Granular Origins of Agglomeration *

Shinnosuke Kikuchi
UCSD

Daniel G. O'Connor
Princeton

October 23, 2025

Abstract

A few large firms dominate many local labor markets. How does that granularity affect the geography of economic activity and optimal place-based policy? To answer this question, we propose a new economic geography model featuring granular firms subject to idiosyncratic shocks. We show that average wages increase with the size of the local labor market due to that granularity, and we derive the optimal place-based policy. Using Japanese administrative data on manufacturing, we estimate our model and provide evidence consistent with our mechanism. Our mechanism implies that the smallest commuting zones have an elasticity of wages to population of 0.005. In large markets like Tokyo, the elasticity is around 0.001. Enacting optimal place-based industrial and wage policy would increase the number of people in the smallest cities, but the effect on the number of firms depends on firm conduct.

*We thank Daron Acemoglu, Treb Allen, Kosuke Aoki, David Atkin, David Autor, Arnaud Costinot, Dave Donaldson, Ryungha Oh, Yuhei Miyauchi, Enrico Moretti, Tomoya Mori, Kentaro Nakajima, Bradley Setzler, Bob Staiger, David Weinstein, Iván Werning, and Nathan Zorzi for their helpful comments. We also thank seminar participants at Chicago, Dartmouth, Hitotsubashi, NBER Japan Project Meeting, Minnesota, MIT, RIETI, and UTokyo. We thank Satoshi Ichikawa and Tomoko Yamaguchi for excellent research assistance. This study is a part of the project "Macroeconomy and Automation" undertaken at the Research Institute of Economy, Trade and Industry (RIETI) of Japan. Under that project, Kikuchi acknowledges permission to access the microdata of the Census of Manufacture (CoM) and the Economic Census for Business Activity (ECBA), granted by the Ministry of Economy, Trade and Industry and the Ministry of Internal Affairs and Communications in Japan.

1 Introduction

A few large firms dominate many local labor markets. Kodak accounted for almost a quarter of Rochester, New York's city payroll at its peak. Toyota employs a large proportion of the workforce in its headquarter-city, Toyota. Even in a large city like Seattle, software engineers are at the mercy of Microsoft and Amazon. With these giants, a shock to a single firm can hurt the entire market. If Microsoft has a bad year and lays off a large proportion of its software engineers, those workers might end up unemployed or at a low-wage job in another industry. How does this exposure to firm-specific shocks shape where workers and firms are located? And is there room for place-based policies to insulate labor markets from a single firm's influence?

In this paper, we study these questions theoretically, empirically, and quantitatively. Our analysis builds on the basic idea that in a small market, if an individual firm has a bad shock, the workers have nowhere else to go. They are stuck at that unproductive firm. In contrast, when a single firm becomes less productive in a large market, workers can move to another firm that is doing better and use their skills more productively. Thus, large markets provide a "constant market for skill" as Marshall (1920) said and Krugman (1992) formalized. This labor market pooling mechanism implies that larger markets use labor more effectively than small markets, so there are benefits to agglomeration when individual firms matter.

Our primary contribution is a new, tractable model of granular local labor markets that is still rich enough to allow for normative, empirical, and quantitative analysis. The model remains tractable because we introduce a continuum of sectors in each location and allow each sector to have a finite number of (ex-ante heterogeneous) firms. This ensures that workers who cannot move sectors in the short run are exposed to individual firm shocks. However, firm entry does not devolve into intractable inequality constraints because entry is not targeted towards a particular sector, but towards an entire location.

We start by showing that, in a granular world, there are benefits to agglomeration because firms use labor more effectively in large labor markets. In particular, firms in large markets expand their employment more in response to productivity shocks, and so they use more workers while they are productive. Therefore, average labor productivity is higher. To see why, consider a local sectoral labor market with only one firm. To expand after a good productivity shock, that firm would need to attract workers from other sectors, but that is difficult, as many workers have sector-specific skills. In contrast, if a firm hires a small share of the local sectoral labor market because the market is very large, it can expand by poaching workers from other firms in the same sector more easily. We

also show that the marginal benefits of increasing the size of the region disappear as the market becomes very large. That is because once a market is sufficiently large, firms no longer have any difficulty finding the workers they need. Therefore, increasing the size will not further improve productivity.

We go beyond this intuition and show that a statistic for how effectively a single firm uses labor is the covariance of that firm's log productivity and its log employment. That is to say, a firm uses labor more productively if it increases its employment when it has high productivity. Then the productivity of the entire market is just the average (employment-weighted) value of that covariance across firms. Thus, we can confirm our intuition of how the mechanism operates and quantify the contribution of granularity to the wage premium of large cities in a transparent, theory-consistent way.

Our last theoretical result considers the implications for policy. Our mechanism implies that too few firms enter in equilibrium. That is because, when firms are granular, they know that their entry affects wages. Not only will their entry increase average wages in a region, but it will increase wages precisely when the firm would like to hire more workers because their own attempt to increase their employment drives up the wages. That induces a correlation between wages and idiosyncratic firm shocks, which further depresses firm profits conditional on entry since profit functions are convex in wages and productivity. Therefore, firms will "under supply" their entry. A social planner can achieve a Pareto improvement using place-based subsidies on firm entry, especially in small locations where granularity matters.

We then enrich the model to test some of the predictions and to quantify the importance of our theoretical mechanism. We introduce imperfect mobility of labor across both firms and sectors, and we allow firms to internalize their market power, competing against each other à la Cournot. We then embed this model into a standard model of economic geography in which workers are mobile across locations. Our main theoretical results hold in this richer setting when firms are competitive, but they need to be adjusted when firms are imperfectly competitive. As we show, imperfect competition strengthens the agglomeration benefits of granularity. However, the optimal policy is different. The social planner still provides larger subsidies to smaller locations, but that policy might be targeted towards workers in terms of wage subsidies rather than firms in terms of entry subsidies.

Our quantitative and empirical analysis focuses on Japan, where we have a panel of all manufacturing establishments with at least 4 employees every year.[1] These data include

[1]For the analyses using establishment-level data, we restrict our samples to establishments with at least 10 employees.

employment, payroll, and shipments by 6-digit product category. We define a sector in the model as a 3-digit JSIC industry and a location as a commuting zone. We then estimate the key parameters of the model using these data.

Before we use the estimated model to quantify our mechanism, we look to validate the model, both qualitatively and quantitatively, by considering some reduced-form evidence of our mechanism. We start by providing evidence that granularity matters. We show that the variance of log payroll in a local sectoral labor market is decreasing in the number of firms in that sector, suggesting, consistent with Gabaix (2011), that individual firms are subject to idiosyncratic shocks. And those shocks average out in larger markets.

We then provide evidence consistent with our mechanism as suggested by the covariance statistic. We begin by showing that the variance of log employment at a single firm in a large local sectoral labor market is larger on average than that of a similarly situated firm in a small local sectoral labor market. That suggests that if those firms are subject to similar shocks, the firm in the larger market has an easier time expanding in response to good shocks and shrinking in response to bad ones. We also provide evidence that this is due to our mechanism rather than another mechanism, for instance, increasing returns to scale in the matching function. To provide more direct evidence, we construct revenue productivity shocks to each firm using that firm's exposure to national changes in demand for the products it produces. Consistent with our mechanism, both qualitatively and quantitatively, firms that already hire a large portion of the local sectoral labor market expand their employment less in response to these shocks.

Finally, we turn to demonstrating the quantitative importance of the mechanism. If firms are perfectly competitive, we find that the implied elasticity of wages to population gets as high as 0.005 in the smallest locations, and is a much smaller 0.001 in Tokyo. In total, the mechanism implies that Tokyo is 2.1% more productive than the smallest city in Japan. These numbers are larger if firms are imperfectly competitive. Combes et al. (2011) find that most causal estimates of the urban wage premium find an elasticity between 0.02 and 0.05 when pooling across locations of all sizes. Therefore, granularity could explain as much as 25% of the urban wage premium.

We then quantify the implications for optimal policy. We define the marginal product of labor and firms at the commuting zone level as the marginal contribution of another worker or firm to the output of the entire commuting zone. We then compare those values to the wages and profits in each location to see if workers and owners are properly compensated for their contributions. The results depend on assumptions about firm conduct. If firms are perfectly competitive, workers are paid their marginal product, and firm profits are only 97% of their marginal product in small locations. In contrast, in large loca-

tions, firms capture 99% of their contribution to commuting zone production. Putting in place the optimal firm entry subsidies would increase the number of firms in the smallest locations by 3.2% and in Tokyo by 1.2%. This reallocation of firms leads a small number of workers to move, increasing population in small areas by about 0.3%.

If firms are Cournot competitive in labor markets, workers in small locations see an average wage markdown of 3%. That markdown implies that firms' profits are actually 17% higher than their marginal contribution to production in small locations. This leads to over-entry of firms. Therefore, putting in place the optimal policies will actually lead to almost a 17% reduction in the number of firms in the smallest locations. Even Tokyo sees a 7% reduction in the number of firms. However, the optimal policy also features a wage subsidy in the smallest locations to undo the wage markdown. On net, these policies end up increasing the population in those small areas by more than 1%.

The rest of the paper is organized as follows. We give a short review of the literature below. In Section 2, we present the baseline model of a single location and show our theoretical results. We enrich the model in Section 3 and estimate it in Section 4. We validate the model and our mechanism in Section 5 before presenting the quantitative results in Section 6. Section 7 concludes and suggests ways the model could be enriched to capture important real-world features of granular markets.

Related Literature. The literature on the spatial agglomeration of economic activity is rich. In an early contribution, Marshall (1920) proposes three reasons why firms might locate around other firms: labor market pooling, access to intermediates, and the sharing of ideas. Subsequent theory papers have formalized these ideas and offered other potential mechanisms (Miyauchi, 2024; Davis and Dingel, 2019). Duranton and Puga (2004) provide a new way to classify these mechanisms in their review. Many empirical studies have shown that there are benefits from agglomeration (Andersson et al., 2014; Kline and Moretti, 2014; Greenstone et al., 2010) and have also analyzed the coagglomeration patterns of sectors to infer the relative importance of different theoretical mechanisms (Ellison and Glaeser, 1997; Ellison et al., 2010). Rosenthal and Strange (2004) review the evidence.

Our paper focuses on a particular mechanism that falls under the broad umbrella of labor market pooling. There are many microfoundations with different mechanisms of how labor market pooling can lead to agglomeration benefits (Andersson et al., 2007; Papageorgiou, 2022). Our model builds on the basic theoretical framework of Krugman (1992), which considers a setting with a finite number of ex-ante identical firms where labor is perfectly mobile within a labor market but not across. Overman and Puga (2010)

extend Krugman's model to include multiple sectors and test the predictions about where those sectors should be located. Other papers test these predictions in different settings (de Almeida and de Moraes Rocha, 2018; Nakajima and Okazaki, 2012). We provide a new model with ex-ante heterogeneous firms and a continuum of sectors that yields new theoretical results clarifying how granular firms generate agglomeration. We also derive the normative implications. Beyond the theoretical contribution, our model can be taken to the data to provide direct evidence of the mechanism and quantify its importance.

More recent work looks for direct evidence of the labor market pooling mechanism. Moretti and Yi (2024) show that workers who are laid off in large labor markets have an easier time finding work as compared to workers in small markets. This is consistent with the theory we lay out, though our evidence focuses on the firm response rather than the worker side. Conte et al. (2024) show that when firms in large markets can more easily expand in response to productivity shocks, more volatile firms will sort into larger markets. We abstract from firm sorting but demonstrate how granularity could explain why firms in larger markets can expand more in response to productivity shocks.

We build on a large literature recently inspired by Gabaix (2011) that looks to quantify what the granular nature of firms means for economic activity and optimal policy. Gabaix (2011) shows that shocks to individual firms could explain nationwide fluctuations. Bernard et al. (2018) give a framework for thinking about how a few important firms could shape the nature of international trade. Gaubert and Itskhoki (2021) discuss what granularity means for the observed comparative advantage of countries, and Gaubert et al. (2021) studies what that implies for optimal policy. We contribute to this literature by considering granularity's implications for the geography of economic activity and optimal place-based policy.

2 How Does Granularity Drive Agglomeration?

The goal of this section is to demonstrate how granularity leads to higher wages in larger markets through the labor market pooling mechanism discussed by Marshall (1920). In order to isolate the effects of granularity most transparently, we focus on the sectoral labor markets of a single location and consider a simplified environment that is neoclassical conditional on firm entry.

We will enrich the model by endogenizing where people live, allowing for imperfect mobility across firms and sectors, and introducing imperfect competition in Section 3.

