What Causes Agglomeration of Services? Theory and Evidence from Seoul*

Ryungha Oh† Jaeeun Seo‡

Yale University MIT

October 2023

Click here for the latest version

Abstract

Why are economic activities concentrated in space? What are the policy implications of this concentration? And how do we expect it to change in the future? We revisit these classic questions in the context of non-tradable services, 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 an otherwise standard quantitative spatial model that endogenously determines 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 it exacerbates 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.

Keywords: Agglomeration, spillover, services travel, trip chaining, services market access.

*We thank Costas Arkolakis and Arnaud Costinot for their guidance and support. We also thank David Atkin, George-Marios Angeletos, Jonathan Dingel, Dave Donaldson, Gilles Duranton, Ana Cecilia Fieler, Simon Fuchs, Edward Glaeser, Jessie Handbury, Ilse Lindenlaub, Yuhei Miyauchi, Giuseppe Moscarini, Michael Peters, Lindsay Relihan, Nick Tsivanidis, and participants at the MIT Trade Lunch, the Yale International and Spatial Economics Lunch, UEA 2022 London, CURE 2022 Philadelphia, and NBER SI 2023 Urban Economics for their helpful comments. 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.

†*Email:* ryungha.oh@yale.edu

‡*Email:* jaeeuns@mit.edu

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.

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

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

[2] 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.

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

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.

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., Leonardi and Moretti, 2023; Koster, Pasidis and van Ommeren, 2019). 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.

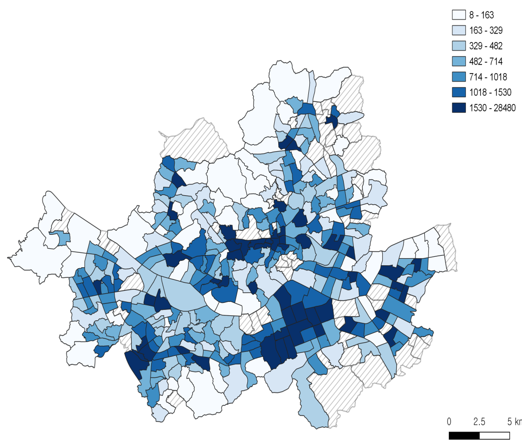


Figure 1. Number of Services Stores Per Area

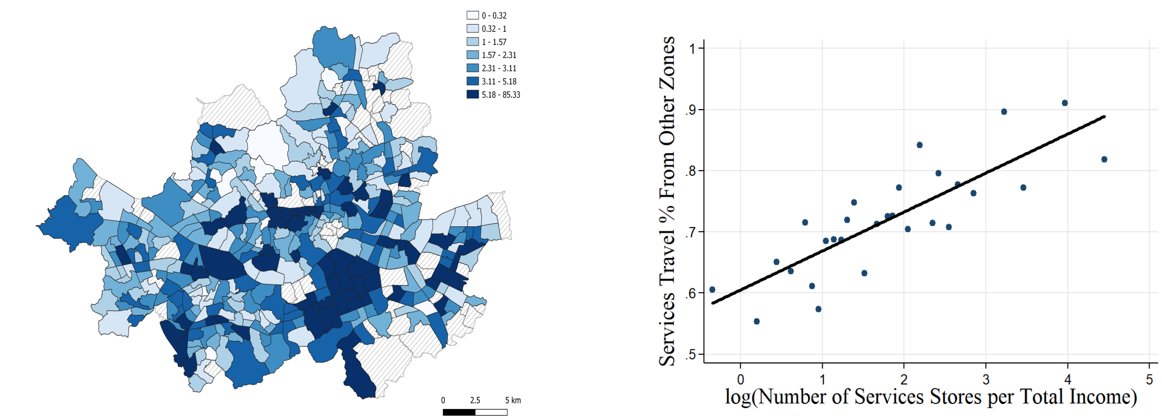
The rest of the paper is organized as follows. Section 2 discusses the background and data, provides reduced-form evidence on the spillover mechanism, and presents stylized facts on services travel. Section 3 develops a structural model that features the spillover mechanism, which is estimated in Section 4. In Section 5, we use the estimated model to investigate the importance of spillovers that arise from the trip-chaining mechanism. Finally, in Section 6, we perform counterfactual exercises on the urban structure in the future.

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



(a) Number of Services Stores Per Total Income

(b) Share of Consumers from Other Zones

Figure 2. Spatial Distribution of Services in Seoul

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 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 all 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.

2.1 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 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.

[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%.

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 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 E. 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 total

[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.

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, which are discussed in more detail in Appendix C.3.

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

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

[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.

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] We report citywide subsector trends in Table A.5 and show in Figure A.2 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, 2013), the exclusion restriction requires orthogonality between a sector's trend and *its* initial composition (in our context, this can be written as $\varepsilon_{jsd} \perp \{s_{j,s,d,0}\}_d$). 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

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

[11] This approach is widely used in the trade and urban literature; examples include Topalova (2010), Autor, Dorn and Hanson (2013), 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).

Table 1: Estimation Results

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

	(1)	(2)	(3)	(4)
	OLS	IV	IV	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			
$\Delta \log N^{\text{Bartik}}_{j,s'}$		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.

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 A.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]

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

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

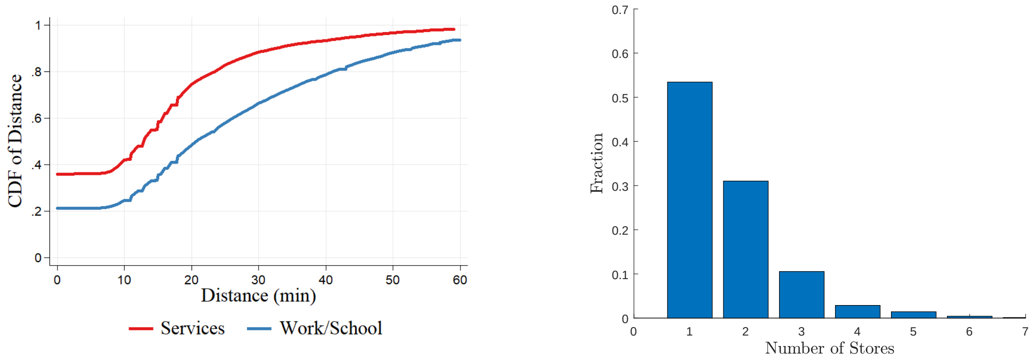
In Appendix A.3, we examine the *within*-sector effect of an exogenous increase in the number of stores. In particular, we estimate the effect on the number of stores in a subsector when there is an exogenous change in the number of stores in other subsectors in the *same* sector. As in the across-sector 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 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 3, we propose a novel mechanism that can explain them.

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

[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.



(a) CDF of Travel Distance

(b) The Number of Stores Per Travel

Figure 3. Stylized Facts

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 on 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 one-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

[17] We define the end of travel 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.

51% if they decide what to buy independently across purchases. These results guide our modeling choices for trip chaining in Section 3. More details on the results of the online survey can be found in Appendix A.1.

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 (1987) and Aguirregabiria and Mira (2010). 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 bedcause 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.2, 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 3.3, we show that consumers' trip-chaining behavior is not a source of inefficiency. In addition, we show in Section 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 are generically inefficient and always lead to greater regional inequality. In Section 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. Appendix B.5 describes how we endogenize input prices and the spatial distribution of consumers—who optimally decide where to live

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

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.

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. In this section, we focus on the services market and assume that the distribution of workers is fixed. However, in Appendix B.5, we close the model by allowing workers to choose their places of residence and work.

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. All omitted derivations and proofs in this section can be found in Appendix B.

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]

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

where $\tilde{C}$ and C_ℓ are the consumption of tradable goods and floor space, which we discuss in more detail in Appendix B.5. Workers consume services goods through their services 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 to inform our modeling choices. Our priority is to maintain

[19] The final utility depends on both consumption utility and residential amenity. See Appendix B.5.

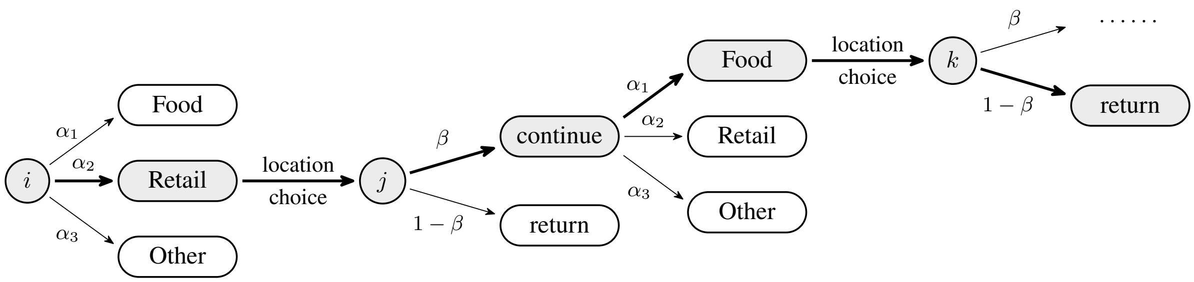


Figure 4. Timeline of Services Travel

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

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

$$\text{s.t.} \qquad \sum_{d\in\mathcal{D}_s} \int_0^{N_{jsd}} p_{jsd}(\omega) q_{jsd}(\omega)\, \mathrm{d}\omega \le e,$$

[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 documented in Section 2.3. In Appendix A.1, we demonstrate that other assumptions on services travel we make in this section are also supported by the data.

[22] Although consumers usually 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.

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, $1 < \sigma \leq \rho$. 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, $\vec{\varepsilon}_t = (\varepsilon_t^j)^j$, and a realization of the sector, s_t, together define the state variable $\sigma_t = (\vec{\varepsilon}_t, s_t)$. 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

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

$$\text{s.t.} \quad \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} q_t(\sigma^t) p_{j_t(\sigma^t) s(\sigma^t)} \pi(\sigma^t) \leq E \tag{5}$$

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 $\vec{\varepsilon}_t$, 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 distance between two regions, and the second term captures the border effect. In Appendix B.2, 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.1, 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

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

[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.

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 starting 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:

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

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 $N_{jsd} = 0$.

[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.

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.

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 4 and 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.

In Appendix B.1, we show that the total consumer spending in region j and sector s is given by

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

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 $c_{jsd} = \left(\frac{r_j}{\gamma}\right)^{\gamma} \left(\frac{w_j}{1-\gamma}\right)^{1-\gamma}$. See Appendix B.1 for details.

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:*

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

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 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.2

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, 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. In Section 4, we further show that this model can quantitatively match these patterns.

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, 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 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.[26] Importantly, the proof 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 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 $\chi > 1$ 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

[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.

external economies of scale. We define external economies of scale as *isoelastic* when they take the form of either

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

or

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

where $\Upsilon_j = \Upsilon_{1j} + \Upsilon_{2j}$. In Appendix B.4, 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, in which 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 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.

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 to match residual spillovers. In particular, we introduce a

[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.

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.

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 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 introduced in Appendix B.5. In this section, we focus on the first three steps of the estimation procedure. See Appendix C.3 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.

[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 2.1. We use the same distance measure for commuting decisions in the general equilibrium model.

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.

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 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. In addition, we present the fit of the estimated model in Figure A.3, which shows that services travel flows are well approximated by this gravity equation.

Parameter Calibration. We calibrate five types of parameters: $\{\alpha_s\}_s, \{\mu_{\tilde{c}}, \mu_r, \mu_w, \mu_\ell\}, \gamma, \rho$, 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

[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.

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) and Davis and Ortalo-Magné (2011). 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 C.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.[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, but using composite productivity instead of the number of stores. Our first set of moment conditions is given by

$$\Delta \log \tilde{A}_{jsd} \;\perp_j\; \Delta \log \tilde{A}_{js'}^{Bartik} \quad \text{for all } (s, s', d) \text{ with } s \neq s'. \tag{11}$$

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.

[30] A comparison with the approach in Section 2 is in order. In Section 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).

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:

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

The second set of moment conditions is

$$\Delta \log \tilde{A}_{jsd} \ \perp_j \ \Delta \log \tilde{A}_{js,-d}^{Bartik} \quad \text{for all } (s,d) \tag{12}$$

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:

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

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

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

We estimate three parameters using these three types of moment conditions, (11)–(13), 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 $\Delta \log \tilde{A}_{js'}^{\text{Bartik}}$. 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 (Duranton and Puga, 2004), 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, 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 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. 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.

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,

[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.

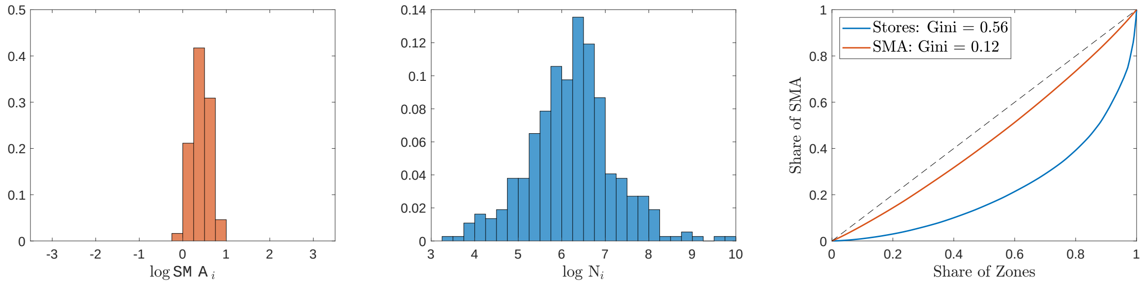


Figure 5. Spatial Disparity in the Number of Stores and SMA

both of which are substantially larger than 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 D.1, 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.

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, in which we focus on the distinctive features of trip chaining.

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

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

deviation 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.

5.2 Welfare Implications of the Trip-chaining Mechanism

In Section 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 5.1, trip chaining increases the concentration of services stores and thus contributes to higher SMA inequality. However, its 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

[37] In Sections 5 and 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 presented in Appendix B.5 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.

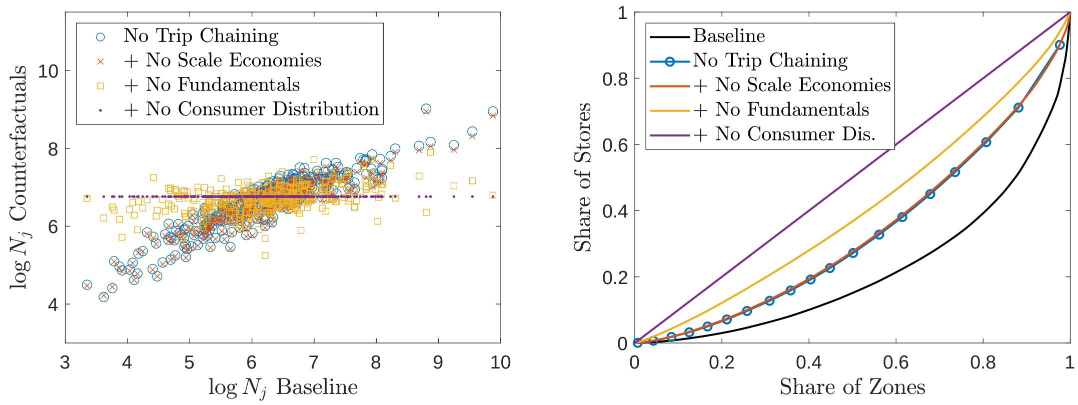


Figure 6. Importance of Trip Chaining in Agglomeration

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.

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

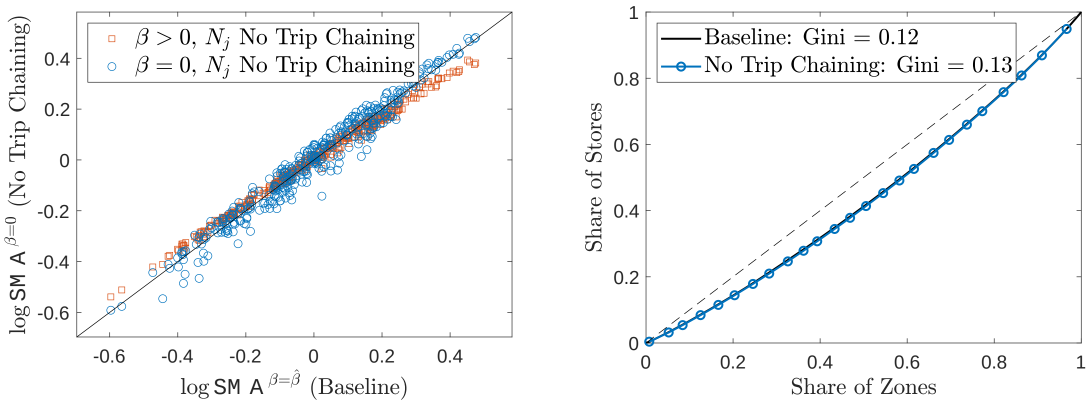


Figure 7. Importance of Trip Chaining in SMA

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.

In this section, we focus on the impact of the rise of work from home and delivery technology, which have unexpectedly been accelerated by the COVID-19 pandemic. In Appendix D.2, we also discuss the effect of transportation improvements.

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

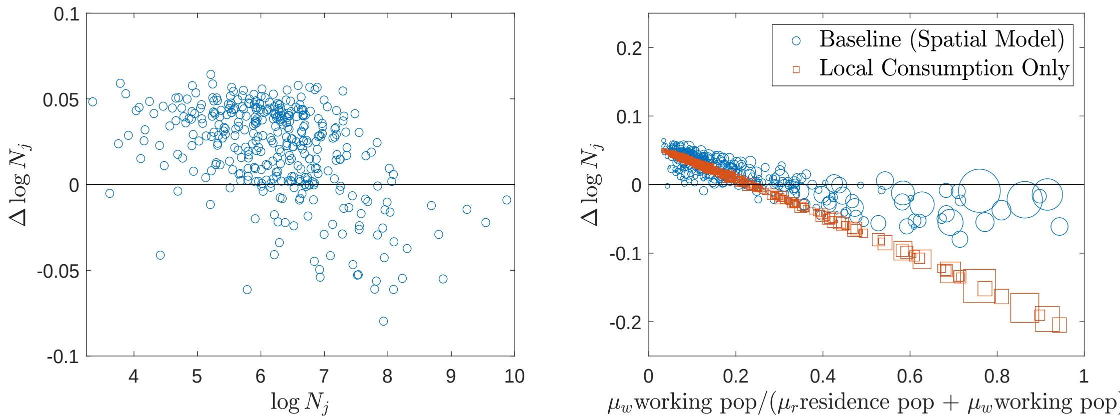


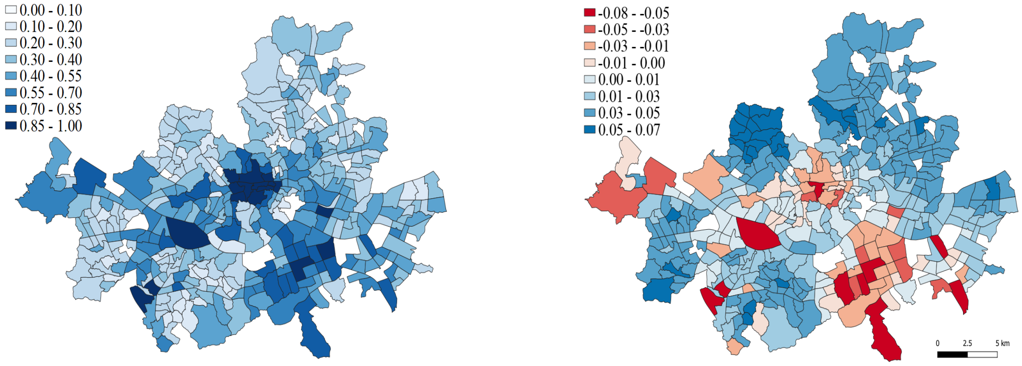
Figure 8. Changes in the Number of Stores after Work from Home

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.

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



(a) Share of the Workplace Population

(b) Change in the Number of Stores

Figure 9. Work from Home: Map

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 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.

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

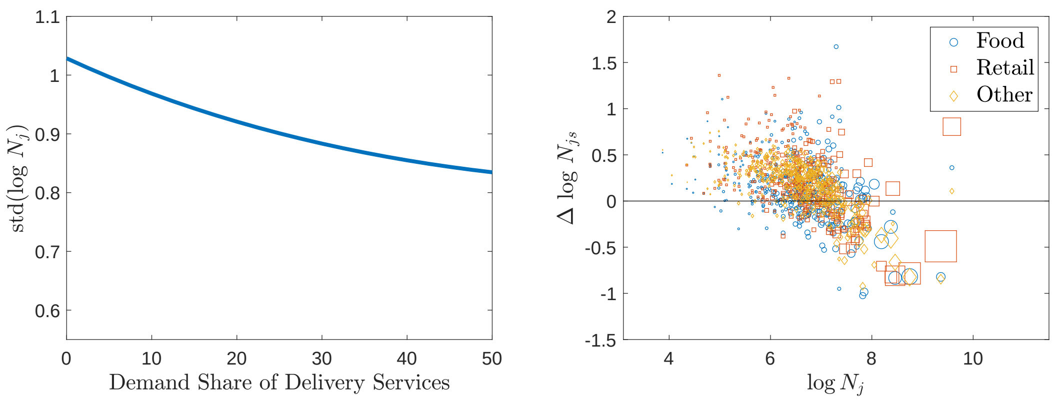


Figure 10. Delivery Services: Concentration of Services Stores

frictions are completely eliminated in the Retail sector and reduced by 50% in the Food sector when services are delivered, which is modeled by adjusting distance disutility parameters (φ, τ) accordingly.[38]

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

[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. In contrast, for the retail sector, shipping costs are flat within a city.

