# CHAPTER 4

# Agglomeration Theory with Heterogeneous Agents

**Kristian Behrens**[*,†,‡,§], **Frédéric Robert-Nicoud**[§,¶,‖]
[*]Department of Economics, Université du Québec à Montréal, Montréal, QC, Canada
[†]National Research University, Higher School of Economics, Moscow, Russia
[‡]CIRPÉE, Université du Québec à Montréal, Montréal, QC, Canada
[§]CEPR, London, UK
[¶]Geneva School of Economics and Management, Université de Genève, Genève, Switzerland
[‖]SERC, The London School of Economics and Political Science, London, UK

## Contents



*Handbook of Regional and Urban Economics*, Volume 5A
ISSN 1574-0080, http://dx.doi.org/10.1016/B978-0-444-59517-1.00004-0

## Abstract

This chapter surveys recent developments in agglomeration theory within a unifying framework. We highlight how locational fundamentals, agglomeration economies, the spatial sorting of heterogeneous agents, and selection effects affect the size, productivity, composition, and inequality of cities, as well as their size distribution in the urban system.

## Keywords

Agglomeration, Heterogeneous agents, Selection, Sorting, Inequality, City size distribution

## JEL Classification Codes

R12, D31

## 4.1. INTRODUCTION

Cities differ in many ways. A myriad of small towns coexist with medium-sized cities and a few urban giants. Some cities have a diversified economic base, whereas others are specialized by industry or by the functions they perform. A few large cities attract the brightest minds, while many small ones can barely retain their residents. Most importantly, however, cities differ in productivity: large cities produce more output per capita than small cities do. This *urban productivity premium* may occur because of locational fundamentals, because of agglomeration economies, because more talented individuals sort into large cities, or because large cities select the most productive entrepreneurs and firms. The literature from Marshall (1890) on has devoted most of its attention to agglomeration economies, whereby a high density of firms and workers generates positive externalities to other firms and workers. It has done so almost exclusively within a representative agent framework. That framework has proved extremely useful for analyzing many different microeconomic foundations for the urban productivity premium. It is, however, ill-suited to study empirically relevant patterns such as the over representation of highly

educated workers and highly productive firms in large cities. It has also, by definition, very little to say on distributional outcomes in cities.

Individual-level and firm-level data have revealed that the broad macro relationships among urban aggregates reflect substantial heterogeneity at the micro level. Theorists have started to build models to address these issues and to provide microeconomic foundations explaining this heterogeneity in a systematic manner. This chapter provides a *unifying framework of urban systems* to study recent developments in agglomeration theory. To this end, we extend the canonical model developed by Henderson (1974) along several dimensions, in particular to heterogeneous agents.[1] Doing so allows us to analyze urban macro outcomes in the light of microheterogeneity, and to better understand the patterns substantiated by the data. We also show how this framework can be used to study under-researched issues and how it allows us to uncover some caveats applying to extant theoretical work. One such caveat is that sorting and selection are intrinsically linked, and that assumptions which seem reasonable in partial equilibrium are inconsistent with the general equilibrium logic of an urban systems model.

This chapter is organized as follows. Section 4.2 uses a cross section of US cities to document the following set of stylized facts that we aim to make sense of within our framework:

- Fact 1 (size and fundamentals): the population size and density of a city are positively correlated with the quality of its fundamentals.
- Fact 2 (urban premiums): the unconditional elasticity of mean earnings and city size is about 8%, and the unconditional elasticity of median housing rents and city size is about 9%.
- Fact 3 (sorting): the share of workers with at least a college degree increases with city size.
- Fact 4 (selection): the share of self-employed is negatively correlated with urban density and with net entry rates of new firms, so selection effects may be at work.
- Fact 5 (inequality): the Gini coefficient of urban earnings is positively correlated with city size and the urban productivity premium increases with the education level.
- Fact 6 (Zipf's law): the size distribution of US places follows closely a log-normal distribution and that of US metropolitan statistical areas (MSAs) follows closely a power law (aka Zipf's law).

The rest of this chapter is devoted to theory. Section 4.3 sets the stage by introducing the canonical model of urban systems with homogeneous agents. We extend it to allow for

[1] Worker and firm heterogeneity has also sparked new theories in other fields. See, for example, the reviews by Grossman (2013) and Melitz and Redding (2014) of international trade theories with heterogeneous workers and heterogeneous firms, respectively.

heterogeneous fundamentals across locations and show how the equilibrium patterns that emerge are consistent with facts 1 (size and fundamentals), 2 (urban premiums), and, under some assumptions, 6 (Zipf's law). We also show how cities differ in their industrial and functional specialization. Section 4.4 introduces heterogeneous agents and shows how the model with sorting replicates facts 2 (urban premiums), 3 (sorting), and 6 (Zipf's law). The latter result is particularly striking since it arises in a static model and relies solely on the sorting of heterogeneous agents across cities. We also show under what conditions the model with heterogeneous agents allows for selection effects, as in fact 4 (selection), what their citywide implications are, and how they are linked to sorting. Section 4.5 builds on the previous developments to establish fact 5 (inequality). We show how worker heterogeneity, sorting, and selection interact with agglomeration economies to deliver a positive equilibrium relationship between city size and urban inequality. This exercise also reveals that few general results are known, and much work remains to be done in this area.

Before proceeding, we stress that our framework is purely static. As such, it is ill-equipped to study important fluctuations in the fate of cities such as New York, which has gone through periods of stagnation and decline before emerging, or more recently Detroit and Pittsburgh. Housing stocks and urban infrastructure depreciate only slowly, so housing prices and housing rents swing much more than city populations do (Henderson and Venables, 2009). The chapter by Desmet and Henderson (2015) in this handbook provides a more systematic treatment of the dynamic aspects and evolution of urban systems.

We further stress that the content of this chapter reflects the difficult and idiosyncratic choices that we made in the process of writing it. We have opted to study a selective set of topics in depth rather than cast a wide but shallow net. We have, for instance, limited ourselves to urban models and largely omitted "regional science" and "new economic geography" contributions. Focusing on the macro aspects and on heterogeneity, we view this chapter as a natural complement to the chapter by Duranton and Puga (2004) on the microfoundations for urban agglomeration economies in volume 4 of this handbook series. Where Duranton and Puga (2004) take city sizes mostly as given to study the microeconomic mechanisms that give rise to agglomeration economies, we take the existence of these citywide increasing returns for granted. Instead, we consider the urban system and allow for worker and firm mobility across cities to study how agglomeration economies, urban costs, heterogeneous locational fundamentals, heterogeneous workers and firms, and selection effects interact to shape the size, composition, productivity, and inequality of cities. In that respect, we build upon and extent many aspects of urban systems that have been analyzed before without paying much attention to micro level heterogeneity (see Abdel-Rahman and Anas, 2004 for a survey).

## 4.2. FOUR CAUSES AND TWO MOMENTS: A GLIMPSE AT THE DATA

To set the stage and organize our thoughts, we first highlight a number of key stylized facts.[2] We keep this section brief on purpose and paint only the big picture related to the four fundamental causes that affect the first two moments of the income, productivity, and size distributions of cities. We report more detailed results from empirical studies as we go along.

The *four fundamental causes* that we focus on to explain the sizes of cities, their composition, and the associated productivity gains are (a) locational fundamentals, (b) agglomeration economies, (c) the spatial sorting of heterogeneous agents, and (d) selection effects. These four causes influence—either individually or jointly—the spatial distribution of economic activity and the *first moments* of the productivity and wage distributions within and across cities. They also affect—especially jointly—the *second moments* of those distributions. The latter effect, which is important from a normative perspective, has received little attention until now.

### 4.2.1 Locational fundamentals

Locations are heterogeneous. They differ in endowments (natural resources, constructible area, soil quality, etc.), in accessibility (presence of infrastructures, access to navigable rivers and natural harbors, relative location in the urban system, etc.), and in many other first- and second-nature characteristics (climate, consumption and production amenities,

[2] *Data sources:* The "places" data come from the "Incorporated Places and Minor Civil Divisions Datasets: Subcounty Resident Population Estimates: April 1, 2010 to July 1, 2012" file from the US Census Bureau (SUB-EST2012.csv). It contains 81,631 places. For the big cities, we use 2010 Census and 2010 American Community Survey 5-year estimates (US Census Bureau) data for 363 continental US MSAs. The 2010 data on urban clusters come from the Census Gazetteer file (Gaz_ua_national.txt). We aggregate up urban clusters at the metropolitan and micropolitian statistical area level using the "2010 Urban Area to Metropolitan and Micropolitan Statistical Area (CBSA) Relationship File" (ua_cbsa_rel_10.txt). From the relationship file, we compute MSA density for the 363 continental MSAs (excluding Alaska, Hawaii, and Puerto Rico). We also compute "cluster density" at the MSA level by keeping only the urban areas within an MSA and by excluding MSA parts that are not classified as urban areas (variable ua = 99999). This yields two density measures per MSA: overall density, $D$, and cluster density, $b$. We further have the total MSA population and "cluster" population. We also compute an "urban cluster" density measure in the spirit of Wheeler (2004), where the cluster density of an MSA is given by the population-weighted average density of the individual urban clusters in the MSA. The "MSA geological features" variable is constructed using the same US Geological Survey data as in Rosenthal and Strange (2008b): seismic hazard, landslide hazard, and sedimentary bedrock. For illustrative purposes, we take the logarithm of the sum of the three measures. The data on firm births, firm deaths, and the number of small firms come from the County Business Patterns (files msa_totals_emplchange_2009-2010.xls and msa_naicssector_2010.xls) of the US Census Bureau. The data on natural amenities come from the US Department of Agriculture (file natamenf_1_.xls). Lastly, the data on state-level venture capital come from the National Venture Capital Association (file RegionalAggregateData42010FINAL.xls).

geological and climatic hazards, etc.). We regroup all these factors under the common header of *locational fundamentals*. The distinctive characteristics of locational fundamentals are that they are exogenous to our static economic analysis and that they can either attract population and economic activity (positive fundamentals such as a mild climate) or repulse them (negative fundamentals such as exposure to natural hazards). The left panel in Figure 4.1 illustrates the statistical relationship between a particular type of (positive) amenities and the size of US MSAs. The MSA amenity score—constructed by the US Department of Agriculture—draws on six underlying factors: mean January temperature; mean January hours of sunlight; mean July temperature; mean July relative humidity; the percentage of water surface; and a topography index.[3] Higher values of the score are associated with locations that display better amenities—for example, sunny places with a mild climate, both of which are valued by residents.

As can be seen from the left panel in Figure 4.1, locations well endowed with (positive) amenities are, on average, larger. As can be seen from the right panel in Figure 4.1, locations with worse geological features (higher seismic or landslide hazard, and a larger share of sedimentary bedrock) are, on average, smaller after partialling out the effect of amenities.[4]

While empirical work on city sizes and productivity suggests that locational fundamentals may explain about one-fifth of the observed geographical concentration (Ellison and Glaeser, 1999), theory has largely ignored them. Locational fundamentals do, however, interact with other agglomeration mechanisms to shape economic outcomes. They pin down city locations and explain why those locations and city sizes are fairly resilient to large shocks or technological change (Davis and Weinstein, 2002; Bleakley and Lin, 2012). As we show later, they may also serve to explain the size distribution of cities.

### 4.2.2 Agglomeration economies

Interactions within and between industries give rise to various sorts of complementarities and indivisibilities. We regroup all those mechanisms under the common header

[3] Higher mean January temperature and more hours of sunlight are positive amenities, whereas higher mean July temperature and greater relative humidity are disamenities. The topography index takes higher values for more difficult terrain (ranging from 1 for flat plains to 21 for high mountains) and thus reflects, on the one hand, the scarcity of land (Saiz, 2010). On the other hand, steeper terrain may offer positive amenities such as unobstructed views. Lastly, a larger water surface is a consumption amenity but a land supply restriction. Its effect on population size is *a priori* unclear.

[4] The right panel in Figure 4.1 shows that worse geological features are positively associated with population size when one does not control for amenities. The reason is that certain amenities (e.g., temperature) are valued more highly than certain disamenities (e.g., seismic risk). This is especially true for California and the US West Coast, which generate a strong positive correlation between seismic and landslide hazards and climate variables.

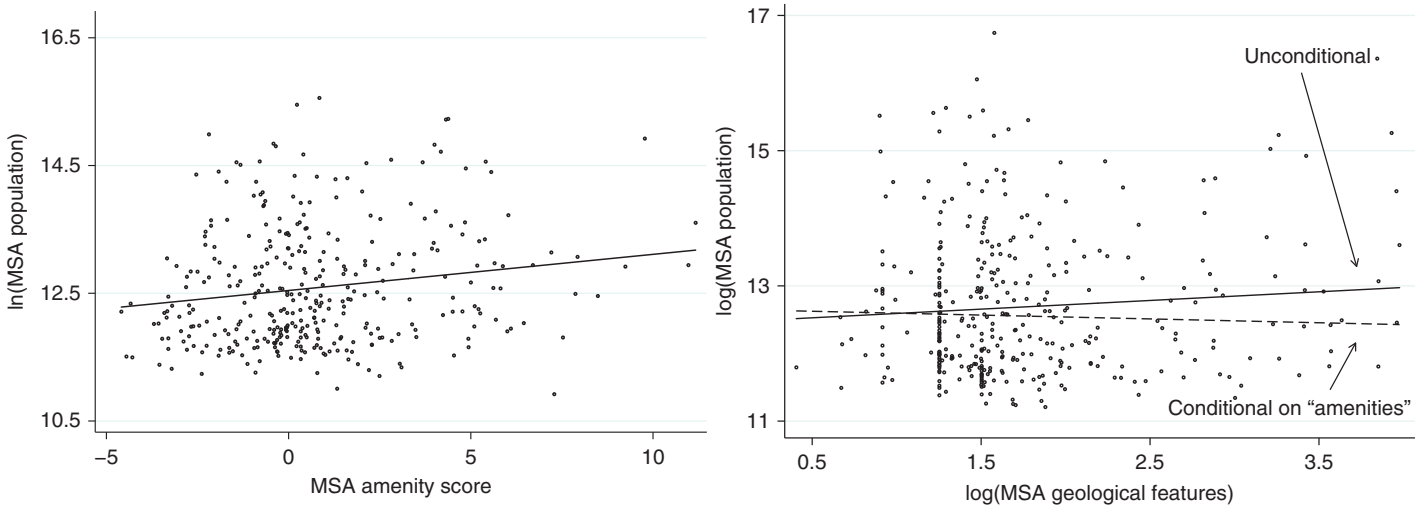


**Figure 4.1** *Fundamentals*. MSA population, climatic amenities, and geological disamenities. *Notes: Authors' calculations based on US Census Bureau, US Department of Agriculture, and US Geological Survey data for 343 and 340 MSAs in 2010 and 2007. See footnote 2 for details. The "MSA geological features" is the product of landslide, seismic hazard, and the share of sedimentary bedrock. The slope in the left panel is 0.057 (standard error 0.019). The unconditional slope in the right panel is 0.059 (standard error 0.053), and the conditional slope is −0.025 (standard error 0.047).*

*agglomeration economies*. These include matching, sharing, and learning externalities (Duranton and Puga, 2004) that can operate either within an industry (localization economies) or across industries (urbanization economies). Labor market pooling, input-output linkages, and knowledge spillovers are the most frequently invoked Marshallian mechanisms that justify the existence of citywide increasing returns to scale.

The left panel in Figure 4.2 illustrates the presence of agglomeration economies for our cross section of US MSAs. The unconditional size elasticity of mean household income with respect to urban population is 0.081 and statistically significant at 1%. This estimate falls within the range usually found in the literature: the estimated elasticity of income or productivity with respect to population (or population density) is between 2% and 10%, depending on the method and the data used (Rosenthal and Strange, 2004; Melo et al., 2009). The right panel in Figure 4.2 depicts the corresponding urban costs ("congestion" for short), with the median gross rent in the MSA as a proxy. The estimated elasticity of urban costs with respect to urban population is 0.088 in our sample and is statistically significant at 1%. Observe that the two estimates are very close: the difference of 0.007 is statistically indistinguishable from zero.[5] Though the measurement of the urban congestion elasticity has attracted much less attention than that of agglomeration economies in the literature, so that it is too early to speak about a consensual range for estimates, recent studies suggest that the gap between urban congestion and agglomeration elasticities is positive yet tiny (Combes et al., 2014). We show later that this has important implications for the spatial equilibrium and the size distribution of cities.

### 4.2.3 Sorting of heterogeneous agents

Though cross-city differences in size, productivity, and urban costs may be the most visible ones, cities also differ greatly in their composition. Most basically, cities differ in their industrial structure: diversified and specialized cities coexist, with no city being a simple replica of the national economy (Helsley and Strange, 2014). Cities may differ both horizontally, in terms of the *set* of industries they host, and vertically, in terms of the *functions* they perform (Duranton and Puga, 2005). Cities also differ fundamentally in their human capital, the set of workers and skills they attract, and the "quality" of their entrepreneurs and firms. These relationships are illustrated in Figure 4.3, which shows that the share of the highly skilled in an MSA is strongly associated with the MSA's size (left panel) and density (right panel). We group under the common header *sorting* all mechanisms that imply that heterogeneous workers, firms, and industries make heterogeneous location choices.

[5] The estimated standard deviation of the difference is 0.011, with a $t$ statistic of 0.63 and a $p$ value of 0.53.

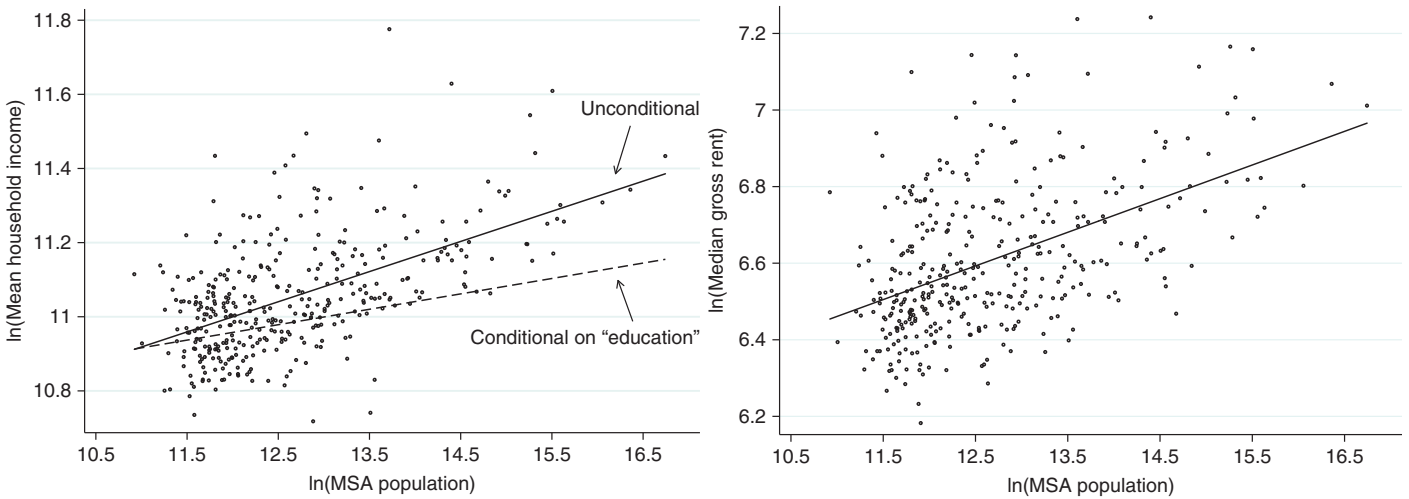


**Figure 4.2** *Agglomeration*. MSA population, mean household income, and median rent. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The unconditional slope in the left panel is 0.081 (standard error 0.006), and the conditional slope is 0.042 (standard error 0.005). The slope in the right panel is 0.088 (standard error 0.008).*

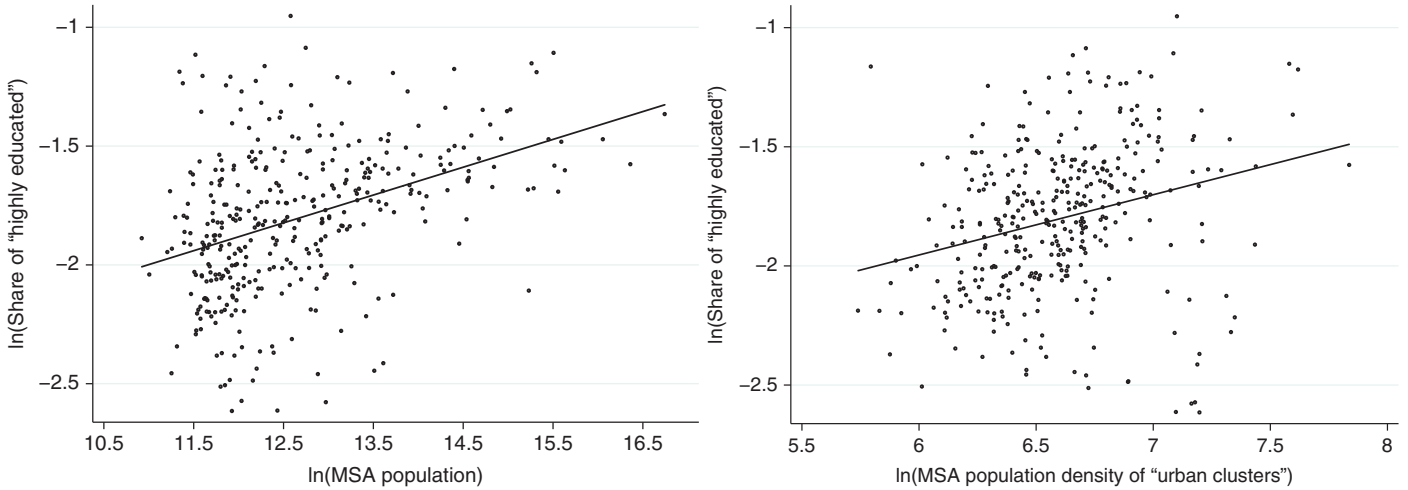


**Figure 4.3** *Sorting*. MSA population, cluster density, and share of "highly educated" workers. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The slope in the left panel is 0.117 (standard error 0.014). The slope in the right panel is 0.253 (standard error 0.048).*

The consensus in the recent literature is that sorting is a robust feature of the data and that differences in worker "quality" across cities explain up to 40–50% of the measured size-productivity relationship (Combes et al., 2008). This is illustrated in the left panel in Figure 4.2, where the size elasticity of wages falls from 0.081 to 0.049 once the share of "highly skilled" is introduced as a control.[6] Although there are some sectoral differences in the strength of sorting, depending on regional density and specialization (Matano and Naticchioni, 2012), sorting is essentially a broad-based phenomenon that cuts across industries: about 80% of the skill differences in larger cities occur within industries, with only 20% accounted for by differences in industrial composition (Hendricks, 2011).

### 4.2.4 Selection effects

The size, density, industrial composition, and human capital of cities affect entrepreneurial incentives and the relative profitability of different occupations. Creating a firm and running a business also entails risks that depend, among other factors, on city characteristics. Although larger cities provide certain advantages for the creation of new firms (Duranton and Puga, 2001), they also host more numerous and better competitors, thereby reducing the chances of success for budding entrepreneurs and nascent firms. They also increase wages, thus changing the returns of salaried work relative to self-employment and entrepreneurship. We group under the common header *selection* all mechanisms that influence agents' occupational choices and the choice of firms and entrepreneurs to operate in the market.

Figure 4.4 illustrates selection into entrepreneurship across US MSAs. Although there is no generally agreed upon measure of "entrepreneurship," we use the share of self-employed in the MSA, or the average firm size, or the net entry rate (firm births minus firm deaths over total number of firms), which are standard proxies in the literature (Glaeser and Kerr, 2009).[7] As can be seen from the left panel in Figure 4.4, there is no clear relationship between MSA size and the share of self-employed in the United States. However, Table 4.1 shows that there is a negative and significant relationship

[6] How to conceive of "skills" or "talent" is a difficult empirical question. There is a crucial distinction to be made between horizontal skills and vertical talent (education), as emphasized by Bacolod et al. (2009a,b, 2010). That distinction is important for empirical work or for microfoundations of urban agglomeration economies, but less so for our purpose of dealing with cities from a macro perspective. We henceforth use the terms "skills," "talent," and "education" interchangeably and mostly conceive of skills, talent, or education as being vertical in nature.

[7] Glaeser and Kerr (2009, pp. 624–627) measure entrepreneurship by "new entry of stand-alone plants." They focus on "manufacturing entrepreneurship" only, whereas our data contain all firms. They note that their "entry metric has a 0.36 and 0.66 correlation with self-employment rates in the year 2000 at the city and state levels, respectively. Correlation with average firm size is higher at −0.59 to −0.80." Table 4.1 shows that our correlations have the same sign, though the correlation with average size is lower.