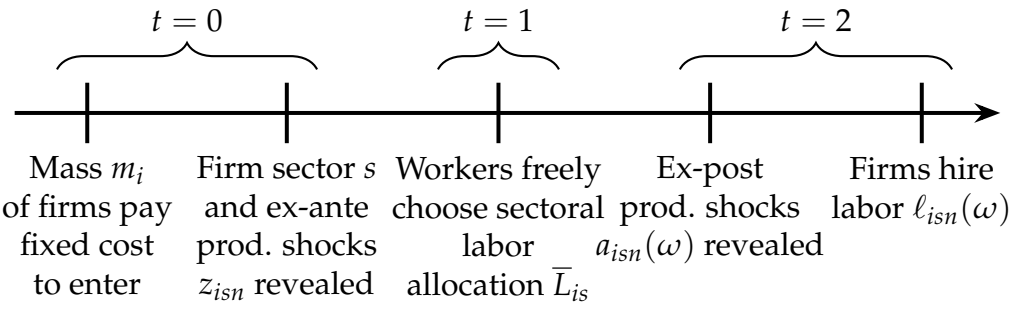


Figure 1: Timing of model

2.1 Environment

There is a region i with a mass ℓ_i of workers and a continuum of sectors $s \in [0, 1]$. The sectors produce perfectly substitutable goods but hire in distinct sectoral labor markets.

Timing. There are three periods $t \in \{0, 1, 2\}$. In period 0, a mass m_i of firms pay a fixed cost of the traded final good in order to enter. Each firm is randomly assigned a sector s and then gets an ex-ante productivity draw z from some known distribution. Thus, each sector ends up with a finite number of firms N_{is} that differ in their productivity even though a non-integer mass of firms might have entered the location i. This captures the long-run differences in firm size and will determine how exposed different sectoral markets are to short-run idiosyncratic firm shocks in period 2.

After observing those initial productivity draws, a representative worker freely allocates her labor $\overline{L}_{is}$ across sectors s in period 1. This captures the long-run decision of a worker free to direct her search or make training choices toward certain sectors in her region.

Then, in period 2, the state of the world $\omega \in \Omega$ is revealed. This determines the short-run productivity shocks to each firm. The worker cannot move labor across sectors at this time.[2] Instead, she supplies her sectoral labor inelastically, and firms hire labor in the sectoral labor market, taking wages as given. Firms then produce and sell their goods. The fact that firms are better able to respond to these ex post shocks in larger markets will imply the agglomeration benefits. The model timing is summarized in Figure 1.

Workers. In location i, there is a mass ℓ_i of risk-neutral representative agents. She gets utility from consuming a freely traded final good c_i and is endowed with one unit of labor that she supplies to the market inelastically. In period 1, the worker freely allocates her

[2]This assumption is done for clarity; none of the key results depend on it. We allow for imperfect movement across both firms and sectors in period 2 in Section 3.

units of labor across sectors $s \in [0,1]$, taking as given the number of firms and each firm's ex-ante productivity z_{isn}. In particular, she chooses her vector of labor supply $\boldsymbol{L_i} \equiv \{L_{is}\}_s$ in the set of feasible labor allocations $\mathcal{L}$, i.e.

$$\boldsymbol{L_i} \in \mathcal{L} \equiv \left\{ \boldsymbol{L_i'} | \int_0^1 L_{is}' ds \leq 1 \right\}.$$

In period 2, the state of the world ω is revealed. This determines ex-post productivity shocks for all firms. The worker is unable to adjust labor at this point, and so inelastically supplies L_{is} labor to sector s.

Firms. There is a continuum of potential firm entrants. To enter, a firm must pay a fixed cost $\psi_i > 0$ in terms of the freely traded final good in period 0. Those firms are then randomly assigned a sector. We denote by $\mathcal{N}_{is}$ the set of firms operating in region i sector s and $N_{is} \equiv |\mathcal{N}_{is}|$ the (finite) number of firms. We assume that firms enter according to the "ball-and-urn model" so that N_{is} is distributed Poisson with mean m_i. That is, the probability mass function for the number of firms in a sector is $m_i^N e^{-m_i}/N!$.

Firm n in sector s then gets an ex-ante productivity draw z_{isn} from a distribution $F_{iz}(\cdot)$ which we assume is continuous and regularly varying.[3] We further assume that the expected HHI of a sector is decreasing and convex in the number of firms when weighted by the average productivity of those firms.[4] This assumption rules out certain distributions for F_{iz} that imply HHI increases with more firms on average. This could happen if there is some probability that a new firm would dominate the market, as it is so much more productive than the other firms.

In period 2, each firm n gets an ex-post idiosyncratic productivity shock, $\tilde{a}_{isn}(\omega)$, a sector-wide productivity shock, $\tilde{A}_{is}(\omega)$, and produces a final good, $y_{isn}(\omega)$, according to,

$$y_{isn}(\omega) = z_{isn} a_{isn}(\omega) \ell_{isn}(\omega)^{1-\eta},$$

where $a_{isn}(\omega) \equiv \tilde{a}_{isn}(\omega)\tilde{A}_{is}(\omega)$ is the total productivity shock to firm n, $\ell_{isn}(\omega)$ is the total amount of labor firm n hires, and $\eta \in (0,1)$.

We assume that $\log \tilde{a}_{isn}(\omega)$ are iid with mean zero and a finite second moment σ_N^2. Similarly, $\log \tilde{A}_{is}(\omega)$ are iid with mean zero and a finite second moment σ_S^2. We further

[3] Formall,y $L : (0,\infty) \to (0,\infty)$ is regularly varying if $\lim_{x\to\infty} \frac{L(ax)}{L(x)} \in \mathbb{R}^+$ for all $a > 0$. In section 3, we will assume the ex-ante distribution is Pareto, which satisfies this condition.

[4] Formally, we assume that $\mathbb{E}\left[\frac{\sum_{n' \in \mathcal{N}} z_{isn'}^{1/\eta}}{N} \frac{\sum_{n' \in \mathcal{N}} \left(z_{isn'}^{1/\eta}\right)^2}{\left(\sum_{n' \in \mathcal{N}} z_{isn'}^{1/\eta}\right)^2} \middle| N \right]$ is decreasing in N and convex for sufficiently large N where η is defined below.

assume that the idiosyncratic and sector-wide shocks are independent of each other.

Market Clearing. Total expected production in the location is

$$Y_i = \mathbb{E}\left[\int_0^1 \sum_{n \in \mathcal{N}_{is}} z_{isn} a_{isn}(\omega) \ell_{isn}(\omega)^{1-\eta} ds\right],$$

where the expectation is taken with respect to ω, the number of firms in each sector, and the ex-ante productivity draws. We assume that the law of large numbers applies across this continuum of sectors so that realized production is always Y_i. Therefore, goods market clearing requires that total consumption plus the amount of the final good used for investment must equal expected production,

$$c_i \ell_i + \psi_i m_i = Y_i. \tag{1}$$

In the labor market, labor demanded needs to equal the individual labor supplied by each worker multiplied by the number of workers,

$$\sum_{n \in \mathcal{N}_{is}} \ell_{isn}(\omega) = L_{is} \ell_i, \quad \forall s, \omega. \tag{2}$$

2.2 Market Structure and Equilibrium

Labor Supply. Workers choose their labor allocation across sectors in period 1 to maximize their expected utility, taking wages as given. We normalize the price of the final good to 1, so workers solve the problem.

$$\boldsymbol{L_i} \in \underset{\boldsymbol{L_i'} \in \mathcal{L}}{\operatorname{argmax}} \quad \mathbb{E}\left[\int_0^1 w_{is}(\omega) L_{is}' ds\right], \tag{3}$$

where $w_{is}(\omega)$ is the equilibrium wage for sector s in state of the world ω. We will denote the maximum of (3) by w_i.

Labor Demand. We normalize productivity so that the price of each good is 1. Then each active firm maximizes profits, taking wages and prices as given,

$$\ell_{isn}(\omega) \in \underset{\ell'}{\operatorname{argmax}} \quad z_{isn} a_{isn}(\omega) (\ell')^{1-\eta} - w_{is}(\omega) \ell'. \tag{4}$$

We will denote the maximum of (4) by $\pi_{isn}(\omega)$.

Entry. We assume that firms enter up to the point that expected profits are equal to the fixed cost of entering. After those entry decisions are made, all firms are randomly assigned to their sector and get their productivity draws. We can write this,

$$\psi_i = \frac{\mathbb{E}\left[\int_0^1 \sum_{n \in \mathcal{N}_{is}} \pi_{isn}(\omega) ds\right]}{m_i}. \tag{5}$$

The numerator is the total amount of profits earned by firms in location i. The denominator is the total measure of firms that enter, so the ratio is a firm's expected profit before any firm realizes its sector or productivity shocks.[5]

Definition 1. A *local equilibrium* consists of wages $w_{is}(\omega)$, labor supply decisions L_{is}, entry decisions m_i, and labor demand decisions $\ell_{isn}(\omega)$, such that

- Workers make labor supply decisions to maximize utility, taking wages as given, (3);
- Conditional on entry, firms maximize profits taking prices and wages as given, (4);
- Firms enter up to the point that expected profits are equal to the fixed cost of entering, (5); and
- Goods and labor markets clear, (1) and (2).

2.3 Labor Market Pooling and Agglomeration

We now demonstrate how granularity implies that average wages increase with the number of workers. We will proceed in two steps. First, in this subsection, we will show that average wages increase with the number of workers if firms adjust their employment more in response to productivity shocks in larger markets. Throughout this paper, we will refer to this mechanism as labor market pooling, as it implies firms have an easier time finding the necessary workers to fill openings in larger markets.

Second, in Section 2.4, we will demonstrate how the existence of granular firms implies that firms can more easily expand their employment in response to shocks in larger markets. Combining this with Section 2.3, we conclude that granularity implies agglomeration benefits. Finally, in Section 2.5, we present the implications for optimal policy.

[5] An alternate entry decision would allow the marginal firm to observe the distribution of firms across sectors and their ex-ante productivity shocks before deciding whether or not to enter. We show that the free entry condition implied by that alternate entry game is also (5) if firms internalize that, were they to enter a sector with N firms, the sector would have $N+1$ firms in the online supplement.

For tractability, we will derive all of our results using a log second-order approximation to expected production around the point with no ex-post shocks. We will use $\overline{x}$ to denote the value of a variable x in the absence of any ex-post shocks.

We start by introducing the *Regional Production Function* as it will be useful to organize the discussion. The regional production function gives the maximum possible production for location i, taking as given the mass of firms m and the number of workers ℓ:

$$Y_i(\ell, m) \equiv \max_{\ell'_{sn}(\omega), L_s} \left\{ \mathbb{E}\left[\int_0^1 \sum_{n \in \mathcal{N}_s} z_{sn} a_{sn} \ell'_{sn}(\omega)^{1-\eta} \middle| m \right] \middle| \boldsymbol{L}' \in \mathcal{L}, \sum_{n \in \mathcal{N}_s} \ell'_{sn}(\omega) = L'_s \ell \right\}.$$

Since labor markets are perfectly competitive, this is not only the maximum level of production, but it also corresponds with the equilibrium production at the equilibrium levels of m and ℓ. The only difference between the regional production function and equilibrium production is that the number of firms is endogenously determined in equilibrium, while they are taken as given in the regional production function.

In the following Lemma, we characterize $Y(\ell, m)$ up to log second order.

Lemma 1. *The regional production function is* $Y_i(\ell, m) = z_i m^\eta \ell^{1-\eta} \Phi(m)$, *where* $z_i \equiv \mathbb{E}[z_{isn}^{1/\eta}]^\eta$ *and* $\Phi(m)$ *is given by,*

$$\Phi(m) \equiv \mathbb{E}[a_{sn}(\omega)] + \frac{1-\eta}{2} \int_0^1 \frac{\overline{\ell}_s}{\ell} \sum_{n \in \mathcal{N}_s} \frac{\overline{\ell}_{sn}}{\overline{\ell}_s} \operatorname{Cov}\left(\log a_{sn}(\omega), \log \ell_{sn}(\omega)\right) ds, \tag{6}$$

where the covariance is for a single firm across different states of the world.

The formal proof of Lemma 1 is in the online supplement, but we will provide a quick sketch here. We start by solving the maximization problem with no ex-post shocks. We then do a second-order approximation to the maximand, which implies $Y_i(\ell, m) = z_i \ell^{1-\eta} m^\eta \tilde{\Phi}(m)$ where

$$\begin{aligned} \tilde{\Phi}(m) \;\equiv\; & \mathbb{E}[a_{sn}(\omega)] + (1-\eta) \int_0^1 \frac{\overline{\ell}_s}{\ell} \sum_{n \in \mathcal{N}_s} \frac{\overline{\ell}_{sn}}{\overline{\ell}_s} \operatorname{Cov}\left(\log a_{sn}(\omega), \log \ell'_{sn}(\omega)\right) ds \\ & -\eta \frac{1-\eta}{2} \int_0^1 \frac{\overline{\ell}_s}{\ell} \sum_{n \in \mathcal{N}_s} \frac{\overline{\ell}_{sn}}{\overline{\ell}_s} \operatorname{Var}\left(\log \ell'_{sn}(\omega)\right) ds. \end{aligned} \tag{7}$$

That is, relative to keeping labor at its no-shock level in every state of the world, there are gains from increasing labor at firms that have good productivity shocks and decreasing labor at firms with bad productivity shocks. Those gains are tempered by the fact that there are decreasing returns to scale, so that productivity decreases if there is a very high

variance of $\log \ell'_{sn}(\omega)$. When labor is efficiently allocated, either because the market is competitive or there is a planner,

$$\eta \int_0^1 \frac{\overline{\ell}_s}{\ell} \sum_{n \in \mathcal{N}_s} \frac{\overline{\ell}_{sn}}{\overline{\ell}_s} \operatorname{Var}\left(\log \ell'_{sn}(\omega)\right) ds = \int_0^1 \frac{\overline{\ell}_s}{\ell} \sum_{n \in \mathcal{N}_s} \frac{\overline{\ell}_{sn}}{\overline{\ell}_s} \operatorname{Cov}\left(\log a_{sn}(\omega), \log \ell'_{sn}(\omega)\right) ds, \quad (8)$$

proving the result. This leads to our first substantive result.