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 4.2.

References

Adão, Rodrigo, Costas Arkolakis, and Federico Esposito (2020) "General Equilibrium Effects in Space: Theory and Measurement."

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 (2010) "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.

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.

Anas, Alex (2007) "A unified theory of consumption, travel and trip chaining," *Journal of Urban Economics*, 62 (2), 162–186.

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.

Autor, David H, David Dorn, and Gordon H Hanson (2013) "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 103 (6), 2121–2168.

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.

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.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2020) "Quasi-Experimental Shift-Share Research Designs."

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.

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.

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.

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*, 4: Elsevier Inc. 2063–2117.

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.

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, 27–50.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020) "Bartik Instruments: What, When, Why, and How," *American Economic Review*, 110 (8), 2586–2624.

Handbury, J., and D. E. Weinstein (2015) "Goods Prices and Availability in Cities," *The Review of Economic Studies*, 82 (1), 258–296.

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.

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.

Manski, Charles F (1993) "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 60 (3), 531–542.

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."

Muñoz, Mathilde (2023) "Trading Non-Tradables: The Implications of Europe's Job Posting Policy."

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.

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 (1987) "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

Silva, JMC Santos, and Silvana Tenreyro (2006) "The log of gravity," *The Review of Economics and statistics*, 88 (4), 641–658.

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.

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 Bogotá's TransMilenio," *Working Paper* (November 2017).

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.

Appendix

Contents

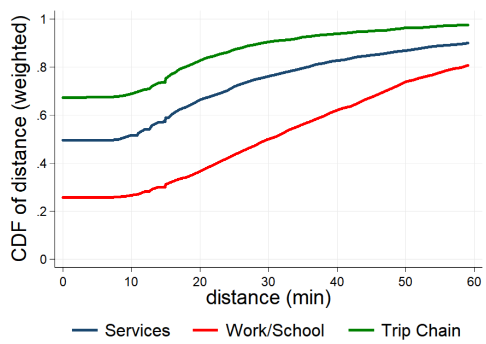
A. Appendix for Section 2

A.1 Online Travel Survey

In this section, we present additional results from the online survey. Although a few recent studies document travel patterns using credit card transaction data or smartphone data, there is still limited understanding of the details of services travel. 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. In particular, these findings allow us to assess whether our assumptions on services travel in Section 3.1 are realistic.

Travel distances. In Figure A.1, we plot the distribution of travel distances for commuting, services travel for the first purchase, and services travel for trip chaining. First, comparing the travel distances for commuting and the first purchase with those from the Korean Household Travel Survey in Figure 3a, we observe that they are comparable, 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.

Figure A.1. CDF of Travel Distance (Online Survey)



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 across types of travel. Table A.1 compares the average number of purchases based on origin and destination types, while Table A.2 compares the average number of purchases based on the sectors of the first purchase.[39] These results suggest that it is reasonable to model the trip chaining parameter as common for all types of travel.

In Table A.3, we present the correlation between the number of purchases and the number of stores in the zone where the first store is located. The two numbers show a positive correlation, indicating that consumers tend to buy more goods when they visit more concentrated areas. While we do not endogenize the decision to continue travel or not, our estimate of the contribution of the trip-chaining mechanism to agglomeration is conservative. If consumers are more inclined to make consecutive purchases in concentrated areas, spillovers would be stronger in these areas, further amplifying the role of trip chaining in services agglomeration.

Sector choices of purchases. We find that sectors of subsequent purchases do not critically depend on the sector of the previous purchase. In Table A.4, we calculate the conditional probability of

[39] 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.

Table A.1: Average Number of Purchases by Origin/Destination Types

	home	school/workplace	others
home	1.90	1.90	1.90
	(0.02)	(0.07)	(0.07)
school/workplace	2.01	1.83	2.73
	(0.07)	(0.05)	(0.41)
others	2.08	2.04	2.21
	(0.07)	(0.17)	(0.09)

Notes: Column and row represent the type of the origin and destination respectively. We report the standard deviation of the average in the parenthesis.

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.

Recursive structure and travel distances. We model services travel in a recursive manner, which has several advantages as outlined in Section 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.[40] 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.

[40] Because we only ask about locations up to the second purchase, we ignore any additional distances that may have resulted from additional purchases.

Table A.2: Average Number of Purchases by the First Sector

Sector	Food	Retail	Other
Average number of purchases	2.00	1.77	1.88
	(0.02)	(0.03)	(0.08)

Notes: We report the standard deviation of the average in the parenthesis.

A.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, which are documented in Table A.5. We first confirm that our instruments have enough variation in the Figure A.2. Below, we perform a few diagnostic tests that are suggested by Goldsmith-Pinkham, Sorkin and Swift (2020).

A.2.1 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 A.6 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 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.

A.2.2 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

Table A.3: Average Number of Purchases across Locations

	(1)	(2)	(3)
$\log N_j$	0.043^{**}	0.037^{**}	0.043^{**}
	(0.018)	(0.018)	(0.019)
Observations	3475	3475	3475
Sector FE		✓	✓
Location type FE			✓

Notes: $\log N_j$ is the total number of stores in a zone j where a consumer made the first purchase. Sector FE refers to the sector of the first purchase. Location type FE includes categories (home, school/workplace, others) of both origin and destination zones.

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 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 A.7, 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 A.7, 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.

A.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

Table A.4: 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.*.

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. To quantify this effect, we run the following regression:[41]

$$\Delta \log N_{jsd} = \alpha_2 + \beta_2 \Delta \log N_{js,-d} + \boldsymbol{X}'_{jsd}\gamma_2 + u_{jsd}$$

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

[41] 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 A.5: City-wide growth rates of subsectors

Subsector	$\mathrm{d}\log N_{sd}$
Top 3	
Japanese restaurants	0.20
Café	0.19
Street foods (*Bunsik*)	0.19
Bottom 3	
Computers	-0.13
Clothing	-0.15
Bars	-0.31

Notes: Data source: Seoul Commercial area data. We compute the growth rates of the total number of stores of the top and the bottom 5 subsectors in Seoul between 2015–2019.

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, and it yields similar results.

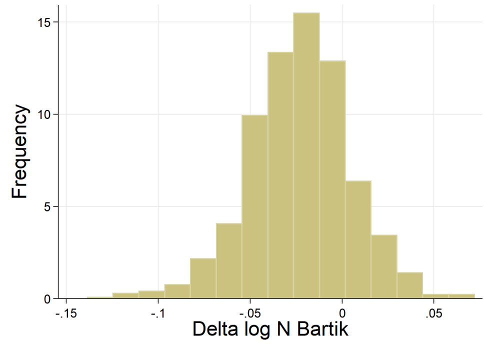


Figure A.2. Histogram of $\log N^{\text{Bartik}}_{js}$

Table A.6: Correlations with the Instruments

	$\log(\text{rent})_j$	$\log(\text{pop density})_j$	$\log(\text{income})_j$
$\Delta \log N^{\text{Bartik}}_{js}$	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 * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Pre-Trends: Number of Stores

	pre-trend (2009–2013)		trend (2015–2019)	
Instruments constructed from Seoul Commercial Dataset				
$\Delta \log N^{\text{Bartik}}_{j}$	-0.084	-0.145	0.192	0.296
	(0.230)	(0.283)	(0.177)	(0.208)
District FE		✓		✓
Observations	365	365	365	365
Instruments constructed from Seoul Business Survey				
$\Delta \log N^{\text{Bartik}}_{j}$	-0.264	-0.105	0.296**	0.285**
	(0.251)	(0.262)	(0.123)	(0.132)
District FE		✓		✓
Observations	365	365	365	365

Notes: Data source are Seoul Commercial Dataset and Seoul Business Survey. Robust standard errors are shown in parentheses, with * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: 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. Appendix for Section 3

B.1 Omitted Derivations in Section 3.1

This section provides derivations of the results presented in Section 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

$$\begin{aligned} C(i,e)^{1/(1-\beta)} &= \prod_s \left(\sum_j \left(\exp\left(U(e/p_{js}) - \tilde{d}(i,j) \right)^{1/\nu} \cdot C(j,e)^{\beta/(\nu(1-\beta))} \right) \right)^{\alpha_s \nu} \\ &= e \cdot \prod_s \left(\sum_j \left(p_{js}^{-1/\nu} \cdot \exp(-\tilde{d}(i,j))^{1/\nu} \cdot C(j,e)^{\beta/(\nu(1-\beta))} \right) \right)^{\alpha_s \nu}. \end{aligned}$$

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_{ij}^s. Thus, the spending on (j,s) is

$$\sum_{i \in \mathcal{J}} (1-\beta)\alpha_s \pi_{ij}^s E_i \equiv (1-\beta)\alpha_s \boldsymbol{E}^\top \boldsymbol{\pi}_j^s$$

where $\boldsymbol{E} = (E_1, \cdots, E_J)^\intercal$ and $\boldsymbol{\pi}_j^s = (\pi_{1j}^s, \cdots, \pi_{Jj}^s)^\intercal$. Likewise, in their second purchases, the spending on (j, s) is given by

$$\sum_{i \in \mathcal{J}} \sum_{i' \in \mathcal{J}} (1-\beta)\beta\alpha_s E_i \pi_{ii'} \pi_{i'j}^s \equiv (1-\beta)\beta\alpha_s \boldsymbol{E}^\intercal \Pi \boldsymbol{\pi}_j^s$$

where $\pi_{ii'} = \sum_s \alpha_s \pi_{ii'}^s$ 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

$$\begin{aligned} c_{jsd} \;&=\; \min_{H_{jsd}(\omega), L_{jsd}(\omega)} \quad r_j H_{jsd}(\omega) + w_j L_{jsd}(\omega) \\ &\qquad\quad \text{s.t.} \qquad\quad H_{jsd}(\omega)^\gamma L_{jsd}(\omega)^{1-\gamma} \leq 1 \\ &= \left(\frac{r_j}{\gamma}\right)^\gamma \left(\frac{w_j}{1-\gamma}\right)^{1-\gamma}. \end{aligned}$$

B.2 First-Order Approximation (Section 3.2)

In this section, we take the first-order approximations to the equilibrium conditions in Section 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.

Lemma A.1.

$$N_{jsd}^{1-\frac{1-\sigma}{1-\rho}} = C_{jsd}^{-1} \tilde{A}_{jsd}^{-(1-\sigma)} c_{jsd}^{1-\sigma} p_{js}^{-(1-\sigma)} R_{js} \tag{A.1}$$

$$p_{js} = \left(\sum_d N_{jsd}^{\frac{1-\sigma}{1-\rho}} \tilde{A}_{jsd}^{-(1-\sigma)} \left(\frac{\rho}{\rho-1}\right)^{1-\sigma} c_{jsd}^{1-\sigma}\right)^{1/(1-\sigma)} \tag{A.2}$$

$$V(i) \equiv \exp(\mathbb{E}v(i,1)) = \prod_s \left(\sum_j p_{js}^{-1/\nu} \exp(-\tilde{\tau}d(i,j) - \tilde{\varphi}\mathbb{1}_{i \neq j}) \cdot V(j)^{\tilde{\beta}}\right)^{\alpha_s \nu} \tag{A.3}$$

$$\begin{aligned} R_{js} &= (1-\beta)\alpha_s \boldsymbol{E}^\intercal (I - \beta\Pi)^{-1} \boldsymbol{\pi}_j^s \\ &= \alpha_s \left((1-\beta)\sum_i E_i \pi_{ij}^s + \beta \sum_i \sum_k E_i \pi_{ik} \pi_{kj}^s\right) + o(\beta) \end{aligned} \tag{A.4}$$

$$\pi_{ij}^s = \frac{p_{js}^{-1/\nu} \exp(-\tilde{\tau}d(i,j) - \tilde{\varphi}\mathbb{1}_{i \neq j}) \cdot V(j)^{\tilde{\beta}}}{\sum_{j'} p_{j's}^{-1/\nu} \exp(-\tilde{\tau}d(i,j') - \tilde{\varphi}\mathbb{1}_{i \neq j'}) \cdot V(j')^{\tilde{\beta}}}. \tag{A.5}$$

where $\pi_{ij} = \sum_s \alpha_s \pi_{ij}^s$, $\boldsymbol{\Pi} = (\pi_{ij})$, *and* $\boldsymbol{\pi}_j^s = (\pi_{1j}^s, \cdots, \pi_{Jj}^s)^\intercal$.

Assumption A.1 (Parameter Restriction). We assume $1 + \frac{1}{\nu}, \sigma \in (1, \rho)$.

We first log linearize the equilibrium conditions, (A.1)–(A.5). Then, we view these linearized equations as exact equilibrium conditions and take the first-order approximation up to o($\boldsymbol{x}$) where $\boldsymbol{x} = (\beta, (\pi^s_{ik})_{i \neq k, s})'$, i.e., we assume that β and π^s_{ij} are small for $i \neq j$ and ignore second-order terms.[42] To simplify notation, we write log deviations as $n_{jsd} = \mathrm{d}\log N_{jsd}$, $\hat{C}_{jsd} = \mathrm{d}\log C_{jsd}$, $\tilde{a}_{jsd} = \mathrm{d}\log \tilde{A}_{jsd}$, $\hat{c}_{jsd} = \mathrm{d}\log c_{jsd}$, $\hat{p}_{js} = \mathrm{d}\log p_{js}$, $r_{js} = \mathrm{d}\log R_{js}$, $\hat{\pi}^s_{ij} = \mathrm{d}\log \pi^s_{ij}$, 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. Finally, we define two more variables $\breve{a}_{js} \equiv \tilde{a}_{js} - \hat{c}_{js}$, which captures the combined effect of changes in productivity and input costs, and $\breve{a}^*_{js} = \breve{a}_{js} - \frac{1}{\rho-1}\hat{C}_{js}$, which additionally captures the effect of changes in operating costs. Equilibrium in terms of log deviations is characterized in the following lemma.

Lemma A.2 (Log linearization). *When $e_i = 0$ for all i, we have*

$$\begin{aligned}
\hat{p}_{js} &= -\breve{a}^*_{js} - \tfrac{1}{\rho-1} r_{js} \\
r_{js} &= \sum_i (1-\beta)\lambda^s_{ij}\hat{\pi}^s_{ij} + \sum_i \sum_k \sum_{s'} \beta \lambda^{s's}_{ikj} (\hat{\pi}^{s'}_{ik} + \hat{\pi}^s_{kj}) \\
\hat{\pi}^s_{ij} &= \sum_{j'} (\mathbb{1}_{j'=j} - \pi^s_{ij'}) \left(-\tfrac{1}{\nu}\hat{p}_{j's} - \tilde{\beta} \sum_{s'} \alpha_{s'} \sum_{j''} \pi^{s'}_{j'j''} \hat{p}_{j''s'} \right) + o(\boldsymbol{x}) \\
n_{jsd} &= \tfrac{\rho-1}{\rho-\sigma} \big(-\hat{C}_{jsd} + (\sigma-1)\breve{a}_{jsd} + (\sigma-1)\hat{p}_{js} + r_{js} \big)
\end{aligned}$$

where $\lambda^s_{ij} = \dfrac{E_i \pi^s_{ij}}{(1-\beta)\sum_i E_i \pi^s_{ij} + \beta \sum_{iks'} E_i \alpha_{s'} \pi^{s'}_{ik} \pi^s_{kj}}$ *and* $\lambda^{s's}_{ikj} = \dfrac{E_i \alpha_{s'} \pi^{s'}_{ik} \pi^s_{kj}}{(1-\beta)\sum_i E_i \pi^s_{ij} + \beta \sum_{iks'} E_i \alpha_{s'} \pi^{s'}_{ik} \pi^s_{kj}}$.

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 A.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:

$$R_{i \to js} \equiv (1-\beta)\alpha_s \big(0, \cdots, 0, \overbrace{E_i}^{i\text{th}}, 0, \cdots, 0\big)(I - \beta\Pi)^{-1} \boldsymbol{\pi}^s_j.$$

We assume that zone j_0 is *small* in the sense that

$$\pi^s_{jj_0}, \frac{R_{j_0 \to js}}{R_{js}} = o(\boldsymbol{x}) \text{ for } j \neq j_0 \text{ and } \sum_{j \neq j_0} \pi^s_{jj_0} = O(1).$$

[42] 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^s_{ij} converges to zero. We write $A = B + o(\boldsymbol{x}^k)$ hereafter when $\lim_{J \to \infty} \frac{A_J - B_J}{\|\boldsymbol{x}^k_J\|} = 0$ for $k \in \mathbb{N}_0$ and write $A = B + O(\boldsymbol{x}^k)$ when $\limsup_{J \to \infty} \frac{A_J - B_J}{\|\boldsymbol{x}^k_J\|} < \infty$.

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 A.2.

Proposition A.1. *Under Assumption A.2, equilibrium is characterized by (up to $o(\boldsymbol{x})$)*

$$\begin{aligned}
\hat{p}_{js}, r_{js} &= 0 \quad \textit{for } j \neq j_0 \\
\hat{p}_{j_0 s} &= -\check{a}^*_{j_0 s} - \tfrac{1}{\rho-1} r_{j_0 s} \\
r_{j_0 s} &= \Psi_{j_0 s} \cdot \hat{q}_{j_0,s} + \sum_{s'} \Phi^{s'}_{j_0 s} \cdot \hat{q}_{j_0 s'} \\
\hat{\pi}^s_{ij} &= \begin{cases} 0 & \textit{if } i, j \neq j_0 \\ -\pi^s_{j_0 j_0} \hat{q}_{j_0 s} & \textit{if } i = j_0, j \neq j_0 \\ \hat{q}_{j_0 s} & \textit{if } i \neq j_0, j = j_0 \\ (1 - \pi^s_{j_0 j_0}) \hat{q}_{j_0 s} & \textit{if } i = j = j_0 \end{cases}
\end{aligned}$$

where $\hat{q}_{j_0 s} = -\frac{1}{\nu}\hat{p}_{j_0 s} - \tilde{\beta} \sum_{s'} \alpha_{s'} \pi^{s'}_{j_0 j_0} \hat{p}_{j_0 s'}$ *and*

$$\begin{aligned}
\Psi_{j_0 s} &= \left(1 - \left((1-\beta)\lambda^s_{j_0 j_0} + \sum_{is'} \beta \lambda^{s's}_{i j_0 j_0}\right) \pi^s_{j_0 j_0}\right) \\
\Phi^{s'}_{j_0 s} &= \left(\sum_i \beta \lambda^{s's}_{i j_0 j_0} (1 - \pi^s_{j_0 j_0} \cdot \mathbb{1}_{i = j_0})\right).
\end{aligned}$$

For future reference, note that $\Psi_{j_0 s}, \Phi^{s'}_{j_0 s} \in (0,1)$ *and that* $\Phi^{s'}_{j_0 s} = o(1)$.

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.[43]

Corollary A.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*

$$\hat{p}_{j_0 s} < 0, r_{j_0 s} > 0, \hat{\pi}^s_{j_0 j_0} > 0, \hat{\pi}^s_{i j_0} > 0, \hat{\pi}^s_{j_0 j} < 0, \text{ and } n_{j_0 s} > 0 \tag{A.6}$$

for $i, j \neq j_0$.