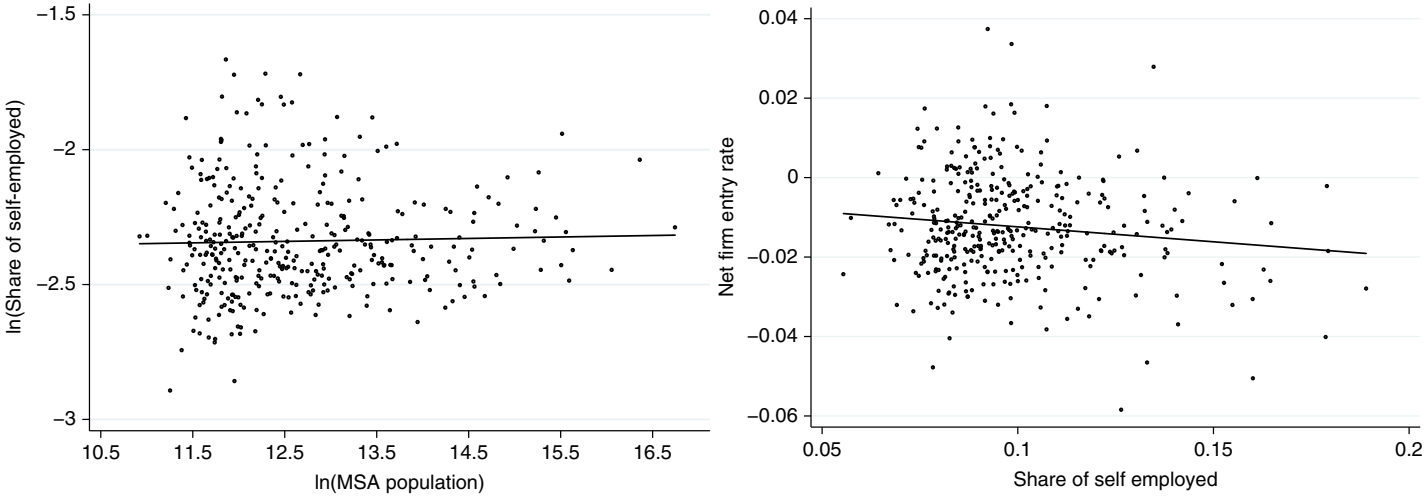


**Figure 4.4** *Selection*. MSA population, share of self-employed, and net entry rates. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The slope in the left panel is 0.005 (standard error 0.010). The slope in the right panel is −0.075 (standard error 0.031).*

**Table 4.1** Correlations between alternative measures of "entrepreneurship" and MSA size

| | "Entrepreneurship" measures | | | |
|---|---|---|---|---|
| **Variables** | **Self-employed (share)** | **log (Average firm employment)** | **Entry rate** | **log (MSA population)** |
| log (MSA population) | 0.0062 | 0.3502* | 0.5501* | – |
| log (MSA density) | −0.1308* | 0.3359* | 0.2482* | 0.6382* |
| log (Average firm employment) | −0.7018* | – | −0.1394* | 0.3502* |
| Exit rate | 0.3979* | −0.2019* | 0.7520* | 0.5079* |
| Entry rate | 0.3498* | −0.1394* | – | 0.5501* |
| Net entry rate | −0.1258* | 0.1144* | 0.2119* | −0.0231 |
| Churning | 0.4010* | −0.1826* | 0.9193* | 0.5664* |
| Venture capital deals (number per capita) | 0.1417* | −0.1396* | −0.0197 | 0.1514* |
| Venture capital invest ($ per capita) | 0.0791 | −0.1028 | 0.0314 | 0.1403* |
| Venture capital invest ($ per deal) | 0.1298* | −0.1366* | 0.1139 | 0.0871 |
| Share of highly educated | 0.2006* | 0.0104 | 0.2414* | 0.4010* |

See footnote 2 for information on the data used. The three venture capital variables are constructed at the state level only (using state-level population for per capita measures). Multistate MSA values are averaged across states. We indicate by asterisks correlations that are significant at the 5% level.

between MSA density and the share of self-employed.[8] Furthermore, as can be seen from the right panel of Figure 4.4 and from the last column of Table 4.1, the net entry rate for firms is lower in larger MSAs. Also, larger cities or cities with more self-employment have smaller average firm sizes, and the latter two characteristics are positively associated with firm churning and different measures of venture capital investment.[9]

The right panel in Figure 4.4 and some correlations in Table 4.1 are suggestive of the possible existence of "selection effects." For example, firm (churning) turnover is substantially higher in bigger cities. We will show that the existence and direction of selection effects with respect to market size or density is theoretically ambiguous: whether more or fewer firms survive or whether the share of entrepreneurs increases or decreases strongly depends on modeling choices. This finding may explain why the current empirical evidence is inconclusive.

[8] The estimated density elasticity from a simple ordinary least squares regression is −0.032 and statistically significant at 1%.

[9] A word of caution is in order. The venture capital data are available only at the state level, and per capita figures are relative to state population. Hence, we cannot account for within-state variation in venture capital across MSAs.

### 4.2.5 Inequality and city size

The size and density of cities are correlated with their composition, with the occupational choices of their residents, and with the success probabilities of businesses. They are also correlated with inequality in economic outcomes. That larger cities are more unequal places is a robust feature of the data (Glaeser et al., 2010; Baum-Snow and Pavan, 2014). This is illustrated in Figure 4.5.

The left panel depicts the relationship between MSA size and inequality as measured by the Gini coefficient of income. The human capital composition of cities has a sizable effect on inequality: the size elasticity of the Gini coefficient falls from 0.011 to 0.008 once education (as measured by the share of college graduates) is controlled for. Size, however, also matters for inequality beyond the sorting of the most educated agents to the largest cities. One of the reasons is that agglomeration interacts with human capital sorting and with selection to "dilate" the income distribution (Combes et al., 2012; Baum-Snow and Pavan, 2014). As can be seen from the right panel in Figure 4.5, the size elasticity of income increases across the income distribution, thus suggesting that agglomeration economies disproportionately accrue to the top of the earnings or productivity distribution of workers and firms.

### 4.2.6 City size distribution

The spatial distribution of population exhibits strong empirical regularities in many countries of the world. Figure 4.6 illustrates these strong patterns for the US data. Two aspects are worth mentioning. First, as can be seen from the left panel in Figure 4.6, the distribution of populated places in the United States is well approximated by a log-normal distribution (Eeckhout, 2004). As is well known, the upper tail of that distribution is difficult to distinguish from a Pareto distribution. Hence, the size distribution of the largest cities in the urban system approximately follows a power law. That this is indeed a good approximation can be seen from the right panel in Figure 4.6: the size distribution of large US cities follows Zipf's law—that is, it follows a Pareto distribution with a unitary shape parameter (Gabaix and Ioannides, 2004; Gabaix, 1999).[10]

### 4.2.7 Assembling the pieces

The foregoing empirical relationships point toward the key ingredients that agglomeration models focusing on citywide outcomes should contain. While prior work has essentially focused on those ingredients individually, we argue that looking at them jointly is important, especially if distributional issues are of concern. To

[10] Rozenfeld et al. (2011) have shown that even the distribution of US "places" follows Zipf's law when places are constructed as geographically connected areas from satellite data. This finding suggests that the distribution is sensitive to the way space is (or is not) partitioned when constructing "places," which is reminiscent of the classic "modifiable areal unit problem" that plagues spatial analysis at large.

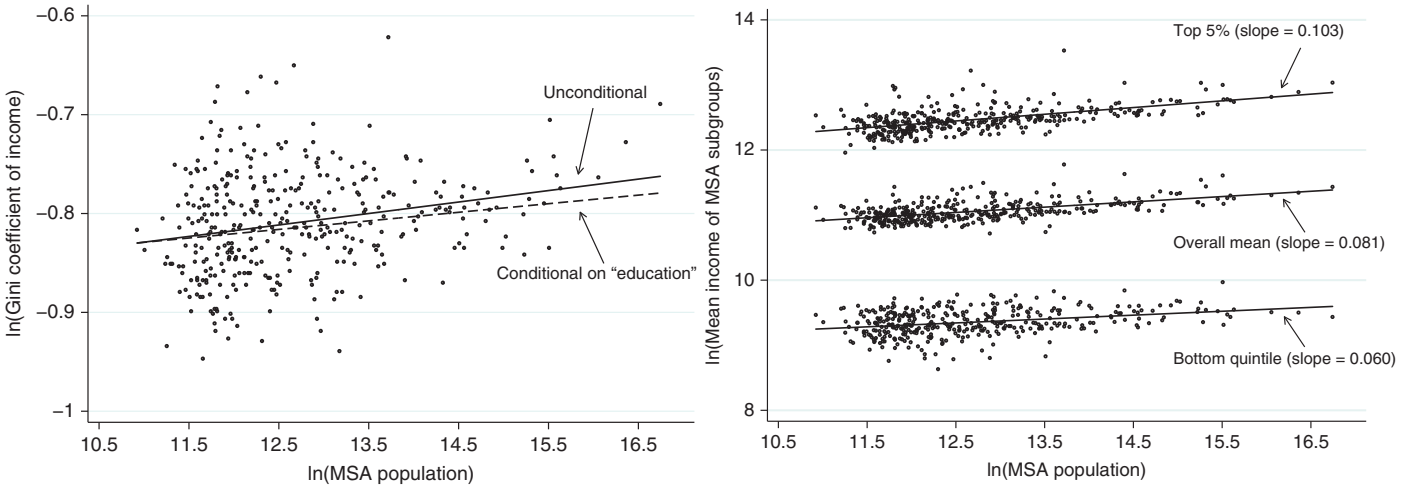


**Figure 4.5** *Inequality*. MSA population, Gini coefficient, and mean incomes by groups. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The unconditional slope in the left panel is 0.012 (standard error 0.003), and the conditional slope is 0.009 (standard error 0.002). The slopes in the right panel are provided in the figure, and they are all significant at 1%.*

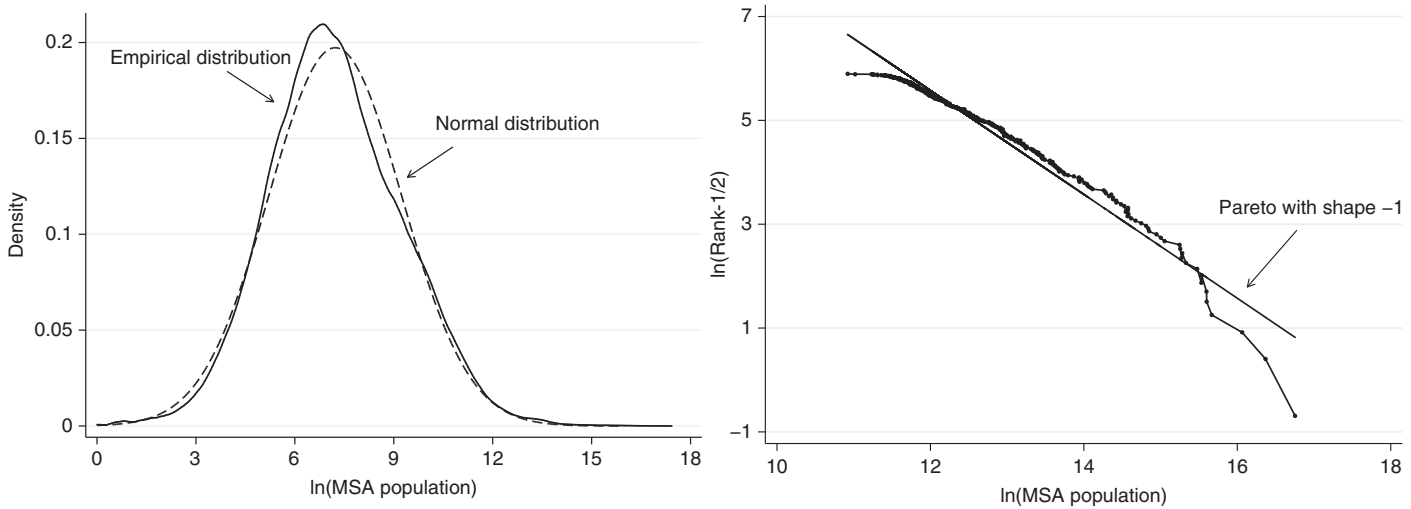


**Figure 4.6** *Size distribution*. Size distribution of places and the rank-size rule of cities. *Notes: Authors' calculations based on US Census Bureau data for 81,631 places in 2010 (left panel) and 363 MSAs in 2010 (right panel). See footnote 2 for details. The estimated slope coefficient in the right panel is −0.922 (standard error 0.009). We subtract 1/2 from the rank as in Gabaix and Ibragimov (2011).*

understand how the four causes (heterogeneous fundamentals, agglomeration economies, and the sorting and selection of heterogeneous agents) interact to shape the two moments (average and dispersion) of the productivity and income distributions, consider the following simple example. Assume that more talented individuals, or individuals with better cognitive skills, gain more from being located in larger cities (Bacolod et al., 2009a). The reasons may be that larger cities are places of intense knowledge exchange, that better cognitive skills allow individuals to absorb and process more information, that information is more valuable in bigger markets, or any combination of these. The complementarity between agglomeration economies—knowledge spillovers in our example—and agents' talent leads to the sorting of more able agents into larger cities. Then, more talented agents make those cities more productive. They also make them places where it is more difficult to succeed in the market—as in the lyrics of Scorsese's eponymous movie "New York, New York, if I can make it there, I'll make it anywhere." Selection effects and increasing urban costs in larger cities then discourage less able agents from going there in the first place, or "fail" some of them who are already there. Those who do not fail, however, reap the benefits of larger urban size. Thus, the interactions between sorting, selection, and agglomeration economies shape the wage distribution and exacerbate income inequality across cities of different sizes. They also largely contribute to shaping the equilibrium size distribution of cities.

## 4.3. AGGLOMERATION

We start by laying out the framework upon which we build throughout this chapter. That framework is flexible enough to encompass most aspects linked to the size, composition, and productivity of cities. It can also accommodate the qualitative relationships in the data we have highlighted, and it lends itself quite naturally to empirical investigation. We are not interested in the precise microeconomic mechanisms that give rise to citywide increasing returns; we henceforth simply assume their existence. Doing so greatly eases the exposition and the quest for a unified framework. We enrich the canonical model as we go along and as required by the different aspects of the theory. Whereas we remain general when dealing with agglomeration economies throughout this chapter, we impose more structure on the model when analyzing sorting, selection, and inequality. We first look at agglomeration theory when agents are homogeneous in order to introduce notation and establish a (well-known) benchmark.

### 4.3.1 Main ingredients

The basic ingredients and notation of our theoretical framework are the following. First, there is set $\mathcal{C}$ of sites. Without loss of generality, one site hosts at most one city. We index cities—and the sites at which they are developed—by $c$ and we denote by $C$ their

endogenously determined number, or mass. Second, there is a (large) number $I$ of perfectly competitive industries, indexed by $i$. Each industry produces a homogeneous final consumption good. For simplicity, we stick to the canonical model of Henderson (1974) and we abstract from intercity trade costs for final goods. We later also introduce nontraded goods specific to some cities.[11] Production of each good requires labor and capital, both of which are freely mobile across cities. Workers are hired locally and paid city-specific wages, whereas capital is owned globally and fetches the same price everywhere. We assume that total output, $Y_{ic}$, of industry $i$ in city $c$ is given by

$$Y_{ic} = \mathbb{A}_{ic}\mathbb{L}_{ic}K_{ic}^{1-\theta_i}L_{ic}^{\theta_i}, \tag{4.1}$$

where $\mathbb{A}_{ic}$ is an industry- and city-specific productivity shifter, which we refer to as "total factor productivity" (TFP); $K_{ic}$ and $L_{ic}$ denote the capital and labor inputs, respectively, with economy-wide labor share $0 < \theta_i \leq 1$; and $\mathbb{L}_{ic}$ is an agglomeration effect external to firms in industry $i$ and city $c$.

Since final goods industries are perfectly competitive, firms in those industries choose labor and capital inputs in Equation (4.1) taking the TFP term, $\mathbb{A}_{ic}$, and the agglomeration effect, $\mathbb{L}_{ic}$, as given. In what follows, bold capitals denote aggregates that are external to individual economic agents. For now, think of them as black boxes that contain standard agglomeration mechanisms (see Duranton and Puga, 2004 and Puga, 2010 for surveys on the microfoundations of urban agglomeration economies). We later open those boxes to look at their microeconomic contents, especially in connection with the composition of cities and the sorting and selection of heterogeneous agents.

## 4.3.2 Canonical model

To set the stage, we build a simple model of a system of cities in the spirit of the canonical model of Henderson (1974). In that canonical model, agglomeration and the size distribution of cities are driven by some external agglomeration effect and the unexplained distribution of TFP across sites. We assume for now that there is no heterogeneity across agents, but locational fundamentals are heterogeneous.

### *4.3.2.1 Equilibrium, optimum, and maximum city sizes*

Consider an economy with a single industry and labor as the sole primary input ($I = 1$ and $\theta_i = 1$). The economy is endowed with $\overline{L}$ *homogeneous* workers who distribute themselves across cities. City formation is endogenous. All cities produce the same homogeneous final good, which is freely tradeable and used as the numeraire. Each city has an exogenous TFP $\mathbb{A}_c > 0$. These city-specific TFP terms are the locational

[11] A wide range of nontraded consumer goods in larger cities are clearly a force pushing toward agglomeration. In recent years, the literature has moved away from the view whereby cities are exclusively places of production to conceive of "consumer cities" as places of consumption of local amenities, goods, and services (Glaeser et al., 2001; Lee, 2010; Couture, 2014).

fundamentals linked to the sites at which the cities are developed. In a nutshell, $\mathbb{A}_c$ captures the comparative advantage of site $c$ to develop a city: sites with a high TFP are particularly amenable to hosting a city. Without loss of generality, we index cities in decreasing order of their TFP: $\mathbb{A}_1 \geq \mathbb{A}_2 \geq \cdots \geq \mathbb{A}_C$.

For cities to arise in equilibrium, we further assume that production exhibits increasing returns to scale at the city level. From (4.1), aggregate output $Y_c$ is such that

$$Y_c = \mathbb{A}_c \mathbb{L}_c L_c. \tag{4.2}$$

Perfect competition in the labor market and zero profits yield a citywide wage that increases with city size: $w_c = \mathbb{A}_c \mathbb{L}_c$. The simplest specification for the external effect $\mathbb{L}_c$ is that it is governed by city size only: $\mathbb{L}_c = L_c^{\epsilon}$. We refer to $\epsilon \geq 0$, a mnemonic for "$\epsilon$xternal," as the elasticity of agglomeration economies with respect to urban population. Many microeconomic foundations involving matching, sharing, or learning externalities give rise to such a reduced-form external effect (Duranton and Puga, 2004). Workers spend their wage net of urban costs on the numeraire good. We assume that per capita urban costs are given by $L_c^{\gamma}$, where the parameter $\gamma$ is the congestion elasticity with respect to urban size. This can easily be microfounded with a monocentric city model in which $\gamma$ is the elasticity of the commuting cost with respect to commuting distance (Fujita, 1989). We could also consider that urban costs are site specific and given by $\mathbb{B}_c L_c^{\gamma}$. If sites differ both in productivity $\mathbb{A}_c$ and in urban costs $\mathbb{B}_c$, most of our results go through by redefining the net advantage of site $c$ as $\mathbb{A}_c/\mathbb{B}_c$. We henceforth impose $\mathbb{B}_c = 1$ for all $c$ for simplicity. Assuming linear preferences for consumers, the utility level associated with living in city $c$ is

$$u_c(L_c) = \mathbb{A}_c L_c^{\epsilon} - L_c^{\gamma}. \tag{4.3}$$

Throughout this chapter, we focus our attention on either of two types of allocation, depending on the topic under study. We characterize the allocation that prevails with welfare-maximizing local governments when studying the composition of cities in Section 4.3.3. We follow this normative approach for the sake of simplicity. In all other cases, we characterize an equilibrium allocation. We also impose the "full-employment condition"

$$\sum_{c \in \mathcal{C}} L_c \leq \overline{L}. \tag{4.4}$$

When agents are homogeneous and absent any friction to labor mobility, a *spatial equilibrium* requires that there exists some *common* equilibrium utility level $u^* \geq 0$ such that

$$\forall c \in \mathcal{C}: \quad (u_c - u^*) L_c = 0, \quad u_c \leq u^*, \tag{4.5}$$

and (4.4) holds. That is to say, all nonempty sites command the same utility level at equilibrium. The spatial equilibrium is "the single most important concept in regional and

urban economics . . . the bedrock on which everything else in the field stands" (Glaeser, 2008, p. 4). We will see later that this concept needs to be modified in a fundamental way when agents are heterogeneous. We maintain the free-mobility assumption throughout the chapter unless otherwise specified. The utility level (4.3) and the indifference conditions (4.5) can be expressed as follows:

$$u_c = \mathbb{A}_c L_c^{\epsilon}\left(1 - \frac{L_c^{\gamma-\epsilon}}{\mathbb{A}_c}\right) = u^*, \tag{4.6}$$

which can be solved for the equilibrium city size $L_c^*$ as a function of $u^*$. This equilibrium is stable only if the marginal utility decreases with city size for all cities with a positive equilibrium population, which requires that

$$\frac{\partial u_c}{\partial L_c} = \epsilon \mathbb{A}_c L_c^{\epsilon-1}\left(1 - \frac{\gamma}{\epsilon}\frac{L_c^{\gamma-\epsilon}}{\mathbb{A}_c}\right) < 0 \tag{4.7}$$

holds at the equilibrium city size $L_c^*$. It is easy to show from Equations (4.6) and (4.7) that a stable equilibrium necessarily requires $\gamma > \epsilon$—that is, urban costs rise faster than urban productivity as the urban population grows. In that case, city sizes are bounded so that not everybody ends up living in a single megacity. We henceforth impose this parameter restriction. Empirically, $\gamma - \epsilon$ seems to be small, and this has important theoretical implications as shown later.

There exist many decentralized equilibria that simultaneously satisfy the full-employment condition (4.4), the indifference condition (4.6), and the stability condition (4.7). The existence of increasing returns to city size for low levels of urban size is the source of potential coordination failures in the absence of large agents able to coordinate the creation of new cities, such as governments and land developers.[12] The precise equilibrium that will be selected—both in terms of sites and in terms of city sizes—is undetermined, but it is *a priori* constrained by the distribution of the $\mathbb{A}_c$ terms, by the number of sites at which cities can be developed, and by the total population of the economy. Figure 4.7 illustrates a decentralized equilibrium with three cities with different underlying TFPs, $\mathbb{A}_1 > \mathbb{A}_2 > \mathbb{A}_3$. This equilibrium satisfies (4.4), (4.6), and (4.7) and yields utility $u^*$ to all urban dwellers in the urban system. Other equilibria may be possible, with fewer or more cities (leading to, respectively, higher and lower equilibrium utility). To

[12] The problem of coordination failure stems from the fact that the utility of a single agent starting a new city is zero, so there is no incentive to do so. Henderson and Venables (2009) develop a dynamic model in which forward-looking builders supply nonmalleable housing and infrastructure, which are sunk investments. In such a setting, either private builders or local governments can solve the coordination problem, and the equilibrium city growth path of the economy becomes unique. Since we do not consider dynamic settings and we focus on static equilibria, we require "static" mechanisms that can solve the coordination problem. Heterogeneity of sites and agents will prove useful here. In particular, heterogeneous agents and sorting along talent across cities may serve as an equilibrium refinement (see Section 4.4). Also, adding a housing market as in Lee and Li (2013) allows one to pin down city sizes.

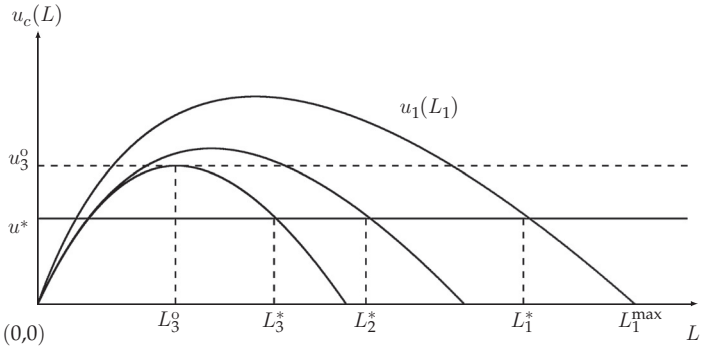


**Figure 4.7** City sizes with heterogeneous $\mathbb{A}_c$ terms.

solve the equilibrium selection problem, the literature has often relied on the existence of large-scale, competitive land developers. When sites are homogeneous, the equilibrium with land developers is both unique and (generally) efficient, arguably two desirable properties (see Henderson, 1988, and Desmet and Henderson, 2015; see also Becker and Henderson 2000b, on the political economy of city formation). When sites are heterogeneous, any decentralized equilibrium (absent transfers across sites) will generally be inefficient though the equilibrium with land developer may be efficient. Providing a full characterization of such an equilibrium is beyond the scope of this chapter.[13] Equilibria feature cities that are larger than the size that a utility-maximizing local government

[13] In Behrens and Robert-Nicoud (2014a), we show that the socially optimal allocation of people across cities and the (unique) equilibrium allocation with perfectly competitive land developers coincide and display the following features: (a) only the most productive sites are developed and more productive sites host larger cities; (b) (gross) equilibrium utility increases with $\mathbb{A}_c$ and equilibrium utility net of equilibrium transfers to competitive land developers is equalized across cities and is weakly smaller than $u_C^o$, where $u_C^o$ is the maximum utility that can be achieved at the least productive populated urban site (thus all developers owning inframarginal sites make pure profits); (c) the socially optimal size of any city $c$ is strictly lower than $L_c^{max}$; and (d) the socially optimal size of any city $c$ is strictly larger than the size chosen by local governments $L_c^o$ for all cities but the smallest, for which the two may coincide. If $\mathcal{C} \subseteq \mathbb{R}$ and if $\mathbb{A}(c)$ is a continuous variable, then $u^* \leq u_C^o$ and $L_C^* \geq L_C^o$. Note that the allocation associated with *local governments that can exclude people* (implementing zoning restrictions, greenbelt policies, or city boundaries) and that maximize the welfare of their current residents violates the indifference condition (4.6) of the standard definition of the urban equilibrium because

$$u\left(L_c^o\right) = \frac{\gamma - \epsilon}{\epsilon}\left(\frac{\epsilon}{\gamma}\mathbb{A}_c\right)^{\frac{\gamma}{\gamma-\epsilon}}$$

increases with $\mathbb{A}_c$. That is, residents of high-amenity places are more fortunate than others because their local authorities do not internalize the adverse effects of restricting the size of their community on others. This raises interesting public policy and political economy questions—for example, whether high-amenity places should implement tax and subsidy schemes to attract certain types of people and to expand beyond the size $L_c^o$ chosen in the absence of transfers. Albouy and Seegert (2012) make several of the same points and analyze under what conditions the market may deliver too many and too small cities when land is heterogeneous and when there are cross-city externalities due to land ownership and federal taxes.

would choose. From a national perspective, some cities may be oversized and some undersized when sites are heterogeneous.[14] In order to characterize common properties of decentralized equilibria, we first derive bounds on feasible city sizes. Let $L_c^{\max}$ denote the maximum size of a city, which is determined by the utility that can be secured by not residing in a city and which we normalize to zero for convenience. Hence, plugging $u^* = 0$ into (4.6) and solving for $L_c$ yields

$$L_c^{\max} = \mathbb{A}_c^{\frac{1}{\gamma-\epsilon}}. \tag{4.8}$$

Let $L_c^o$ denote the size that would be implemented by a *local* government in city $c$ that can restrict entry but cannot price discriminate between current and potential residents, and that maximizes the welfare of its residents. This provides a lower bound to equilibrium city sizes by (4.7) and $\gamma > \epsilon$. Maximizing (4.3) with respect to $L_c$ and solving for $L_c^o$ yields

$$L_c^o = \left(\frac{\epsilon}{\gamma}\mathbb{A}_c\right)^{\frac{1}{\gamma-\epsilon}}. \tag{4.9}$$

Equations (4.8) and (4.9) establish that the lower and upper bounds of city sizes are both proportional to $\mathbb{A}_c^{1/(\gamma-\epsilon)}$. At any spatial equilibrium, the utility level $u^*$ is in $[0, u_C^o]$, where $u_C^o$ is the maximum utility that can be achieved in the city with the smallest $\mathbb{A}_c$ (in the decentralized equilibrium with three cities illustrated in Figure 4.7, $u_C^o$ is $u_3^o$). Cities are oversized in any equilibrium such that $u^* < u_C^o$ because individuals do not take into account the negative impact they impose on other urban dwellers at the margin when making their location decisions. This coordination failure is especially important when thinking about the efficiency of industrial coagglomeration (Helsley and Strange, 2014), as we discuss in Section 4.3.3.1.