Corollary 1. *The regional production function features increasing returns to scale, i.e.* $\frac{dY_i(\alpha\ell,\alpha m)}{d\alpha} > Y_i$*, if and only if the employment weighted covariance between productivity and employment is increasing in the number of firms, i.e.* $\frac{\partial \Phi(m)}{\partial m} > 0$.

This result follows immediately from Lemma 1 since

$$\frac{dY_i(\alpha\ell, \alpha m)}{d\alpha} = \frac{d}{d\alpha}\left[\alpha z_i m^{\eta} \ell^{1-\eta} \Phi(\alpha m)\right] = Y_i(m, \ell)\left(1 + \frac{\partial \log \Phi(m)}{\partial \log m}\right).$$

Corollary 1 provides a clear interpretation of how labor market pooling implies that the regional production function features increasing returns to scale. Consider a firm n that never adjusts its workforce in response to idiosyncratic productivity shocks so that the covariance is 0. That firm would hold onto a large number of workers when it has bad productivity shocks and not expand to take advantage of good productivity shocks. Therefore, its average labor productivity would depend solely on its average productivity shock. By contrast, if the firm were to expand after a good productivity shock and shrink after a bad shock, its average labor productivity would increase because it hires more workers when it is more productive. Thus, that firm could produce more goods on average while hiring the same average number of workers simply by increasing its covariance.

Corollary 1 then says that there are increasing returns to scale if increasing the size of the market improves how much labor reallocates across firms in response to productivity shocks, properly weighted by the importance of each firm. Thus, if firms are better able to expand their employment in response to good productivity shocks in a larger market, then larger markets will be more productive.

Relation to Misallocation. This mechanism is closely related to the misallocation literature (Hsieh and Klenow, 2009). To see this in the math, note that for any firm on its labor demand curve, $w_s(\omega) = (1-\eta) z_{sn} a_{sn}(\omega) \ell_{sn}(\omega)^{-\eta}$. Therefore, looking at the variance of

log wages,

$$\int_0^1 \frac{\overline{\ell}_s}{\overline{\ell}} \operatorname{Var}(\log w_s(\omega))ds = \operatorname{Var}(\log a_{sn}(\omega)) - \eta \int_0^1 \frac{\overline{\ell}_s}{\overline{\ell}} \sum_{n \in \mathcal{N}_s} \frac{\overline{\ell}_{sn}}{\overline{\ell}_s} \operatorname{Cov}(\log a_{sn}(\omega), \log \ell_{sn}(\omega))ds,$$

using equation (8). That is, the employment weighted variance of log wages is decreasing in the employment weighted covariance of productivity and labor. This immediately implies the next corollary.

Corollary 2. *The regional production function features increasing returns to scale* $\frac{dY_i(\alpha\ell,\alpha m)}{d\alpha} > Y_i$ *if and only if the employment weighted variance of log wages is decreasing in the number of firms, i.e.* $\frac{\partial \Phi(m)}{\partial m} > 0$.

Corollary 2 says that larger markets are more productive than smaller markets if the average variance of log wages is lower in large markets. Loosely speaking, this is because there is less "misallocation" of labor. More precisely, the variance of log wages in sector s is a measure of how much more productive labor in sector s is in some states of the world than others. Therefore, it can be thought of as a measure of how unproductively labor is used on average.

Importantly, unlike misallocation, this variance is not a sign of inefficiency. Markets are perfectly competitive and are therefore efficient. Workers would love to reallocate some of their labor from states of the world in which the sectoral wage is low to states of the world where wages are high if they could. But that is not possible because the labor is stuck in each sector across all states of the world.

Agglomeration Benefits. We next turn to show how the increasing returns to scale of the regional production function $Y_i(\ell, m)$ imply that there are benefits to agglomeration. To do that, we need to relate the equilibrium number of firms to $Y_i(\ell, m)$. This is straightforward as wages are set competitively, so workers are paid their marginal product. That is,

$$w_i = \frac{\partial Y_i}{\partial \ell} = \frac{(1-\eta)Y_i}{\ell_i}. \tag{9}$$

Therefore, total profits are simply $Y_i - w_i\ell_i = \eta Y_i$, so that free entry (5) can be rewritten,

$$\psi_i = \frac{\eta Y_i}{m_i}. \tag{10}$$

Combining equations (9) and (10) with Corollary 1 then implies the next proposition.

Proposition 1. *In any stable equilibrium, the average wage is increasing in the number of workers, i.e.* $\frac{d \log w_i}{d \log \ell_i} > 0$*, if and only if the employment weighted covariance between log firm productivity and log firm employment is increasing in m, i.e.* $\frac{\partial \log \Phi(m)}{\partial \log m}\Big|_{m=m_i} > 0$.

Proposition 1 summarizes the basic argument of labor market pooling. If firms have an easier time expanding in response to good productivity shocks in larger markets than in smaller markets, then average wages will increase with the size of the market.

2.4 Granularity and Labor Market Pooling

The previous subsection demonstrated how labor market pooling mechanisms can make larger markets more productive than smaller markets. In this subsection, we complete the argument that granularity implies agglomeration benefits by demonstrating how granularity implies those labor market pooling forces.

Proposition 2. *The average wage is increasing in the number of workers, i.e.,* $\frac{d \log w_i}{d \log \ell_i} > 0$*, if and only if idiosyncratic shocks have a positive variance,* $\sigma_N^2 > 0$*. Furthermore, the agglomeration benefits converge to zero as the size of the market goes to infinity, i.e.* $\frac{d \log w_i}{d \log \ell_i} \to 0$ *as* $m_i \to \infty$.

We start by giving a basic intuition for why there are increasing returns to scale using Proposition 1. Consider how firm n in sector s responds to an idiosyncratic, ex-post productivity shock $\Delta \log a_{isn}(\omega)$. To first order, for a firm on its labor demand curve,

$$\Delta \log \ell_{isn}(\omega) = \frac{1}{\eta} \left(\Delta \log a_{isn}(\omega) - \Delta \log w_{is}(\omega) \right) = \frac{1}{\eta} \left(1 - \frac{\bar{\ell}_{isn}}{\bar{\ell}_{is}} \right) \Delta \log a_{isn}(\omega).$$

In a small region, where the mass of firms m_i is small, chances are that there are very few other firms in sector s. Therefore, firm n hires a large share of the labor force in sector s, i.e. $\bar{\ell}_{isn}/\bar{\ell}_{is}$ is large. In that case, the firm's employment does not respond much after a productivity shock because it already hires a large proportion of the sectoral labor force, and it cannot attract workers from other sectors. Instead, it drives up the wages in the sector. Therefore, the firm does not effectively scale up in response to a productivity shock and ends up using labor unproductively. In a market with a large mass of firms m, firm n's share of the sector s labor force is smaller. Therefore, firm n's labor responds more in response to productivity shocks because it can poach workers from other firms in the sector, using that labor more productively.

The formal proof for Proposition 2 is in Appendix A, but we give a sketch of the proof here. Allowing for productivity shocks to every firm in sector s implies that, to first order, labor at firm n in sector s is given by

$$\Delta \log \ell_{isn}(\omega) = \frac{1}{\eta}\left(\Delta \log a_{isn}(\omega) - \sum_{n' \in \mathcal{N}_{is}} \frac{\bar{\ell}_{isn}}{\bar{\ell}_{is}} \Delta \log a_{isn'}(\omega)\right).$$

Then, some straightforward algebra reveals that the employment weighted average covariance is given by

$$\int_0^1 \frac{\bar{\ell}_{is}}{\bar{\ell}_i} \sum_{n \in \mathcal{N}_{is}} \frac{\bar{\ell}_{isn}}{\bar{\ell}_{is}} \text{Cov}(\log a_{isn}(\omega), \log \ell_{isn}(\omega)) ds = \frac{1}{\eta}\left(1 - \int_0^1 \frac{\bar{\ell}_{is}}{\bar{\ell}_i} HHI_{is} ds\right) \sigma_N^2,$$

where $HHI_{is} \equiv \sum_{n \in \mathcal{N}_{is}} \left(\frac{\bar{\ell}_{isn}}{\bar{\ell}_{is}}\right)^2$. This covariance depends on the number of firms only through the HHI_{is}. Thus, the degree of increasing returns to scale is

$$\frac{\partial \log \Phi(m)}{\partial \log m} = -\frac{1}{2}\frac{1-\eta}{\eta}\frac{\frac{\partial}{\partial \log m}\left[\int_0^1 \frac{\bar{\ell}_{is}}{\bar{\ell}_i} HHI_{is} ds\right]}{\Phi(m)} \sigma_N^2. \tag{11}$$

If the average *HHI* across sectors decreases as the number of firms m increases, Proposition 1 is satisfied, and there are agglomeration benefits. By assumption, the expected *HHI* of a given sector s is decreasing in the number of firms N_{is}, so the last thing we need to check is that firm entry across sectors is not too strange so that average HHI across sectors in location i is decreasing in the mass of firm entrants, m_i. For example, suppose that there were only two sectors with firms in them: one with two firms and another with one firm. If firm entry goes into unoccupied sectors, the average *HHI* would actually increase as the average number of firms in occupied sectors decreases. This, and other pathological cases, are ruled out with Poisson entry.

The speed with which the average *HHI* decreases depends on the distribution of entrants across sectors and the ex-ante productivity distribution $F_{is}(z)$. Entry especially matters for low m. As m becomes larger, the ex-ante productivity distribution matters more. As discussed in Gabaix (2011), if $F_{is}(z)$ has thin tails, HHI_{is} decreases approximately at the rate of N_{is}^{-1}. Then the granular agglomeration forces are strong when there is a small number of firms and the *HHI* falls quickly with new entrants. However, the *HHI* quickly approaches zero, at which point the average *HHI* cannot fall further. For example, if a sector has one firm and adds another ex-ante identical firm, the HHI drops from 1 to 0.5. If a sector already has 100 identical firms, the *HHI* is 0.01, doubling the number of firms

only decreases it to 0.005, not increasing productivity much. Intuitively, that is because if there are already 100 firms in a sector, it is easy for any firm to expand by attracting workers from the other 99 firms. Adding more firms does not have much effect.

If $F_{is}(z)$ has thick tails, HHI_{is} decreases at a rate of $N_{is}^{-\xi}$ where $\xi \in (0,1)$.[6] In that case, the weighted *HHI* does not fall as quickly, and large firms continue to be constrained in their ability to respond to productivity shocks. Thus, the externality is not as strong for small markets, but matters for medium-sized and even large cities. Nonetheless, the externality becomes less important as the number of firms increases, and the average *HHI* approaches zero.

2.5 Optimal Policy

In this subsection, we consider what this force for agglomeration means for policy. Conditional on firm entry, the model is competitive, so the only possible source of inefficiency is firm entry. Rewriting the goods market clearing condition (1), the first-best level of entry m_i^{FB} maximizes the total production net of firm entry costs,

$$m_i^{FB} \in \underset{m'}{\text{argmax}} \quad Y_i(\ell_i, m') - \psi_i m'. \tag{12}$$

Taking the first order condition, we find that the first-best entry must satisfy

$$\psi_i = \left(1 + \frac{1}{\eta}\frac{\partial \log \Phi\left(m_i^{FB}\right)}{\partial \log m}\right)\frac{\eta Y_i}{m_i^{FB}}.$$

In contrast, equilibrium entry satisfies equation (10). Therefore, to implement the first best entry, the planner needs to enact a firm entry subsidy proportional to profits given by $\tau_i = \frac{1}{\eta}\frac{\partial \log \Phi(m)}{\partial \log m}$. And as $\frac{\partial \log \Phi(m)}{\partial \log m} \to 0$ as $m \to \infty$, the size of the optimal subsidy will also converge to zero in large markets. This implies our normative result.

