a}^*_{jsd} - \check{a}^*_{js}) - (\hat{C}_{jsd} - \hat{C}_{js}),$$

or equivalently,

$$n_{js} = \gamma^1_{js} \tilde{a}_{js} - \gamma^1_{js} \hat{c}_{js} - \left(1 + \tfrac{\gamma^1_{js}}{\rho - 1}\right) \hat{C}_{js} + \sum_{s' \neq s} \tau^{2s'}_{js} \left(\tilde{a}_{js'} - \tfrac{1}{\rho - 1} \hat{C}_{js'} - \hat{c}_{js'}\right)$$
$$n_{jsd} = n_{js} + \kappa(\tilde{a}_{jsd} - \tilde{a}_{js}) - \tfrac{\kappa}{\rho - 1}(\hat{C}_{jsd} - \hat{C}_{js}) - \kappa(\hat{c}_{jsd} - \hat{c}_{js})$$

hold for all zones $j \in \mathcal{J}$. We implicitly ignore spatial linkages as in Goldsmith-Pinkham, Sorkin, and Swift (2020) or Borusyak, Hull, and Jaravel (2020).[7] In Section 2.4, we propose an estimation method that exploits exogenous

[7] When there are shocks also to other regions, the combined effect of them can be greater than $o(\boldsymbol{x})$.

variations from shift-share design, while taking into account spatial linkages. We start with a formal definition of Bartik instruments.

Definition B.1. *For some given weights* $\{\varphi_{js}\}_{j,s}$ *with* $\sum_j \varphi_{js} = 1$*, we can define* Bartik Instruments *as*

$$n_{js}^{Bartik} \equiv \sum_d \theta_{jsd,0} n_{sd} \quad \text{where } n_{sd} = \sum_j \varphi_{js} n_{jsd}.$$

For future reference, we also define quasi-Bartik instruments $\tilde{n}_{js}^{Bartik} \equiv \sum_d \theta_{jsd} n_{sd}$*, which uses* θ_{jsd} *instead of* $\theta_{jsd,0}$*.*

We first aggregate the changes in the number of stores, n_{jsd}, across regions using the weights $\{\varphi_{js}\}_j$ to calculate the citywide change in the number of stores for each subsector, n_{sd}.[8] We then interact these subsector-level changes with the subsector composition of (j, s) to obtain n_{js}^{Bartik}.

To show consistency of the IV estimators, we make three sets of assumptions. First, unit costs and operating costs may depend on regions and sectors, but not on subsectors. As a result, we can use Bartik instruments to isolate the effect of productivity shocks, $\tilde{a}$. Second, we make assumptions for the relevance condition and the exclusion restriction, similar to those in GSS. Finally, we impose symmetry across regions at the sector level and symmetry across sectors so as to restrict our attention to a single coefficient instead of region- and sector-specific coefficients. This final assumption is not strictly necessary but simplify propositions.[9] Formally, our assumptions are as follows.

Assumption B.3.

(i) (Costs) Unit costs and operating costs do not depend on d*:* $\hat{C}_{jsd} = \hat{C}_{js}$ *and* $\hat{c}_{jsd} = \hat{c}_{js}$ *for all* d*.*

(ii) (GSS) Assume the following probabilistic structure:

$$\tilde{a}_{jsd} = \tilde{a}_{sd} + \tilde{\varepsilon}_{jsd}$$
$$\hat{C}_{js} = \hat{C}_s + \hat{\varepsilon}_{js}$$

where $\tilde{a}_{sd} = \sum_j \varphi_{js} \tilde{a}_{jsd}$*. We assume that* $\{\tilde{\varepsilon}_j, \hat{\varepsilon}_j, \theta_j, \theta_{j0}\} \sim$ *i.i.d. across* j *and view* $\tilde{a}_{sd}$ *and* $\hat{C}_s$ *as fixed.*

(ii′) (GSS: Exogeneity) Assume $\theta_{jsd,0} \perp_j \begin{pmatrix} \tilde{\varepsilon}_{js'} & \hat{\varepsilon}_{js'} & \tilde{\varepsilon}_{js'd'} \end{pmatrix}' | c_j$*, for all* s, s', d, d' *where* $c_j = \begin{pmatrix} \hat{c}_{j1} & \hat{c}_{j2} & \hat{c}_{j3} & 1 \end{pmatrix}'$*.*[10]

(ii″) (Relevance) The 3×3 *matrix whose* (s, s')*-element is* $\sum_{d,d'} \tilde{a}_{sd} \tilde{a}_{s'd'} \mathbb{E}_j \left[\theta_{jsd}^{\perp} \theta_{js'd',0}^{\perp}\right]$ *is invertible.*[11]

(iii) (Homogeneous Effect) Initially all regions are symmetric at the sector level and sectors are symmetric so that the coefficients γ_{js}^1 *and* $\tau_{js}^{2s'}$ *are no longer* j*- or* s*-specific, and we write them as* γ *and* τ*, respectively.*

[8] A more natural choice of weights is $\varphi'_{jsd} = \frac{N_{jsd}}{\sum_{j'} N_{j'sd}}$ because $N_{sd} = \sum_j N_{jsd}$ implies $n_{sd} = \sum_j \varphi'_{jsd} n_{jsd}$. If the weights are subsector-dependent, however, n_{sd} not only captures the aggregate trend of sector d, but it is also contaminated by regional shocks. To see this, suppose $\varphi'_{j_1sd_1} > \varphi'_{j_2sd_1}$ and $\varphi'_{j_1sd_2} < \varphi'_{j_2sd_2}$. Then, if there is a positive regional shock to j_1, this would increase n_{sd_1} relative to n_{sd_2}.

[9] Without this assumption, propositions in this section should be written in terms of weighted averages across regions and across sectors, as in Section IV of GSS.

Under these assumptions, we first prove two useful representations that relate changes in the number of stores with quasi-Bartik instruments and error terms that are orthogonal to Bartik instruments.

Lemma B.3. *Under Assumption B.3, we can write*

$$n_{js} = \tfrac{\gamma}{\kappa}\tilde{n}_{js}^{\textit{Bartik}} + \sum_{s' \neq s} \tfrac{\tau}{\kappa}\tilde{n}_{js'}^{\textit{Bartik}} - \gamma\hat{c}_{js} - \sum_{s' \neq s} \tau\hat{c}_{js'} + \mathsf{FE}_s + \varepsilon_{js}$$

$$n_{jsd} = \tfrac{\gamma-\kappa}{\kappa}\tilde{n}_{js}^{\textit{Bartik}} + \sum_{s' \neq s} \tfrac{\tau}{\kappa}\tilde{n}_{js'}^{\textit{Bartik}} - \gamma\hat{c}_{js} - \sum_{s' \neq s} \tau\hat{c}_{js'} + \mathsf{FE}_{sd} + \check{\varepsilon}_{jsd}.$$

where $n_{js''}^{\textit{Bartik}} \perp_j \begin{pmatrix} \varepsilon_{js} & \check{\varepsilon}_{jsd} \end{pmatrix}^{\mathsf{T}} | c_j$, *for all* s, s'', d.

Corollary B.3. *Under Assumption B.3, we can write*

$$n_{jsd} = \tfrac{\tau}{\gamma}n_{js'} + \left(\tfrac{\gamma}{\kappa} - 1 - \tfrac{\tau^2}{\kappa\gamma}\right)\tilde{n}_{js}^{\textit{Bartik}} + \left(\tfrac{\tau}{\kappa} - \tfrac{\tau^2}{\kappa\gamma}\right)\tilde{n}_{js''}^{\textit{Bartik}} + \left(\tfrac{\tau^2}{\gamma} - \gamma\right)\hat{c}_{js} + \left(\tfrac{\tau^2}{\gamma} - \tau\right)\hat{c}_{js''} + \mathsf{FE}_{sd} + \mathsf{FE}_{s'} + \grave{\varepsilon}_{jsd,s'} \tag{B.12}$$

$$n_{jsd} = \tfrac{\gamma-\kappa}{\gamma}n_{js} + \tfrac{\tau}{\gamma}\tilde{n}_{js'}^{\textit{Bartik}} + \tfrac{\tau}{\gamma}\tilde{n}_{js''}^{\textit{Bartik}} - \kappa\hat{c}_{js} - \tfrac{\tau\kappa}{\gamma}\hat{c}_{js'} - \tfrac{\tau\kappa}{\gamma}\hat{c}_{js''} + \mathsf{FE}_{sd} + \grave{\varepsilon}_{jsd} \tag{B.13}$$

where $n_{js''}^{\textit{Bartik}} \perp_j \begin{pmatrix} \grave{\varepsilon}_{jsd,s'} & \grave{\varepsilon}_{jsd} \end{pmatrix}^{\mathsf{T}} | c_j$, *for all* s, d, s', s''.

These representations (B.12–B.13) directly yield the following two propositions.

Proposition B.4 (IV−Across Sector). *If we regress* n_{jsd} *on* $(n_{js'}, \tilde{n}_{js}^{\textit{Bartik}}, \tilde{n}_{js''}^{\textit{Bartik}})$, *instrumented by* $(n_{js}^{\textit{Bartik}}, n_{js'}^{\textit{Bartik}}, n_{js''}^{\textit{Bartik}})$, *controlling for* c_j, FE_{sd}, *and* $\mathsf{FE}_{s'}$, *then the IV coefficient on* $n_{js'}$ *converges in probability to* $\frac{\tau}{\gamma}$, *which is strictly positive and vanishes as* $\beta \to 0$.

Proposition B.5 (IV−Within Sector). *If we regress* n_{jsd} *on* $(n_{js}, \tilde{n}_{js'}^{\textit{Bartik}}, \tilde{n}_{js''}^{\textit{Bartik}})$, *instrumented by* $(n_{js}^{\textit{Bartik}}, n_{js'}^{\textit{Bartik}}, n_{js''}^{\textit{Bartik}})$, *controlling for* c_j *and* FE_{sd}, *then the IV coefficient on* n_{js} *converges in probability to* $\frac{\gamma-\kappa}{\gamma}$, *which is negative if and only if* (B.9) *holds.*

If we have external economies of scale, the probability limits become $\frac{\tilde{\tau}}{\tilde{\gamma}}$ and $\frac{\tilde{\gamma}-\kappa}{\tilde{\gamma}}$. In particular, $\frac{\tilde{\tau}}{\tilde{\gamma}}$ is strictly positive and vanishes as $(\beta, \varepsilon_c, \varepsilon_a) \to (0, 0, 0)$.

[10] In terms of the framework of GSS, what we need is orthogonality between θ_0 and the structural error terms. The proof of Lemma B.3 reveals that the structural error terms contain $\tilde{\varepsilon}_{js} \equiv \sum_d \theta_{jsd}\tilde{\varepsilon}_{jsd}$, $\hat{\varepsilon}_{js}$, and $\tilde{\varepsilon}_{jsd}$. The key difference from GSS is that the observables (n_{jsd}) we use to construct Bartik instruments are endogenous in this paper.

[11] For a variable x_j that has a region index j, we define the residualized version $x_j^{\perp}$ as the j-th residual from the regression of x_j on c_j. Using this notation, we can rewrite the exogeneity assumption as $\theta_{jsd,0}^{\perp} \perp_j \begin{pmatrix} \tilde{\varepsilon}_{js'}^{\perp} & \hat{\varepsilon}_{js'}^{\perp} & \tilde{\varepsilon}_{js'd'}^{\perp} \end{pmatrix}^{\mathsf{T}}$.

B.4 Efficiency Properties: Proofs

B.4.1 Efficiency Properties

B.4.1.1 Preliminary Results: CES Efficiency

We first prove some preliminary results that will be used to study the efficiency of decentralized equilibrium with trip chaining and/or external economies of scales. These results extend the CES efficiency result of Dixit and Stiglitz (1977) by allowing nested aggregation, heterogeneous consumers, and external economies of scale. We first consider nested utility functions and show that the decentralized equilibrium is efficient when the lowest nest features constant elasticity of substitution. We then consider heterogeneous agents and external economies of scale and provide the condition on the social welfare function and on the elasticity of external economies of scale under which the decentralized equilibrium attains the social optimum.[12]

Homogeneous agent. Consider a set of nests, each of which is indexed by $j \in \mathcal{J}$. Utility is given by arbitrary aggregation, $U = U(\{q_j\}_j)$. For example, in our application j indexes regions, sectors, and subsectors. Within each nest, we assume that quantities are aggregated by

$$q_j = \int_0^{N_j} f_j(q_j(\omega))\,\mathrm{d}\omega \quad \text{(e.g., CES: } f_j(q) = q^{1-1/\rho_j}\text{)}$$

The cost of producing one unit of goods in nest j is given by c_j, and the cost of increasing the number of variety for nest j is E_j. In the decentralized equilibrium, consumers solve the utility maximization problem,

$$\max_{\{q_j(\omega)\}_{j,\omega}} \quad U = U(\{q_j\}_{j,}) \quad \text{s.t.} \sum_j \int_0^{N_j} q_j(\omega) p_j(\omega)\,\mathrm{d}\omega \le w$$

where $p_j(\omega)$ is the price and w is income. The first-order condition is given by

$$\frac{\partial U}{\partial q_j} \cdot f_j'(q_j(\omega)) = \lambda p_j(\omega) \tag{B.14}$$

where λ is the Lagrange multiplier associated with the budget constraint. Thus, monopolistically competitive firms solve the profit maximization problem,

$$\max_{q_j(\omega)} \left\{ \frac{1}{\lambda} \frac{\partial U}{\partial q_j} f_j'(q_j(\omega)) q_j(\omega) - c_j q_j(\omega) \right\}.$$

[12] The proofs of lemmas, Proposition B.10, and Proposition B.11 are omitted and are available upon request.

Note that the structure of the economy is characterized by a mapping from $\{q_j(\omega)\}_{\omega\in[0,N_j]}$ to q_j and another mapping from $\{q_j\}_j$ to U. When there is only one nest and the latter is an identity map, it is well known (e.g., Dixit and Stiglitz, 1977) that the decentralized allocation is constrained efficient when the former features constant elasticity of substitution. We further show in the following proposition that this result holds for any mapping from $\{q_j\}_j$ to U. We postpone a proof of this result to the end of this section, where we prove a more general result with heterogeneous agents and external economies of scale.

Proposition B.6. *When $f_j(\cdot)$ is CES with elasticity of substitution not being dependent on j, the decentralized equilibrium coincides with the solution of the centralized welfare maximization.*[13]

Heterogeneous agent. There are I different agents types indexed by $i \in \mathcal{I}$. For example, in our application consumers are different in terms of their income and the regions they start their services travel. In the decentralized equilibrium, an agent of type i solves

$$\max_{\{q_j^i(\omega)\}_{j,\omega}} U^i = U^i(\{q_j^i\}_j) \quad \text{where } q_j^i = \int_0^{N_j} f_j(q_j^i(\omega))\,\mathrm{d}\omega \quad \text{s.t. } \sum_j \int_0^{N_j} p_j(\omega) q_j^i(\omega)\,\mathrm{d}\omega \le w^i \tag{B.15}$$

The first-order condition is given by

$$\frac{\partial U^i}{\partial q_j^i} \cdot f_j'(q_j^i(\omega)) = \lambda^i \cdot p_j(\omega), \tag{B.16}$$

where λ^i is the Lagrange multiplier associated with the budget constraint of agents with type i. This condition characterizes the demand $q_j^i(\omega) = q_j^i(p_j(\omega))$. Thus, monopolistically competitive firms solve

$$\max_{p_j(\omega)} \left\{ (p_j(\omega) - c_j) \sum_i q_j^i(p_j(\omega)) \right\}. \tag{B.17}$$

In this economy, the following proposition extends the result of Proposition B.6, characterizing the conditions on the social welfare function under which decentralized allocation solves the social planner problem. Again, we postpone a proof to the end of this section.

Proposition B.7. *When (i) $f_j(\cdot)$ is CES with elasticity of substitution ρ not being dependent on j, and (ii) the mapping $\{q_j^i(\omega)\}_{j,\omega} \mapsto U^i$ is homogeneous of degree s for some $s > 0$, the decentralized allocation solves the following social planner problem:*

$$\max_{\{N_j\}_j, \{q_j^i(\omega)\}_{j,\omega,i}} \sum_i w^i \cdot \log U^i \qquad \textit{(w^i acts as a Pareto weight)} \tag{SP}$$

[13] Elasticity of substitution should not be j-specific. See the proof of Lemma B.6.

$$\text{s.t.} \qquad \sum_i \sum_j \int_0^{N_j} c_j q_j^i(\omega)\,\mathrm{d}\omega + \sum_j E_j N_j \leq \sum_i w^i,$$

hence, it is Pareto efficient with Pareto weights independent of the structure of the economy.

External Economies of Scale. Again, there are I different agents types, where an agent of type i solves the utility maximization problem (B.15), and firms solve the profit maximization problem (B.17). Suppose that the set of nests $\mathcal{J}$ is partitioned into $\mathcal{J} = \biguplus_\iota \mathcal{J}_\iota$, and $J : \mathcal{J} \to \{J_\iota\}$ assigns each element of $\mathcal{J}$ to the partition that contains it. The only difference is that now the unit cost and entry cost are endogenously determined by

$$c_j = c_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) \text{ and } E_j = E_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)})$$

where $\Upsilon_{1J(j)} = \sum_{j' \in J(j)} \sum_i \int_0^{N_{j'}} c_{j'} q_{j'}^i(\omega)\,\mathrm{d}\omega$ and $\Upsilon_{2J(j)} = \sum_{j' \in J(j)} E_{j'} N_{j'}$ are the total resource spent on production and variety creation for partition $J(j)$, respectively.[14]

As an intermediate step, we follow Dhingra and Morrow (2019) and first characterize the conditions under which the decentralized allocation solves the centralized *revenue* maximization problem:

$$\max_{\{N_j\}_j, \{q_j^i(\omega)\}_{j,\omega,i}} \quad \sum_{j,i} \int_0^{N_j} \underbrace{\frac{1}{\lambda^i} \frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}} \cdot f_j'(q_j^i(\omega))}_{\text{from (B.16)}} \cdot q_j^i(\omega)\,\mathrm{d}\omega \tag{CRM}$$

$$\text{s.t.} \qquad \sum_{j,i} \int_0^{N_j} c_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) q_j^i(\omega)\,\mathrm{d}\omega + \sum_j E_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) N_j \leq \sum_i w_i$$

where the value of $\frac{1}{\lambda^i}\frac{\partial U^i}{\partial q_j^i}$ is evaluated at the decentralized allocation so that the term $\frac{1}{\lambda^i}\frac{\partial U^i}{\partial q_j^i}\Big|_{\text{de}} \cdot f_j'(q_j^i(\omega))$ captures the residual demand firm j faces in the decentralized equilibrium, and we take this value as given when solving problem (CRM). Lemmas B.4 and B.5 summarize the results.