Proposition A.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 A.2. *Under Assumption A.2 and up to $o(\boldsymbol{x})$, we have*

$$n_{j_0 s} = \gamma^1_{j_0 s} \check{a}^*_{j_0 s} - \hat{C}_{j_0 s} + \sum_{s' \neq s} \tau^{2s'}_{j_0 s} \check{a}^*_{j_0 s'} \tag{A.7}$$

$$n_{j_0 sd} = \gamma^1_{j_0 s} \check{a}^*_{j_0 s} - \hat{C}_{j_0 sd} + \kappa(\check{a}^*_{j_0 sd} - \check{a}^*_{j_0 s}) + \sum_{s' \neq s} \tau^{2s'}_{j_0 s} \check{a}^*_{j_0 s'} \tag{A.8}$$

[43] 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.

where
$$\kappa = \frac{(\sigma-1)(\rho-1)}{(\rho-\sigma)}, \quad \gamma^1_{j_0 s} = \Theta_{j_0 s}\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}}, \quad \text{and} \quad \tau^{2s'}_{j_0 s} = \Lambda^{s'}_{j_0 s}\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}}\frac{1}{1-\frac{\Theta_{j_0 s'}}{\rho-1}}.$$

The shocks $\{\breve{a}^*_{j_0 sd}\}$ and $\{\hat{C}_{j_0 sd}\}$ 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 A.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*

$$\nu(\sigma-1) > (1 - \lambda^s_{j_0 j_0}\pi^s_{j_0 j_0}). \tag{A.9}$$

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, we can write

$$\begin{aligned} a_{jsd} &= \bar{a}_{jsd} + \varepsilon_a \cdot \mathrm{d}\log\Upsilon_{2j} \\ \hat{C}_{jsd} &= \bar{\hat{C}}_{jsd} - \varepsilon_c \cdot \mathrm{d}\log\Upsilon_{2j} \end{aligned}$$

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 A.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 A.3. *Under Assumption A.2 and up to $o(\boldsymbol{x})$, we have*

$$n_{j_0 s} = \tilde{\gamma}^1_{j_0 s}\bar{\breve{a}}^*_{j_0 s} + \check{\gamma}^1_{j_0 s}\bar{\hat{C}}_{j_0 s} + \sum_{s' \neq s}\tilde{\tau}^{2s'}_{j_0 s}\bar{\breve{a}}^*_{j_0 s'} + \sum_{s' \neq s}\check{\tau}^{2s'}_{j_0 s}\bar{\hat{C}}_{j_0 s'} \tag{A.10}$$

$$n_{j_0 sd} = n_{j_0 s} - (\bar{\hat{C}}_{j_0 sd} - \bar{\hat{C}}_{j_0 s}) + \kappa(\bar{\breve{a}}^*_{j_0 sd} - \bar{\breve{a}}^*_{j_0 s}). \tag{A.11}$$

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 Justification of IV Specification

In this section, we interpret the reduced-form estimates in Section 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

$$\begin{aligned} n_{js} &= \gamma^1_{js}\breve{a}^*_{js} - \hat{C}_{js} + \sum_{s' \neq s}\tau^{2s'}_{js}\breve{a}^*_{js'} \\ n_{jsd} &= n_{js} + \kappa(\breve{a}^*_{jsd} - \breve{a}^*_{js}) - (\hat{C}_{jsd} - \hat{C}_{js}), \end{aligned}$$

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).[44] In Section 4, we propose an estimation method that exploits exogenous variations from shift-share design, while taking into account spatial linkages. We start with a formal definition of Bartik instruments.

Definition A.1. For some given weights $\{\varphi_{js}\}_{j,s}$ with $\sum_j \varphi_{js} = 1$, we can define *Bartik Instruments* as

$$n^{\text{Bartik}}_{js} \equiv \sum_d \theta_{jsd,0} n_{sd} \quad \text{where } n_{sd} = \sum_j \varphi_{js} n_{jsd}.$$

For future reference, we also define quasi-Bartik instruments $\tilde{n}^{\text{Bartik}}_{js} \equiv \sum_d \theta_{jsd} n_{sd}$, 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}.[45] We then interact these subsector-level changes with the subsector composition of (j, s) to obtain n^{Bartik}_{js}.

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.[46] Formally, our assumptions are as follows.

Assumption A.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 $\theta_{jsd,0} \perp_j \begin{pmatrix}\tilde{\varepsilon}_{js'} & \hat{\varepsilon}_{js'} & \tilde{\varepsilon}_{js'd'}\end{pmatrix}^{\mathsf{T}} \big| \vec{c}_j$, for all s, s', d, d' where $\vec{c}_j = \begin{pmatrix}\hat{c}_{j1} & \hat{c}_{j2} & \hat{c}_{j3} & 1\end{pmatrix}^{\mathsf{T}}$.[47]

[44] When there are shocks also to other regions, the combined effect of them can be greater than $o(\boldsymbol{x})$.

[45] 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 $\varphi'_{j_1 sd_1} > \varphi'_{j_2 sd_1}$ and $\varphi'_{j_1 sd_2} < \varphi'_{j_2 sd_2}$. Then, if there is a positive regional shock to j_1, this would increase n_{sd_1} relative to n_{sd_2}.

[46] 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.

(ii″) (Relevance) The 3×3 matrix whose (s, s')-element is $\sum_{d,d'} \tilde{a}_{sd}\tilde{a}_{s'd'}\mathbb{E}_j\left[\theta^{\perp}_{jsd}\theta^{\perp}_{js'd',0}\right]$ is invertible.[48]

(iii) (Homogeneous Effect) Initially all regions are symmetric *at the sector level* and sectors are symmetric so that the coefficients γ^{1}_{js} and $\tau^{2s'}_{js}$ are no longer j- or s-specific, and we write them as γ and τ, respectively.

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 A.3. *Under Assumption A.3, we can write*

$$n_{js} = \tfrac{\gamma}{\kappa}\tilde{n}^{\textit{Bartik}}_{js} + \sum_{s'\neq s} \tfrac{\tau}{\kappa}\tilde{n}^{\textit{Bartik}}_{js'} - \gamma\hat{c}_{js} - \sum_{s'\neq s}\tau\hat{c}_{js'} + \mathsf{FE}_s + \varepsilon_{js}$$
$$n_{jsd} = \tfrac{\gamma-\kappa}{\kappa}\tilde{n}^{\textit{Bartik}}_{js} + \sum_{s'\neq s} \tfrac{\tau}{\kappa}\tilde{n}^{\textit{Bartik}}_{js'} - \gamma\hat{c}_{js} - \sum_{s'\neq s}\tau\hat{c}_{js'} + \mathsf{FE}_{sd} + \check{\varepsilon}_{jsd}.$$

where $n^{\textit{Bartik}}_{js''} \perp_j \begin{pmatrix}\varepsilon_{js} & \check{\varepsilon}_{jsd}\end{pmatrix}^{\mathsf{T}} | \vec{c}_j$*, for all* s, s'', d.

Corollary A.3. *Under Assumption A.3, we can write*

$$n_{jsd} = \tfrac{\tau}{\gamma}n_{js'} + \left(\tfrac{\gamma}{\kappa} - 1 - \tfrac{\tau^2}{\kappa\gamma}\right)\tilde{n}^{\textit{Bartik}}_{js} + \left(\tfrac{\tau}{\kappa} - \tfrac{\tau^2}{\kappa\gamma}\right)\tilde{n}^{\textit{Bartik}}_{js''} + \left(\tfrac{\tau^2}{\gamma} - \gamma\right)\hat{c}_{js} + \left(\tfrac{\tau^2}{\gamma} - \tau\right)\hat{c}_{js''} + \mathsf{FE}_{sd} + \mathsf{FE}_{s'} + \grave{\varepsilon}_{jsd,s'} \tag{A.12}$$

$$n_{jsd} = \tfrac{\gamma-\kappa}{\gamma}n_{js} + \tfrac{\tau}{\gamma}\tilde{n}^{\textit{Bartik}}_{js'} + \tfrac{\tau}{\gamma}\tilde{n}^{\textit{Bartik}}_{js''} - \kappa\hat{c}_{js} - \tfrac{\tau\kappa}{\gamma}\hat{c}_{js'} - \tfrac{\tau\kappa}{\gamma}\hat{c}_{js''} + \mathsf{FE}_{sd} + \grave{\varepsilon}_{jsd} \tag{A.13}$$

where $n^{\textit{Bartik}}_{js''} \perp_j \begin{pmatrix}\grave{\varepsilon}_{jsd,s'} & \grave{\varepsilon}_{jsd}\end{pmatrix}^{\mathsf{T}} | \vec{c}_j$*, for all* s, d, s', s''.

These representations (A.12−A.13) directly yield the following two propositions.

Proposition A.4 (IV−Across Sector). *If we regress* n_{jsd} *on* $(n_{js'}, \tilde{n}^{\textit{Bartik}}_{js}, \tilde{n}^{\textit{Bartik}}_{js''})$*, instrumented by* $(n^{\textit{Bartik}}_{js}, n^{\textit{Bartik}}_{js'}, n^{\textit{Bartik}}_{js''})$*, controlling for* $\vec{c}_j$*,* 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 A.5 (IV−Within Sector). *If we regress* n_{jsd} *on* $(n_{js}, \tilde{n}^{\textit{Bartik}}_{js'}, \tilde{n}^{\textit{Bartik}}_{js''})$*, instrumented by* $(n^{\textit{Bartik}}_{js}, n^{\textit{Bartik}}_{js'}, n^{\textit{Bartik}}_{js''})$*, controlling for* $\vec{c}_j$ *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* (A.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 $(\beta, \varepsilon_c, \varepsilon_a) \to (0, 0, 0)$.

B.4 Efficiency Properties of Trip Chaining and External Economies of Scale (Section 3.3)

B.4.1 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

[47] In terms of the framework of GSS, what we need is orthogonality between θ_0 and the structural error terms. The proof of Lemma A.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.

[48] For a variable x_j that has a region index j, we define the residualized version $x^{\perp}_j$ as the j-th residual from the regression of x_j on $\vec{c}_j$. Using this notation, we can rewrite the exogeneity assumption as $\theta^{\perp}_{jsd,0} \perp_j \begin{pmatrix}\tilde{\varepsilon}^{\perp}_{js'} & \hat{\varepsilon}^{\perp}_{js'} & \tilde{\varepsilon}^{\perp}_{js'd'}\end{pmatrix}^{\mathsf{T}}$.

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.

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

$$q_j = \int_0^{N_j} f_j(q_j(\omega))\,\mathrm{d}\omega \quad \text{(e.g., CES: } f_j(q) = q^{1-1/\rho_j})$$

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,

$$\begin{aligned} \max_{\{q_j(\omega)\}_{j,\omega}} \quad & U = U(\{q_j\}_{j,}) \\ \text{s.t.} \quad & \sum_j \int_0^{N_j} q_j(\omega) p_j(\omega)\,\mathrm{d}\omega \leq w \end{aligned}$$

where $p_j(\omega)$ is the price and w is income. The first-order condition is given by

$$\frac{\partial U}{\partial q_j} \cdot f_j'(q_j(\omega)) = \lambda p_j(\omega) \tag{A.14}$$

where λ is the Lagrange multiplier associated with the budget constraint. Thus, monopolistically competitive firms solve the profit maximization problem,

$$\max_{q_j(\omega)} \left\{ \frac{1}{\lambda} \frac{\partial U}{\partial q_j} f_j'(q_j(\omega)) q_j(\omega) - c_j q_j(\omega) \right\}.$$

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 A.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.*[49]

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

[49] Elasticity of substitution should not be j-specific. See the proof of Lemma A.6.

equilibrium, an agent of type i solves

$$\max_{\{q_j^i(\omega)\}_{j,\omega}} \quad U^i = U^i(\{q_j^i\}_j) \quad \text{where } q_j^i = \int_0^{N_j} f_j(q_j^i(\omega))\, \mathrm{d}\omega \tag{A.15}$$
$$\text{s.t.} \quad \sum_j \int_0^{N_j} p_j(\omega) q_j^i(\omega)\, \mathrm{d}\omega \leq w^i$$

The first-order condition is given by

$$\frac{\partial U^i}{\partial q_j^i} \cdot f_j'(q_j^i(\omega)) = \lambda^i \cdot p_j(\omega), \tag{A.16}$$

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

$$\max_{p_j(\omega)} \left\{ (p_j(\omega) - c_j) \sum_i q_j^i(p_j(\omega)) \right\}. \tag{A.17}$$

In this economy, the following proposition extends the result of Proposition A.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 A.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:*

$$\max_{\{N_j\}_j, \{q_j^i(\omega)\}_{j,\omega,i}} \quad \sum_i w^i \cdot \log U^i \qquad (w^i \textit{ acts as a Pareto weight}) \tag{SP}$$
$$\text{s.t.} \quad \sum_i \sum_j \int_0^{N_j} c_j q_j^i(\omega)\, \mathrm{d}\omega + \sum_j E_j N_j \leq \sum_i w^i,$$

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 (A.15), and firms solve the profit maximization problem (A.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

$$\Upsilon_{1J(j)} = \sum_{j' \in J(j)} \sum_i \int_0^{N_{j'}} c_{j'} q_{j'}^i(\omega)\, \mathrm{d}\omega \ \text{ and } \ \Upsilon_{2J(j)} = \sum_{j' \in J(j)} E_{j'} N_{j'}$$

are the total resource spent on production and variety creation for partition $J(j)$, respectively.[50]

[50] 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 A.4 or A.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 A.4. This gives $\Upsilon_{1J(j)} = \sum_{j' \in J(j)} c_{j'} q_{j'} = \left(\sum_{j' \in J(j)} \bar{c}_{j'} q_{j'}\right) \Upsilon_{1J(j)}^{\varepsilon_c} = \left(\sum_{j' \in J(j)} \bar{c}_{j'} q_{j'}\right)^{1/(1-\varepsilon_c)}$.

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:

$$\max_{\{N_j\}_j,\{q^i_j(\omega)\}_{j,\omega,i}} \quad \sum_{j,i}\int_0^{N_j} \underbrace{\frac{1}{\lambda^i}\frac{\partial U^i}{\partial q^i_j}\bigg|_{\text{de}} \cdot f'_j(q^i_j(\omega))}_{\text{from (A.16)}} \cdot q^i_j(\omega)\,\mathrm{d}\omega \tag{CRM}$$

$$\text{s.t.} \quad \sum_{j,i}\int_0^{N_j} c_j(\Upsilon_{1J(j)},\Upsilon_{2J(j)})q^i_j(\omega)\,\mathrm{d}\omega + \sum_j E_j(\Upsilon_{1J(j)},\Upsilon_{2J(j)})N_j \le \sum_i w_i$$

where the value of $\frac{1}{\lambda^i}\frac{\partial U^i}{\partial q^i_j}$ is evaluated at the decentralized allocation so that the term $\frac{1}{\lambda^i}\frac{\partial U^i}{\partial q^i_j}\Big|_{\text{de}} \cdot f'_j(q^i_j(\omega))$ captures the residual demand firm j faces in the decentralized equilibrium, and we take this value as given when solving problem (CRM). Lemmas A.4 and A.5 summarize the results. The proofs are given in Appendix F.

Lemma A.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.*[51]

Lemma A.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.[52,53]

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 A.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 $\{q^i_j(\omega)\}_{j,\omega} \mapsto U^i$ is homogeneous of degree s for some $s > 0$, the solution of (CRM) coincides*

[51] If ρ_j is the same across all j, we only need $\varepsilon_{c1} + \frac{1}{\rho-1}\varepsilon_{E1} = (\rho-1)\varepsilon_{c2} + \varepsilon_{E2}$.

[52] If ρ_j is the same across all j, we do not need any condition on ε_c and ε_E.

[53] 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 $p^i_j(\omega)$ 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 $p^i_j(\omega) = p^{i'}_j(\omega)$ 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).

with the solution of the following social planner problem:

$$\max_{\{N_j\}_j, \{q_j^i(\omega)\}_{j,\omega,i}} \quad \sum_i w^i \cdot \log U^i \qquad (w^i \textit{ acts as a Pareto weight}) \tag{SP-EES}$$

$$\text{s.t.} \quad \sum_{j,i} \int_0^{N_j} c_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) q_j^i(\omega)\, \mathrm{d}\omega + \sum_j E_j(\Upsilon_{1J(j)}, \Upsilon_{2J(j)}) N_j \leq \sum_i w^i.$$

Lemmas A.4–A.6 immediately prove Proposition A.8, which characterizes conditions under which the decentralized allocation is efficient. Propositions A.6 and A.7 are special cases of Proposition A.8 with homogeneous agent, $I = 1$, and without external economies of scale.

Proposition A.8. *When (i) c_j and E_j satisfy the conditions of either Lemma A.4 or A.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).*

B.4.2 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:

$$\max_{\{x_j(\omega;\sigma)\}_{j,\omega,\sigma}} \quad \tilde{U}(\{x_j(\sigma)\}_{j,\sigma}) \quad \text{where } \ x_j(\sigma) = F(\{x_j(\omega;\sigma)\}_\omega) \tag{A.18}$$

$$\text{s.t.} \quad \sum_j \sum_\sigma \pi(\sigma) \sum_\omega c_j(\omega; \Upsilon_{1J(j)}) \cdot x_j(\omega;\sigma) \leq I$$

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 Appendix F.

Lemma A.7 (Two-step Maximization). *We can solve problem* (A.18) *in two steps. First, we solve the problem for given values of $\{x_j\}_j$:*

$$U(\{x_j\}_j) = \max_{\{x_j(\sigma)\}_{j,\sigma}} \quad \tilde{U}(\{x_j(\sigma)\}_{j,\sigma})$$

$$\text{s.t.} \quad \sum_\sigma \pi(\sigma) x_j(\sigma) \leq x_j, \quad \forall j.$$

Second, we choose $\{x_j\}_j$ that maximize U subject to the resource constraint:

$$\begin{aligned} \max_{\{x_j\}_j} \quad & U(\{x_j\}_j) \\ \text{s.t.} \quad & \sum_j c(\{c_j(\omega; \Upsilon_{1J(j)})\}_\omega) \cdot x_j \leq I. \end{aligned}$$

We can easily see that this is equivalent to

$$\begin{aligned} \max_{\{x_j(\omega)\}_{j,\omega}} \quad & U(\{x_j\}_j) \quad \text{where } \ x_j = F(\{x_j(\omega)\}_\omega) \\ \text{s.t.} \quad & \sum_j \sum_\omega c_j(\omega; \Upsilon_{1J(j)}) \cdot x_j(\omega) \leq I. \end{aligned}$$

Trip Chaining. To apply Propositions A.6–A.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

$$\begin{aligned} \mathcal{U}^C_i = \max_{\tilde{C}_i, C^r_i} \quad & \left(\frac{\tilde{C}_i}{1-\mu}\right)^{1-\mu} \left(\frac{C^r_i}{\mu}\right)^{\mu} \\ \text{s.t.} \quad & \tilde{C}_i + P_i C^r_i \leq I_i \end{aligned} \tag{A.19}$$

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

$$\begin{aligned} & \max_{\{j^i_t(\sigma^t)\}, \{q^i(\sigma^t)\}, \{q^i_d(\sigma^t)\}, \{q^i_d(\omega;\sigma^t)\}, \tilde{C}_i} \quad \mathcal{U}^C_i \\ & \text{where} \quad C^r_i = \exp((1-\beta)V_i) \\ & \qquad V_i = \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U\big(q^i(\sigma^t)\big) - \tau d(j_{t-1}(\sigma^{t-1}), j^i_t(\sigma^t)) - \varphi \mathbb{1}_{j_{t-1}(\sigma^{t-1}) \neq^i j^i_t(\sigma^t)} + \nu \varepsilon_t^{j^i_t(\sigma^t)} \Big) \pi(\sigma^t) \\ & \qquad q^i(\sigma^t) = \left(\sum_d \phi^{1/\sigma}_{j^i_t(\sigma^t) s(\sigma^t) d} q^i_d(\sigma^t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}} \\ & \qquad q^i_d(\sigma^t) = \left(\int_0^{N_{j^i_t(\sigma^t) s(\sigma^t) d}} q^i_d(\omega; \sigma^t)^{1-1/\rho} \, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\ & \text{s.t.} \quad \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \left(\sum_d \int_0^{N_{j^i_t(\sigma^t) s(\sigma^t) d}} p_{j^i_t(\sigma^t) s(\sigma^t) d}(\omega) \cdot q^i_d(\omega; \sigma^t) \, \mathrm{d}\omega \right) \pi(\sigma^t) + \tilde{C}_i \leq I_i. \end{aligned} \tag{DE}$$

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, $(1-\mu)\sum_i I_i L_i$.

$$\max_{\{N_{jsd}\},\{j_t^i(\sigma^t)\},\{q^i(\sigma^t)\},\{q_d^i(\sigma^t)\},\{q_d^i(\omega;\sigma^t)\},\{\tilde{C}_i\}} \quad \sum_i \theta_i \log \mathcal{U}_i^C L_i \tag{SP}$$

$$\begin{aligned}
\text{where} \quad & C_i^r = \exp((1-\beta)V_i) \\
& V_i = \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U\big(q^i(\sigma^t)\big) - \tau d(j_{t-1}^i(\sigma^{t-1}), j_t^i(\sigma^t)) - \varphi \mathbb{1}_{j_{t-1}^i(\sigma^{t-1}) \neq j_t^i(\sigma^t)} + \nu \varepsilon_t^{j_t(\sigma^t)} \Big) \pi(\sigma^t) \\
& q^i(\sigma^t) = \left(\sum_d \phi_{j_t^i(\sigma^t)s(\sigma^t)d}^{1/\sigma} q_d^i(\sigma^t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}} \\
& q_d^i(\sigma^t) = \left(\int_0^{N_{j_t^i(\sigma^t)s(\sigma^t)d}} q_d^i(\omega;\sigma^t)^{1-1/\rho}\, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\
\text{s.t.} \quad & \sum_i \left(\sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \sum_d \int_0^{N_{j_t^i(\sigma^t)s(\sigma^t)d}} \frac{c_{j_t^i(\sigma^t)s(\sigma^t)d}}{A_{j_t^i(\sigma^t)s(\sigma^t)d}} \cdot q_d^i(\omega;\sigma^t)\, \mathrm{d}\omega \pi(\sigma^t) \right) L_i + \sum_{j,s,d} N_{jsd} C_{jsd} + \sum_i \tilde{C}_i L_i \leq \sum_i I_i L_i \\
& \sum_i \tilde{C}_i L_i = (1-\mu) \sum_i I_i L_i
\end{aligned}$$

The unconstrained social planner solves the same problem without the last constraint.

