Review of Economic Studies (2007) **74**, 597–624 0034-6527/07/00210597$02.00

Urban Structure and Growth

ESTEBAN ROSSI-HANSBERG
Princeton University

and

MARK L. J. WRIGHT
University of California, Los Angeles

First version received May 2005; *final version accepted October* 2006 (*Eds.*)

Most economic activity occurs in cities. This creates a tension between local increasing returns, implied by the existence of cities, and aggregate constant returns, implied by balanced growth. To address this tension, we develop a general equilibrium theory of economic growth in an urban environment. In our theory, variation in the urban structure through the growth, birth, and death of cities is the margin that eliminates local increasing returns to yield constant returns to scale in the aggregate. We show that, consistent with the data, the theory produces a city size distribution that is well approximated by Zipf's law, but that also displays the observed systematic underrepresentation of both very small and very large cities. Using our model, we show that the dispersion of city sizes is consistent with the dispersion of productivity shocks found in the data.

1. INTRODUCTION

Aggregate economic activity is primarily *urban* economic activity. For example, in the U.S. in 2000, 80% of the population lived in urban agglomerations and they earned around 85% of income. This fact creates a tension. On the one hand, the organization of economic activity in cities is evidence for the presence of scale effects: there are economic rewards to the agglomeration of firms and individuals in a city. On the other hand, scale does not appear to be rewarded in the aggregate, as suggested by the evidence on balanced growth. In this paper, we argue that it is the urban structure—the number and size of cities—that resolves this tension.

We begin by constructing a theory of economic growth in an urban environment. In our theory, the size of cities is determined by the trade-off between agglomeration effects and congestion costs with the strength of these forces implying equilibrium city sizes that vary with the stock of factors and the level of productivity. As the economy expands keeping factor proportions and productivity levels constant, each city operates at the equilibrium size and the economy behaves as if using a constant returns to scale technology by varying the number of cities. In this way, it is the endogenous evolution of the urban structure that produces the linear aggregate production functions necessary for balanced growth in a world with urban scale effects.[1] Hence, the first contribution of this paper is to provide a tractable general equilibrium growth theory that incorporates urban structure.

We then show that this theory is also able to generate a number of well-established empirical regularities about the size distribution of cities. Perhaps the best known of these regularities is that the size distribution of cities is well approximated by a Pareto distribution with coefficient 1, also known as Zipf's law. First discovered by Auerbach (1913), this regularity has since been

1. Specifically, the production set of the aggregate economy is, asymptotically, a convex cone. In both exogenous and endogenous growth models such as Lucas (1988), scale economies at the industry level are transformed into constant returns at the aggregate by assuming linear factor accumulation technologies (see also Jones, 1999).

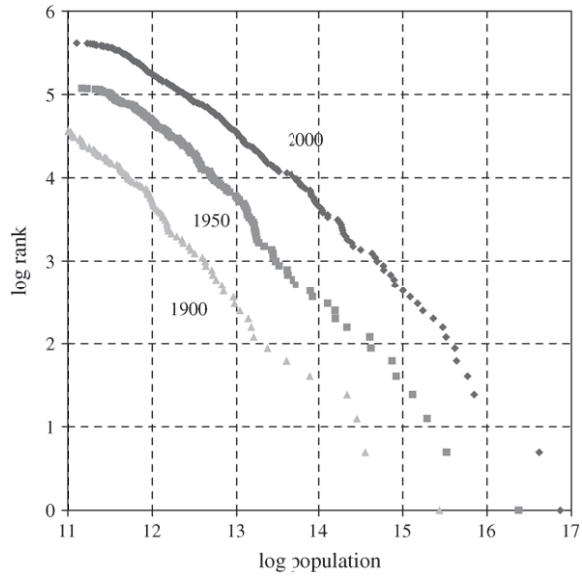

FIGURE 1
Zipf's law for the U.S.

documented using modern data for a wide range of countries by many authors including Rosen and Resnick (1980), Dobkins and Ioannides (2001), Ioannides and Overman (2003), and Soo (2005). We can illustrate this regularity graphically by noting that under a Pareto distribution with coefficient 1, the proportion of cities larger than a given size x is inversely proportional to that size, or $P(\text{City}_i > x) = M/x$ for some constant M. As a result, if Zipf's law holds exactly and we plot $\ln P(\text{City}_i > x)$ against the the natural logarithm of a city's size, we should observe a straight line with a slope of -1. As shown in Figure 1 for the U.S., Zipf's law is a good approximation and indeed appears to be as good a description of the size distribution of cities at the turn of the 21st century as it was at the turn of the 20th century.[2] Furthermore, as illustrated in Figure 2(a) and (b), Zipf's law also appears to be a good description of the size distribution of cities across a broad range of countries today.

Much of the recent empirical work on the size distribution of cities (*e.g.* Eeckhout, 2004 or the survey of Gabaix and Ioannides, 2003) has emphasized a number of systematic and significant deviations from Zipf's law. One of the most robust is the underrepresentation of small cities and the absence of very large ones, relative to Zipf's law, which is illustrated in Figure 1 for the U.S. and again in Figure 2 for a wide range of countries. In the left tail, the plots for all countries appear concave reflecting the underrepresentation of small cities. In the upper tail, where there are less observations, some segments of the plot appear locally convex, especially in the neighbourhood of the capital city, which is often an outlier. However, the tendency for an approximately concave relationship remains, reflecting the absence of very large cities relative to Zipf's law. Figure 2 also displays a second commonly observed deviation from Zipf's law: some countries have a size distribution that is more or less dispersed than that predicted by Zipf's law, which is reflected in flatter or steeper plots of log-rank against log-size.[3]

2. We thank Yannis Ioannides and Linda Dobkins for providing historical data on the U.S. size distribution of cities.

3. Some studies, like Soo (2005), find that the slope of Zipf's relationship is also correlated to income: less developed countries have a tendency towards flatter plots reflecting a more dispersed size distribution than predicted by Zipf's law.

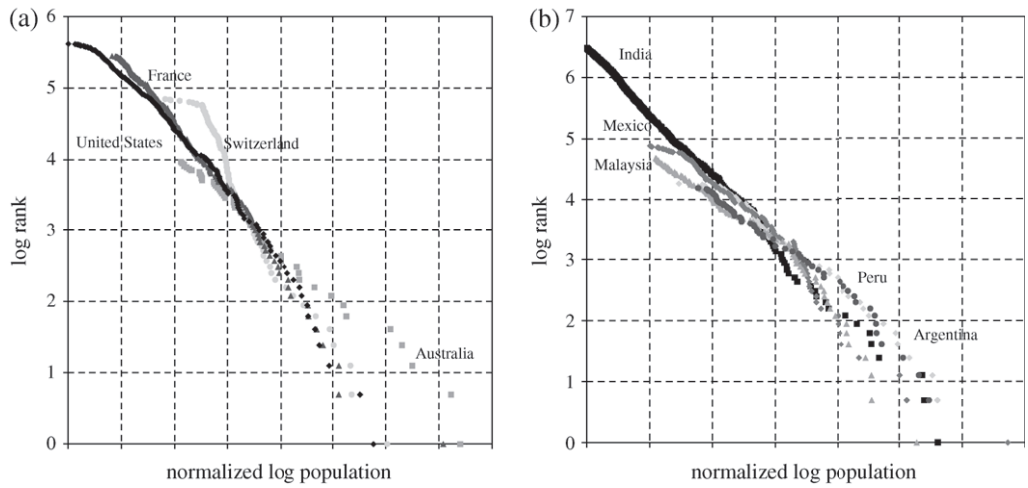

FIGURE 2
(a) Zipf's law: developed countries; (b) Zipf's law: developing countries

Our theory can match all of these facts. In particular, Zipf's law of cities emerges exactly from some special assumptions on our model. Outside these special cases, the city size distribution will tend towards Zipf's law, but will always display relatively thin tails. The overall dispersion of city sizes may be more or less than that predicted by Zipf's law. To see why this is true, note that in our set-up, cities result out of the trade-off between commuting costs and local production externalities in human capital and labour. Industry-specific externalities imply that cities specialize in an industry, and so the evolution of industry productivity shocks and the way they are propagated through the accumulation of industry-specific factors, as summarized by changes in the average product of labour in an industry, drive variations in the urban structure. For example, in response to a positive productivity shock, cities grow, and the number of cities in which that industry operates falls.

Under two polar sets of strong assumptions on technology and the evolution of productivity, this mechanism produces Gibrat's law of cities: the mean and variance of the growth rate of a city are independent of its size. For example, if the only factors of production are labour and human capital both growing at constant rates, the growth rate of the average product of labour, and hence the size of cities, is driven by the growth rate of total factor productivity. Therefore, if shocks are permanent, the growth process of cities is scale independent. Conversely, in an economy in which human capital and labour do not grow and the production function is linear in capital (an AK model), temporary productivity shocks imply permanent changes in the capital stock, the average product of labour, and hence also in the size of cities. For these special cases, we provide a proof that the size distribution of cities converges to Zipf's law. Unlike previous efforts in this regard our approach works through the endogenous variation in the number of cities.

Aside from these polar cases there will be mean reversion in the sizes of cities resulting from the process of factor accumulation. Essentially, industries with small stocks of specific capital operate in small cities. For these industries, diminishing returns to physical capital lead to high rates of return, high incentives to accumulate industry-specific capital, and hence a high growth rate for cities operating in this industry. This logic implies that growth rates decrease with

size, which almost immediately produces the first deviation—the fact that, relative to Zipf's law, both very small and very large cities are underrepresented. Intuitively, because small cities grow faster, and large cities slower, than required to produce Zipf's law, mass is shifted away from the tails of the distribution. The second deviation—variation in the dispersion of the size distributions of cities or variation in Zipf's coefficients—can be accounted for by variation in the volatility of productivity shocks across countries.[4]

Importantly, and unlike previous theoretical studies of the evolution of the size distribution of cities, endogenous city creation and destruction is one of the key mechanisms that generates the results of this paper. This mechanism appears to be very important in practice. For example, Henderson and Wang (2005) show that the margin of city creation and destruction has been very active and volatile over the past four decades, even as the relative size distribution has stayed remarkably constant. In this period, the number of cities has grown by more than 120%, accounting for 26% of the increase in the world urban population. They also show that the S.D. of the growth in the number of cities is more than twice as large as the mean (0·3 and 0·13 respectively). Note also that, for our purposes, these numbers are a lower bound for the activity in the margin of city creation and destruction, as they do not capture the reallocation of industries across existing sites. This is important since our theory accounts for these reallocations as changes in the number of cities operating in a particular industry.

This paper draws from four related literatures. The first is the extensive literature on endogenous growth spawned by Lucas (1988) and Romer (1990). In this literature, as emphasized by Jones (1999), the treatment of scale effects is crucial, as it is the imposition of linearity in the aggregate production technology that is necessary for the existence of balanced growth. Where our paper differs is in its utilization of the urban structure as the vehicle for obtaining this linearity.