Lemma B.4 (External Economies of Scale I). *Assume that (i) c_j and E_j feature isoelastic external economies of scale:*

$$\frac{\partial \ln c_j}{\partial \ln \Upsilon_{\ell J(j)}} = \varepsilon_{c\ell} \text{ and } \frac{\partial \ln E_j}{\partial \ln \Upsilon_{\ell J(j)}} = \varepsilon_{E\ell}$$

for $\ell = 1, 2$ and that (ii) $f_j(\cdot)$ is CES with elasticity of substitution ρ_j possibly being different across j. A sufficient condition for the decentralized equilibrium to solve the centralized revenue maximization problem is $\varepsilon_{c1} = \varepsilon_{E2}$ and $\varepsilon_{c2} = \varepsilon_{E1} = 0$. Unless ρ_j is the same across all j, this is also a necessary condition.[15]

[14] In general, $\Upsilon_{1J(j)}$ and $\Upsilon_{2J(j)}$ might not be well-defined because they depend on c_j and E_j, which are functions of $\Upsilon_{1J(j)}$ and $\Upsilon_{2J(j)}$. We implicitly restrict $c_j(\cdot)$ and $E_j(\cdot)$ so that $\Upsilon_{1J(j)}$ and $\Upsilon_{2J(j)}$ are well-defined. Note that under the conditions of Lemma B.4 or B.5, $\Upsilon_{1J(j)}$ and $\Upsilon_{2J(j)}$ are indeed well-defined. For example, we assume $c_j = \bar{c}_j \cdot \Upsilon_{1J(j)}^{\varepsilon_c}$ in Lemma B.4. This gives $\Upsilon_{1J(j)} = \sum_{j' \in J(j)} c_{j'} q_{j'} = \left(\sum_{j' \in J(j)} \bar{c}_{j'} q_{j'}\right) \Upsilon_{1J(j)}^{\varepsilon_c} = \left(\sum_{j' \in J(j)} \bar{c}_{j'} q_{j'}\right)^{1/(1-\varepsilon_c)}$.

Lemma B.5 (External Economies of Scale II). *Assume that (i) c_j and E_j are functions of $\Upsilon_{J(j)} = \Upsilon_{1J(j)} + \Upsilon_{2J(j)}$, and they feature isoelastic external economies of scale:*

$$\frac{\partial \ln c_j}{\partial \ln \Upsilon_{J(j)}} = \varepsilon_c \text{ and } \frac{\partial \ln E_j}{\partial \ln \Upsilon_{J(j)}} = \varepsilon_E$$

and that (ii) $f_j(\cdot)$ is CES with elasticity of substitution ρ_j possibly being different across j. A sufficient condition for the decentralized equilibrium to solve the centralized revenue maximization problem is $\varepsilon_c = \varepsilon_E$. Unless ρ_j is the same across all j, this is also a necessary condition.[16,17]

Finally, we characterize additional conditions needed to show that the solution of the centralized revenue maximization problem and that of the social planner problem coincide.

Lemma B.6 (CRM $\to$ SP). *When (i) $f_j(\cdot)$ is CES with elasticity of substitution ρ not being dependent on j, and (ii) the mapping $\{q_j^i(\omega)\}_{j,\omega} \mapsto U^i$ is homogeneous of degree s for some $s > 0$, the solution of (CRM) coincides with the solution of the following social planner problem:*

$$\max_{\{N_j\}_j, \{q_j^i(\omega)\}_{j,\omega,i}} \quad \sum_i w^i \cdot \log U^i \qquad (w^i \text{ acts as a Pareto weight}) \qquad \text{(SP-EES)}$$

$$\text{s.t.} \quad \sum_{j,i} \int_0^{N_j} c_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) q_j^i(\omega)\, \mathrm{d}\omega + \sum_j E_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) N_j \leq \sum_i w^i.$$

Lemmas B.4–B.6 immediately prove Proposition B.8, which characterizes conditions under which the decentralized allocation is efficient. Propositions B.6 and B.7 are special cases of Proposition B.8 with homogeneous agent, $I = 1$, and without external economies of scale.

Proposition B.8. *When (i) c_j and E_j satisfy the conditions of either Lemma B.4 or B.5, (ii) $f_j(\cdot)$ is CES with elasticity of substitution ρ not being dependent on j, and (iii) the mapping $\{q_j^i(\omega)\}_{j,\omega} \mapsto U^i$ is homogeneous of degree s for some $s > 0$, the decentralized allocation solves the social planner problem (SP-EES).*

[15] If ρ_j is the same across all j, we only need $\varepsilon_{c1} + \frac{1}{\rho-1}\varepsilon_{E1} = (\rho-1)\varepsilon_{c2} + \varepsilon_{E2}$.

[16] If ρ_j is the same across all j, we do not need any condition on ε_c and ε_E.

[17] Even when there is no external economies of scale, the decentralized allocation does not solve (CRM) in genereal. The necessary and sufficient condition for it to solve (CRM) for arbitrary $\{w^i\}_i$ is that f_j is CES with elasticity of substitution possibly being different across j. In contrast to Dhingra and Morrow (2019), we need CES assumption to prove that the decentralized allocation solves the (CRM). To understand this, suppose that firms can price discriminate agents with different types (i.e., type-specific price $p_j^i(\omega)$ instead of $p_j(\omega)$). We can show that the decentralized allocation with price discrimination always solves (CRM). Thus, the decentralized allocation without price discrimination solves (CRM) if and only if we have $p_j^i(\omega) = p_j^{i'}(\omega)$ for all $i \neq i'$ under the decentralized equilibrium with price discrimination This requires markups to be uniform across i. Since consumers with different types will generically consume different quantities, uniform markup in turn requires f_j to be constant elasticity of substitution (CES).

B.4.1.2 Preliminary Results: Two-Step Maximization

In this section, we prove that a certain class of utility maximization problems can be solved in two steps. This includes both the decentralized utility maximization problem and social planner problem with trip-chaining and external economies of scale. We will use this result to show that the possibility of trip chaining only affects the mapping from underlying quantities to utility, not affecting the efficiency property of the decentralized equilibrium. Consider a utility maximization problem subject to a resource constraint:

$$\begin{aligned} \max_{\{x_j(\omega;\sigma)\}_{j,\omega,\sigma}} \quad & \tilde{U}(\{x_j(\sigma)\}_{j,\sigma}) \quad \text{where } \; x_j(\sigma) = F(\{x_j(\omega;\sigma)\}_\omega) \\ \text{s.t.} \quad & \sum_j \sum_\sigma \pi(\sigma) \sum_\omega c_j(\omega; \Upsilon_{1J(j)}) \cdot x_j(\omega;\sigma) \leq I \end{aligned} \tag{B.18}$$

where j again indexes different nests, σ indexes separate purchases from a given nest, and ω indexes different variety in nest j. For example, in our application j indexes regions, sectors, and subsectors, σ indexes individual purchases of services goods, and ω indexes individual stores. $\pi(\sigma)$ is the weight, and the aggregation $F(\cdot)$ is constant return to scale.

To encompass the case with external economies of scale, we allow the cost $c_j(\cdot)$ to negatively depend on the resource spent on production for partition $J(j)$,

$$\Upsilon_{1J(j)} = \sum_{j' \in J(j)} \sum_\sigma \pi(\sigma) \sum_\omega c_{j'}(w; \Upsilon_{1J(j)}) x_{j'}(\omega;\sigma).$$

The key idea of two-step maximization is that when the minimized cost of producing $\{x_j(\sigma)\}_{j,\sigma}$ only depends on the values of $\{x_j\}_j$ where $x_j = \sum_\sigma \pi(\sigma) x_j(\sigma)$, we can solve the utility maximization problem by first maximizing the utility for given values of $\{x_j\}_j$, and then maximizing it over possible values of $\{x_j\}_j$. The following lemma formalizes this idea. The proof is given in Oh and Seo (2023).

Lemma B.7 (Two-step Maximization). *We can solve problem* (B.18) *in two steps. First, we solve the problem for given values of* $\{x_j\}_j$*:*

$$U(\{x_j\}_j) = \max_{\{x_j(\sigma)\}_{j,\sigma}} \tilde{U}(\{x_j(\sigma)\}_{j,\sigma}) \;\; \text{s.t.} \sum_\sigma \pi(\sigma) x_j(\sigma) \leq x_j, \quad \forall j.$$

Second, we choose $\{x_j\}_j$ *that maximize* U *subject to the resource constraint:*

$$\max_{\{x_j\}_j} U(\{x_j\}_j) \;\; \text{s.t.} \sum_j c(\{c_j(\omega; \Upsilon_{1J(j)})\}_\omega) \cdot x_j \leq I.$$

We can easily see that this is equivalent to

$$\max_{\{x_j(\omega)\}_{j,\omega}} U(\{x_j\}_j) \quad \text{where } x_j = F(\{x_j(\omega)\}_\omega) \quad \text{s.t. } \sum_j \sum_\omega c_j(\omega; \Upsilon_{1J(j)}) \cdot x_j(\omega) \leq I.$$

Trip Chaining. To apply Propositions B.6–B.8 to our model, we need to reformulate the decentralized utility maximization problem and social planner problem to two-step maximization problems. Recall that the consumption utility $\mathcal{U}^C_{i^r i^w}(o; I_{i^w}(o))$ for a worker o who live in zone i^r and work in zone i^w with income $I_{i^w}(o)$ is given by (3). For expositional simplicity, we assume throughout this section that consumers start their services travel only from their resident zone, that they only consume tradable goods and services, and that consumers who live in the same zone have the same income. But the results in this section hold without these assumptions. Under these assumptions, (3) is simplified to

$$\mathcal{U}^C_i = \max_{\tilde{C}_i, C^r_i} \left(\frac{\tilde{C}_i}{1-\mu}\right)^{1-\mu} \left(\frac{C^r_i}{\mu}\right)^{\mu} \text{ s.t. } \tilde{C}_i + P_i C^r_i \leq I_i$$

where $\mu = \mu^r_c$. The subscript i denotes the resident zone. $\tilde{C}_i$ and C^r_i denote the consumption indices for tradable goods and services travel. I_i is income of workers who live in region i. In the decentralized equilibrium, these workers solve the utility maximization problem

$$\begin{aligned}
&\max_{\{j^i_t(\sigma^t)\},\{q^i(\sigma^t)\},\{q^i_d(\sigma^t)\},\{q^i_d(\omega;\sigma^t)\},\tilde{C}_i} \quad \mathcal{U}^C_i \qquad \text{(DE)}\\
&\text{where} \quad C^r_i = \exp((1-\beta)V_i)\\
&\qquad V_i = \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U\big(q^i(\sigma^t)\big) - \tau d(j_{t-1}(\sigma^{t-1}), j^i_t(\sigma^t)) - \varphi \mathbb{1}_{j_{t-1}(\sigma^{t-1}) \neq^i j^i_t(\sigma^t)} + \nu \varepsilon^{j^i_t(\sigma^t)}_t \Big) \pi(\sigma^t)\\
&\qquad q^i(\sigma^t) = \left(\sum_d \phi^{1/\sigma}_{j^i_t(\sigma^t) s(\sigma^t) d} q^i_d(\sigma^t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}}, q^i_d(\sigma^t) = \left(\int_0^{N_{j^i_t(\sigma^t) s(\sigma^t) d}} q^i_d(\omega;\sigma^t)^{1-1/\rho} \, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}}\\
&\text{s.t.} \quad \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \left(\sum_d \int_0^{N_{j^i_t(\sigma^t) s(\sigma^t) d}} p_{j^i_t(\sigma^t) s(\sigma^t) d}(\omega) \cdot q^i_d(\omega;\sigma^t) \, \mathrm{d}\omega \right) \pi(\sigma^t) + \tilde{C}_i \leq I_i.
\end{aligned}$$

Consider a constrained social planner problem that maximizes the utility of the representative consumer by choosing resource allocation within services market. The social planner is constrained in the sense that she cannot change the resource allocation between tradable goods consumption and services market. Thus, the resource allocated to tradable goods is given by the amount chosen in the decentralized equilibrium, $(1-\mu)\sum_i I_i L_i$.

$$\begin{aligned}
&\max_{\{N_{jsd}\},\{j^i_t(\sigma^t)\},\{q^i(\sigma^t)\},\{q^i_d(\sigma^t)\},\{q^i_d(\omega;\sigma^t)\},\{\tilde{C}_i\}} \sum_i \theta_i \log \mathcal{U}^C_i L_i \qquad \text{(SP)}\\
&\text{where} \quad C^r_i = \exp((1-\beta)V_i)\\
&\qquad V_i = \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U\big(q^i(\sigma^t)\big) - \tau d(j^i_{t-1}(\sigma^{t-1}), j^i_t(\sigma^t)) - \varphi \mathbb{1}_{j^i_{t-1}(\sigma^{t-1}) \neq j^i_t(\sigma^t)} + \nu \varepsilon^{j_t(\sigma^t)}_t \Big) \pi(\sigma^t)
\end{aligned}$$

$$
\begin{aligned}
& q^i(\sigma^t) = \left(\sum_d \phi^{1/\sigma}_{j^i_t(\sigma^t)s(\sigma^t)d} q^i_d(\sigma^t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}}, q^i_d(\sigma^t) = \left(\int_0^{N_{j^i_t(\sigma^t)s(\sigma^t)d}} q^i_d(\omega;\sigma^t)^{1-1/\rho} \, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\
\text{s.t.} \quad & \sum_i \left(\sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \sum_d \int_0^{N_{j^i_t(\sigma^t)s(\sigma^t)d}} \frac{c_{j^i_t(\sigma^t)s(\sigma^t)d}}{A_{j^i_t(\sigma^t)s(\sigma^t)d}} \cdot q^i_d(\omega;\sigma^t) \, \mathrm{d}\omega \pi(\sigma^t) \right) L_i + \sum_{j,s,d} N_{jsd} C_{jsd} + \sum_i \tilde{C}_i L_i \leq \sum_i I_i L_i \\
& \sum_i \tilde{C}_i L_i = (1-\mu) \sum_i I_i L_i
\end{aligned}
$$

The unconstrained social planner solves the same problem without the last constraint.

We can reformulate (DE) and (SP) to two-step maximization problems:[18] First, we compute the maximized utility from non-tradable services for given values of $\{q^i_{jsd}\}$, and the maximized value is denoted by $C^i(\{q^i_{jsd}\})$:

$$
\begin{aligned}
C^i(\{q^i_{jsd}\}) = \max_{\{j^i_t(\sigma^t)\},\{q^i(\sigma^t)\},\{q^i_d(\sigma^t)\}} \quad & C^r_i = \exp((1-\beta)V_i) \\
\text{where} \quad & V_i = \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U\big(q^i(\sigma^t)\big) - \tau d(j^i_{t-1}(\sigma^{t-1}), j^i_t(\sigma^t)) - \varphi \mathbb{1}_{j^i_{t-1}(\sigma^{t-1}) \neq j^i_t(\sigma^t)} + \nu \varepsilon^{j^i_t(\sigma^t)}_t \Big) \pi(\sigma^t) \\
& q^i(\sigma^t) = \left(\sum_d \phi^{1/\sigma}_{j^i_t(\sigma^t)s(\sigma^t)d} q^i_d(\sigma^t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}} \\
\text{s.t.} \quad & \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \mathbb{1}_{j^i_t(\sigma^t)=j, s(\sigma^t)=s} \cdot q^i_d(\sigma^t) \pi(\sigma^t) \leq q^i_{jsd}, \quad \forall j,s,d.
\end{aligned}
$$

Think of problem (DE) as maximizing the objective function for a given value of $\tilde{C}_i$ and then maximizing over possible values of it. An equivalent formulation of (DE) is

$$
\begin{aligned}
\max_{\{q^i_{jsd}\},\{q^i_{jsd}(\omega)\},\tilde{C}_i} \quad & \log \tilde{C}_i^{1-\mu} \cdot (C^i(\{q^i_{jsd}\})^{\mu} \\
\text{where} \quad & q^i_{jsd} = \left(\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho} \, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\
\text{s.t.} \quad & \sum_{j,s,d} \int_0^{N_{jsd}} p_{jsd}(\omega) q^i_{jsd}(\omega) \, \mathrm{d}\omega + \tilde{C}_i \leq I_i.
\end{aligned}
\tag{DE'}
$$

Similarly, think of problem (SP) as maximizing the objective function for given values of $\{\tilde{C}_i\}$ and $\{N_{jsd}\}$ and then maximizing over possible values of them. We then have an equivalent formulation of (SP),

$$
\begin{aligned}
& \max_{\{N_{jsd}\},\{q^i_{jsd}\},\{q^i_{jsd}(\omega)\},\{\tilde{C}_i\}} \quad \sum_i \theta_i \log U^C_i L_i \\
\text{where} \quad & U^C_i = \tilde{C}_i^{1-\mu} (C^i(\{q^i_{jsd}\})^{\mu} \\
& q^i_{jsd} = \left(\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho} \, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\
\text{s.t.} \quad & \sum_i \left(\sum_{j,s,d} \int_0^{N_{jsd}} \tilde{c}_{jsd}(\Upsilon_{1j}, \Upsilon_{2j}) q^i_{jsd}(\omega) \, \mathrm{d}\omega \right) L_i + \sum_{j,s,d} N_{jsd} C_{jsd}(\Upsilon_{1j}, \Upsilon_{2j}) + \sum_i \tilde{C}_i L_i \leq \sum_i I_i L_i \\
& \sum_i \tilde{C}_i L_i = (1-\mu) \sum_i I_i L_i
\end{aligned}
\tag{SP'}
$$

[18] The index σ in this section corresponds to σ^t here, ω to ω, j to (j,s,d), and $\pi(\sigma)$ to $\beta^t \cdot \pi(\sigma^t)$.

where $\tilde{c}_{jsd}(\Upsilon_{1j}, \Upsilon_{2j}) = \dfrac{c_{jsd}(\Upsilon_{1j}, \Upsilon_{2j})}{A_{jsd}(\Upsilon_{1j}, \Upsilon_{2j})}$.