We can apply the result of Appendix B.4.2 to reformulate (DE) and (SP) to two-step maximization problems:[54] First, we compute the maximized utility from non-tradable services for given values of $\{q_{jsd}^i\}$, and the maximized value is denoted by $C^i(\{q_{jsd}^i\})$:

$$\begin{aligned}
C^i(\{q_{jsd}^i\}) = \max_{\{j_t^i(\sigma^t)\},\{q^i(\sigma^t)\},\{q_d^i(\sigma^t)\}} \quad & C_i^r = \exp((1-\beta)V_i) \\
\text{where} \quad & V_i = \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \Big(U\big(q^i(\sigma^t)\big) - \tau d(j_{t-1}^i(\sigma^{t-1}), j_t^i(\sigma^t)) - \varphi \mathbb{1}_{j_{t-1}^i(\sigma^{t-1}) \neq j_t^i(\sigma^t)} + \nu \varepsilon_t^{j_t^i(\sigma^t)} \Big) \pi(\sigma^t) \\
& q^i(\sigma^t) = \left(\sum_d \phi_{j_t^i(\sigma^t)s(\sigma^t)d}^{1/\sigma} q_d^i(\sigma^t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}} \\
\text{s.t.} \quad & \sum_{t=0}^{\infty} \beta^t \sum_{\sigma^t} \mathbb{1}_{j_t^i(\sigma^t)=j, s(\sigma^t)=s} \cdot q_d^i(\sigma^t) \pi(\sigma^t) \leq q_{jsd}^i, \quad \forall j, s, d.
\end{aligned}$$

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. Applying the result of Appendix B.4.2, an equivalent formulation of (DE) is

$$\begin{aligned}
\max_{\{q_{jsd}^i\},\{q_{jsd}^i(\omega)\},\tilde{C}_i} \quad & \log \tilde{C}_i^{1-\mu} \cdot (C^i(\{q_{jsd}^i\})^{\mu} \qquad \text{(DE}'\text{)} \\
\text{where} \quad & q_{jsd}^i = \left(\int_0^{N_{jsd}} q_{jsd}^i(\omega)^{1-1/\rho}\, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\
\text{s.t.} \quad & \sum_{j,s,d} \int_0^{N_{jsd}} p_{jsd}(\omega) q_{jsd}^i(\omega)\, \mathrm{d}\omega + \tilde{C}_i \leq I_i.
\end{aligned}$$

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),

[54] The index σ in Appendix B.4.2 corresponds to σ^t here, ω to ω, j to (j,s,d), and $\pi(\sigma)$ to $\beta^t \cdot \pi(\sigma^t)$.

$$
\begin{aligned}
&\max_{\{N_{jsd}\},\{q^i_{jsd}\},\{q^i_{jsd}(\omega)\},\{\tilde{C}_i\}} \quad \sum_i \theta_i \log U_i^C L_i && \text{(SP}'\text{)}\\
&\text{where} \quad U_i^C = \tilde{C}_i^{1-\mu}(C^i(\{q^i_{jsd}\})^{\mu} \\
&\qquad\quad q^i_{jsd} = \left(\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho}\,\mathrm{d}\omega\right)^{\frac{\rho}{\rho-1}} \\
&\quad \text{s.t.} \quad \sum_i \left(\sum_{j,s,d}\int_0^{N_{jsd}} \tilde{c}_{jsd}(\Upsilon_{1j},\Upsilon_{2j}) q^i_{jsd}(\omega)\,\mathrm{d}\omega\right) L_i + \sum_{j,s,d} N_{jsd} C_{jsd}(\Upsilon_{1j},\Upsilon_{2j}) + \sum_i \tilde{C}_i L_i \le \sum_i I_i L_i \\
&\qquad\quad \sum_i \tilde{C}_i L_i = (1-\mu)\sum_i I_i L_i
\end{aligned}
$$

where $\tilde{c}_{jsd}(\Upsilon_{1j},\Upsilon_{2j}) = \dfrac{c_{jsd}(\Upsilon_{1j},\Upsilon_{2j})}{A_{jsd}(\Upsilon_{1j},\Upsilon_{2j})}$.

B.4.3 Application: Efficiency Properties of Trip Chaining

In this section, we apply Propositions A.6 and A.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 A.9. *In the economy with homogeneous consumers, the decentralized allocation solves the constrained social planner problem (SP).*

This is a direct application of Proposition A.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.

Proposition 2–Part 1. *When trip chaining is not allowed ($\beta = 0$), the decentralized equilibrium solves the constrained social planner problem with Pareto weights $\theta_i = I_i$, for all i. Furthermore, even when trip chaining is allowed ($\beta > 0$), the decentralized equilibrium solves the constrained social planner problem with the same Pareto weights.*

To prove this result, we can once again apply Proposition A.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:

$$
\begin{aligned}
\max_{\{q^i_{jsd}\},\{q^i_{jsd}(\omega)\},\tilde{C}_i} \quad & \log\, \tilde{C}_i^{1-\mu} \cdot (C^i(\{q^i_{jsd}\})^{\mu} \qquad \text{(DE}'\text{)} \\
\text{where} \quad & q^i_{jsd} = \left(\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho}\, \mathrm{d}\omega \right)^{\frac{\rho}{\rho-1}} \\
\text{s.t.} \quad & \sum_{j,s,d} \int_0^{N_{jsd}} p_{jsd}(\omega) q^i_{jsd}(\omega)\, \mathrm{d}\omega + p^{\text{tradable}}\tilde{C}_i \leq I_i + T_i
\end{aligned}
$$

and the free-entry condition is given by:

$$
\bar{C}_{jsd} = \sum_i (\bar{p}_{jsd}(\omega) - \tilde{c}_{jsd}) q^i_{jsd}(\omega) L_i.
$$

The next proposition shows 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. 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.

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

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

where N^{LF}_{jsd} represents the number of stores in the laissez-faire equilibrium and $\chi = \frac{\rho}{\rho-(1-\mu)} > 1$ is a constant that remains unaffected by the presence of trip chaining. This optimal allocation can be implemented through a combination of a tax on tradable goods and a subsidy for non-tradable services: $t^{tradable} = \frac{\mu}{\rho-1}$, $s_{jsd}(\omega) = \frac{1-\mu}{\rho}$, $S_{jsd} = 0$, and $T_i = 0$.

B.4.4 Application: Efficiency Properties of External Economies of Scale

Constrained-Efficient Specification. In this section, we first apply Proposition A.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 A.8 to show the following result.

Proposition 3–Part 1. *Assume isoelastic external economies of scale of the form either*

$$\tilde{c}_{jsd} = \bar{c}_{jsd} \cdot \Upsilon_{1j}^{-\varepsilon} \text{ and } C_{jsd} = \bar{C}_{jsd} \cdot \Upsilon_{2j}^{-\varepsilon}$$

or

$$\tilde{c}_{jsd} = \bar{c}_{jsd} \cdot \Upsilon_{j}^{-\varepsilon} \text{ and } C_{jsd} = \bar{C}_{jsd} \cdot \Upsilon_{j}^{-\varepsilon}$$

where $\Upsilon_{1j} = \sum_{s,d} \tilde{c}_{jsd} \left(\sum_i \int_0^{N_{jsd}} q^i_{jsd}(\omega) \, \mathrm{d}\omega \right)$ and $\Upsilon_{2j} = \sum_{s,d} C_{jsd} N_{jsd}$ represent the total resources expended on production and variety creation for region j, respectively, and $\Upsilon_j = \Upsilon_{1j} + \Upsilon_{2j}$. Under these conditions, the decentralized equilibrium solves the social planner problem.

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 A.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 Appendix B.4.4. We can prove the following result, where Proposition 2−Part 2 is a special case with $\varepsilon = 0$.

Proposition A.11 (Unconstrained Social Planner). *Assume isoelastic external economies of scale of the form either*

$$\tilde{c}_{jsd} = \bar{c}_{jsd} \cdot \Upsilon_{1j}^{-\varepsilon} \text{ and } C_{jsd} = \bar{C}_{jsd} \cdot \Upsilon_{2j}^{-\varepsilon}$$

or

$$\tilde{c}_{jsd} = \bar{c}_{jsd} \cdot \Upsilon_{j}^{-\varepsilon} \text{ and } C_{jsd} = \bar{C}_{jsd} \cdot \Upsilon_{j}^{-\varepsilon}.$$

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.5 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.[55] Once workers determine a pair of districts (d^r, d^w),

[55] We use d to index districts to avoid confusion with i and j, which are indices for zones.

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

$$\mathcal{U}(o) = \max_{d^r, d^w} \ \mathbb{E}[B_{d^r} \cdot z_{d^r}(o) \cdot \mathcal{U}_{i^r i^w}(o)|d^r, d^w]$$

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

$$\tilde{d}_{i^r i^w} = \exp\big(\kappa d_{d(i^r)d(i^w)} + \varphi_\kappa \mathbb{1}_{d(i^r) \neq d(i^w)}\big)$$

where $d(i)$ denotes a district to which zone i belongs.[56] 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.[57]

For simplicity, we follow Tsivanidis (2019) to assume that workers first choose residential districts, and then choose business districts.[58] 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.

[56] 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).

[57] 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}.

[58] 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.

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.

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

$$\mathcal{U}_{d^r d^w}(o) = \sum_{i^r i^w} \Pr(i^r, i^w | d^r, d^w) P_{i^r}^{-\mu_c^r} P_{i^w}^{-\mu_c^w} \cdot (p^{tradable})^{-\mu_{\tilde{C}}} \cdot r_{d^r}^{-\mu_\ell} \cdot \frac{\xi_{d^w} \cdot w_{d^w} \cdot v_{d^w}(o)}{\exp(\kappa d_{d^r d^w} + \varphi_\kappa \mathbb{1}_{d^r \neq d^w})}. \quad \text{(A.20)}$$

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,

$$\begin{aligned}
\mathcal{U}_{d^r}(o) &= \mathbb{E}\Big[\max_{d^w}\{\mathcal{U}_{d^r d^w}(o)\}\Big| d^r\Big] \\
&= B_{d^r} \cdot z_{d^r}(o) \cdot \bar{P}_{d^r}^{-\mu_c^r} (p^{tradable})^{-\mu_{\tilde{C}}} r_{d^r}^{-\mu_\ell} \cdot \left(\sum_d \left(\bar{P}_d^{-\mu_c^w} \xi_d w_d \exp(-\kappa d_{d^r d} + \varphi_\kappa \mathbb{1}_{d^r \neq d})\right)^{\varepsilon_v}\right)^{1/\varepsilon_v} \\
&\equiv \mathcal{U}_{d^r} \cdot z_{d^r}(o)
\end{aligned}$$

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 A.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 $M_{i^r i^w}$, 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.

C. Appendix for Section 4

C.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 $\mathbb{S}_i^{\text{unobs}}$.

$$
\begin{aligned}
&\mathbb{P}^{\text{cal}} = \{\boldsymbol{\mu}, \boldsymbol{\alpha}, \rho, \gamma, \beta\} && \mathbb{P}^{\text{PE}} = \{\tilde{\tau}, \tilde{\varphi}, \varepsilon, \sigma, \nu\} && \mathbb{P}^{\text{GE}} = \{\kappa, \varepsilon_z, \varepsilon_v, \theta, \eta_A, \eta_B\} \\
&\mathbb{S}_{\text{exo}}^{\text{cal}} = \{\mathcal{J}, \mathcal{S}, \boldsymbol{\mathcal{D}}, \boldsymbol{d}, \boldsymbol{T}\} && \mathbb{S}_{\text{exo}}^{\text{PE}} = \{\tilde{\boldsymbol{A}}, \boldsymbol{C}\} && \mathbb{S}_{\text{exo}}^{\text{GE}} = \{\bar{\boldsymbol{B}}, \bar{\boldsymbol{A}}^{tradable}, \boldsymbol{\xi}\} \\
&\mathbb{S}_{\text{PE}}^{\text{obs}} = \{\boldsymbol{N}, \boldsymbol{R}\} && \mathbb{S}_{\text{PE}}^{\text{unobs}} = \{\boldsymbol{P}, \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{H}^{service}, \boldsymbol{L}^{service}, \Pi\} \\
&\mathbb{S}_{\text{GE}}^{\text{obs}} = \{\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{M}, \boldsymbol{I}, \boldsymbol{E}, M\} && \mathbb{S}_{\text{GE}}^{\text{unobs}} = \{\boldsymbol{H}^{r}, \boldsymbol{H}^{tradable}, \boldsymbol{L}^{tradable}\}
\end{aligned}
$$

where the bold letters denote vectors. The services market equilibrium model is essentially a mapping from $\mathbb{P}^{\text{cal}} \cup \mathbb{P}^{\text{PE}} \cup \mathbb{S}_{\text{exo}}^{\text{cal}} \cup \mathbb{S}_{\text{exo}}^{\text{PE}} \cup \mathbb{S}_{\text{GE}}^{\text{obs}}$ 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 A.8 (Equilibrium Inversion). *Given $\mathbb{P}^{cal}$, $\mathbb{P}^{PE}$, $\mathbb{S}_{exo}^{cal}$, and $\mathbb{S}_{GE}^{obs}$, there exist unique values of the location characteristics $\mathbb{S}_{exo}^{PE}$ that rationalize the observed data $\mathbb{S}_{PE}^{obs}$. Given $\mathbb{P}$, $\mathbb{S}_{exo}^{cal}$, and $\mathbb{S}_{exo}^{PE}$, there exist unique values of the location characteristics $\mathbb{S}_{exo}^{GE}$ that rationalize the observed data $\mathbb{S}_{PE}^{obs}$ and $\mathbb{S}_{GE}^{obs}$.*

Proof. Lemma A.8 is a direct application of Proposition 2 in Ahlfeldt et al. (2015). □

The first part of Lemma A.8 implies that, once we have data on $\mathbb{S}_{\text{GE}}^{\text{obs}}$, we can back out the composite productivity and operating costs without relying on the general equilibrium component of the model. In Section 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 $\mathbb{S}_{\text{GE}}^{\text{obs}}$. Thus, the estimation procedure remains robust to any potential misspecification of our general equilibrium model.

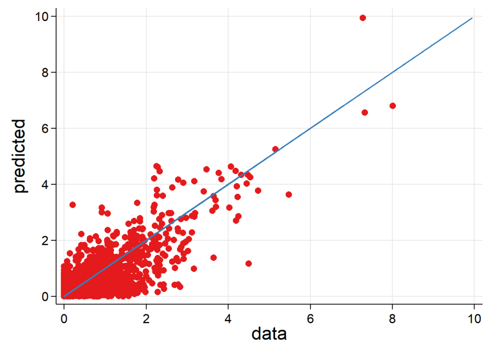


Figure A.3. Gravity fit

Notes: Data source: Korean Household Travel Survey (2016). The blue line is the 45-degree line.

C.2 Fit of the Estimated Gravity Equation

C.3 Parameter Estimation: General Equilibrium Model

In this section, we discuss the estimation procedure of general equilibrium. In Figure A.3, we plot the predicted the number of services travel against the observed number of services travel. It demonstrates that the flow of services travel is well approximated by the gravity equation.

We calibrate the GE parameters mostly from the literature or directly from the moments in the data. We summarize the parameter values in Table A.9.

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, $\mu_c^r + \mu_c^w$, 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 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

Table A.9: 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).

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.

Table A.10: 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.

D. Appendix for Sections 5 and 6

D.1 SMA and (Real) Income Inequality

Does SMA inequality exacerbate or alleviate welfare inequality between the rich and the poor? 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 A.10, 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, 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%.[59] Finally, in the last column of Table A.10, we compute real income, which additionally adjusts the differences in SMA—the inverse of the price index of services.[60] 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

[59] 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.

[60] We compute district-level SMA by taking an average using the population distribution,

$$\text{SMA}_{d^r} = \sum_{i^r,i^w} \Pr(i^r, i^w | d^r)\, \text{SMA}_{i^r}^{\frac{\mu_c^r}{\mu_c^r+\mu_c^w}}\, \text{SMA}_{i^w}^{\frac{\mu_c^w}{\mu_c^r+\mu_c^w}}.$$

this case, the price index for non-tradable goods is typically higher in urban areas, dampening the real income dispersion.[61] 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.

D.2 Transportation Improvement

The magnitude of spatial frictions is critical for the dispersion of SMA. For example, if one incorrectly assumes that a consumer can only purchase services goods from her own region—i.e., travel costs are infinite $\tau = \infty$—then SMA inequality would be 90% higher than the baseline estimate. This observation suggests that improving transportation infrastructure can be an effective measure to reduce SMA inequality. In this section, we explore how the distribution of services and SMA inequality change when transportation improves. We show that transportation improvement is an effective way to reduce SMA inequality, but the effectiveness is muted due to the endogenous response of services distribution.

To quantify these effects, we consider a counterfactual exercise in which we gradually reduce travel costs. We decrease travel costs parameters from the current levels (τ, φ) to $(\zeta\tau, \zeta\varphi), \zeta \in [0, 1]$. We decompose changes in SMA into two parts: direct effects and indirect effects. Direct effects isolate impacts of the decrease in travel costs, holding the distribution of services stores fixed. As a consumer can more easily access other regions, her current region becomes less important in determining SMA, leading to a decrease in SMA inequality. Indirect effects measure impacts through the endogenous change in the distribution of services stores.

We define two counterfactual measures of SMA. First, for direct effects, we recursively define

$$\mathrm{SMA}_i^D(\zeta) = \exp\left(\sum_s \nu\alpha_s \log\left(\sum_j \mathrm{e}^{-\zeta\frac{\tau}{\nu}d(i,j)-\zeta\frac{\varphi}{\nu}\mathbb{1}_{i\neq j}} p_{js}^{-1/\nu} \cdot (\mathrm{SMA}_j^D(\zeta))^{\frac{\beta}{\nu}}\right)\right)$$

where we fix p_{js} at the estimated levels, i.e., holding fixed the distribution of services. Second, for indirect effect, we define $\mathrm{SMA}_i^P(\zeta)$,

$$\mathrm{SMA}_i^P(\zeta) = \exp\left(\sum_s \nu\alpha_s \log\left(\sum_j \mathrm{e}^{-\frac{\tau}{\nu}d(i,j)-\frac{\varphi}{\nu}\mathbb{1}_{i\neq j}} p_{js}(\zeta)^{-1/\nu} \cdot (\mathrm{SMA}_j^P(\zeta))^{\frac{\beta}{\nu}}\right)\right)$$

where we fix travel costs but use counterfactual prices $p_{js}(\zeta)$ computed under the new equilibrium distribution of services stores.

In Figure A.4, we plot SMA inequality, measured by standard deviation of $\log(\mathrm{SMA})$, while varying transportation costs. From the left panel, we find that transportation improvement significantly decreases SAM inequality. About 30% decline in travel costs results in about an 18% decline in inequality. In the middle panel, we isolate direct effects and plot the standard deviation of $\log\mathrm{SMA}_i^D$. As transportation improves, the direct effect lowers SMA inequality as expected. Finally, in the right panel, we plot the standard deviation of $\log\mathrm{SMA}_i^P$. The endogenous response of the distribution of services has a

[61] 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.

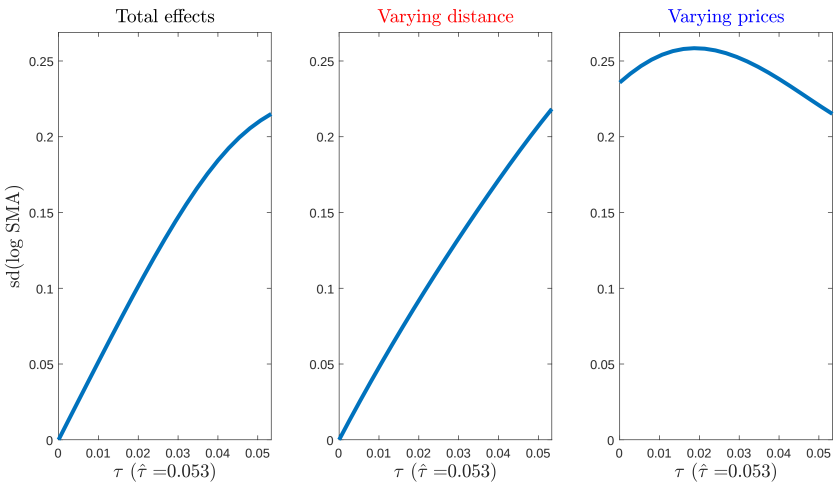


Figure A.4. Transportation improvement: SMA inequality

non-monotonic impact on inequality. Transportation improvement initially increases SMA dispersion. As travel becomes easier, consumers are more willing to travel to distant regions, further increasing services concentration and resulting in increased SMA inequality even without direct impacts. However, if we further decrease travel costs, geographic locations become less important, and SMA inequality starts to decline. This discussion highlights the importance of indirect general equilibrium effects for valid policy evaluation which can weaken the impact of transportation improvements on SMA inequality. For example, after a 30% decline in travel costs, if the distribution of services stores remains unchanged, the inequality declines by 25%, which is much larger than the total effects, 18%.