Proposition 3. *If idiosyncratic shocks have a positive variance, $\sigma_N^2 > 0$, the optimal policy features a subsidy on firm entry proportional to firm profits given by $\tau_i = \frac{1}{\eta}\frac{\partial \log \Phi(m)}{\partial \log m}\big|_{m=m_i} > 0$. Furthermore, the optimal subsidy converges to zero as the size of the market goes to infinity, i.e. $\tau_i \to 0$ as $m_i \to \infty$.*

One might be confused as to why there is any reason for policy at all, since wages are competitively set in our setting. The first welfare theorem breaks down because firm

[6] Formally, If $F_{is}(z)$ is a pareto distribution, i.e. $F_{is}(z) = 1 - az^{-\lambda}$, it is thick-tailed if $1 < \lambda\eta < 2$. Then $\xi = 2\left(1 - \frac{1}{\lambda\eta}\right)$.

entry is not Walrasian. In a competitive equilibrium, firms need to take prices and wages as given, but in our model, we allow firms to ask what their profits would be conditional on entry. That requires a firm to internalize the effect that its entry will have on wages. Firms know that by entering, wages will be higher. Not only will wages be higher on average, but they will be higher precisely when the firm would like to hire more workers because their own idiosyncratic shocks affect wages. This leads firms to not enter when a similarly situated competitive firm would enter because the competitive firm would take as given the observed, uncorrelated distribution of wages. This will be especially relevant in small markets where adding one more firm will have a big effect on the distribution of wages.

A skeptical reader might wonder why we depart from competitive entry. The reason is simple. An equilibrium with competitive entry does not exist. To see this, note that solving for the optimal labor choice and plugging in for profits, we get that

$$\pi_{isn}(\omega) = \eta(1-\eta)^{\frac{1-\eta}{\eta}} a_{isn}(\omega)^{\frac{1}{\eta}} w_{is}(\omega)^{-\frac{1-\eta}{\eta}}.$$

Taking a log second-order approximation to profits around the no ex-post shock equilibrium implies that

$$\mathbb{E}[\pi_{isn}(\omega)] \approx \overline{\pi}_{isn}\left[\zeta_s - \frac{1-\eta}{\eta^2}\operatorname{Cov}(\log a_{isn}(\omega), \log w_{is}(\omega))\right]$$

where ζ_s is some sector wide constant and $\overline{\pi}_{isn}$ are firm profits when the ex-post shocks are 1. That is, profits are decreasing in the covariance between firm productivity and equilibrium wages. This is because profit functions are convex, so firms appreciate the variance that comes when productivity and wages are not correlated.

For a firm that is currently operating, wages are correlated with their shocks because their attempt to hire more workers drives up wages. On the other hand, a potential entrant's productivity shocks are not correlated with wages. Therefore, given an equilibrium distribution of wages, the potential entrant expects strictly higher profits than the operating firm. But in any competitive equilibrium, operating firms must expect weakly positive profits while potential entrants must expect weakly negative profits. This is not possible when potential entrants expect strictly higher profits than current entrants, and therefore, no competitive equilibrium exists.

One can also think of this result using the regional production function $Y_i(\ell, m)$. In proving Proposition 2, we showed that the production function has increasing returns to

scale, i.e. $\frac{dY_i(\alpha\ell,\alpha m)}{d\alpha} > Y_i$. One can rewrite this as,

$$\frac{\partial Y_i}{\partial \ell}\ell_i + \frac{\partial Y_i}{\partial m}m_i > Y_i.$$

Therefore, if workers and firms were both paid their marginal product, that is $w_i = \frac{\partial Y_i}{\partial \ell}$ and $\pi_i = \frac{\partial Y_i}{\partial m}$, the firms in the location would need to pay out more than the revenue they are earning. This is clearly not possible, so some factor of production must be underpaid. And because we assume that the labor market is competitive, workers are paid their marginal product, and firms must be underpaid.

3 A Quantitative, Granular Model of Economic Geography

In Section 2, we presented a stylized model of sectoral labor markets within a single location where workers were freely mobile within a sector and unable to move across sectors. In this section, we model a country made up of I locations indexed by $i \in \mathcal{I} \equiv \{1, \ldots, I\}$. We do this by introducing a migration decision at time $t = -1$, allowing workers to choose where to live. We then assume that workers are stuck in their location for the remaining periods.

We extend the model of the labor market in Section 2 so that in period 1, workers allocate their labor across sectors and firms in their location. Then, in period 2, workers can move their labor across both firms and sectors, subject to movement frictions. We further allow firms to internalize their labor market power so that they compete with other firms in their sector according to Cournot competition. We show that the main results remain unchanged in this extended model with perfect competition, and the results only need to be adjusted slightly with imperfect competition. A more detailed description of the model, a characterization of the equilibrium, and proofs of the results are in the online supplement.

3.1 Extending the Baseline Model

Migration Across Regions. There is a mass ℓ of workers in the country. The fundamental utility of living in location i is $u_i = \overline{u}_i w_i$, where $\overline{u}_i$ is the local amenities. Each worker has an idiosyncratic preference for each location ε_i so that the utility the worker gets from living in location i is $u_i\varepsilon_i$.

We assume that ε_i are distributed Fréchet with shape parameter $\theta > 0$. Therefore, when people are free to live where they would like, the number of people who live in

location i is given by $\ell_i = (u_i/u)^{\theta} \ell$, where $u = \left(\sum_{i \in \mathcal{I}} (u_i)^{\theta}\right)^{\frac{1}{\theta}}$.

Imperfect Mobility across Firms and Sectors. In period 1, the representative worker freely allocates her units of labor across sectors $s \in [0,1]$ and firms $n \in \mathcal{N}_{is}$, taking as given each firm's ex-ante productivity z_{isn}. In particular, she chooses her vector of labor supply $\boldsymbol{L_i} \equiv \{L_{isn}\}_{s,n}$ in the set of feasible labor allocations $\mathcal{L}$,

$$\boldsymbol{L_i} \in \mathcal{L} \equiv \left\{ \boldsymbol{L_i'} | \int_0^1 \sum_{n \in \mathcal{N}_{is}} L'_{isn} ds \leq 1 \right\}.$$

In period 2, the state of the world ω is revealed. This determines the ex-post productivity shocks for all firms. The worker then reallocates labor across firms, choosing a vector of labor supply $\boldsymbol{L_i(\omega)} \equiv \{L_{isn}(\omega)\}_{s,n}$ in the set of feasible labor allocations $\mathcal{L}_\Omega(\boldsymbol{L_i})$ which depends on the worker's labor choices in period 1. The set is given by

$$\mathcal{L}_\Omega(\boldsymbol{L_i}) \equiv \left\{ \boldsymbol{L_i(\omega)} | 1 = \left(\int_0^1 L_{is}^{-\frac{1}{\nu}} L_{is}(\omega)^{\frac{1+\nu}{\nu}} ds \right)^{\frac{\nu}{1+\nu}}, \right.$$
$$\left. L_{is}(\omega) = \left(\sum_{n \in \mathcal{N}_{is}} \left(\frac{L_{isn}}{L_{is}} \right)^{-\frac{1}{\kappa}} L_{isn}(\omega)^{\frac{1+\kappa}{\kappa}} \right)^{\frac{\kappa}{1+\kappa}} \right\},$$

where $L_{is} = \sum_{n \in \mathcal{N}_{is}} L_{isn}$.

This set implies that $L_{is}(\omega) = L_{is}$ and $L_{isn}(\omega) = L_{isn}$ is feasible, but the more that the worker deviates from period 1 labor choices, the more labor is lost in transition. $\kappa > 0$ is the short-run elasticity of substitution across firms within a sector, and $\nu > 0$ is the short-run elasticity of substitution across sectors. The worker in location i chooses $\boldsymbol{L_i}$ and $\{\boldsymbol{L_i(\omega)}\}_\omega$ to maximize expected utility taking wages as given.

Ex-ante Productivity Distribution. The ex-ante productivity distribution is distributed Pareto with shape parameter λ and location parameter z_i that can differ across locations. We assume that $\lambda\eta > 1$ so that the expected size of each firm is finite.

Labor Market Competition. For the rest of the paper, we will present results under alternate assumptions on firm behavior. Under perfect competition, firms continue to maximize profits, taking wages as given. However, now that workers are imperfectly substitutable across firms, each firm has its own wage $w_{isn}(\omega)$ that it takes as given.

Under imperfect competition, firms internalize that they can affect the wages they face. We assume that the firm can commit to a wage schedule in period 1, so that the

firm internalizes how wage choices in state of the world ω will affect how much period 1 labor the worker supplies to the firm. This implies that, in the absence of ex-post shocks, wages will correspond with competitive wages as workers are freely mobile in period 1. With ex-post shocks, firms will sometimes hire a larger portion of the market and have more market power. Therefore, the firm can vary its markdown across different states of the world to increase profits. We assume Cournot competition, so that each firm takes as given the labor hired by the other firms in its sector and the work opportunities in the other sectors.

3.2 Theoretical Results with Perfect Competition

When firms engage in perfect competition, the key results, Lemma 1, Proposition 1, Proposition 2, and Proposition 3 do not change at all as long as workers are more mobile across firms within a sector than across sectors, $\kappa > \nu$. The only qualitative change is that Corollary 2, relating the mechanism to misallocation, no longer holds. That is because misallocation only measures how efficiently existing labor is used. With movement frictions, the worker also needs to consider how much labor is lost in moving labor across firms and sectors.

Quantitatively, numbers change as the covariance, determining the strength of agglomeration, is given by

$$\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i} \sum_{n \in \mathcal{N}_{is}} \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}} \operatorname{Cov}(\log a_{isn}(\omega), \log \ell_{isn}(\omega)) ds$$
$$= \frac{1}{\eta + \frac{1}{\nu}} \sigma_S^2 + \frac{\eta + \frac{1}{\nu} - \left(\frac{1}{\nu} - \frac{1}{\kappa}\right) \int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i} HHI_{is} ds}{\left(\eta + \frac{1}{\kappa}\right)\left(\eta + \frac{1}{\nu}\right)} \sigma_N^2.$$

The size of the externality then depends on how much harder it is for a firm to attract workers from other sectors compared to workers from other firms in its own sector, $\frac{1}{\nu} - \frac{1}{\kappa}$.

3.3 Theoretical Results with Imperfect Competition

In this subsection, we discuss how the theoretical results change when firms compete á la Cournot.

Most importantly, Lemma 1 does not hold with Cournot competition as the covariance expression relied on the market efficiently trading off the gains from moving workers to more productive firms against the loss of labor and decreasing returns to scale. However,

the production function still has a simplified form to log second-order, as we confirm in the next lemma. We will use $f \in \{p, c\}$ to denote firm conduct, with p denoting perfect competition and c denoting Cournot.

Lemma 2. *For firms competing $f \in \{p, c\}$, the regional production function is $Y_i^f(\ell, m) = z_i m^\eta \ell^{1-\eta} \Phi^f(m)$, where $z_i \equiv \mathbb{E}[z_{isn}^{1/\eta}]^\eta$. Furthermore, $\Phi^c(m) \leq \Phi^p(m)$.*

Generically, $\Phi^c(m) < \Phi^p(m)$ because when firms are imperfectly competitive, misallocation due to varying wage markdowns decreases the productivity of the market relative to the efficient benchmark.

In practice, including imperfect competition increases the agglomeration benefits. The reason is that with imperfect competition, larger locations have two advantages: it is easier for a theoretical planner to move workers to their most productive use, and the equilibrium gets closer to the planner's solution. In the limit with an infinite number of firms, firms do not vary their wage markdown so that workers move to their efficient use. By contrast, in small locations, firms do not raise their wage as much as a planner would like in response to a good productivity shock, so firms expand too little. Thus, Proposition 2 only becomes stronger.

In contrast, the optimal policy changes with imperfect competition. The first-best policy would exactly replicate the perfect competition benchmark with firm- and state-dependent wage subsidies that exactly undo the wage markdown. Throughout this paper, we will assume that this policy is infeasible. Instead, we assume that the planner can give place-based transfers to workers and place-based entry subsidies to firms. Therefore, conditional on the firms and workers in a location, the planner takes as given how competition continues. In math, this implies that the planner will take as given the functions $\Phi^f(m)$.

Then, to explore optimal policy, we need to know what workers and firms are paid in equilibrium, and how that differs from their marginal contribution to production. With perfect competition, workers are paid their marginal product, so the optimal policy features firm entry subsidies, as firms cannot get paid their marginal product. If firms exploit their market power, there are wage markdowns, and workers are not paid their marginal product. Instead, their earnings are distorted downwards. The following lemma summarizes how much.

Lemma 3. *For firms competing Cournot, there exists a function $\psi^C(m) \leq 0$ such that total wage compensation in location i can be written,*

$$w_i \ell_i = (1-\eta) z_i (m_i)^\eta (\ell_i)^{1-\eta} \left(\Phi^c(m_i) + \Psi^c(m_i)\right).$$

We derive $\Psi^c(m)$ in the online supplement and use it when considering our mechanism quantitatively in Section 6. Comparing the marginal product of labor and firms to their payments then implies our next result.