What can the foregoing results for the bounds of equilibrium city sizes teach us about the equilibrium city size distribution? Rearranging (4.6) yields

$$L_c^* = \left(\mathbb{A}_c - \frac{u^*}{L_c^{*\epsilon}}\right)^{\frac{1}{\gamma-\epsilon}}. \tag{4.10}$$

Equation (4.10) shows that $L_c^*$ is smaller than but gets closer to $\mathbb{A}_c^{1/(\gamma-\epsilon)}$ when $L_c^*$ becomes large (to see this, observe that $\lim_{L_c^* \to \infty} u^*/L_c^{*\epsilon} = 0$). Therefore, *the upper tail* of the equilibrium city size distribution $L_c^*$ inherits the properties of the TFP distribution in the same way as $L_c^o$ and $L_c^{\max}$ do. In other words, the distribution of $\mathbb{A}_c$ is crucial for determining the distribution of equilibrium sizes of large cities. We trace out implications of that property in the next section.

[14] The optimal allocation requires one to equalize the net marginal benefits across all occupied sites. Henderson (1988) derives several results with heterogeneous sites, some of them heuristically. See also Vermeulen (2011), Albouy and Seegert (2012), and Albouy et al. (2015).

We can summarize the properties of the canonical model, characterized by Equations (4.7)–(4.10), as follows:

**Proposition 4.1** (equilibrium size). *Let $\gamma > \epsilon > 0$ and assume that the utility level enjoyed outside cities is zero. Then any stable equilibrium features city sizes $L_c^* \in \left[L_c^o, L_c^{\max}\right]$ and a utility level $u^* \in \left[0, u_C^o\right]$. Equilibrium city sizes are larger than the sizes chosen by local governments and both $L_c^o$ and $L_c^{\max}$ are proportional to $\mathbb{A}_c$. Finally, in equilibrium the upper tail of the size distribution of cities follows the distribution of the TFP parameters $\mathbb{A}_c$.*

Four comments are in order. First, although all agents are free to live in cities, some agents may opt out of the urban system. This may occur when the outside option of not living in cities is large and/or when the number of potential sites for cities is small compared with the population. Second, not all sites need to develop cities. Since both $L_c^o$ and $L_c^{\max}$ increase with $\mathbb{A}_c$, this is more likely to occur for any given number of sites if locational fundamentals are good, since $L_c^*$ is bounded by two terms that both increase with $\mathbb{A}_c$.[15] Third, the empirical link between city size and $\mathbb{A}_c$ (with an index of natural amenities or with geological features as a proxy) is borne out in the data, as illustrated by the two panels in Figure 4.1. Regressing the logarithm of the population on the MSA amenity score yields a positive size elasticity of 0.057, statistically significant at the 1% level. Lastly, we argued in Section 4.2.2 that $\gamma - \epsilon$ is small in the data. From Proposition 4.1 and from Equation (4.10), we thus obtain that small differences in the underlying $\mathbb{A}_c$ terms can map into large equilibrium size differences between cities. In other words, we may observe cities of vastly different sizes even in a world where locational fundamentals do not differ much across sites.

#### 4.3.2.2 Size distribution of cities

One well-known striking regularity in the size distribution of cities is that it is roughly log-normal, with an upper tail that is statistically indistinguishable from a Pareto distribution with unitary shape parameter: Zipf's law holds for (large) cities (Gabaix, 1999; Eeckhout, 2004; Gabaix and Ioannides, 2004).[16] Figure 4.6 depicts those two properties.

[15] It is reasonable to assume that sites are populated in decreasing order of productivity. Bleakley and Lin (2012, p. 589) show that "locational fundamentals" are good predictors of which sites develop cities. Focusing on "breaks" in navigable transportation routes (portage sites; or hubs in Behrens, 2007), they find that the "footprint of portage is evident today [since] in the south-eastern United States, an urban area of some size is found nearly every place a river crosses the fall line." Those sites are very likely places to develop cities. One should keep in mind, however, that with sequential occupation of sites in the presence of taste heterogeneity, path dependence is an issue (Arthur, 1994). In other words, the most productive places need not be developed first, and depending on the sequence of site occupation, there is generally a large number of equilibrium development paths.

[16] The log-normal and the Pareto distributions theoretically have very different tails, but those are arguably hard to distinguish empirically. The fundamental reason is that, by definition, we have to be "far" in the tail, and any estimate there is quite imprecise owing to small sample size (especially for cities, since there are only very few very large ones).

The canonical model has been criticized for not being able to deliver empirically plausible city size distributions other than if ad hoc assumptions are made on the distribution of $\mathbb{A}_c$. Recent progress has been made, however, and the model can generate such distributions on the basis of fairly weak assumptions on the heterogeneity of sites.[17] Proposition 4.1 reveals that the size distribution of cities inherits the properties of the distribution of $\mathbb{A}_c$, at least in the upper tail of that distribution. In particular, if $\mathbb{A}_c$ follows a power law (or a log-normal distribution), then $L_c$ also follows a power law (or a log-normal distribution) in the upper tail. The question then is why $\mathbb{A}_c$ should follow such a specific distribution. Lee and Li (2013) have shown that if $\mathbb{A}_c$ consists of the product of a large number of underlying factors $a_{fc}$ (where $f = 1,2,\ldots,F$ indexes the factors) that are randomly distributed and not "too strongly correlated," then the size distribution of cities converges to a log-normal distribution and is generally consistent with Zipf's law in its upper tail. Formally, this result is the static counterpart of random growth theory that has been widely used to generate city size distributions in a dynamic setting (Gabaix, 1999; Eeckhout, 2004; Duranton, 2006; Rossi-Hansberg and Wright, 2007). Here, the random shocks (the factors) are stacked in the cross section instead of occurring through time. The factors can be viewed broadly as including consumption amenities, production amenities, and elements linked to the land supply in each location. Basically, they may subsume all characteristics that are positively associated with the desirability of a location. Each factor can also depend on city size—that is, it can be subject to agglomeration economies as captured by $a_{fc}L_c^{\epsilon_f}$. Let

$$\mathbb{A}_c \equiv \prod_f a_{fc} \quad \text{and} \quad \mathbb{L}_c \equiv \prod_f L_c^{\epsilon_f} \tag{4.11}$$

and assume that production is given by (4.2). Let $\epsilon \equiv \sum_f \epsilon_f$ subsume the agglomeration effects generated by all the underlying factors. Consistent with the canonical model, we assume that congestion economies dominate agglomeration economies at the margin—that is, $\gamma > \epsilon$. Plugging $\mathbb{A}_c$ and $\mathbb{L}_c$ into (4.8), and assuming that the outside option leads to a utility of zero so that $u^* = 0$, we find the equilibrium city size is $L_c^* = \mathbb{A}_c^{1/(\gamma-\epsilon)}$. Letting $a_{fc} \equiv \ln a_{fc}$ and taking the logarithm, we then can rewrite this as

$$\ln L_c^* = \frac{1}{\gamma - \epsilon}\left(\sum_{f=1}^{F} \hat{\alpha}_{fc} + \sum_{f=1}^{F} \overline{\alpha}_{fc}\right), \tag{4.12}$$

where we denote by $\hat{\alpha}_{fc} = \ln a_{fc} - \ln \overline{a_{fc}}$ the demeaned log factor, and where $\overline{a_{fc}}$ is the geometric mean of the $a_{fc}$ terms. As shown by Lee and Li (2013), one can then apply a particular variant of the central-limit theorem to the sum of centered random variables $\sum_{f=1}^{F} \hat{\alpha}_{fc}$ in (4.12) to show that the city size distribution converges asymptotically to a

[17] As shown in Section 4.4.1, there are other mechanisms that may serve the same purpose when heterogeneous agents sort across cities. Hsu (2012) proposes yet another explanation, based on differences in fixed costs across industries and central place theory, to generate Zipf's law.

log-normal distribution $\ln\mathcal{N}\left(\frac{1}{\gamma-\epsilon}\sum_{j=1}^{J}\overline{\alpha}_{fc}, \frac{\sigma^2 F}{(\gamma-\epsilon)^2}\right)$, where $\sigma^2$ is the limit of the variance of the partial sums.[18]

As with any asymptotic result, the question arises as to how close one needs to get to the limit for the approximation to be reasonably good. Lee and Li (2013) use Monte Carlo simulations with randomly generated factors to show that (a) the size distribution of cities converges quickly to a log-normal distribution, and (b) Zipf's law holds in the upper tail of the distribution even when the number of factors is small and when they are quite highly correlated. One potential issue is, however, that the random factors do not correspond to anything we can observe in the real world. To gauge how accurate the foregoing results are when we consider "real factors" and not simulated ones, we rely on US Department of Agriculture county-level amenity data to approximate the $a_{fc}$ terms. We use the same six factors as for the amenity score in Section 4.2.1 to construct the corresponding $\mathbb{A}_c$ terms.[19]

The distribution of the $\mathbb{A}_c$ terms is depicted in the left panel in Figure 4.8, which contrasts it with a normal distribution with the same mean and standard deviation. As can be seen, even a number of observable factors as small as six may deliver a log-normal distribution.[20] However, even if the distribution of factors is log-normal, they should be *strongly and positively associated with city size* for the theory to have significant explanatory power. In words, large values of $\mathbb{A}_c$ should map into large cities. As can be seen from the right panel in Figure 4.8, although there is a positive and statistically significant association between locational fundamentals and city sizes, that relationship is very fuzzy. The linear correlation for our 363 MSAs of the logarithm of the population and the amenity terms is only 0.147, whereas the Spearman rank correlation is 0.142. In words, only about 2.2% of the size distribution of MSAs in the United States is explained by the factors underlying our $\mathbb{A}_c$ terms, even if the latter are log-normally distributed.[21]

[18] As shown by expression (4.12), a key requirement for the result to hold is that the functional forms are all multiplicatively separable. The ubiquitous Cobb–Douglas and constant elasticity of substitution (CES) specifications satisfy this requirement.

[19] The factors are mean January temperature, mean January hours of sunlight, the inverse of mean July temperature, the inverse of mean July relative humidity, the percentage of water surface, and the inverse of the topography index. We take the logarithm of each factor, center the values, and sum them up to generate a county-specific value. We then aggregate these county-specific values by MSA, weighting each county by its land-surface share in the MSA. This yields MSA-specific factors $\mathbb{A}_c$ which map into an MSA size distribution.

[20] Using either the Shapiro–Wilk, the Shapiro–Francia, or the skewness and kurtosis tests for normality, we cannot reject at the 5% level (and almost at the 10% level) the null hypothesis that the distribution of our MSA amenity factors is log-normal.

[21] This may be because we focus on only a small range of consumption amenities, but those at least do not seem to matter that much. This finding is similar to the that of Behrens et al. (2013), who use a structural model to solve for the logit choice probabilities that sustain the observed city size distribution. Regressing those choice probabilities on natural amenities delivers a small positive coefficient, but which does not explain much of the city size distribution either.

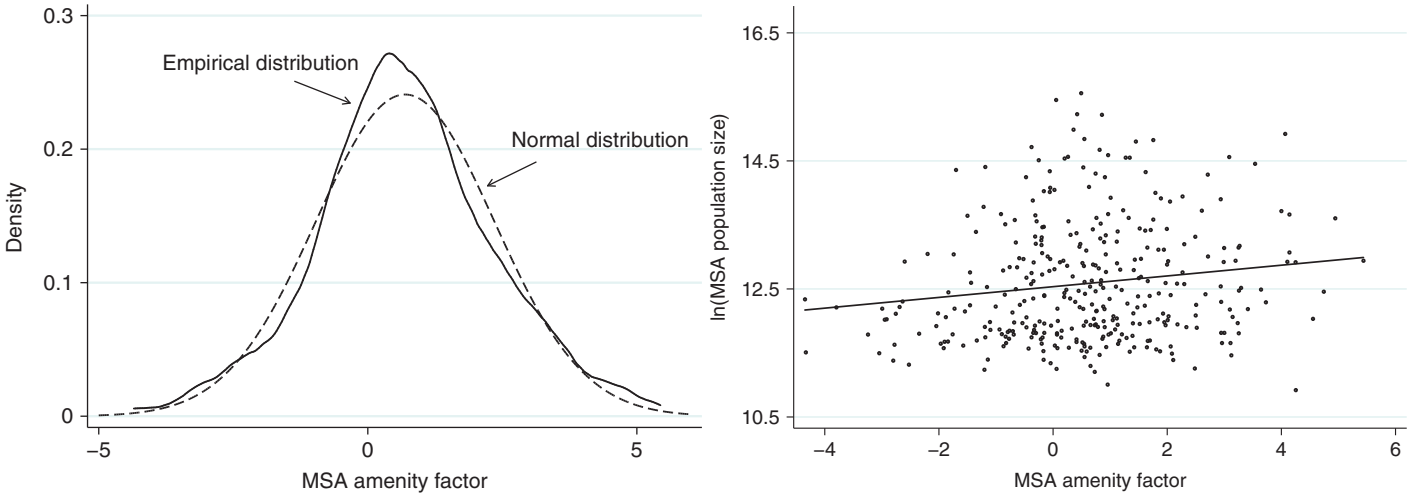


**Figure 4.8** Log-normal distribution of MSA amenity factors $\mathbb{A}_c$, and factors-city size plot. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. The MSA amenity factors are constructed using US Department of Agriculture amenity data. See footnotes 2 and 19 for details. The estimated slope coefficient in the right panel is 0.083 (standard error 0.031).*

Log-normality of $\mathbb{A}_c$ does not by itself guarantee that the resulting distribution matches closely with the ranking of city sizes, which thus breaks the theoretical link between the distribution of amenities and the distribution of city sizes. This finding also suggests that, as stated in Section 4.2.1, locational fundamentals are no longer a major determinant of observed city size distributions in modern economies. We thus have to find alternative explanations for the size distribution of cities, a point we come back to in Section 4.4.1.4.

#### 4.3.2.3 *Inside the "black boxes": extensions and interpretations*

We now use the canonical model to interpret prior work in relation to its key parameters $\epsilon$, $\gamma$, and $\mathbb{A}_c$. To this end, we take a look inside the "black boxes" of the model.

##### Inside $\epsilon$

The literature on agglomeration economies, as surveyed in Duranton and Puga (2004) and Puga (2010), provides microeconomic foundations for $\epsilon$. For instance, if agglomeration economies arise as a result of input sharing, where $Y_c$ is a CES aggregate of differentiated intermediate inputs produced under increasing returns to scale (as in Ethier, 1982), using local labor only, then $\epsilon = 1/(\sigma - 1)$, where $\sigma > 1$ is the elasticity of substitution between any pair of inputs. If, instead, production of $Y_c$ requires the completion of an exogenous set of tasks and urban dwellers allocate their time between learning, which raises their effective amount of productive labor with an elasticity of $\theta \in (0,1)$, and producing (as in Becker and Murphy, 1992; Becker and Henderson, 2000a), then larger cities allow for a finer division of labor and this gives rise to citywide increasing returns, with $\epsilon = \theta$.[22] The same result is obtained in a model where workers have to allocate a unit of time across tasks, and where learning-by-doing increases productivity for a task with an elasticity of $\theta$. What is remarkable in all these models is that, despite having very different underlying microeconomic mechanisms, they generate a reduced-form citywide production function given by (4.2), where only the structural interpretation of $\epsilon$ changes. The empirical literature on the estimation of agglomeration economies, surveyed by Rosenthal and Strange (2004) and Melo et al. (2009), estimates this parameter to be in the range from 0.02 to 0.1 for a variety of countries and using a variety of econometric techniques. The consensus among urban economists nowadays is that the "true" value of $\epsilon$ is closer to the lower bound, especially when unobserved heterogeneity is controlled for using individual data and when different endogeneity concerns are properly addressed (see the chapter by Combes and Gobillon, 2015 in this handbook).

[22] Agglomeration economies may stem from investment in either vertical talent or horizontal skill (Kim, 1989). Larger markets favor investment in horizontal skills (which are useful in specific occupations) instead of vertical talent (which is useful in any occupation) because of better matching in thicker markets.

### Inside $\gamma$

The literature on the microeconomic foundations of urban costs, $\gamma$, is much sparser than the literature on the microeconomic foundations of agglomeration economies. In theory, $\gamma$ equals the elasticity of the cost per unit distance of commuting to the central business district in the one-dimensional Alonso–Muth–Mills model (see also Fujita and Ogawa, 1982; Lucas and Rossi-Hansberg, 2002). It also equals the elasticity of utility with respect to housing consumption in the Helpman (1998) model with an exogenous housing stock. The empirical literature on the estimation of $\gamma$ is scarcer still: we are aware of only Combes et al. (2014). This is puzzling since the relative magnitude of urban costs, $\gamma$, and of agglomeration economies, $\epsilon$, is important for understanding a variety of positive and normative properties of the spatial equilibrium. Thus, precise estimates of *both elasticities* are fundamental. The simplest models with linear cities and linear commuting costs suggest a very large estimate of $\gamma = 1$. This is clearly much too large compared with the few available estimates, which are also close to 2%.

### Inside $\mathbb{A}_c$

The TFP parameters $\mathbb{A}_c$ are related to the industrial or functional composition of cities, the quality of their sites, and their commuting infrastructure. We have seen that heterogeneity in site-specific underlying factors may generate Zipf's law. However, just as the random growth version of Zipf's law, that theory has nothing to say about the microeconomic contents of the $\mathbb{A}_c$ terms. Heterogeneity in sites may stem from many underlying characteristics: production and consumption amenities, endowments, natural resources, and *locational advantage* in terms of transportation access to markets. This issue has received some attention in the new economic geography literature, but multiregion models are complex and thus have been analyzed only sparsely. The reason is that with multiple cities or regions, the relative position matters for access to demand (a positive effect) and exposure to competition (a negative effect). The urban literature has largely ignored costly trade between cities: trade costs are usually either zero or infinite, just as in classical trade theory.

Behrens et al. (2009) extend the "home market effect" model of Krugman (1980) to many locations. There is a mobile increasing returns to scale sector that produces differentiated varieties of a good that can be traded across space at some cost, and there is an immobile constant returns to scale sector that produces some freely traded good. The latter sector differs exogenously by productivity across sites, with productivity $1/z_c$ at site $c$. Sites also differ in their relative advantage for the mobile sector as compared with the outside sector: $a_c = (1/m_c)/(1/z_c)$. Finally, locations differ in access to each other: transportation costs across all sites are of the iceberg type and are represented by some $C \times C$ matrix $\mathbf{\Phi}$, where the element $\phi_{c,c'}$ is the freeness of trade between sites $c$ and $c'$. Specifically, $\phi_{c,c'} \in [0,1]$, with $\phi_{c,c'} = 0$ when trade between sites $c$ and $c'$ is prohibitively costly and $\phi_{c,c'} = 1$ when bilateral trade is costless. Behrens et al. (2009)

show that the equilibrium per capita output of site $c$ is given by $y_c = \mathbb{A}_c$, with $\mathbb{A}_c \equiv A_c(\mathbf{\Phi}, \{a_c\}_{c\in C}, 1/z_c)$. Per capita output increases with the site's productivity, which is a complex combination of its own productivity parameters ($1/z_c$ and $a_c$) and some spatially weighted combination of the productivity parameters of all other sites, and interacts with the spatial transportation cost structure of the economy. Intuitively, sites that offer better access to markets—that are closer to more productive markets, where incomes are higher—have a locational advantage in terms of access to consumers. However, those markets are also exposed to more competition from more numerous and more productive competitors, which may partly offset that locational advantage. The spatial allocation of firms across sites, and the resulting productivity distribution, crucially depends on the equilibrium trade-off between these two forces.[23]

Another model that can be cast into our canonical mold is that of Desmet and Rossi-Hansberg (2013). In their model, per capita output of the homogeneous numeraire good in city $c$ is given by

$$y_c = A_c \mathbb{L}_c k_c^{1-\theta} h_c^{\theta}, \tag{4.13}$$

where $k_c$ and $h_c$ are per capita capital and hours worked, respectively, $A_c$ is a city-specific productivity shifter, and $\mathbb{L}_c = L_c^{\epsilon}$ is the agglomeration externality. Observe that Equation (4.13) is identical to our expression (4.1), except for the endogenous labor-leisure choice: consumers are endowed with one unit of time that can be used for work, $h_c$, or leisure, $1 - h_c$. They have preferences $v_c = \ln u_c + \psi \ln(1 - h_c) + a_c$ that are log-linear in consumption of the numeraire, $u_c$ (which is, as before, income net of urban costs), leisure, and consumption amenities $a_c$.

In each city $c$ of size $L_c$, a local government levies a tax $\tau_c$ on total labor income $L_c w_c h_c$ to finance infrastructure that is used for commuting. A consumer's consumption of the numeraire good is thus given by $u_c = w_c h_c(1 - \tau_c) - R_c$, where $R_c$ is the per capita urban costs (commuting plus land rents) borne by a resident of city $c$. Assuming that cities are monocentric, and choosing appropriate units of measurement, we obtain per capita urban costs $R_c = L_c^{\gamma}$.

Consumers choose labor and leisure time to maximize utility and producers choose labor and capital inputs to minimize costs. Using the optimal choice of inputs, as well as the expression for urban costs $R_c$, we obtain per capita consumption and production as follows:

$$u_c = \theta(1 - \tau_c) y_c - L_c^{\gamma} \quad \text{and} \quad y_c = \kappa A_c^{\frac{1}{\theta}} L_c^{\frac{\epsilon}{\theta}} h_c,$$

[23] The same holds in the model of Behrens et al. (2013). In that model, cross-city differences in market access are subsumed by the selection cutoff for heterogeneous firms. We deal more extensively with selection effects in Section 4.4.2.

where $\kappa > 0$ is a bundle of parameters. Desmet and Rossi-Hansberg (2013) show that $h_c \equiv h_c(\tau_c, A_c, L_c)$ is a monotonically increasing function of $L_c$: agents work more in bigger cities (Rosenthal and Strange, 2008a). Thus $u_c = \mathbb{A}_c h_c(\tau_c, A_c, L_c) L_c^{\epsilon/\theta} - L_c^{\gamma}$, where $\mathbb{A}_c \equiv \mathbb{A}_c(\tau_c, A_c) = \kappa\theta(1-\tau_c)A_c^{1/\theta}$. If utility were linear in consumption and labor supply were fixed (as we have assumed so far), we would obtain an equilibrium relationship that is structurally identical to Equation (4.3). The cross-city heterogeneity in taxes, $\tau_c$, and productivity parameters, $A_c$, serves to shift up or down the equilibrium city sizes via the TFP term $\mathbb{A}_c$.[24] However, labor supply is variable and utility depends on income, leisure, and consumption amenities. Hence, the spatial equilibrium condition requiring the equalization of utility is slightly more complex and is given by

$$\ln\left[\mathbb{A}_c h_c(\tau_c, A_c, L_c) L_c^{\frac{\epsilon}{\theta}} - L_c^{\gamma}\right] + \psi \ln\left[1 - h_c(\tau_c, A_c, L_c)\right] + a_c = u^*, \tag{4.14}$$

for some $u^*$ that is determined in general equilibrium by the mobility of agents. The equilibrium allocation of homogeneous agents across cities depends on the cross-city distribution of three elements: (a) local taxes, $\tau_c$, also referred to as "labor wedges"; (b) exogenous productivity differences, $A_c$; and (c) differences in exogenous consumption amenities, $a_c$. Quite naturally, the equilibrium city size $L_c^*$ increases with $A_c$ and $a_c$, and decreases with $\tau_c$.

The key contribution of Desmet and Rossi-Hansberg (2013) is to apply their spatial general equilibrium model (4.14) in a structural way to the data.[25] To this end, they first estimate the productivity shifters $A_c$ and the labor wedges $\tau_c$ from their structural equations, and infer the amenities $a_c$ such that—conditional on the labor wedges and productivity shifters—the model replicates the observed distribution of city sizes for 192 US cities in 2005–2008. They then evaluate the correlation between the implied $a_c$ and a variety of quality-of-life measures usually used in the literature. Having thus calibrated the model, they finally perform an "urban accounting" exercise. The objective is to quantify the respective contribution of the different wedges—labor $\tau_c$, productivity

[24] The full model of Desmet and Rossi-Hansberg (2013) is more complicated since they also make taxes endogenous. To pin them down, they assume that the local government must provide a quantity of infrastructure proportional to the product of wages and total commuting costs in the city, scaled by some city-specific government inefficiency $g_c$. Assuming that the government budget is balanced then requires that $\tau_c \propto g_c L_c^{\gamma}$—that is, big cities with inefficient governments have higher tax rates.

[25] For more information on the use of structural methods in urban economics, see the chapters by Holmes and Sieg (2014) in this volume of the handbook. Behrens et al. (2013) perform a similar analysis in a very different setting. They use a multicity general equilibrium model that builds on the monopolistic competition framework developed by Behrens and Murata (2007). In that framework, heterogeneous firms produce differentiated varieties of a consumption good that can be traded at some cost across all cities. The key objective of Behrens et al. (2013) is to quantify how trade frictions and commuting costs affect individual city sizes, the size distribution of cities, and aggregate productivity. They find that the city size distribution is fairly stable with respect to trade frictions and commuting costs.

$A_c$, and amenities $a_c$—to city sizes, to welfare, and to the city size distribution. This is achieved by simulating counterfactual changes when one of the three channels—$\tau_c$, $a_c$, or $A_c$—is shut down—that is, what happens if "we eliminate differences in a particular characteristic by setting its value to the population weighted average"? (Desmet and Rossi-Hansberg, 2013, p. 2312). They obtain large population reallocations but small welfare effects.[26] In words, the movement of agents across cities in response to possibly large shocks yields only fairly small welfare gains (see also Behrens et al. 2014a). These results are quite robust to the inclusion of consumption and production externalities in the US data. By contrast, applying their model to Chinese data, Desmet and Rossi-Hansberg (2013) obtain fewer population movements but larger welfare effects.