E. 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.
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.

This survey 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.

Example 1) Buying coffee at a Café on the way back from a restaurant near your workplace at lunchtime: 2 stores

Example 2) 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

Example 3) 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.

What is your gender?

a. Male
b. Female
c. Others, or refuse to answer

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?

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 your 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 your 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 your 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 your 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 Sat, 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 your 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 your 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 your 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 your 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.)

F. Omitted Proofs

Proof of Proposition 1. The free entry condition equates profits and operating costs:

$$\begin{aligned}\frac{(\rho-1)^{\sigma-1}}{\rho^{\sigma}}\cdot C_{jsd} &= \frac{1}{\rho}\cdot p_{jsd}(\omega)\cdot q_{jsd}(\omega)\\ &= \frac{1}{\rho}\cdot p_{jsd}(\omega)\cdot\left(\frac{p_{jsd}(\omega)}{p_{jsd}}\right)^{-\rho}\cdot\phi_{jsd}\cdot\left(\frac{p_{jsd}}{p_{js}}\right)^{-\sigma}\cdot\frac{R_{js}}{p_{js}}\\ &= \frac{1}{\rho}\cdot p_{jsd}(\omega)^{1-\sigma}\cdot N_{jsd}^{\frac{\rho-\sigma}{1-\rho}}\cdot\phi_{jsd}\cdot p_{js}^{-(1-\sigma)}\cdot R_{js}\\ &= \frac{1}{\rho}\cdot\left(\frac{\rho}{\rho-1}\right)^{1-\sigma}c_{jsd}^{1-\sigma}\cdot A_{jsd}^{-(1-\sigma)}\cdot N_{jsd}^{\frac{\rho-\sigma}{1-\rho}}\cdot\phi_{jsd}\cdot p_{js}^{-(1-\sigma)}\cdot(1-\beta)\alpha_s\boldsymbol{E}^{\intercal}(I-\beta\Pi)^{-1}\boldsymbol{\pi}_j^s\end{aligned}$$

where the third equality uses the fact that, imposing symmetry across individual stores, we have $p_{jsd} = p_{jsd}(\omega)\cdot N_{jsd}^{\frac{1}{1-\rho}}$. Rearranging proves the proposition. □

Proof of Lemma A.2. The first two conditions can be linearized as

$$(1-\frac{1-\sigma}{1-\rho})n_{jsd} = -\hat{C}_{jsd}-(1-\sigma)\tilde{a}_{jsd}+(1-\sigma)\hat{c}_{jsd}-(1-\sigma)\hat{p}_{js}+r_{js} \tag{A.1*}$$

$$\hat{p}_{js} = \sum_d \theta_{jsd}(\frac{1}{1-\rho}n_{jsd}-\tilde{a}_{jsd}+\hat{c}_{jsd}) \equiv \hat{c}_{js}-(\frac{1}{\rho-1}n_{js}+\tilde{a}_{js}) \equiv -(\frac{1}{\rho-1}n_{js}+\check{a}_{js}). \tag{A.2*}$$

Aggregating (A.1*), we have

$$n_{js} = \frac{\rho-1}{\rho-\sigma}\big(-\hat{C}_{js}-(1-\sigma)\tilde{a}_{js}+(1-\sigma)\hat{c}_{js}-(1-\sigma)\hat{p}_{js}+r_{js}\big) \tag{A.1**}$$

so that

$$\hat{p}_{js} = -\check{a}_{js}+\tfrac{1}{\rho-1}\hat{C}_{js}-\tfrac{1}{\rho-1}r_{js} \equiv -\check{a}^*_{js}-\tfrac{1}{\rho-1}r_{js}.$$

The third condition becomes

$$\begin{aligned}\hat{V}(i) &= \sum_s \alpha_s\cdot\sum_j \pi_{ij}^s\big(-\hat{p}_{js}+\beta\hat{V}(j)\big) = -\underbrace{\sum_s \alpha_s\cdot\sum_j \pi_{ij}^s\hat{p}_{js}}_{\equiv\hat{P}_i}+\beta\sum_j \pi_{ij}\hat{V}(j)\\ &= -\hat{P}_i-\beta\sum_j \pi_{ij}\hat{P}_j+o(\boldsymbol{x}).\end{aligned} \tag{A.3*}$$

Next, we have

$$r_{js} = \sum_i (1-\beta)\lambda_{ij}^s(e_i+\hat{\pi}_{ij}^s)+\sum_i\sum_k\sum_{s'}\beta\lambda_{ikj}^{s's}(e_i+\hat{\pi}_{ik}^{s'}+\hat{\pi}_{kj}^s) \tag{A.4*}$$

where $\lambda_{ij}^s = \dfrac{E_i\pi_{ij}^s}{\Lambda}$ and $\lambda_{ikj}^{s's} = \dfrac{E_i\alpha_{s'}\pi_{ik}^{s'}\pi_{kj}^s}{\Lambda}$ with $\Lambda = (1-\beta)\sum_i E_i\pi_{ij}^s+\beta\sum_{iks'}E_i\alpha_{s'}\pi_{ik}^{s'}\pi_{kj}^s$. Finally, we have

$$\begin{aligned}\hat{\pi}_{ij}^s &= \sum_{j'}(\mathbb{1}_{j'=j}-\pi_{ij'}^s)\big(-\tfrac{1}{\nu}\hat{p}_{j's}+\tilde{\beta}\hat{V}(j')\big) = \sum_{j'}(\mathbb{1}_{j'=j}-\pi_{ij'}^s)\big(-\tfrac{1}{\nu}\hat{p}_{j's}-\tilde{\beta}\hat{P}_{j'}\big)+o(\boldsymbol{x})\\ &= \sum_{j'}(\mathbb{1}_{j'=j}-\pi_{ij'}^s)\left(-\tfrac{1}{\nu}\hat{p}_{j's}-\tilde{\beta}\sum_{s'}\alpha_{s'}\cdot\sum_{j''}\pi_{j'j''}^{s'}\hat{p}_{j''s'}\right)+o(\boldsymbol{x})\end{aligned} \tag{A.5*}$$

□

Proof of Proposition A.1. Let us *suppose* for the moment that $\hat{p}_{js} = o(\boldsymbol{x})$ and $r_{js} = o(\boldsymbol{x})$ for $j \neq j_0$. Then for any j, we have

$$\begin{aligned}
\hat{\pi}^s_{ij} &= \sum_{j'} (\mathbb{1}_{j'=j} - \pi^s_{ij'}) \left(-\tfrac{1}{\nu} \hat{p}_{j's} - \tilde{\beta} \sum_{s'} \alpha_{s'} \cdot \pi^{s'}_{j'j_0} \hat{p}_{j_0 s'} \right) + o(\boldsymbol{x}) \\
&= (\mathbb{1}_{j_0=j} - \pi^s_{ij_0}) \underbrace{\left(-\tfrac{1}{\nu} \hat{p}_{j_0 s} - \tilde{\beta} \sum_{s'} \alpha_{s'} \cdot \pi^{s'}_{j_0 j_0} \hat{p}_{j_0 s'} \right)}_{\equiv \hat{q}_{j_0 s}} + o(\boldsymbol{x}).
\end{aligned}$$

Under Assumption A.2, this can be further simplified as

$$\hat{\pi}^s_{ij} = \begin{cases} 0 & \text{if } i, j \neq j_0 \\ -\pi^s_{j_0 j_0} \hat{q}_{j_0 s} & \text{if } i = j_0, j \neq j_0 \\ \hat{q}_{j_0 s} & \text{if } i \neq j_0, j = j_0 \\ (1 - \pi^s_{j_0 j_0}) \hat{q}_{j_0 s} & \text{if } i = j_0, j = j_0 \end{cases} + o(\boldsymbol{x})$$

Note also the following lemma:

Lemma A.9. *For $j \neq j_0$, Assumption A.2 implies*

$$(1 - \beta) \lambda^s_{j_0 j} + \beta \sum_{k, s'} \lambda^{s's}_{j_0 k j} = o(\boldsymbol{x}) \tag{A.21}$$

and

$$\beta \sum_{i \neq j_0, s'} \lambda^{s's}_{i j_0 j} = o(\boldsymbol{x}).$$

We thus have, for $j \neq j_0$,

$$\begin{aligned}
r_{js} &= \sum_i (1 - \beta) \lambda^s_{ij} \hat{\pi}^s_{ij} + \sum_i \sum_k \sum_{s'} \beta \lambda^{s's}_{ikj} (\hat{\pi}^{s'}_{ik} + \hat{\pi}^s_{kj}) \\
&= (1 - \beta) \lambda^s_{j_0 j} \hat{\pi}^s_{j_0 j} + \beta \sum_{k \neq j_0, s'} \lambda^{s's}_{j_0 k j} \hat{\pi}^{s'}_{j_0 k} + \beta \sum_{i, s'} \lambda^{s's}_{i j_0 j} (\hat{\pi}^{s'}_{i j_0} + \hat{\pi}^s_{j_0 j}) + o(\boldsymbol{x}) \quad (\because \text{ ignore } \hat{\pi}\text{'s without } j_0) \\
&\leq \left[(1 - \beta) \lambda^s_{j_0 j} + \beta \sum_{k, s'} \lambda^{s's}_{j_0 k j} + \beta \sum_{i \neq j_0, s'} \lambda^{s's}_{i j_0 j} \right] \cdot \max_{i,k,s,s'} \{ \hat{\pi}^{s'}_{i j_0} + \hat{\pi}^s_{j_0 k} \} \\
&= o(\boldsymbol{x})
\end{aligned}$$

where the last inequality follows from Lemma A.9. Thus, we indeed have $r_{js} = o(\boldsymbol{x})$ and, hence, $\hat{p}_{js} = o(\boldsymbol{x})$ for $j \neq j_0$. In this case, we can also simplify $r_{j_0 s}$ and get the following result: We have, up to $o(\boldsymbol{x})$,

$$\begin{aligned} r_{j_0 s} &= \sum_i (1-\beta)\lambda^s_{ij_0}\hat{\pi}^s_{ij_0} + \sum_{iks'} \beta\lambda^{s's}_{ikj_0}(\hat{\pi}^{s'}_{ik} + \hat{\pi}^s_{kj_0}) \\ &= \sum_i (1-\beta)\lambda^s_{ij_0}\hat{q}_{j_0 s} - (1-\beta)\lambda^s_{j_0 j_0}\pi^s_{j_0 j_0}\hat{q}_{j_0 s} + \sum_{iks'}\beta\lambda^{s's}_{ikj_0}(\hat{\pi}^{s'}_{ik} + \hat{q}_{j_0 s}) - \sum_{is'}\beta\lambda^{s's}_{ij_0 j_0}\pi^s_{j_0 j_0}\hat{q}_{j_0 s} \\ &= \hat{q}_{j_0 s}\left(1 - \left((1-\beta)\lambda^s_{j_0 j_0} + \sum_{is'}\beta\lambda^{s's}_{ij_0 j_0}\right)\pi^s_{j_0 j_0}\right) + \sum_{i=j_0 \text{ or } k=j_0, s'} \beta\lambda^{s's}_{ikj_0}\hat{\pi}^{s'}_{ik} \\ &= \hat{q}_{j_0 s}\left(1 - \left((1-\beta)\lambda^s_{j_0 j_0} + \sum_{is'}\beta\lambda^{s's}_{ij_0 j_0}\right)\pi^s_{j_0 j_0}\right) + \sum_{is'}\beta\lambda^{s's}_{ij_0 j_0}\hat{\pi}^{s'}_{ij_0} + \sum_{k\neq j_0, s'}\beta\lambda^{s's}_{j_0 k j_0}\hat{\pi}^{s'}_{j_0 k} \\ &= \hat{q}_{j_0 s}\left(1 - \left((1-\beta)\lambda^s_{j_0 j_0} + \sum_{is'}\beta\lambda^{s's}_{ij_0 j_0}\right)\pi^s_{j_0 j_0}\right) + \sum_{s'}\hat{q}_{j_0 s'}\left(\sum_i \beta\lambda^{s's}_{ij_0 j_0}(1 - \pi^s_{j_0 j_0}\cdot \mathbb{1}_{i=j_0})\right) + \sum_{k\neq j_0, s'}\beta\lambda^{s's}_{j_0 k j_0}\hat{\pi}^{s'}_{j_0 k}. \end{aligned}$$

Note that

$$\begin{aligned} \sum_{k\neq j_0, s'} \lambda^{s's}_{j_0 k j_0} &= \frac{\sum_{k\neq j_0, s'} E_{j_0}\alpha_{s'}\pi^{s'}_{j_0 k}\pi^s_{kj_0}}{(1-\beta)\sum_i E_i \pi^s_{ij_0} + \cdots} \\ &\leq \frac{\max_{s'}\left\{E_{j_0}\sum_{k\neq j_0}\pi^{s'}_{j_0 k}\pi^s_{kj_0}\right\}}{(1-\beta)E_{j_0}\pi^s_{j_0 j_0}} = \frac{\max_{s'}\left\{\sum_{k\neq j_0}\pi^{s'}_{j_0 k}\pi^s_{kj_0}\right\}}{(1-\beta)\pi^s_{j_0 j_0}} \leq \frac{\max_{s'}\left\{\sum_{k\neq j_0}\pi^s_{kj_0}\right\}}{(1-\beta)\pi^s_{j_0 j_0}} \\ &= o(1). \end{aligned}$$

Finally, $\Psi_{j_0 s}, \Phi^{s'}_{j_0 s} \in (0,1)$ is immediate. $\square$

Proof of Lemma A.9. We have

$$\begin{aligned} R_{j_0 \to js} &= (1-\beta)\alpha_s E_{j_0}\pi^s_{j_0 j} + (1-\beta)\beta\alpha_s\sum_k E_{j_0}\pi_{j_0 k}\pi^s_{kj} + (1-\beta)\beta^2\alpha_s\sum_{k,l}E_{j_0}\pi_{j_0 k}\pi_{kl}\pi^s_{lj} + \cdots \\ R_{js} &= (1-\beta)\alpha_s\sum_i E_i\pi^s_{ij} + (1-\beta)\beta\alpha_s\sum_{iks'}E_i\alpha_{s'}\pi^{s'}_{ik}\pi^s_{kj} + o(\boldsymbol{x}) \\ &< 2\left((1-\beta)\alpha_s\sum_i E_i\pi^s_{ij} + \beta\alpha_s\sum_{iks'}E_i\alpha_{s'}\pi^{s'}_{ik}\pi^s_{kj}\right). \end{aligned}$$

Thus,

$$(1-\beta)\lambda^s_{j_0 j} + \beta\sum_{k,s'}\lambda^{s's}_{j_0 kj} = \frac{(1-\beta)E_{j_0}\pi^s_{j_0 j} + \beta\sum_{ks'}E_{j_0}\alpha_{s'}\pi^{s'}_{j_0 k}\pi^s_{kj}}{(1-\beta)\sum_i E_i\pi^s_{ij} + \beta\sum_{iks'}E_i\alpha_{s'}\pi^{s'}_{ik}\pi^s_{kj}} < \frac{R_{j_0\to js}/(1-\beta)}{R_{js}/2} = o(\boldsymbol{x})$$

which proves the first claim. Second,

$$\begin{aligned} \beta\sum_{i\neq j_0, s'}\lambda^{s's}_{ij_0 j} &= \beta\frac{\sum_{i\neq j_0, s'}E_i\alpha_{s'}\pi^{s'}_{ij_0}\pi^s_{j_0 j}}{(1-\beta)\sum_i E_i\pi^s_{ij} + \beta\sum_{iks'}E_i\alpha_{s'}\pi^{s'}_{ik}\pi^s_{kj}} \\ &= \beta\,\underbrace{\max_{s'}\left\{\sum_{i\neq j_0}E_i\pi^{s'}_{ij_0}\right\}}_{o(1)}\frac{\pi^s_{j_0 j}}{(1-\beta)\sum_i E_i\pi^s_{ij} + \beta\sum_{iks'}E_i\alpha_{s'}\pi^{s'}_{ik}\pi^s_{kj}} + o(\boldsymbol{x}) \\ &= o(\boldsymbol{x}). \end{aligned}$$

$\square$

Proof of Corollary A.1. First, note that, up to $o(1)$, we have

$$r_{j_0 s} = \Psi_{j_0 s}\hat{q}_{j_0 s} = \Psi_{j_0 s}(-\tfrac{1}{\nu}\hat{p}_{j_0 s}) = \Psi_{j_0 s}\Big(\tfrac{1}{\nu}\breve{a}^*_{j_0 s} + \tfrac{1}{\nu}\tfrac{1}{\rho-1}r_{j_0 s}\Big)$$

hence

$$r_{j_0 s} = \frac{\Psi_{j_0 s}\frac{1}{\nu}\breve{a}^*_{j_0 s}}{1 - \Psi_{j_0 s}\frac{1}{\nu}\frac{1}{\rho-1}} > 0$$

where we use $\frac{1}{\rho-1}\frac{1}{\nu} < 1$ from Assumption A.1 and $\Psi_{j_0 s} < 1$. This highlights that we need to assume $\frac{1}{\rho-1}\frac{1}{\nu} < 1$ in order to guarantee stable equilibrium. This in turn implies the rest of inequalities in (A.6), except for that of n_{js}, which follows from

$$\begin{aligned}
n_{j_0 s} &\overset{\text{sgn}}{=} -\hat{C}_{j_0 s} + (\sigma-1)\breve{a}_{j_0 s} + (\sigma-1)\hat{p}_{j_0 s} + r_{j_0 s} \\
&= -\hat{C}_{j_0 s} + (\sigma-1)\breve{a}_{j_0 s} - (\sigma-1)\breve{a}^*_{j_0 s} + \tfrac{\rho-\sigma}{\rho-1}r_{j_0 s} \\
&= -\tfrac{\rho-\sigma}{\rho-1}\hat{C}_{j_0 s} + \tfrac{\rho-\sigma}{\rho-1}r_{j_0 s} \\
&> 0.
\end{aligned}$$

□

Proof of Proposition A.2. First, we rewrite the change in revenue as

$$\begin{aligned}
r_{j_0 s} &= \Psi_{j_0 s}\hat{q}_{j_0 s} + \sum_{s'}\Phi^{s'}_{j_0 s}\hat{q}_{j_0 s'} + o(\boldsymbol{x}) \\
&= \Psi_{j_0 s}(-\tfrac{1}{\nu}\hat{p}_{j_0 s} - \tilde{\beta}\sum_{s'}\alpha_{s'}\pi^{s'}_{j_0 j_0}\hat{p}_{j_0 s'}) + \sum_{s'}\Phi^{s'}_{j_0 s}(-\tfrac{1}{\nu}\hat{p}_{j_0 s'}) + o(\boldsymbol{x}) \\
&= -\Theta_{j_0 s}\hat{p}_{j_0 s} - \sum_{s'\neq s}\Lambda^{s'}_{j_0 s}\hat{p}_{j_0 s'} + o(\boldsymbol{x}) \\
&= \Theta_{j_0 s}\Big(\breve{a}_{j_0 s} + \tfrac{1}{\rho-1}n_{j_0 s}\Big) + \sum_{s'\neq s}\Lambda^{s'}_{j_0 s}\Big(\breve{a}_{j_0 s'} + \tfrac{1}{\rho-1}n_{j_0 s'}\Big) + o(\boldsymbol{x})
\end{aligned}$$

where

$$\begin{aligned}
\Theta_{j_0 s} &= \Psi_{j_0 s}\cdot\tfrac{1}{\nu} + \Phi^{s}_{j_0 s}\cdot\tfrac{1}{\nu} + \tilde{\beta}\Psi_{j_0 s}\alpha_s\pi^{s}_{j_0 j_0} > 0, \\
\Lambda^{s'}_{j_0 s} &= \Phi^{s'}_{j_0 s}\cdot\tfrac{1}{\nu} + \tilde{\beta}\Psi_{j_0 s}\alpha_{s'}\pi^{s'}_{j_0 j_0} > 0.
\end{aligned}$$

Using this result, we first prove the following intermediate result.

Lemma A.10. *Under Assumption A.2 and up to $o(\boldsymbol{x})$, we have*

$$\begin{aligned}
n_{j_0 s} &= -\hat{C}_{j_0 s} + \Theta_{j_0 s}\Big(\breve{a}_{j_0 s} + \tfrac{1}{\rho-1}n_{j_0 s}\Big) + \sum_{s'\neq s}\Lambda^{s'}_{j_0 s}\Big(\breve{a}_{j_0 s'} + \tfrac{1}{\rho-1}n_{j_0 s'}\Big) \\
&\equiv \tilde{\gamma}^1_{j_0 s}\breve{a}_{j_0 s} + \sum_{s'\neq s}\tilde{\gamma}^{2s'}_{j_0 s}\Big(\breve{a}_{j_0 s'} + \tfrac{1}{\rho-1}n_{j_0 s'}\Big) - \tilde{\gamma}^3_{j_0 s}\hat{C}_{j_0 s} \\
n_{j_0 sd} &= n_{j_0 s} + \kappa(\breve{a}^*_{j_0 sd} - \breve{a}^*_{j_0 s}) - (\hat{C}_{j_0 sd} - \hat{C}_{js})
\end{aligned}$$

where

$$\kappa = \tfrac{(\sigma-1)(\rho-1)}{\rho-\sigma}, \quad \tilde{\gamma}^1_{j_0 s} = \Theta_{j_0 s}\cdot\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}}, \quad \tilde{\gamma}^{2s'}_{j_0 s} = \Lambda^{s'}_{j_0 s}\cdot\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}}, \quad \tilde{\gamma}^3_{j_0 s} = \frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}}.$$

In particular, $\tilde{\gamma}$'s and κ are positive.