A second related literature is the small number of papers on urban growth. Eaton and Eckstein (1997) and Black and Henderson (1999) both present deterministic urban growth models with two types of cities in which, along the balanced growth path, both cities grow at the same rate. Unlike both of these papers, ours focuses on a stochastic environment and introduces a rich industrial structure, which allows us to characterize the evolution of the entire size distribution of cities over time. In addition, both of these papers obtain the linearity of the aggregate production process by assuming knife-edge conditions on production and externality parameters. In contrast, in our theory the urban structure produces this linearity without any further conditions on parameter values.

This paper is related to the large number of papers that propose statistical explanations of Zipf's law. Including most notably Kalecki (1945), Champernowne (1953), Levy and Solomon (1996), Gabaix (1999*a*), Malcai, Biham and Solomon (1999), Blank and Solomon (2000), and Cordoba (2004), these papers have focused on the role of Gibrat's law of city growth in producing a size distribution of cities that satisfies Zipf's law for an economy in which there is a fixed number of cities. Much attention has also been devoted to the role of various frictions, or lower bounds, on city sizes in ensuring a thick lower tail of the distribution of city sizes. In contrast, this paper considers an environment with an endogenous number of cities and establishes scale-independent variation in this extensive margin as a new way of producing Zipf's law.

Finally, the paper is related to the growing economic literature on the size distribution of cities. Gabaix (1999*a*,*b*) and Cordoba (2004) present models in which cities grow as labour migrates between a fixed number of cities in direct response to city amenity, taste, or productivity shocks, and show that they can produce Zipf's law exactly. In contrast, this paper generates the existence of cities endogenously and focuses on the role of factor accumulation in producing

4. This could explain the positive relationship between Zipf coefficients and output, found by Soo (2005), if more developed countries experience less volatile shocks.

some of the robustly documented deviations from Zipf's law. In two recent papers, Eeckhout (2004) and Duranton (2007) present models with fixed numbers of cities that generically produce a size distribution of firms with thinner tails than that predicted by Zipf's law. In contrast, our theory focuses upon the relationship between factor accumulation, productivity shocks, and the urban structure in producing size distributions that either exactly match or approximate Zipf's law, and produces testable implications for the way in which observed size distributions should differ from Zipf's law. In this sense, it is the economic structure of our framework that pushes us away from Zipf's law and allows our theory to rationalize a wider set of phenomena on the size distribution of cities and growth. Importantly, and in contrast to all of these papers, it is endogenous city formation that both eliminates scale effects in growth and provides a novel theory of the size distribution of cities.

The rest of the paper is organized as follows. The next section presents the model. Section 3 derives the main results of the paper on growth, Zipf's law, and deviations from Zipf's law. Section 4 illustrates the results of the model numerically and compares them to data from several countries. Section 5 concludes. The Appendix contains the basic elements of the decentralization and proofs of the main propositions.

2. AN URBAN GROWTH MODEL

Consider an economy in which production occurs at specific locations that we call cities. Firms set up in a city, hiring capital and employing workers. Agglomeration results from a positive production externality on labour and human capital. Agents reside in cities and commute to work. Households are made up of workers who consume, accumulate industry-specific physical capital to be used in each industry, and devote their time to working and learning so as to accumulate industry-specific human capital. We follow Henderson (1974) and postulate the existence of a class of competitive property developers that own each potential city site and compete to attract workers and firms. These property developers serve to internalize the local production externality and guarantee that market outcomes are efficient.[5] Throughout, we assume log-linear preferences and Cobb–Douglas production functions so that we can solve for the growth path of cities in closed form. The advantage of closed-form solutions is that it allows us to make analytical statements about the evolution of individual cities, as well as about the long-run size distribution of cities.

2.1. *Cities*

Our approach to modelling cities follows the classic paper of Henderson (1974) and has been used in the urban growth model of Black and Henderson (1999). We consider a world in which there are a large number of potential city sites. Cities are monocentric, with all production occurring at the single, exogenously given, point, which we refer to as the central business district (CBD). It is assumed that every agent who works at the CBD must reside in the area surrounding the city, so that locations closer to the CBD are more desirable because they involve a shorter commute to work. We assume that the cost of commuting is linear in the distance travelled, and let τ be the cost per mile of commuting. For simplicity we denominate commuting costs solely in terms of the output of the city. However, our results can be generalized to the case of both monetary and time costs (we return to this in Section 3 below).

All agents consume the services of one unit of land per period. In order for agents to be indifferent about where to live in the city, rents differ by the amount of commuting costs, with

5. Since property developers make zero profits in equilibrium, we would obtain the same results if we allowed households to own the land.

rents on the city edge equal to 0. Therefore, in a city of radius $\bar{z}$, rents at a distance z from the centre must be given by $R(z) = \tau(\bar{z} - z)$. Hence, total rents in a city of radius $\bar{z}$ are given by

$$\text{TR} = \int_0^{\bar{z}} 2\pi z R(z) dz = \frac{\pi\tau}{3}\bar{z}^3.$$

Since everyone in the city lives in one unit of land, a city of population $\tilde{N}$ has a radius of $\bar{z} = (\tilde{N}/\pi)^{1/2}$ and so

$$\text{TR} = \frac{\pi\tau}{3}\left(\frac{\tilde{N}}{\pi}\right)^{3/2} = \frac{b}{2}\tilde{N}^{3/2},$$

where $b \equiv 2\pi^{(-1/2)}\tau/3$. Total commuting costs are given by

$$\text{TCC} = \int_0^{\bar{z}} 2\pi z \tau z dz = b\tilde{N}^{3/2},$$

with each resident of the city paying a total of $3b\tilde{N}^{(1/2)}/2$ in terms of rents and commuting costs. Note that both total and average commuting costs are increasing in city population.

2.2. *Firms*

Production occurs in firms that face a constant returns to scale technology. The production of a representative firm in industry j located in an arbitrary city at any point in time t has the Cobb–Douglas form

$$\tilde{A}_{tj} k_{tj}^{\beta_j} h_{tj}^{\alpha_j} (u_{tj} n_{tj})^{1-\alpha_j-\beta_j},$$

where $\tilde{A}_{tj}$ is the total factor productivity of a firm in this particular city given that good j is produced there, k_{tj} is the amount of industry j-specific capital used by that firm, h_{tj} is the amount of human capital, and n_{tj} is the number of workers employed in a firm, each of whom spends a fraction u_{tj} of his or her time at work.

There is a local industry-specific externality in the labour input, so that the productivity of any firm in the city depends upon the number of workers in a city and the amount of human capital they have:

$$\tilde{A}_{tj} = A_{tj} \tilde{H}_{tj}^{\gamma_j} \tilde{N}_{tj}^{\varepsilon_j},$$

where A_{tj} is an industry-specific productivity shock and $\tilde{H}_{tj}$ and $\tilde{N}_{tj}$ represent the total stock of human capital and the total amount of labour in the city. Increasing returns at the city level cause agglomeration in the model. Firms are assumed to be small, taking the size of the externality as given. The industry-specific productivity shock is assumed to be first-order Markov with bounded support, and the distribution of its innovations can vary arbitrarily across industries.

Two comments regarding external effects are in order. First, the external effect is related to human capital and the number of workers, but not to physical capital. This allows us to interpret the external effects as knowledge spillovers. Adding an external effect from physical capital would not change our qualitative results for city dynamics or the shape of the size distribution, although it would change its mean. We also assume that the external effect depends on the number of workers and not on the fraction of time they devote to work, u. This assumption is also for interpretative purposes only, but has no effect on our results. Second, we assumed the externality to be city and industry specific. Therefore, it is efficient for cities to specialize in one industry,

as this allows agents to economize on commuting costs. In the unpublished Appendix, we show how the model can be generalized to produce diversified cities.

The problem of the firm is to hire labour and human and physical capital to maximize profits taking as given the total amount of labour input in the city (and hence the size of the externality term), factor prices, and subsidies. As there are constant returns to scale within the firm, we can treat each city as though it had a representative firm. If we let p_{tj}, w_{tj}, r_{tj}, and s_{tj} be the prices and rental rates of labour and physical and human capital, respectively, written in terms of some numeraire commodity and τ_{tj}^k and τ_{tj}^h be the subsidies to physical and human capital paid by property developers to attract firms to a particular city, then the firm's optimization problem yields

$$w_{tj}/p_{tj} = (1-\alpha_j-\beta_j)\tilde{Y}_{tj}/(u_{tj}\tilde{N}_{tj}), \tag{1}$$

$$(1-\tau_{tj}^k)r_{tj}/p_{tj} = \beta_j\tilde{Y}_{tj}/\tilde{K}_{tj}, \tag{2}$$

$$(1-\tau_{tj}^h)s_{tj}/p_{tj} = \alpha_j\tilde{Y}_{tj}/\tilde{H}_{tj}, \tag{3}$$

where $\tilde{Y}_{tj}$ and $\tilde{K}_{tj}$ denote production and physical capital stock in the city.

It will be convenient for the analysis to divide the original set of J industries into groups with the same basic technology parameters. In particular, within a group, firms in each industry produce using exactly the same technology, but use industry-specific human and physical capital and receive industry-specific productivity shocks. That is, within a group all factor shares, and in general, all parameters, are identical. Across groups, all aspects of the technology may differ. In line with much of the literature, we see this as a natural way of organizing the set of products observed in the economy. Some products are distinguished because they are produced with fundamentally different technologies, while others embody different designs or fulfil different purposes, but are produced with the same *ex ante* technology. This assumption, which ensures the homogeneity of technology within a group, implies that the city evolution process will be the same for all cities within a group. We will then use this to establish the conditions under which Zipf's law holds for each of these groups, before finally aggregating across groups to obtain Zipf's law for the entire economy.

2.3. *Households*

The economy is populated by a unit measure of identical small households. The initial number of people per household is N_0, and we assume that the population of each household grows exogenously at rate g_N. Each household starts with the same strictly positive endowments of industry j-specific physical (K_{0j}) and human (H_{0j}) capital, as well as identical initial financial wealth. The distinction between households and members is useful so that each household can have members working in all industries, which allows the household to diversify all risk. Hence, even though there are multiple industries there is a representative household. As a result, in what follows, we focus on allocations in which all households are treated symmetrically.

Households order preferences over stochastic sequences of the consumption good according to

$$(1-\delta)E_0\left[\sum_{t=0}^{\infty}\delta^t N_t\left(\sum_{j=1}^{J}\ln\left(\frac{C_{tj}}{N_t}\right)\right)\right],$$

where δ is a discount factor that lies strictly between 0 and $1/(1+g_N)$, and C_{tj} is a sequence of state-contingent consumption of each good j. Here E_0 is an expectation operator conditional on information available to the household at time 0.

Capital services in industry j are proportional to the stock of industry j-specific capital, which is accumulated according to the log-linear equation

$$K_{t+1j} = K_{tj}^{\omega_j} X_{tj}^{1-\omega_j}.$$

Investment in industry j, X_j, is assumed to be denominated in terms of that industry's consumption good. The assumption of log-linearity for the capital accumulation equation is introduced so as to allow us to solve for the entire time path of capital accumulation in closed form. This specification was first introduced by Lucas and Prescott (1971) and has since been exploited by many other authors including, for example, Hercowitz and Sampson (1991). This accumulation equation also implies that the higher the capital stock the larger the needed investments to increase (or maintain) the capital stock. Thus, the capital accumulation equation is one source of reversion to the mean for the capital stocks in the model, although not the only one (as we discuss in Section 3).