### 4.3.3 The composition of cities: industries, functions, and skills

Until now, cities differ only in terms of exogenous fundamentals. That cities also differ in their industrial structure is probably the most obvious difference that meets the eye. Cities differ further in many other dimensions, especially in the functions they perform and in whom inhabits them. In this section, we cover recent studies that look at the interactions between agglomeration economies and the industrial, functional, and skill composition of cities. Abdel-Rahman and Anas (2004) and Duranton and Puga (2000) offer comprehensive treatments of the earlier literature, and many of the results we derive on industry composition belong to it. With respect to industry composition, the production mix of large cities is more diversified than that of small ones (Henderson, 1997; Helsley and Strange, 2014). Also, large and small cities do not specialize in the same sectors, and their industrial composition can change rapidly as there is substantial churning of industries (Duranton, 2007).[27] Regarding functional composition, large firms increasingly slice up the value chain and outsource tasks to independent suppliers. Cities of different sizes specialize in different tasks or functions along the value chain, with larger cities attracting the headquarters and small cities hosting production and routine tasks (Duranton and Puga, 2005; Henderson and Ono, 2008). Finally, cities differ in terms of their skill composition. Large cities attract a larger fraction of highly skilled workers than small cities do (Combes et al., 2008; Hendricks, 2011).

[26] Behrens et al. (2013) reach the opposite conclusion in a model with heterogeneous agents. Shutting down trade frictions and urban frictions, they find that population reallocations are rather small, but that welfare and productivity gains may be substantial. As pointed out by Behrens et al. (2013), the rather small welfare effects in their model are driven by their assumption of homogeneous agents.

[27] Smaller cities usually produce a subset of the goods produced in larger cities. See the "number-average size rule" put forward in the empirical work of Mori et al. (2008).

#### 4.3.3.1 Industry composition

We modify Equation (4.1) as follows. Consider an economy with $I$ different industries. Let $p_i$ denote the price of good $i$, which is freely traded, and let $Y_i$ denote physical quantities. Then the value of output of industry $i$ in city $c$ is

$$p_i Y_{ic} = p_i \mathbb{J}_c \mathbb{U}_c \mathbb{L}_{ic} \mathbb{A}_{ic} L_{ic}, \tag{4.15}$$

where $\mathbb{L}_{ic}$ now captures the extent of *localization economies* (namely, to what extent local employment in a given industry contributes to scale economies external to individual firms belonging to that industry), $\mathbb{U}_c$ captures the extent of *urbanization economies* (namely, to what extent local employment, whatever its industry allocation, contributes to external scale economies), and $\mathbb{J}_c$ captures the external effects of industry diversity, following Jacobs (1969). In (4.15), we have made the assumption that urbanization and Jacobs externalities affect all sectors in the same way; this is for simplicity and to avoid a proliferation of cases.

An equilibrium in this model requires that (a) workers of any city $c$ earn the same nominal wage in all active industries in that city—that is, $w_c \geq p_i \mathbb{J}_c \mathbb{U}_c \mathbb{L}_{ic} \mathbb{A}_{ic}$ with equality for all $i$ such that $L_{ic} > 0$—and (b) that they achieve the same utility in all populated cities—that is, $u_c = w_c - L_c^{\gamma} = u^*$ for some $u^*$, if $L_c > 0$. The simplest functional forms consistent with localization economies and urbanization economies are $\mathbb{L}_{ic} = L_{ic}^{\nu}$ and $\mathbb{U}_c = L_c^{\epsilon}$, respectively. A simple functional form for Jacobs externalities that enables us to encompass several cases studied by the literature is given by

$$\mathbb{J}_c = \left[ \sum_{i=1}^{I} \left( \frac{L_{ic}}{L_c} \right)^{\rho} \right]^{\frac{1}{\rho}}, \tag{4.16}$$

where $\rho < 1$ is a parameter governing the complementarity among the different industries: $\rho$ is negative when employment levels in various industries are strongly complementary, positive when they are substitute, and tends to unity when variety does not matter (since $\lim_{\rho \to 1} \mathbb{J}_c = 1$).[28] In (4.16), diversification across industries brings external benefits to urban labor productivity. To see this, note that $\mathbb{J}_c \in \{0, 1\}$ if $c$ is fully specialized in some industry, and $\mathbb{J}_c = I^{-1+(1/\rho)}$ when all industries are equally represented.[29] In the latter case, $\mathbb{J}_c > 1$ (diversification raises urban productivity) because $\rho < 1$. Observe also that (4.16) is homogeneous of degree zero by construction so that it is a pure measure of the industrial diversity of cities (size effects are subsumed in $\mathbb{U}_c$ and $\mathbb{L}_{ic}$).

##### Specialization

Consider first the model of Fujita and Thisse (2013, Chapter 4). In this case, Jacobs and urbanization economies are absent ($\rho = 1$ and $\nu = 0$) and there are no exogenous

[28] See Helsley and Strange (2011) for recent microeconomic foundations to Jacobs externalities.
[29] If $L_{ic} = L_c$ for some $i$, then $\mathbb{J}_c = 0$ if $\rho \leq 0$ and $\mathbb{J}_c = 1$ if $\rho > 0$.

differences across sites ($\mathbb{A}_{ic} = \mathbb{A}_i$, for all $c$). Output of any industry is freely traded among all cities. Thus, there is no benefit in bringing two or more different industries to the same city (Henderson, 1974). A simple proof of this is by contradiction. Assume that an arbitrary city of size $L_c$ is hosting at least two different industries. The per capita urban cost is $L_c^\gamma$. Per capita gross income of workers in industry $i$ is equal to $\mathbb{A}_i L_{ic}^\epsilon$. The fact that there is more than one industry in city $c$ implies $L_{ic} < L_c$. Consider next another city $c'$ specialized in industry $i$, with employment $L_{c'} = L_{ic'} = L_{ic}$. Then, per capita income of workers in industry $i$ net of urban costs is equal to $\mathbb{A}_i L_{ic'}^\epsilon - L_{ic}^\gamma$, which is strictly larger than $\mathbb{A}_i L_{ic}^\epsilon - L_c^\gamma$ because $L_{ic'} = L_{ic}$ and $L_{ic} < L_c$. Hence, a competitive land developer could profitably enter and create a specialized city $c'$ and attract the workers of industry $i$ who are located in city $c$. No diversified city exists in equilibrium. The unique spatial equilibrium of this model of urban systems has cities specialized by industry, and their (optimal) sizes depend only on the industry in which they specialize. We can therefore label cities by their industry subscripts only and write

**Proposition 4.2** (industrial specialization). *Assume that $\rho = 1$, $\nu = 0$, and $\mathbb{A}_{ic} = \mathbb{A}_i$ for all $i$ and all $c$. Then all cities are specialized by industry at the unique spatial equilibrium with competitive land developers, and their size is optimal:*

$$L_i = \left( p_i \frac{\epsilon}{\gamma} \mathbb{A}_i \right)^{\frac{1}{\gamma - \epsilon}}. \tag{4.17}$$

The proof of the first part (specialization) is given in the text above. The second part follows from the fact that competitive land developers create cities that offer the largest possible equilibrium utility to agents, which, given specialization, yields the same result as in the foregoing section where we considered a single industry. Note that the distribution of $L_c^{\gamma - \epsilon}$ need no longer follow the distribution of $\mathbb{A}_c$ in a multi-industry environment; (endogenous) prices in (4.17) may break the link between the two that Proposition 4.1 emphasizes. Note that cities are fully specialized and yet their size distribution approximately follows Zipf's law in the random growth model of Rossi-Hansberg and Wright (2007).

### Industry assignment

The literature on the assignment of industries, occupations, and/or skills to cities dates back to Henderson (1974, 1988). Ongoing work by Davis and Dingel (2014) does this in a multidimensional environment using the tools of assignment theory (Sattinger, 1993; Costinot, 2009).[30] Here, we are interested in the assignment of industries to urban sites. In order to connect tightly with the framework we have developed so far, we assume that

[30] See also Holmes and Stevens (2014) for an application to the spatial patterns of plant-size distributions, and Redding (2012) for an application to regional inequality and welfare.

industries are distinct in their degree of localization economies, now given by $\epsilon_i$. Furthermore, the suitability of each site for an industry may differ, and there is a large finite set $\mathcal{C} = \{1, 2, \ldots, C\}$ of sites. We maintain $\nu = 0$ and $\rho = 1$. We denote by $\mathbb{A}_{ic}$ the site-specific TFP shifter for industry $i$. Assume that all goods can be traded at no cost, so nominal wage net of urban cost provides a measure of utility. We further assume that all goods are essential—that is, they must be produced in some city. There are local city governments that create cities in order to maximize utility of their residents. Agents are mobile between sectors within each city. We disregard integer constraints and assume that all cities are fully specialized (this is literally true if $\mathcal{C}$ is a continuum).

We solve the problem in three steps. First, we solve for the city size chosen by each local government $c$ conditional on industry $i$. As shown by Proposition 4.2, if cities are fully specialized then the size chosen by the local government of a city developed at site $c$ and specialized in industry $i$ is given by (4.17). It offers utility

$$u_{ic} = \left(\frac{\gamma}{\epsilon_i} - 1\right)\left(p_i \frac{\epsilon_i}{\gamma}\mathbb{A}_{ic}\right)^{\frac{\gamma}{\gamma-\epsilon_i}} \tag{4.18}$$

to its residents. Second, local governments choose to specialize their city in the industry that yields the highest utility—namely, they solve $\max_i u_{ic}$. Cities thus specialize according to their comparative advantage. The nature of this comparative advantage is a mixture of Ricardian technology and external scale economies. To see the first part of this statement, let us get rid of differences in external scale economies and temporarily impose $\epsilon_i = \epsilon$ for all $i$. Consider two cities, $c$ and $d$. City $c$ specializes in the production of good $i$ and city $d$ specializes in the production of good $j$ if the following chain of comparative advantage holds:

$$\frac{A_{cj}}{A_{ci}} < \frac{p_i}{p_j} < \frac{A_{dj}}{A_{di}}.$$

This is the well-known chain of Ricardian comparative advantage, as was to be shown.

It is not possible to write such an expression for the more interesting case $\epsilon_i \neq \epsilon_j$. The solution here is to tackle the problem as an assignment problem where we match industries to cities following the method developed by Costinot (2009). This is our third and final step. Taking logarithms and differentiating (4.18), one can easily verify that

$$\frac{\partial^2 \ln u_{ic}}{\partial \epsilon_i \partial \mathbb{A}_{ic}} = \frac{\gamma}{(\gamma - \epsilon_i)^2}\frac{1}{\mathbb{A}_{ic}} > 0;$$

that is, utility is log-supermodular in industry-site characteristics $\mathbb{A}_{ic}$ and agglomeration economies $\epsilon_i$. The outcome is then an allocation with positive assortative matching (PAM) between industries and cities. The quality of urban sites and the strength of agglomeration economies are complements: high-$\mathbb{A}_{ic}$ cities specialize in the production of high-$\epsilon_i$ goods.

The results above crucially hinge on the complementarity between industries and sites, the presence of local governments (which can exclude migrants from joining a city), and the absence of Jacobs externalities. When agents are free to migrate across cities, and in the presence of cross-industry externalities, Helsley and Strange (2014) show that inefficient coagglomeration of industries generally takes place. Migration is a very weak disciplining device for efficiency. Specialized cities are generally too big, whereas coagglomerated cities are generally too big and do not contain the right mix of industries.[31] Part of the problem with multiple industries and cross-industry externalities stems from the fact that *distributions matter*—that is, the optimal location of one industry is conditional on the distribution of industries across cities. In that case, (log)-supermodularity may fail to hold, which can lead to many patterns that do not display regular assignments of industries to sites. A similar issue arises in the context of the sorting of heterogeneous workers that we study in Section 4.4.

Urban sectoral specialization fully accounts for city size differences in this model. However, that cities are fully specialized is counterfactual, and so industry specialization cannot be the main ingredient of a reasonable static explanation for Zipf's law (fact 6). The model would at least need to be combined with a "random growth component" in the spirit of Lee and Li (2013), as discussed in Section 4.3.2.2, or some self-selection constraints of heterogeneous workers in the presence of sorting, as discussed in Section 4.4.1.4. Alternatively, we can consider under what conditions cities end up with a diversified industrial structure in equilibrium.

### Diversification

In general, the optimal industry composition of urban employment depends on the tension between foregone localization economies and higher urban costs, on the one hand, and the Jacobian benefits of diversity—or citywide "economies of scope" to use the terminology of Abdel-Rahman and Anas (2004)—on the other hand.[32] To see this, assume that all industries are symmetric and all sites are homogeneous ($\mathbb{A}_{ic} = \mathbb{A} > 0$, for all $c$ and all $i$). Then the optimal allocation implies $p_i = p$ for all $i$. Without further loss of generality, we choose units so that $p\mathbb{A} = 1$. Consider two cities of equal size $L$. City $c$ is fully specialized ($L_{ic} = L$ for some $i$, and $L_{jc} = 0$, for all $j \neq i$) and city $c'$ is fully diversified ($L_{ic'} = L/I$ for all $i$). Urban costs are the same in both cities under our working

[31] The result regarding the inefficiency of coagglomeration has important implications for empirical research. Indeed, empirical work on agglomeration economies increasingly looks at coagglomeration patterns (Ellison et al., 2010) to tease out the relative contribution of the different Marshallian mechanisms for agglomeration. The underlying identifying assumption is that the observed coagglomeration is "efficient" so that nominal factor returns fully reflect the presence and strength of agglomeration economies. As shown by Helsley and Strange (2014), this will unfortunately not be the case.

[32] See also Abdel-Rahman and Fujita (1993). By assuming free trade among cities, we omit another potential reason for the diversification of cities: to save on transportation costs (Abdel-Rahman, 1996).

assumption. The nominal wage in city $c$ is equal to $w_c = L^{\epsilon+\nu}$, whereas the nominal wage in city $c'$ is equal to $w_{c'} = L^{\epsilon+\nu} I^{-\epsilon} I^{-1+1/\rho}$ by inserting $\mathbb{J}_{c'} = I^{-1+1/\rho}$ and $L_{ic'} = L/I$ into (4.15). It immediately follows that $w_{c'} > w_c$ if and only if $1 + \epsilon < 1/\rho$—that is, the optimal city is diversified if the benefits from diversification, $1/\rho$, are large relative to the scope of localization economies, $\epsilon$. Since $\epsilon > 0$, the foregoing case arises only if $\rho < 1$—that is, if there is complementarity among sectors.[33]

#### 4.3.3.2 Functional composition

The slicing up of the value chain across space (offshoring) and beyond firm boundaries (outsourcing) also has implications for the composition of cities (Ota and Fujita, 1993; Rossi-Hansberg et al., 2009). Duranton and Puga (2005) and Henderson and Ono (2008) report that cities are increasingly specialized by function, whereas Rossi-Hansberg et al. (2009) report a similar pattern *within* cities: urban centers specialize in complex tasks and the suburbs specialize in the routine (back office) tasks.

In this subsection, we are interested in the location of the various activities of firms and no longer in the industrial composition of cities. We thus start by considering a single, representative industry. We briefly turn to the multi-industry case at the end of this subsection.

##### Representative industry

We follow Duranton and Puga (2005) and Ota and Fujita (1993) and consider the location decisions of a firm regarding its various tasks in light of the proximity-localization trade-off. These authors adopt a technological view of the firm in which the costs of coordinating a firm's headquarter and production facilities increase with the geographical distance separating them. Henderson and Ono (2008) report empirical evidence that is consistent with this view. We encapsulate these models into our framework as follows. Each firm conducts headquarter and manufacturing activities, and each activity benefits from its own localization economies. That is to say, the proximity of the headquarters of other firms enhances the productivity of the headquarters of a typical firm, and the proximity of the manufacturing plants of other firms enhances the productivity of its own manufacturing plant. There are two types of tasks, $M$ (for "manufacturing") and $H$ (for "headquarter"), each being specific to one type of activity. All workers in the economy are equally able to perform either task. Let the subscripts v and f pertain to vertically integrated and to functionally specialized cities, respectively. The output of the representative firm of a typical industry is equal to

[33] The assumption $\rho > 1$ is the opposite to the assumption made by Jane Jacobs and is consistent with Sartre's view that "Hell is other people"—namely, diversity lowers the productivity of everybody. In this case, $\mathbb{J}_c = I^{-1+1/\rho} < 1$ if $c$ is fully diversified and $\mathbb{J}_c = 1$ if $c$ is fully specialized. Clearly, urban labor productivity is higher in the former case than in the latter case. This force comes in addition to urban congestion forces and, therefore, also leads to specialized cities.

$$Y_v = \mathbb{A}(\mathbb{M}M)^{\lambda}(\mathbb{H}H)^{1-\lambda} \tag{4.19}$$

if this firm locates its headquarter and manufacturing tasks in the same city (i.e., this city is vertically integrated), and $Y_f = Y_v/\tau$ if it locates these units in two distinct cities (i.e., cities are vertically disintegrated). In expression (4.19), $0 < \lambda < 1$ is the share of manufacturing labor in production, $M$ and $H$ are manufacturing and headquarter employment of the representative firm, $\mathbb{M}$ and $\mathbb{H}$ denote localization economies specific to each type of task, and $\tau > 1$ is a Samuelson "iceberg" cost of coordinating remote headquarter and manufacturing activities. As before, the simplest specification for localization economies is $\mathbb{M} = M^{\epsilon}$ and $\mathbb{H} = H^{\nu}$, where $\epsilon$ and $\nu$ are the size elasticities of agglomeration economies specific to plants and to headquarters, respectively. To stress the main insights of the model in the simplest possible way, we impose symmetry between tasks by assuming $\nu = \epsilon$ and $\lambda = 1/2$.[34] Let $h \equiv H/(H + M)$ denote the share of workers performing headquarter tasks in production, and let $L \equiv H + M$ denote the size of the workforce. The model being symmetric in $H$ and $M$, we can anticipate that the optimal allocation is symmetric too. We may write per capita (average) utility as

$$\begin{aligned} u(\mathbb{I}_v) = {} & \tau^{\mathbb{I}_v - 1}\mathbb{A}\left[(1-h)h\right]^{\frac{1+\epsilon}{2}}L^{\epsilon} \\ & - \mathbb{I}_v L^{\gamma} - (1-\mathbb{I}_v)L^{\gamma}\left[(1-h)^{1+\gamma} + h^{1+\gamma}\right], \end{aligned} \tag{4.20}$$

where $\mathbb{I}_v = 1$ if firms are spatially vertically integrated and $\mathbb{I}_v = 0$ if headquarter and manufacturing activities are located in distinct, functionally specialized cities. The key trade-off between proximity (due to $\tau > 1$) and local congestion (due to $h^{1+\gamma} + (1-h)^{1+\gamma} < 1$) is clearly apparent in (4.20).

Consider first the case of a *vertically integrated* city—namely, a city that contains vertically integrated firms only ($\mathbb{I}_v = 1$). The optimal size and composition of that city are

$$L_v = \left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}}{2^{1+\epsilon}}\right)^{\frac{1}{\gamma-\epsilon}} \quad \text{and} \quad h_v = \frac{1}{2}, \tag{4.21}$$

respectively. Observe that the expression characterizing the optimal integrated city size in (4.21) is structurally identical to (4.9) in the canonical model.

Turning to the case $\mathbb{I}_v = 0$ of *functional cities*—namely, of cities that specialize fully in either headquarter or manufacturing activities—we again have $h_f = 1/2$, so the optimal headquarter-city and manufacturing-city sizes are given by

[34] In practice, agglomeration effects are stronger for high-end services (Combes et al., 2008; Davis and Henderson, 2008; Dekle and Eaton, 1999). Note that $\upsilon > \epsilon$ would imply that service cities are larger than manufacturing cities, in line with the evidence. It can also explain part of the painful adjustment of many former manufacturing powerhouses such as Detroit and Sheffield. We thank Gilles Duranton for pointing this out to us.

$$H_f = M_f = \left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}}{2\tau}\right)^{\frac{1}{\gamma-\epsilon}}. \tag{4.22}$$

We next compare the normative properties of the allocations in (4.21) and (4.22) by plugging the relevant values into the expressions for $u(\mathbb{I}_v)$ in (4.20). In both cases, congestion costs are equal to a fraction $\epsilon/\gamma$ of output at the optimal allocations. Both output and congestion costs are lower in the allocation with functional cities than in the allocation with vertically integrated cities. Which of the two dominates depends on the parameters of the model. Specifically, average utility (consumption of the numeraire good $Y$) with vertically integrated cities and cities specialized by function is given by

$$u_v \equiv u(1) = \frac{\gamma-\epsilon}{\epsilon}\left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}}{2^{1+\epsilon}}\right)^{\frac{\gamma}{\gamma-\epsilon}} \quad \text{and} \quad u_f \equiv u(0) = \frac{\gamma-\epsilon}{\epsilon}\left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}}{2\tau}\right)^{\frac{\gamma}{\gamma-\epsilon}}, \tag{4.23}$$

respectively. The following results then directly follow by inspection of (4.21), (4.22), and (4.23):

**Proposition 4.3** (functional specialization). *Functional cities are larger than vertically integrated cities and yield higher utility if and only if coordination costs are low enough and/or localization economies are strong enough:*

$$u_f > u_v \quad \text{and} \quad H_f = M_f > L_v \quad \text{if and only if} \quad 1 \leq \tau < \tau_{vf} \equiv 2^{\epsilon}. \tag{4.24}$$

When coordination costs are low, the output forgone by coordinating manufacturing activities from a remote headquarters is low. If we keep in mind that the congestion cost is a constant proportion of output, it then follows that the size of functional cities, and the per capita consumption of the numeraire good, decreases with the coordination costs. Strong agglomeration economies by function magnify the level of output lost or saved relative to the allocation with vertically integrated cities.

Duranton and Puga (2005) insist on the time-series implication of Proposition 4.3 (see also the chapter by Desmet and Henderson, 2015 in this volume): cities increasingly specialize by function as coordination costs fall over time owing to technical changes in communication technologies. We can also stress the following cross-sectional implication of Proposition 4.3 when industries differ in the scope of agglomeration economies: given $\tau$, an industry with little scope for localization economies (a low $\epsilon$) is more likely to be vertically integrated and to form vertically integrated cities than an industry with a higher $\epsilon$.

### Functional composition with several industries

We encapsulate (4.15) and (4.16) into (4.19) in order to study the determinants of the localization of headquarter and manufacturing services of *different* industries in the presence of urbanization and Jacobs externalities. Specifically, consider $I$ symmetric industries with production functions

$$Y_i(\mathbb{I}_v) = \tau^{\mathbb{I}_v - 1}\mathbb{A}(\mathbb{M}M_i)^{\frac{1}{2}}(\mathbb{H}H_i)^{\frac{1}{2}}, \quad \text{where} \quad \mathbb{M} = \left(\sum_{j=1}^{I} M_j^{\rho}\right)^{\frac{\epsilon}{\rho}} \quad \text{and} \quad \mathbb{H} = \left(\sum_{j=1}^{I} H_j^{\rho}\right)^{\frac{\epsilon}{\rho}}.$$

We make two observations about this specification. First, the model is symmetric across industries and production factors. We readily anticipate that any optimal allocation will be symmetric in these variables too. Second, this specification assumes away localization economies. Urbanization economies operate if $\epsilon > 0$ and so do Jacobs economies if $\rho < 1$. Assuming these inequalities hold implies that all industries will be represented in all optimal cities. Then the only relevant question is whether the planner creates vertically integrated cities or functionally specialized cities.

Assume that preferences are symmetric in all goods, so $p_i = p$ for all $i$. Let $p \equiv 1$ by choice of the numeraire. Output in a vertically integrated city of size $L$ is given by

$$Y_v \equiv \sum_{i=1}^{I} Y_i(1) = I\mathbb{A}\left[I\left(\frac{L}{2I}\right)^{\rho}\right]^{\frac{\epsilon}{\rho}}\frac{L}{2I} = \mathbb{A}I^{\left(\frac{1}{\rho}-1\right)\epsilon}\left(\frac{L}{2}\right)^{1+\epsilon},$$

where the first equality makes use of the symmetry of the model (and of $M_i = H_i = L/(2I)$ for all $i$ in particular), and the second equality simplifies the expressions. Maximizing per capita output net of urban costs $u = Y/L - L^{\gamma}$ with respect to $L$ and solving for $L$ yields

$$L_v = \left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}I^{\left(\frac{1}{\rho}-1\right)\epsilon}}{2^{1+\epsilon}}\right)^{\frac{1}{\gamma-\epsilon}},$$

which is identical to (4.21) for $I = 1$. We turn now to the joint output of a pair of functional cities (a manufacturing and a headquarter city). Let $M = H = L/2$ denote the (common) size of these cities. Then the joint output is given by

$$Y_f \equiv \sum_{i=1}^{I} Y_i(0) = \frac{\mathbb{A}}{\tau}I^{\left(\frac{1}{\rho}-1\right)\epsilon}\left(\frac{L}{2}\right)^{1+\epsilon}.$$

Maximizing per capita output net of urban costs $u = Y/L - 2(L/2)^{\gamma}$ with respect to $L$ and solving for $L/2$ yields

$$M_f = H_f = \left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}I^{\left(\frac{1}{\rho}-1\right)\epsilon}}{2\tau}\right)^{\frac{1}{\gamma-\epsilon}},$$

which is again identical to (4.22) for $I = 1$. The per capita utility levels $u_v$ and $u_f$ evaluated at the optimal city sizes are proportional to the expressions in (4.23), namely,

$$u_v \equiv u(1) = \frac{\gamma-\epsilon}{\epsilon}\left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}I^{\left(\frac{1}{\rho}-1\right)\epsilon}}{2^{1+\epsilon}}\right)^{\frac{\gamma}{\gamma-\epsilon}} \quad \text{and} \quad u_f \equiv u(0) = \frac{\gamma-\epsilon}{\epsilon}\left(\frac{\epsilon}{\gamma}\frac{\mathbb{A}I^{\left(\frac{1}{\rho}-1\right)\epsilon}}{2\tau}\right)^{\frac{\gamma}{\gamma-\epsilon}}.$$

It then immediately follows that the conditions in (4.24) hold in the current setting too. We conclude that cities specialize by function if and only if coordination costs are low enough and/or if urbanization economies are strong enough.