Thus, we can write

$$
\begin{aligned}
\begin{pmatrix} n_{j_0 1} \\ n_{j_0 2} \\ n_{j_0 3} \end{pmatrix} &= \frac{1}{\rho-1}\begin{pmatrix} 0 & \tilde{\gamma}^{22}_{j_0 1} & \tilde{\gamma}^{23}_{j_0 1} \\ \tilde{\gamma}^{21}_{j_0 2} & 0 & \tilde{\gamma}^{23}_{j_0 2} \\ \tilde{\gamma}^{21}_{j_0 3} & \tilde{\gamma}^{22}_{j_0 3} & 0 \end{pmatrix}\begin{pmatrix} n_{j_0 1} \\ n_{j_0 2} \\ n_{j_0 3} \end{pmatrix} + \begin{pmatrix} \tilde{\gamma}^{1}_{j_0 1} & \tilde{\gamma}^{22}_{j_0 1} & \tilde{\gamma}^{23}_{j_0 1} \\ \tilde{\gamma}^{21}_{j_0 2} & \tilde{\gamma}^{1}_{j_0 2} & \tilde{\gamma}^{23}_{j_0 2} \\ \tilde{\gamma}^{21}_{j_0 3} & \tilde{\gamma}^{22}_{j_0 3} & \tilde{\gamma}^{1}_{j_0 3} \end{pmatrix}\begin{pmatrix} \breve{a}_{j_0 1} \\ \breve{a}_{j_0 2} \\ \breve{a}_{j_0 3} \end{pmatrix} - \begin{pmatrix} \tilde{\gamma}^{3}_{j_0 1} & 0 & 0 \\ 0 & \tilde{\gamma}^{3}_{j_0 2} & 0 \\ 0 & 0 & \tilde{\gamma}^{3}_{j_0 3} \end{pmatrix}\begin{pmatrix} \hat{C}_{j_0 1} \\ \hat{C}_{j_0 2} \\ \hat{C}_{j_0 3} \end{pmatrix} \\
\boldsymbol{n}_{j_0} &= \frac{1}{\rho-1}\Gamma_2 \boldsymbol{n}_{j_0} + (\Gamma_1+\Gamma_2)\breve{\boldsymbol{a}}_{j_0} - \Gamma_3 \hat{\boldsymbol{C}}_{j_0} \\
\boldsymbol{n}_{j_0} &= \left(I - \frac{1}{\rho-1}\Gamma_2\right)^{-1}\Big((\Gamma_1+\Gamma_2)\breve{\boldsymbol{a}}_{j_0} - \Gamma_3 \hat{\boldsymbol{C}}_{j_0}\Big) \quad \text{where } \Gamma_2 = o(1) \\
&= \left(I + \frac{1}{\rho-1}\Gamma_2\right)\Big((\Gamma_1+\Gamma_2)\breve{\boldsymbol{a}}_{j_0} - \Gamma_3 \hat{\boldsymbol{C}}_{j_0}\Big) + o(\boldsymbol{x}) \\
&= \left(\Gamma_1 + \Gamma_2 + \frac{1}{\rho-1}\Gamma_2\Gamma_1\right)\breve{\boldsymbol{a}}_{j_0} - \left(\Gamma_3 + \frac{1}{\rho-1}\Gamma_2\Gamma_3\right)\hat{\boldsymbol{C}}_{j_0} + o(\boldsymbol{x}) \\
&= \begin{pmatrix} \tilde{\gamma}^{1}_{j_0 1} & (1+\frac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 2})\tilde{\gamma}^{22}_{j_0 1} & (1+\frac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 3})\tilde{\gamma}^{23}_{j_0 1} \\ (1+\frac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 1})\tilde{\gamma}^{21}_{j_0 2} & \tilde{\gamma}^{1}_{j_0 2} & (1+\frac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 3})\tilde{\gamma}^{23}_{j_0 2} \\ (1+\frac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 1})\tilde{\gamma}^{21}_{j_0 3} & (1+\frac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 2})\tilde{\gamma}^{22}_{j_0 3} & \tilde{\gamma}^{1}_{j_0 3} \end{pmatrix}\breve{\boldsymbol{a}}_{j_0} - \begin{pmatrix} \tilde{\gamma}^{3}_{j_0 1} & \frac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 2}\tilde{\gamma}^{22}_{j_0 1} & \frac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 3}\tilde{\gamma}^{23}_{j_0 1} \\ \frac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 1}\tilde{\gamma}^{21}_{j_0 2} & \tilde{\gamma}^{3}_{j_0 2} & \frac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 3}\tilde{\gamma}^{23}_{j_0 2} \\ \frac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 1}\tilde{\gamma}^{21}_{j_0 3} & \frac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 2}\tilde{\gamma}^{22}_{j_0 3} & \tilde{\gamma}^{3}_{j_0 3} \end{pmatrix}\hat{\boldsymbol{C}}_{j_0} \\
\therefore\ n_{j_0 s} &= \tilde{\gamma}^{1}_{j_0 s}\breve{a}_{j_0 s} - \tilde{\gamma}^{3}_{j_0 s}\hat{C}_{j_0 s} + \sum_{s' \neq s}\Big(\tilde{\tau}^{2s'}_{j_0 s}\breve{a}_{j_0 s'} - \tilde{\kappa}^{2s'}_{j_0 s}\hat{C}_{j_0 s'}\Big) + o(\boldsymbol{x})
\end{aligned}
$$

where

$$
\tilde{\tau}^{2s'}_{j_0 s} = \left(1 + \tfrac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 s'}\right)\tilde{\gamma}^{2s'}_{j_0 s} \quad \text{and} \quad \tilde{\kappa}^{2s'}_{j_0 s} = \tfrac{1}{\rho-1}\tilde{\gamma}^{3}_{j_0 s'}\tilde{\gamma}^{2s'}_{j_0 s}.
$$

This can be rewritten as

$$
n_{j_0 s} = \gamma^{1}_{j_0 s}\breve{a}^{*}_{j_0 s} - \gamma^{3}_{j_0 s}\hat{C}_{j_0 s} + \sum_{s' \neq s}\Big(\tau^{2s'}_{j_0 s}\breve{a}^{*}_{j_0 s'} - \kappa^{2s'}_{j_0 s}\hat{C}_{j_0 s'}\Big) + o(\boldsymbol{x})
$$

where $\gamma^{1}_{j_0 s} = \tilde{\gamma}^{1}_{j_0 s}$, $\gamma^{3}_{j_0 s} = \tilde{\gamma}^{3}_{j_0 s} - \frac{\tilde{\gamma}^{1}_{j_0 s}}{\rho-1}$, $\tau^{2s'}_{j_0 s} = \tilde{\tau}^{2s'}_{j_0 s}$, and $\kappa^{2s'}_{j_0 s} = \tilde{\kappa}^{2s'}_{j_0 s} - \frac{\tilde{\tau}^{2s'}_{j_0 s}}{\rho-1}$. It turns out that, these four coefficients can be simplified as

$$
\begin{aligned}
\gamma^{1}_{j_0 s} &= \tilde{\gamma}^{1}_{j_0 s} = \Theta_{j_0 s}\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}} \\
\gamma^{3}_{j_0 s} &= \tilde{\gamma}^{3}_{j_0 s} - \frac{\tilde{\gamma}^{1}_{j_0 s}}{\rho-1} = \left(1 - \frac{\Theta_{j_0 s}}{\rho-1}\right)\cdot\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}} = 1 \\
\tau^{2s'}_{j_0 s} &= \tilde{\tau}^{2s'}_{j_0 s} = \left(1 + \tfrac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 s'}\right)\tilde{\gamma}^{2s'}_{j_0 s} = \Lambda^{s'}_{j_0 s}\frac{1}{1-\frac{\Theta_{j_0 s}}{\rho-1}}\frac{1}{1-\frac{\Theta_{j_0 s'}}{\rho-1}} \\
\kappa^{2s'}_{j_0 s} &= \tilde{\kappa}^{2s'}_{j_0 s} - \frac{\tilde{\tau}^{2s'}_{j_0 s}}{\rho-1} = \tfrac{1}{\rho-1}\left(\tilde{\gamma}^{3}_{j_0 s'}\tilde{\gamma}^{2s'}_{j_0 s} - \left(1 + \tfrac{1}{\rho-1}\tilde{\gamma}^{1}_{j_0 s'}\right)\tilde{\gamma}^{2s'}_{j_0 s}\right) \\
&= \tfrac{1}{\rho-1}\gamma^{2s'}_{j_0 s}\left(\frac{1}{1-\frac{\Theta_{j_0 s'}}{\rho-1}} - 1 - \frac{\Theta_{j_0 s'}}{\rho-1}\frac{1}{1-\frac{\Theta_{j_0 s'}}{\rho-1}}\right) \\
&= 0.
\end{aligned}
$$

Finally, it is clear from its definition that $\tau^{2s'}_{j_0 s}$ vanishes as $\beta \downarrow 0$. □

Proof of Lemma A.10. We have

$$
\begin{aligned}
\underbrace{\tfrac{\rho-\sigma}{\rho-1}}_{>0} n_{j_0 s d} &= -\hat{C}_{j_0 s d} + (\sigma-1)\breve{a}_{j_0 s d} + (\sigma-1)\hat{p}_{j_0 s} + r_{j_0 s} \\
&= -\tfrac{\rho-\sigma}{\rho-1}\hat{C}_{j_0 s d} + (\sigma-1)\breve{a}^{*}_{j_0 s d} - (\sigma-1)\breve{a}^{*}_{j_0 s} + \tfrac{\rho-\sigma}{\rho-1}r_{j_0 s}
\end{aligned}
$$

$$\implies n_{j_0sd} = -\hat{C}_{j_0sd} + \kappa(\check{a}^*_{j_0sd} - \check{a}^*_{j_0s}) + r_{j_0s}$$

$$\implies n_{j_0s} = -\hat{C}_{j_0s} + r_{j_0s}$$

$$= -\hat{C}_{j_0s} + \Theta_{j_0s}\Big(\check{a}_{j_0s} + \tfrac{1}{\rho-1}n_{j_0s}\Big) + \sum_{s'\neq s}\Lambda^{s'}_{j_0s}\Big(\check{a}_{j_0s'} + \tfrac{1}{\rho-1}n_{j_0s'}\Big)$$

$$\implies (1 - \tfrac{\Theta_{j_0s}}{\rho-1})n_{j_0s} = -\hat{C}_{j_0s} + \Theta_{j_0s}\check{a}_{j_0s} + \sum_{s'\neq s}\Lambda^{s'}_{j_0s}\Big(\check{a}_{j_0s'} + \tfrac{1}{\rho-1}n_{j_0s'}\Big)$$

$$\implies n_{j_0s} = \tilde{\gamma}^1_{j_0s}\check{a}_{j_0s} + \sum_{s'\neq s}\tilde{\gamma}^{2s'}_{j_0s}\Big(\check{a}_{j_0s'} + \tfrac{1}{\rho-1}n_{j_0s'}\Big) - \tilde{\gamma}^3_{j_0s}\hat{C}_{j_0s}.$$

Note that $1 - \frac{\Theta_{j_0s}}{\rho-1} = 1 - \frac{1}{\nu}\frac{1}{\rho-1}(1 - \lambda^s_{j_0j_0}\pi^s_{j_0j_0}) + o(1) \in (0,1)$ under Assumption A.1. □

Proof of Proposition A.3. Define $\gamma_{2j} \equiv \mathrm{d}\log\Upsilon_{2j}$, which can be simplified to

$$\gamma_{2j} = \sum_s \varphi_{js}\sum_d \theta_{jsd}(\hat{C}_{jsd} + n_{jsd})$$

$$= \sum_s \varphi_{js}(\hat{C}_{js} + n_{js})$$

$$= \sum_s \varphi_{js}(\bar{\hat{C}}_{js} + n_{js}) - \varepsilon_c\cdot\gamma_{2j}$$

$$= \tfrac{1}{1+\varepsilon_c}\sum_s \varphi_{js}(\bar{\hat{C}}_{js} + n_{js})$$

where $\varphi_{js} = \frac{R_{js}}{\sum_{s'}R_{js'}}$. Then, we have

$$\check{a}^*_{j_0sd} = \bar{\check{a}}^*_{j_0sd} + \varepsilon\cdot\gamma_{2j}$$

where $\varepsilon \equiv \varepsilon_a + \frac{\varepsilon_c}{\rho-1}$. Thus, equation (A.7) becomes

$$n_{j_0s} = \gamma^1_{j_0s}(\bar{\check{a}}^*_{j_0s} + \varepsilon\gamma_{2j_0}) - \bar{\hat{C}}_{j_0s} + \varepsilon_c\gamma_{2j_0} + \sum_{s'\neq s}\tau^{2s'}_{j_0s}(\bar{\check{a}}^*_{j_0s'} + \varepsilon\gamma_{2j_0})$$

$$= \gamma^1_{j_0s}\bar{\check{a}}^*_{j_0s} - \bar{\hat{C}}_{j_0s} + \sum_{s'\neq s}\tau^{2s'}_{j_0s}\bar{\check{a}}^*_{j_0s'} + \underbrace{\frac{\varepsilon\gamma^1_{j_0s} + \varepsilon_c + \varepsilon\sum_{s'\neq s}\tau^{2s'}_{j_0s}}{1+\varepsilon_c}}_{\equiv\varepsilon_{j_0s}}\sum_{s'}\varphi_{j_0s'}(\bar{\hat{C}}_{j_0s'} + n_{j_0s'})$$

$$= \gamma^1_{j_0s}\bar{\check{a}}^*_{j_0s} - (1 - \varepsilon_{j_0s}\varphi_{j_0s})\bar{\hat{C}}_{j_0s} + \sum_{s'\neq s}\tau^{2s'}_{j_0s}\bar{\check{a}}^*_{j_0s'} + \varepsilon_{j_0s}\sum_{s'\neq s}\varphi_{j_0s'}\bar{\hat{C}}_{j_0s'} + \varepsilon_{j_0s}\sum_{s'}\varphi_{j_0s'}n_{j_0s'}$$

$$\equiv x_{j_0s} + \varepsilon_{j_0s}\sum_{s'}\varphi_{j_0s'}n_{j_0s'}. \tag{A.22}$$

Aggregating this equation across sectors, we have

$$\sum_s \varphi_{j_0s}n_{j_0s} = \sum_s \varphi_{j_0s}x_{j_0s} + \underbrace{(\sum_s \varphi_{j_0s}\varepsilon_{j_0s})}_{\equiv\varepsilon_{j_0}}(\sum_s \varphi_{j_0s}n_{j_0s})$$

$$= \tfrac{1}{1-\varepsilon_{j_0}}\sum_s \varphi_{j_0s}x_{j_0s}. \tag{A.23}$$

Combining equations (A.22) and (A.23), we have

$$
\begin{aligned}
n_{j_0 s} &= x_{j_0 s} + \underbrace{\tfrac{\varepsilon_{j_0 s}}{1-\varepsilon_{j_0}}}_{\tilde{\varepsilon}_{j_0 s}} \sum_{s'} \varphi_{j_0 s'} x_{j_0 s'} \\
&= (1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s}) x_{j_0 s} + \tilde{\varepsilon}_{j_0 s} \sum_{s' \neq s} \varphi_{j_0 s'} x_{j_0 s'} \\
&= \left((1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s})\gamma^1_{j_0 s} + \tilde{\varepsilon}_{j_0 s}\left(\varphi_{j_0 s'}\tau^{2s}_{j_0 s'} + \varphi_{j_0 s''}\tau^{2s}_{j_0 s''}\right)\right)\bar{\tilde{a}}^*_{j_0 s} \\
&+ \left((1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s})\tau^{2s'}_{j_0 s} + \tilde{\varepsilon}_{j_0 s}\left(\varphi_{j_0 s'}\gamma^1_{j_0 s'} + \varphi_{j_0 s''}\tau^{2s'}_{j_0 s''}\right)\right)\bar{\tilde{a}}^*_{j_0 s'} \\
&+ \left((1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s})\tau^{2s''}_{j_0 s} + \tilde{\varepsilon}_{j_0 s}\left(\varphi_{j_0 s'}\tau^{2s''}_{j_0 s'} + \varphi_{j_0 s''}\gamma^1_{j_0 s''}\right)\right)\bar{\tilde{a}}^*_{j_0 s''} \\
&+ \left(-(1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s})(1-\varepsilon_{j_0 s}\varphi_{j_0 s}) + \tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s}(\varphi_{j_0 s'}\varepsilon_{j_0 s'} + \varphi_{j_0 s''}\varepsilon_{j_0 s''})\right)\bar{\hat{C}}_{j_0 s} \\
&+ \left((1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s})\varepsilon_{j_0 s}\varphi_{j_0 s'} + \tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s'}(-(1-\varepsilon_{j_0 s'}\varphi_{j_0 s'}) + \varphi_{j_0 s''}\varepsilon_{j_0 s''})\right)\bar{\hat{C}}_{j_0 s'} \\
&+ \left((1+\tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s})\varepsilon_{j_0 s}\varphi_{j_0 s''} + \tilde{\varepsilon}_{j_0 s}\varphi_{j_0 s''}(\varphi_{j_0 s'}\varepsilon_{j_0 s'} - (1-\varepsilon_{j_0 s''}\varphi_{j_0 s''}))\right)\bar{\hat{C}}_{j_0 s''} \\
&\equiv \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'}.
\end{aligned}
$$

It is immediate from this equation that $\tilde{\gamma}^1_{j_0 s}$ and $\tilde{\tau}^{2s'}_{j_0 s}$ are positive, and $\lim_{(\beta,\varepsilon_a,\varepsilon_c)\to(0,0,0)} \tilde{\tau}^{2s'}_{j_0 s} = 0$. □

Proof of Lemma A.3. Thanks to the fixed effect FE_s, we can ignore components of n^{Bartik}_{js} that are invariant across j. This means that we can ignore components of n_{sd} invariant across d; i.e., can ignore components of n_{jsd} that are invariant across d. Thus, we can write

$$
n_{jsd} = \kappa\tilde{a}_{jsd} + \mathsf{FE}_{js}, \;\; n_{sd} = \kappa\tilde{a}_{sd} + \mathsf{FE}_s, \;\; n^{\text{Bartik}}_{js} = \kappa\sum_d \theta_{jsd,0}\tilde{a}_{sd} + \mathsf{FE}_s, \;\; \text{and} \;\; \tilde{n}^{\text{Bartik}}_{js} = \kappa\sum_d \theta_{jsd}\tilde{a}_{sd} + \mathsf{FE}_s\,.
$$

Note that we have

$$
\begin{aligned}
n_{js} &= \gamma\tilde{a}_{js} - \gamma\hat{c}_{js} - \left(1+\tfrac{\gamma}{\rho-1}\right)\hat{C}_{js} + \sum_{s'\neq s}\tau\left(\tilde{a}_{js'} - \tfrac{1}{\rho-1}\hat{C}_{js'} - \hat{c}_{js'}\right) \\
&= \gamma\sum_d \theta_{jsd}\tilde{a}_{jsd} - \gamma\hat{c}_{js} - \left(1+\tfrac{\gamma}{\rho-1}\right)\hat{\varepsilon}_{js} + \sum_{s'\neq s}\left(\tau\sum_d \theta_{js'd}\tilde{a}_{js'd} - \tfrac{\tau}{\rho-1}\hat{\varepsilon}_{js'} - \tau\hat{c}_{js'}\right) + \mathsf{FE}_s \\
&= \gamma\sum_d \theta_{jsd}\tilde{a}_{sd} + \gamma\tilde{\varepsilon}_{js} - \gamma\hat{c}_{js} - \left(1+\tfrac{\gamma}{\rho-1}\right)\hat{\varepsilon}_{js} + \sum_{s'\neq s}\left(\tau\sum_d \theta_{js'd}\tilde{a}_{s'd} + \tau\tilde{\varepsilon}_{js'} - \tfrac{\tau}{\rho-1}\hat{\varepsilon}_{js'} - \tau\hat{c}_{js'}\right) + \mathsf{FE}_s \\
&= \gamma\sum_d \theta_{jsd}\tilde{a}_{sd} + \sum_{s'\neq s}\left(\tau\sum_d \theta_{js'd}\tilde{a}_{s'd}\right) - \gamma\hat{c}_{js} - \sum_{s'\neq s}\tau\hat{c}_{js'} + \mathsf{FE}_s + \varepsilon_{js} \\
&= \tfrac{\gamma}{\kappa}\tilde{n}^{\text{Bartik}}_{js} + \sum_{s'\neq s}\tfrac{\tau}{\kappa}\tilde{n}^{\text{Bartik}}_{js'} - \gamma\hat{c}_{js} - \sum_{s'\neq s}\tau\hat{c}_{js'} + \mathsf{FE}_s + \varepsilon_{js}
\end{aligned}
$$

where $\varepsilon_{js} = \gamma\tilde{\varepsilon}_{js} - \left(1+\frac{\gamma}{\rho-1}\right)\hat{\varepsilon}_{js} + \sum_{s'\neq s}\left(\tau\tilde{\varepsilon}_{js'} - \frac{\tau}{\rho-1}\hat{\varepsilon}_{js'}\right)$. We also have

$$
\begin{aligned}
n_{jsd} &= n_{js} + \kappa(\tilde{a}_{jsd} - \tilde{a}_{js}) \\
&= n_{js} + \kappa(\tilde{a}_{sd} + \tilde{\varepsilon}_{jsd}) - \kappa\sum_d \theta_{jsd}(\tilde{a}_{sd} + \tilde{\varepsilon}_{jsd})
\end{aligned}
$$

$$= \tfrac{\gamma}{\kappa}\tilde{n}_{js}^{\text{Bartik}} + \sum_{s'\neq s}\tfrac{\tau}{\kappa}\tilde{n}_{js'}^{\text{Bartik}} - \gamma\hat{c}_{js} - \sum_{s'\neq s}\tau\hat{c}_{js'} + \mathsf{FE}_s + \varepsilon_{js} + \kappa(\tilde{a}_{sd} + \tilde{\varepsilon}_{jsd}) - \kappa\sum_d \theta_{jsd}\tilde{a}_{sd} - \kappa\tilde{\varepsilon}_{js}$$

$$= \tfrac{\gamma-\kappa}{\kappa}\tilde{n}_{js}^{\text{Bartik}} + \sum_{s'\neq s}\tfrac{\tau}{\kappa}\tilde{n}_{js'}^{\text{Bartik}} - \gamma\hat{c}_{js} - \sum_{s'\neq s}\tau\hat{c}_{js'} + \mathsf{FE}_{sd} + \check{\varepsilon}_{jsd}$$

where $\check{\varepsilon}_{jsd} = \varepsilon_{js} + \kappa\tilde{\varepsilon}_{jsd} - \kappa\tilde{\varepsilon}_{js}$.