B.4.1.3 Application: Efficiency Properties of Trip Chaining

In this section, we apply Propositions B.6 and B.7 to demonstrate that trip chaining does not give rise to any inefficiencies. First, we show that if all consumers are identical, the decentralized resource allocation *within* services market—across regions and between production and variety creation—is efficient. Second, we consider a general case where consumers differ in terms of their income levels and the origins of their services travel. We demonstrate that even in this case, trip chaining does not introduce any inefficiencies. Finally, we show that while an unconstrained social planner would reallocate resources from tradable goods to nontradable services, this inefficiency does not interact with the presence of trip chaining.

Homogeneous Consumer, Constrained Social Planner. We assume for the moment that consumers are homogeneous and reside in i, with an income of I_i. In this economy, there are two potential inefficiencies in the resource allocation within services market. First, resources can be allocated inefficiently across consumption regions. Second, within a consumption region, resources can be inefficiently allocated between production and store creation. This is the quantity-diversity trade-off discussed by Dixit and Stiglitz (1977). The following proposition shows that these inefficiencies do not arise in the decentralized equilibrium, regardless of the presence of trip chaining.

Proposition B.9. *In the economy with homogeneous consumers, the decentralized allocation solves the constrained social planner problem (SP).*

This is a direct application of Proposition B.6 to (DE′) and (SP′), where the index j in the proposition corresponds to each region-sector-subsector (j, s, d).

Heterogeneous Consumer, Constrained Social Planner. Let us now assume that different consumers start their services travel from different regions, indexed by i, and have different incomes denoted as I_i.

We refer to the social planner problem with Pareto weights $\theta_i = I_i$ for all i as the benchmark social planner problem, as the decentralized allocation is shown to solve it when trip chaining is not allowed ($\beta = 0$). By considering this benchmark social planner problem, we can concentrate solely on the potential inefficiency arising from the trip-chaining mechanism. The following proposition reveals that trip chaining, in fact, does not generate inefficiency.

To prove Proposition 2 (1), we can once again apply Proposition B.7 to (DE′) and (SP′), with the index j in the proposition corresponding to each region-sector-subsector (j, s, d). Notably, the presence of trip chaining does not alter the homogeneous function condition.

Unconstrained Social Planner. Now suppose that the social planner has the flexibility to reallocate resources between tradable goods consumption and the services market. In this case, the social planner solves problem (SP′), but under a single resource constraint that applies to both tradable goods consumption and the services market. To understand how we can implement the socially optimal allocation, we introduce three types of taxes: subsidies for non-tradable services $\{s_{jsd}(\omega)\}$, a tax on tradable goods t^{tradable}, and entry subsidies $\{S_{jsd}\}$. These taxes create wedges between the prices faced by consumers $\{p_{jsd}(\omega)\}$ and those faced by firms $\{\bar{p}_{jsd}(\omega)\}$, between the price of tradable goods faced by consumers p^{tradable} and its competitive price, which is normalized to 1, and between the entry cost faced by firms $\bar{C}_{jsd}$ and the resource cost of entry C_{jsd},

$$\begin{aligned} p_{jsd}(\omega) &= (1 - s_{jsd}(\omega))\bar{p}_{jsd}(w) \\ p^{\text{tradable}} &= 1 + t^{\text{tradable}} \\ \bar{C}_{jsd} &= (1 - S_{jsd})C_{jsd}. \end{aligned}$$

Net revenues are rebated back to consumers through a lump-sum transfer T_i. Under these taxes, the decentralized utility maximization problem can be expressed as follows:

$$\begin{aligned} \max_{\{q^i_{jsd}\},\{q^i_{jsd}(\omega)\},\tilde{C}_i} \quad & \log \tilde{C}_i^{1-\mu} \cdot (C^i(\{q^i_{jsd}\})^{\mu} \\ \text{where} \quad & q^i_{jsd} = \left(\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho} \, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\ \text{s.t.} \quad & \sum_{j,s,d} \int_0^{N_{jsd}} p_{jsd}(\omega) q^i_{jsd}(\omega) \, \mathrm{d}\omega + p^{\text{tradable}} \tilde{C}_i \leq I_i + T_i \end{aligned} \tag{DE′}$$

and the free-entry condition is given by:

$$\bar{C}_{jsd} = \sum_i (\bar{p}_{jsd}(\omega) - \tilde{c}_{jsd}) q^i_{jsd}(\omega) L_i.$$

Proposition 2 (2) shows that that the unconstrained social planner would reallocate resources from tradable goods consumption to non-tradable services consumption, achieved through taxing the former and subsidizing the latter: $t^{\text{tradable}} = \frac{\mu}{\rho-1}$, $s_{jsd}(\omega) = \frac{1-\mu}{\rho}$, $S_{jsd} = 0$, and $T_i = 0$. In particular, the social planner increases the number of non-tradable services stores proportionally more than the decentralized number of stores. Importantly, this proportionality remains unchanged regardless of the presence of trip chaining. A formal proof of this result is deferred to the next section, where we prove a general result with external economies of scale.

B.4.1.4 Application: Efficiency Properties of External Economies of Scale

Constrained-Efficient Specification. In this section, we first apply Proposition B.8 to characterize the conditions on the form of external economies of scales under which the decentralized equilibrium achieves constrained efficiency. Subsequently, we show that the economy is generically constrained inefficient and that the presence of external economies of scale exacerbates the inefficient allocation of resources between tradable goods and non-tradable services. We can apply Proposition B.8 to show Proposition 3 (1).

General Constrained Inefficiency. To illustrate the generic inefficiency associated with non-tradable services market with external economies of scale, we consider a general case with

$$\tilde{c}_{jsd} = c_{jsd}(\Upsilon_{1j}, \Upsilon_{2j}) \text{ and } C_{jsd} = C_{jsd}(\Upsilon_{1j}, \Upsilon_{2j})$$

with possibly varying elasticities,

$$\varepsilon_{c\ell j} = \frac{\partial \ln \tilde{c}_{jsd}}{\partial \ln \Upsilon_{\ell j}} \text{ and } \varepsilon_{E\ell j} = \frac{\partial \ln C_{jsd}}{\partial \ln \Upsilon_{\ell j}}.$$

The following proposition underscores that, in the presence of external economies of scale, the non-tradable services market is generically constrained inefficient.

Proposition B.10 (Constrained Inefficiency). *The constrained-efficient allocation can be implemented through a combination of nontradable services subsidies:*

$$S_{jsd} = \frac{\rho(\mathcal{E}_{2j} - \mathcal{E}_{1j})}{1 - \mathcal{E}_{1j} + (\rho - 1)(\mathcal{E}_{2j} - \mathcal{E}_{1j})} \text{ and } s_{jsd}(\omega) = \mathcal{E}_{1j}.$$

where

$$1 - \mathcal{E}_{1j} = \frac{1 - \varepsilon_{E2j} + \frac{1}{\rho - 1}\varepsilon_{E1j}}{(1 - \varepsilon_{c1j})(1 - \varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}} \text{ and } 1 - \mathcal{E}_{2j} = \frac{1 - \frac{\rho-1}{\rho}\varepsilon_{E2j} + \frac{1}{\rho}\varepsilon_{E1j} + \frac{\rho-1}{\rho}\varepsilon_{c2j} - \frac{1}{\rho}\varepsilon_{c1j}}{(1 - \varepsilon_{c1j})(1 - \varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}.$$

This proposition clearly demonstrates the inefficiency of the decentralized equilibrium in terms of both interregional and intraregional allocation. For instance, if $\mathcal{E}_{1j}$ and $\mathcal{E}_{2j}$ tend to increase with Υ_{1j} and Υ_{2j}, the social planner would subsidize concentrated regions with higher values of Υ_{1j} and Υ_{2j}. If $\mathcal{E}_{2j}$ tend to be higher than $\mathcal{E}_{1j}$, it is optimal to reallocate resources from production to variety creation.

Unconstrained Inefficiency. To simplify the exposition, we consider isoelastic external economies of scale as assumed in Proposition 3.

Proposition B.11 (Unconstrained Social Planner). *The unconstrained social planner chooses the number of non-tradable services stores* $\{N^*_{jsd}\}$ *given by*

$$N^*_{jsd} = \chi(\varepsilon) \cdot N^{LF}_{jsd},$$

where N^{LF}_{jsd} *represents the number of stores in the laissez-faire equilibrium. The constant* $\chi(\varepsilon) = \frac{\tilde{\rho}(\varepsilon)}{\tilde{\rho}(\varepsilon)-(1-\mu)} > 1$*, where* $\tilde{\rho}(\varepsilon) = \frac{\rho(1+\varepsilon)}{1+\rho\varepsilon}$*, remains unaffected by the presence of trip chaining, but positively depends on* ε*. The optimal allocation can be implemented through a combination of a tax on tradable goods and a subsidy for non-tradable services:* $t^{tradable} = \frac{\mu}{\tilde{\rho}(\varepsilon)-1}$*,* $s_{jsd}(\omega) = \frac{1-\mu}{\tilde{\rho}(\varepsilon)}$*,* $S_{jsd} = 0$*, and* $T_i = 0$*.*

B.4.2 A General Equilibrium Model

We close the model by specifying the remaining parts of the city structure. In particular, we endogenize the spatial distribution of consumers, wages, and rent prices that we take as given in equilibrium of the services market. We mainly follow Ahlfeldt et al. (2015) and Tsivanidis (2019) with a few key modifications. Each location differs in terms of their productivity, amenities, wages, and land supply. We first consider workers' location choice problems. Then, we describe how decisions of firms—both tradable or non-tradable—and market clearing conditions determine wages and rent prices.

In this section, our spatial unit is a district, which is a larger unit than a zone. In particular, each worker chooses districts to live and work, and the labor and land markets clear at the district level. This is mainly due to data limitations: for a few variables including average wages by residential region, we only have data at the district level. We continue to assume that the spatial unit of the services market is a zone. The main focus of this paper is the services distribution and its efficiency and welfare consequences. Thus, as long as counterfactual exercises affect zones within a district similarly in terms of general equilibrium outcomes, our assumption is not overly restrictive.

Workers' Location Choice. A city is populated with a fixed measure of workers M. Workers, indexed by o, choose where to live d^r and where to work d^w.[19] Once workers determine a pair of districts (d^r, d^w), they are randomly allocated to a residential zone $i^r \in d^r$ and a business zone $i^w \in d^w$, according to probability $\Pr(i^r, i^w | d^r, d^w)$. For simplicity, we assume that a residential zone and a business zone are independently determined, i.e., $\Pr(i^r, i^w | d^r, d^w) = \Pr^r(i^r | d^r) \Pr^w(i^w | d^w)$. The value of worker o is given by

$$\mathcal{U}(o) = \max_{d^r, d^w} \ \mathbb{E}[B_{d^r} \cdot z_{d^r}(o) \cdot \mathcal{U}_{i^r i^w}(o) | d^r, d^w]$$

[19] We use d to index districts to avoid confusion with i and j, which are indices for zones.

where B_{d^r} is a regional residential amenity, and $z_{d^r}(o)$ is an idiosyncratic residential amenity. The term $\mathcal{U}_{i^r i^w}(o)$ summarizes the utility associated with to the workplace and consumption decisions.

$$\mathcal{U}_{i^r i^w}(o) = \frac{\xi_{d^w}}{\tilde{d}_{i^r i^w}} \mathcal{U}^C_{i^r i^w}(o; I_{i^w}(o))$$

where ξ_{d^w} is a workplace amenity, and $\mathcal{U}^C_{i^r i^w}(\cdot)$ is the consumption utility defined in (3). Finally, $\tilde{d}_{i^r i^w}$ is the commuting disutility, given by

$$\tilde{d}_{i^r i^w} = \exp\left(\kappa d_{d(i^r)d(i^w)} + \varphi_\kappa \mathbb{1}_{d(i^r)\neq d(i^w)}\right)$$

where $d(i)$ denotes a district to which zone i belongs.[20] The distance between two districts d^r and d^r is defined as a weighted average of the distance between zones in two districts, $d_{d^r,d^w} = \sum_{i^r,i^w} d_{i^r,i^w} \cdot P(i^r, i^w | d^r, d^w)$. Note that the parameter κ can be different from its counterpart τ, which governs the disutility of services travel. We also include a border effect φ_κ as before.

We assume that labor income is the only source of income, and factor payments to capital or land go to absentee owners. Thus, income $I_{i^w}(o)$ in the budget constraint in (3) is given by

$$I_{i^w}(o) = w_{d(i^w)} \cdot v_{d(i^w)}(o)$$

where w_{d^w} is the workplace-specific wage, and $v_{d^w}(o)$ is the idiosyncratic component of the wage. Finally, we assume that the idiosyncratic components $z_{d^r}(o)$ and $v_{d^w}(o)$ follow Fréchet distributions:

$$F_z(z) = \exp(-z^{-\varepsilon_z}) \text{ and } F_v(v) = \exp(-v^{-\varepsilon_v})$$

where $\varepsilon_z, \varepsilon_v > 1$ are the shape parameters. Higher values imply that idiosyncratic components have less importance in decisions.[21]

For simplicity, we follow Tsivanidis (2019) to assume that workers first choose residential districts, and then choose business districts.[22] This allows a simple equilibrium characterization using backward induction. We can summarize the timeline as follows. First, a worker o observes realizations of $\{z_{d^r}(o)\}_{d^r}$. Second, she chooses her residential district d^r that gives her the highest expected utility. Third, she observes realizations of $\{v_{d^w}(o)\}_{d^w}$. Fourth, she optimally chooses business district d^w. Fifth, she is randomly allocated to (i^r, i^d). Finally, she makes consumption decisions for tradable goods, non-tradable services, and residential floor space.

[20] Alternatively, we can assume that the commuting disutility is defined between zones: $\tilde{d}_{i^r i^w} = \exp(\kappa d_{i^r i^w} + \varphi_\kappa \mathbb{1}_{i^r \neq i^w})$. This specification, however, does not allow us to use a gravity equation to estimate (κ, φ_κ).

[21] It is well known in the literature that it is without loss to assume unit scale parameters because they can be isomorphically captured by the terms B_{d^r} and ξ_{d^w}.

[22] In Ahlfeldt et al. (2015), they assume that workers draw idiosyncratic component of utility for all pairs (i^r, i^w) from the independent Fréchet distribution, but this approach is computationally burdensome.