#### Nursery cities and the life cycle of products

Our framework is also useful to link the life cycle of products to the location of tasks along the value chain. Duranton and Puga (2001) provide evidence from France and the United States that firms locate their innovation activities in *large and diverse* "nursery cities" and afterward relocate the production tasks to smaller manufacturing cities specialized by industry. The reason is that firms face uncertainty and need to discover their optimal production process in the early stages of the product life cycle and afterward want to exploit localization economies in production once they have discovered and mastered the optimal mass production process.

Duranton and Puga (2001) propose a dynamic model with microeconomic foundations that accounts for these facts. It is, however, possible to distill the spirit of their approach using our static framework. The development phase of a product consists of trials and errors and the local experiences of all industries are useful to any other industry: everybody learns from the errors and successes of everyone else.[35] Thus, at the innovation stage urbanization and Jacobs economies dominate, while localization economies are relatively unimportant. In the context of Equations (4.15) and (4.16), the presence of urbanization and Jacobs economies at the development stage implies $\nu^{\mathrm{I}} > 0$ (*size* matters) and $\rho^{\mathrm{I}} < 1$ (*diversity* matters), where the superscript I stands for "innovation." Conversely, localization economies prevail for manufacturing tasks, implying $\epsilon^{\mathrm{M}} > 0$, while urbanization and Jacobs externalities are relatively unimportant at the production stage: $\nu^{\mathrm{M}} = 0$ and $\rho^{\mathrm{M}} = 1$, where the superscript M stands for "manufacturing."

### 4.3.3.3 Skill composition

Hendricks (2011) reports that large US cities are relatively skill abundant and that 80% of the skill abundance of a city is unrelated to its industry composition. Put differently, all industries are more skill intensive in large cities than in small cities. Furthermore, the urban premium of skilled workers is unrelated to the industry that employs them, which is suggestive of the existence of human capital externalities that operate broadly across industries in the city (see Moretti, 2004 for a survey of the empirical evidence).

To see how our framework can make sense of these patterns, assume that there are two types of labor in the economy, unskilled workers and skilled workers. Let $L_c$ denote

[35] Using a model where the success or failure of firms shapes the beliefs of entrants as to how suitable a region is for production, Ossa (2013) shows that agglomeration may take place even when there are no external effects in production. Large cities may in part be large because they signal to potential entrants that they provide an environment amenable to the successful development of new products.

the size of a city, and $h_c$ denote its fraction of skilled workers. Assume that the per capita output of a representative industry net of urban costs is given by

$$u_c = \mathbb{A}_c \left[\mathbb{L}_c h_c^\rho + (1-h_c)^\rho\right]^{\frac{1}{\rho}} - L_c^\gamma,$$

where $\rho < 1$ and $\mathbb{L}_c = L_c^\epsilon$. This expression assumes skill-biased scale effects, whereas local production amenities $\mathbb{A}_c$ are Hicks neutral as before. Maximizing per capita output net of urban costs with respect to the composition and the size of an arbitrary city yields

$$L_c = \left(\frac{h_c}{1-h_c}\right)^{\frac{1-\rho}{\epsilon}} \quad \text{and} \quad L_c^{\gamma-\epsilon} = \frac{\epsilon}{\gamma}\frac{\mathbb{A}_c}{\rho} h_c^\rho (1-h_c)^{-\frac{(1-\rho)^2}{\rho}}, \tag{4.25}$$

respectively. City size, $L_c$, and city skill abundance, $h_c$, are positively correlated by the first expression in (4.25), and both increase with local amenities $\mathbb{A}_c$ under some regularity condition.[36] This generates the positive correlation between skill abundance and city size uncovered by Hendricks (2011).

While the foregoing mechanism relies on the heterogeneity in the TFP terms, $\mathbb{A}_c$, and skill-biased scale effects to generate the positive correlation between size and skills, we now show that the sorting of heterogeneous individuals across cities generates the same relationship without imposing such assumptions.

## 4.4. SORTING AND SELECTION

Our objective in this section is to propose a framework of sorting of heterogeneous agents *across* cities and selection of heterogeneous agents *within* cities. In what follows, we refer to *sorting* as the heterogeneous location choices of heterogeneous workers or firms. We refer to *selection* as either an occupational choice (workers) or a market-entry choice (firms). Our framework is simple enough to highlight the key issues and problems associated with those questions and to encompass recent models that look at them in greater detail. We also highlight two fundamental difficulties that plague sorting and selection models: the general equilibrium feedbacks that arise in cities and the choice of functional forms. In sorting models, general equilibrium feedbacks preclude in many cases supermodularity, thus making the problem of assignment of heterogeneous agents to cities a fairly complicated one. In selection models, selection effects can go in general

[36] Using both expressions to eliminate $L_c$ yields the following implicit equation for $h_c$ as a function of $\mathbb{A}_c$ and of the other parameters of the model:

$$\frac{h_c^{(1-\rho)\frac{\gamma}{\epsilon}-1}}{(1-h_c)^{(1-\rho)\left(\frac{\gamma}{\epsilon}-\frac{1}{\rho}\right)}} = \mathbb{A}_c \frac{\epsilon}{\rho\gamma}.$$

If $\frac{\gamma}{\epsilon} > \min\left\{\frac{1}{1-\rho}, \frac{1}{\rho}\right\}$ then $h_c$ increases with $\mathbb{A}_c$.

either way, thereby precluding clear comparative static results in the absence of specific functional forms. Although several tricks have been used in the literature to cope with both issues, we argue that any analysis of sorting across cities and selection within cities is complicated and unlikely to yield very robust theoretical results. It is here that interactions between theory and empirical analysis become important to select (no pun intended) the "correct" models.

### 4.4.1 Sorting

We first analyze sorting and show that it is closely related to selection in general equilibrium. This will serve as a basis for the analysis of selection in the next subsection.

#### *4.4.1.1 A simple model*

We develop a simple reduced-form extension of the canonical model of Henderson (1974) in which individuals are endowed with heterogeneous ability. Within that model, we then derive (a) a spatial equilibrium with sorting, (b) limiting results when the size elasticity of agglomeration economies, $\epsilon$, and the size elasticity of urban costs, $\gamma$, are small, as vindicated by the data, and (c) limiting results on the city size distribution when $\gamma/\epsilon$ is close to 1. We then show how our model encompasses or relates to recent models in the literature that have investigated either the sorting of workers (Behrens et al., 2014a; Davis and Dingel, 2013; Eeckhout et al., 2014) or the sorting of firms (Baldwin and Okubo, 2006; Forslid and Okubo, 2014; Gaubert, 2014; Nocke, 2006) across locations. Let $t \in [\underline{t}, \overline{t}]$ denote some individual characteristic that is distributed with probability distribution function $f(\cdot)$ and cumulative distribution function $F(\cdot)$ in the population. For short, we refer to $t$ as "talent." More able workers have higher values of $t$. As in the canonical urban model, workers are free to move to the city of their choice. We assume that total population is fixed at $\overline{L}$. The number $C$ of cities, as well as their sizes $L_c$, are as before endogenously determined by workers' location choices. Yet, the *talent composition* of each city is now endogenous and determined by the location choices of heterogeneous individuals. Each worker chooses one city in equilibrium, so $\overline{L} = \sum_c L_c$.

We assume that a worker with talent $t$ supplies $t^a$ efficiency units of labor, with $a > 0$. Labor in city $c$ is used to produce a freely traded homogeneous final consumption good under the constant returns to scale technology (4.2). We ignore site heterogeneity by letting $\mathbb{A}_c = \mathbb{A}$ for all $c$. Hence, $w_c = \mathbb{A}\mathbb{L}_c$ is the wage per efficiency unit of labor. Assuming that agglomeration economies depend solely on city size and are given by $\mathbb{L}_c \equiv L_c^\epsilon$, and that preferences are linear, the utility of a type $t$ agent in city $c$ is given by

$$u_c(t) = \mathbb{A} L_c^\epsilon t^a - L_c^\gamma. \tag{4.26}$$

Note the complementarity between talent and agglomeration economies in (4.26): a larger city size $L_c$ disproportionately benefits the most talented agents. This is the basic force pushing toward the sorting of more talented agents into larger cities, and it

constitutes the "micro-level equivalent" of (4.25) in the previous section. Observe that there are no direct interactions between the talents of agents: the sorting of one type into a location does not depend on the other types present in that location. This assumption, used for example in Gaubert (2014) in the context of the spatial sorting of firms, is restrictive yet simplifies the analysis greatly.[37] When the payoff to locating in a city depends on the *composition* of that city—which is itself based on the choices of all other agents—things become more complicated. We return to this point in Section 4.4.1.6.

Using (4.26), one can readily verify that the single-crossing property

$$\frac{\partial^2 u_c}{\partial t \partial L_c}(t) > 0 \tag{4.27}$$

holds. Hence, utility is *supermodular* in talent and city size, which implies that there will be PAM in equilibrium (Sattinger, 1993). In a nutshell, agents will sort themselves across cities according to their talent. As can be anticipated from (4.26) and (4.27), not all types of agents will choose the same city in equilibrium. The reason is that urban costs are not type specific, unlike urban premia. Hence, only the more talented agents are able to pay the higher urban costs of larger cities, because they earn more, whereas the less talented agents choose to live in smaller cities, where urban costs are also lower.[38]

#### 4.4.1.2 Spatial equilibrium with a discrete set of cities

Let $\mathcal{C} = \{1, 2, \ldots, C\}$ be an exogenously determined set of cities. Because of PAM in (4.27), we know that agents of similar talent will end up locating in similar cities. Hence, we can look at equilibria that induce a partition of talent across cities. Denote by $t_c$ the talent thresholds that pin down the marginal agent who is indifferent between two consecutive cities $c$ and $c+1$. By definition of those thresholds, it must be that

[37] Gaubert (2014) uses a setting similar to ours yet focuses on the sorting of heterogeneous firms. In her model, trade is costless, which implies that the spatial distribution of firms across cities has no impact on the industry price index. Thus, the location choices of firms are driven by city sizes, and not by the composition of cities in terms of the productivity of the firms they host or the overall spatial distribution of the industry.

[38] PAM need not hold in sorting models, especially in general equilibrium. For example, in Mori and Turrini (2005), who build on the work of Krugman (1991), more skilled agents are *less sentitive to market size* because they can more easily absorb the extra costs incurred for trading their good across regions. When trade costs are high enough, this effect may imply that there is a (rather counterfactual) negative relationship between market size and sorting along skills: the more skilled may actually concentrate in the smaller region. Wrede (2013) extends the work of Mori and Turrini (2005) to include housing à la Helpman (1998) and by dropping communication costs. His model is then close to ours and predicts that there is sorting along talent across regions, with the more talented region being larger and commanding higher wages and housing prices. Venables (2011) develops a model of imperfect information in which the most talented workers signal their ability by living in large, expensive cities.

$$\mathbb{A}L_c^{\epsilon}t_c^a - L_c^{\gamma} = \mathbb{A}L_{c+1}^{\epsilon}t_c^a - L_{c+1}^{\gamma}, \quad \text{so} \quad t_c^a = \frac{1}{\mathbb{A}}\frac{1-\left(\frac{L_c}{L_{c+1}}\right)^{\gamma}}{1-\left(\frac{L_c}{L_{c+1}}\right)^{\epsilon}}L_{c+1}^{\gamma-\epsilon}. \tag{4.28}$$

As in the canonical model in Section 4.3.2, expressions (4.28) provide only bounds on the distribution of talent and the corresponding city sizes that can be sustained as equilibria. Any equilibrium must exhibit a partition of talent and a monotonic increase in city sizes associated with higher talent because of PAM. Without any coordinating device such as local developers or local governments, a large number of equilibria can be potentially sustained under sorting.

For expositional purposes, let us assume $\epsilon, \gamma \to 0$ and $\gamma/\epsilon \to 1$. In words, we assume that the size elasticity of agglomeration economies, $\epsilon$, and the size elasticity of urban costs, $\gamma$, are both "small" and of similar magnitude. Although it is debatable what "small" means in numerical terms, the empirical partial correlations of $\hat{\epsilon} = 0.081$ and $\hat{\gamma} = 0.088$ in our data (see Section 4.2) imply that $\hat{\gamma}/\hat{\epsilon} = 1.068$, which is close to 1, and that the gap $\hat{\gamma} - \hat{\epsilon} = 0.007$ is small and statistically indistinguishable from zero. Recent estimates of $\gamma$ and $\epsilon$ using microdata and a proper identification strategy find even smaller values and a tiny gap $\gamma - \epsilon$ between them (Combes et al., 2008, 2014). Using the foregoing limit for the ratio on the left-hand side of (4.28), relationship (4.28) can be rewritten as follows:

$$t_c^a \approx \frac{1}{\mathbb{A}}L_{c+1}^{\gamma-\epsilon}\lim_{\epsilon,\gamma\to 0}\frac{1-\left(\frac{L_c}{L_{c+1}}\right)^{\gamma}}{1-\left(\frac{L_c}{L_{c+1}}\right)^{\epsilon}} = \frac{1}{\mathbb{A}}\frac{\gamma}{\epsilon}L_{c+1}^{\gamma-\epsilon}. \tag{4.29}$$

Taking ratios, we can express condition (4.29) in $c$ and $c-1$ as follows:

$$\left(\frac{t_c}{t_{c-1}}\right)^a = \left(\frac{L_{c+1}}{L_c}\right)^{\gamma-\epsilon} \quad \Rightarrow \quad L_{c+1} = L_c\left(\frac{t_c}{t_{c-1}}\right)^{\gamma-\epsilon} > L_c, \tag{4.30}$$

where the last inequality comes from $\gamma > \epsilon$ and $t_c > t_{c-1}$. Under our approximation, city size can be directly expressed as a function of the talent of its least talented resident:

$$L_c = L(t_c) = \left(\frac{\epsilon}{\gamma}\mathbb{A}t_c^a\right)^{\frac{1}{\gamma-\epsilon}}. \tag{4.31}$$

Clearly, equilibrium city sizes increase with the talent threshold: more talented cities, with a larger $t_c$, are bigger in equilibrium.[39] Recalling that available estimates of $\gamma - \epsilon$

[39] This holds for any partition of talents across cities. Even when there are multiple equilibria, every equilibrium is such that an upward shift of any threshold is accompanied by an increase in city sizes. Clearly, (4.31) depends strongly on the limits. Yet, when the city size distribution has a sufficiently fat upper tail, $L_c/L_{c+1}$ rapidly becomes small, and thus (4.28) implies that $t_c^a \approx L_{c+1}^{\gamma-\epsilon}/\mathbb{A}$. The qualitative implications of (4.31) then approximately carry over to that case.

are a fraction of a percentage point, we find the elasticity $1/(\gamma - \epsilon)$ in the expression above is extremely large: small cross-city differences in talent translate into huge differences in city sizes. More talented cities also have a higher average productivity. Let

$$\overline{t}_c \equiv \left( \int_{t_c}^{t_{c+1}} t^a \mathrm{d}F_c(t) \right)^{\frac{1}{a}} \tag{4.32}$$

denote the city's average talent, where $F_c(\cdot)$ is the city-specific talent distribution. We then have $y_c = \mathbb{A}_c L_c^{\epsilon}$, where $\mathbb{A}_c \equiv \mathbb{A}\overline{t}_c^a$ is the city-specific TFP term, which depends on site characteristics $\mathbb{A}$—common to all sites in the simple model—and the sites' endogenously determined composition in terms of human capital, $\overline{t}_c$. Hence, productivity gains depend on agglomeration economies in a classical sense (via $L_c^{\epsilon}$) and via a human capital composition effect (via $\overline{t}_c^a$). The latter accounts for about 40–50% of the observed differences in wages between cities of different sizes (Combes et al., 2008). Turning to utility, from (4.26) we have

$$u_c(t) = \left( \frac{\epsilon}{\gamma} \mathbb{A} t_c^a \right)^{\frac{\gamma}{\gamma-\epsilon}} \left[ \frac{\gamma}{\epsilon} \left( \frac{t}{t_c} \right)^a - 1 \right], \quad \text{so} \quad \overline{u}_c = y_c - L_c^{\gamma} = \left( \frac{\epsilon}{\gamma} \mathbb{A} t_c^a \right)^{\frac{\gamma}{\gamma-\epsilon}} \left[ \frac{\gamma}{\epsilon} \left( \frac{\overline{t}_c}{t_c} \right)^a - 1 \right].$$

The utility in the first expression is increasing in own talent and ambiguous in the city's minimum talent $t_c$. On the one hand, a more talented city means more effective units of labor and thus higher productivity ceteris paribus, and this benefits all urban dwellers and especially the more talented; see Moretti (2004) for a comprehensive review of the literature on human capital externalities in cities. On the other hand, talented cities are bigger by (4.31) and congestion costs larger, which hurts all urban dwellers equally. The second expression reveals that in the limiting case where $\overline{t}_c/t_c$ is approximately constant across cities (as in Behrens et al. 2014a), average utility is convex in $t_c$: more talented agents are able to leverage their talent by forming larger cities. We have thus established the following result:

**Proposition 4.4** (sorting and city size). *In the simple sorting model, equilibrium city size, $L_c$, and per capita output, $y_c$, are increasing functions of the average talent, $\overline{t}_c$, of the agents located in the city. The equilibrium utility of an agent t located in city c is increasing in own talent t and ambiguous in $t_c$.*

Figure 4.9 illustrates the sorting of agents across three cities. Agents with the lowest talent pick cities of type 1, which are small. Agents with intermediate talent pick cities of type 2, which are larger. Agents with the highest talent pick cities of type 3, which are larger still. As shown before, the equilibrium relationship between talent and utility—and between talent and city size—is convex. More talented agents gain the most from being in large cities, and large cities must be "sufficiently larger" to discourage less talented agents from going there.

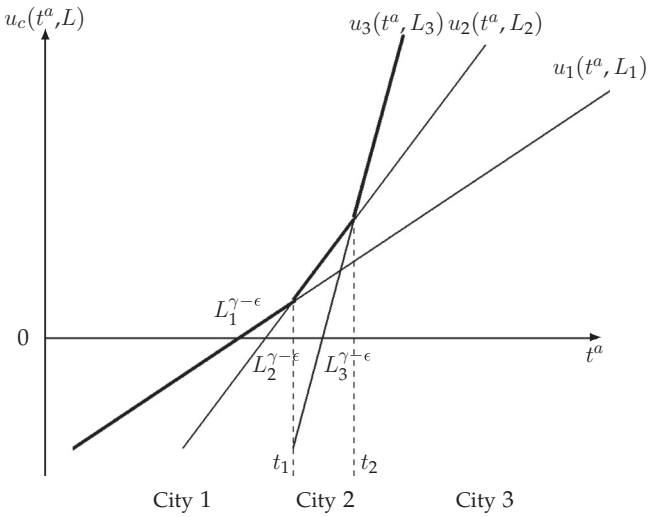


**Figure 4.9** Sorting of heterogeneous agents across three cities.

Three remarks are in order. First, the least talented agent pins down the city size that makes that agent indifferent. Any increase in the size of the city would lead the agent to deviate to a smaller city in order to save on urban costs. In each city, more talented individuals naturally receive higher utility. Second, and as a direct consequence of the previous point, the standard condition for a spatial equilibrium in the absence of mobility frictions—namely, the equalization of utility across all locations—breaks down *since no type is generically represented in all cities*. Except for the marginal types who are indifferent between exactly two cities, all agents are strictly better off in the city of their choice.[40] In words, the ubiquitous condition of equal utility across all populated places naturally ceases to hold in a world where agents differ by type and where different types opt for different locations. The formulation of the spatial equilibrium in (4.6)—"the field's central theoretical tool" (Glaeser and Gottlieb, 2009, p. 984)—must be modified. This has fundamental theoretical and empirical implications.[41] Lastly, the positive correlation between "talent" and city size is strongly borne out in the data, as can be seen from the left panel in Figure 4.3. Sorting matters!

[40] Much of the literature has recently moved away from the idea of a simple spatial equilibrium without frictions or heterogeneity and with equalization of utilities across locations. Behrens et al. (2013), Diamond (2013), Gaubert (2014), and Kline and Moretti (2014) all relax this condition either by introducing mobility frictions explicitly or by assuming that agents have locational taste differences. The latter has been previously applied to new economic geography models by, for example, Murata (2003) and Tabuchi et al. (2002) in order to obtain equilibria that vary smoothly with the parameters of the models.

[41] For instance, regressing individual earnings on a measure of citywide average human capital leads to biased results in the presence of self-selection of agents across locations (this bias is positive if agents with similar abilities make similar choices because the error term is positively correlated with $\bar{t}^a$).

In the foregoing, we looked at "discrete cities,"—that is, cities that span some talent range $[t_c, t_{c+1}]$. Discrete cities induce a discrete partition of the talent space. Though this is empirically relevant because cities host agents of multiple talents, the downside is that the model is quite hard to work with since there is a continuum of equilibria. To solve the model implies specifying a partition, solving for relative city sizes, and choosing a scale for absolute city sizes (by specifying the outside option). Depending on the choice of partition and scale, a multitude of equilibria may be sustained. Part of the problem comes from the fact that we assign a predetermined city structure to agents and then check the equilibrium conditions. Alternatively, we may consider a setting without any predetermined structure in which agents can form any type of city in terms of size and composition.

#### 4.4.1.3 Spatial equilibrium with a continuum of cities

Assume next that agents can choose cities optimally in the sense that they decide—conditional on their talent—which city size they prefer to live in. Formally, an agent with talent $t$ maximizes his or her utility with respect to city size—that is, the agent picks one city size from the menu of all possible city sizes. Here, we assume that the set of cities $\mathcal{C} = [0, C]$ is a continuum. All cities can potentially be formed and the mass (number) of cities $C$ is an endogenous variable. This is essentially the model developed by Behrens et al. (2014a). The first-order condition of that problem is given by[42]

$$\max_{L_c} \; u_c(t) \quad \Rightarrow \quad \mathbb{A}\epsilon L_c^{\epsilon-1} t^a - \gamma L_c^{\gamma-1} = 0, \tag{4.33}$$

which yields the preferred city size of agents with talent $t$:

$$L_c(t) = \left(\frac{\epsilon}{\gamma}\mathbb{A}t^a\right)^{\frac{1}{\gamma-\epsilon}}. \tag{4.34}$$

It is easily verified that the second-order condition holds at the equilibrium city sizes.

Five comments are in order. First, comparing Equations (4.31) and (4.34) reveals that they have the same structure. The difference is that (4.31) applies to the marginal agent, whereas (4.34) applies to any agent. The equilibrium with a large number of discrete cities approaches the one where agents can sort across a continuum of cities.

[42] It is here that the assumption that the city composition does not matter becomes important. In general, the problem of an agent would involve two dimensions: the choice of a city size, and the choice of a city composition. The latter makes matters complicated. Behrens et al. (2014a) simplify the problem by focusing on "talent-homogeneous" cities—that is, cities which host only one type of talent. In that case, solving for $L_c(t)$ involves solving a differential equation. In our simple model, the talent composition does not matter, so size is the only choice variable and cities will trivially be "talent homogeneous," as shown by (4.34).

The intuition is that in the continuous model, all agents are almost indifferent between cities of similar sizes. Yet, every agent has his or her own preferred size, depending on his or her talent.

Second, (4.34) gives a relationship that uniquely maps talents into city size: two different agents would optimally choose to not live in a city of the same size. This significantly narrows down the composition of cities in terms of talents: cities are talent homogeneous, and PAM implies that more talented agents choose to live in larger cities. We trace out the implications of this for the city size distribution in the next subsection. Since every agent picks his or her preferred city, this is a stable equilibrium in the sense that no one can profitably deviate. There are potentially many equilibria with a partition of talent across cities (see the discrete setting in the previous subsection), but in that case not all agents live in a city of the size they would prefer had they the choice of city size. How such an equilibrium, where agents can form the number of cities they wish and each agent chooses to live in a city with his or her preferred size, is actually implemented in the static model is an open question.

Third, having talent heterogeneity and a continuum of cities convexifies the problem of allocating agents to cities. We can think about this convexification as follows. In the discrete case, the utility of type $t$ in city $c$ is $u_c(t) = \mathbb{A}L_c^{\epsilon}(t^a - t_c^a\epsilon/\gamma)$, which is a linear function of $t^a$ (recall that $L_c$ depends only on the marginal type $t_c$). A change in $L_c$ in city $c$ will change the talent composition of that city (see Figure 4.9), yet can be sustained as an equilibrium if the change in $L_c$ is not too large: city sizes are not uniquely determined. In the continuous case, the utility of type $t$ in a city of optimal size is $u_c(t) = \mathbb{A}L_c^{\epsilon}t^a(1-\epsilon/\gamma) = (\epsilon/\gamma)^{\epsilon/(\gamma-\epsilon)}(\mathbb{A}t^a)^{\gamma/(\gamma-\epsilon)}(1-\epsilon/\gamma)$, which is a strictly convex function of $t^a$. The convexification stems from the fact that an increase in talent raises utility more than linearly as city size changes with the talent of its representative urban dweller. Contrary to the discrete case, the size–talent relationship is uniquely determined. Intuitively, a city cannot grow larger or smaller than (4.34) because of the existence of arbitrarily similar cities in terms of size and talent to which agents could deviate to get higher utility.

Fourth, per capita output in a type $t$ city is given by $y_c = \mathbb{A}L_c^{\epsilon}t^a$. If we take logarithms, this becomes either

$$\ln y_c = \kappa_1 + \epsilon \ln L_c + a \ln t_c \tag{4.35}$$

or

$$\ln y_c = \kappa_2 + \gamma \ln L_c, \tag{4.36}$$

where (4.36) is obtained by making use of (4.34). Hence, a log–log regression of productivity $y_c$ on size $L_c$ yields either the elasticity of agglomeration economies in (4.35), where sorting is controlled for, or the elasticity of urban costs in (4.36), where sorting is not controlled for.

Last, taking logarithms of (4.34), we obtain $\ln t_c = \kappa + \frac{\gamma-\epsilon}{a} \ln L_c$, where $\kappa$ is some constant term. When $\gamma - \epsilon$ is small, the elasticity of talent with respect to city size is small: the size elasticity of "education" with respect to city size is 0.117 in our US data (see the left panel in Figure 4.3). The fact that large cities are only slightly more "talented"—as measured by educational attainment of the city population—is the mirror image of the property that small differences in education have to be offset by large differences in city sizes. Thus, a small elasticity of talent with respect to city size is in no way indicative that sorting is unimportant, as some authors have sometimes argued.