To show the orthogonality, note that, for any s, d, s'', d'', we have

$$\theta_{js''d'',0} \perp_j \begin{pmatrix}\tilde{\varepsilon}_{js} & \hat{\varepsilon}_{js} & \tilde{\varepsilon}_{jsd}\end{pmatrix}^{\mathsf{T}} | \vec{c}_j.$$

Because $n_{js''}^{\text{Bartik}} = \kappa\sum_{d''}\theta_{js''d'',0}\tilde{a}_{s''d''} + \mathsf{FE}_{s''}$ and $\tilde{a}_{s''d''}$ is fixed, we in turn have

$$n_{js''}^{\text{Bartik}} \perp_j \begin{pmatrix}\tilde{\varepsilon}_{js} & \hat{\varepsilon}_{js} & \tilde{\varepsilon}_{jsd}\end{pmatrix}^{\mathsf{T}} | \vec{c}_j$$

and

$$n_{js''}^{\text{Bartik}} \perp_j \begin{pmatrix}\varepsilon_{js} & \check{\varepsilon}_{jsd}\end{pmatrix}^{\mathsf{T}} | \vec{c}_j. \qquad \square$$

Proof of Proposition A.4. From (A.12), the exclusion restrictions directly come from the orthogonalities in Corollary A.3. The relevance condition holds because

$$\begin{aligned}
&\text{rank}\Big(\mathbb{E}_j\Big[\begin{pmatrix}\tilde{n}_{js}^{\text{Bartik}\perp} & n_{js'}^{\perp} & \tilde{n}_{js''}^{\text{Bartik}\perp}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}n_{js}^{\text{Bartik}\perp} & n_{js'}^{\text{Bartik}\perp} & n_{js''}^{\text{Bartik}\perp}\end{pmatrix}\Big]\Big)\\
&= \text{rank}\Big(\mathbb{E}_j\Big[\begin{pmatrix}\tilde{n}_{js}^{\text{Bartik}\perp} & \tilde{n}_{js'}^{\text{Bartik}\perp} & \tilde{n}_{js''}^{\text{Bartik}\perp}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}n_{js}^{\text{Bartik}\perp} & n_{js'}^{\text{Bartik}\perp} & n_{js''}^{\text{Bartik}\perp}\end{pmatrix}\Big]\Big)\\
&= \text{rank}\Big(\mathbb{E}_j\Big[\begin{pmatrix}\sum_d \theta_{jsd}^{\perp}\check{a}_{sd}^{*} & \sum_d \theta_{js'd}^{\perp}\check{a}_{s'd}^{*} & \sum_d \theta_{js''d}^{\perp}\check{a}_{s''d}^{*}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\sum_{d'} \theta_{jsd',0}^{\perp}\check{a}_{sd'}^{*} & \sum_{d'} \theta_{js'd',0}^{\perp}\check{a}_{s'd'}^{*} & \sum_{d'} \theta_{js''d',0}^{\perp}\check{a}_{s''d'}^{*}\end{pmatrix}\Big]\Big)\\
&= \text{rank}\left(\left[\sum_d\sum_{d'}\check{a}_{sd}^{*}\check{a}_{s'd'}^{*}\mathbb{E}_j\left[\theta_{jsd}^{\perp}, \theta_{js'd',0}^{\perp}\right]\right]_{s,s'}\right)\\
&= 3
\end{aligned}$$

where the first equality comes from the fact that the rank of a matrix is invariant to elementary row operations. Thus, the IV estimator is consistent, i.e., the IV coefficient on $n_{js'}$ converges in probability to

$$\frac{\tau}{\gamma} = \frac{\Lambda_s^{s'}}{\Theta_s'}\cdot\frac{1}{1-\frac{\Theta_s}{\rho-1}}.$$

We can see that the last formula is strictly positive and vanishes as $\beta \to 0$. $\square$

Proof of Proposition A.5. The proof is the same as that for Proposition A.4. $\square$

Proof of Lemmas A.4 and A.5. The first-order condition (A.16) characterizes the demand $q_j^i(\omega) = q_j^i(p_j(\omega))$. Differentiating it with respect to $p_j(\omega)$, we have

$$\frac{\partial U^i}{\partial q_j^i} f_j''(q_j^i(\omega))\frac{\mathrm{d}q_j^i(p_j(\omega))}{\mathrm{d}p_j(\omega)} = \lambda^i. \tag{A.24}$$

Firms solve

$$\max_{p_j(\omega)}\left\{(p_j(\omega) - c_j)\sum_i q_j^i(p_j(\omega))\right\}.$$

The first-order condition is given by

$$
\begin{aligned}
0 \quad &= \quad \sum_i q_j^i(p_j(\omega)) + (p_j(\omega) - c_j) \sum_i \frac{\mathrm{d}q_j^i(p_j(\omega))}{\mathrm{d}p_j(\omega)} \\
&\overset{\text{(A.16), (A.24)}}{=} \sum_i q_j^i(p_j(\omega)) + \frac{p_j(\omega) - c_j}{p_j(\omega)} \sum_i \frac{f_j'(q_j^i(\omega))}{f_j''(q_j^i(\omega))}.
\end{aligned} \tag{A.25}
$$

The zero-profit condition is given by

$$
E_j = \pi_j(\omega) = (p_j(\omega) - c_j) \sum_i q_j^i(p_j(\omega)). \tag{A.26}
$$

These conditions characterize the decentralized equilibrium.

Now, consider the centralized revenue maximization problem. To compute the first-order conditions, we first prove a lemma. We define $q_j = \sum_i \int_0^{N_j} q_j^i(\omega)\,\mathrm{d}\omega$.

Lemma A.11. *View $\Upsilon_{iJ(j)}$ as a function of $c_{j'}, q_{j'}^{i'}(\omega'), E_{j'}$ and $N_{j'}$, we have*[62]

$$
\begin{aligned}
\frac{\partial \Upsilon_{J(j)}}{\partial q_j^i(\omega)} &= \frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} + \frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} = \frac{1 - \varepsilon_{E2j} + \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}} c_j \\
\frac{\partial \Upsilon_{J(j)}}{\partial N_j} &= \frac{\partial \Upsilon_{1J(j)}}{\partial N_j} + \frac{\partial \Upsilon_{2J(j)}}{\partial N_j} = \frac{\left(1 - \varepsilon_{E2j} + \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}\right)\frac{c_j q_j}{N_j} + \left(1 - \varepsilon_{c1j} + \varepsilon_{c2j}\frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}}\right) E_j}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}
\end{aligned}
$$

where $\varepsilon_{c\ell j} = \frac{\partial \ln c_j}{\partial \ln \Upsilon_{\ell J(j)}}$ and $\varepsilon_{E\ell j} = \frac{\partial \ln E_j}{\partial \ln \Upsilon_{\ell J(j)}}$.

Proof.

$$
\begin{aligned}
\frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} &= c_j + \sum_{j' \in J(j)} \frac{\partial c_{j'}}{\partial \Upsilon_{1J(j)}} \frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} q_{j'} + \sum_{j' \in J(j)} \frac{\partial c_{j'}}{\partial \Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} q_{j'} \\
&= c_j + \varepsilon_{c1j} \frac{\sum_{j' \in J(j)} c_{j'} q_{j'}}{\Upsilon_{1J(j)}} \frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} + \varepsilon_{c2j} \frac{\sum_{j' \in J(j)} c_{j'} q_{j'}}{\Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} \\
&= c_j + \varepsilon_{c1j} \frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} + \varepsilon_{c2j} \frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} \\
\frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} &= \sum_{j' \in J(j)} \frac{\partial E_{j'}}{\partial \Upsilon_{1J(j)}} \frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} N_{j'} + \sum_{j' \in J(j)} \frac{\partial E_{j'}}{\partial \Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} N_{j'} \\
&= \varepsilon_{E1j} \frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}} \frac{\partial \Upsilon_{1J(j)}}{\partial q_j^i(\omega)} + \varepsilon_{E2j} \frac{\partial \Upsilon_{2J(j)}}{\partial q_j^i(\omega)} \\
\frac{\partial \Upsilon_{1J(j)}}{\partial N_j} &= \frac{c_j q_j}{N_j} + \sum_{j' \in J(j)} \frac{\partial c_{j'}}{\partial \Upsilon_{1J(j)}} \frac{\partial \Upsilon_{1J(j)}}{\partial N_j} q_{j'} + \sum_{j' \in J(j)} \frac{\partial c_{j'}}{\partial \Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial N_j} q_{j'} \\
&= \frac{c_j q_j}{N_j} + \varepsilon_{c1j} \frac{\partial \Upsilon_{1J(j)}}{\partial N_j} + \varepsilon_{c2j} \frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial N_j} \\
\frac{\partial \Upsilon_{2J(j)}}{\partial N_j} &= E_j + \sum_{j' \in J(j)} \frac{\partial E_{j'}}{\partial \Upsilon_{1J(j)}} \frac{\partial \Upsilon_{1J(j)}}{\partial N_j} N_{j'} + \sum_{j' \in J(j)} \frac{\partial E_{j'}}{\partial \Upsilon_{2J(j)}} \frac{\partial \Upsilon_{2J(j)}}{\partial N_j} N_{j'}
\end{aligned}
$$

[62] $\partial q_j^i(\omega)$ means a unit increase in $q_j^i(\omega)$ for $\omega \in \Omega$ for a unit measure set Ω.

$$= E_j + \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}\frac{\partial\Upsilon_{1J(j)}}{\partial N_j} + \varepsilon_{E2j}\frac{\partial\Upsilon_{2J(j)}}{\partial N_j}.$$

Thus,

$$\frac{\partial\Upsilon_{1J(j)}}{\partial q_j^i(\omega)} = \frac{1-\varepsilon_{E2j}}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}c_j$$
$$\frac{\partial\Upsilon_{2J(j)}}{\partial q_j^i(\omega)} = \frac{\varepsilon_{E1j}}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}c_j$$

and

$$\begin{pmatrix} 1-\varepsilon_{c1j} & -\varepsilon_{c2j}\frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}} \\ -\varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}} & 1-\varepsilon_{E2j} \end{pmatrix}\begin{pmatrix} \frac{\partial\Upsilon_{1J(j)}}{\partial N_j} \\ \frac{\partial\Upsilon_{2J(j)}}{\partial N_j} \end{pmatrix} = \begin{pmatrix} \frac{c_j q_j}{N_j} \\ E_j \end{pmatrix}.$$

Thus, we have

$$\begin{pmatrix} \frac{\partial\Upsilon_{1J(j)}}{\partial N_j} \\ \frac{\partial\Upsilon_{2J(j)}}{\partial N_j} \end{pmatrix} = \frac{1}{N_j}\frac{1}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}\begin{pmatrix} 1-\varepsilon_{E2j} & \varepsilon_{c2j}\frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}} \\ \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}} & 1-\varepsilon_{c1j} \end{pmatrix}\begin{pmatrix} c_j q_j \\ E_j N_j \end{pmatrix}$$
$$= \frac{1}{N_j}\frac{1}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}\begin{pmatrix} (1-\varepsilon_{E2j})c_j q_j + \varepsilon_{c2j}\frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}}E_j N_j \\ \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}c_j q_j + (1-\varepsilon_{c1j})E_j N_j \end{pmatrix}. \qquad \square$$

Using Lemma A.11, we can compute the first-order conditions for the centralized revenue maximization problem:

$$(\partial q_j^i(\omega)) \quad \frac{1}{\lambda^i}\frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}}(f_j''(q_j^i(\omega))q_j^i(\omega) + f_j'(q_j^i(\omega))) = \tilde{\lambda}c_j\frac{1}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}}\left(1-\varepsilon_{E2j} + \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}\right)$$
$$(\partial N_j) \quad \sum_i \frac{1}{\lambda^i}\frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}} \cdot f_j'(q_j^i(\omega))q_j^i(\omega) = \frac{\tilde{\lambda}}{N_j}\frac{1}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j}) - \varepsilon_{c2j}\varepsilon_{E1j}} \cdot \bigg(c_j q_j + E_j N_j - \varepsilon_{E2j}c_j q_j$$
$$+ \varepsilon_{E1j}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}c_j q_j + \varepsilon_{c2j}\frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}}E_j N_j - \varepsilon_{c1j}E_j N_j\bigg).$$

First, consider the specification of Lemma A.4. The first-order conditions are simplified to

$$(\partial q_j^i(\omega)) \quad \cdots = \tilde{\lambda}c_j\frac{1}{(1-\varepsilon_{c1})(1-\varepsilon_{E2}) - \varepsilon_{c2}\varepsilon_{E1}}\left(1-\varepsilon_{E2} + \varepsilon_{E1}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}\right) \tag{A.27}$$
$$(\partial N_j) \quad \cdots = \frac{\tilde{\lambda}}{N_j}\frac{1}{(1-\varepsilon_{c1})(1-\varepsilon_{E2}) - \varepsilon_{c2}\varepsilon_{E1}} \cdot \left(c_j q_j + E_j N_j - \varepsilon_{E2}c_j q_j + \varepsilon_{E1}\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}c_j q_j + \varepsilon_{c2}\frac{\Upsilon_{1J(j)}}{\Upsilon_{2J(j)}}E_j N_j - \varepsilon_{c1}E_j N_j\right). \tag{A.28}$$

Unless ρ_j is the same across all j, the term $\frac{\Upsilon_{2J(j)}}{\Upsilon_{1J(j)}}$ varies across different j, so we should have $\varepsilon_{E1} = \varepsilon_{c2} = 0$ and $\varepsilon_{c1} = \varepsilon_{E2}$. When $f_j(\cdot)$ is CES with elasticity of substitution ρ_j, condition (A.25) implies $p_j(\omega) = \frac{\rho_j}{\rho_j-1}c_j$. With $\tilde{\lambda} = 1-\varepsilon_{c1}$, we can check that condition (A.27) holds and condition (A.28) coincides with condition (A.26). When $\rho_j = \rho$ for all j, we only need $\varepsilon_{c1} + \frac{1}{\rho-1}\varepsilon_{E1} = (\rho-1)\varepsilon_{c2} + \varepsilon_{E2}$, and conditions (A.27) and (A.28) hold with

$$\tilde{\lambda} = \frac{(1-\varepsilon_{c1})(1-\varepsilon_{E2}) - \varepsilon_{c2}\varepsilon_{E1}}{1 - \frac{\rho-1}{\rho}\varepsilon_{E2} + \frac{1}{\rho}\varepsilon_{E1} + \frac{\rho-1}{\rho}\varepsilon_{c2} - \frac{1}{\rho}\varepsilon_{c1}}.$$

Second, consider the specification of Lemma A.5. Note that we have

$$\begin{aligned}\frac{\ln c_j}{\ln \Upsilon_{\ell J(j)}} &= \frac{\ln c_j}{\ln \Upsilon_{J(j)}} \frac{\ln \Upsilon_{J(j)}}{\ln \Upsilon_{\ell J(j)}} = \frac{\ln c_j}{\ln \Upsilon_{J(j)}} \frac{\Upsilon_{\ell J(j)}}{\Upsilon_{J(j)}} \\ \frac{\ln E_j}{\ln \Upsilon_{\ell J(j)}} &= \frac{\ln E_j}{\ln \Upsilon_{J(j)}} \frac{\ln \Upsilon_{J(j)}}{\ln \Upsilon_{\ell J(j)}} = \frac{\ln E_j}{\ln \Upsilon_{J(j)}} \frac{\Upsilon_{\ell J(j)}}{\Upsilon_{J(j)}}.\end{aligned}$$

Thus, the first-order conditions are simplified to

$$(\partial q_j^i(\omega)) \qquad \cdots = \tilde{\lambda} c_j \frac{1}{1-\varepsilon_{J(j)}} \tag{A.29}$$

$$(\partial N_j) \qquad \cdots = \tilde{\lambda} \frac{c_j q_j + E_j N_j}{N_j} \frac{1}{1-\varepsilon_{J(j)}} \quad . \tag{A.30}$$

where

$$\varepsilon_{J(j)} = \varepsilon_c \frac{\Upsilon_{1J(j)}}{\Upsilon_{J(j)}} + \varepsilon_E \frac{\Upsilon_{2J(j)}}{\Upsilon_{J(j)}}.$$

By the same logic as of Lemma A.4, we need $\varepsilon_c = \varepsilon_E$ unless ρ_j is the same across all j. □

Proof of Lemma A.6. Aggregate revenue (i.e., the objective function of CRM) is given by

$$\sum_i \sum_j \int_0^{N_j} \frac{1}{\lambda^i} \frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}} \cdot f_j'(q_j^i(\omega)) \cdot q_j^i(\omega)\, \mathrm{d}\omega = \tfrac{\rho-1}{\rho} \sum_i \sum_j \frac{1}{\lambda^i} \frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}} q_j^i.$$

Therefore, the marginal change in the objective function of problem (CRM) is proportional to

$$\mathrm{d}\left(\sum_i \sum_j \frac{1}{\lambda^i} \frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}} q_j^i \right) = \sum_i \sum_j \frac{1}{\lambda^i} \frac{\partial U^i}{\partial q_j^i}\bigg|_{\text{de}} \mathrm{d}q_j^i, \tag{A.31}$$

while that of the social planner problem is

$$\mathrm{d}\left(\sum_i w^i \log U^i \right) = \sum_i \frac{w^i}{U^i} \mathrm{d}U^i = \sum_i \frac{w^i}{U^i} \sum_j \frac{\partial U^i}{\partial q_j^i} \mathrm{d}q_j^i. \tag{A.32}$$

Note that the budget constraint is homogeneous of degree 1 with respect to $\{q_j^i(\omega)\}_{j,\omega}$. Thus, when the mapping $\{q_j^i(\omega)\}_{j,\omega} \mapsto U^i$ is homogeneous of degree s, we can easily show that $\lambda^i|_{\text{de}} = s \cdot \frac{U^i|_{\text{de}}}{w^i}$ when evaluating at the decentralized allocation. Thus, equations (A.31) and (A.32) are proportional when evaluated at the decentralized allocation. This proves Lemma A.6. □

Proof of Lemma A.7. To show that the the minimized cost of producing $\{x_j(\sigma)\}_{j,\sigma}$ only depends on $\{x_j\}_j$, consider the cost-minimization problem of producing $\{x_j(\sigma)\}_{j\in J,\sigma}$ for a given partition J:

$$\begin{aligned}\min_{\{x_j(\omega;\sigma)\}_{j\in J,\sigma,\omega}} \quad & \sum_{j\in J} \sum_\sigma \pi(\sigma) \sum_\omega c_j(\omega; \Upsilon_{1J}) \cdot x_j(\omega;\sigma) \\ \text{s.t.} \quad & F(\{x_j(\omega;\sigma)\}_\omega) \geq x_j(\sigma), \quad \forall j \in J, \sigma\end{aligned} \tag{A.33}$$

Suppose that $\{x_j^*(\omega;\sigma)\}_{j\in J,\sigma,\omega}$ solve this problem and the minimized cost is Υ_{1J}^*. Then, $\{x_j^*(\omega;\sigma)\}_{j\in J,\sigma,\omega}$ also solve the following problem:[63]

$$\begin{aligned}\min_{\{x_j(\omega;\sigma)\}_{j\in J,\sigma,\omega}} \quad & \sum_{j\in J}\sum_{\sigma}\pi(\sigma)\sum_{\omega}c_j(\omega;\Upsilon_{1J}^*)\cdot x_j(\omega;\sigma) \\ \text{s.t.} \quad & F(\{x_j(\omega;\sigma)\}_{\omega}) \geq x_j(\sigma), \ \ \forall j\in J,\sigma\end{aligned} \tag{A.34}$$

Let $c(\{c_j(\omega;\Upsilon_{1J}^*)\}_{\omega})$ be the minimized cost of

$$\begin{aligned}\min_{\{x_j(\omega;\sigma)\}_{\omega}} \quad & \sum_{\omega}c_j(\omega;\Upsilon_{1J}^*)\cdot x_j(\omega;\sigma) \\ \text{s.t.} \quad & F(\{x_j(\omega;\sigma)\}_{\omega}) \geq 1,\end{aligned}$$

then, since $F(\cdot)$ is constant return to scale, the minimized cost of problem (A.34) is $\sum_{j\in J}c(\{c_j(\omega;\Upsilon_{1J}^*)\}_{\omega})\cdot(\sum_{\sigma}\pi(\sigma)x_j(\sigma))$. Thus, we can write

$$\Upsilon_{1J}^* = \sum_{j\in J}c(\{c_j(\omega;\Upsilon_{1J}^*)\}_{\omega})\cdot x_j,$$

and this implicit definition of Υ_{1J}^* implies that it only depends on $\{x_j\}_{j\in J}$. This means that the minimized cost of producing $\{x_j(\sigma)\}_{j,\sigma}$ only depends on $\{x_j\}_j$.