Let us start with the business area decision of workers who chose to live in d^r. From (3), the utility when working in d^w is given by

$$\mathcal{U}_{d^r d^w}(o) = \sum_{i^r i^w} \Pr(i^r, i^w | d^r, d^w) P_{i^r}^{-\mu_c^r} P_{i^w}^{-\mu_c^w} \cdot (p^{tradable})^{-\mu_{\tilde{C}}} \cdot r_{d^r}^{-\mu_\ell} \cdot \frac{\xi_{d^w} \cdot w_{d^w} \cdot v_{d^w}(o)}{\exp(\kappa d_{d^r d^w} + \varphi_\kappa \mathbb{1}_{d^r \neq d^w})}. \tag{B.19}$$

Workplace d^w affects not only the wage but also the price index for services travel starting from the workplace. This together with the Fréchet assumption gives the probability of workers, who live in d^r, choosing to work in d^w:

$$\Pr(d^w | d^r) = \frac{\left(\bar{P}_{d^w}^{-\mu_c^w} \xi_{d^w} w_{d^w} \exp(-\kappa d_{d^r d^w} - \varphi_\kappa \mathbb{1}_{d^r \neq d^w})\right)^{\varepsilon_v}}{\sum_d \left(\bar{P}_d^{-\mu_c^w} \xi_d w_d \exp(-\kappa d_{d^r d} - \varphi_\kappa \mathbb{1}_{d^r \neq d})\right)^{\varepsilon_v}}$$

where $\bar{P}_{d^w}^{-\mu_c^w} = \sum_{i^w \in d^w} \Pr^w(i^w | d^w) P_i^{-\mu_c^w}$ is the expected price index of services travel that starts from the workplace. We now turn to the residential district choice. The expected indirect utility from choosing a residential district d^r has a simple expression,

$$\begin{aligned}
\mathcal{U}_{d^r}(o) &= \mathbb{E}\left[\max_{d^w} \{\mathcal{U}_{d^r d^w}(o)\} \middle| d^r\right] \\
&= B_{d^r} \cdot z_{d^r}(o) \cdot \bar{P}_{d^r}^{-\mu_c^r} (p^{tradable})^{-\mu_{\tilde{C}}} r_{d^r}^{-\mu_\ell} \cdot \left(\sum_d \left(\bar{P}_d^{-\mu_c^w} \xi_d w_d \exp(-\kappa d_{d^r d} + \varphi_\kappa \mathbb{1}_{d^r \neq d})\right)^{\varepsilon_v} \right)^{1/\varepsilon_v} \\
&\equiv \mathcal{U}_{d^r} \cdot z_{d^r}(o)
\end{aligned}$$

where expectation is taken over $\{v_{d^w}(o)\}_{d^w}$, and $\bar{P}_{d^r}^{-\mu_c^r} = \sum_{i^r \in d^r} \Pr^r(i^r | d^r) P_{i^r}^{-\mu_c^r}$ is the expected price index, defined analogously to $P_{d^w}^{-\mu_c^w}$. From the Fréchet assumption, workers choose a district d^r with probability

$$\Pr(d^r) = \frac{(\mathcal{U}_{d^r})^{\varepsilon_z}}{\sum_d (\mathcal{U}_d)^{\varepsilon_z}}.$$

In sum, the probability of workers choosing a residential district d^r and a business district d^w is given by

$$\Pr(d^r, d^w) = \frac{(\mathcal{U}_{d^r})^{\varepsilon_z}}{\sum_d (\mathcal{U}_d)^{\varepsilon_z}} \cdot \frac{\left(\bar{P}_{d^w}^{-\mu_c^w} \xi_{d^w} w_{d^w} \exp(-\kappa d_{d^r d^w} - \varphi_\kappa \mathbb{1}_{d^r \neq d^w})\right)^{\varepsilon_v}}{\sum_d \left(\bar{P}_d^{-\mu_c^w} \xi_d w_d \exp(-\kappa d_{d^r d} - \varphi_\kappa \mathbb{1}_{d^r \neq d})\right)^{\varepsilon_v}}.$$

From the discussion so far, the spatial distribution of workers is determined by

$$M_{i^r i^w} = M \cdot \Pr(i^r, i^w | d^r, d^w) \cdot \Pr(d^r, d^w), \quad M_{i^r}^r = \sum_{i^w} M_{i^r i^w}, \quad \text{and} \quad M_{i^w}^w = \sum_{i^r} M_{i^r i^w}.$$

Lastly, average income of consumers who work in zone i^w is $I_{i^w} = \mathbb{E}[w_{d(i^w)} v_{d(i^w)}] = \Gamma\big(1 - \frac{1}{\varepsilon_v}\big) w_{d(i^w)}$, and the expected welfare of consumers in the city is $\bar{U} = \Gamma\big(1 - \frac{1}{\varepsilon_z}\big) \cdot \big(\sum_d (\mathcal{U}_d)^{\varepsilon_z}\big)^{1/\varepsilon_z}$.

Tradable Goods Sector. In each district d, there is a representative firm in the tradable goods sector that produces a homogeneous final good. This final good is freely traded within the city at the price $p^{tradable}$. The production technology combines labor and floor space and features constant returns to scale:

$$y_d = \theta^\theta (1-\theta)^{1-\theta} A_d^{tradable} (L_d^{tradable})^\theta (H_d^{tradable})^{1-\theta}$$

where $A_d^{tradable}$ is district-specific productivity, $L_d^{tradable}$ and $H_d^{tradable}$ are labor and floor space inputs respectively, and θ is the labor share. A representative firm decides how to combine inputs and how much to produce in a competitive manner. Perfect competition implies that marginal cost equals to the price, $p^{tradable} = w_{d^w}^\theta r_{d^w}^{1-\theta} / A_{d^w}^{tradable}$.

Labor Market. Firms in both the tradable goods and services sectors demand labor, while workers' workplace choices determine labor supply. The equilibrium wage clears the district-level labor markets

$$M_d^w = L_d^{tradable} + L_d^{services} \quad \text{for all } d.$$

Land Market. Floor space in each district, H_d, is supplied by the competitive construction sector using the production function, $K_d^\mu T_d^{1-\mu}$, where K_d is capital and T_d is land. We assume that the cost of capital r_K is exogenously determined in the world capital market, and the land price is determined in the local land market. Then, floor space supply, $H_d = \bar{H} T_d r_d^{\frac{\mu}{1-\mu}}$, increases in the floor space price r_d where $\bar{H} = (\mu / r_k)^{\frac{\mu}{1-\mu}}$ is a constant. Floor space can be used for residential purposes H_d^r, producing the tradable goods $H_d^{tradable}$, or producing consumption services $H_d^{service}$. We assume that there exist zone-specific land use regulations for services production. We summarize this regulation with the floor space wedge, ϱ_i. Then, the floor space price for services sector is given by $r_i^s = \varrho_i r_{d(i)}$. We assume zero net transfers to land usage of services sector on average, i.e., $\frac{1}{N_d} \sum_{i \in d} r_i^s = r_{d(i)}$ where N_d is the number of zones in district d. The land market clearing condition is given by

$$\bar{H} T_d r_d^{\frac{\mu}{1-\mu}} = H_d^r + H_d^{tradable} + H_d^{service}, \quad \text{for all } d$$

where

$$H_d^r = \frac{1}{r_d} \mu_\ell \sum_d \frac{M_{d^r d}}{M_{d^r}^r} I_d, \;\; H_d^{tradable} = \frac{1}{r_d} \frac{1-\theta}{\theta} w_d L_d^{tradable}, \;\; \text{and} \;\; H_d^{service} = \frac{\rho - 1}{\rho} \gamma \sum_{i \in d} \frac{1}{r_i^s} \left(\sum_{sd} R_{isd} \right).$$

Agglomeration. Following the literature, we assume that local productivity of the tradable goods, $A_d^{tradable}$, and residential amenity, B_d, feature externalities. In particular, they are increasing in local residential or working population density,

$$A_d^{tradable} = \bar{A}_d^{tradable} \cdot (M_d^w / T_d)^{\eta_A},$$
$$B_d = \bar{B}_d \cdot (M_d^r / T_d)^{\eta_B},$$

where $\bar{A}_d^{tradable}$ and $\bar{B}_d$ are exogenous fundamentals, and η_A and η_B are the degree of agglomeration. We assume there exists no spillovers across districts.

Equilibrium Conditions. We define the general equilibrium.

Definition B.2 (General Equilibrium). *Given exogenous values of productivity $\{\bar{A}_d^{tradable}\}$, amenity $\{\bar{B}_d\}$, land supply $\{T_d\}$, rent wedges $\{\varrho_i\}$, the total number of workers M, and distance $\{d(\cdot,\cdot)\}$,* general equilibrium *consists of the worker distribution $M_{ir\,iw}$, rent prices $\{r_d\}$, and wages $\{w_d\}$ such that (i) workers optimally choose their resident and workplace districts; (ii) firms in the tradable goods sector maximize their profits; (iii) floor space suppliers maximize their profits; (iv) all conditions for the services market equilibrium hold; and (v) all markets clear.*

B.5 Estimation Details

B.5.1 Model Inversion: Identifying Productivity

We can invert the model to obtain a mapping from the observed data to the unobserved variables. To formalize the idea, we divide the variables and parameters into four sets. The first set $\mathbb{P}$ contains the parameters of the model. The second set $\mathbb{S}_{\text{exo}}$ contains variables exogenous to our model. The third set $\mathbb{S}_{\text{PE}}$ contains variables that are endogenously determined in the services market equilibrium. Finally, the fourth set $\mathbb{S}_{\text{GE}}$ contains variables that are taken as given in the services market but are determined in the general equilibrium.

We further partition $\mathbb{P}$ into three subsets: $\mathbb{P}^{\text{cal}}$ for externally calibrated parameters, $\mathbb{P}^{\text{PE}}$ for parameters estimated from the services market equilibrium model, and $\mathbb{P}^{\text{GE}}$ for parameters calibrated or estimated from the general equilibrium model. Similarly, we split the exogenous variables in $\mathbb{S}_{\text{exo}}$ into $\mathbb{S}_{\text{exo}}^{\text{cal}}$, $\mathbb{S}_{\text{exo}}^{\text{PE}}$, and $\mathbb{S}_{\text{exo}}^{\text{GE}}$. Moreover, within each set $\mathbb{S}_{\text{PE}}$ and $\mathbb{S}_{\text{GE}}$, we further divide the variables into those that we can observe in the data, $\mathbb{S}_i^{\text{obs}}$, and unobserved variables $\mathbb{S}_i^{\text{unobs}}$.

$$\mathbb{P}^{\text{cal}} = \{\boldsymbol{\mu}, \boldsymbol{\alpha}, \rho, \gamma, \beta\} \qquad \mathbb{P}^{\text{PE}} = \{\tilde{\tau}, \tilde{\varphi}, \varepsilon, \sigma, \nu\} \qquad \mathbb{P}^{\text{GE}} = \{\kappa, \varepsilon_z, \varepsilon_v, \theta, \eta_A, \eta_B\}$$

$$\mathbb{S}^{\text{cal}}_{\text{exo}} = \{\mathcal{J}, \mathcal{S}, \boldsymbol{\mathcal{D}}, \boldsymbol{d}, \boldsymbol{T}\} \qquad \mathbb{S}^{\text{PE}}_{\text{exo}} = \{\tilde{\boldsymbol{A}}, \boldsymbol{C}\} \qquad \mathbb{S}^{\text{GE}}_{\text{exo}} = \{\bar{\boldsymbol{B}}, \bar{\boldsymbol{A}}^{tradable}, \boldsymbol{\xi}\}$$

$$\mathbb{S}^{\text{obs}}_{\text{PE}} = \{\boldsymbol{N}, \boldsymbol{R}\} \qquad \mathbb{S}^{\text{unobs}}_{\text{PE}} = \{\boldsymbol{P}, \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{H}^{service}, \boldsymbol{L}^{service}, \Pi\}$$

$$\mathbb{S}^{\text{obs}}_{\text{GE}} = \{\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{M}, \boldsymbol{I}, \boldsymbol{E}, M\} \qquad \mathbb{S}^{\text{unobs}}_{\text{GE}} = \{\boldsymbol{H}^{r}, \boldsymbol{H}^{tradable}, \boldsymbol{L}^{tradable}\}$$

where the bold letters denote vectors. The services market equilibrium model is essentially a mapping from $\mathbb{P}^{\text{cal}} \cup \mathbb{P}^{\text{PE}} \cup \mathbb{S}^{\text{cal}}_{\text{exo}} \cup \mathbb{S}^{\text{PE}}_{\text{exo}} \cup \mathbb{S}^{\text{obs}}_{\text{GE}}$ to $\mathbb{S}_{\text{PE}}$, and the general equilibrium model is a mapping from $\mathbb{P} \cup \mathbb{S}_{\text{exo}}$ to $\mathbb{S}_{\text{PE}} \cup \mathbb{S}_{\text{GE}}$. The following lemma shows that we can invert these mappings to back out location characteristics from the observed data.

Lemma B.8 (Equilibrium Inversion). *Given $\mathbb{P}^{cal}$, $\mathbb{P}^{PE}$, $\mathbb{S}^{cal}_{exo}$, and $\mathbb{S}^{obs}_{GE}$, there exist unique values of the location characteristics $\mathbb{S}^{PE}_{exo}$ that rationalize the observed data $\mathbb{S}^{obs}_{PE}$. Given $\mathbb{P}$, $\mathbb{S}^{cal}_{exo}$, and $\mathbb{S}^{PE}_{exo}$, there exist unique values of the location characteristics $\mathbb{S}^{GE}_{exo}$ that rationalize the observed data $\mathbb{S}^{obs}_{PE}$ and $\mathbb{S}^{obs}_{GE}$.*

The above lemma is a direct application of Proposition 2 in Ahlfeldt et al. (2015). The first part of Lemma B.8 implies that, once we have data on $\mathbb{S}^{\text{obs}}_{\text{GE}}$, we can back out the composite productivity and operating costs without relying on the general equilibrium component of the model. In Section 2.4.1, we use this lemma to estimate the services market equilibrium parameters, $\mathbb{P}^{\text{PE}}$, separately from the general equilibrium parameters, $\mathbb{P}^{\text{GE}}$. The advantage of this approach is that we can estimate the services market equilibrium model without explicitly specifying the general equilibrium structure, as long as we have access to data on $\mathbb{S}^{\text{obs}}_{\text{GE}}$. Thus, the estimation procedure remains robust to any potential misspecification of our general equilibrium model.

B.5.2 Parameter Estimation: General Equilibrium Model

In this section, we discuss the estimation procedure of general equilibrium.

We calibrate the GE parameters mostly from the literature or directly from the moments in the data. We summarize the parameter values in Table B.5.

We estimate commuting disutility parameters by running a gravity equation using commuting flows from Household Travel Survey. To do this, we first aggregate the data to the district level and then employ the PPML estimator with border effects, similar to our estimation approach for services travel parameters.

Next, we calculate the total expenditure share on services, $\mu^r_c + \mu^w_c$, using the total revenue of services sectors and the total income. We first rescale the total revenue in 2019 from the Commercial District dataset. Although the dataset covers most of the regions of Seoul, it is not universal since it only includes the stores in the commercial areas. Comparing the total number of restaurants in Seoul in 2017, our dataset includes 96.28% of restaurants in Seoul (Seoul Business Survey, 2017). Thus, we increase the total revenue in 2019 by 3.86% and divide it with the total population to get the average monthly total services revenue per capita. We take the average income per capita in

Table B.5: Calibration Results: GE

Parameter	Description	Value	Source
$\kappa\varepsilon_v$	Commuting Travel Elasticity	0.0605	Gravity
φ_κ	Commuting Travel Border Effects	1.266	Gravity
$\mu_c^r + \mu_c^w$	Total services spending share	0.27	Total revenue, Income[a]
μ_ℓ	Housing spending share	0.25	Davis and Ortalo-Magné (2011)
$1-\theta$	Share on floor space of tradable goods	0.2	Valentinyi and Herrendorf (2008)
$1-\mu$	Share on land of floor space production	0.25	Combes, Duranton, and Gobillon (2012)
$\varepsilon_v, \varepsilon_z$	Preference scale	6, 6	Ahlfeldt et al. (2015)
η_A, η_B	Agglomeration	0.07, 0.15	Ahlfeldt et al. (2015)[b]

[a] Source: Seoul Commercial Area Data and Statistics of Korea.
[b] Tsivanidis (2019) estimates $\eta_A = 0.212$ and $\eta_B \in [0.419, 0.576]$, which are larger than the estimates obtained by Ahlfeldt et al. (2015) using a German dataset. As Korea is a developed country, we take the results of Ahlfeldt et al. (2015).

Seoul from Statistics Korea (2019). We then adjust the average income to take into account that 50% of households in Seoul live in their own house and do not pay rents (Korea Housing Survey, 2019). Thus, we scale up their income by $\frac{1}{1-\mu_\ell}$. Dividing the revenue per capita with the average income, we obtain that the spending share on services, $\mu_c^r + \mu_c^w$, equals 27%.

We use residential population data at the zone level provided by Seoul Metropolitan Government (2019) and calculate the conditional probability $\Pr^r(i^r|d^r)$ for each district. Similarly, we use Seoul Business Survey which provides the total number of workers of each zone and calculate the conditional probability $\Pr^w(i^w|d^w)$ for each district.

B.6 SMA and (Real) Income Inequality

In this section, we show that SMA inequality worsens real income inequality and the magnitude of its impact is larger than that of housing rents. We define high-skilled workers as college graduates. We use income data from Korean Labor Panel (2019) and focus on workers who live in Seoul with positive income.

Based on the equation (3), we can define real incomes of workers who live in district d^r as below.

$$\text{real income} = \text{nominal income} \cdot \text{rent}_{d^r}^{-\mu_\ell} \cdot \text{SMA}_{d^r}^{\mu_c^r + \mu_c^w}.$$

In the first column of Table B.6, we first report the college premium on nominal income. On average, high-skilled workers earn 47.7% higher income than low-skilled workers. It is smaller than that of US, which is about 79% (BLS,

Table B.6: College Premium

	Income	Income, Rent	Income, Rent, SMA
Low-skilled	220.7	221.3	220.77
High-skilled	325.9	324.8	325.65
College premium	47.7%	46.1%	47.5%

Notes: Data source: Korean Labor Panel (2019). We drop outliers with the top and the bottom 1% of income to remove noises in data.

2020), but still the magnitude is very large. In the next column, we report the college premium after adjusting the differences in housing rents. College graduates who earn higher income tend to live in regions with higher housing rents. Thus, the college premium decreases to 46.1%.[23] Finally, in the last column of Table B.6, we compute real income, which additionally adjusts the differences in SMA—the inverse of the price index of services.[24] The college premium increases from 46.1% to 47.5%. High-skilled workers live in regions with higher housing rents, but at the same time, they enjoy higher SMA. This better access to the services market widens the real income gap between high- and low-skilled workers.

It is striking that effects of SMA inequality on real income gap is as important as those of rents dispersion. Most of the literature has focused on how housing prices affect spatial inequality, but the result shows that services markets are as important as housing markets. Although some studies consider the implications of non-tradable goods, they typically focus on the price dispersion across cities. In this case, the price index for non-tradable goods is typically higher in urban areas, dampening the real income dispersion.[25] On the contrary, we claim that SMA differentials exacerbate income inequality within cities as high skilled workers can enjoy better access to the services market.

B.7 Survey Questions

This survey is for the study of consumers' offline consumption spending behavior. Please carefully read the explanation below before taking the survey. In this survey, the term "purchase" refers to spending on the following three categories, made through in-person transactions (i.e., excluding online shopping).

1. Foods: restaurant, Café, bakery, bars, etc.

[23] One limitation is that we only have residence information at the district level. This can potentially lead to downward bias to the importance of housing rents.