#### 4.4.1.4 Implications for city sizes

As shown before, the sorting of heterogeneous individuals across cities gives rise to cities of different equilibrium sizes. What does the theory imply for the *size distribution* of cities? We now use the model with a continuum of cities to show that the implications for that distribution are striking. Observe first that the "number" of agents of talent $t$ in the population is given by $\overline{L}f(t)$. As shown before, agents of talent $t$ prefer cities of size $L(t)$ as given by (4.34). Assume that $n(t)$ of such cities form. Since all agents choose a city in equilibrium, it must be the case that $\overline{L}f(t) = n(t)L(t)$ or, equivalently,

$$n(t) = \frac{\overline{L}f(t)}{L(t)}. \tag{4.37}$$

Let $C$ denote the total mass of cities in the economy. The cumulative distribution $N(\cdot)$ of cities is then given by

$$N(\tau) = \frac{\overline{L}}{C}\int_0^\tau \frac{f(t)}{L(t)}\mathrm{d}t.$$

Using the relationship between talent and size (4.34), we have

$$\frac{f(t)}{L(t)} = \frac{f\left(\xi L(t)^{\frac{\gamma-\epsilon}{a}}\right)}{L(t)} \quad \text{and} \quad \mathrm{d}L = \frac{a}{\xi(\gamma-\epsilon)}L(t)^{1-\frac{\gamma-\epsilon}{a}}\mathrm{d}t,$$

where $\xi \equiv \left(\frac{\epsilon}{\gamma}\mathbb{A}\right)^{-\frac{1}{a}}$ is a positive bundle of parameters. With use of the distribution of talent and the change in variable from talent to city size, the density and the cumulative distribution of city sizes are given by

$$n(L) = \frac{\overline{L}\eta\xi}{C}f(\xi L^\eta)L^{\eta-2} \quad \text{and} \quad N(L) = \frac{\overline{L}\eta\xi}{C}\int_0^\ell f(\xi\ell^\eta)\ell^{\eta-2}\mathrm{d}\ell, \tag{4.38}$$

with $\eta \equiv \frac{\gamma-\epsilon}{a}$. The first-order approximation of (4.38) around $\eta = 0$ is given by

$$n(L) = \kappa L^{-2}, \tag{4.39}$$

where $\kappa \equiv \frac{\overline{L}\eta\xi}{C} f(\xi) > 0$ is a positive constant (recall that $\eta$ remains positive). Using this expression and the full-employment condition, $\overline{L} = \int_{L(\underline{t})}^{L(\overline{t})} n(L) L \mathrm{d}L$, and solving for the equilibrium mass of cities yields

$$C = \eta\xi f(\xi)[\ln L(\overline{t}) - \ln L(\underline{t})]\overline{L};$$

that is, the number of cities is proportional to the size of the population. The urban system displays constant returns to scale in equilibrium. Thus, by inspection of Equation (4.39), we can show (Behrens et al., 2014a).

**Proposition 4.5** (Zipf's law). *Assume that agents sort across cities according to* (4.34). *Then the size distribution of cities follows a Pareto distribution with shape parameter* $-1$ *in the limit* $\eta \equiv \frac{\gamma - \epsilon}{a} \to 0$.

The right panel in Figure 4.6 illustrates that relationship. That Zipf's law holds in this model is remarkable because it *does not depend on the underlying distribution of talent in the population*. In other words, when $\gamma - \epsilon$ is small—as seems to be the case in the data—the city size distribution in the model converges to Zipf's law irrespective of the underlying talent distribution.[43] Crucial for obtaining this result are two relatively reasonable requirements. First, the "number" of cities—more precisely the mass of cities—associated with each level of talent is endogenously determined. Second, city sizes are also endogenously determined and agents can sort themselves across cities of their preferred type. Since agents of any type $t$ have a preferred city size that is a continuous function of their talent, taking that talent to a sufficiently large power implies that the resulting city size distribution is of the Zipf type.

Random growth models also (approximately) generate Zipf's law in the steady state if Gibrat's law holds. The latter has been challenged lately on empirical grounds (see Michaels et al., 2012). Desmet and Rappaport (2013) show that Gibrat's law appears to settle once the distribution is of the Zipf type (and not the other way round). The model in this subsection displays one possible mechanism to generate Zipf's law, like the models in Hsu (2012) and Lee and Li (2013).[44] One distinct advantage of our model is that it generates Zipf's law for plausible values of the parameters irrespective of the underlying distribution of talent (which we do not observe).

#### 4.4.1.5 Some limitations and extensions

The model developed in Section 4.4.1.1 has the virtue of simplicity. The flip side is that it naturally has a number of shortcomings. Firstly, like almost any model in the literature

[43] Behrens et al. (2014a) show that convergence to Zipf's law is very fast as $\eta$ gets smaller. For empirically plausible values of $\eta$, the simulated city size distribution is indistinguishable from a Pareto distribution with unitary shape parameter.

[44] Hsu (2012) also generates Zipf's law using a static framework. The mechanism, based on central place theory and fixed costs, is however very different from the other two models reviewed here.

(e.g., Mori and Turrini, 2005; Nocke, 2006; Baldwin and Okubo, 2006; Okubo et al., 2010), it predicts *strict sorting* along a single dimension. Yet, it is well known that there is a significant overlap of productivities in cities. Larger cities host, on average, more able agents, yet there is nothing close to a clear partition along firm productivity and individual education across cities in the data (Combes et al., 2012; Eeckhout et al., 2014; Forslid and Okubo, 2014). For example, although the correlation between the share of highly skilled workers and city size in the United States is statistically very significant (see the left panel in Figure 4.3), the associated $R^2$ in the log–log regression is only 0.161.[45]

Our simple model with a continuum of cities can easily be extended in the spirit of Behrens et al. (2014a) to allow for incomplete sorting along productivity. The idea is to have a two-stage process, where agents sort on an *ex ante* signal (their talent), but where *ex post* productivity is uncertain. Assume that after choosing a city $c$, each agent gets hit by a random productivity shock $s \in [0, \overline{s}_c]$, with cumulative distribution function $G_c(\cdot)$. We can think about $s$ as being luck or "serendipity"—the agent is in the right place at the right time. The efficiency units of labor the agent can supply depend on the agent's talent $t$ and the shock $s$ in a multiplicative way: $\varphi \equiv s \times t$. Denote by $\Phi_c(\cdot)$ the distribution of *productivity* in city $c$. Clearly, even two cities with similar yet different talent compositions will end up having largely overlapping productivity distributions. We then have the following expected wage in city $c$ with average talent $\overline{t}_c$ defined in (4.32):

$$\mathbb{E}w_c(t) = \mathbb{A}L_c^{\epsilon}\int_0^{\overline{t}_c\overline{s}_c}\varphi^a \mathrm{d}\Phi_c(\varphi) = \underbrace{\mathbb{A}\left(\int_0^{\overline{s}_c} s^a \mathrm{d}G_c(s)\right)\overline{t}_c^{\,a}}_{=\mathbb{A}_c(\mathbb{A},\overline{t}_c,G_c(\cdot))} L_c^{\epsilon}.$$

Clearly, the TFP term $\mathbb{A}_c$ is city specific and a function of sorting and of a city-specific distribution of shocks, and there is a nondegenerate distribution of wages and productivities in all cities. The distribution of productivity of cities endowed with highly talented individuals stochastically dominates the distribution of less talented cities.[46]

Another way to generate incomplete sorting is to assume that agents choose locations on the basis of a random component in their objective function, as in Behrens et al. (2013) or Gaubert (2014). The idea is that the location choices of consumers and firms have a deterministic component (profit or indirect utility) as well as a probabilistic component. Under standard assumptions on the distribution of the probabilistic component—if it

[45] Sorting by skills in the United States increased between 1980 and 2000. Diamond (2013) studies its consequences for welfare inequality.

[46] It may be reasonable to assume that the shocks may be, on average, better in larger cities as the result of various insurance mechanisms, better opportunities, etc. This is an additional force pushing toward sorting through the TFP terms: more talented agents will go to places with better shocks since they stand to gain more from good shocks and to lose less from bad shocks.

follows a type I extreme value distribution—location choice *probabilities* are then of the logit form and allow for incomplete sorting across locations: observationally identical agents need not make the same location decisions. More talented agents will, on average, pick larger cities, but the distribution of types is fuzzy across cities. The same result can be achieved by including a deterministic type-independent "attachment to home" component as in Wrede (2013).

Finally, the foregoing models predict PAM: larger cities host, on average, more talented individuals, and the productivity distribution in larger cities first-order stochastically dominates that in smaller cities. However, some recent empirical evidence documents that the right *and* the left tails for the productivity distributions of French workers (Combes et al., 2012), US workers (Eeckhout et al., 2014), and Japanese firms (Forslid and Okubo, 2014) are both fatter in larger cities. In other words, larger markets seem to attract both the most and the least productive workers and firms. Large cities are thus more unequal since they host a disproportionate share of both highly productive and poorly productive agents. While the empirical evidence on two-way sorting is certainly intriguing and points to the existence of some nontrivial complementarities, existing models of two-way sorting still fall short of providing either theoretically plausible or empirically testable mechanisms.[47] The over representation of the left tail of skills in larger cities could be due to many things, including more generous welfare policies, complementarities between skilled and unskilled workers (e.g., rich households employing unskilled workers for housekeeping and child care activities), greater availability of public housing, effects of migrants, or the presence of public transportation as pointed out by Glaeser et al. (2008). As we argue in the next section, complex general equilibrium effects in the presence of selection effects can generate supermodularity for the upper tail and submodularity for the lower tail of the skill distribution. While the jury is not yet in as to what may drive two-way sorting, we believe that more work is needed in that direction.

#### 4.4.1.6 Sorting when distributions matter (a prelude to selection)

In the simple model in Section 4.4.1.1, individuals make location choices by looking at the sizes and average talent of cities only: a more talented city is a city endowed with more efficiency units of labor per capita. *Per se*, there are no benefits or drawbacks associated with living in a talented city. Yet, there are a number of reasons to believe that the talent composition of a city directly matters for these choices in subtler ways. On the one hand,

[47] Whether or not the patterns in the data are due to "two-way sorting" or "sorting and selection" is a priori unclear, as we will emphasize in the next section. There may be one-way sorting—larger markets attract more able agents—but selection afterward fails a certain share of them. Those agents end up as low-productivity ones, a pattern that we see in the data.

locating in a city with more talented entrepreneurs may provide a number of upsides, such as access to cheaper intermediates or higher wages for workers. It may also allow more productive interactions among workers, who learn from each other, especially when the quality of learning depends on the talent of the other agents (Davis and Dingel, 2013). Locating in a place with many talented people may, on the other hand, also have its downsides. Most notably, it toughens up competition since any agent has to compete against more numerous and more talented rivals. Whatever the net effect of the pros and cons, it should be clear that, in general, the location decision of any agent is at least partly based on where other agents go—that is, sorting is endogenous to the whole distribution of talent across cities. Sorting when the whole distribution of talent matters is formalized in both Behrens et al. (2014a) and Davis and Dingel (2013). Behrens et al. (2014a) consider that agents sort across cities on the basis of their talent. As in Section 4.4.1.5, productivity $\varphi$ is the product of "talent" and "luck." Agents who are productive enough—their productivity exceeds some endogenous city-specific *selection cutoff* $\underline{\varphi}_c$—become entrepreneurs and produce local intermediates that are assembled at the city level by some competitive final sector using a CES aggregator. They earn profits $\pi_c(\varphi)$. The remaining agents become workers and supply $\varphi^a$ units of efficient labor, as in our simple model, and earn $w_c\varphi^a \leq \pi_c(\varphi)$. In that context, wages and per capita output in city $c$ are, respectively, given by

$$w_c = \frac{1}{1+\epsilon}\left(\int_{\underline{\varphi}_c}^{\infty} \varphi^{\frac{1}{\epsilon}} \mathrm{d}\Phi_c(\varphi)\right)^{\epsilon} L_c^{\epsilon} \quad \text{and} \quad y_c = \underbrace{\left(\int_{\underline{\varphi}_c}^{\infty} \varphi^{\frac{1}{\epsilon}} \mathrm{d}\Phi_c(\varphi)\right)^{\epsilon}\left(\int_0^{\underline{\varphi}_c} \varphi^{a} \mathrm{d}\Phi_c(\varphi)\right)}_{=\mathbb{A}_c(\underline{\varphi}_c, \Phi_c)} L_c^{\epsilon}, \tag{4.40}$$

where $\Phi_c(\cdot)$ is the city-specific productivity distribution. Observe that the TFP term $\mathbb{A}_c$ is endogenous and depends on sorting (via the productivity distribution $\Phi_c$) and selection (via the cutoff $\underline{\varphi}_c$). The same holds true for wages. This affects the location decisions of heterogeneous agents in nontrivial ways. In the model of Behrens et al. (2014a), the random shocks $s$ occur after a city has been chosen. Individuals' location decisions are thus based on the expected utility that an agent with talent $t$ obtains in all cities. For some arbitrary city $c$, this expected utility is given by

$$\mathbb{E}u_c(t) = \int_0^{\bar{s}_c} \max\{\pi_c(st), w_c(st)^a\} \mathrm{d}G_c(s) - L_c^{\gamma}.$$

It should be clear from the foregoing expression that a simple single-crossing property $\frac{\partial^2 \mathbb{E}u_c}{\partial t \partial L_c}(t) > 0$ need not generally hold. The reason is that both the selection cutoff $\underline{\varphi}_c$ and the whole productivity distribution $\Phi_c(\cdot)$ depend on the city size $L_c$ in general equilibrium. As shown in Section 4.4.2, it is generally not possible to assess whether larger

markets have tougher selection ($\partial \underline{\varphi}_c / \partial L_c > 0$) or not. Thus, it is also *a priori* not possible to make clear statements about sorting: PAM does not hold in general.

Another way in which the talent composition of a city may matter for sorting is when there are learning externalities. Consider the following simplified variant of the model of Davis and Dingel (2013). There are two types of workers. The first type produces nontradable goods under constant returns to scale and no externalities. The second type produces some costlessly traded good. Productivity in that sector is subject to learning externalities. Each worker has $t$ units of efficient labor, which can be used either for work or for learning from others. In equilibrium, workers with $t \geq \underline{t}_c$ engage in the production of traded goods in city $c$, whereas the others produce nontraded goods. In other words, the model features occupational selection. Let $\beta \in (0,1)$ denote the share of time a worker devotes to learning (this is a choice variable). The output of a type $t$ worker in city $c$ employed in the traded sector is given by[48]

$$y_c(t) = (\beta t)^{\alpha_c} [(1-\beta) t \mathbb{L}_c]^{1-\alpha_c}, \tag{4.41}$$

where the first part is the output from allocating time to work, and where the second part is the productivity-enhancing effect of learning. Here, $\alpha_c \in (1/3, 1/2)$ is a city-specific parameter that subsumes how important learning is for an agent's productivity. Expression (4.41) reveals the basic force pushing toward ability sorting: more talented agents benefit more from larger learning externalities.

Maximizing (4.41) with respect to $\beta$ yields $\beta^* = \frac{\alpha_c}{1-2\alpha_c}$, which increases with $\alpha_c$ and is independent of talent.[49] The learning externality, $\mathbb{L}_c$, depends on the time that all agents in the city allocate to that activity (a scale effect), and to the average talent of agents in the city (a composition effect). Let us assume that

$$\mathbb{L}_c = \mathbb{I}_c^{\epsilon} \cdot \bar{t}_c, \quad \text{where} \quad \mathbb{I}_c = L_c \int_{t \geq \underline{t}_c} (1-\beta_c) \mathrm{d}F_c(t) \quad \text{and} \quad \bar{t}_c = \frac{1}{1 - F_c(\underline{t}_c)} \int_{t \geq \underline{t}_c} t \mathrm{d}F_c(t) \tag{4.42}$$

are the scale and the composition effects, respectively. The former effect can be computed as $\mathbb{I}_c = L_c \frac{1-3\alpha_c}{1-2\alpha_c} [1 - F_c(\underline{t}_c)]$ and implies that there is greater potential for spillovers when more agents engage in learning. The second effect implies that the quality of learning increases with the average talent of those who are engaged in learning. Both depend on the selection of agents, as captured by the selection threshold $\underline{t}_c$.

Substituting $\beta^*$ and expressions (4.42) into (4.41), we obtain the average productivity in city $c$:

[48] This specification rules out the "no learning" equilibria that arise in Davis and Dingel (2013). Those equilibria are of no special interest.

[49] Although it may seem reasonable to consider that more talented workers stand to gain more from learning as in Davis and Dingel (2013) and should thus choose higher $\beta$ values in equilibrium, our assumption simplifies the model while still conveying its key insights.

$$y_c = \underbrace{\kappa_c \bar{t}_c^{2-\alpha_c}\left[1 - F_c\left(\underline{t}_c\right)\right]^{\epsilon(1-\alpha_c)+1}}_{=\mathbb{A}_c\left(\underline{t}_c, F_c\right)} L_c^{\epsilon(1-\alpha_c)}, \tag{4.43}$$

where $\kappa_c$ is a term that depends on $\alpha_c$, $\beta$, and $\epsilon$. The TFP term $\mathbb{A}_c$ again depends on the endogenous allocation of talents across cities, $F_c(\cdot)$, and selection into occupations within cities (as captured by $\underline{t}_c$). In general, the threshold is itself a function of city size and the distribution of talent across cities. In a nutshell, $\underline{t}_c$, $F_c(\cdot)$, and $L_c$ are simultaneously determined at the city level, and the locational equilibrium condition, whereby each agent picks his or her preferred location, must hold. Note the similarity between (4.40) and (4.43). Both models predict that sorting and selection interact to determine the productivity advantage of cities. We return to this point below.

Although the sorting of workers across cities has attracted the most attention, a growing literature looks at the sorting of firms (see, e.g., Baldwin and Okubo, 2006; Forslid and Okubo, 2014; Nocke, 2006; Okubo et al., 2010). In a subnational context, we can think about the sorting of firms in the same way as we think about the sorting of entrepreneurs since it is fair to say that most firms move with the people running them.[50] Gaubert (2014) assumes that a firm's realized productivity is given by $\psi(t,L_c)$, where $t$ is the firm's intrinsic productivity. The latter interacts, via $\psi$, with agglomeration economies with city size $L_c$ as a proxy. With use of a simple single-sector variant of Gaubert's multi-industry CES model, the profit of a firm with productivity $t$ is given by

$$\pi_c(t) = \mathbb{A}_c \mathbb{P}_c^{\sigma-1}\left(\frac{\psi(t, L_c)}{w_c}\right)^{\sigma-1}, \tag{4.44}$$

where $\mathbb{A}_c$ is a city-specific TFP shifter, $\mathbb{P}_c$ is the city-specific CES price aggregator, $w_c$ is the city-specific wage, and $\sigma > 1$ is the demand elasticity. As can be seen from (4.44), the firm-level productivity $t$ interacts with city size $L_c$ both directly, via the reduced-form function $\psi$, and indirectly via the citywide variables $\mathbb{A}_c$, $\mathbb{P}_c$, and $w_c$. Taking logarithms of (4.44) and differentiating, and noting that none of the citywide variables $\mathbb{A}_c$, $\mathbb{P}_c$, and $w_c$ depend on a firm's individual $t$, we see that the profit function is *log-supermodular* in $t$ and $L_c$ if and only if $\psi$ is log-supermodular:

[50] Empirical evidence suggests that the bulk of the spatial differences in wages is due to the sorting of workers (Combes et al., 2008), with only a minor role for the sorting of firms by size and productivity (Mion and Naticchioni, 2009). Furthermore, it is difficult to talk about the sorting of firms since, for example, less than 5% of firms relocate in France over a 4-year period (Duranton and Puga, 2001). Figures for other countries are fairly similar, and most moves are short distance moves within the same metro area. Entry and exit dynamics thus drive observed patterns, and those are largely due to selection effects.

$$\frac{\partial^2 \ln \pi_c(t)}{\partial L_c \partial t} > 0 \quad \Leftrightarrow \quad \frac{\partial^2 \ln \psi(t, L_c)}{\partial L_c \partial t} > 0.$$

In words, the profit function inherits the log-supermodularity of the reduced-form productivity function $\psi$, which then implies that more productive firms sort into larger cities.

Four comments are in order. First, this sorting result generically holds only if profits are log-linear functions of citywide aggregates and $\psi$. The latter is the case with CES preferences. Relaxing CES preferences implies that individual profit is generically not multiplicatively separable in $\psi$ and $L_c$; in that case, log-supermodularity of $\psi$ is neither necessary nor sufficient to generate log-supermodularity of $\pi$. Second, log-linearity of profits implies that only the direct interactions between $t$ and $L_c$ matter for the sorting of firms. If we relax the (relatively strong) assumption of log-supermodularity of $\psi$, the model by Gaubert (2014) would also be a model of sorting where the (endogenous) productivity distribution of cities influences location choices in a nontrivial way. As such, it would be extremely hard to solve as we argue in the next subsection. Third, with proper microeconomic foundations for sorting and selection (more on this below), it is not clear at all that $\psi$ is log-supermodular in $t$ and $L_c$ in equilibrium. Fourth, in general equilibrium, the indirect interactions of city size via $\mathbb{P}_c$ and $w_c$ with the individual $t$ may suffice to induce sorting. For example, in the model with an inelastic housing stock as in Helpman (1998), $w(L_c)$ is an increasing function of $L_c$ to compensate mobile workers for higher housing costs. This has opposite effects on profits (higher costs reduce profits, but there are citywide income effects) which may make larger cities more profitable for more productive agents and thereby induce sorting. How these general equilibrium effects influence occupational choice and interact with sorting is the focus of the next subsection.

### 4.4.2 Selection

We now touch upon an issue that has rightly started attracting attention in recent years: selection. Before proceeding, it is useful to clarify the terminology. We can think of two types of selection: *survival selection* and *occupational selection*. Survival selection refers to a *stochastic selection* of the Hopenhayn–Melitz type where entrants have to pay some sunk entry cost, then discover their productivity, and finally decide whether or not to stay in the market (Hopenhayn, 1992; Melitz, 2003; Melitz and Ottaviano, 2008; Zhelobodko et al., 2012). Occupational selection refers to a *deterministic selection* where agents decide whether to run firms or to be workers, depending on their talent (Lucas, 1978).[51] For

[51] In a spatial context, the former has been investigated by Ottaviano (2012), Behrens et al. (2014b), and Behrens and Robert-Nicoud (2014b). The latter has been analyzed by Davis and Dingel (2013), Behrens et al. (2014a), and Behrens et al. (2014c).

simplicity, we deal only with occupational selection in what follows.[52] The selection cutoff $t_c$ for talent in city $c$ then determines how agents are split among different occupational groups (firms or entrepreneurs vs. workers).

Our aim is not to provide a full-fledged model of selection, but rather to distill some key insights. Our emphasis is on the interactions between selection, sorting, and agglomeration. We show in this section that selection and sorting are causally linked, observationally equivalent, and, therefore empirically very difficult to disentangle (Combes et al., 2012). We also show that the impact of market size on selection is generally ambiguous in economic models—that is, it is unclear whether larger markets have more or fewer firms (entrepreneurs) and whether market size is associated with a procompetitive effect. This result is largely due to the general equilibrium interactions between selection, sorting, and agglomeration.

#### 4.4.2.1 A simple model

While sorting can be studied under fairly general assumptions, studying selection requires imposing more structure on the model. More precisely, we need a model in which the *relative position* of an agent—as compared with the other agents in the market—matters. Models of imperfect competition with heterogeneous agents usually satisfy that requirement. Selection can thus be conveniently studied in general equilibrium models of monopolistic competition with heterogeneity, where the payoff to one agent depends on various characteristics such as market size, the skill composition of the market, and the number of competitors. Developing a full model is beyond the scope of this chapter, but a simple reduced-form version will allow us to highlight the key issues at hand.

Consider a set of heterogeneous producers (entrepreneurs) who produce differentiated varieties of some nontraded consumption good or service in city $c$. We denote by $F_c(\cdot)$ the cumulative distribution of talent in city $c$, with support $[\underline{t}_c, \overline{t}_c]$. To make our point clearly, we take that distribution, and especially $\overline{t}_c$, *as given* here—that is, we ignore sorting across cities. The reason is that sorting and selection are difficult to analyze jointly. We discuss the difficulties of allowing for an endogenous talent distribution $F_c(\cdot)$, as well as the interaction of that distribution with selection, later in this section.

Workers earn $w_c$ per efficiency unit of labor, and workers with talent $t$ supply $t^a$ efficiency units. We assume that entrepreneurial productivity increases with talent. We further assume that talented individuals have a comparative advantage in becoming entrepreneurs (this requires entrepreneurial earnings to increase with $t$ at a rate higher than $a$), so the more talented agents (with $t > t_c$) operate firms as entrepreneurs in

[52] See Melitz and Redding (2014) for a recent review of survival selection in international trade. Mrázová and Neary (2012) provide additional details on selection effects in models with heterogeneous firms.

equilibrium. We refer to $t_c$ as the *occupational selection cutoff* (or cutoff, for short). An entrepreneur with talent $t$ hires $1/t$ efficiency units of labor to produce a unit of output. Entrepreneurs maximizes profits, which we assume are given by

$$\pi_c(t) = \left(p_c(t) - \frac{w_c}{L_c^\epsilon t}\right) L_c x_c(t), \tag{4.45}$$

where $p_c(t)$ is the price of the variety sold by the entrepreneurs, $L_c^\epsilon$ is a reduced-form agglomeration externality, and $L_c x_c(t)$ is the total demand faced by the entrepreneur in city $c$, $x_c(t)$ being the per capita demand.[53] Observe from expression (4.45) the complementarity between entrepreneurial talent, $t$, and the agglomeration externality, $L_c^\epsilon$. As argued before, this is a basic force pushing toward sorting along skills into larger cities. However, in the presence of selection, things are more complicated since profits depend in a nontrivial way on market size in general equilibrium. As shown in the next section, the complementarity is also a basic force that dilates the income distribution of entrepreneurs and, therefore, leads to larger income inequality in bigger cities.