Each member of the household is endowed with one unit of time in each period, which can be devoted to either the accumulation of human capital or the provision of labour services in each of the j industries. In order to work in industry j, a member of the household must be physically present (at the start of the period) at a location that produces good j. It is convenient to think of the household as first distributing N_j of its members to each industry j subject to $\sum_j N_{tj} \leq N_t$, in each period, and then allocating the workers across cities producing each good. In equilibrium all cities end up specializing in the production of one industry, and all cities in an industry are identical, so that in the symmetric equilibrium the household allocates an equal amount of labour to each city in an industry.

Each worker spends u_{tj} amount of time working, with the remainder of each worker's time used to produce new human capital according to

$$H_{t+1j} = H_{tj}[B_j^0 + (1-u_{tj})B_j^1],$$

where B_j^0 and B_j^1 are non-negative constants. This specification allows us to nest both endogenous and exogenous growth within the same framework. If $B_j^1 = 0$, then human capital evolves exogenously at a constant rate B_j^0 and we have an exogenous growth model. If B_j^1 is positive, then the time allocation of a worker affects the growth rate of the economy, which results in an endogenous growth model. The assumption of linearity is made for simplicity, but is not necessary to generate balanced growth in this model since, as we will show below, the economy exhibits constant returns to scale in the aggregate.

Clearly, households will allocate their labour and human and physical capital services to the cities with the highest wages and rental rates net of commuting costs, so that in equilibrium these must be equal across all cities producing a given good. The household's sequences of flow budget constraints are given by

$$\sum_{j=1}^{J} p_{tj}[C_{tj} + X_{tj} + [\mathrm{ACC}_{tj} + \mathrm{AR}_{tj}]N_{tj}]$$
$$\leq \sum_{j=1}^{J}[w_{tj}N_{tj}u_{tj} + r_{tj}K_{tj} + s_{tj}H_{tj} + p_{tj}T_{tj}N_{tj}], \quad (4)$$

where ACC_{tj} and AR_{tj} represent average commuting costs and average rents, and T_{tj} denotes transfers to households from property developers.

2.4. *Property developers*

Property developers own land and aim to maximize total rents from their land. In order to attract firms and workers to the city, developers may subsidize the employment of all factors of production. Agents derive utility out of consumption of goods that are costlessly tradable, and so they live in the city if their income, net of commuting costs, is at least as large as what they could obtain elsewhere. Firms produce in the city as long as profits are non-negative. Free entry implies that developers earn zero profits in equilibrium, and so patterns of land ownership are irrelevant for all our results. Solving this problem results in city sizes that are optimal. Given the size of the industry, this means that we must allow for the possibility of a non-integer number of cities, all of which are identical in size within an industry. Since developers are fully internalizing the external effect, the equilibrium allocation is efficient.

It is important to stress that in this formulation developers choose to subsidize human capital independently of the subsidy to labour and that this subsidy is on the employment, but not the accumulation, of human capital. This distinction is important, since free mobility restricts the ability of developers to extract the benefits of subsidies to human capital accumulation. In practice, some policies that may achieve this goal are subsidies to firms that employ highly skilled workers or the provision of local public goods preferred by highly educated agents (*e.g.* fine arts).[6]

Property developers aim to maximize rents net of subsidies paid to firms in order to attract them, as well as factors of production, to the city. In order for workers to live in the city, they must receive large enough wages w_{tj}/p_{tj}, such that, net of commuting costs, their income I_{tj} is at least as large as what they could obtain in any other city producing in this industry. Free mobility then guarantees that all cities in an industry are identical. In order to attract firms, the returns to all factors have to be at least as large as the rental rates of these factors after subsidies.

Let μ_{tj} denote the number of cities in industry j at time t. All cities producing good j will be identical, so that if there are μ_{tj} cities producing good j, the amounts of labour and human capital in any one city are given by H_{tj}/μ_{tj} and N_{tj}/μ_{tj}. Then the problem of a property developer is to choose factor inputs in the city $N_{tj}/\mu_{tj}, K_{tj}/\mu_{tj}$, and H_{tj}/μ_{tj} and subsidies to factors of production, $T_{tj}, \tau^k_{tj}, \tau^h_{tj}$, to maximize

$$\Pi = \max\left[\frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{3/2} - T_{tj}\frac{N_{tj}}{\mu_{tj}} - \tau^k_{tj}\frac{r_{tj}}{p_{tj}}\frac{K_{tj}}{\mu_{tj}} - \tau^h_{tj}\frac{s_{tj}}{p_{tj}}\frac{H_{tj}}{\mu_{tj}}\right],$$

subject to

$$(1-\tau^k_{tj})r_{tj}/p_{tj} = \beta_j Y_{tj}/K_{tj},$$

$$(1-\tau^h_{tj})s_{tj}/p_{tj} = \alpha_j Y_{tj}/H_{tj},$$

$$I_{tj} = (1-\alpha_j-\beta_j)\frac{Y_{tj}}{N_{tj}} + T_{tj} - \frac{3b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{1/2}.$$

Competition from other developers ensures that profits are 0, so

$$T_{tj} = \frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{1/2} - \tau^k_{tj}\frac{r_{tj}}{p_{tj}}\frac{K_{tj}}{N_{tj}} - \tau^h_{tj}\frac{s_{tj}}{p_{tj}}\frac{H_{tj}}{N_{tj}}.$$

As total commuting costs are a convex function of the number of residents, all cities will be completely specialized in one industry. This follows from the fact that a developer could increase profits just by splitting a diversified city into several smaller specialized cities: total production would be identical, but residents would face lower commuting costs.

6. See Black and Henderson (1999) for a discussion of the difficulties in implementing this type of subsidy.

Note also that we are giving developers all the necessary instruments to internalize the externality, which allows us to show below that the equilibrium allocation will be optimal. If we were to restrict the instruments that developers can use (*e.g.* allow them only to provide transfers to agents but no human capital subsidies) the equilibrium allocation will no longer be efficient. For our purposes and the results that follow, what is important is that developers play a coordination role and form cities by maximizing rents. Whether cities are efficient or not is unimportant for the dynamics of the model and the characteristics of the size distribution of cities. By postulating the existence of these developers we eliminate equilibria in which, for example, all agents in an industry agglomerate in one city even though they could benefit from coordinating and forming more cities. Without property developers, this could occur given that there are no individual agent incentives to deviate to a new city since the productivity of an empty site is 0. Property developers serve to eliminate this multiplicity.

2.5. *Equilibrium*

Before we proceed to formally define an equilibrium allocation it is perhaps useful to discuss some of the features of the set-up we have introduced above. Instead of choosing a general specification of technology, factor accumulation, and utility, we have introduced a particular log-linear specification. This allows us to solve the model in closed form. Almost any deviation from this specification cannot be solved analytically and so we would need to rely completely on numerical simulations. This particular set-up is simple to solve because the log-linear specification ensures that income and substitution effects exactly balance. As a result, investments in physical capital in each industry turn out to be just a constant fraction of output, while employment in each industry turns out to be a constant fraction of total population. As we discuss below, the main forces that lead to our main results do not rely heavily on these features, although the precise formulas do. We prove in the next section that given the presence of property developers—that internalize external effects within a city—the equilibrium allocation exists, is unique, and efficient. None of these properties relies on the particular log-linear specification of the model.

We are now in a position to define a competitive equilibrium for this economy. Given that all cities in an industry are specialized and identical, we specify the equilibrium in terms of industry aggregates.

Definition 1. A competitive equilibrium for this economy is a set of state-contingent sequences $C_{tj}, X_{tj}, u_{tj}, N_{tj}, \mu_{tj}, H_{tj}, K_{tj}$ for each industry j and each period t and a price system $p_{tj}, w_{tj}, r_{tj}, s_{tj}$ and transfers and subsidies $T_{tj}, \tau^k_{tj}, \tau^h_{tj}$ for each industry j at each period t, such that

1. given $p_{tj}, w_{tj}, r_{tj}, s_{tj}$, and T_{tj}, households optimize,
2. given $p_{tj}, w_{tj}, r_{tj}, s_{tj}$, and τ^k_{tj}, τ^h_{tj}, firms hire K_{tj}, H_{tj}, and $N_{tj}u_{tj}$ so as to maximize profits,
3. given $p_{tj}, w_{tj}, r_{tj}, s_{tj}$, developers choose $T_{tj}, \tau^k_{tj}, \tau^h_{tj}$ and $N_{tj}/\mu_{tj}, K_{tj}/\mu_{tj}, H_{tj}/\mu_{tj}$ to maximize profits,
4. aggregate and individual decisions are consistent,
5. free entry implies zero profits for developers, and
6. markets for goods and factors clear:

$$C_{tj} + X_{tj} + bN_{tj}^{3/2}\mu_{tj}^{-1/2} = Y_{tj},$$

$$\sum_{j=1}^{J} N_{tj} = N_t.$$

2.6. *Efficient allocations*

As it is efficient for all cities to be identical and specialized, all Pareto-efficient allocations are the solution of the following *Social Planning Problem*: choose state-contingent sequences $\{C_{tj}, X_{tj}, N_{tj}, \mu_{tj}, u_{tj}, K_{tj}, H_{tj}\}_{t=0, j=1}^{\infty, J}$ to maximize

$$(1-\delta)E_0\left[\sum_{t=0}^{\infty}\delta^t N_t\left(\sum_{i=1}^{J}\ln C_{ti}/N_t\right)\right], \tag{5}$$

subject to, for all t and j,

$$C_{tj} + X_{tj} + b\left(\frac{N_{tj}}{\mu_{tj}}\right)^{3/2}\mu_{tj} \leq A_{tj}\left(\frac{K_{tj}}{\mu_{tj}}\right)^{\beta_j}\left(\frac{H_{tj}}{\mu_{tj}}\right)^{\alpha_j+\gamma_j}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{1-\alpha_j-\beta_j+\varepsilon_j} u_{tj}^{1-\alpha_j-\beta_j}\mu_{tj}, \tag{6}$$

$$N_t = \sum_{j=1}^{J} N_{tj}, \tag{7}$$

$$K_{t+1j} = K_{tj}^{\omega_j} X_{tj}^{1-\omega_j}, \tag{8}$$

$$H_{t+1j} = H_{tj}[B_j^0 + (1-u_{tj})B_j^1], \tag{9}$$

given H_{0j} and K_{0j}. The first constraint states that consumption plus investment plus commuting costs has to be less than or equal to production in all cities in the industry.

This is not a convex dynamic optimization problem. However, since the problem of determining the optimal city size is static, we can solve it separately. This transforms the problem into a convex dynamic optimization problem. We can then show that the optimal allocation exists and is unique. The rest of the proof shows that the conditions that determine an equilibrium allocation are identical to the ones that determine the unique efficient allocation. This implies that there exists a unique equilibrium allocation and that the equilibrium is efficient. Apart from the developers problem, the proof of this proposition is standard. Note that, because all households are identical, we focus on the symmetric allocation.