Thus the original problem (A.18) can be written as

$$\begin{aligned}\max_{\{x_j(\sigma)\}_{j,\sigma}} \quad & \tilde{U}(\{x_j(\sigma)\}_{j,\sigma}) \\ \text{s.t.} \quad & \sum_{\sigma}\pi(\sigma)x_j(\sigma) \leq x_j, \ \ \forall j \\ & \sum_{j}c(\{c_j(\omega;\Upsilon_{1J(j)})\}_{\omega})\cdot x_j \leq I.\end{aligned}$$

This means that we can solve the problem in two steps. First, we solve the problem for given values of $\{x_j\}_j$:

$$\begin{aligned}U(\{x_j\}_j) = \max_{\{x_j(\sigma)\}_{j,\sigma}} \quad & \tilde{U}(\{x_j(\sigma)\}_{j,\sigma}) \\ \text{s.t.} \quad & \sum_{\sigma}\pi(\sigma)x_j(\sigma) \leq x_j, \ \ \forall j.\end{aligned}$$

Second, we choose $\{x_j\}_j$ that maximize U subject to the resource constraint:

$$\max_{\{x_j\}_j} \quad U(\{x_j\}_j)$$

[63] Suppose to the contrary that $\{x_j^{**}(\omega;\sigma)\}_{j\in J,\sigma,\omega}$ solve problem (A.34) with the minimized objective function $\Upsilon_{1J}^{**} < \Upsilon_{1J}^*$. Let Υ_{1J}^{***} be the value of the objective function of (A.33) when $x_j(\omega;\sigma) = x_j^{**}(\omega;\sigma)$. Because $\{x_j^*(\omega;\sigma)\}_{j\in J,\sigma,\omega}$ solve problem (A.33), we must have $\Upsilon_{1J}^* \leq \Upsilon_{1J}^{***}$. However, we have

$$\begin{aligned}\Upsilon_{1J}^* &> \Upsilon_{1J}^{**} \\ &= \sum_{j\in J}\sum_{\sigma}\pi(\sigma)\sum_{\omega}c_j(\omega;\Upsilon_{1J}^*)\cdot x_j^{**}(\omega;\sigma) \\ &\geq \sum_{j\in J}\sum_{\sigma}\pi(\sigma)\sum_{\omega}c_j(\omega;\Upsilon_{1J}^{***})\cdot x_j^{**}(\omega;\sigma) \\ &= \Upsilon_{1J}^{***},\end{aligned}$$

which is a contradiction. The weak inequality comes from the fact that c_j is decreasing in $\Upsilon_{1J(j)}$.

$$\text{s.t.} \quad \sum_j c(\{c_j(\omega; \Upsilon_{1J(j)})\}_\omega) \cdot x_j \leq I.$$

We can easily see that this is equivalent to

$$\begin{aligned} \max_{\{x_j(\omega)\}_{j,\omega}} \quad & U(\{x_j\}_j) \quad \text{where } \; x_j = F(\{x_j(\omega)\}_\omega) \\ \text{s.t.} \quad & \sum_j \sum_\omega c_j(\omega; \Upsilon_{1J(j)}) \cdot x_j(\omega) \leq I. \end{aligned}$$

□

Proof of Proposition A.10. Using Lemma A.4, the first-order conditions of the constrained social planner problem are given by

$$\begin{aligned} &(\partial q^i_{jsd}(\omega)) && \frac{\mu\theta_i L_i}{C^i}\frac{\partial C^i}{\partial q^i_{jsd}}(q^i_{jsd})^{1/\rho} q^i_{jsd}(\omega)^{-1/\rho} = \tilde{\lambda}\tilde{c}_{jsd}L_i \cdot (1-\mathcal{E}_{1j}) \\ &(\partial N_{jsd}) && \sum_i \frac{\mu\theta_i L_i}{C^i}\frac{\partial C^i}{\partial q^i_{jsd}}\frac{\rho}{\rho-1}(q^i_{jsd})^{1/\rho} q^i_{jsd}(\omega)^{1-1/\rho} = \tilde{\lambda}\left(\sum_i \tilde{c}_{jsd}q^i_{jsd}(\omega)L_i + C_{jsd}\right)\cdot(1-\mathcal{E}_{2j}) \end{aligned}$$

where

$$\begin{aligned} 1-\mathcal{E}_{1j} &= \frac{1-\varepsilon_{E2j}+\frac{1}{\rho-1}\varepsilon_{E1j}}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j})-\varepsilon_{c2j}\varepsilon_{E1j}} \\ 1-\mathcal{E}_{2j} &= \frac{1-\frac{\rho-1}{\rho}\varepsilon_{E2j}+\frac{1}{\rho}\varepsilon_{E1j}+\frac{\rho-1}{\rho}\varepsilon_{c2j}-\frac{1}{\rho}\varepsilon_{c1j}}{(1-\varepsilon_{c1j})(1-\varepsilon_{E2j})-\varepsilon_{c2j}\varepsilon_{E1j}}. \end{aligned}$$

Using the first condition, we can simplify the second condition to

$$\frac{\rho}{\rho-1}\tilde{\lambda}\sum_i \tilde{c}_{jsd}q^i_{jsd}(\omega)L_i \cdot (1-\mathcal{E}_{1j}) = \tilde{\lambda}\left(\sum_i \tilde{c}_{jsd}q^i_{jsd}(\omega)L_i + C_{jsd}\right)\cdot(1-\mathcal{E}_{2j}).$$

Comparing these conditions with conditions (A.38) and (A.40), we can see that

$$S_{jsd} = \frac{\rho(\mathcal{E}_{2j}-\mathcal{E}_{1j})}{1-\mathcal{E}_{1j}+(\rho-1)(\mathcal{E}_{2j}-\mathcal{E}_{1j})} \quad \text{and} \quad s_{jsd}(\omega) = \mathcal{E}_{1j}$$

implement the optimal allocation. □

Proof of Proposition A.11. The first-order conditions of problem (DE′) are given by

$$\begin{aligned} &(\partial q^i_{jsd}(\omega)) && \frac{\mu}{C^i}\frac{\partial C^i}{\partial q^i_{jsd}}(q^i_{jsd})^{1/\rho} q^i_{jsd}(\omega)^{-1/\rho} = \lambda^i p_{jsd}(\omega) \\ &(\partial \tilde{C}_i) && \frac{1-\mu}{\tilde{C}_i} = \lambda^i p^{\text{tradable}}. \end{aligned}$$

Plugging these conditions into the consumer's budget constraint, we have

$$I_i + T_i = \frac{\mu}{\lambda^i C^i}\sum_{j,s,d}\frac{\partial C^i}{\partial q^i_{jsd}}(q^i_{jsd})^{1/\rho}\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho}\,\mathrm{d}\omega + \frac{1-\mu}{\lambda^i}$$

$$= \frac{\mu}{\lambda^i C^i} \sum_{j,s,d} \frac{\partial C^i}{\partial q^i_{jsd}} q^i_{jsd} + \frac{1-\mu}{\lambda^i}$$
$$= 1/\lambda^i$$

where the last equality comes from the fact that $C^i(\cdot)$ is homogeneous of degree one. The first-order condition with $\lambda^i = (I_i + T_i)^{-1}$ characterizes the demand

$$q^i_{jsd}(\omega) = \left(\frac{\mu(I_i+T_i)}{C^i} \frac{\partial C^i}{\partial q^i_{jsd}} (q^i_{jsd})^{1/\rho} \right)^{\rho} p_{jsd}(\omega)^{-\rho}$$
$$= \left(\frac{\mu(I_i+T_i)}{C^i} \frac{\partial C^i}{\partial q^i_{jsd}} (q^i_{jsd})^{1/\rho} \right)^{\rho} \cdot (1 - s_{jsd}(\omega))^{-\rho} \cdot \bar{p}_{jsd}(\omega)^{-\rho}.$$

This is isoelastic, so firms optimally set $\bar{p}_{jsd}(\omega) = \frac{\rho}{\rho-1}\tilde{c}_{jsd}$. Thus, the free-entry condition is given by

$$\bar{C}_{jsd} = \sum_i (\bar{p}_{jsd}(\omega) - \tilde{c}_{jsd}) q^i_{jsd}(\omega) L_i$$
$$= \frac{1}{\rho-1} \sum_i \tilde{c}_{jsd} q^i_{jsd}(\omega) L_i.$$

The laissez-faire equilibrium allocation $\{q^{i,\text{LF}}_{jsd}(\omega)\}, \{\tilde{C}^{\text{LF}}_i\}, \{N^{\text{LF}}_{jsd}\}$ should satisfy

$$\frac{\mu I_i}{C^{i,\text{LF}}} \frac{\partial C^i}{\partial q^i_{jsd}}\bigg|_{\text{LF}} (q^{i,\text{LF}}_{jsd})^{1/\rho} q^{i,\text{LF}}_{jsd}(\omega)^{-1/\rho} = \tfrac{\rho}{\rho-1}\tilde{c}_{jsd} \tag{A.35}$$
$$(1-\mu) I_i = \tilde{C}^{\text{LF}}_i \tag{A.36}$$
$$C_{jsd} = \tfrac{1}{\rho-1} \sum_i \tilde{c}_{jsd} q^{i,\text{LF}}_{jsd}(\omega) L_i. \tag{A.37}$$

The decentralized equilibrium allocation $\{q^{i,\text{DE}}_{jsd}(\omega)\}, \{\tilde{C}^{\text{DE}}_i\}, \{N^{\text{DE}}_{jsd}\}$ should satisfy

$$\frac{\mu (I_i + T_i)}{C^{i,\text{DE}}} \frac{\partial C^i}{\partial q^i_{jsd}}\bigg|_{\text{DE}} (q^{i,\text{DE}}_{jsd})^{1/\rho} q^{i,\text{DE}}_{jsd}(\omega)^{-1/\rho} = \tfrac{\rho}{\rho-1}(1 - s_{jsd}(\omega))\tilde{c}_{jsd} \tag{A.38}$$
$$(1-\mu)(I_i + T_i) = (1 + t^{\text{tradable}}) \tilde{C}^{\text{DE}}_i \tag{A.39}$$
$$(1 - S_{jsd}) C_{jsd} = \tfrac{1}{\rho-1} \sum_i \tilde{c}_{jsd} q^{i,\text{DE}}_{jsd}(\omega) L_i. \tag{A.40}$$

Now, consider the unconstrained social planner problem (SP′). Following the proof of Lemma A.4, the first-order conditions are

$$(\partial q^i_{jsd}(\omega)) \qquad \frac{\mu \theta_i L_i}{C^i} \frac{\partial C^i}{\partial q^i_{jsd}} (q^i_{jsd})^{1/\rho} q^i_{jsd}(\omega)^{-1/\rho} = \tilde{\lambda} \tilde{c}_{jsd} L_i \cdot \tfrac{1}{1+\varepsilon}$$
$$(\partial \tilde{C}_i) \qquad \frac{(1-\mu)\theta_i L_i}{\tilde{C}_i} = \tilde{\lambda} L_i$$
$$(\partial N_{jsd}) \qquad \sum_i \frac{\mu \theta_i L_i}{C^i} \frac{\partial C^i}{\partial q^i_{jsd}} \frac{\rho}{\rho - 1} (q^i_{jsd})^{1/\rho} q^i_{jsd}(\omega)^{1-1/\rho} = \tilde{\lambda} \left(\sum_i \tilde{c}_{jsd} q^i_{jsd}(\omega) L_i + C_{jsd} \right) \cdot \tfrac{1}{1+\varepsilon}$$

Using the first condition, we can simplify the third condition to

$$\tfrac{1}{\rho-1}\sum_i \tilde{c}_{jsd} q^i_{jsd}(\omega) L_i = C_{jsd}.$$

Plugging these conditions into the resource constraint of the social planner, we get

$$\begin{aligned}
\sum_i I_i L_i &= \tfrac{1+\varepsilon}{\tilde{\lambda}}\tfrac{\rho}{\rho-1}\sum_i \frac{\mu\theta_i L_i}{C^i}\sum_{jsd}\frac{\partial C^i}{\partial q^i_{jsd}}(q^i_{jsd})^{1/\rho}\int_0^{N_{jsd}} q^i_{jsd}(\omega)^{1-1/\rho}\,\mathrm{d}\omega + \tfrac{1}{\tilde{\lambda}}\sum_i(1-\mu)\theta_i L_i \\
&= \tfrac{1}{\tilde{\lambda}}\left(\tfrac{\rho}{\rho-1}(1+\varepsilon)\mu\sum_i \theta_i L_i + (1-\mu)\sum_i \theta_i L_i\right) \\
&= \tfrac{1}{\tilde{\lambda}}\left(1+\tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}\right)\sum_i \theta_i L_i.
\end{aligned}$$

where $\tilde{\rho}(\varepsilon) = \frac{\rho(1+\varepsilon)}{1+\rho\varepsilon}$ is decreasing in ε. Thus, with $\theta_i = I_i$ for all i, we need $\tilde{\lambda} = 1 + \frac{\mu}{\tilde{\rho}(\varepsilon)-1}$. In sum, the socially optimal allocation $\{q^{i,*}_{jsd}(\omega)\}, \{\tilde{C}^*_i\}, \{N^*_{jsd}\}$ should satisfy

$$\frac{\mu I_i}{C^{i*}}\frac{\partial C^i}{\partial q^i_{jsd}}\bigg|_{\text{SP}}(q^{i*}_{jsd})^{1/\rho} q^{i*}_{jsd}(\omega)^{-1/\rho} = \left(1+\tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}\right)\tilde{c}_{jsd}\cdot\tfrac{1}{1+\varepsilon} \tag{A.41}$$

$$(1-\mu)I_i = \left(1+\tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}\right)\tilde{C}^*_i \tag{A.42}$$

$$C_{jsd} = \tfrac{1}{\rho-1}\sum_i \tilde{c}_{jsd} q^{i*}_{jsd}(\omega) L_i. \tag{A.43}$$

We first show that $q^i_{jsd}(\omega) = q^{i,\text{LF}}_{jsd}(\omega), N_{jsd} = \chi(\varepsilon)\cdot N^{\text{LF}}_{jsd}, \tilde{C}_i = \frac{1}{1+\frac{\mu}{\tilde{\rho}(\varepsilon)-1}}\tilde{C}^{\text{LF}}_i$ where $\chi(\varepsilon) = \frac{\tilde{\rho}(\varepsilon)}{\tilde{\rho}(\varepsilon)-(1-\mu)}$ solve the unconstrained social planner problem. We then study how to implement this allocation using taxes and subsidies. Conditions (A.42) and (A.43) are immediate from the laissez-faire counterparts (A.36) and (A.37). Under this allocation, we can check that condition (A.41) also holds:

$$\begin{aligned}
\frac{\mu I_i}{C^i}\frac{\partial C^i}{\partial q^i_{jsd}}\bigg|_{\text{SP}}(q^i_{jsd})^{1/\rho} q^i_{jsd}(\omega)^{-1/\rho} &= \frac{\mu I_i}{\chi(\varepsilon)^{\frac{\rho}{\rho-1}} C^{i,\text{LF}}}\frac{\partial C^i}{\partial q^i_{jsd}}\bigg|_{\text{LF}}\chi(\varepsilon)^{\frac{1}{\rho-1}}(q^{i,\text{LF}}_{jsd})^{1/\rho} q^{i,\text{LF}}_{jsd}(\omega)^{-1/\rho} \\
&\overset{\text{(A.35)}}{=} \chi(\varepsilon)^{-1}\tfrac{\rho}{\rho-1}\tilde{c}_{jsd} \\
&= \left(1+\tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}\right)\tilde{c}_{jsd}\cdot\tfrac{1}{1+\varepsilon}.
\end{aligned}$$

Note that individual stores produce the same amount, but we have different number of stores, and hence different number for $q^i_{jsd} = \chi(\varepsilon)^{\frac{\rho}{\rho-1}} q^{i,\text{LF}}_{jsd}$. The first equality uses the fact that the mapping $\{q^i_{jsd}\}_{j,s,d} \mapsto C^i$ is homogeneous of degree one. Finally, the resource constraint holds,

$$\begin{aligned}
&\sum_i\left(\sum_{j,s,d}\int_0^{N_{jsd}}\tilde{c}_{jsd} q^i_{jsd}(\omega)\,\mathrm{d}\omega\right)L_i + \sum_{j,s,d} N_{jsd} C_{jsd} + \sum_i \tilde{C}_i L_i \\
&= \sum_i\left(\sum_{j,s,d}\int_0^{\chi(\varepsilon)N^{\text{LF}}_{jsd}}\tilde{c}_{jsd} q^{i,\text{LF}}_{jsd}(\omega)\,\mathrm{d}\omega\right)L_i + \chi(\varepsilon)\sum_{j,s,d} N^{\text{LF}}_{jsd} C_{jsd} + \frac{1}{1+\frac{\mu}{\tilde{\rho}(\varepsilon)-1}}\sum_i \tilde{C}^{\text{LF}}_i L_i \\
&= \mu\chi(\varepsilon)\tfrac{\rho-1}{\rho}\sum_i I_i L_i + \chi(\varepsilon)\tfrac{1}{\rho}\mu\sum_i I_i L_i + \frac{1}{1+\frac{\mu}{\tilde{\rho}(\varepsilon)-1}}(1-\mu)\sum_i I_i L_i \\
&= \sum_i I_i L_i.
\end{aligned}$$

This allocation can be implemented by the following taxes and transfers:

$$t^{\text{tradable}} = \tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}, \;\; s_{jsd}(\omega) = \tfrac{1-\mu}{\tilde{\rho}(\varepsilon)}, \;\; S_{jsd} = 0, \;\; \text{and} \;\; T_i = 0.$$

Comparing with conditions (A.41)−(A.43), we can easily check that conditions (A.38)−(A.40) hold. Finally, the consumer's budget constraint holds with $T_i = 0$:

$$\begin{aligned}
&\sum_{j,s,d} \int_0^{N_{jsd}} p_{jsd}(\omega) q^i_{jsd}(\omega)\,\mathrm{d}\omega + p^{\text{tradable}} \tilde{C}_i \\
&= \sum_{j,s,d} \int_0^{\chi(\varepsilon)\cdot N^{\text{LF}}_{jsd}} p^{\text{LF}}_{jsd} q^{i,\text{LF}}_{jsd}(\omega)\,\mathrm{d}\omega \cdot \Big(1 - \tfrac{1-\mu}{\tilde{\rho}(\varepsilon)}\Big) + \Big(1 + \tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}\Big) \frac{1}{1 + \tfrac{\mu}{\tilde{\rho}(\varepsilon)-1}} \tilde{C}^{\text{LF}}_i \\
&= \sum_{j,s,d} \int_0^{N^{\text{LF}}_{jsd}} p^{\text{LF}}_{jsd} q^{i,\text{LF}}_{jsd}(\omega)\,\mathrm{d}\omega + \tilde{C}^{\text{LF}}_i \\
&= I_i. \qquad \square
\end{aligned}$$