Maximizing profits (4.45) with respect to prices yields the standard condition

$$p_c(t) = \frac{\mathcal{E}_{x,p}}{\mathcal{E}_{x,p} - 1} \frac{w_c}{L_c^\epsilon t}, \tag{4.46}$$

where $\mathcal{E}_{x,p} = 1/r(x_c(t))$ is the price elasticity of per capita demand $x_c(t)$, which can be expressed using the "relative love for variety" (RLV), $r(\cdot)$ (Zhelobodko et al., 2012).[54] The profit of an agent who produces a variety with talent $t \geq t_c$ located in a city of size $L_c$, is then given by

$$\pi_c(t) = \underbrace{\frac{r(x_c(t))}{1 - r(x_c(t))} \frac{w_c}{t}}_{=\mu(t, t_c, L_c)} L_c^{1-\epsilon} x_t, \tag{4.47}$$

where $\mu(t,t_c,L_c)$ denotes the *profit margin* of a type $t$ agent in a city with cutoff $t_c$ and size $L_c$.

The set of entrepreneurs who produce differentiated varieties is endogenously determined by the cutoff $t_c$. More formally, agents self-select into occupations (entrepreneurs

[53] For simplicity, we assume that aggregate demand $X_c(t) = L_c x_c(t)$. This will hold true in quasi-linear settings or when preferences are such that aggregate demand depends on some summary statistic (a "generalized Lagrange multiplier"). The latter property amounts to imposing some form of quasi separablility on the inverse of the subutility function as in Behrens and Murata (2007).

[54] In additively separable models, where utility is given by $U = \int u(x_t) \mathrm{d}F_c(t)$, we have $\mathcal{E}_{x,p} = 1/r(x_t)$, where $r(x) = -xu''(x)/u'(x) \in (0,1)$. Condition (4.46) links the firms' markups solely to the properties of the subutility function $u$ (via the RLV). The way that market size affects selection crucially depends on the properties of $r(\cdot)$ and, therefore, on the properties of preferences. Note that $r(\cdot)$ is a function of individual consumption $x_t$ and that it will, in general, be neither a constant nor a monotonic function.

vs. workers) on the basis of the maximum income they can secure. The *selection condition* that pins down the marginal entrepreneur is as follows:

$$\pi_c(t_c) - w_c t_c^a L_c^{\xi} = 0, \tag{4.48}$$

where $L_c^{\xi}$ is an agglomeration externality that makes workers more productive (increases their effective labor). In words, the marginal entrepreneur earns profits equal to the wage he or she could secure as a worker, whereas all agents with talent $t$ such that $\pi_c(t) > w_c t^a L_c^{\xi}$ choose to become entrepreneurs and the others become workers.

The key questions to be addressed are the following. What is the impact of city size $L_c$ on the occupational structure via $t_c$, and how does the talent composition of the city, $F_c(\cdot)$, and various agglomeration externalities, interact with selection? We look at the distribution of incomes within and across groups in the next section.

#### 4.4.2.2 CES illustration

To keep things simple, let us start with the well-known case of CES preferences: $u(x) = x^{\rho}$. In that case $r(x_c(t)) = 1 - \rho$ is constant and independent of individual consumption (and thus of city size). Aggregate CES demand can be expressed as $L_c x_c(t) = L_c[\mathbb{A}_c/p_c(t)]^{1/(1-\rho)}$, where $\mathbb{A}_c$ is some city-specific market aggregate that depends on the distribution of income in the city but that is taken as given by each entrepreneur. From (4.46), we have constant markup pricing: $p_c(t) = w_c/(\rho L_c^{\epsilon} t)$.

Plugging $x_c(t)$ and $p_c(t)$ into profits yields

$$\pi_c(t) = \rho^{\frac{\rho}{1-\rho}}(1-\rho) L_c^{1+\epsilon\frac{\rho}{1-\rho}} \mathbb{A}_c^{\frac{1}{1-\rho}} \left(\frac{w_c}{t}\right)^{\frac{\rho}{\rho-1}}.$$

The occupational selection condition $\pi_c(t_c) = w_c t_c^a L_c^{\xi}$ can then be written as

$$L_c^{1+\epsilon\frac{\rho}{1-\rho}-\xi} \left(\frac{\mathbb{A}_c}{w_c}\right)^{\frac{1}{1-\rho}} = t_c^{a-\frac{\rho}{1-\rho}} \rho^{\frac{\rho}{\rho-1}} \frac{1}{1-\rho}. \tag{4.49}$$

In general equilibrium, the term $\mathbb{A}_c/w_c$ is pinned down by the citywide market clearing condition. Consider the labor market clearing condition: agents who do not become entrepreneurs are workers who will be hired by the entrepreneurs. That condition is given by

$$\int_{\underline{t}_c}^{t_c} t^a L_c^{\xi} \mathrm{d}F_c(t) = \int_{t_c}^{\bar{t}_c} \frac{L_c x_c(t)}{L_c^{\epsilon} t} \mathrm{d}F_c(t). \tag{4.50}$$

Inserting the expression $L_c x_c(t) = L_c(\mathbb{A}_c/p_c(t))^{1/(1-\rho)}$ and simplifying, we obtain the relationship

$$\underbrace{L_c^{1+\epsilon\frac{\rho}{1-\rho}-\xi\left(\frac{\mathbb{A}_c}{w_c}\right)^{\frac{1}{1-\rho}}}\rho^{\frac{1}{1-\rho}}}_{\text{ZPC}}\int_{t_c}^{\bar{t}_c} t^{\frac{\rho}{1-\rho}}\mathrm{d}F_c(t) = \int_{\underline{t}_c}^{t_c} t^a \mathrm{d}F_c(t)$$

$$\Rightarrow \quad t_c^{a-\frac{\rho}{1-\rho}}\frac{\rho}{1-\rho}\int_{t_c}^{\bar{t}_c} t^{\frac{\rho}{1-\rho}}\mathrm{d}F_c(t) = \int_{\underline{t}_c}^{t_c} t^a \mathrm{d}F_c(t),$$

where we have replaced ZPC by the selection condition (4.49). As can be seen, the last condition depends only on the selection cutoff $t_c$. Hence, conditional on the distribution of skills—as captured by the distribution $F_c(\cdot)$ and the support $[\underline{t}_c, \bar{t}_c]$—the selection cutoff $t_c$ is independent of city size, although profits are increasing as the direct effect of $L_c$. The reason is that $\mathbb{A}_c/w_c$ is endogenously determined in the citywide general equilibrium. Any increase in $L_c$ triggers an inverse fall in $\mathbb{A}_c/w_c$, so profits and workers' wages increase in the same proportion in equilibrium. Consequently, city size $L_c$ has no bearing on selection when preferences are of the CES type. Two cities with different sizes but identical skill composition have the same selection cutoff and the same share of entrepreneurs. These findings seem to be in line with the empirical results obtained by Combes et al. (2012) and with the observation that the share of self-employed (a proxy for "entrepreneurship") is independent of city size in the United States (see the left panel in Figure 4.4). Observe though that there is still an effect of sorting on selection: a city $c$ with a better underlying skill distribution than a city $c'$—for example, because $F_c(\cdot)$ first-order stochastically dominates $F_{c'}(\cdot)$—has a larger $t_c$ in equilibrium.

There are two main take-away messages from the foregoing analysis. First, selection effects are inherently a general equilibrium phenomenon. Since large cities (especially MSAs) can be viewed as large economic systems, taking into account general equilibrium effects strikes us as being important. Disregarding those effects may lead to erroneous assessments as to the impacts of market size and talent composition on economic outcomes. Larger cities may be tougher markets, but they are also bigger and richer markets. Taking into account income effects and resource constraints is an important part of the analysis. Second, sorting induces selection. Once sorting has been controlled for, there may or may not be an additional effect of market size on selection. In other words, larger markets may or may not have "tougher selection" (conditional on sorting). The absence of selection effects due to market size in the above example is an artifice of the CES structure where markups are constant (Zhelobodko et al., 2012; Behrens et al., 2014a,c). Yet, selection is still influenced by the talent composition of the city. General equilibrium effects matter.

#### 4.4.2.3 Beyond the CES

The CES structure is arguably an extremely special one. Unfortunately, little is known about selection with more general preferences and demands. What is known is that the selection cutoff $t_c$ usually depends on $L_c$ in general equilibrium, essentially since markups

are variable and a function of $L_c$. Two models where market size matters for the selection of heterogeneous producers are those of Ottaviano (2012) and Behrens and Robert-Nicoud (2014b). They build on the Melitz and Ottaviano (2008) quadratic preferences model to study the relationship between market size and selection in a new economic geography and in a monocentric city setting, respectively. However, sorting along skills is absent in those models. The same holds true for the models building on constant absolute risk aversion preferences (Behrens et al., 2013, 2014b). We are not aware of any model displaying between-city sorting in the presence of nontrivial selection effects.

Behrens et al. (2014c) use general additive preferences in a quasi-linear setting to show that larger markets may have either tougher selection (fewer entrepreneurs) or weaker selection (more entrepreneurs), depending crucially on the properties of preferences.[55] In specifications that many consider as being the normal case (e.g., Vives, 2001), demands become less elastic with consumption levels, so larger cities have tougher selection and fewer entrepreneurs.[56] We suspect that models where larger markets put downward pressure on prices and markups may yield additional effects of selection on sorting. However, to the best of our knowledge, little progress has been made in that direction to date.

#### 4.4.2.4 Selection and sorting

How do selection and sorting interact? In the foregoing, we developed a simple example that shows that sorting induces selection, even when market size does not matter directly. Clearly, selection also has an impact on sorting by changing the payoff structure for agents. The basic question for sorting is always whether larger markets are more profitable places for more talented entrepreneurs. From (4.47), the single-crossing condition can be expressed as follows (recall that we hold the distribution of talent $F_c(\cdot)$ in the city fixed):

$$\frac{\partial^2 \pi_c(t)}{\partial L_c \partial t} = (1-\epsilon)L^{-\epsilon}\left(\frac{\partial x}{\partial t}\mu + \frac{\partial \mu}{\partial t}x\right) + L^{1-\epsilon}\left(\frac{\partial^2 \mu}{\partial t \partial L_c}x + \frac{\partial \mu}{\partial t}\frac{\partial x}{\partial L_c} + \frac{\partial^2 x}{\partial t \partial L_c}\mu + \frac{\partial x}{\partial t}\frac{\partial \mu}{\partial L_c}\right)$$
$$+ \frac{\partial t_c}{\partial L_c}L^{1-\epsilon}\left(\frac{\partial^2 \mu}{\partial t \partial t_c}x + \frac{\partial^2 x}{\partial t \partial t_c}\mu + \frac{\partial \mu}{\partial t}\frac{\partial x}{\partial t_c} + \frac{\partial x}{\partial t}\frac{\partial \mu}{\partial t_c}\right).$$

The first term on the right-hand-side above is the "profit margin effect," which depends on how markups and output change with productivity. First, more productive firms sell larger quantities ($\partial x/\partial t > 0$; Zhelobodko et al., 2012). Second, the effect of productivity on profit margins ($\partial \mu/\partial t$) is generally ambiguous and depends on whether the RLV $r(\cdot)$ is

[55] The impact of a change in city size $L_c$ on the selection cutoff $t_c$—and thus on the share of entrepreneurs and the range of varieties—can go either way, depending on the scale elasticity of $u(\cdot)$ and its RLV.

[56] This class of preferences includes the quasi-linear quadratic model of Melitz and Ottaviano (2008), Ottaviano (2012), and Behrens and Robert-Nicoud (2014b), as well as the constant absolute risk aversion specification of Behrens and Murata (2007) and Behrens et al. (2013, 2014b).

an increasing or decreasing function of productivity. In the CES case, the first term is unambiguously positive, but this is not a general result.

The second term captures the interactions between talent and size that influence the entrepreneur's profits. This term cannot be unambiguously signed either. Whereas the terms $\partial x/\partial t$ and $\partial x/\partial L_c$ are generally positive and negative, respectively, the other terms cannot be signed a priori. For example, per unit profit may increase or decrease with market size and with productivity under reasonable specifications for preferences.

The last term, which we call the *selection effect* ($\partial t_c/\partial L_c$), is also ambiguous. The basic selection term $\partial t_c/\partial L_c$ cannot be signed in general, as we have argued above. The reason is that it depends on many features of the model, in particular on preferences.

To summarize, even in simple models of selection with heterogeneous agents, little can be said a priori on how agents sort across cities in general equilibrium. The main reason for this negative result is that sorting induces selection (via $F_c(\cdot)$ and $L_c$), and that selection changes the payoffs to running firms. Depending on whether those payoffs rise or fall with city size for more talented agents, we may or may not observe PAM sorting across cities. Supermodularity may fail to hold, and analyzing sorting in the absence of supermodularity is a difficult problem. Many equilibria involving nontrivial patterns of sorting may in principle be sustained.

#### 4.4.2.5 Empirical implications and results

Distinguishing between sorting and selection has a strong conceptual basis: it is location choice versus occupation (either as a choice or as an outcome). Distinguishing between the two is hard empirically. The key difficulties are illustrated in Figure 4.10. The arrows labeled (a) in Figure 4.10 show that there is a causal relationship from the talent composition to the size of a city: tougher cities repel agents. Ceteris paribus, people rather want to be "first in the village rather than second in Rome." We refer to this as

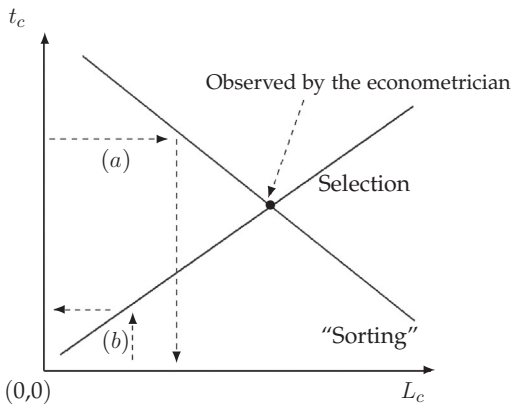


**Figure 4.10** Interactions between sorting and selection.

*sorting*. The arrows labeled (b) in Figure 4.10 show that there is also a causal relationship in the opposite direction, from city size to talent: the talent composition of a city changes with its size. We refer to this as *selection*. The econometrician observes the equilibrium tuples $(t_c,L_c)$ across the urban system. To identify selection, it is necessary to have exogenous shifts in sorting and vice versa. This is difficult, since sorting is itself endogeneous. In the end, distinguishing sorting from selection *ex post* is very difficult since both are observationally equivalent and imply that the productivity composition varies systematically across markets.[57]

The empirical evidence on selection effects to date is mixed. This may be a reflection of their theoretical ambiguity, or of their intrinsic relationship with sorting effects. Di Addario and Vuri (2010) find that the share of entrepreneurs increases with population and employment density in Italian provinces. However, once individual characteristics and education are controlled for, the share of entrepreneurs decreases with market size. The probability of young Italian college graduates being entrepreneurs 3 years after graduation decreases by 2–3 percentage points when the population density of a province doubles. About one-third of this "selection effect" seems to be explained by increased competition among entrepreneurs within industries. However, conditional on survival, successful entrepreneurs in dense provinces reap the benefits of agglomeration: their income elasticity with respect to city size is about 2–3%. Sato et al. (2012) find similar results for Japanese cities. Using survey data, they document that the *ex ante* share of individuals who desire to become entrepreneurs is higher in larger and denser cities: a 10% increase in density increases the share of prospective entrepreneurs by about 1%. It, however, reduces it *ex post* by more than that, so the observed rate of entrepreneurship is lower in denser Japanese cities.

To summarize, the empirical evidence suggests that larger markets have more prospective entrepreneurs (more entrants), but only a smaller share of those entrants survive (tougher selection).[58] Those who do survive in larger markets perform, however, significantly better, implying that denser markets will also be more unequal. Additional evidence for positive selection effects in larger markets in the United States is provided by Syverson (2004, 2007) and by Campbell and Hopenhayn (2005). By contrast, Combes et al. (2012) find no evidence for selection effects—defined as the left truncation of the productivity distribution of firms—when comparing large and small French cities. This finding relies on the identifying assumption that the underlying (unobserved) productivity distributions are the same in small and large cities, and the results are consistent with the CES model.

[57] Okubo et al. (2010) refer to the "spatial selection" of heterogeneous agents when talking about "sorting." That terminology clearly reveals how intrinsically linked sorting and selection really are.

[58] The theoretical predictions of the model of Behrens and Robert-Nicoud (2014b) are consistent with this finding.

## 4.5. INEQUALITY

Heterogeneous agents face heterogeneous outcomes. Hence, it is natural to study issues related to the second moments of the distributions of outcomes. Specifically, one may ask if larger cities are more unequal places than small towns? What mechanisms drive the dispersion of income in large cities? And how does inequality depend on sorting and selection?

We have seen in the previous sections how the size (agglomeration economies) and composition (selection and sorting) of cities influence occupational choices and individual earnings. They thus naturally influence the distribution of earnings within cities. Figure 4.5 reports that large cities are more unequal than smaller ones and suggests that this effect is the joint outcome of composition and size effects (left panel) and an urban premium that varies across the wage distribution (right panel). Indeed, the partial correlation between city size and city Gini coefficient is positive, whether we control for the talent composition of cities (using the share of college graduates as a proxy) or not, and it is larger when we control for it (dashed line) than when we do not (solid line).

Studying the causes and effects of urban inequality is important for at least two reasons. First, earning and wealth inequality seems to be on the rise in many countries (Piketty, 2014), and understanding this rise at the country level requires at least a partial understanding of the positive relationship between city size and earnings inequality. Indeed, Baum-Snow and Pavan (2014) report that at least a quarter of the overall increase in earnings inequality in the United States over the period 1979–2007 is explained by the relatively high growth of earnings inequality in large urban areas.[59] Second, earnings inequality at the local level matters per se: people perceive inequality more strongly when they see it at close range, and cities are not only the locus where inequality materializes, but they are also hosts to mechanisms (sorting and selection) that contribute to changes in that inequality. As such, focusing on cities is of primary interest when designing policies that aim at reducing inequality and its adverse social effects. This is a complex issue because ambitious redistributive policies at the local level may lead to outflow of wealthy taxpayers and an inflow of poor households, a phenomenon that is thought to have contributed to the financial crisis that hit New York City in the 1970s.

Let $y(t,L_c,F_c)$ denote the earnings of an individual with talent $t$ who lives in city $c$ of population size $L_c$ and talent composition $F_c$. It immediately follows that the earnings distribution in any city inherits some properties of its talent distribution, and also that its size and its composition both affect its shape. In this section, we consider two modifications of (4.27) to study how the composition and the size of cities are related to urban inequality as measured by the Gini coefficient of city earnings. We start with sorting.

[59] The measure of earnings inequality in Baum-Snow and Pavan (2014) is the variance of the logarithm of hourly wages.

### 4.5.1 Sorting and urban inequality

Consider first the following slightly generalized version of (4.26):

$$y(t, L_c, F_c) = \mathbb{A}_c t^a L_c^{\epsilon}, \tag{4.51}$$

where $\mathbb{A}_c$ is the usual TFP shifter and $F_c$ is the talent composition of $c$. To fix ideas, assume that the distribution of talent $F_c$ is city specific and log-normal with[60]

$$\ln t \sim \mathcal{N}(\mu_{tc}, \sigma_{tc}^2). \tag{4.52}$$

Assumptions (4.51) and (4.52) together imply that earnings $y$ in city $c$ are also log-normally distributed and the Gini coefficient is a function of the standard deviation of the logarithm of earnings in city $c$ only (Aitchison and Brown, 1963):

$$\text{Gini}(L_c, F_c) = 2\Phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right) - 1, \tag{4.53}$$

where $\Phi(\cdot)$ is the cumulative of the normal distribution and $\sigma_{yc} = a\sigma_{tc}$ is the standard deviation of the logarithm of earnings. It immediately follows from $\Phi'(\cdot) > 0$ and the definition of $\sigma_{yc}$ that earnings inequality increases with talent inequality (a composition effect)—namely,

$$\frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{tc}} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{yc}} \frac{\partial \sigma_{yc}}{\partial \sigma_{tc}} = a\sqrt{2}\phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right) > 0, \tag{4.54}$$

where $\phi(\cdot)$ is the density of the normal distribution, and the second equality follows from the definition of $\sigma_{yc}$. Observe that city size has no direct effect on the Gini coefficient of earnings.[61] This is because agglomeration economies benefit all talents in the same proportion in (4.51).

We know from the previous section that sorting and selection effects imply that the composition of large cities differs systematically from the composition of smaller ones. That is to say, $L_c$ and $F_c$ are jointly determined in general equilibrium. We may thus write

$$\frac{\text{dGini}(L_c, F_c)}{\text{d}L_c} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{tc}} \frac{\text{d}\sigma_{tc}}{\text{d}L_c},$$

where the partial derivative is from (4.54). This simple framework is consistent with the positive partial correlation between the urban Gini coefficient and city size in the left panel in Figure 4.5 if and only if $\text{d}\sigma_{tc}/\text{d}L_c > 0$. If urban talent heterogeneity increases with city size, as in Combes et al. (2012) and Eeckhout et al. (2014), or if large cities

[60] This convenient assumption allows us to parameterize the whole distribution of talents with only two parameters, $\mu_{tc}$ and $\sigma_{tc}$, which simplifies the analysis below.

[61] Note that urban size has a positive effect on the *variance* of earnings, $\text{var}_{yc} = \exp(2\mu_{yc} + \sigma_{yc}^2)\left[\exp(\sigma_{yc}^2) - 1\right]$, where $\mu_{yc} = \mu_{tc} + \ln \mathbb{A}_c + \epsilon \ln L_c$.

attract a disproportionate share of talented workers (so the variance of talents increases with city size), then this inequality holds. Glaeser et al. (2009) report that differences in the skill distribution across US MSAs explain one-third of the variation in Gini coefficients. Variations in the returns to skill may explain up to half of the cross-city variation in income inequality according to the same authors. We turn to this explanation next.

### 4.5.2 Agglomeration and urban inequality

Agglomeration economies affect all talents to the same degree in the previous subsection. This is counterfactual. Using individual data, Wheeler (2001) and Baum-Snow and Pavan (2012) estimate that the skill premium and the returns to experience of US workers increase with city size.[62] A theoretical framework that delivers a positive relationship between city size and the returns to productivity is provided in Davis and Dingel (2013) and Behrens and Robert-Nicoud (2014b). We return to the latter in some detail in Section 4.5.3. To the best of our knowledge, the assignment mechanism similar to Rosen's 1981 "superstar effect" of the former—with markets suitably reinterpreted as urban markets—and the procompetitive effects that skew market shares toward the most productive agents of the latter are the only mechanisms to deliver this theoretical prediction.

To account for this, we now modify (4.26) as follows:

$$y(t, L_c, F_c) = \mathbb{A}_c L_c^{a+\epsilon t}, \quad \text{where} \quad t \sim \mathcal{N}(\mu_t, \sigma_t). \tag{4.55}$$

These expression differ from (4.51) and (4.52) in two ways. First, $y$ is log-supermodular in size and talent in (4.55) but it is only supermodular in (4.51): "simple" supermodularity is not enough to drive complementarity between individual talent and city size. Second, talent is normally distributed and we assume that the composition of talent is constant across cities—that is, $F_c = F$ for all $c$.

As before, our combination of functional forms for earnings and the distribution of talent implies that the distribution of earnings is log-normal and that the city Gini coefficient is given by (4.53). The novelty is that the standard deviation of the logarithm of earnings increases with city size, which is consistent with the empirical finding of Baum-Snow and Pavan (2014):

$$\sigma_{yc} = \sigma_t \epsilon \ln L_c. \tag{4.56}$$

Combining (4.53) and (4.56) implies that urban inequality increases with city size:

[62] See also Baum-Snow and Pavan (2014) for evidence consistent with this mechanism. These authors also report that the positive relationship between urban inequality and city size strengthened between 1979 and 2007, explaining a large fraction of the rise in within-group inequality in the United States.

$$\frac{\partial \text{Gini}(L_c, F_c)}{\partial \ln L_c} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{yc}} \frac{\partial \sigma_{yc}}{\partial \ln L_c} = \sigma_t \epsilon \sqrt{2} \phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right) > 0, \tag{4.57}$$

where the second expression follows from (4.56). From an urban economics perspective, agglomeration economies disproportionately benefit the most talented individuals: the urban premium increases with talent. From a labor economics perspective, and assuming that observed skills are a good approximation for unobserved talents, this result means that the skill premium increases with city size.

Putting the pieces together, we assume finally that city size and individual talent are log-supermodular as in (4.55) and that the talent distribution is city specific as in Section 4.5.1:

$$y(t, L_c, F_c) = \mathbb{A}_c L_c^{a+\epsilon t}, \quad \text{where} \quad t \sim \mathcal{N}(\mu_{tc}, \sigma_{tc}). \tag{4.58}$$

Then the relationship between urban inequality and city size is the sum of the size and composition effects:

$$\frac{\text{dGini}(L_c, F_c)}{\text{d}L_c} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial L_c} + \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{ct}} \frac{\text{d}\sigma_{ct}}{\text{d}L_c} = \sqrt{2} \epsilon \frac{L_c}{\sigma_{tc}} \left(1 + \ln L_c \frac{\text{d} \ln \sigma_{tc}}{\text{d} \ln L_c}\right) \phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right),$$

where the second equality follows from (4.54), (4.57), and (4.58). Both terms are positive if $\text{d}\sigma_{tc}/\text{d}L_c > 0$. The solid line in the left panel in Figure 4.5 reports the empirical counterpart to this expression.[63]

### 4.5.3 Selection and urban inequality

So far, we have allowed urban inequality to depend on the talent composition of cities, city size, or both. There was no selection. In order to study the relationship between selection and urban inequality, we introduce selection in a simple way by imposing the following set of assumptions. Assume first that selection takes a simple form, where the earnings of agents endowed with a talent above some threshold $t_c$ take the functional form in (4.51) and are zero otherwise:

$$y(t, t_c, L_c) = \begin{cases} 0 & \text{if} \quad t \leq t_c \\ \mathbb{A}_c t^a L_c^{\epsilon} & \text{if} \quad t > t_c. \end{cases} \tag{4.59}$$

We refer to the fraction of the population earning zero, $\Phi_c(t_c)$, as the "failure rate" in city $c$. Second, we rule out sorting and assume that the composition of talent is invariant across cities—that is, $F_c = F$, for all $c$—and that talents are log-normally distributed as in

[63] The empirical relationship between urban density and inequality is less clear. Using worker micro data and different measures of earnings inequality from 1970 to 1990—including one that corrects for observable individual characteristics—Wheeler (2004) documents a robust and significantly *negative* association between MSA density and inequality, even when controlling for a number of other factors. This suggests that workers in the bottom income quintile benefit more from *density* than workers in the top income quintile, which maps into smaller earnings inequality in denser cities.