[24] We compute district-level SMA by taking an average using the population distribution,

$$\text{SMA}_{d^r} = \sum_{i^r, i^w} \Pr(i^r, i^w | d^r) \text{SMA}_{i^r}^{\frac{\mu_c^r}{\mu_c^r + \mu_c^w}} \text{SMA}_{i^w}^{\frac{\mu_c^w}{\mu_c^r + \mu_c^w}}.$$

[25] One exception is Handbury and Weinstein (2015) who focus on heterogeneity and the variety of goods. They argue that large cities have wider variety of goods which can contribute to a lower price index for food products in large cities.

2. Retail: convenience stores, groceries, clothing, shoes, cosmetics, books, furniture, home appliances, gas stations, etc.
3. Other services: gym, beauty salon, skincare, car repair, laundry, billiard room, golf practice center, etc.

We will ask you about the number of times you went out during the day and the stores you visited for each trip. Please count the number of stores you made a purchase for an instance of travel as follows.

- Buying coffee at a Café on the way back from a restaurant near your workplace at lunchtime: 2 stores
- Visiting a department store, buying three items of clothing at the first clothing store and vising a second shoe store without making a purchase, then buying lotion at the cosmetics section. After that, buying groceries for the week at a supermarket when returning home: 3 stores
- On the way to work from home, stopping at a convenience store for buying snacks and drinks: 1 store

If you have already returned to your home, work, or your next destination unrelated to your purchase, and have gone out again on the same day after completing the travel, please write it as separate travel.

[Personal Information]

Q. What is your gender? a. Male b. Female c. Others, or refuse to answer

Q. What is the year of your birth?

Questions 1 to 4 are about basic personal information.

1. What is your final educational background? (If you are currently a student, respond to the educational institution you are attending) a. Elementary/middle/high school b. College and above
2. What is your occupation?
   a. Office/expert/manager (e.g., teachers, public officials)
   b. Services/sales/function/agriculture/other (e.g., cook, hairdresser, sales profession)
   c. Looking for a job/unemployment, etc.
   d. Housewives
   e. Students
3. Which “Dong” does your home belong to?
4. Which “Dong” does your workplace or school belong to?

[Offline Consumption Expenditure]

Questions 5 to 8 are questions about offline consumption expenditure during the week.

5. Have you been out to eat, shop, or buy other services during the most recent Thursday (except today)? If you don’t, please choose a day when you were out.

a. Yes. I went out on Thursday.

b. No. Not on Thu, but on Wednesday.

c. No. Not on Wed/Thu, but on Tuesday.

d. No. Not on Tue/Wed/Thu, but on Monday.

e. No. Not on Mon/Tue/Wed/Thu, but on Friday.

f. No, I haven't been out for purchase in the past week (please recall the most recent weekday outing).

6. How many purchases did you make in total during the first trip on the day of the week selected in question 5? Please write down the total number of stores you visited for making any (non-zero) purchases.

7. How many purchases did you make in total during the second trip on the day of the week selected in question 5? Please write down the total number of stores you visited for making any (non-zero) purchases.

8. How many purchases did you make in total during the third trip on the day of the week selected in question 5? Please write down the total number of stores you visited for making any (non-zero) purchases.

9. How many purchases did you make in total during the fourth trip on the day of the week selected in question 5? Please write down the total number of stores you visited for making any (non-zero) purchases.

From now on, please answer about the travel with the highest total number of purchases (select outings in front of you in case of redundancy) among responses in questions 6-8 above.

10. Please answer the location and types of the departure area where you started the travel.

a. "Dong" b. Home/School/Workplace/Others

11. Please answer the location and types of the destination where you ended the travel.

a. "Dong" b. Home/School/Workplace/Others

12. Please write down about your first purchase of travel.

a. Location: "Dong"

b. Purpose: (1) restaurants (2) shopping/retail (3) other services

c. Details (e.g., Chinese restaurants, convenience stores, bookstores, car repairs, etc.):

13. Is there any additional purchase after the first purchase for this travel? If so, please respond.

a. Travel ended after the first purchase.

b. Location: "Dong"

c. Purpose: (1) restaurants (2) shopping/retail (3) other services

d. Details (e.g., Chinese restaurants, convenience stores, bookstores, car repairs, etc.):

From now on, the same question is about going out on weekends, not weekdays. Please respond in the same way as before.

14. Have you been out to eat, shop, or buy other services during the most recent Saturday (except today)? If you don't have one, please write down your travel on Sunday or the previous weekend.

a. Yes. I went out on Saturday.

b. No. Not on Thu, but on Sunday.

c. No. I haven't been out on Saturday/Sunday most recently, but I went out on Saturday/Sunday the week before.

d. No. I haven't been out on weekends for the last two weeks (please recall the most recent weekend outing).

15. How many purchases did you make in total during the first trip on the day of the weekend selected in question 14? Please write down the total number of stores you visited for making any (non-zero) purchases.

16. How many purchases did you make in total during the second trip on the day of the weekend selected in question 14? Please write down the total number of stores you visited for making any (non-zero) purchases.

17. How many purchases did you make in total during the third trip on the day of the weekend selected in question 14? Please write down the total number of stores you visited for making any (non-zero) purchases.

18. How many purchases did you make in total during the fourth trip on the day of the weekend selected in question 14? Please write down the total number of stores you visited for making any (non-zero) purchases.

From now on, please answer about the travel with the highest total number of purchases (select outings in front of you in case of redundancy) among responses in questions 6-8 above.

19. Please answer the location and types of the departure area where you started the travel.

a. "Dong" b. Home/School/Workplace/Others

20. Please answer the location and types of the destination where you ended the travel.

a. "Dong" b. Home/School/Workplace/Others

21. Please write down about your first purchase of travel.

a. Location: "Dong"

b. Purpose: (1) restaurants (2) shopping/retail (3) other services

c. Details (e.g., Chinese restaurants, convenience stores, bookstores, car repairs, etc.):

22. Is there any additional purchase after the first purchase for this travel? If so, please respond.

a. Travel ended after the first purchase.

b. Location: "Dong"

c. Purpose: (1) restaurants (2) shopping/retail (3) other services

d. Details (e.g., Chinese restaurants, convenience stores, bookstores, car repairs, etc.):

Appendix C

Persistent Noise, Feedback, and Endogenous Optimism: A Rational Theory of Overextrapolation

C.1 Proofs for Sections 3.2–3.4

Proof of Lemma 1. Signals of the form $s_t = a_t + \sigma \cdot \tilde{\xi}_t$ give a strictly positive object function unless $\sigma = 0$, which violates the information constraint. Consider a signal of the form $s_t = a_t + \sigma \cdot \tilde{\xi}$ where σ is sufficiently high to satisfy the information constraint. Since we know unconditional distribution of a_t, with an infinite number of realizations of s_t, we can exactly learn the realization of $\tilde{\xi}$, thereby having $\mathbb{E}[a_t|s^t] = a_t$ for all t. □

Proof of Lemma 2. We have

$$\left(1 - \frac{1}{\eta} - \frac{1}{\theta}\right) y_{ijt} = \mathbb{E}_{it}\left[-\frac{1}{\eta} y_t + \gamma y_t - \frac{1}{\theta} a_{it} + (1 - \frac{\sigma}{\eta})\frac{1}{\sigma}(y_{ijt} - y_t)\right]$$

or

$$\left(1 - \frac{1}{\sigma} - \frac{1}{\theta}\right) y_{ijt} = \mathbb{E}_{it}\left[\left(\gamma - \frac{1}{\sigma}\right) y_t - \frac{1}{\theta} a_{it}\right]$$

or equivalently

$$\begin{aligned} y_{ijt} &= \left(\frac{1}{\theta} + \frac{1}{\sigma} - 1\right)^{-1} \mathbb{E}_{it}\left[\frac{1}{\theta} a_{it} + \left(\frac{1}{\sigma} - \gamma\right) y_t\right] \\ &\equiv \mathbb{E}_{it}[(1 - \alpha)\tilde{a}_{it} + \alpha y_t] \end{aligned}$$

which gives the desired result. □

Proof of Proposition 1. This result can be seen as a special case of Theorem 1 with $\rho = 0$. □

Proof of Lemma 3. Right before observing s_{it}, we have $\xi_{it-1}|\tilde{\Omega}_{it} \sim \mathcal{N}(m_{it-1}, V_{t-1})$, hence

$$\xi_{it}|\tilde{\Omega}_{it} \sim \mathcal{N}\big(\rho m_{it-1}, \rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2\big).$$

On the other hand, the prior belief of $\tilde{a}_{it} \equiv \rho_a a_{it-1} + \varepsilon_{it}^p$ is $\mathcal{N}\big(\rho_a a_{it-1}, \sigma_p^2\big)$. Thus, Bayesian updating gives

$$\mathbb{E}_{it}[a_{it}] = \mathbb{E}_{it}[\tilde{a}_{it}] = \rho_a \cdot a_{it-1} + K_t(s_{it} - \rho_a a_{it-1} - \rho m_{it-1})$$

where $K_t = \frac{\sigma_p^2}{\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2 + \sigma_p^2} \in (0,1)$. Also note that

$$s_{it} - \rho_a a_{it-1} - \rho m_{it-1} = \varepsilon_{it}^p + \xi_{it} - \rho m_{it-1} = \varepsilon_{it}^p + \tilde{\mathcal{O}}_{it}.$$

Finally, we have

$$\mathbb{E}_{it}[a_{it}] = \mathbb{E}_{it}[\tilde{a}_{it}] = \mathbb{E}_{it}[s_{it} - \xi_{it}] = s_{it} - \mathbb{E}_{it}[\xi_{it}] = \rho_a a_{it-1} + \varepsilon_{it}^p + \mathcal{O}_{it}.$$ □

Proof of Lemma 4. Consider the following state-space representation.

$$\boldsymbol{x}_t \equiv \begin{pmatrix} \tilde{a}_{it} \\ \xi_{it} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \rho_a a_{it-1} \\ \rho m_{t-1} \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma_p^2 & 0 \\ 0 & \rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2 \end{pmatrix}}_{\Sigma} \right)$$

$$\boldsymbol{y}_t \equiv \begin{pmatrix} s_{it} \\ a_{it} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}}_{G} \boldsymbol{x}_t + \begin{pmatrix} 0 \\ e_t^u \end{pmatrix}$$

The Kalman filter gives

$$\boldsymbol{x}_t|\boldsymbol{y}_t \sim \mathcal{N}\left(\begin{pmatrix} \rho_a a_{it-1} \\ \rho m_{t-1} \end{pmatrix} + K\left(\boldsymbol{y}_t - G\begin{pmatrix} \rho_a a_{it-1} \\ \rho m_{t-1} \end{pmatrix} \right), KRK' + (I - KG)\Sigma(I - KG)' \right)$$

where $K = \Sigma G'(G\Sigma G + R)^{-1}$ and $R = \text{Var}\begin{pmatrix} 0 \\ \varepsilon_t^u \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}$. This gives $\xi_{it}|\tilde{\Omega}_{it+1} \sim \mathcal{N}(m_{it}, V_t)$ with

$$m_{it} = (\gamma_1 + \gamma_2)s_{it} + \gamma_3 \rho m_{it-1} - \rho_1 \rho_a a_{it-1} - \gamma_2 a_{it}$$
$$V_t = \gamma_3(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2)$$

where

$$\gamma_1 = \frac{\sigma_u^2(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2)}{(\sigma_p^2+\sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}$$
$$\gamma_2 = \frac{\sigma_p^2(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2)}{(\sigma_p^2+\sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}$$
$$\gamma_3 = \frac{\sigma_p^2\sigma_u^2}{(\sigma_p^2+\sigma_u^2)(\rho^2 V_{t-1} + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}$$

Thus, $\gamma_1, \gamma_2, \gamma_3 \in (0,1)$ and $\gamma_1 + \gamma_2 + \gamma_3 = 1$. □

Proof of Proposition 2. The law of motion for ex-ante optimism directly follows from Lemma 4. In the proof of Lemma 3, we have shown that

$$\mathcal{O}_{it} = K\tilde{\mathcal{O}}_{it} - (1-K)\varepsilon_{it}^p,$$

which gives the law of motion for ex-post optimism. □

Proof of Lemma 5. We have

$$V = \frac{\sigma_p^2\sigma_u^2(\rho^2 V + (1-\rho^2)\sigma_\eta^2)}{(\sigma_p^2+\sigma_u^2)(\rho^2 V + (1-\rho^2)\sigma_\eta^2) + \sigma_p^2\sigma_u^2}$$

or equivalently

$$\frac{1}{V} - \frac{1}{\rho^2 V + (1-\rho^2)\sigma_\eta^2} = \frac{1}{\sigma_p^2} + \frac{1}{\sigma_u^2}.$$

Since $\gamma_1 = \frac{V}{\sigma_p^2}$, we can write

$$\frac{1}{\gamma_1} - \frac{1}{\rho^2\gamma_1 + (1-\rho^2)\frac{\sigma_\eta^2}{\sigma_p^2}} = 1 + \frac{\sigma_p^2}{\sigma_u^2}.$$

The left hand side is then increasing in σ_η^2 and decreasing in σ_p^2. Moreover, as it can be alternatively written as

$$\frac{(1-\rho^2)\left(\frac{\sigma_\eta^2}{\sigma_u^2} - \gamma_1\right)}{\gamma_1\left(\rho^2\gamma_1 + (1-\rho^2)\frac{\sigma_\eta^2}{\sigma_p^2}\right)},$$

the left hand side is also decreasing in γ_1. Therefore, we can conclude that γ_1 is increasing in σ_u^2 and σ_η^2 while decreasing in σ_p^2. In a similar way, we can show that γ_2 is increasing in σ_p^2 and σ_η^2 while decreasing in σ_u^2. For the comparative statics for γ_3, define $W = \rho^2 V + (1-\rho^2)\sigma_\eta^2$. This implies $V = \rho^{-2}W - (\rho^{-2}-1)\sigma_\eta^2$ and

$\gamma_3 \equiv \frac{V}{W} = \rho^{-2} - (\rho^{-2}-1)\frac{\sigma_\eta^2}{W}$. The last term $\frac{\sigma_\eta^2}{W}$ satisfies

$$\frac{(\rho^{-2}-1)\left(1-\frac{W}{\sigma_\eta^2}\right)}{\frac{W}{\sigma_\eta^2}\left(\rho^{-2}\frac{W}{\sigma_\eta^2}-(\rho^{-2}-1)\right)} = \frac{\sigma_\eta^2}{\sigma_p^2}+\frac{\sigma_\eta^2}{\sigma_u^2}.$$

The left hand side is increasing in $\frac{\sigma_\eta^2}{W}$. Thus, γ_3 is decreasing in σ_η^2 while increasing in σ_p^2 and σ_u^2. □

Proof of Theorem 1. From Proposition 2 and the definition of optimism, we have

$$\begin{aligned} y_{it+1} &= \rho_a a_{it} + K(s_{it+1} - \rho_a a_{it} - \rho m_{it}) \\ &= \rho_a a_{it} + K(\varepsilon_{it+1}^p + \xi_{it+1} - \rho m_{it}) \\ &= \rho_a^2 a_{it-1} + (\rho_a - \rho K\gamma_1)\varepsilon_{it}^p + (\rho_a + \rho K\gamma_2)\varepsilon_{it}^u + K\varepsilon_{it+1}^p + K\eta_{it+1} + \rho\gamma_3 K\tilde{\mathcal{O}}_{it}. \end{aligned}$$

We can use the relationship between ex-ante and ex-post optimism to derive the second result. □

Proof of Lemma 6. Suppose that all agents except for i use a strategy of the form $y_{j0} = \theta s_{j0}$. Then, we can calculate the best response of i as

$$\begin{aligned} y_{i0} &= (1-\alpha)\,\mathbb{E}_{i0}[\varepsilon_0^p] + \alpha\,\mathbb{E}_{i0}[y_0] \\ &= (1-\alpha+\alpha\theta)\,\mathbb{E}_{i0}[\varepsilon_0^p] \\ &= (1-\alpha+\alpha\theta)\frac{\sigma_p^2}{\sigma_p^2+\sigma_\xi^2}s_{i0}. \end{aligned}$$

Thus, the unique linear equilibrium is given by $y_{i0} = \theta s_{i0}$ where $\theta = \frac{(1-\alpha)\sigma_p^2}{(1-\alpha)\sigma_p^2+\sigma_\xi^2} \in (0,1)$. For the general uniqueness, see Morris and Shin (2002). □

Proof of Lemma 7. Suppose that all agents except for i use a strategy of the form $y_{j1} = \theta_1 s_{j0} + \theta_2 a_0 + \theta_3 s_{j1}$, then these decisions aggregate into

$$y_1 = \theta_1\varepsilon_0^p + \theta_2 a_0 + \theta_2\varepsilon_1^p.$$

Thus, the the best response of i in period 1 is given by

$$\begin{aligned} y_{i1} &= (1-\alpha)\,\mathbb{E}_{i1}\,\varepsilon_1^p + \alpha\,\mathbb{E}_{i1}\,y_1 \\ &= (1-\alpha+\alpha\theta_3)\,\mathbb{E}_{i1}\,\varepsilon_1^p + \alpha\theta_2 a_0 + \alpha\theta_1\,\mathbb{E}_{i1}\,\varepsilon_0^p. \end{aligned}$$

Consider the following state-space representation.

$$\boldsymbol{x} \equiv \begin{pmatrix} \varepsilon_0^p \\ \varepsilon_1^p \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma) \quad \text{where } \Sigma = \begin{pmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_p^2 \end{pmatrix}$$
$$\boldsymbol{y} \equiv \begin{pmatrix} s_{i0} \\ a_0 \\ s_{i1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{G} \begin{pmatrix} \varepsilon_0^p \\ \varepsilon_1^p \end{pmatrix} + \begin{pmatrix} \xi_i \\ \varepsilon_0^u \\ \xi_i \end{pmatrix}.$$