Proposition 1. *There exists a unique symmetric Pareto-efficient allocation. There exists a unique and efficient symmetric competitive equilibrium.*

As we argued above, efficiency is the result of the existence of property developers who are able to offer a rich set of subsidies. The main implications of our model are unaffected by the number of instruments possessed by property developers and hence also by whether the equilibrium allocation is efficient. Hence, we proceed with the case where developers can fully internalize the externality.

3. CHARACTERIZATION

With these results in hand, we are free to make use of the solution to the social planning problem in order to characterize the competitive equilibrium of the model. We now proceed to derive several properties of the equilibrium allocation. Due to our functional form assumptions, we are able to solve for the entire equilibrium growth path and size distribution of cities in closed form.

3.1. *Aggregate constant returns*

As there are no adjustment costs at the city level, the problem of choosing the optimal sizes of cities is static. In each period, the planner sets the city size to maximize output net of commuting costs. We solve this problem first and then, imposing the solution, we solve for the dynamics. Towards this, we can rewrite the resource constraint in an industry j at time t as a function of industry-wide variables and the number of cities in an industry:

$$C_{tj} + X_{tj} + bN_{tj}^{3/2}\mu_{tj}^{-1/2} \leq A_{tj}K_{tj}^{\beta_j}H_{tj}^{\alpha_j+\gamma_j}N_{tj}^{1-\alpha_j-\beta_j+\varepsilon_j}u_{tj}^{1-\alpha_j-\beta_j}\mu_{tj}^{-\gamma_j-\varepsilon_j} \equiv Y_{tj}.$$

The first-order condition with respect to μ_{tj} yields the optimal number of cities in industry j, as a function of output and employment in that industry:

$$\mu_{tj} = \left[\frac{2(\gamma_j+\varepsilon_j)}{b}\frac{Y_{tj}}{N_{tj}}\right]^{-2}N_{tj}, \tag{10}$$

and so total commuting costs satisfy

$$\mathrm{TCC}_{tj} = 2(\gamma_j+\varepsilon_j)Y_{tj}. \tag{11}$$

Notice that we need to impose $\gamma_j+\varepsilon_j < 1/2$, in order to guarantee that the solution of the first-order condition attains a maximum (if this condition is not satisfied, the first-order conditions yield a minimum at which total commuting costs are larger than total output). To interpret this restriction, write industry output net of total commuting cost as

$$A_{tj}K_{tj}^{\beta_j}H_{tj}^{\alpha_j+\gamma_j}N_{tj}^{1-\alpha_j-\beta_j+\varepsilon_j}u_{tj}^{1-\alpha_j-\beta_j}\mu_{tj}^{-\gamma_j-\varepsilon_j} - bN_{tj}^{3/2}\mu_{tj}^{-1/2},$$

and notice that if the above condition is not satisfied, as the number of cities decreases, given industry aggregates, the value of the expression increases unboundedly. This implies that the above problem has no internal solution: the planner would like to make cities as large as possible.

Substituting the results for the optimal number of cities and total commuting costs in the resource constraint implies that

$$C_{tj} + X_{tj} \leq F_j\hat{A}_{tj}H_{tj}^{\hat{\alpha}_j}K_{tj}^{\hat{\beta}_j}N_{tj}^{1-\hat{\alpha}_j-\hat{\beta}_j}u_{tj}^{\hat{\phi}_j} \equiv \hat{Y}_{tj}, \tag{12}$$

where

$$F_j = (1-2(\gamma_j+\varepsilon_j))\left[\frac{2(\gamma_j+\varepsilon_j)}{b}\right]^{[2(\gamma_j+\varepsilon_j)]/[1-2(\gamma_j+\varepsilon_j)]},$$

$$\hat{A}_{tj} = A_{tj}^{1/[1-2(\gamma_j+\varepsilon_j)]}, \quad \hat{\alpha}_j = \frac{\alpha_j+\gamma_j}{1-2(\gamma_j+\varepsilon_j)}$$

$$\hat{\beta}_j = \frac{\beta_j}{1-2(\gamma_j+\varepsilon_j)}, \quad \text{and} \quad \hat{\phi}_j = \frac{1-\alpha_j-\beta_j}{1-2(\gamma_j+\varepsilon_j)}.$$

Since u_{tj} is constant in equilibrium, output net of commuting costs for the optimal city structure ($\hat{Y}_{tj}$) is constant returns to scale in industry aggregates. Notice that by equation (11) output in the industry is also a constant returns to scale function of inputs in the industry.

The constraint in (12) contains the first main result of our paper: introducing the margin of the creation of new cities eliminates increasing returns at the urban level from the aggregate problem. We summarize this result in the following proposition.

Proposition 2 (Aggregate constant returns to scale). *Output in industry j, Y_{tj}, and industry output net of commuting costs, $\hat{Y}_{tj}$, are constant returns to scale functions of industry-specific capital K_{tj}, industry-specific human capital H_{tj}, and labour N_{tj}.*

The result in this proposition has implications for the way in which we view the growth process. First, it allows us to reconcile the coexistence of cities, which implies the existence of scale economies, with balanced growth. Second, it shows that it is inappropriate to test for the existence of increasing returns with aggregate data even though increasing returns are, in fact, present in the production technology. Third, the observed level of aggregate productivity (the magnitude of F_j in equation (12)) is determined by the way production is organized in cities, as well as the parameters governing externalities and commuting costs. This suggests the possibility that differences in the pattern of urbanization are the source of differences in total factor productivity across countries.[7]

3.2. *City sizes*

To understand the process of city size determination, rewrite the first-order condition for the number of cities, μ_{tj}, as

$$\frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{-1/2} = (\gamma_j + \varepsilon_j)\frac{Y_{tj}/N_{tj}}{N_{tj}/\mu_{tj}}.$$

That is, the planner increases the number of people in the city until the change in commuting costs per person for current residents (L.H.S.) is equal to the change in earnings per person for current residents (R.H.S.).

From this equation it is easy to see that anything that increases the level of the average product of labour increases the average size of the city. For example, consider the effect of an increase in productivity. Everything else equal, output per worker increases, and the planner finds it optimal to attract more workers to the city. If the productivity increase is permanent, the city will be permanently larger. The growth model presented above is, in essence, a mechanism for producing persistence in the average product of labour in a city, while at the same time remaining consistent with aggregate growth facts.

Our mechanism relies on city sizes that respond to factor accumulation and productivity shocks. This is the case as long as average commuting costs do not rise by exactly the same amount as the average product of labour. If commuting costs were to rise by less, or even more, than the average product of labour, the basic result that productivity shocks are translated into fluctuations in city sizes remains. However, one combination of assumptions that does not work is if commuting costs are denominated *purely* in units of time, *and* workers supply labour inelastically, *and* the production function is Cobb–Douglas. In this knife-edge case, marginal and average products are proportional and hence commuting costs, measured as forgone wages, rise at exactly the same rate as the average product of labour. More generally, any combination of time and material costs of commuting yields the necessary response of city sizes to productivity shocks. In the model above, we focus on a simple case in which commuting costs within a city are denominated in terms of the output of that city. The results are analogous if we include time costs of commuting as well.

7. Au and Henderson (2006) examine this possibility for the particular case of China.

3.3. *Growth rates*

To solve for the dynamics of factor accumulation, note that after substituting for the optimal number of cities we obtain a standard dynamic problem with constant returns to scale production technology. In particular, our problem becomes one of choosing $\{C_{tj},\ X_{tj},\ N_{tj},\ u_{tj},\ K_{tj},\ H_{tj}\}_{t=0,j=1}^{\infty,J}$ so as to maximize (5) subject to (12) and (7)–(9). The value function of the planner has the form

$$V(\{H_{tj}, K_{tj}, A_{tj}\}_{j=1}^{J}) = D_0 + \sum_{j=1}^{J}[D_j^H \ln(H_{tj}) + D_j^K \ln(K_{tj}) + D_j^A \ln(A_{tj})], \tag{13}$$

which is the result of the particular log-linear specification we have assumed. We could set up a more general model at the cost of losing the ability to solve the model analytically. The details of the solution are entirely standard and so are relegated to the Appendix. Three basic results come out of solving the planner's problem. The share of population working in each industry is constant. Investment is a constant share of output net of commuting costs $X_{tj} = x_j \hat{Y}_{tj}$, for some constant x_j, and the fraction of time used for production is constant at u_j^*.

Note that the model is capable of producing growth, either exogenously or endogenously. More importantly, the model delivers two properties not present in most other *urban* growth models: a balanced growth path exists without knife-edge assumptions on the size of externalities, and growth is positive even in the absence of population growth. On the balanced growth path (with no uncertainty) we know that the growth rates of capital (g_{K_j}), human capital (g_{Hj}), and output net of commuting costs ($g_{\hat{Y}_j}$) are constant, so

$$g_{K_{t+1j}} \equiv \ln K_{t+1j} - \ln K_{tj} = (1-\omega_j)[\ln x_j + \ln \hat{Y}_{tj}] - (1-\omega_j)\ln K_{tj}.$$

Hence, on the balanced growth path $\ln \hat{Y}_{tj} - \ln K_{tj}$ is constant. The growth rate of human capital is given by $g_{H_j} = B_j^0 + (1-u_j^*)B_j^1$. For income, when $\hat{\beta}_j < 1$, on the balanced growth path[8] (with no uncertainty),

$$g_{\hat{Y}_j} = \frac{\hat{\alpha}_j g_{H_j} + (1-\hat{\alpha}_j - \hat{\beta}_j) g_N}{1-\hat{\beta}_j}.$$

That is, in the long run, growth is driven by endogenous human capital accumulation (if $B_j^1 > 0$) and exogenous population growth.

Notice that in this model linearity in human capital accumulation implies that growth rates are constant in the long run, even with increasing returns in the aggregate production function. In general, this type of linearity plays two different roles in growth models: it is a source of endogenous growth, and it prevents growth rates from diverging to infinity. In this paper, this linearity serves the first, and not the second, purpose. We use it to show that our results do not depend on the source of growth and, in particular, whether it is exogenous or endogenous. To illustrate this point, suppose we set $1 < \alpha_j + \beta_j + \gamma_j$ for all j, and we let human capital accumulate exactly as physical capital. Then, without cities, due to the presence of aggregate increasing returns, growth rates would diverge to infinity. However, with this type of increasing returns at the city level, the mechanism we have introduced in this paper would yield constant returns in the aggregate and therefore a balanced growth path in which $g_{\hat{Y}_j} = g_N$.

8. For the case when $\hat{\beta}_j = 1$, $g_N = g_H = 0$, and $\omega = 0$ (the AK model), $g_{\hat{Y}_{t+1j}} = \ln x_j + \ln(F_j A_{tj})$.