Proposition 4. *The optimal policy features a subsidy on earnings in location i proportional to average wages of τ_i^w and a subsidy (or tax) on firm entry proportional to profits τ_i^π that satisfies,*

$$1+\tau_i^w = \frac{1}{1+\frac{\Psi^f(m_i)}{\Phi^f(m_i)}}; \qquad 1+\tau_i^\pi = \frac{1+\frac{1}{\eta}\frac{\partial \log \Phi^f(m_i)}{\partial \log m}}{1-\frac{1-\eta}{\eta}\frac{\Psi^f(m_i)}{\Phi^f(m_i)}}. \tag{13}$$

Without any more information, we know that the optimal policy features a subsidy on wages, $\tau_i^w > 0$. Those subsidies will undo the average markdown workers face in location i. The sign of τ_i^π, on the other hand, is ambiguous. As we showed in Proposition 3, granularity implies that there is under-entry. This force is represented in the numerator of equation (13) as $\frac{\partial \log \Phi^f(m)}{\partial \log m} > 0$ and pushes for the optimal subsidy to be positive. However, with imperfect competition, there is another force in the denominator. As discussed by Mankiw and Whinston (1986), when firms are imperfectly competitive, there is a tendency for too many firms to enter. In our model, this arises because firms markdown wages, biasing profits above their competitive levels, increasing the incentive to enter. The quantitative importance of this force is summarized by the size of the wage markdown, $\frac{\Psi^f(m)}{\Phi^f(m)}$. We return to how these forces interact quantitatively in Section 6.

4 Estimation of Granular Driven Agglomeration

In Section 3, we presented a quantitative model of economic geography with granular firms. In this section, we estimate the model in order to quantify how important the granular mechanism is for agglomeration in Section 6. We focus on Japan, where high-quality firm-level data across years enables us to examine the extent to which firms experience idiosyncratic shocks.

4.1 Data

Data Sources. Our primary data source is the Census of Manufacture (CoM) in Japan, conducted annually by the Ministry of Economy, Trade, and Industry (METI). The census covers all manufacturing establishments in years ending with 0, 3, 5, or 8, and those with at least four employees in other years. The CoM was not conducted in 2012 and 2016;

instead, the Economic Census for Business Activity (ECBA) by METI and the Ministry of Internal Affairs and Communications provides data for 2011 and 2015.[7] We use the ECBA to substitute for CoM in 2011 and 2015.

We use data from 2002–2019, as product classification changed discontinuously in 2002. For establishment-level analysis, we restrict the sample to establishments with at least 10 employees and those surveyed in at least 10 of the 18 years (2002–2019).

These data have two advantages. First, we observe panels of all the establishments with at least 4 employees. This feature allows us to compute a variety of volatility measures at the establishment level as well as commuting zone and sector-level variables.[8] Second, we observe yearly shipment values by detailed product categories for each establishment. This enables us to construct establishment-level exposure to product-level demand shocks at an annual frequency.

Mapping from Model to Data. We interpret region i as a commuting zone and denote by $\mathcal{I}$ the set of 225 commuting zones in Japan.[9] We map sectors $s \in \mathcal{S}$ onto 148 manufacturing industries at the 3-digit level.[10] Each establishment in the data is treated as an independent firm, as multi-establishment firms are not modeled. We refer to each establishment as a "firm" throughout.

4.2 Labor Supply

Short Run Labor Supply Elasticity across Firms. We first estimate a short-run labor supply elasticity across firms, κ. As we show in the online supplement, the labor supply equation in period 2 is given by

$$\log \ell_{isn}(\omega) - \log \ell_{isn} = \kappa \log w_{isn}(\omega) - (\kappa - \nu) \log w_{is}(\omega) - \nu \log w_i(\omega) + \tilde{\epsilon}^w_{isn}(\omega). \quad (14)$$

[7] The ECBA covers all establishments, including non-manufacturing ones, but we focus on manufacturing to maintain consistency with the CoM.

[8] One further advantage, compared to the US LBD data, is that we can separately identify single establishments within each of the 47 prefectures.

[9] To construct time-consistent commuting zones, we follow Kondo (2023) to convert municipalities into time-consistent groups. Japan has 1,724 municipalities as of June 2023, including 6 in the Northern Territories, which we drop because the CoM does not cover them. We then use the converter in Adachi et al. (2020) to map these groups into commuting zones and retain those with at least 10 manufacturing establishments in 2019.

[10] We use a RIETI crosswalk to convert all categories into 2011 codes. In theory, sectors form a continuum, but we use 148 finite sectors because workers move freely within these broad groups, which best represent labor markets. This abstraction omits sectoral granularity, underestimating individual firm influence.

We interpret the short run of period 2 as one year. Therefore, to estimate κ, we take a one-year change of these variables and get the following equation:

$$\Delta \log \ell_{isnt} = \kappa \Delta \log w_{isnt} + \gamma_{ist} + \tilde{\epsilon}^{w}_{isnt}, \tag{15}$$

where γ_{ist} is a market-time fixed effect.

The key threat to the identification of κ is that changes in wages of firm n in location i and sector s might be correlated with the changes in workers' taste for firm n. To address this concern, we instrument $\Delta \log w_{isnt}$ with a shift-share IV, Δd_{isnt}, constructed as follows.

$$\Delta d_{isnt} \equiv \sum_{p} \overline{s^{p}_{isn}} \cdot \Delta d^{\text{national}}_{pt}, \tag{16}$$

where $\overline{s^{p}_{isn}}$ is a time-invariant share of product p in total shipment from firm n with $\sum_p \overline{s^{p}_{isn}} = 1$ for all firms n. We take the average over our sample periods between 2002 and 2019. $\Delta d^{\text{national}}_{pt}$ is a one-year log change in the shipment of product p at a national level. We interpret this shift in demand as a shock to the revenue productivity of the firm, as that shift increases the effective price of the firm's output. As we have normalized the price of each good to one, this shift in price is loaded on the effective productivity $a_{isn}(\omega)$ in the model.

The results are in Columns (1) and (2) of Table 1. Column (2) weighs samples by the median of market-level total employment. We take the estimate in Column (2), $\hat{\kappa} = 2.48$, as our baseline estimate. This is in line with many of the estimates in the literature. Lamadon et al. (2022) estimate this elasticity to be 6.52 in the United States, while Berger et al. (2022) find 10.85. In Brazil, Felix (2024) determines $\hat{\kappa} = 1.02$

Short Run Labor Supply Elasticity across Markets. Next, we estimate a short-run labor supply elasticity across markets, ν. As we derive in the Supplementary Material, the labor supply to sector s in period 2 is written,

$$\log \ell_{is}(\omega) - \log \ell_{is} = \nu \log w_{is}(\omega) - \nu \log w_{i}(\omega) + \varepsilon^{w}_{is}(\omega), \tag{17}$$

where

$$w_{is}(\omega) \equiv \left[\sum_{n \in \mathcal{N}_{is}} \frac{\ell_{isn}}{\ell_{is}} w_{isn}(\omega)^{1+\kappa} \right]^{\frac{1}{1+\kappa}}.$$

We then estimate ν following a very similar procedure to how we estimate κ. Taking a

Table 1: Estimation of Short-run Labor Supply Elasticity

Dep. Var.: Log Employment Growth				
	Across Firms		Across Markets	
	(1)	(2)	(3)	(4)
Log Wage Growth	2.34	2.48	1.78	1.46
	(0.50)	(0.73)	(0.18)	(0.07)
Observations	1,519,077	1,519,077	186,148	186,148
1st Stage F-Stat.	27.85	14.26	110.11	482.06
Covariates	✓	✓	✓	✓
Weighted		✓		✓

Note: This table shows the estimates of short-run labor supply elasticity. Columns (1) and (2) report estimates of the elasticity across firms within markets, κ, following (15). Columns (3) and (4) report estimates of elasticities across markets, ν, following (18). All columns include lagged log employment growth as a covariate. Columns (2) and (4) weigh samples by the median of market-level total employment.

one-year difference of these variables, we get the following equation:

$$\Delta \log \ell_{ist} = \nu \Delta \log w_{ist} + \gamma_{it} + \gamma_{st} + \varepsilon^{w}_{ist}, \tag{18}$$

where γ_{it} is a location-time fixed effect, and γ_{st} is a sector-time fixed effect. To get theory-consistent measures of the market-level labor and market-level wages, we use the estimate $\hat{\kappa} = 2.48$, to construct the variables

$$\ell_{ist} = \left[\sum_{n \in \mathcal{N}_{is}} \left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}} \right)^{-\frac{1}{\hat{\kappa}}} \ell_{isnt}^{\frac{1+\hat{\kappa}}{\hat{\kappa}}} \right]^{\frac{\hat{\kappa}}{1+\hat{\kappa}}}, \quad w_{ist} = \left[\sum_{n \in \mathcal{N}_{is}} \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}} w_{isnt}^{1+\hat{\kappa}} \right]^{\frac{1}{1+\hat{\kappa}}}.$$

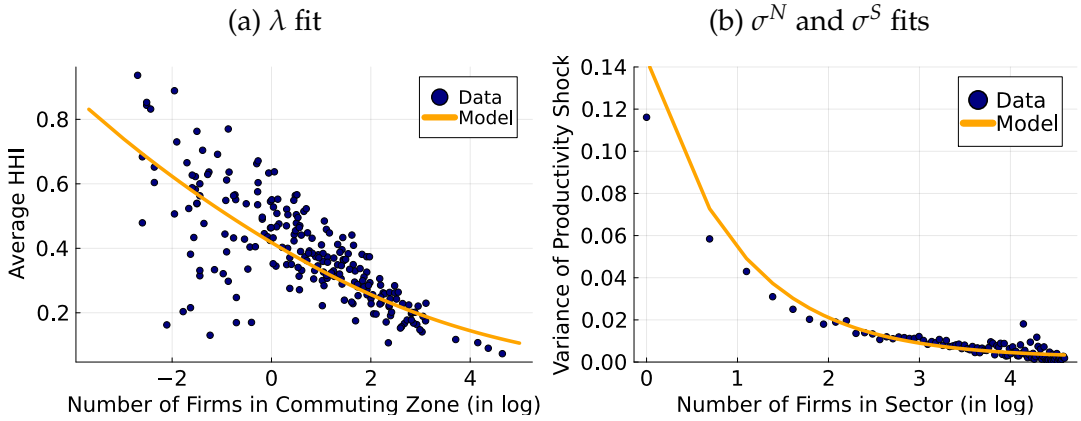
We then use equation (18) to estimate ν. Just as in the case of κ, we instrument for wages to avoid endogeneity issues. We construct market-wide versions of the same instrument used to estimate κ,

$$\Delta d_{ist} \equiv \sum_{n} \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}} \cdot \Delta d_{isnt}, \tag{19}$$

where Δd_{isnt} is the shift-share IV constructed in (16), $\overline{\ell}_{isn}$ is the median of employment of firm n in location i and sector s, and $\overline{\ell}_{is}$ is the median of total employment in location i and sector s. In essence, Δd_{ist} is a weighted average of firm-level demand shock at a market level.

The results are in Columns (3) and (4) of Table 1. Column (4) weighs samples by the median of total market-level employment. We take the estimate in Column (4), $\hat{\nu} = 1.46$,

Figure 2: Supply Side - Model Fit



Note: The figures show the fits of the model to the moments we target. The left panel shows the average HHI across sectors and the mass of firms in each commuting zone in log. The right panel shows the variance of the average productivity shock to sectors with N firms defined by (21) and the number of firms per sector in log. In both of the figures, the blue dots show the data for each commuting zone, and the orange line shows the model-implied relationship.

as our baseline estimate. In the US setting, Lamadon et al. (2022) and Berger et al. (2022) estimate this elasticity to be 4.57 and 0.42, respectively. Felix (2024) finds $\nu = 0.80$ in Brazil.

4.3 Labor Demand

Ex-Ante Productivity Distribution. We could estimate the shape parameter of the ex-ante productivity distribution λ in a couple of different ways. However, Equation (11) suggests that the key moment determining the strength of the externality is the average HHI across sectors within a location. Therefore, we estimate λ to match the average HHI across sectors for each location i with a quadratic loss function. The fit is depicted in Figure 2.

Ex-Post Productivity Distribution. To estimate the distribution of ex-post productivity shocks, we look at the distribution of the average shock to a sectoral labor market as a function of the number of firms in that market. Recall that production is $y_{isn}(\omega) = z_{isn}a_{isn}(\omega)\ell_{isn}(\omega)^{1-\eta}$. As we interpret the time scale of period 2 as one year, we can take first differences to get

$$\Delta \log a_{isnt} = \Delta \log y_{isnt} - (1-\eta)\Delta \log \ell_{isnt}. \tag{20}$$

Table 2: Summary of Estimated and Calibrated Parameters

Description	Parameter	Value	Source
A. Labor Supply			
Short run labor elasticity across firms	κ	2.48	Estimated (CoM)
Short run labor elasticity across markets	ν	1.46	Estimated (CoM)
B. Labor Demand			
Returns to scale	η	0.13	Profit Share (Data, FSSC)
Ex-ante firm prod. tail	λ	10.5	Estimated SMM (CoM)
Variance of sector shocks	σ_S^2	2.0×10^{-3}	Estimated GMM (CoM)
Variance of idiosyncratic shocks	σ_N^2	0.14	Estimated GMM (CoM)
C. Economic Geography Parameters			
Migration elasticity	θ	3	Redding (2016)
Avg. Productivity, Amenity, Entry Cost	$z_i, \overline{u}_i, \psi_i$		Exact hat algebra (2019)

Note: This table summarizes where the key parameters of the model come from. See the discussion in Section 4 for a more in-depth description.