The Kalman filter gives

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = K\boldsymbol{y} \quad \text{where } K = \Sigma G'(G\Sigma G' + R)^{-1} \quad \text{with } R = \begin{pmatrix} \sigma_\xi^2 & 0 & \sigma_\xi^2 \\ 0 & \sigma_u^2 & 0 \\ \sigma_\xi^2 & 0 & \sigma_\xi^2 \end{pmatrix}.$$

Thus, we have

$$y_{i1} = \alpha\theta_1 \begin{pmatrix} 1 & 0 \end{pmatrix} K\boldsymbol{y} + (1 - \alpha + \alpha\theta_3) \begin{pmatrix} 0 & 1 \end{pmatrix} K\boldsymbol{y} + \alpha\theta_2 a_0$$

Matching coefficient, we have

$$\theta_1 = \begin{pmatrix} \alpha\theta_1 & 1 - \alpha + \alpha\theta_3 \end{pmatrix} K \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$
$$\theta_2 = \frac{1}{1-\alpha} \begin{pmatrix} \alpha\theta_1 & 1 - \alpha + \alpha\theta_3 \end{pmatrix} K \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$
$$\theta_3 = \begin{pmatrix} \alpha\theta_1 & 1 - \alpha + \alpha\theta_3 \end{pmatrix} K \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Let $B = \begin{pmatrix} \alpha\theta_1 & 1 - \alpha + \alpha\theta_3 \end{pmatrix}$, then we can obtain B by

$$B = \begin{pmatrix} 0 & 1 - \alpha \end{pmatrix} + BK \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \alpha & 0 \end{pmatrix} + BK \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & \alpha \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 1 - \alpha \end{pmatrix} \left(I - K \begin{pmatrix} \alpha & 0 \\ 0 & 0 \\ 0 & \alpha \end{pmatrix} \right)^{-1},$$

which in turn gives the values for θ_1, θ_2 and θ_3. Thus, we can write

$$y_1 = \theta_1 \varepsilon_0^p + \theta_2(\varepsilon_0^p + \varepsilon_0^u) + \theta_3 \varepsilon_1^p \equiv \gamma_p \varepsilon_0^p + \gamma_u \varepsilon_0^u + \gamma_p' \varepsilon_1^p$$

where

$$\begin{aligned}
\gamma_p &= -\frac{\sigma_u^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2} \\
\gamma_u &= \frac{\sigma_p^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2} \\
\gamma_p' &= \frac{(1-\alpha)\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + \sigma_u^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2}.
\end{aligned}$$

□

Proof of Lemma 8. From the proof of Lemma 7, we can get

$$\begin{aligned}
\mathbb{E}_{it}[\xi_i] &\equiv \mathbb{E}\left[\xi_i \middle| \begin{pmatrix} s_{i0} \\ a_0 \\ s_{i1} \end{pmatrix}\right] = \mathbb{E}\left[s_{i0} - \varepsilon_0^p \middle| \begin{pmatrix} s_{i0} \\ a_0 \\ s_{i1} \end{pmatrix}\right] \\
&= \left[\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 1 & 0 \end{pmatrix} K\right] \begin{pmatrix} s_{i0} \\ a_0 \\ s_{i1} \end{pmatrix} \\
&= \underbrace{\left[\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 1 & 0 \end{pmatrix} K\right] \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}}_{L} \boldsymbol{z}_i \quad \text{where } \boldsymbol{z}_i \equiv \begin{pmatrix} \varepsilon_0^p \\ \xi_i \\ \varepsilon_0^u \\ \varepsilon_1^p \end{pmatrix}.
\end{aligned}$$

Note first that

$$\mathcal{O}_{i1} \equiv \xi_i - \mathbb{E}_{i1}\,\xi_i = \underbrace{\left[\begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} - L\right]}_{Q} \boldsymbol{z}_i.$$

Also note that

$$\mathbb{E}_{i1}\left[\int_0^1 \boldsymbol{z}_j\, dj\right] = \mathbb{E}_{i1}\left[\begin{pmatrix} \varepsilon_0^p \\ 0 \\ \varepsilon_0^u \\ \varepsilon_1^p \end{pmatrix}\right] = \mathbb{E}_{i1}\left[\begin{pmatrix} s_{i0} - \xi_i \\ 0 \\ a_0 - s_{i0} + \xi_i \\ s_{i1} - \xi_i \end{pmatrix}\right]$$

$$= \underbrace{\begin{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} - L \\ \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & -1 & 1 & 0 \end{pmatrix} + L \\ \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix} - L \end{pmatrix}}_{T} \boldsymbol{z}_i.$$

Suppose that $\mathcal{O}_{i1}^{h-1} = QT^{h-2}\boldsymbol{z}_i$ holds, which indeed holds for $h = 2$. Then, we have

$$\begin{aligned} \mathcal{O}_{i1}^{h} &= \mathbb{E}_{i1}\left[QT^{h-2}\int_0^1 \boldsymbol{z}_j \, dj\right] \\ &= QT^{h-1}\boldsymbol{z}_i. \end{aligned}$$

Thus, we can inductively show that

$$\mathcal{O}_{i1}^{h} = QT^{h-1}\boldsymbol{z}_i.$$

After some algebra, we can write Q and QT^{h-1} as functions of underlying parameters:

$$Q = \begin{pmatrix} -\frac{\sigma_u^2\,\sigma_\xi^2}{\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} & \frac{\sigma_p^2\,\sigma_u^2}{\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} & \frac{\sigma_p^2\,\sigma_\xi^2}{\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} & -\frac{\sigma_u^2\,\sigma_\xi^2}{\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} \end{pmatrix}$$

$$QT^{h-1} = \begin{pmatrix} -\frac{(\sigma_p^2)^{h-1}\,(\sigma_u^2)^{h}\,\sigma_\xi^2}{\left(\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2\right)^h} & -\frac{(\sigma_p^2)^{h-1}\,(\sigma_u^2)^{h-1}\,\sigma_\xi^2\left(\sigma_p^2+2\,\sigma_u^2\right)}{\left(\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2\right)^h} & \frac{(\sigma_p^2)^{h}\,(\sigma_u^2)^{h-1}\,\sigma_\xi^2}{\left(\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2\right)^h} & -\frac{(\sigma_p^2)^{h-1}\,(\sigma_u^2)^{h}\,\sigma_\xi^2}{\left(\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2\right)^h} \end{pmatrix}.$$

Note that

$$\begin{aligned} Q_{1,2} &= \frac{\sigma_p^2\,\sigma_u^2}{\sigma_p^2\,\sigma_u^2+\sigma_p^2\,\sigma_\xi^2+2\,\sigma_u^2\,\sigma_\xi^2} \\ &= \frac{1}{1+\frac{\sigma_\xi^2}{\sigma_u^2}+2\frac{\sigma_\xi^2}{\sigma_p^2}} \end{aligned}$$

is decreasing in σ_ξ^2 and increasing in σ_u^2 and σ_p^2. □

Proof of Lemma 9. First, we have

$$\mathbb{E}_{i1}[a_1] = \mathbb{E}_{i1}[\varepsilon_{i1}^p] = \mathbb{E}_{i1}[s_{i1} - \xi_i] = \varepsilon_{i1}^p + \xi_i - \mathbb{E}_{i1}[\xi_i] = \varepsilon_{i1}^p + \mathcal{O}_{i1}.$$

Thus,

$$\bar{\mathbb{E}}_1\varepsilon_1^p = \bar{\mathcal{O}}_1.$$

Suppose that we have

$$\mathbb{E}_{i1}\,\bar{\mathbb{E}}_1^{h-1}[a_1] = \varepsilon_1^p + \mathcal{O}_{i1} + \mathcal{O}_{i1}^2 + \cdots + \mathcal{O}_{i1}^{h-1}$$
$$\bar{\mathbb{E}}_1^{h}[a_1] = \varepsilon_1^p + \bar{\mathcal{O}}_1 + \bar{\mathcal{O}}_1^2 + \cdots + \bar{\mathcal{O}}_1^{h}$$

for a given h. Then, we can obtain

$$\begin{aligned}\mathbb{E}_{i1}\,\bar{\mathbb{E}}_1^{h}[a_1] &= \mathbb{E}_{i1}\left[\varepsilon_1^p + \bar{\mathcal{O}}_1 + \bar{\mathcal{O}}_1^2 + \cdots + \bar{\mathcal{O}}_1^{h}\right] \\ &= \varepsilon_1^p + \mathcal{O}_{i1} + \mathcal{O}_{i1}^2 + \cdots + \mathcal{O}_{i1}^{h+1}.\end{aligned}$$

hence

$$\bar{\mathbb{E}}_1^{h+1}[a_1] = \varepsilon_1^p + \bar{\mathcal{O}}_1 + \bar{\mathcal{O}}_1^2 + \cdots + \bar{\mathcal{O}}_1^{h+1}.$$

Thus, we can inductively show Lemma 9. □

Proof of Lemma 10. We have shown in the proof of Lemma 8 that

$$\frac{\partial \bar{\mathcal{O}}_1^h}{\partial \varepsilon_0^u} = \frac{(\sigma_p^2)^h\,(\sigma_u^2)^{h-1}\,\sigma_\xi^2}{\left(\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2\right)^h}, \quad \text{for all } h \geq 1$$

and that

$$\frac{\partial \bar{\mathcal{O}}_1^h}{\partial \varepsilon_0^p} = -\frac{(\sigma_p^2)^{h-1}\,(\sigma_u^2)^{h}\,\sigma_\xi^2}{\left(\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2\right)^h}, \quad \text{for all } h \geq 1.$$

After some algebra, we can obtain the results. □

Proof of Lemma 11. Recall that, in the proof of Lemma 7, we have

$$y_1 = \gamma_p \varepsilon_0^p + \gamma_u \varepsilon_0^u + \gamma_p' \varepsilon_1^p$$

where

$$\gamma_p = -\frac{\sigma_u^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2}$$
$$\gamma_u = \frac{\sigma_p^2\,\sigma_\xi^2}{(1-\alpha)\sigma_p^2\,\sigma_u^2 + \sigma_p^2\,\sigma_\xi^2 + 2\,\sigma_u^2\,\sigma_\xi^2}$$

Thus, the effect of ε_0^u on y_1 (i.e., γ_u) is increasing in σ_p^2, decreasing in σ_u^2, and increasing in σ_ξ^2. Likewise, the effect of ε_0^p on y_1 (i.e., $|\gamma_p|$) is increasing in σ_u^2, decreasing in σ_p^2, and increasing in σ_ξ^2. □

C.2 Details of the Numerical Exercise

We utilize the method of Woodford (2003) to solve for the equilibrium dynamics of the aggregate output, which exploits the fact that firms only need to track particular linear combinations of higher-order beliefs. The absence of endogenous signals permits us to do so; see Huo and Takayama (2015). We start from a guess that the relevant aggregate state can be summarized in

$$\mathbf{x}_t = \begin{pmatrix} \varepsilon_t^p & \varepsilon_t^u & F_t & y_t \end{pmatrix}'$$

where[1]

$$F_t = \sum_{k=1}^{\infty} (1-\alpha)\alpha^{k-1} \overline{\mathbb{E}_{it}\, \xi_{it}}^k = (1-\alpha)\overline{\mathbb{E}_{it}\, \xi_{it}} + \alpha \bar{\mathbb{E}}_t F_t$$
$$y_t = \sum_{k=1}^{\infty} (1-\alpha)\alpha^{k-1} \overline{\mathbb{E}_{it}\, \varepsilon_{it}^p}^k = (1-\alpha)\overline{\mathbb{E}_{it}\, \varepsilon_{it}^p} + \alpha \bar{\mathbb{E}}_t y_t$$

with (and similarly for ε_{it}^p)

$$\overline{\mathbb{E}_{it}\, \xi_{it}} = \overline{\mathbb{E}_{it}\, \xi_{it}}^1 = \int_0^1 \mathbb{E}_{jt}\, \xi_{jt}\, \mathrm{d}j \text{ and } \overline{\mathbb{E}_{it}\, \xi_{it}}^k = \int_0^1 \mathbb{E}_{jt}\, \overline{\mathbb{E}_{it}\, \xi_{it}}^{k-1}\, \mathrm{d}j$$

and the expectation operators are based on the information set $\Omega_{it} = (\cdots, s_{it-1}, a_{t-1}, s_{it})$. Firms in island i then observe

$$a_t = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \end{pmatrix} \mathbf{x}_{it} \text{ and } s_{it+1} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \end{pmatrix} \mathbf{x}_{it+1}.$$

We will guess and verify that $\mathbf{x}_t$ evolves according to the following law of motion

$$\mathbf{x}_t = \mathbf{M}\mathbf{x}_{t-1} + \mathbf{m} \begin{pmatrix} \varepsilon_t^p \\ \varepsilon_t^u \end{pmatrix}$$

for some matrices $\mathbf{M} \in \mathbb{R}^{4\times 4}$ and $\mathbf{m} \in \mathbb{R}^{4\times 2}$. We can then solve for firms' signal extraction problem to obtain how firms update $\bar{\mathbb{E}}_{t+1}[\mathbf{x}_{t+1}]$ from $\bar{\mathbb{E}}_t[\mathbf{x}_t]$, taking the *perceived* law of motion assumed above as given. It turns out that

[1] Note that $\mathbb{E}_{it}\left[\int_0^1 \xi_{jt}\, \mathrm{d}j\right]$ is always zero, but $\int_0^1 \mathbb{E}_{jt}\, \xi_{it}\, \mathrm{d}j$ is not.

$\begin{pmatrix} F_t \\ y_t \end{pmatrix}$ is a linear combination of $\bar{\mathbb{E}}_t[\mathbf{x}_t]$, thus we can calculate $\begin{pmatrix} F_t \\ y_t \end{pmatrix}$ as a function of the previous aggregate state $\mathbf{x}_{t-1}$ and innovation ε_t^p. This gives the *actual* law of motion of $\mathbf{x}_t$. The equilibrium is then characterized by a fixed point of mapping from the perceived law of motion to the actual law of motion.

C.3 Proofs for Section 3.5

For future reference, we start with the following three lemmas.

Lemma C.1 (Covariance). *We have* $\mathrm{Cov}(\bar{\tilde{\mathcal{O}}}_t, a_{t-1}) = -\dfrac{\rho V}{1-\rho\gamma_3\rho_a}\left(\dfrac{\bar{\sigma}_p^2}{\sigma_p^2} - \dfrac{\bar{\sigma}_u^2}{\sigma_u^2}\right)$.

Proof of Lemma C.1.

$$\begin{aligned}
\mathrm{Cov}(\bar{\tilde{\mathcal{O}}}_t, a_{t-1}) &= \mathrm{Cov}\big((1-\rho\gamma_3 L)^{-1}(-\rho\gamma_1\varepsilon_{t-1}^p + \rho\gamma_2\varepsilon_{t-1}^u), (1-\rho_a L)^{-1}(\varepsilon_{t-1}^p + \varepsilon_{t-1}^u)\big) \\
&= \frac{-\rho\gamma_1\bar{\sigma}_p^2 + \rho\gamma_2\bar{\sigma}_u^2}{1-\rho\gamma_3\rho_a} \\
&= -\frac{\rho V}{1-\rho\gamma_3\rho_a}\left(\frac{\bar{\sigma}_p^2}{\sigma_p^2} - \frac{\bar{\sigma}_u^2}{\sigma_u^2}\right).
\end{aligned}$$

□

Lemma C.2 (Coefficients). *Let* $\Sigma \equiv \rho^2 V + (1-\rho^2)\sigma_\eta^2$, *then*

$$\frac{1}{V} = \frac{1}{\Sigma} + \frac{1}{\sigma_p^2} + \frac{1}{\sigma_u^2}.$$

Moreover, we have

$$K = \frac{\sigma_p^2}{\Sigma + \sigma_p^2} \quad \gamma_1 = \frac{\sigma_u^2\Sigma}{\Psi} \quad \gamma_2 = \frac{\sigma_p^2\Sigma}{\Psi} \quad \gamma_3 = \frac{\sigma_p^2\sigma_u^2}{\Psi}$$

where $\Psi \equiv \Sigma(\sigma_p^2 + \sigma_u^2) + \sigma_p^2\sigma_u^2 = \frac{\sigma_p^2\sigma_u^2\Sigma}{V}$. *Thus,*

$$\gamma_1 = \frac{V}{\sigma_p^2} \quad \gamma_2 = \frac{V}{\sigma_u^2} \quad \gamma_3 = \frac{V}{\Sigma}.$$

Lemma C.3 (Variables). *We can write our variables of interest in terms of innovations and states:*

$$\begin{aligned}
Y_1 &\equiv a_{t+1} - \bar{\mathbb{E}}_t a_{t+1} &&= \rho_a\big((1-K)\varepsilon_t^p + \varepsilon_t^u - K\rho\gamma_3\bar{\tilde{\mathcal{O}}}_{t-1} + \rho\gamma_1 K\varepsilon_{t-1}^p - \rho\gamma_2 K\varepsilon_{t-1}^u\big) + \varepsilon_{t+1}^p + \varepsilon_{t+1}^u \\
Y_2 &\equiv a_{t+1} - \bar{\mathbb{E}}_{t+1} a_{t+1} &&= \varepsilon_{t+1}^u + (1-K)\varepsilon_{t+1}^p + \rho K\gamma_1\varepsilon_t^p - \rho K\gamma_2\varepsilon_t^u - \rho\gamma_3 K(\gamma_3\rho\bar{\tilde{\mathcal{O}}}_{t-1} - \rho\gamma_1\varepsilon_{t-1}^p + \rho\gamma_2\varepsilon_{t-1}^u) \\
X^{cg} &\equiv \bar{\mathbb{E}}_t a_{t+1} - \bar{\mathbb{E}}_{t-1} a_{t+1} \overset{\text{sgn}}{=} (\rho_a(1-K) - \rho K\gamma_1)\varepsilon_{t-1}^p + (\rho_a + \rho K\gamma_2)\varepsilon_{t-1}^u + K\varepsilon_t^p + (\rho\gamma_3 - \rho_a)K\bar{\tilde{\mathcal{O}}}_{t-1}
\end{aligned}$$

$$X^{kw} \equiv \qquad a_t \qquad = \rho_a^2 a_{t-2} + \varepsilon_t^u + \varepsilon_t^p + \rho_a \varepsilon_{t-1}^p + \rho_a \varepsilon_{t-1}^u$$