3.4. *Gibrat's and Zipf's laws*

Given the evolution of output in each industry, we can study the evolution of the size distribution of cities. The growth rate of a city in industry j is given by

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right)-\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)=2[\ln(A_{t+1j})-\ln(A_{tj})]-2(\hat{\alpha}_j+\hat{\beta}_j)[\ln(N_{t+1})-\ln(N_t)]$$
$$+2\hat{\alpha}_j\ln(B_j^0+(1-u_j^*)B_j^1)+2\hat{\beta}_j[\ln(K_{t+1j})-\ln(K_{tj})].$$

Recursively substituting for capital growth, we get an expression for the long-run growth rate of cities:[9]

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right)-\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)=\frac{2\hat{\alpha}_j}{1-\hat{\beta}_j}[g_{Hj}-g_N]+2[\ln(A_{t+1j})-\ln(A_{tj})]$$
$$+2(1-\omega_j)\hat{\beta}_j\left[\ln(A_{tj})-\sum_{s=1}^{\infty}\frac{(\omega_j+(1-\omega_j)\hat{\beta}_j)^{s-1}}{(1-(\omega_j+(1-\omega_j)\hat{\beta}_j))^{-1}}\ln(A_{t-sj})\right]. \quad (14)$$

Equation (14) is the key equation for characterizing city dynamics. From this equation we can deduce conditions under which Gibrat's law is guaranteed for each group of industries. In order to generate Gibrat's law we need the growth processes at the city level to be independent of scale. As labour is perfectly mobile across cities and industries, this in turn requires that the marginal product of labour be independent of scale. The proposition below outlines two scenarios in which this is exactly the case: the first is one in which current productivity shocks are the only stochastic force in growth and are permanent, thus producing permanent increases in the level of the marginal product of labour, so that the growth rate of the marginal product is independent of scale. These assumptions eliminate the third term in equation (14) and therefore all scale dependence. This result is invariant to whether the engine of growth is endogenous or exogenous. The second case is one in which productivity shocks are temporary, but have a permanent effect on the marginal product of labour through the linear accumulation of physical capital. This amounts to transforming the model into an AK model with no human capital and 100% depreciation. In this context, both last period output and capital react linearly to last period shocks. These two effects cancel out, and the only remaining source of uncertainty is the contemporaneous productivity shock.[10]

We then show that Gibrat's law implies Zipf's law in our framework. Note that in our framework we need to consider the entry and exit of new cities. As a consequence we cannot simply apply previous results in the literature, which rely on a fixed number of cities (Gabaix, 1999*a*,*b*; Cordoba, 2004; Eeckhout, 2004).[11] The next proposition provides a new proof of the relationship between Gibrat's and Zipf's laws in our framework, based on some results by Levy and Solomon (1996) and Malcai *et al.* (1999). In the proof of the proposition, we use the assumption that our industries can be divided into groups with similar technologies to first prove that Zipf's law holds

9. For the details on how to derive this expression see the solution to the planner's problem in the Appendix.

10. Note that if we were to allow infinite order Markov processes for A_j, we could fine-tune the specification of the process so as to yield Gibrat's law exactly for any parameter set.

11. Gabaix (1999*a*) and Cordoba (2004) impose lower bounds on city sizes and a particular structure on the shocks that leads to an urban structure described by a Pareto distribution with coefficient 1. Eeckhout (2004) has permanent productivity shocks that lead, without a lower bound via the Central Limit Theorem, to a log-normal distribution for city sizes. The economic interpretation of the shocks differs in all three cases.

for each group. We then aggregate across groups to show Zipf's law for the entire economy. The proof of this result requires us to impose an arbitrarily small lower bound on the size of a city (as in Gabaix, 1999*a*). All proofs are relegated to the Appendix.

Proposition 3 (Exact Gibrat's law and Zipf's law). *The growth process of city sizes satisfies Gibrat's law if and only if one of the following two conditions is satisfied:*

1. *(No physical capital) There is no physical capital ($\hat{\beta}_j = 0$ or $\omega_j = 1$), and productivity shocks are permanent.*
2. *(AK model) City production is linear in physical capital and there is no human capital ($\hat{\alpha}_j = 0$, $\hat{\beta}_j = 1$), depreciation is* 100% *($\omega_j = 0$), and productivity shocks are temporary.*

If the growth process satisfies Gibrat's law and city sizes are bounded below by $\underline{s}$, the invariant distribution for city sizes satisfies Zipf's law as $\underline{s} \to 0$.

3.5. *Scale dependence*

Obviously, the conditions outlined in Proposition 3 are restrictive. Reality surely lies between these two extremes: capital is a factor of production, but not the only one. The question that arises is between these two extremes, how close are the predictions of the model to observed urban structures? As mentioned in the introduction, an extensive empirical literature (surveyed in Gabaix and Ioannides, 2003) has uncovered two systematic departures from Zipf's law. First, plots of log-rank against log-size are concave, reflecting the fact that small cities are underrepresented and that big cities are not "big enough". Second, there is variation in cross-country estimates of Zipf's coefficients (Soo, 2005).

In the next two propositions we argue that, in general, the model can account for these same deviations from Zipf's law. First we show that if a city is relatively large because it operates in an industry that experienced a history of above-average productivity shocks, it can be expected to grow slower than average in the future, while the opposite is true of small cities. Intuitively, since $\hat{\beta} < 1$, diminishing returns to capital imply that industries with high capital stocks have a lower return to capital than industries with low capital stocks, and so cities in industries with relatively low stocks of physical capital grow faster. This effect is emphasized by the fact that when $\omega_j > 0$ for all j, in order to keep physical capital constant, industry investments have to be higher in industries with large capital stocks and lower in industries with low capital stocks. Urban growth rates exhibit reversion to the mean. This implies that the log rank–size relationship will in general (apart from particular realizations of the shocks) be concave or, in other words, that the invariant distribution for city sizes has thinner tails than a Pareto distribution with coefficient 1. Eeckhout (2004) emphasizes exactly this feature of the data.

Proposition 4 (Concavity). *If conditions* 1 *and* 2 *in Proposition* 3 *are not satisfied, the growth rate of cities exhibits reversion to the mean. If productivity levels are bounded for all industries (there exist uniform bounds such that $A_{tj} \in [\underline{A}_j, \overline{A}_j]$ for all t, j), then there exists a unique invariant distribution of city sizes with thinner tails than a Pareto distribution with coefficient* 1.

Unless the conditions of Proposition 3 are satisfied, variation in the S.D. of productivity shocks affects the distribution of city sizes. Intuitively, given capital stocks, a larger S.D. of shocks directly implies a larger S.D. of city sizes. Moreover, it also implies a larger S.D. of investments, which in turn implies a more dispersed distribution of capital stocks. We formalize this intuition in the following proposition.

Proposition 5. *If conditions* 1 *and* 2 *in Proposition* 3 *are not satisfied, the S.D. of city sizes increases with the S.D. of industry shocks.*

Proposition 5 points to the S.D. of productivity shocks as the key parameter linking our model with the observed urban structure. In the next section we explore whether the international evidence on Zipf's coefficients is consistent with the evidence on the volatility of industry productivity shocks.

4. NUMERICAL EXERCISES

In the previous section, we established analytically the conditions under which our theory produces Zipf's law exactly, and showed that away from these special cases the city size distribution always has thinner tails than is implied by Zipf's law, and that the amount of dispersion in the distribution increases with the volatility of industry shocks. In this section, we supplement these results with numerical simulations designed to show that the theory can robustly generate size distributions in line with the data for a wide range of plausible parameter values.

To illustrate the deviations from Zipf's law that result in our model when we move away from the assumptions in Proposition 3 and how they relate to the data, we begin with Figure 3 that contains data on the city size distribution, defining cities as Metropolitan Statistical Areas (MSAs), in 2002 for the U.S., Belgium, and Saudi Arabia. Alongside the actual data for these countries, we also present the results of numerical simulations of the model. Each simulation has been run for 10,000 periods, after which the simulated distribution of city sizes is not changing significantly through time. We used relatively standard values for most parameters. The discount factor δ was set to 0·95, which is consistent with annual rates of return. The production parameters for the firm were all set to one-third, or $\alpha = \beta = \phi = 1/3$, while the externality parameters were set to one-tenth, or $\gamma = \varepsilon = 0{\cdot}1$. Human capital accumulation is parameterized so that there is no exogenous accumulation of human capital, or $B^0 = 1$ with $B^1 = 0{\cdot}2$, while population growth is 2%, or $g_N = 1{\cdot}02$. We set commuting costs to ten per unit of distance $\tau = 10$. One non-standard parameter is ω, which governs the importance of existing capital to the accumulation of future capital. This parameter controls the extent to which capital depreciates and so we set it to 0·9. However, it turns out that changing the level of ω has only a modest effect on the quantitative behaviour of the model. To see this, note that as we increase ω, on the one hand, the mean reversion caused by the capital accumulation equation increases since the share of investment in tomorrow's capital production $(1-\omega)$ decreases. On the other hand, less capital is accumulated and so the stock of human capital is lower, which reduces this effect. Finally, we need to set the stochastic process for productivity shocks. Given that we are interested in the long-run distribution, we directly parameterize the long-run shock process to fit the U.S. urban structure. In particular, we set $m = 0$ and $\text{sd} = 0{\cdot}5$, which denote the mean and S.D. of the normal distribution from which the logarithms of the long-run transitory shocks are drawn.

As one can see in Figure 3, the model does very well—arguably better than Zipf's law—in matching the U.S. data. In particular, and as expected given Proposition 4, the curve is slightly concave, as in the data. That is, large cities are too small and there are not enough small cities. Note also, given that this plot is the result of one particular sample of shocks, at the top end there are some portions of local convexity. Nonetheless, the overall picture is of an approximately concave plot.

Empirical studies have found that Zipf's law fits the data well across a wide variety of countries and over long periods of time. Therefore, fitting the distribution for one particular country at a single point in time is not helpful in explaining this general phenomenon. Instead, we want to focus on the robustness of the model's predictions to variations in the underlying key parameters.

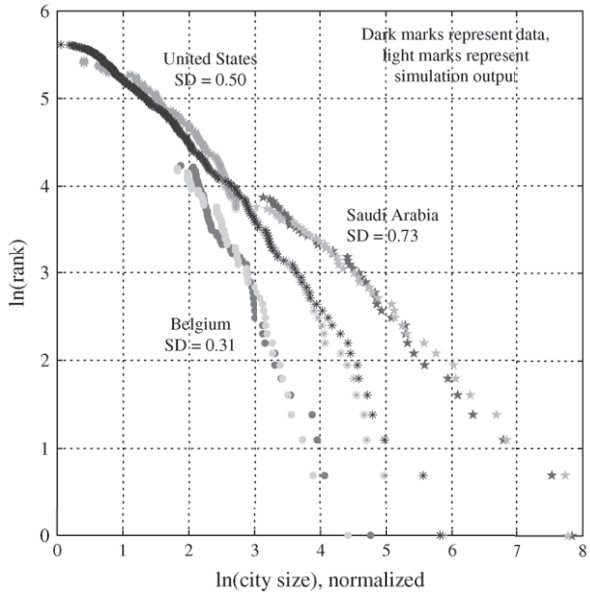

FIGURE 3
Model vs. data

Proposition 5 tells us that one key parameter is the S.D. of industry shocks. Otherwise, the model seems to be relatively robust (although not invariant) to all other parameter values.[12] This justifies our focus on the S.D.: the model has identified this parameter as the main source of variation in Zipf's law coefficients. We illustrate the urban distributions resulting from different assumptions on the S.D. by fitting the distribution of Belgium and Saudi Arabia. These two countries exhibit city size distributions that are either extremely concentrated or extremely dispersed, relative to other countries. The rank–size relationship in Belgium is very steep with a Zipf's coefficient of 1·59. The S.D. of long-run shocks that yields a city size distribution consistent with the Belgian data is 0·31. We perform the same exercise for Saudi Arabia that exhibits a very flat rank–size relationship. Saudi Arabia's cities are very distinct in terms of population sizes, with a Zipf's coefficient of 0·78. The S.D. used in the numerical simulation is $\text{sd} = 0{\cdot}73$.[13]

To assess whether or not this range of S.D. is plausible, we can use the model to compare the international evidence on urban structures with evidence on the range of observed industry productivity shock variances. To begin, note that Belgium and Saudi Arabia provide a range of S.D. that would imply city size distributions consistent with what we observe in the data. We first ask whether this range is in line with measures of productivity shocks by industry. Horvath (2000) measures the S.D. and persistence of industry shocks in the U.S. for 36 industries, from which we calculate long-run S.D. As the U.S. is the world's largest economy and is relatively diversified across industries, we will take these data to represent the universe of possible productivity shock processes. We then ask what proportion of industries have productivity shock

12. Except the discount factor, δ, that is related to the S.D., sd, via the period length, which is calibrated to one year.