Using our estimate of η, we can estimate the productivity shock to firm n from equation (20).[11] We then create a measure of the variance of the average productivity shock to sectors with N firms, $\sigma^2(N)$, defined as

$$\sigma^2(N) = \text{Var}\left(\frac{\sum_{n \in \mathcal{N}_{is}} \Delta \log a_{isnt}}{N} \middle| N_{is} = N \right). \tag{21}$$

We then estimate σ_S^2 and σ_N^2 to match $\sigma^2(N)$ for $N = \{1, \ldots, 99\}$. The fit is depicted in Figure 2.

4.4 Economic Geography Parameters

We take the migration elasticity $\theta = 3$ from the literature (Redding, 2016). Similar migration elasticity estimates have been found in the United States (Hornbeck and Moretti, 2024) and Indonesia (Bryan and Morten, 2019). That leaves parameters summarizing the average firm productivity z_i, amenity $\overline{u}_i$, and fixed cost of opening a firm ψ_i in each location i. We calibrate these parameters to exactly match 2019 data on the population, average wages, and number of firms in each location.

[11] We take η from the average firm profit share in 2019 using Financial Statements Statistics of Corporations (FSSC).

5 Validation of the Mechanism

We presented a quantitative model of economic geography with granular firms in Section 3 and estimated it in Section 4. Before we show the quantitative importance of the mechanism, we present some reduced-form evidence of the mechanism and demonstrate that it is consistent with the model we have just quantified.

Our reduced form evidence focuses on two key predictions intimately tied to the mechanism. First, we provide evidence that firms are subject to idiosyncratic shocks that "average out" in larger markets by looking at the variance of the wage bill in labor markets of different sizes. Second, we provide evidence that firms in larger markets expand their employment more than firms in small markets in response to positive productivity shocks.

5.1 Shocks Average Out in Large Markets

Shocks to firm revenue are passed through to wages paid to workers. Thus, we can use the total payments to workers in a sectoral labor market as a measure of the shocks to the market.

If our theory is correct, firms are subject to different shocks, so the variance of the log wage bill at the sectoral market should be lower for markets with more firms. In math, workers are paid their marginal product of labor so that $w_{isn}(\omega) = (1-\eta) z_{isn} a_{isn}(\omega) \ell_{isn}(\omega)^{-\eta}$. And, the short-run labor supply is given by (14). Some straightforward algebra combining these equations then implies our first empirical prediction.

Proposition 5. *If $\sigma_N^2 > 0$, the variance of the log wage bill of a sectoral labor market is decreasing in the number of firms. In particular,*

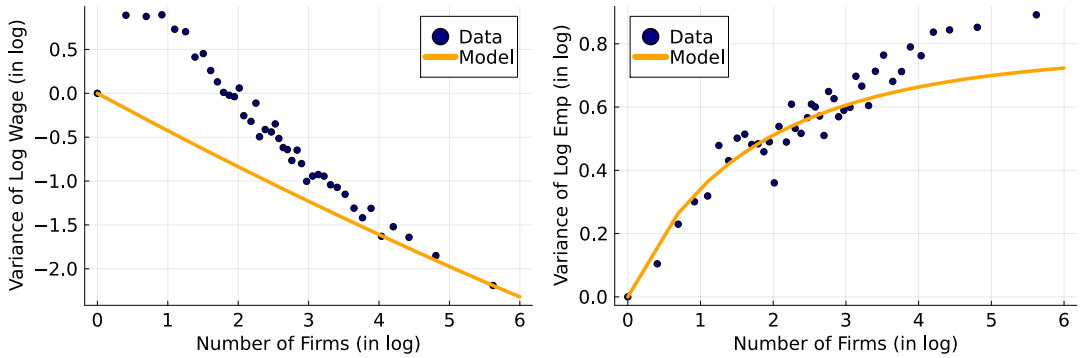
$$\text{Var}\left(\log\left(\sum_{n \in \mathcal{N}_{is}} w_{isn}(\omega)\ell_{isn}(\omega)\right)\right) = \frac{1+\nu}{1+\eta\nu}\left(\sigma_S^2 + HHI_{is}\sigma_N^2\right).$$

We present a binned scatter plot of the log variance of the log wage bill by sectoral local labor market against the log number of firms in Figure 3(a). We include the variance of the log wage bill implied by the model as well. We normalize the scale so the value is 0 with 1 firm.

Just as predicted by the model, the variance of the log wage bill is decreasing in the number of firms. The data suggests that the rate of that decrease is slightly faster than that suggested by the model. This is likely because our model assumes that every firm has the same variance of log productivity shocks, whereas in practice, larger firms have

Figure 3: Model Validation

(a) Variance of Log Wage Bill (b) Variance of Log Employment



Note: The figures show the fits of the model to the moments we do not target. The left panel shows the variance of the log wage bill for different sectoral local labor markets. The right panel shows the variance of firm-level employment aggregated to the sectoral local labor market level. Both statistics are plotted against the number of firms across local labor markets and in log units. In both of the figures, the blue dots show the binned data of each local labor market, and the orange line shows the model-implied relationship. The y-axis for both data and model is normalized so that the values for $x = 0$, which is at the local labor market with only one firm, are zero.

less volatile productivity. Therefore, in the real world, the market becomes less exposed to individual firm shocks at a faster rate than that suggested by the model.[12]

5.2 Firms Expand More in Larger Markets

We next turn to validating the model's key prediction for the mechanism: firms expand more in response to productivity shocks in larger markets. We provide two separate pieces of evidence for this prediction.

Cross-Sectional Evidence. We first give cross-sectional evidence for the prediction. In particular, we look at the variance of the log employment of individual firms in different-sized markets. If our theory is correct, and firms in larger markets are subject to similar shocks as those in smaller markets, then firms in larger markets should have a higher variance of log employment as they respond more to the same shocks.

In math, we can find how employment at an individual firm n responds to its own

[12]Estimating λ, σ_N^2, and σ_S^2 to match this bin scatter rather than how we estimate them in Section 4 does not significantly change any results.

productivity shock and the shocks to every other firm to first order,

$$\Delta \log \ell_{isn}(\omega) = \frac{1}{\eta + \frac{1}{\kappa}} \left(\Delta \log a_{isn}(\omega) - \frac{\frac{1}{\nu} - \frac{1}{\kappa}}{\eta + \frac{1}{\nu}} \sum_{n' \in \mathcal{N}_{is}} \frac{\ell_{isn'}}{\ell_{is}} \Delta \log a_{isn'} \right). \tag{22}$$

Then, with some algebra, we can solve for the average variance of log employment, weighted by the average labor share. This implies our second empirical prediction.

Proposition 6. *If workers are more mobile across firms within a sector than across sectors, i.e. $\kappa > \nu$, and $\sigma_N^2 > 0$, then the weighted average variance of log employment increases with the number of firms in a sectoral labor market. In particular,*

$$\begin{aligned} \sum_{n \in \mathcal{N}_{is}} \frac{\ell_{isn}}{\ell_{is}} \operatorname{Var}\left(\log \ell_{isn}(\omega)\right) = & \left(\frac{1}{\eta + \frac{1}{\nu}} \right)^2 \sigma_S^2 + \left(\frac{1}{\eta + \frac{1}{\kappa}} \right)^2 \sigma_N^2 \\ & - \left(\frac{1}{\nu} - \frac{1}{\kappa} \right) \frac{\eta + \frac{1}{\kappa} + \eta + \frac{1}{\nu}}{\left(\eta + \frac{1}{\kappa}\right)^2 \left(\eta + \frac{1}{\nu}\right)^2} HHI_{is} \sigma_N^2. \end{aligned}$$

We plot the bin scatter for the log weighted variance of log employment for each firm against the log number of firms in Figure 3(b). We include the model-implied value as well.

Just as predicted by the model, the variance of log employment for individual firms is increasing in the number of firms. Furthermore, although we do not target it, our model matches the rate of that increase well. We even see evidence that the variance of log employment stops increasing as much for sectoral labor markets with a large number of firms as would be suggested by the model.

One thing we do not hit is the level of the variance of log employment. The figure normalizes the value to be 0 when the number of firms is one, but our model implies a higher variance in log employment than what we see in the data. This could be for a variety of reasons. One possible reason is that the employment data measures employment in June, while firms can adjust their employment throughout the year in response to shocks.

Employment-Response to Demand Shocks. We conclude with a more direct test of the mechanism by examining whether firms that hire only a small share of the sectoral labor market expand employment more in response to a revenue productivity shock.

Table 3: Responses of Employment to Product-level Shocks

	Dep. Var.: Log employment growth			
	(1)	(2)	(3)	(4)
Shock	0.059	0.060	0.048	0.048
	(0.002)	(0.002)	(0.002)	(0.002)
Shock x Payroll Share	-0.026	-0.028	-0.021	-0.023
	(0.005)	(0.005)	(0.005)	(0.005)
Payroll Share	0.006	0.006	-0.081	-0.070
	(0.001)	(0.001)	(0.003)	(0.003)
Lag. Log Emp. Growth		-0.205		-0.261
		(0.002)		(0.002)
Implied Ratio	-0.435	-0.463	-0.428	-0.477
95% CI	[-0.589, -0.281]	[-0.606, -0.319]	[-0.615, -0.240]	[-0.651, -0.303]
Observations	1,747,629	1,516,621	1,740,782	1,511,376
Year FE	✓	✓	✓	✓
Firm FE			✓	✓

Note: This table shows the estimates of the impact of firm-level demand shock on employment, following (23). All columns include year fixed effects. Columns (2) and (4) include lagged log employment growth as a covariate. Columns (3) and (4) include firm fixed effects. Standard errors are clustered at the firm level. The implied ratio is the ratio of the coefficient on the interaction term between the shock and the lagged payroll share within local labor markets, β_3, to the coefficient on the shock, β_1. The 95% confidence interval of that implied ratio is also reported.

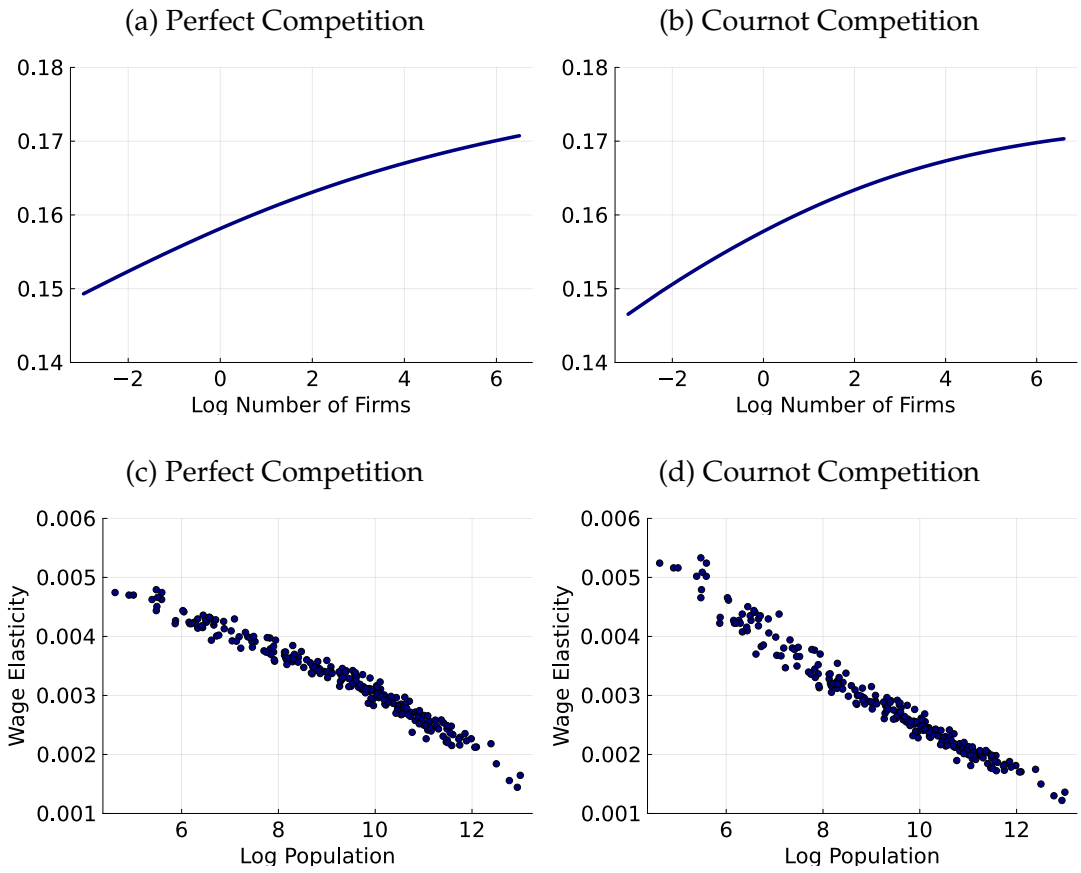
To test this prediction, we run the regression suggested by equation (22),

$$\begin{aligned}\Delta \log \ell_{i,s,n,t} =& \beta_1 \Delta d_{i,s,n,t} + \beta_2 \left(\frac{\ell_{i,s,n,t}}{\ell_{i,s,t}} \right) + \beta_3 \left(\Delta d_{i,s,n,t} \times \frac{\ell_{i,s,n,t}}{\ell_{i,s,t}} \right) \\ &+ \beta_4 \Delta \log \ell_{i,s,n,t-1} + \zeta_t + \zeta_n + \varepsilon_{i,s,n,t} \end{aligned} \tag{23}$$

where $\Delta d_{i,s,n,t}$ is the firm-level shock constructed as in (16). The shock implied by the theory is in terms of revenue productivity, while the shock $\Delta d_{i,s,n,t}$ is in terms of total sales of products produced by the firm. Therefore, without an elasticity relating the price of the products produced by firm n to a shift in demand for the products, there are no predictions on the size of β_1 or β_3. However, the theory predicts that the ratio is given by $\frac{\beta_1}{\beta_3} = -\frac{\frac{1}{\nu}-\frac{1}{\kappa}}{\eta+\frac{1}{\nu}} \approx -0.35$.