Proof of Lemma C.3.

$$\begin{aligned}
Y_1 \equiv a_{t+1} - \bar{\mathbb{E}}_t a_{t+1} &= \rho_a(a_t - \bar{\mathbb{E}}_t a_t) + \varepsilon_{t+1}^p + \varepsilon_{t+1}^u \\
&= \rho_a((1-K)\varepsilon_t^p + \varepsilon_t^u - K\bar{\tilde{\mathcal{O}}}_t) + \varepsilon_{t+1}^p + \varepsilon_{t+1}^u \\
&= \rho_a\big((1-K)\varepsilon_t^p + \varepsilon_t^u - K\rho\gamma_3\bar{\tilde{\mathcal{O}}}_{t-1} + \rho\gamma_1 K\varepsilon_{t-1}^p - \rho\gamma_2 K\varepsilon_{t-1}^u\big) + \varepsilon_{t+1}^p + \varepsilon_{t+1}^u \\
Y_2 \equiv a_{t+1} - \bar{\mathbb{E}}_{t+1} a_{t+1} &= \varepsilon_{t+1}^u + (1-K)\varepsilon_{t+1}^p + \rho K\gamma_1\varepsilon_t^p - \rho K\gamma_2\varepsilon_t^u - \rho\gamma_3 K\bar{\tilde{\mathcal{O}}}_t \\
&= \varepsilon_{t+1}^u + (1-K)\varepsilon_{t+1}^p + \rho K\gamma_1\varepsilon_t^p - \rho K\gamma_2\varepsilon_t^u - \rho\gamma_3 K(\gamma_3\rho\bar{\tilde{\mathcal{O}}}_{t-1} - \rho\gamma_1\varepsilon_{t-1}^p + \rho\gamma_2\varepsilon_{t-1}^u) \\
X^{cg} \equiv \bar{\mathbb{E}}_t a_{t+1} - \bar{\mathbb{E}}_{t-1} a_{t+1} &= \rho_a(\bar{\mathbb{E}}_t a_t - \rho_a\bar{\mathbb{E}}_{t-1}a_{t-1}) \\
&\overset{\text{sgn}}{=} \Big(\rho_a^2 a_{t-2} + (\rho_a - \rho K\gamma_1)\varepsilon_{t-1}^p + (\rho_a + \rho K\gamma_2)\varepsilon_{t-1}^u + K\varepsilon_t^p + \rho\gamma_3 K\bar{\tilde{\mathcal{O}}}_{t-1}\Big) \\
&\quad - \rho_a\Big(\rho_a a_{t-2} + K(\varepsilon_{t-1}^p + \bar{\tilde{\mathcal{O}}}_{t-1})\Big) \\
&= (\rho_a(1-K) - \rho K\gamma_1)\varepsilon_{t-1}^p + (\rho_a + \rho K\gamma_2)\varepsilon_{t-1}^u + K\varepsilon_t^p + (\rho\gamma_3 - \rho_a)K\bar{\tilde{\mathcal{O}}}_{t-1} \\
X^{kw} \equiv a_t &= \rho_a^2 a_{t-2} + \varepsilon_t^u + \varepsilon_t^p + \rho_a\varepsilon_{t-1}^p + \rho_a\varepsilon_{t-1}^u
\end{aligned}$$

□

Proof of Proposition 3.

- Case 1: Common ε_t^p: $\mathrm{Var}(\varepsilon_t^p) = \sigma_p^2$

$$\begin{aligned}
\mathrm{Cov}(Y_2, X^{cg}) &\overset{\text{sgn}}{=} \rho K^2\gamma_1\sigma_p^2 - \rho\gamma_3 K\Big(\rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \rho\gamma_1(\rho_a(1-K) - \rho K\gamma_1)\sigma_p^2 + \rho\gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2\Big) \\
&\overset{\text{sgn}}{=} K\Sigma - \Big(\rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \rho\gamma_1(\rho_a(1-K) - \rho K\gamma_1)\sigma_p^2 + \rho\gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2\Big) \\
&\overset{\text{sgn}}{=} K\Sigma - \Big(\rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \rho V(\rho_a(1-K) - \rho K\gamma_1) + \rho V(\rho_a + \rho K\gamma_2)\Big) \\
&\overset{\text{sgn}}{=} K\Sigma - \rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \rho VK(\rho_a + \rho\gamma_2 + \rho\gamma_1) \\
&\overset{\text{sgn}}{=} \Sigma - \rho\gamma_3(\rho\gamma_3 - \rho_a)\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \rho V(\rho_a + \rho\gamma_2 + \rho\gamma_1)
\end{aligned}$$

where $\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) = \dfrac{\rho^2(\gamma_1^2\sigma_p^2 + \gamma_2^2\sigma_u^2)}{1 - \gamma_3^2\rho^2}$. Since we always have $\Sigma > \rho V(\rho_a + \rho\gamma_2 + \rho\gamma_1)$, or

$$1 > \rho\gamma_3(\rho_a + \rho - \rho\gamma_3),$$

a sufficient condition for $\mathrm{Cov}(Y, X) > 0$ is to have $\rho_a > \rho\gamma_3$.

Moreover, we can show that $Cov(Y_2, X^{cg}) > 0$ always holds.

- Case 2: Fully Idiosyncratic ε_t^p: $\mathrm{Var}(\varepsilon_t^p) = 0$ Then, since $\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_t) = \frac{\rho^2\gamma_2^2\sigma_u^2}{1-\rho^2\gamma_3^2}$,

$$\begin{aligned}
\mathrm{Cov}(Y_2, X^{cg}) &= -\rho\gamma_3 K\Big(\gamma_3\rho(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) + \rho\gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2\Big) \\
&\overset{\text{sgn}}{=} \gamma_3(\rho_a - \rho\gamma_3)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2 \\
&\overset{\text{sgn}}{=} (\rho_a - \rho\gamma_3)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \Sigma(\rho_a + \rho K\gamma_2).
\end{aligned}$$

This is linear in ρ_a, so it suffices to show $\mathrm{Cov} < 0$ when $\rho_a = 0$ and $\rho_a = 1$. The former is obvious, the latter is:

$$\begin{aligned}
\mathrm{Cov}(Y_2, X^{cg}) &\overset{\text{sgn}}{=} (1 - \rho\gamma_3)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \Sigma(1 + \rho K\gamma_2) \\
&\overset{\text{sgn}}{=} \frac{\rho^2\gamma_2^2\sigma_u^2 K}{1 + \rho\gamma_3} - \Sigma(1 + \rho K\gamma_2) \\
&< 0.
\end{aligned}$$

- For the cases with Y_1

$$\begin{aligned}
\mathrm{Cov}(Y_1, X^{cg}) &\overset{\text{sgn}}{=} (1-K)K\bar{\sigma}_p^2 - K\rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) + K\rho\gamma_1(\rho_a(1-K) - \rho K\gamma_1)\bar{\sigma}_p^2 - K\rho\gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2 \\
&\overset{\text{sgn}}{=} (1-K)\bar{\sigma}_p^2 - \rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) + \rho\gamma_1(\rho_a(1-K) - \rho K\gamma_1)\bar{\sigma}_p^2 - \rho\gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2
\end{aligned}$$

For $\bar{\sigma}_p^2 = \sigma_p$, we have

$$\begin{aligned}
\mathrm{Cov}(Y_1, X^{cg}) &\overset{\text{sgn}}{=} (1-K)\sigma_p^2 - \rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) + \rho\gamma_1(\rho_a(1-K) - \rho K\gamma_1)\sigma_p^2 - \rho\gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2 \\
&= (1-K)\sigma_p^2 - \rho\gamma_3(\rho\gamma_3 - \rho_a)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) + \rho V(\rho_a(1-K) - \rho K\gamma_1) - \rho V(\rho_a + \rho K\gamma_2) \\
&\overset{\text{sgn}}{=} \frac{1-K}{K}\sigma_p^2 - \rho\gamma_3(\rho\gamma_3 - \rho_a)\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \rho V(\rho_a + \rho\gamma_1 + \rho\gamma_2) \\
&\overset{\text{sgn}}{=} \mathrm{Cov}(Y_2, X^{cg})
\end{aligned}$$

For $\bar{\sigma}_p^2 = 0$, we have

$$\begin{aligned}
\mathrm{Cov}(Y_1, X^{cg}) &\overset{\text{sgn}}{=} \gamma_3(\rho_a - \rho\gamma_3)K\,\mathrm{Var}(\bar{\tilde{\mathcal{O}}}_{t-1}) - \gamma_2(\rho_a + \rho K\gamma_2)\sigma_u^2 \\
&\overset{\text{sgn}}{=} \mathrm{Cov}(Y_2, X^{cg})
\end{aligned}$$

□

Proof of Proposition 4.

$$\begin{aligned}
\operatorname{Cov}(Y_2, X^{kw}) &\overset{\text{sgn}}{=} \gamma_1\bar\sigma_p^2 - \gamma_2\bar\sigma_u^2 - \gamma_3(-\rho\gamma_1\rho_a\bar\sigma_p^2 + \rho\gamma_2\rho_a\bar\sigma_u^2) - \rho\gamma_3^2\rho_a^2 \operatorname{Cov}(\bar{\bar{\mathcal{O}}}_{t-1}, a_{t-2}) \\
&\overset{\text{sgn}}{=} \frac{\bar\sigma_p^2}{\sigma_p^2} - \frac{\bar\sigma_u^2}{\sigma_u^2} < 0. \\
\operatorname{Cov}(Y_1, X^{kw}) &\overset{\text{sgn}}{=} (1-K)\bar\sigma_p^2 + \bar\sigma_u^2 - K\rho\gamma_3\rho_a^2 \operatorname{Cov}(\bar{\bar{\mathcal{O}}}_{t-1}, a_{t-2}) + K\rho\gamma_1\rho_a\bar\sigma_p^2 - K\rho\gamma_2\rho_a\bar\sigma_u^2 \\
&\overset{\text{Lemma C.1}}{=} (1-K)\bar\sigma_p^2 + \bar\sigma_u^2 + K\rho\gamma_3\rho_a^2\frac{\rho V}{1-\rho\gamma_3\rho_a}\left(\frac{\bar\sigma_p^2}{\sigma_p^2} - \frac{\bar\sigma_u^2}{\sigma_u^2}\right) + K\rho\rho_a V\left(\frac{\bar\sigma_p^2}{\sigma_p^2} - \frac{\bar\sigma_u^2}{\sigma_u^2}\right) \\
&= (1-K)\bar\sigma_p^2 + \bar\sigma_u^2 + \frac{K\rho\rho_a V}{1-\rho\gamma_3\rho_a}\left(\frac{\bar\sigma_p^2}{\sigma_p^2} - \frac{\bar\sigma_u^2}{\sigma_u^2}\right).
\end{aligned}$$

This is linear in $\bar\sigma_p^2$ (note: V and γ's depend on σ_p^2, not $\bar\sigma_p^2$), so it suffices to show $\operatorname{Cov} > 0$ when $\bar\sigma_p^2 = 0$ and $\bar\sigma_p^2 = \sigma_p^2$. The latter is obvious, the former is:

$$\begin{aligned}
\operatorname{Cov}(Y_1, X^{kw}) &\overset{\text{sgn}}{=} \bar\sigma_u^2 - \frac{K\rho\rho_a V}{1-\rho\gamma_3\rho_a}\frac{\bar\sigma_u^2}{\sigma_u^2} \\
&\overset{\text{sgn}}{=} 1 - \frac{K\rho\rho_a\gamma_2}{1-\rho\gamma_3\rho_a} \\
&\overset{\text{sgn}}{=} 1 - \rho\rho_a(\gamma_3 + K\gamma_2) > 0.
\end{aligned}$$

Finally, $a_{t+1} - \bar{\mathbb{E}}_t a_{t+1} = \rho_a(a_t - \bar{\mathbb{E}}_t a_t) + \varepsilon^p_{t+1} + \varepsilon^u_{t+1}$, so $\operatorname{Cov}(Y_1, X^{kw}) \overset{\text{sgn}}{=} \operatorname{Cov}(LY_2, X^{kw})$. □

Proposition C.1 (Misspecification). *When agent i thinks that her noise term follows an AR(1) process, $\xi_{it} = \hat\rho\xi_{it-1} + \eta_{it}$ where $\eta_{it} \sim \mathcal{N}\big(0, (1-\hat\rho^2)\sigma_\xi^2\big)$, while the truth is $\xi_{it} = \rho\xi_{it-1} + \eta_{it}$ where $\eta_{it} \sim \mathcal{N}\big(0, (1-\rho^2)\sigma_\xi^2\big)$, we have*[2]

$$\operatorname{Cov}(a_{it+h} - \mathbb{E}_{it}\, a_{it+h}, \mathbb{E}_{it}\, a_{it+h} - \mathbb{E}_{it-1}\, a_{it+h}) < 0 \iff \rho < \hat\rho.$$

Proof. We can ignore the volatility from $\varepsilon^p, \varepsilon^u$. Modulo this volatility, we have

$$\begin{aligned}
Y_t &= a_{it+h} - \mathbb{E}_{it}\, a_{it+h} \\
&= \rho_a^h(a_t - \mathbb{E}_{it}\, a_{it}) + \varepsilon_{t,t+h} \\
&= \rho_a^h\Big(\big(\rho_a a_{it-1} + \varepsilon^p_{it} + \varepsilon^u_{it}\big) - \big(\rho_a a_{it-1} + K(\varepsilon^p_{it} + \xi_{it} - \hat\rho m_{t-1})\big)\Big) + \varepsilon_{t,t+h} \\
&= \rho_a^h\Big(- K\xi_{it} + K\hat\rho((\gamma_1+\gamma_2)\xi_{it-1} + \gamma_3\hat\rho m_{t-2})\Big) \\
X_t &= \mathbb{E}_{it}\, a_{it+h} - \mathbb{E}_{it-1}\, a_{it+h}
\end{aligned}$$

[2] We have normalized the variance of innovation to have $\operatorname{Var}(\xi_{it}) = \sigma_\xi^2$.

$$
\begin{aligned}
&= \rho_a^h(\mathbb{E}_{it}\, a_{it} - \rho_a\, \mathbb{E}_{it-1}\, a_{it-1}) \\
&= \rho_a^h\Big(\big(\rho_a a_{it-1} + K(\varepsilon_{it}^p + \xi_{it} - \hat{\rho} m_{t-1})\big) - \rho_a\big(\rho_a a_{it-2} + K(\varepsilon_{it-1}^p + \xi_{it-1} - \hat{\rho} m_{t-2})\big)\Big) \\
&= \rho_a^h\Big(\rho_a(\varepsilon_{it-1}^p + \varepsilon_{it-1}^u) + K\varepsilon_{it}^p + K\xi_{it} - K\hat{\rho} m_{t-1} - \rho_a K\varepsilon_{it-1}^p - \rho_a K\xi_{it-1} + \rho_a K\hat{\rho} m_{t-2}\Big) \\
&= \rho_a^h\Big(K\xi_{it} - K\hat{\rho}((\gamma_1+\gamma_2)\xi_{it-1} + \gamma_3\hat{\rho} m_{t-2}) - \rho_a K\xi_{it-1} + \rho_a K\hat{\rho} m_{t-2}\Big)
\end{aligned}
$$

Thus, as $m_t = \gamma_3\hat{\rho} m_{t-1} + (\gamma_1+\gamma_2)\xi_{it}$,

$$
\begin{aligned}
\tilde{Y}_t \equiv \frac{Y_t}{K\rho_a^h\hat{\rho}(\gamma_1+\gamma_2)} &= -\frac{1}{\hat{\rho}(\gamma_1+\gamma_2)}\xi_{it} + \xi_{it-1} + \frac{\gamma_3\hat{\rho}}{\gamma_1+\gamma_2} m_{t-2} \\
&= -\frac{1}{\hat{\rho}(\gamma_1+\gamma_2)}\xi_{it} + \xi_{it-1} + \gamma_3\hat{\rho}\xi_{it-2} + (\gamma_3\hat{\rho})^2\xi_{it-3} + \cdots \\
\tilde{X}_t \equiv \frac{X_t}{K\rho_a^h\hat{\rho}(\gamma_1+\gamma_2)} &= \frac{1}{\hat{\rho}(\gamma_1+\gamma_2)}\xi_{it} + \frac{\rho_a - \hat{\rho}\gamma_3}{\gamma_1+\gamma_2} m_{t-2} - \left(1 + \frac{\rho_a}{\hat{\rho}(\gamma_1+\gamma_2)}\right)\xi_{it-1}
\end{aligned}
$$

so, for $\theta \equiv \frac{1}{\hat{\rho}(\gamma_1+\gamma_2)}$, $\beta = \gamma_3\hat{\rho}$, and $\delta = \rho_a - \hat{\rho}\gamma_3$,

$$
\frac{\mathbb{E}\big[\tilde{Y}_t\tilde{X}_t\big]}{\sigma_\eta^2} = \begin{pmatrix} -\theta & 1 & \beta & \beta^2 & \beta^3 & \cdots \end{pmatrix}
\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots \\ \rho & 1 & \rho & \rho^2 & \cdots \\ \rho^2 & \rho & 1 & \rho & \cdots \\ \rho^3 & \rho^2 & \rho & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}
\begin{pmatrix} \theta \\ -1-\rho_a\theta \\ \delta \\ \delta\beta \\ \delta\beta^2 \\ \delta\beta^3 \\ \vdots \end{pmatrix} = (1) + (2) + (3)
$$

$$
\begin{aligned}
\text{where } \ (1) &= \delta\beta\left(\sum_{j,k\geq 0}\rho^{|j-k|}\beta^{j+k}\right) = \delta\beta\left(\frac{1}{1-\beta^2} + \sum_{t\geq 1}\rho^t 2\frac{\beta^t}{1-\beta^2}\right) = \delta\beta\left(\frac{1}{1-\beta^2} + \frac{2}{1-\beta^2}\frac{\rho\beta}{1-\rho\beta}\right) \\
&= \delta\beta\frac{1+\rho\beta}{(1-\beta^2)(1-\rho\beta)} \\
(2) &= -\theta\left(\theta - \rho(1+\rho_a\theta) + \frac{\delta\rho^2}{1-\beta\rho}\right) + \left(\rho\theta - 1 - \rho_a\theta + \frac{\delta\rho}{1-\beta\rho}\right) \\
(3) &= \theta\frac{\beta\rho^2}{1-\beta\rho} - (1+\rho_a\theta)\frac{\beta\rho}{1-\beta\rho}
\end{aligned}
$$

Finally, we can show that

$$
\frac{\partial\Big((1)+(2)+(3)\Big)}{\partial\rho} \geq 0
$$

□

Proof of Proposition 5.