13. There are a few countries that exhibit Zipf's coefficients that are higher or lower than Belgium and Saudi Arabia. The reason we do not use them is that typically they have only very few cities. For example, Guatemala, with 13 cities, has a Zipf's coefficient of 0·728, while Kuwait, with 28 cities, has a Zipf's coefficient of 1·720. Using these countries would only improve the performance of the model in the comparisons that follow.

processes within the implied interval. In interpreting the result of this test, it is important to stress that this comparison puts a heavy burden on our theory. To understand this, consider a situation where all of the S.D. of productivity shocks are inside the intervals implied by the range of Zipf's coefficients. That would mean that if a country were to have industries that faced *only* the least variable productivity shocks, it would still exhibit a Zipf's coefficient within the range of international evidence. However, we know that *all* countries produce in a variety of industries that face shocks that differ in their S.D. That is, there is no country that produces only in the most volatile industry. Therefore, it is impossible for *all* industries' volatilities to be inside the implied range. Conversely, if none of the S.D. were inside the implied range, it would be evidence against our theory. It turns out that, under this calculation, 50% of the industry long-run S.D. estimated by Horvath lie within the bounds implied by our model. We interpret this is as a substantial success for our model. We can also do the reverse exercise and calculate the range of Zipf's coefficients that our model can produce given Horvath's productivity numbers. The resulting range is wide enough to include the urban size distribution of all countries. Thus, variation in the S.D. of industry productivity shocks can go a long way in explaining the observed variation in Zipf's coefficients.[14]

5. CONCLUSIONS

We have proposed a tractable general equilibrium urban growth theory. It emphasizes the role of the accumulation of specific factors across industries in determining the evolution of the urban structure. In this theory, cities arise endogenously out of a trade-off between agglomeration forces and congestion costs. It is the size distribution of cities itself, and it is evolution through the birth, growth, and death of cities, which leads to a reconciliation between increasing returns at the local level and constant returns at the aggregate level. The urban structure of the economy prevents growth rates from diverging. Moreover, this same urban structure displays many of the features observed in actual city size distributions across countries and over time.

An advantage of the simple specification we adopted above is that it allows us to identify analytically the S.D. of industry productivity shocks as the crucial factor determining cross-country differences in urban structure. An empirical analysis of this parameter is, we believe, an important part of any systematic empirical evaluation of cross-country differences in the size distribution of cities.

One of the limitations of this simple specification is that cities specialize in only one industry. We can introduce diversified cities using either cross-industry spillovers or non-traded consumption goods. In these cases cities will produce goods in multiple industries, but city dynamics and the characteristic of the cross-sectional distribution of cities, as well as the aggregate properties of the model, will remain unchanged.

Finally, our theory points to differences in the efficiency at which cities are organized as a potential explanation of the observed differences in total factor productivity across countries. In our theory, we justified focusing on cities that are organized efficiently by postulating the existence of property developers with access to a sophisticated range of policy instruments. Restricting the range of policy instruments available to these developers, for example by eliminating subsidies on human capital, does not affect the main results of our theory, but translates into lower observed levels of total factor productivity. The varying ability of local governments in different

14. Soo (2005) finds that the coefficients in absolute value tend to be smaller (more unequal distribution of cities) in Africa, South America, and Asia than in Europe, North America, and Oceania. Since most of the developed economies are in the last group of continents and presumably these are the countries that experience less volatility of income (*i.e.* smaller industry shocks), we view the response of the model to changes in sd as potentially identifying the source of the differences in Zipf's coefficients observed in the data.

countries to use these policies is, potentially, an important determinant of income levels. These policies are particularly important for cities, given that urban scale economies are unlikely to have been fully internalized. We hope that future research will examine the empirical relationship between local government policy, urban structure, and aggregate total factor productivity levels across countries.

APPENDIX

A.1. *Solution of social planner's problem (SPP)*

Our first task is to solve the planning problem. This SPP is to choose state-contingent sequences $\{C_{tj}, X_{tj}, N_{tj}, u_{tj}, K_{tj}, H_{tj}\}_{t=0,j=1}^{\infty,J}$ to maximize (5) subject to, for all t and j, (7)–(9), and

$$F_j A_{tj} H_{tj}^{\hat{\alpha}_j} K_{tj}^{\hat{\beta}_j} N_{tj}^{1-\hat{\alpha}_j-\hat{\beta}_j} u_{tj}^{\hat{\phi}_j} = C_{tj} + X_{tj}.$$

To solve this problem, we can verify that the value function of the problem takes the form given by (13). This leads to

$$C_{tj}^* = \frac{(1-\delta)}{\delta D_j^K (1-\omega_j) + (1-\delta)} \hat{Y}_{tj},$$

which implies that

$$X_{tj}^* = \frac{\delta D_j^K (1-\omega_j)}{\delta D_j^K (1-\omega_j) + (1-\delta)} \hat{Y}_{tj} \equiv x_j \hat{Y}_{tj}.$$

We can use this result to obtain expressions for u_{tj} and N_{tj}^*:

$$u_j^* = \frac{\hat{\phi}_j (B_j^0 + B_j^1)[\delta D_j^K (1-\omega_j) + (1-\delta)]}{\delta D_j^H B_j^1 + \hat{\phi}_j B_j^1 [\delta D_j^K (1-\omega_j) + (1-\delta)]},$$

$$N_{tj}^* = \frac{(1-\hat{\alpha}_j - \hat{\beta}_j)(\delta D_j^K (1-\omega_j) + (1-\delta))}{\sum_{j=1}^J [(1-\hat{\alpha}_j - \hat{\beta}_j)(\delta D_j^K (1-\omega_j) + (1-\delta))]} N_t \equiv n_j N_t,$$

where

$$D_j^K = \frac{(1-\delta)\hat{\beta}_j}{1-\delta\omega_j - \delta(1-\omega_j)\hat{\beta}_j}, \text{ and}$$

$$D_j^H = \hat{\alpha}_j + \frac{\delta\hat{\beta}_j(1-\omega_j)\hat{\alpha}_j}{1-\delta\omega_j - \delta(1-\omega_j)\hat{\beta}_j}.$$

We would like to find out what these results imply for the law of motion of physical and human capital. For this, notice that

$$\ln H_{tj} = \ln H_{0j} + t \ln(B_j^0 + (1-u_j^*)B_j^1),$$

$$\ln K_{tj} = \omega_j \ln K_{t-1j} + (1-\omega_j)[\ln x_j + \ln \hat{Y}_{t-1j}].$$

Of course,

$$\ln \hat{Y}_{tj} = \ln(F_j) + \ln(A_{tj}) + \hat{\alpha}_j \ln(H_{tj}) + \hat{\beta}_j \ln(K_{tj}) + (1-\hat{\alpha}_j - \hat{\beta}_j)\ln(N_{tj}^*) + \hat{\phi}_j \ln(u_j^*),$$

so

$$\ln K_{tj} = \omega_j \ln K_{t-1j} + (1-\omega_j)[\ln x_j + \ln(F_j) + \ln(A_{t-1j}) + \hat{\alpha}_j \ln(H_{t-1j}) + \hat{\beta}_j \ln(K_{t-1j}) + (1-\hat{\alpha}_j - \hat{\beta}_j)\ln(N_{t-1j}^*) + \hat{\phi}_j \ln(u_j^*)].$$

Given that we are interested in characterizing the solution with shocks, we want to determine the invariant distribution of the model. For this, we want to characterize first $\lim_{t\to\infty} \ln K_{tj} - \ln K_{t-1j}$. Taking differences, recursively substituting, assuming that $\hat{\beta}_j < 1$ and that population growth is constant, so that $N_t = (g_N)^t N_0$, we obtain

$$\lim_{t\to\infty}[\ln K_{tj} - \ln K_{t-1j}] = (1-\omega_j)\lim_{t\to\infty}\left[\ln(A_{t-1j}) - \sum_{T=1}^{t-1}\frac{(\omega_j+(1-\omega_j)\hat{\beta}_j)^{t-1-T}}{(1-(\omega_j+(1-\omega_j)\hat{\beta}_j))^{-1}}\ln(A_{T-1j})\right]$$

$$+\frac{1}{1-\hat{\beta}_j}[(1-\hat{\alpha}_j-\hat{\beta}_j)g_N + \hat{\alpha}_j\ln(B_j^0+(1-u_j^*)B_j^1)].$$

The size of the city is given by

$$\frac{N_{tj}}{\mu_{tj}} = \left[\frac{2(\varepsilon_j+\gamma_j)}{b}\frac{Y_{tj}}{n_j N_t}\right]^2,$$

so

$$\ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = 2\left[\ln\left(\frac{2(\varepsilon_j+\gamma_j)}{bn_j}\right)+\ln(Y_{tj})-\ln(N_t)\right]$$

$$= 2\left[\ln\left(\frac{n_j F_j 2(\varepsilon_j+\gamma_j)}{bn_j^{\hat{\alpha}_j+\hat{\beta}_j}(1-2(\varepsilon_j+\gamma_j))}\right)+\ln(A_{tj})+\hat{\alpha}_j\ln(H_{tj})\right.$$

$$\left.+\hat{\beta}_j\ln(K_{tj})-(\hat{\alpha}_j+\hat{\beta}_j)\ln(N_t)+\hat{\phi}_j\ln(u_j^*)\right].$$