We report the results in Table 3. The estimated ratios range from -0.43 to -0.48 across specifications, all of which are close to, and not statistically different from, the theoretical prediction of -0.35. The model-implied ratio lies well within the 95 percent confidence intervals in every specification. These results indicate that the theoretical mechanism is both qualitatively and quantitatively consistent with the data.

Figure 4: $\log \Phi^f(m)$ and Wage Elasticities



Note: The figures shows $\log \Phi^f(m)$ against the log number of firms across different numbers of firms and wage elasticity across commuting zones in Japan. The left panel shows the case of perfect competition, and the right panel shows the case of Cournot competition.

6 Quantification of Granular Driven Agglomeration

In this section, we present the results demonstrating the quantitative contribution of granularity to agglomeration. For all of the results, we present the results assuming that firms are perfectly and Cournot competitive.

6.1 Strength of Agglomeration Benefits

As we know from Proposition 1, the strength of the granular origins of agglomeration is determined by the shape of $\Phi^f(m)$. Thus, we plot $\log \Phi^f(m)$ in Figure 4 (a) and (b).

Consistent with Proposition 2, the curve is upward sloping, implying that larger markets are more productive than smaller markets. Also consistent with Proposition 2, the

curve flattens out for larger m, suggesting that the marginal benefit of increasing the number of firms in a location that is already larger is small, while locations with few firms could see large productivity benefits from increasing the number of firms. The implied $\Phi^f(m)$ looks very similar whether firms are competing perfectly or according to Cournot. The Cournot curve is slightly lower than the curve with perfect competition, as there is misallocation when firms compete imperfectly. This gap is most pronounced for locations with a small number of firms, as labor market power shapes the employment response most. Thus, locations with fewer than 0.13 (-2 in log) firms per sector have a $\log \Phi^f$ close to 0.145 if firms behave according to Cournot, while the perfect competition case does not reach such low values. For locations with more than 400 (6 in log) firms per sector, the gap between the two curves mostly disappears, as firms behave almost competitively.

In total, our estimated $\Phi^p(m)$ implies that the largest commuting zone in Japan (Tokyo, with an average of 105.8 firms per sector) is 2.1% more productive than the smallest commuting zone (with an average of 0.07 ($\approx 10/148$) firms per sector). Cournot competition implies the gap is 2.3%.

In Figure 4 (c) and (d), we plot the implied elasticity of wages to population for each of the commuting zones in Japan against the population. In the perfect competition case, we find that the wage elasticity gets as high as 0.0048 for the smallest commuting zone. The implied elasticity is much smaller for the large locations: Tokyo has an implied elasticity of 0.0014. If firms are competing à la Cournot, the elasticities range from 0.0012 to 0.0053.

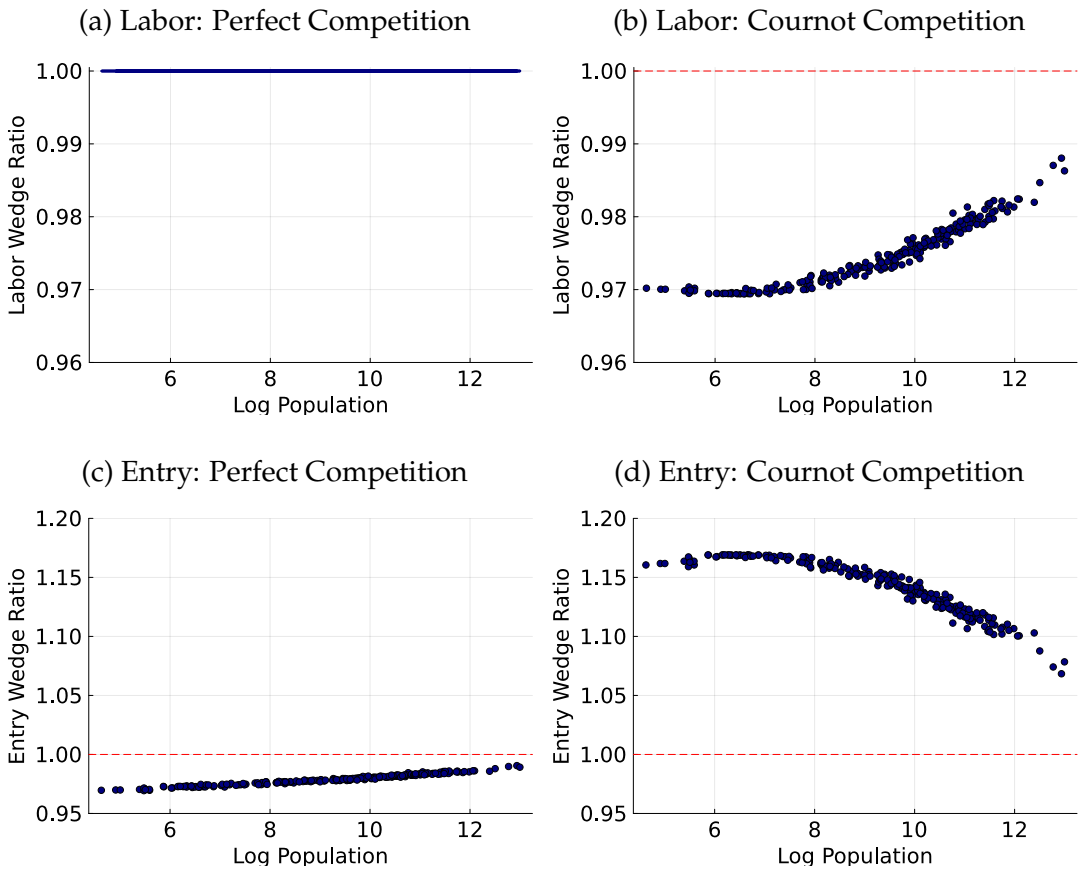
For context, Combes et al. (2011) find that most causal estimates of the urban wage premium find an elasticity between 0.02 and 0.05 when pooling across locations of all sizes. Thus, the granular mechanism could explain as much as 25% of the wage benefits of living in a large location. However, this mechanism cannot explain most of the urban wage premium.

6.2 Factor Wedges

We next turn to quantify what our model implies for the degree of wage markdowns and under-entry.

Labor Wedges. We start by plotting the labor wedge ratio in Figure 5. This is the ratio of average wages to the marginal product of another worker. By assumption, the perfect competition case in Figure 5(a) is 1 for all locations. Under Cournot competition, firms' market power increases following good shocks and decreases following bad shocks, which enables them to markdown wages even though workers are freely mobile

Figure 5: Factor Wedge: Labor and Firm Entry



Note: The figures show the labor wedge ratio and firm entry wedge across commuting zones in Japan. The left panel shows the case of perfect competition, and it is one by construction. The right panel shows the case of Cournot competition, and each dot represents a commuting zone in Japan.

in the long run. This short-run market power implies that workers face an average markdown of around 3% in small and medium-sized cities. Only in larger locations like Tokyo, the wage markdown becomes as small as 1%. Perhaps surprisingly, the smallest commuting zones do not have the largest markdowns. Instead, there is a local maximum around a log population of 6. That is because firms can only exploit their market power if their share of the sectoral local labor market varies. If there is only one firm in the sector, the firm only competes against other firms in the long run, when workers are freely mobile. If there are two firms in the sector, then the firm will see large swings in its own market power and can vary its markdown to take advantage of that.

Firm Entry Wedges. We next turn to the implications for firm entry. As stated in Proposition 3, granularity implies that too few firms enter in equilibrium. In Figure 5, we plot

the ratio of firm profits to the marginal product of another firm for the location's output. Consistent with the proposition, Figure 5(c) shows that every firm is paid less than its marginal product if firms are perfectly competitive. In the smallest locations, firms capture 96.9% of their contribution to production, while in the largest location they capture a full 98.8%.

As we mentioned in Section 3, this prediction of under-entry is not robust to the conduct assumptions. By exploiting their market power, firms distort wages downward and profits up. This can lead profits to increase above the marginal product of another firm, taking as given the conduct of the firms, conditional on firm entry. In Figure 5(d), we plot the ratio of firm profits to the marginal product of firms. We find that simply by varying their wage markdown, leading to an average wage markdown of 3%, firms can, in fact, increase profits so much that too many firms enter in equilibrium.

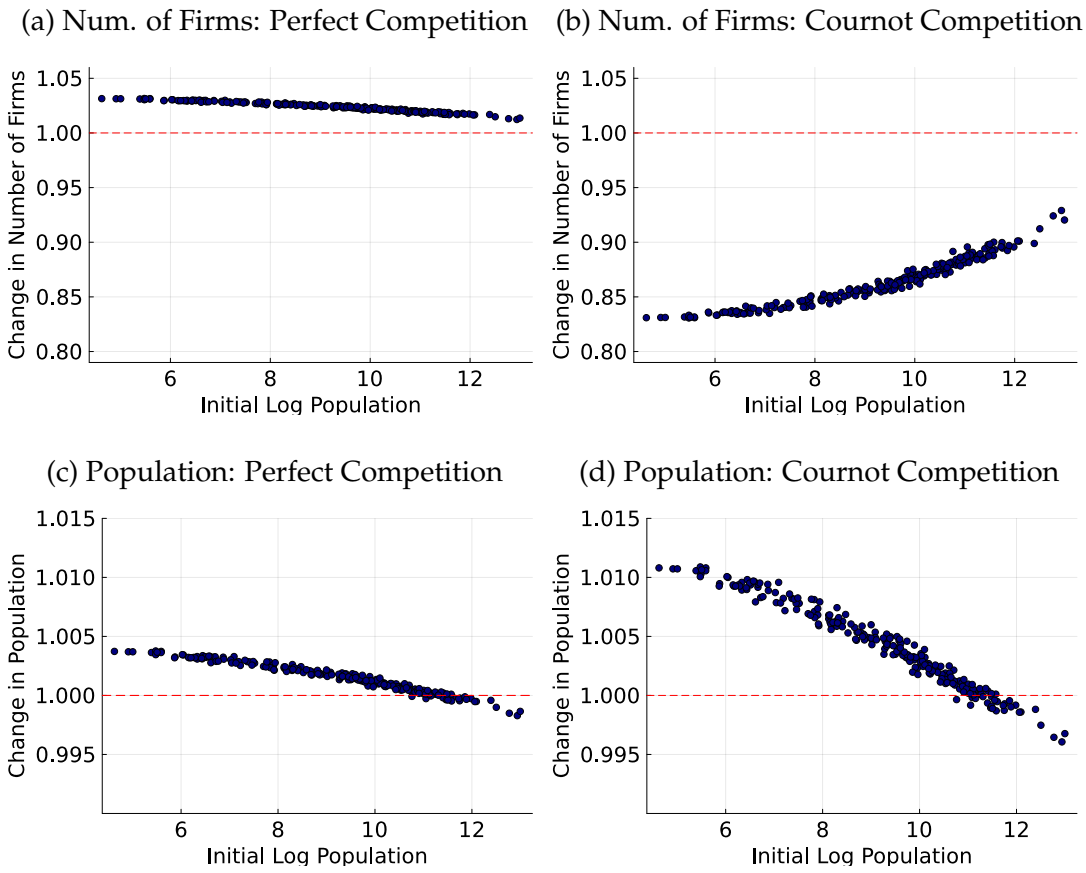
6.3 Implications of Optimal Policy

In the previous subsection, we showed the ratio of factor payments to the marginal product of that factor under different firm conduct assumptions. In this subsection, we consider how the equilibrium levels of population and mass of firms would change if a national government put in place the optimal place-based transfers and firm entry subsidies suggested by Proposition 4, paid for with a proportional increase in income taxes.

We plot how the optimal subsidies change the equilibrium number of firms and population in Figure 6. In Figure 6(a), we show that the equilibrium number of firms increases with the optimal subsidies if firms are competitive. The smallest locations see a 3.2% increase in the number of firms, while in Tokyo, the number of firms only increases by 1.2%. The effect of the optimal policy on the number of firms is very different when firms compete imperfectly. The over-entry implications of imperfect competition overrule the under-entry implications of granularity. Thus, Figure 6(b) shows that the optimal number of firms is 83% of the observed number of firms in the smallest locations, while it is 93% in Tokyo.