Proof for our model. Note first that

$$z_{t+1} \equiv a_{t+1} - \bar{\mathbb{E}}_{t+1} a_{t+1} = \rho\gamma_3 z_t + \kappa_{t+1} \quad \text{where } \kappa_{t+1} = -\rho(\gamma_3 + K\gamma_2)\varepsilon_t^u + \varepsilon_{t+1}^u + (1-K)\varepsilon_{t+1}^p$$
$$a_t = \rho_a a_{t-1} + \mu_t \quad \text{where } \mu_t = \varepsilon_t^p + \varepsilon_t^u.$$

Thus,

$$\begin{aligned}
\operatorname{Cov}(z_{t+1}, a_t) &= \frac{1}{1-\rho\gamma_3\rho_a}\Big(\operatorname{Cov}(\kappa_{t+1}, \mu_t) + \rho\gamma_3 \operatorname{Cov}(z_t, \mu_t) + \rho_a \operatorname{Cov}(\kappa_{t+1}, a_{t-1})\Big) \\
&= \frac{1}{1-\rho\gamma_3\rho_a}\Big(-\rho(\gamma_3 + K\gamma_2)\bar{\sigma}_u^2 + \rho\gamma_3((1-K)\bar{\sigma}_p^2 + \bar{\sigma}_u^2)\Big) \\
&= \frac{\rho}{1-\rho\gamma_3\rho_a}\Big(\gamma_3(1-K)\bar{\sigma}_p^2 - K\gamma_2\bar{\sigma}_u^2\Big) \\
&= \frac{\rho K V}{1-\rho\gamma_3\rho_a}\left(\frac{\bar{\sigma}_p^2}{\sigma_p^2} - \frac{\bar{\sigma}_u^2}{\sigma_u^2}\right)
\end{aligned}$$

Finally, we have $\operatorname{Var}(a_t) = \frac{\bar{\sigma}_p^2 + \bar{\sigma}_u^2}{1-\rho_a^2}$. □

Proof for extrapolation.

$$\operatorname{Cov}(a_{t+1} - \bar{E}_{t+1} a_{t+1}, a_t) = \operatorname{Cov}(\rho a_t + u_{t+1} - \hat{\rho} a_t, a_t) = (\rho - \hat{\rho}) \operatorname{Var}(a_t)$$
$$\operatorname{Cov}(a_{it+1} - E_{it+1} a_{it+1}, a_{it}) = \operatorname{Cov}(\rho a_{it} + u_{it+1} - \hat{\rho} a_{it}, a_{it}) = (\rho - \hat{\rho}) \operatorname{Var}(a_{it}).$$ □

Proof for diagnostic expectations. For diagnostic expectations with Kalman gain g_0, we have

$$\widehat{\beta}_{aggr} = \widehat{\beta}_{ind} = \frac{(1-g_0)\rho(1-\rho^2)(1-K)}{1-\rho^2(1-K)} \overset{\text{sgn}}{=} (1-g_0).$$

We have

$$\begin{aligned}
a_{it+1} - \mathbb{E}_{it+1} a_{it+1} &= a_{it+1} - \mathbb{E}_{it}^{NRE} a_{it+1} - g_0(z_{it+1} - \mathbb{E}_{it}^{NRE} a_{it+1}) \\
&\equiv (1-g_0)(a_{it+1} - \mathbb{E}_{it}^{NRE} a_{it+1})
\end{aligned}$$

Thus,

$$\begin{aligned}
\widehat{\beta}_{ind} &= (1-g_0)\frac{\operatorname{Cov}(a_{it+1} - \mathbb{E}_{it}^{NRE} a_{it+1}, a_{it})}{\operatorname{Var}(a_{it})} \\
&= (1-g_0)\rho\frac{\operatorname{Cov}(a_{it} - \mathbb{E}_{it}^{NRE} a_{it})}{\operatorname{Var}(a_{it})}
\end{aligned}$$

$$= (1-g_0)\rho\left(1 - \frac{K}{1-\rho^2(1-K)}\right)$$
$$= \frac{(1-g_0)\rho(1-\rho^2)(1-K)}{1-\rho^2(1-K)}$$

where the second to the last equality uses the fact that

$$\begin{aligned}\mathbb{E}_{it}^{NRE}\, a_{it} &= \mathbb{E}_{it-1}^{NRE}\, a_{it} + K(z_{it} - \mathbb{E}_{it-1}^{NRE}\, a_{it}) \\ &= Kz_{it} + \rho(1-K)\,\mathbb{E}_{it-1}^{NRE}\, a_{it-1} \\ &= K\sum_{h\geq 0}\rho^h(1-K)^h z_{it-h}\end{aligned}$$

hence

$$\mathrm{Cov}(\mathbb{E}_{it}^{NRE}\, a_{it}, a_{it}) = K\sum_{h\geq 0}\rho^h(1-K)^h \underbrace{\mathrm{Cov}(z_{it-h}, a_{it})}_{=\rho^h \mathrm{Var}(a_{it})}.$$

Second, we have

$$a_{t+1} - \bar{\mathbb{E}}_{t+1}a_{t+1} \equiv (1-g_0)(a_{t+1} - \bar{\mathbb{E}}_{it}^{NRE}a_{t+1})$$

hence

$$\begin{aligned}\widehat{\beta}_{aggr} &= (1-g_0)\frac{\mathrm{Cov}(a_{t+1} - \bar{\mathbb{E}}_t^{NRE}a_{t+1}, a_t)}{\mathrm{Var}(a_t)} \\ &= (1-g_0)\rho\frac{\mathrm{Cov}(a_t - \bar{\mathbb{E}}_t^{NRE}a_t, a_t)}{\mathrm{Var}(a_t)} \\ &= (1-g_0)\rho\left(1 - \frac{K}{1-\rho^2(1-K)}\right) \\ &= \frac{(1-g_0)\rho(1-\rho^2)(1-K)}{1-\rho^2(1-K)}\end{aligned}$$

where the second to the last equality uses the fact that

$$\bar{\mathbb{E}}_t^{NRE}a_t = K\sum_{h\geq 0}\rho^h(1-K)^h a_{t-h}$$

hence

$$\mathrm{Cov}(\bar{\mathbb{E}}_t^{NRE}a_t, a_t) = K\sum_{h\geq 0}\rho^h(1-K)^h \underbrace{\mathrm{Cov}(a_{t-h}, a_t)}_{=\rho^h \mathrm{Var}(a_t)}.$$

□

Proof for AHS . We have

$$a_{it} = \rho a_{it-1} + \varepsilon_{it} \qquad \text{(perceived: } \hat{\rho}\text{)}$$

$$z_{it} = a_{it} + \frac{1}{\sqrt{\tau}} u_{it}, \qquad \text{(perceived: } \hat{\tau}\text{)}$$

we have

$$\widehat{\beta}_{aggr} = \widehat{\beta}_{ind} = \rho - \hat{K}\rho - \frac{\hat{K}\hat{\rho}(1-\hat{K})}{1-\hat{\rho}\rho(1-\hat{K})}.$$

As in above, we have

$$\mathbb{E}_{it}\, a_{it} = \hat{K} \sum_{h \geq 0} \hat{\rho}^h (1-\hat{K})^h z_{it-h}$$

where $\hat{K} \in (0,1)$ is a function of $\hat{\rho}$ and $\hat{\tau}$. We then have

$$a_{it+1} - \mathbb{E}_{it+1}\, a_{it+1} = \sum_{h \geq 0} \rho^h \varepsilon_{i,t+1-h} - \hat{K} \sum_{h \geq 0} \hat{\rho}^h (1-\hat{K})^h z_{i,t+1-h}$$

so

$$\begin{aligned}
\operatorname{Cov}(a_{it+1} - \mathbb{E}_{it+1}\, a_{it+1}, a_{it}) &= \sum_{h \geq 0} \rho^h \operatorname{Cov}(\varepsilon_{i,t+1-h}, a_{it}) - \hat{K} \sum_{h \geq 0} \hat{\rho}^h (1-\hat{K})^h \operatorname{Cov}(z_{i,t+1-h}, a_{it}) \\
&= \rho \operatorname{Var}(\varepsilon_{it}) + \rho^3 \operatorname{Var}(\varepsilon_{it}) + \rho^5 \operatorname{Var}(\varepsilon_{it}) + \cdots \\
&\quad - \hat{K}\rho \operatorname{Var}(a_{it}) - \hat{K}\hat{\rho}(1-\hat{K}) \operatorname{Var}(a_{it}) - \hat{K}\hat{\rho}^2(1-\hat{K})^2 \rho \operatorname{Var}(a_{it}) - \cdots \\
&= \operatorname{Var}(a_{it}) \left(\rho - \hat{K}\rho - \frac{\hat{K}\hat{\rho}(1-\hat{K})}{1-\hat{\rho}\rho(1-\hat{K})} \right).
\end{aligned}$$

where the last equality uses the fact that

$$(1-\rho^2) \operatorname{Var}(a_{it}) = \operatorname{Var}(\varepsilon_{it}).$$

Thus, we have

$$\widehat{\beta}_{ind} = \rho - \hat{K}\rho - \frac{\hat{K}\hat{\rho}(1-\hat{K})}{1-\hat{\rho}\rho(1-\hat{K})}.$$

We have

$$\operatorname{Cov}(a_{t+1} - \bar{\mathbb{E}}_{t+1} a_{t+1}, a_t) = \sum_{h \geq 0} \rho^h \operatorname{Cov}(\varepsilon_{t+1-h}, a_t) - \hat{K} \sum_{h \geq 0} \hat{\rho}^h (1-\hat{K})^h \operatorname{Cov}(a_{t+1-h}, a_t)$$

$$
\begin{aligned}
&= \rho \operatorname{Var}(\varepsilon_t) + \rho^3 \operatorname{Var}(\varepsilon_t) + \rho^5 \operatorname{Var}(\varepsilon_t) + \cdots \\
&\quad - \hat{K}\rho \operatorname{Var}(a_t) - \hat{K}\hat{\rho}(1-\hat{K}) \operatorname{Var}(a_t) - \hat{K}\hat{\rho}^2(1-\hat{K})^2 \rho \operatorname{Var}(a_t) - \cdots \\
&= \operatorname{Var}(a_t)\left(\rho - \hat{K}\rho - \frac{\hat{K}\hat{\rho}(1-\hat{K})}{1-\hat{\rho}\rho(1-\hat{K})}\right).
\end{aligned}
$$

where the last equality uses the fact that

$$
(1-\rho^2)\operatorname{Var}(a_t) = \operatorname{Var}(\varepsilon_t).
$$

Thus, we have

$$
\widehat{\beta}_{aggr} = \rho - \hat{K}\rho - \frac{\hat{K}\hat{\rho}(1-\hat{K})}{1-\hat{\rho}\rho(1-\hat{K})}. \qquad \square
$$

Proof for KW . We have

$$
\begin{aligned}
y_{it} &= \sum_j x_{ijt} \\
x_{ijt} &= a_j \theta_{it} + b_j u_{ijt} \\
\theta_{it} &= \rho \theta_{it-1} + \eta_{it} \\
z_{ijt} &= x_{ijt} + q_j \cdot \varepsilon_{ijt}
\end{aligned}
$$

Then, we have $\widehat{\beta}_{aggr} < \widehat{\beta}_{ind}$ if and only if

$$
\frac{\operatorname{Var}(u_{jt})}{\operatorname{Var}(\eta_t)} > \frac{\operatorname{Var}(u_{ijt})}{\operatorname{Var}(\eta_{it})}.
$$

Starting from (3), we have

$$
\begin{aligned}
\mathbb{E}_{it}[\theta_{it}] &= \mathbb{E}_{it-1}[\theta_{it}] + \sum_j g_j (z_{ijt} - \mathbb{E}_{it-1}\, z_{ijt}) \\
&= \mathbb{E}_{it-1}[\theta_{it}] + \sum_j g_j (z_{ijt} - a_j\, \mathbb{E}_{it-1}\, \theta_{it}) \\
&= \rho(1 - \sum_j g_j a_j)\, \mathbb{E}_{it-1}\, \theta_{it-1} + \sum_j g_j (a_j \theta_{it} + b_j u_{ijt} + q_j \varepsilon_{ijt}).
\end{aligned}
$$

Thus,

$$\begin{aligned}\theta_{it} - \mathbb{E}_{it}\,\theta_{it} &= (1-\sum_j g_j a_j)(\rho\theta_{it-1}+\eta_{it}) - \rho(1-\sum_j g_j a_j)\,\mathbb{E}_{it-1}\,\theta_{it-1} - \sum_j g_j b_j u_{ijt} - \sum_j g_j q_j \varepsilon_{ijt} \\ &= \underbrace{\rho(1-\sum_j g_j a_j)}_{\Gamma}(\theta_{it-1} - \mathbb{E}_{it-1}\,\theta_{it-1}) + \underbrace{(1-\sum_j g_j a_j)\eta_{it} - \sum_j g_j b_j u_{ijt} - \sum_j g_j q_j \varepsilon_{ijt}}_{\zeta_{it}} \\ &= \sum_{h\geq 0}\Gamma^h \zeta_{it-h}.\end{aligned}$$

Thus, we have

$$\begin{aligned}y_{it+1} - \mathbb{E}_{it+1}\,y_{it+1} &\equiv (\sum_j a_j)(\theta_{it+1} - \mathbb{E}_{it+1}\,\theta_{it+1}) \\ &= (\sum_j a_j)\sum_{h\geq 0}\Gamma^h \zeta_{i,t+1-h}.\end{aligned}$$

and

$$y_{it} = (\sum_j a_j)\sum_{h\geq 0}\rho^h \eta_{it-h} + \sum_j b_j u_{ijt}.$$

Thus, the covariance between them is

$$\mathrm{Cov} = -(\sum_j a_j)\Gamma\sum_j g_j b_j^2\,\mathrm{Var}(u_{ijt}) + (\sum_j a_j)^2 \Gamma \underbrace{\mathrm{Cov}\left(\sum_{h\geq 0}\Gamma^h\zeta_{it-h}, \sum_{h\geq 0}\rho^h\eta_{it-h}\right)}_{=(1-\sum_j g_j a_j)\frac{1}{1-\Gamma\rho}\mathrm{Var}(\eta_{it})}$$

Thus,

$$\widehat{\beta}_{ind} = \frac{(\sum_j a_j)\Gamma\left(\frac{(\sum_j a_j)(1-\sum_j g_j a_j)}{1-\Gamma\rho}\mathrm{Var}(\eta_{it}) - \sum_j g_j b_j^2\,\mathrm{Var}(u_{ijt})\right)}{\frac{(\sum_j a_j)^2}{1-\rho^2}\mathrm{Var}(\eta_{it}) + \sum_j b_j^2\,\mathrm{Var}(u_{ijt})}.$$

On the other hand, we have

$$\begin{aligned}y_{t+1} - \bar{\mathbb{E}}_{t+1} y_{t+1} &= (\sum_j a_j)\sum_{h\geq 0}\Gamma^h\zeta_{t+1-h} \\ y_t &= (\sum_j a_j)\sum_{h\geq 0}\rho^h\eta_{t-h} + \sum_j b_j u_{jt}.\end{aligned}$$

hence

$$\widehat{\beta}_{aggr} = \frac{(\sum_j a_j)\Gamma\Big(\frac{(\sum_j a_j)(1-\sum_j g_j a_j)}{1-\Gamma\rho}\operatorname{Var}(\eta_t) - \sum_j g_j b_j^2 \operatorname{Var}(u_{jt})\Big)}{\frac{(\sum_j a_j)^2}{1-\rho^2}\operatorname{Var}(\eta_t) + \sum_j b_j^2 \operatorname{Var}(u_{jt})}.$$

Both beta hats are of the form (impose $\operatorname{Var}(u_{jt}) = \operatorname{Var}(u_{j't})$)

$$\widehat{\beta} = \frac{b - a\kappa}{d + c\kappa} \quad \text{where } \kappa = \frac{\operatorname{Var}(u)}{\operatorname{Var}(\eta)} > 0$$

wherer $a, b, c, d > 0$. We can easily show that $\widehat{\beta}$ is decreasing in κ. So we have $\widehat{\beta}_{aggr} < \widehat{\beta}_{ind}$ if and only if

$$\frac{\operatorname{Var}(u_{jt})}{\operatorname{Var}(\eta_t)} > \frac{\operatorname{Var}(u_{ijt})}{\operatorname{Var}(\eta_{it})}.$$

□