Hence,

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right)-\ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = 2[\ln(A_{t+1j})-\ln(A_{tj})]-2(\hat{\alpha}_j+\hat{\beta}_j)[\ln(N_{t+1})-\ln(N_t)]$$

$$+2\hat{\alpha}_j\ln(B_j^0+(1-u_j^*)B_j^1)+2\hat{\beta}_j[\ln(K_{t+1j})-\ln(K_{tj})],$$

where the expression for $\ln(K_{t+1j})-\ln(K_{tj})$ is given above. Taking limits,

$$\lim_{t\to\infty}\left[\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right)-\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\right] = 2\lim_{t\to\infty}[[\ln(A_{t+1j})-\ln(A_{tj})]-(\hat{\alpha}_j+\hat{\beta}_j)[\ln(N_{t+1})-\ln(N_t)]]$$

$$+2\hat{\alpha}_j\ln(B_j^0+(1-u_j^*)B_j^1)+2\hat{\beta}_j\lim_{t\to\infty}[\ln(K_{t+1j})-\ln(K_{tj})].$$

Imposing constant population growth,

$$\lim_{t\to\infty}\left[\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right)-\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\right] = 2\lim_{t\to\infty}[\ln(A_{t+1j})-\ln(A_{tj})]$$

$$+2(1-\omega_j)\hat{\beta}_j\lim_{t\to\infty}\left[\ln(A_{tj})-\sum_{T=1}^{t}\frac{(\omega_j+(1-\omega_j)\hat{\beta}_j)^{t-T}}{(1-(\omega_j+(1-\omega_j)\hat{\beta}_j))^{-1}}\ln(A_{T-1j})\right]$$

$$-\frac{2\hat{\alpha}_j}{1-\hat{\beta}_j}g_N+\frac{2\hat{\alpha}_j}{1-\hat{\beta}_j}[\ln(B_j^0+(1-u_j^*)B_j^1)].$$

A.2. *Proofs of propositions*

Proof of Proposition 1. We start with the proof that there exists a unique Pareto-efficient allocation. As the number of cities of each type μ_{tj} enters only into the resource constraint, the optimal choice of the number of cities is static and maximizes

$$A_{tj}K_{tj}^{\beta_j}H_{tj}^{\alpha_j+\gamma_j}N_{tj}^{1-\alpha_j-\beta_j+\varepsilon_j}u_{tj}^{1-\alpha_j-\beta_j}\mu_{tj}^{-\varepsilon_j-\gamma_j}-bN_{tj}^{3/2}\mu_{tj}^{-1/2}. \quad (15)$$

We will study the properties of this expression for given strictly positive values of K_{tj}, H_{tj}, u_{tj}, and N_{tj}. Let

$$A(K_{tj}, H_{tj}, u_{tj}, N_{tj}) \equiv A_{tj} K_{tj}^{\beta_j} H_{tj}^{\alpha_j+\gamma_j} N_{tj}^{1-\alpha_j-\beta_j+\varepsilon_j} u_{tj}^{1-\alpha_j-\beta_j}.$$

Then it is easy to see that

$$\frac{A(K_{tj}, H_{tj}, u_{tj}, N_{tj})}{bN_{tj}^{\frac{3}{2}}} \mu_{tj}^{\frac{1}{2}-\gamma_j},$$

under our assumption that $\varepsilon_j + \gamma_j < 1/2$, is strictly increasing in μ_{tj}, equals 0 when $\mu_{tj} = 0$, and is unbounded as μ_{tj} tends to positive infinity. Hence, there exists a μ^* such that for all $\mu \leq \mu^*$, the expression in (15) is negative, while for all other μ it is strictly positive. Moreover, in the limit as μ goes to infinity, the expression in (15) goes to 0. Hence, as the expression is continuous in μ, it possesses a maximum on $[\mu^*, +\infty)$, which from the first-order necessary condition satisfies (11). Rearranging the first-order condition we also find that the optimal number of cities is given as a function of output and employment in the industry, so (10) holds.

If we substitute these expressions into the above optimization problem, we get the augmented social planning problem described above. This problem is convex, and as the objective function is strictly concave, it possesses a unique solution. As a result of the functional form assumptions, the solution has strictly positive levels for physical and human capital, employment, and hours worked at every date and in every state of the world. Hence the solution of the adjusted programming problem also satisfies the constraints of the social planning problem, and hence it is also the unique solution to the social planning problem.

To show the equivalence of the competitive equilibrium and social optimum, we begin with the solution of the SPP. We know that this solution is the unique allocation satisfying the first-order condition of the SPP to choose state-contingent sequences $\{C_{tj}, X_{tj}, N_{tj}, \mu_{tj}, u_{tj}, K_{tj}, H_{tj}\}_{t=0, j=1}^{\infty, J}$ to maximize (5) subject to, for all t and j, (7) and (8). If we let the multipliers on the constraints be denoted respectively by $\lambda_{tj}^{\mathrm{SP}}$, $\gamma_{Ktj}^{\mathrm{SP}}$, $\gamma_{Htj}^{\mathrm{SP}}$, and $\gamma_{Nt}^{\mathrm{SP}}$, the first-order conditions are

$$(1-\delta)\delta^t N_t \frac{1}{C_{tj}} = \lambda_{tj}^{\mathrm{SP}}$$

$$\gamma_{Ktj}^{\mathrm{SP}} (1-\omega_j) K_{tj}^{\omega_j} X_{tj}^{-\omega_j} = \lambda_{tj}^{\mathrm{SP}}$$

$$\lambda_{tj}^{\mathrm{SP}} (1-\alpha_j-\beta_j) \frac{Y_{tj}}{u_{tj}} = \gamma_{Htj}^{\mathrm{SP}} B_j^1 H_{tj}$$

$$\lambda_{tj}^{\mathrm{SP}} \left[(1-\alpha_j-\beta_j+\varepsilon_j) \frac{Y_{tj}}{N_{tj}} - \frac{3b}{2} \left(\frac{N_{tj}}{\mu_{tj}} \right)^{1/2} \right] = \gamma_{Nt}^{\mathrm{SP}}$$

$$\lambda_{tj}^{\mathrm{SP}} \left[\frac{b}{2} \left(\frac{N_{tj}}{\mu_{tj}} \right)^{3/2} - (\varepsilon_j + \gamma_j) \frac{Y_{tj}}{\mu_{tj}} \right] = 0$$

$$E_t \left\{ \lambda_{t+1j}^{\mathrm{SP}} \beta_j \frac{Y_{t+1j}}{K_{t+1j}} + \gamma_{Kt+1j}^{\mathrm{SP}} \omega_j K_{t+1}^{\omega_j - 1} X_{t+1j}^{1-\omega_j} \right\} = \gamma_{Ktj}^{\mathrm{SP}}$$

$$E_t \left\{ \lambda_{t+1j}^{\mathrm{SP}} (\alpha_j + \gamma_j) \frac{Y_{t+1j}}{H_{t+1j}} + \gamma_{Ht+1j}^{\mathrm{SP}} [B_j^0 + (1-u_{t+1j}) B_j^1] \right\} = \gamma_{Htj}^{\mathrm{SP}}.$$

To show that this allocation is equivalent to the one attained in the competitive equilibrium we need to compare this set of conditions with the corresponding set of conditions for the competitive equilibrium. This is what we turn to next.

1. Households optimize: The household's problem is to maximize (5) subject to sequences of flow budget constraints (4), the laws of motion for human and physical capital (8) and (9), and the constraint on labour allocation (7). Letting λ_t^{HH} be the multipliers on the budget constraints, $\gamma_{Ktj}^{\mathrm{HH}}$ and $\gamma_{Htj}^{\mathrm{HH}}$ be those on physical and human capital accumulation, and $\gamma_{Nt}^{\mathrm{HH}}$ be that on labour supply, the first-order conditions of the household are

$$(1-\delta)\delta^t N_t \frac{1}{C_{tj}} = \lambda_t^{\mathrm{HH}} p_{tj}$$

$$\gamma_{Ktj}^{\mathrm{HH}}(1-\omega_j)K_{tj}^{\omega_j} X_{tj}^{-\omega_j} = \lambda_t^{\mathrm{HH}} p_{tj}$$

$$\lambda_t^{\mathrm{HH}} w_{tj} N_{tj} = \gamma_{Htj}^{\mathrm{HH}} B_j^1 H_{tj}$$

$$\lambda_t^{\mathrm{HH}}\{p_{tj}[T_{tj} - \mathrm{ACC}_{tj} - \mathrm{AR}_{tj}] + w_{tj}u_{tj}\} = \gamma_{Nt}^{\mathrm{HH}}$$

$$E_t\left\{\lambda_{t+1}^{\mathrm{HH}} r_{t+1j} + \gamma_{Kt+1j}^{\mathrm{HH}}\omega_j K_{t+1j}^{\omega_j - 1} X_{t+1j}^{1-\omega_j}\right\} = \gamma_{Ktj}^{\mathrm{HH}}$$

$$E_t\left\{\lambda_{t+1}^{\mathrm{HH}} s_{t+1j} + \gamma_{Ht+1j}^{\mathrm{HH}}[B_j^0 + (1-u_{t+1j})B_j^1]\right\} = \gamma_{Htj}^{\mathrm{HH}}.$$

2. Firms optimize: So equations (1)–(3) hold.
3. Developer choices and free entry: The relevant first-order conditions from the developer's problem after some rearranging can be expressed as

$$\tau_{tj}^k \frac{r_{tj}}{p_{tj}} = 0,$$

$$\tau_{tj}^h \frac{s_{tj}}{p_{tj}} = \gamma_j \frac{Y_{tj}}{H_{tj}},$$

$$T_{tj} = \varepsilon_j \frac{Y_{tj}}{N_{tj}}.$$

Notice that, as expected, the subsidy on capital is 0 since there is no externality on capital. The zero profit condition is then given by

$$T_{tj} = \frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{1/2} - \gamma_j \frac{Y_{tj}}{N_{tj}}.$$

Substituting the last first-order condition, we obtain

$$\frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{1/2} = (\varepsilon_j + \gamma_j)\frac{Y_{tj}}{N_{tj}},$$

which is exactly the first-order condition of the SPP with respect to μ_{tj}. Using the second first-order condition and the fact that firms choose human capital optimally, we know that $\tau_{tj}^h = \gamma_j/(\alpha_j + \gamma_j)$.

4. Markets clear: So (7) and

$$C_{tj} + X_{tj} + bN_{tj}^{3/2}\mu_{tj}^{-1/2} = Y_{tj},$$

are satisfied.

In order to establish the equivalence, it is sufficient to establish that the first-order conditions of each set of problems are multiples of each other (*i.e.* it is sufficient to establish the existence of the appropriate set of Lagrãnge multipliers in each case). The equivalences follow easily. Comparing the social planner's first-order condition in C_{tj} with that of the household, we must have $\lambda_{tj}^{\mathrm{SP}} = \lambda_t^{\mathrm{HH}} P_{tj}$. Looking at first-order conditions in investment, we get $\lambda_{tj}^{\mathrm{SP}}/\gamma_{Ktj}^{\mathrm{SP}} = \lambda_t^{\mathrm{HH}} p_{tj}/\gamma_{Ktj}^{\mathrm{HH}}$, which, using the first equivalence, implies $\gamma_{Ktj}^{\mathrm{SP}} = \gamma_{Ktj}^{\mathrm{HH}}$. Looking at the first-order condition in u_{tj} we get from the household's equation $B_j^1 H_{tj} = (\lambda_t^{\mathrm{HH}}/\gamma_{Htj}^{\mathrm{HH}})w_{tj}N_{tj}$. Substituting for w_{tj} and rearranging, this implies $\gamma_{Htj}^{\mathrm{SP}} = \gamma_{Htj}^{\mathrm{HH}}$.