Figure 6 (c) and (d) plot the effect of the optimal policy on the population. If firms compete perfectly, the optimal policy only indirectly affects where people would like to live because the optimal policy will increase the number of firms, especially in the smallest locations. Therefore, the smallest locations see a 0.3% increase in population while Tokyo sees a very small decline in population in Figure 6(c). The effects of optimal policy on population are more pronounced when firms are competing à la Cournot, as shown in

Figure 6: Changes in Number of Firms and Population



Note: The figures show the changes in the number of firms and population in response to the optimal subsidies. Each dot represents a commuting zone in Japan. The left panel shows the case of perfect competition, and the right panel shows the case of Cournot competition.

Figure 6(d). Here, the optimal policy features a large tax on firm entry in the smallest locations, but also a subsidy on wages. And because the aggregate externality is still largest for the smallest locations, the net effect is that the population should increase in the small commuting zones. In fact, the population should increase even more than under perfect competition because the externality is larger with imperfect competition. Thus, the optimal population is more than 1% higher than the observed population in the smallest locations.

7 Concluding Remarks

The world is granular, and local sectoral labor markets are especially so. That granularity has important implications for the geography of economic activity and optimal

policy. Larger locations are more productive because the firms in those areas can expand in response to productivity shocks. A simple model of this labor market pooling mechanism suggests that it could explain as much as 25% of the wage premium of large cities. Thus, while most of the urban wage premium must come from other mechanisms, granularity has important explanatory power.

In formalizing the mechanism, we were forced to ignore certain important features of real firms and labor markets. Firms often look for workers with very particular and rare skills. Those workers are often complementary to other specialized workers in the firm. Including that feature could imply that granular labor market pooling is even more important than what we find in this paper, as not being able to find people with those specialized skills could be more costly. We also abstract from wage rigidity, unemployment, and other inefficient labor market adjustments to shocks. Including these features could also increase the social cost of being over-exposed to individual firm-level shocks. We think our model is tractable enough to incorporate many of these real-world features.

Furthermore, we ignore the fact that workers are often not only exposed to idiosyncratic firm shocks but also sector-wide shocks. As a result, a planner might want to diversify the local labor market by attracting firms from a variety of sectors rather than only increasing the number of firms. There is a lot of work to be done analyzing what this means for economic activity and optimal place-based policies.

A Theory

This Appendix presents the proofs of the theoretical results in Section 2 and states the Technical Lemmas used in those proofs. The proofs of the technical Lemmas and the theoretical results in the more general model of Section 3 are presented in the Supplementary Material. Throughout the appendix, we will use $\overline{x}$ to denote the value of x with no ex-post shocks, and we will use $\hat{x}$ to denote log deviations from that value.

A.1 Statement of Technical Lemmas

Lemma 4. *Expected sectoral HHI, weighted by average productivity shock, converges to 0 as the number of firms goes to infinity. More precisely,* $\psi_N \to 0$ *as* $N \to \infty$ *where* $\psi_N \equiv \mathbb{E}\left[\frac{\sum_{n\in\mathcal{N}} z_{isn}^{1/\eta}}{N z_i^{1/\eta}} \frac{\sum_{n\in\mathcal{N}} (z_{isn}^{1/\eta})^2}{\left(\sum_{n\in\mathcal{N}} z_{isn}^{1/\eta}\right)^2} \middle| N\right]$.

Lemma 5. *Average sectoral HHI has the following properties:*

(i) $\frac{\partial}{\partial \log m}\left[\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)^2 ds\right] < 0$; *and*

(ii) $\frac{\partial}{\partial \log m}\left[\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)^2 ds\right] \to 0$ *as* $m \to \infty$.

A.2 Proofs of Main Propositions

Proof of Proposition 1. This result follows almost immediately from Lemma 1. Implicitly differentiating the free entry condition (10) implies that

$$\frac{d\log m_i}{d\log \ell_i} = \frac{1}{1-\frac{1}{1-\eta}\frac{\partial \log \Phi(m_i)}{\partial \log m_i}}.$$

In any stable equilibrium, the denominator must be positive. Therefore, differentiating the expression for wages gives

$$\begin{aligned}\frac{d\log w_i}{d\log \ell_i} &= \left(\eta + \frac{\partial \log \Phi(m_i)}{\partial \log m_i}\right)\frac{d\log m_i}{d\log \ell_i} - \eta \\ &= \frac{\frac{1}{1-\eta}\frac{\partial \log \Phi(m_i)}{\partial \log m_i}}{1-\frac{1}{1-\eta}\frac{\partial \log \Phi(m_i)}{\partial \log m_i}},\end{aligned}$$

which is greater than zero if and only if $\frac{\partial \log \Phi(m_i)}{\partial \log m_i} > 0$. □

Proof of Proposition 2. By Proposition 1, the average wage is increasing in population if and only if $\frac{\partial \log \Phi(m)}{\partial \log m} > 0$. Thus, we need to find the covariance.

Taking a log first-order approximation to the labor market clearing condition, (2) implies that

$$\sum_{n\in\mathcal{N}_{is}} \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\hat{\ell}_{isn}(\omega) = 0.$$

Firm profit maximization implies that wages are equal to the marginal product of labor, $z_{isn}a_{isn}(\omega)\ell_{isn}(\omega)^{-\eta} = w_{is}(\omega)$. Taking a log first-order approximation to this implies that

$$\hat{a}_{isn}(\omega) - \eta\hat{\ell}_{isn}(\omega) = \hat{w}_{is}(\omega).$$

Solving this system of equations implies that $\hat{w}_{is}(\omega) = \sum_{n\in\mathcal{N}_{is}} \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\hat{a}_{isn}(\omega)$, and

$$\hat{\ell}_{isn}(\omega) = \frac{1}{\eta}\left(\hat{a}_{isn}(\omega) - \sum_{n'\in\mathcal{N}_{is}} \frac{\overline{\ell}_{isn'}}{\overline{\ell}_{is}}\hat{a}_{isn'}(\omega)\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}[\hat{a}_{isn}(\omega)\hat{\ell}_{isn}(\omega)] &= \mathbb{E}\left[\hat{a}_{isn}(\omega)\frac{1}{\eta}\left(\hat{a}_{isn}(\omega) - \sum_{n'\in\mathcal{N}_{is}}\frac{\overline{\ell}_{isn'}}{\overline{\ell}_{is}}\hat{a}_{isn'}(\omega)\right)\right] \\ &= \frac{1}{\eta}\left(\mathbb{E}\left[\hat{a}_{isn}(\omega)^2\right] - \sum_{n'\in\mathcal{N}_{is}}\frac{\overline{\ell}_{isn'}}{\overline{\ell}_{is}}\mathbb{E}\left[\hat{a}_{isn}(\omega)\hat{a}_{isn'}(\omega)\right]\right) \\ &= \frac{1}{\eta}\left(\sigma_N^2 + \sigma_S^2 - \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\sigma_N^2 - \sigma_S^2\right) = \frac{1}{\eta}\left(1 - \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)\sigma_N^2.\end{aligned}$$

Then we can calculate $\Phi(m)$,

$$\begin{aligned}\Phi(m) &= \mathbb{E}[a_{isn}(\omega)] + \frac{1-\eta}{2}\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\mathbb{E}[\hat{a}_{isn}(\omega)\hat{\ell}_{isn}(\omega)]ds \\ &= \mathbb{E}[a_{isn}(\omega)] + \frac{1-\eta}{2}\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\frac{1}{\eta}\left(1 - \frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)\sigma_N^2 ds \\ &= \mathbb{E}[a_{isn}(\omega)] + \frac{1}{2}\frac{1-\eta}{\eta}\left(1 - \int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)^2 ds\right)\sigma_N^2.\end{aligned}$$

Therefore,

$$\frac{\partial \log\Phi(m)}{\partial \log m} = -\frac{1}{2}\frac{1-\eta}{\eta}\frac{\partial}{\partial \log m}\left[\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)^2 ds\right]\frac{\sigma_N^2}{\Phi(m)}$$

By Lemma 5, $\frac{\partial}{\partial \log m}\left[\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)^2 ds\right] < 0$. Then by Proposition 1, the first result will follow. Similarly, by Lemma 5, $\frac{\partial}{\partial \log m}\left[\int_0^1 \frac{\overline{\ell}_{is}}{\overline{\ell}_i}\sum_{n\in\mathcal{N}_{is}}\left(\frac{\overline{\ell}_{isn}}{\overline{\ell}_{is}}\right)^2 ds\right] \to 0$ as $m \to \infty$ so

$$\frac{d\log w_i}{d\log \ell_i} = \frac{\frac{1}{1-\eta}\frac{\partial \log\Phi(m_i)}{\partial \log m_i}}{1 - \frac{1}{1-\eta}\frac{\partial \log\Phi(m_i)}{\partial \log m_i}} \to 0,$$

as $m \to \infty$.

□

References

Adachi, D., Fukai, T., Kawaguchi, D., and Saito, Y. U. (2020). Commuting zones in Japan. Discussion papers 20-E-021, Research Institute of Economy, Trade and Industry (RIETI).

Andersson, F., Burgess, S., and Lane, J. I. (2007). Cities, matching and the productivity gains of agglomeration. *Journal of Urban Economics*, 61(1):112–128.

Andersson, M., Klaesson, J., and Larsson, J. P. (2014). The sources of the urban wage premium by worker skills: Spatial sorting or agglomeration economies? *Papers in Regional Science*, 93(4):727–747.

Berger, D., Herkenhoff, K., and Mongey, S. (2022). Labor market power. *American Economic Review*, 112(4):1147–93.

Bernard, A. B., Jensen, J. B., Redding, S. J., and Schott, P. K. (2018). Global firms. *Journal of Economic Literature*, 56(2):565–619.

Bryan, G. and Morten, M. (2019). The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy*, 127(5):2229–2268.

Combes, P.-P., Duranton, G., and Gobillon, L. (2011). The identification of agglomeration economies. *Journal of Economic Geography*, 11(2):253–266.

Conte, M., Méjean, I., Michalski, T. K., and Schmutz, B. (2024). The volatility advantages of large labor markets. *Available at SSRN*.

Davis, D. R. and Dingel, J. I. (2019). A spatial knowledge economy. *American Economic Review*, 109(1):153–170.

de Almeida, E. T. and de Moraes Rocha, R. (2018). Labor pooling as an agglomeration factor: Evidence from the brazilian northeast in the 2002–2014 period. *EconomiA*, 19(2):236–250.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 4, pages 2063–2117. Elsevier.

Ellison, G. and Glaeser, E. L. (1997). Geographic concentration in us manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105(5):889–927.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

Felix, M. (2024). Trade, labor market concentration, and wages. Technical report, Working paper.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica*, 79(3):733–772.

Gaubert, C. and Itskhoki, O. (2021). Granular comparative advantage. *Journal of Political Economy*, 129(3):871–939.

Gaubert, C., Itskhoki, O., and Vogler, M. (2021). Government policies in a granular global economy. *Journal of Monetary Economics*, 121:95–112.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Hornbeck, R. and Moretti, E. (2024). Estimating who benefits from productivity growth: local and distant effects of city productivity growth on wages, rents, and inequality. *Review of Economics and Statistics*, 106(3):587–607.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics*, 124(4):1403–1448.

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. *The Quarterly Journal of Economics*, 129(1):275–331.

Kondo, K. (2023). Municipality-level Panel Data and Municipal Mergers in Japan. Technical papers 23-T-001, Research Institute of Economy, Trade and Industry (RIETI).

Krugman, P. (1992). *Geography and trade*. MIT press.

Lamadon, T., Mogstad, M., and Setzler, B. (2022). Imperfect competition, compensating differentials, and rent sharing in the us labor market. *American Economic Review*, 112(1):169–212.

Mankiw, N. G. and Whinston, M. D. (1986). Free entry and social inefficiency. *The RAND Journal of Economics*, pages 48–58.

Marshall, A. (1920). *Principles of Economics*. Macmillan.

Miyauchi, Y. (2024). Matching and agglomeration: Theory and evidence from japanese firm-to-firm trade. *Econometrica*, 92(6):1869–1905.

Moretti, E. and Yi, M. (2024). Size matters: Matching externalities and the advantages of large labor markets. Technical report, National Bureau of Economic Research.

Nakajima, K. and Okazaki, T. (2012). Labor pooling as a source of industrial agglomeration—the case of the japanese manufacturing industries—. *Economic Review*, 63(3):227–235.

Overman, H. G. and Puga, D. (2010). Labor pooling as a source of agglomeration: An empirical investigation. In *Agglomeration economics*, pages 133–150. University of Chicago Press.

Papageorgiou, T. (2022). Occupational matching and cities. *American Economic Journal: Macroeconomics*, 14(3):82–132.

Redding, S. J. (2016). Goods trade, factor mobility and welfare. *Journal of International Economics*, 101:148–167.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 4, pages 2119–2171. Elsevier.