(4.52). Third, we assume that the conditional distribution of talent above the survival selection cutoff $t_c$ is reasonably well approximated by a Pareto distribution with shape parameter $k > 1$:

$$F(t|t \geq t_c) = 1 - \left(\frac{t_c}{t}\right)^k. \tag{4.60}$$

We use this approximation for two related reasons. First, a Pareto distribution is a good approximation of the upper tail of the log-normal distribution in (4.52)—and this is precisely the tail of interest here. Second, the Gini coefficient associated with (4.59) and (4.60) obeys a simple functional form,

$$\text{Gini}(t_c, L_c) = \Phi(t_c) + \frac{1}{2ak-1}[1 - \Phi(t_c)] = \frac{1 + 2(ak-1)\Phi(t_c)}{2ak-1}, \tag{4.61}$$

whereas the Gini coefficient associated with the conditional log-normal $\Phi(t|t \geq t_c)$ does not. The first term in (4.61) is the decomposition of the Gini coefficient into the contributions of the zero-earners and of the earners with a talent above the cutoff $t_c$, respectively. The term $1/(2ak - 1)$ is the Gini coefficient computed among the subpopulation of agents with a talent above $t_c$. Note that this formula for the Gini coefficient is valid only if $ak > 1$ because any Gini coefficient belongs to the unit interval by definition. It follows by inspection of the second term of (4.61) that the Gini coefficient increases with the extent of selection as captured by $\Phi(t_c)$.

We propose a model of urban systems that fits the qualitative properties of this reduced-form model in Behrens and Robert-Nicoud (2014b). Preferences are quasi-linear and quadratic and $t$ is Pareto distributed as in Melitz and Ottaviano (2008). *Ex ante* homogeneous workers locate in cities with possibly heterogeneous $\mathbb{A}_c$. Cities endowed with a large $\mathbb{A}_c$ attract more workers in equilibrium. In turn, large urban markets are more competitive and a smaller proportion of workers self-select into entrepreneurship as a result—that is, the failure rate $\Phi(t_c)$ increases with city size. This is related to our fact 4 (selection) for the United States and is consistent with the empirical findings of Di Addario and Vuri (2010) and Sato et al. (2012) for Italy and Japan, respectively. Recalling that workers are homogeneous prior to making their location decision in Behrens and Robert-Nicoud (2014b), we find that returns to successful entrepreneurs increase with city size. This latter effect is absent in (4.59) but is accounted for in the model we develop in Section 4.5.2.

We can finally compute the relationship between urban inequality and city size in the absence of sorting and agglomeration effects as follows:

$$\frac{\text{dGini}(t_c, L_c)}{\text{d}L_c} = \frac{\partial \text{Gini}(t_c, L_c)}{\partial t_c}\frac{\text{d}t_c}{\text{d}L_c} = 2\phi(t_c)\frac{ak-1}{2ak-1}\frac{\text{d}t_c}{\text{d}L_c},$$

which is positive if and only if $dt_c/dL_c > 0$, and where we have made use of the partial derivative of (4.61) with respect to $t_c$. The interaction between selection and size may thus be conducive to the pattern illustrated in Figure 4.5. Behrens et al. (2014c) show that the equilibrium relationship between urban selection and city size depends on the modeler's choice of the functional forms for preferences. It can even be nonmonotonic in theory, thus suggesting that the impacts of size on inequality could also be nonmonotonic.

## 4.6. CONCLUSIONS

We have extended the canonical urban model along several lines to include heterogeneous workers, firms, and sites. This framework can accommodate all key stylized facts in Section 4.2 and it is useful to investigate what heterogeneity adds to the big picture. Two direct consequences of worker and firm heterogeneity are sorting and selection. These two mechanisms—and their interactions with agglomeration economies and locational fundamentals—shape cities' productivity, income, and skill distributions. We have also argued that more work is needed on the general equilibrium aspects of urban systems with heterogeneous agents. Though difficult, making progress here is key to obtaining a full story about how agents sort across cities, select into occupations, and reap the benefits from and pay the costs of urban size. The first article doing so (albeit in a two-city environment) was that of Davis and Dingel (2013). We use this opportunity to point out a number of avenues along which urban models featuring selection and sorting with heterogeneous agents need to be extended. First, we need models where sorting and nontrivial selection effects interact with citywide income effects and income distributions. This is important if we want to understand better how sorting and selection affect inequalities in cities, and how changes in the urban system influence the macro economy at large. Unfortunately, modeling sorting and selection in the presence of income distributions and nontrivial income effects is a notoriously difficult task. This is probably one explanation for the strong reliance on representative agent models, which, despite their convenience, do not teach us much when it comes to sorting, selection, and inequality. A deeper understanding of the interactions between selection and sorting should also allow us to think better about empirical strategies aimed at disentangling them.

Second, in the presence of heterogeneous agents, the *within-city* allocation of those agents becomes an interesting topic to explore. How do agents organize themselves in cities, and how does heterogeneity across and within cities interact to shape the outcomes in the urban system? There is a large literature on the internal structure of cities, but that literature typically deals with representative agents and is only interested in the implications of city structure for agglomeration economies, land rents, and land use (Beckman,

1976; Fujita and Ogawa, 1982; Lucas and Rossi-Hansberg, 2002; Mossay and Picard, 2011). Extending that literature to include heterogeneous agents seems important to us. For example, if agents sort themselves in specific ways across cities—so that richer agents compete more fiercely for good locations and pay higher land rents—real income inequality in cities may be very different from nominal income inequality. The same holds true for different cities in the urban system, and understanding how heterogeneous agents allocated themselves across and within cities is key to understanding the income and inequality patterns we observe. Davis and Dingel (2014) provide a first step in that direction.

Third, heterogeneous firms and workers do not really interact in urban models. Yet, there is a long tradition in labor economics that deals with that interaction (see, e.g., Abowd et al., 1999). There is also a growing literature in international trade that investigates the consequences of the matching between heterogeneous firms and workers (Helpman et al., 2010). Applying firm-worker matching models to an urban context seems like a natural extension, and may serve to understand better a number of patterns we see in the data. For example, Mion and Naticchioni (2009) use matched employer–employee data for Italy and interpret their findings as evidence for assortative matching between firms and workers.[64] Yet, this assortative matching is stronger in smaller and less dense markets, thus suggesting that matching quality is less important in bigger and denser markets. Theory has, to the best of our knowledge, not much to say about those patterns, and models with heterogeneous workers and firms are obviously required to make progress in that direction.

Lastly, the attentive reader will have noticed that our models depart from the canonical framework of Henderson (1974) by not including transportation or trade costs, so the *relative location* of cities is irrelevant. Multicity trade models with heterogeneous mobile agents are difficult to analyze, yet progress needs to be made in that direction to understand better spatial patterns, intercity trade flows, and the evolution of the urban system in a globalizing world. In a nutshell, we need to get away from models where trade is either prohibitively costly or free. We need to bring back space into urban economic theory, just as international trade brought back space in the 1990s. The time is ripe for *new urban economics* featuring heterogeneity and transportation costs in urban systems.

## ACKNOWLEDGMENTS

We thank Bob Helsley for his input during the early stages of the project. Bob should have been part of this venture but was unfortunately kept busy by other obligations. We further thank our discussant, Don Davis, and the editors Gilles Duranton, Vernon Henderson, and Will Strange for extremely valuable comments and suggestions. Théophile Bougna provided excellent research assistance. K. B. and R. -N. gratefully acknowledge financial support from the CRC Program of the Social Sciences and Humanities Research Council of Canada for the funding of the Canada Research Chair in Regional Impacts of Globalization.

[64] The PAM between firms and workers, or its absence, is a difficult and still open issue in labor economics.

## REFERENCES


Abdel-Rahman, H.M., 1996. When do cities specialize in production? Reg. Sci. Urban Econ. 26, 1–22.

Abdel-Rahman, H.M., Anas, A., 2004. Theories of systems of cities. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, North-Holland, pp. 2293–2339.

Abdel-Rahman, H.M., Fujita, M., 1993. Specialization and diversification in a system of cities. J. Urban Econ. 3, 189–222.

Abowd, J.M., Kramarz, F., Margolis, D.N., 1999. High-wage workers and highwage firms. Econometrica 67, 251–333.

Aitchison, J., Brown, J.A.C., 1963. The Lognormal Distribution. Cambridge Univ. Press, Cambridge, UK.

Albouy, D., Seegert, N., 2012. The Optimal Population Distribution Across Cities and the Private-SocialWedge. Univ. of Michigan, processed.

Albouy, D., Behrens, K., Robert-Nicoud, F.L., Seegert, N., 2015. Are cities too big? Optimal city size and the Henry George theorem revisited, in progress.

Arthur, W.B., 1994. Increasing Returns and Path Dependence in the Economy. University of Michigan Press, Ann Arbor, MI.

Bacolod, M., Blum, B.S., Strange, W.C., 2009a. Skills in the city. J. Urban Econ. 65, 136–153.

Bacolod, M., Blum, B.S., Strange, W.C., 2009b. Urban interactions: soft skills vs. specialization. J. Econ. Geogr. 9, 227–262.

Bacolod, M., Blum, B.S., Strange, W.C., 2010. Elements of skill: traits, intelligences, and agglomeration. J. Reg. Sci. 50, 245–280.

Baldwin, R.E., Okubo, T., 2006. Heterogeneous firms, agglomeration and economic geography: spatial selection and sorting. J. Econ. Geogr. 6, 323–346.

Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage gap. Rev. Econ. Stud. 79, 88–127.

Baum-Snow, N., Pavan, R., 2014. Inequality and city size. Rev. Econ. Stat. 95, 1535–1548.

Becker, G.S., Murphy, K.M., 1992. The division of labor, coordination costs, and knowledge. Q. J. Econ. 107, 1137–1160.

Becker, R., Henderson, J.V., 2000a. Intra industry specialization and urban development. In: Huriot, J.M., Thisse, J.F. (Eds.), The Economics of Cities. Cambridge University Press, Cambridge.

Becker, R., Henderson, J.V., 2000b. Political economy of city sizes and formation. J. Urban Econ. 48, 453–484.

Beckman, M.J., 1976. Spatial equilibrium in the dispersed city. In: Papageorgiou, Y.Y. (Ed.), Mathematical Land Use Theory. Lexington Books, Lexington, MA.

Behrens, K., 2007. On the location and lock-in of cities: geography vs transportation technology. Reg. Sci. Urban Econ. 37, 22–45.

Behrens, K., Murata, Y., 2007. General equilibrium models of monopolistic competition: a new approach. J. Econ. Theory 136, 776–787.

Behrens, K., Robert-Nicoud, F.L., 2014a. Equilibrium and optimal urban systems with heterogeneous land, in progress.

Behrens, K., Robert-Nicoud, F.L., 2014b. Survival of the fittest in cities: urbanisation and inequality. Econ. J. 124 (581), 1371–1400.

Behrens, K., Lamorgese, A.R., Ottaviano, G.I.P., Tabuchi, T., 2009. Beyond the home market effect: market size and specialization in a multi-country world. J. Int. Econ. 79, 259–265.

Behrens, K., Mion, G., Murata, Y., Südekum, J., 2013. Spatial frictions. Univ. of Québec at Montréal; Univ. of Surrey; Nihon University; and Univ. of Duisburg-Essen, processed.

Behrens, K., Duranton, G., Robert-Nicoud, F.L., 2014a. Productive cities: sorting, selection and agglomeration. J. Pol. Econ. 122, 507–553.

Behrens, K., Mion, G., Murata, Y., Südekum, J., 2014b. Trade, wages, and productivity. Int. Econ. Rev. (forthcoming).

Behrens, K., Pokrovsky, D., Zhelobodko, E., 2014c. Market size, entrepreneurship, and income inequality. Technical Report, Centre for Economic Policy Research, London, UK Discussion Paper 9831.

Bleakley, H., Lin, J., 2012. Portage and path dependence. Q. J. Econ. 127, 587–644.

Campbell, J.R., Hopenhayn, H.A., 2005. Market size matters. J. Industr. Econ. LIII, 1–25.

Combes, P.P., Gobillon, L., 2015. The empirics of agglomeration economies. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, North-Holland, pp. 247–348.
Combes, P.P., Duranton, G., Gobillon, L., 2008. Spatialwage disparities: sorting matters! J. Urban Econ. 63, 723–742.
Combes, P.P., Duranton, G., Gobillon, L., Puga, D., Roux, S., 2012. The productivity advantages of large cities: distinguishing agglomeration from firm selection. Econometrica 80, 2543–2594.
Combes, P.P., Duranton, G., Gobillon, L., 2014. The Costs of Agglomeration: Land Prices in French Cities. University of Pennsylvania, Wharton School, in progress.
Costinot, A., 2009. An elementary theory of comparative advantage. Econometrica 77, 1165–1192.
Couture, V., 2014. Valuing the Consumption Benefits of Urban Density. University of California Berkeley, processed.
Davis, D.R., Dingel, J.I., 2013. A Spatial Knowledge Economy. Columbia University, processed.
Davis, D.R., Dingel, J.I., 2014. The comparative advantage of cities. NBER Working paper 20602. National Bureau of Economic Research.
Davis, J.C., Henderson, J.V., 2008. The agglomeration of headquarters. Reg. Sci. Urban Econ. 38, 445–460.
Davis, D.R., Weinstein, D.E., 2002. Bones, bombs, and break points: the geography of economic activity. Am. Econ. Rev. 92, 1269–1289.
Dekle, R., Eaton, J., 1999. Agglomeration and land rents: Evidence from the prefectures. J. Urban Econ. 46, 200–214.
Desmet, K., Henderson, J.V., 2015. The geography of development within countries. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, North-Holland, pp. 1457–1517.
Desmet, K., Rappaport, J., 2013. The settlement of the United States, 1800 to 2000: the long transition towards Gibrat's law. Discussion Paper 9353, Centre for Economic Policy Research, London, UK.
Desmet, K., Rossi-Hansberg, E., 2013. Urban accounting and welfare. Am. Econ. Rev. 103, 2296–2327.
Di Addario, S., Vuri, D., 2010. Entrepreneurship and market size: the case of young college graduates in Italy. Labour Econ. 17 (5), 848–858.
Diamond, R., 2013. The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000. Stanford University, processed.
Duranton, G., 2006. Some foundations for zipf 's law: product proliferation and local spillovers. Reg. Sci. Urban Econ. 36, 542–563.
Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. Am. Econ. Rev. 97, 197–221.
Duranton, G., Puga, D., 2000. Diversity and specialisation in cities: why, where and when does it matter? Urban Stud. 37, 533–555.
Duranton, G., Puga, D., 2001. Nursery cities: urban diversity, process innovation, and the life cycle of products. Am. Econ. Rev. 91, 1454–1477.
Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, North-Holland, pp. 2063–2117.
Duranton, G., Puga, D., 2005. From sectoral to functional urban specialisation. J. Urban Econ. 57, 343–370.
Eeckhout, J., 2004. Gibrat's law for (all) cities. Am. Econ. Rev. 94, 1429–1451.
Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. J. Pol. Econ. 122, 554–620.
Ellison, G., Glaeser, E.L., 1999. The geographic concentration of industry: does natural advantage explain agglomeration? Am. Econ. Rev. Pap. Proc. 89, 311–316.
Ellison, G.D., Glaeser, E.L., Kerr, W.R., 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. Am. Econ. Rev. 100, 1195–1213.
Ethier, W., 1982. National and international returns to scale in the modern theory of international trade. Am. Econ. Rev. 72, 389–405.
Forslid, R., Okubo, T., 2014. Spatial relocation with heterogeneous firms and heterogeneous sectors. Reg. Sci. Urban Econ. 46, 42–56.
Fujita, M., 1989. Urban Economic Theory. MIT Press, Cambridge, MA.

Fujita, M., cois Thisse, J.F., 2013. Economics of Agglomeration: Cities, Industrial Location, and Globalization, second ed. Cambridge University Press, Cambrige, MA.
Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. Reg. Sci. Urban Econ. 12, 161–196.
Gabaix, X., 1999. Zipf's law for cities: an explanation. Q. J. Econ. 114, 739–767.
Gabaix, X., Ibragimov, R., 2011. Rank-1/2: a simple way to improve the OLS estimation of tail exponents. J. Bus. Econ. Stat. 29, 24–39.
Gabaix, X., Ioannides, Y.M., 2004. The evolution of city size distributions. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, North-Holland, pp. 2341–2378.
Gaubert, C., 2014. Firm Sorting and Agglomeration. Princeton University, processed.
Glaeser, E.L., 2008. Cities, Agglomeration, and Spatial Equilibrium. Oxford University Press, Oxford, UK.
Glaeser, E.L., Gottlieb, J.D., 2009. The wealth of cities: agglomeration economies and spatial equilibrium in the United States. J. Econ. Liter. 47, 983–1028.
Glaeser, E.L., Kerr, W.R., 2009. Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? J. Econ. Manag. Strateg. 18, 623–663.
Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. J. Urban Econ. 63, 1–24.
Glaeser, E.L., Resseger, M., Tobia, K., 2009. Inequality in cities. J. Reg. Sci. 49 (4), 617–646.
Glaeser, E.L., Kolko, J., Saiz, A., 2001. Consumer city. J. Econ. Geogr. 1, 27–50.
Grossman, G.M., 2013. Heterogeneous workers and international trade. Rev. World Econ. 149, 211–245.
Helpman, E., 1998. The size of regions. In: Pines, D., Sadka, E., Zilcha, I. (Eds.), Topics in Public Economics. Cambridge University Press, Cambridge, UK, pp. 33–54.
Helpman, E., Itskhoki, O., Redding, S.J., 2010. Inequality and unemployment in a global economy. Econometrica 78, 1239–1283.
Helsley, R.W., Strange, W.C., 2011. Entrepreneurs and cities: complexity, thickness, and balance. Reg. Sci. Urban Econ. 44, 550–559.
Helsley, R.W., Strange, W.C., 2014. Coagglomeration, clusters, and the scale and composition of cities. J. Pol. Econ. 122 (5), 1064–1093.
Henderson, J.V., 1974. The sizes and types of cities. Am. Econ. Rev. 64, 640–656.
Henderson, J.V., 1988. Urban Development: Theory, Fact and Illusion. Oxford University Press, New York, NY.
Henderson, J.V., 1997. Medium size cities. Reg. Sci. Urban Econ. 27, 583–612.
Henderson, J.V., Ono, Y., 2008. Where do manufacturing firms locate their headquarters? J. Urban Econ. 63, 431–450.
Henderson, J.V., Venables, A.J., 2009. The dynamics of city formation. Rev. Econ. Dyn. 12, 233–254.
Hendricks, L., 2011. The skill composition of US cities. Int. Econ. Rev. 52, 1–32.
Holmes, T.J., Sieg, H., 2014. Structural estimation in urban economics. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, North-Holland.
Holmes, T.J., Stevens, J.J., 2014. An alternative theory of the plant size distribution, with geography and intra- and international trade. J. Pol. Econ. 122 (2), 369–421.
Hopenhayn, H.A., 1992. Entry, exit, and firm dynamics in long run equilibrium. Econometrica 60, 1127–1150.
Hsu, W.T., 2012. Central place theory and city size distribution. Econ. J. 122, 903–922.
Jacobs, J., 1969. The Economy of Cities. Vintage, New York, NY.
Kim, S., 1989. Labor specialization and the extent of the market. J. Pol. Econ. 97, 692–705.
Kline, P., Moretti, E., 2014. People, places, and public policy: some simple welfare economics of local economic development programs. Ann. Rev. Econ. 6 (1), 629–662.
Krugman, P.R., 1980. Scale economies, product differentiation, and the pattern of trade. Am. Econ. Rev. 70, 950–959.
Krugman, P.R., 1991. Increasing returns and economic geography. J. Pol. Econ. 99, 483–499.
Lee, S., 2010. Ability sorting and consumer city. J. Urban Econ. 68, 20–33.
Lee, S., Li, Q., 2013. Uneven landscapes and city size distributions. J. Urban Econ. 78, 19–29.
Lucas Jr., R.E., 1978. On the size distribution of business firms. Bell J. Econ. 9, 508–523.

Lucas Jr., R.E., Rossi-Hansberg, E., 2002. On the internal structure of cities. Econometrica 70, 1445–1476.
Marshall, A., 1890. Principles of Economics, eighth ed. Macmillan and Co., Ltd, London, UK, (1920) edition.
Matano, A., Naticchioni, P., 2012. Wage distribution and the spatial sorting of workers. J. Econ. Geogr. 12, 379–408.
Melitz, M.J., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71, 1695–1725.
Melitz, M.J., Ottaviano, G.I.P., 2008. Market size, trade and productivity. Rev. Econ. Stud. 75, 295–316.
Melitz, M.J., Redding, S.J., 2014. Heterogeneous firms and trade. In: Helpman, E., Gopinath, G., Rogoff, K. (Eds.), Handbook of International Economics, vol. 4. Elsevier, North-Holland, pp. 1–54.
Melo, P.C., Graham, D.J., Noland, R.B., 2009. A meta-analysis of estimates of urban agglomeration economies. Reg. Sci. Urban Econ. 39, 332–342.
Michaels, G., Rauch, F., Redding, S.J., 2012. Urbanization and structural transformation. Q. J. Econ. 127, 535–586.
Mion, G., Naticchioni, P., 2009. The spatial sorting and matching of skills and firms. Can. J. Econ. 42, 28–55.
Moretti, E., 2004. Human capital externalities in cities. In: Henderson, J.V., cois Thisse, J.F. (Eds.), In: Handbook of Regional and Urban Economics, vol. 4. Elsevier, North-Holland, pp. 2243–2291.
Mori, T., Turrini, A., 2005. Skills, agglomeration and segmentation. Eur. Econ. Rev. 49, 201–225.
Mori, T., Nishikimi, K., Smith, T.E., 2008. The number-average size rule: a new empirical relationship between industrial location and city size. J. Reg. Sci. 48, 165–211.
Mossay, P., Picard, P.M., 2011. On spatial equilibria in a social interaction model. J. Econ. Theory 146, 2455–2477.
Mrázová, M., Neary, J.P., 2013. Selection Effects with Heterogeneous Firms. University of Surrey and Oxford University, processed.
Murata, Y., 2003. Product diversity, taste heterogeneity, and geographic distribution of economic activities: market vs. non-market interactions. J. Urban Econ. 53, 126–144.
Nocke, V., 2006. A gap for me: entrepreneurs and entry. J. Eur. Econ. Assoc. 4, 929–956.
Okubo, T., Picard, P.M., cois Thisse, J.F., 2010. The spatial selection of heterogeneous firms. J. Int. Econ. 82, 230–237.
Ossa, R., 2013. A gold rush theory of economic development. J. Econ. Geogr. 13, 107–117.
Ota, M., Fujita, M., 1993. Communication technologies and spatial organization of multi-unit firms in metropolitan areas. Reg. Sci. Urban Econ. 23, 695–729.
Ottaviano, G.I.P., 2012. Agglomeration, trade, and selection. Reg. Sci. Urban Econ. 42, 987–997.
Piketty, T., 2014. Capital in the 21st Century. Harvard University Press, Cambridge, MA.
Puga, D., 2010. Themagnitude and causes of agglomeration economies. J. Reg. Sci. 50, 203–219.
Redding, S.J., 2012. Goods trade, factormobility and welfare. Technical Report, National Bureau for Economic Research, Cambridge, MA, NBER Discussion Paper.
Rosen, S., 1981. The economics of superstars. Am. Econ. Rev. 71, 845–858.
Rosenthal, S.S., Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., cois Thisse, J.F. (Eds.), In: Handbook of Regional and Urban Economics, vol. 1. Elsevier, North-Holland, pp. 2119–2171.
Rosenthal, S.S., Strange, W.C., 2008a. Agglomeration and hours worked. Rev. Econ. Stat. 90, 105–118.
Rosenthal, S.S., Strange, W.C., 2008b. The attenuation of human capital spillovers. J. Urban Econ. 64, 373–389.
Rossi-Hansberg, E., Wright, M.L.J., 2007. Urban structure and growth. Rev. Econ. Stud. 74, 597–624.
Rossi-Hansberg, E., Sarte, P.D., Owens III, R., 2009. Firm fragmentation and urban patterns. Int. Econ. Rev. 50, 143–186.
Rozenfeld, H.D., Rybski, D., Gabaix, X., Makse, H.A., 2011. The area and population of cities: new insights from a different perspective on cities. Am. Econ. Rev. 101, 2205–2225.
Saiz, A., 2010. The geographic determinants of housing supply. Q. J. Econ. 125, 1253–1296.
Sato, Y., Tabuchi, T., Yamamoto, K., 2012. Market size and entrepreneurship. J. Econ. Geogr. 12, 1139–1166.

Sattinger, M., 1993. Assignments models of the distribution of earnings. J. Econ. Liter. 31, 831–880.
Syverson, C., 2004. Market structure and productivity: a concrete example. J. Pol. Econ. 112, 1181–1222.
Syverson, C., 2007. Prices, spatial competition and heterogeneous producers: an empirical test. J. Ind. Econ. LV. 197–222.
Tabuchi, T., cois Thisse, J.F., 2002. Taste heterogeneity, labor mobility and economic geography. J. Dev. Econ. 69, 155–177.
Venables, A.J., 2011. Productivity in cities: self-selection and sorting. J. Econ. Geogr. 11, 241–251.
Vermeulen, W., 2011. Agglomeration Externalities and Urban Growth Controls. SERB Discussion Paper 0093, Spatial Economics Research Centre, London School of Economics.
Vives, X., 2001. Oligopoly Pricing: Old Ideas and New Tools. MIT Press, Cambridge, MA.
Wheeler, C.H., 2001. Search, sorting, and urban agglomeration. J. Lab. Econ. 19, 879–899.
Wheeler, C.H., 2004. Wage inequality and urban density. J. Econ. Geogr. 4, 421–437.
Wrede, M., 2013. Heterogeneous skills and homogeneous land: segmentation and agglomeration. J. Econ. Geogr. 13, 767–798.
Zhelobodko, E., Kokovin, S., Parenti, M., cois Thisse,, J.F., 2012. Monopolistic competition: beyond the constant elasticity of substitution. Econometrica 80, 2765–2784.