Using these results along with the first-order condition of the firm, we can easily establish the equivalence of the first-order condition with respect to capital. In order to establish the equivalence of the human capital Euler equation of the planner's and household's problem, substitute in the latter the first-order condition of the developer's problem. All that remains is to establish the city part of the problem. From the SPP, we have the first-order conditions in N_{tj} and μ_{tj}. From the competitive problem, we have the household's first-order condition in N_{tj} combined with the developer's free entry and optimality conditions. From the household's first-order condition, imposing free entry of developers, we get

$$\frac{w_{tj}}{p_{tj}}u_{tj} - \mathrm{ACC}_{tj} - \gamma_j \frac{Y_{tj}}{N_{tj}} = \frac{\gamma_{Nt}^{\mathrm{HH}}}{p_{tj}\lambda_t^{\mathrm{HH}}}.$$

Substituting for real wages and the result of the property developer's problem, we obtain

$$(1-\alpha_j-\beta_j-\varepsilon_j)\frac{Y_{tj}}{N_{tj}} - \frac{3b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{1/2} = \frac{\gamma_{Nt}^{\mathrm{HH}}}{p_{tj}\lambda_t^{\mathrm{HH}}}.$$

This latter equation is the same as the first-order condition for N_{tj} from the SPP under the equivalence $\gamma_{Nt}^{\mathrm{HH}}/(p_{tj}\lambda_t^{\mathrm{HH}}) = \gamma_{Nt}^{\mathrm{SP}}/\lambda_{tj}^{\mathrm{SP}}$. ‖

Proof of Proposition 3. To show that the growth process of city sizes satisfies Gibrat's law, note that in the first case, we have that

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = 2[\ln(A_{t+1j}) - \ln(A_{tj})] - 2\hat{\alpha}_j[\ln(N_{t+1}) - \ln(N_t)]$$
$$+2\hat{\alpha}_j \ln(B_j^0 + (1-u_j^*)B_j^1),$$

which varies with j but is independent of city size, as $E[\ln(A_{t+1j}) \mid \ln(A_{tj})]$ is independent of $\ln(A_{tj})$.

In the second case, we have

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = 2[\ln(A_{t+1j}) - \ln(A_{tj})] + 2[\ln(K_{t+1j}) - \ln(K_{tj})],$$

but under these conditions $K_{t+1j} = X_{tj} = x_j Y_{tj} = x_j F_j A_{tj} K_{tj} u_{tj}^{\hat{\phi}_j}$, which implies, as N_{tj} is constant, that

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = 2\ln(A_{t+1j}) + 2\ln(x_j F_j u_{tj}^{\hat{\phi}_j}).$$

This process is independent of city size. Hence, if the conditions in either case one or two are satisfied, city growth satisfies Gibrat's law.

To show the converse, note that if the growth process satisfies Gibrat's law, it must be of form $\ln(N_{t+1j}/\mu_{t+1j}) - \ln(N_{tj}/\mu_{tj}) = C_G + \ln\varepsilon_t$, where ε_t is i.i.d. Given the growth process of city sizes derived in the text, this implies that we must be in one of the two cases described in the proposition.

To show that if the growth process satisfies Gibrat's law the size distribution of cities satisfies Zipf's law, start with the process $\ln(N_{t+1j}/\mu_{t+1j}) - \ln(N_{tj}/\mu_{tj}) = \xi_j$. This summarizes the growth processes derived for both cases above when ξ_j is i.i.d. In order to prove convergence to a unique invariant distribution, we impose a lower bound, f_j, on the normalized process of city growth, s_j (as in Gabaix, 1999*a*, among others). We study the invariant distribution that results as the lower bound tends to 0. Specifically, let

$$s_{t+1j} = \max\left\{\frac{N_{t+1j}}{\mu_{t+1j}}/\bar{s}_{tj},\, f_j\right\},$$

where

$$\bar{s}_{tj} = \frac{1}{G_j}\sum_{i=1}^{G_j}\frac{N_{tj}}{\mu_{tj}},$$

and G_j is the number of industries with the same *ex ante* technology as industry j. Since this argument holds for all industries in this group we suppress j in the notation whenever it is clear by the context. Then $s_{t+1} = s_t\xi$, and letting $\hat{s} = \ln s$, this implies $\hat{s}_{t+1} = \hat{s}_t + \ln\xi$. Hence if $q(s)$ is the stationary probability of a representative city in the industry having size s, the stationary probability of a representative city having log-size $\hat{s}$ is given by $\hat{q}(\hat{s}) = e^{\hat{s}}q(e^{\hat{s}})$.

The master equation for this probability distribution, above the lower bound, is of the form

$$\hat{q}(\hat{s}, t+1) - \hat{q}(\hat{s}, t) = \int_{\xi} q^{\xi}(\xi)\hat{q}(\hat{s} - \ln\xi, t)d\xi - \hat{q}(\hat{s}, t),$$

where $q^{\xi}(\xi)$ denotes the probability of the growth rate taking the value ξ, and $\hat{q}(\hat{s}, t)$ denotes the distribution of $\hat{s}$ at time t. Standard results (see, for example, Levy and Solomon, 1996; Malcai *et al.*, 1999) then imply that the only asymptotic stationary solution of the master equation is of the form $\hat{q}(\hat{s}) = Me^{-\eta\hat{s}}$, for some M and η to be determined. This implies that $q(s) = Ms^{-1-\eta}$. Using the normalization

$$\int_f^G sq(s)ds = 1,$$

and the fact that $q(s)$ is a probability distribution,

$$\int_f^G q(s)ds = 1,$$

we can derive an implicit equation that determines η given by

$$G = \frac{\eta - 1}{\eta}\left[\frac{\left(\frac{f}{G}\right)^{\eta} - 1}{\left(\frac{f}{G}\right)^{\eta} - \left(\frac{f}{G}\right)}\right].$$

For finite G and sufficiently small values of f, the above expression is well approximated by

$$G \simeq \frac{1-\eta}{\eta}\left(\frac{f}{G}\right)^{-\eta}.$$

Taking natural logarithms and rearranging we obtain

$$\eta \simeq \frac{\ln G - \ln\left(\frac{1-\eta}{\eta}\right)}{\ln\left(\frac{G}{f}\right)},$$

and so as the barrier f goes to 0, η converges to 0. To understand the last step suppose, on the contrary, that $\left|\lim_{f\to 0}\eta\right| = \bar{\eta} < \infty$ where $\bar{\eta} \neq 1$. Then $\lim_{f\to 0}\ln\left(\frac{1-\eta}{\eta}\right) = \ln\left(\frac{1-\bar{\eta}}{\bar{\eta}}\right) < \infty$ and so by the previous equation $\lim_{f\to 0}\eta = 0$: a contradiction. Hence, either $\bar{\eta} = 0$ or 1, but for $\bar{\eta} = 1$, $G \simeq \frac{1-\eta}{\eta}\left(\frac{f}{G}\right)^{-\eta}$ implies $G = 0$: a contradiction. Hence, $\bar{\eta} = 0$. Thus, we obtain that $q(s) = M/s$.

So far we have only considered the size distribution of *representative cities* within a group. To get the size distribution of *cities* within a group, we need to consider that each industry may have many cities. In particular, given $\bar{s}_j$ and N_j for a group, an industry with representative city size normalized to s_j has $N_j\bar{s}_j/s_j$ cities. The term $N_j\bar{s}_j$ is constant across industries within a group, and hence the size distribution of cities, not representative cities, is given by $q^{\text{City}}(\varsigma) = \hat{M}/\varsigma^2$, for some $\hat{M}$ finite. The cumulative distribution function is then given by

$$Q^{\text{City}}(\varsigma > \bar{\varsigma}) = \int_0^{\bar{\varsigma}} \hat{M}\frac{1}{\varsigma^2}d\varsigma = \frac{\hat{M}}{\bar{\varsigma}},$$

which is a statement of Zipf's law for that group.

To obtain the size distribution of cities for the economy as a whole, notice first that the argument above implies that the cumulative distribution of cities in that group is given by $Q_i^{\text{City}}(\varsigma > \bar{\varsigma}) = \hat{M}_i/\bar{\varsigma}$, where i indexes industry groups (assume the total number of groups is given by $\bar{G}$). Using this, and if λ_i is the proportion of cities in group i, the cumulative distribution function for the economy is

$$Q^{\text{City}}(\varsigma > \bar{\varsigma}) = \sum_{i=1}^{\bar{G}}\lambda_i\frac{\hat{M}_i}{\bar{\varsigma}} = \left[\sum_{i=1}^{\bar{G}}\lambda_i\hat{M}_i\right]\frac{1}{\bar{\varsigma}},$$

which is a statement of Zipf's law for the economy. ‖

Proof of Proposition 4. We have that city growth rates are given by

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = 2[\ln(A_{t+1j}) - \ln(A_{tj})] - 2(\hat{\alpha}_j + \hat{\beta}_j)[\ln(N_{t+1}) - \ln(N_t)]$$
$$+2\hat{\alpha}_j\ln(B_j^0 + (1-u_j^*)B_j^1) + 2\hat{\beta}_j[\ln(K_{t+1j}) - \ln(K_{tj})].$$

The only places that productivity shocks enter this equation is through their contemporaneous effects on output and through the accumulation of past capital. If we examine the equation for capital accumulation, recursively substituting,

we find, ignoring all other terms, that the effect of productivity shocks is given by

$$2\left[\ln(A_{t+1j})+(\hat{\beta}_j(1-\omega_j)-1)\ln(A_{tj})-\hat{\beta}_j\sum_{T=1}^{t}\frac{(\omega_j+(1-\omega_j)\hat{\beta}_j)^{t-T}}{(1-(\omega_j+(1-\omega_j)\hat{\beta}_j))^{-1}}(1-\omega_j)\ln(A_{T-1j})\right]$$

$$=2\left[\ln(A_{t+1j})+(\hat{\beta}_j(1-\omega_j)-1)\ln(A_{tj})-\hat{\beta}_j(1-(\omega_j+(1-\omega_j)\hat{\beta}_j))(1-\omega_j)\right.$$

$$\left.\times\sum_{T=1}^{t}(\omega_j+(1-\omega_j)\hat{\beta}_j)^{t-T}\ln(A_{T-1j})\right].$$

Now if we examine only the weights on the lagged productivity shocks, we find that

$$\hat{\beta}_j(1-(\omega_j+(1-\omega_j)\hat{\beta}_j))(1-\omega_j)\sum_{T=1}^{t}(\omega_j+(1-\omega_j)\hat{\beta}_j)^{t-T}=\hat{\beta}_j(1-(\omega_j+(1-\omega_j)\hat{\beta}_j)^{t-1})(1-\omega_j).$$

If we take limits into the infinite past, so as to remove the effect of initial conditions, this expression reduces to $\hat{\beta}_j(1-\omega_j)$, so that the weights on past productivity shocks sum to -1.

From this we can conclude that if the city type is of average size, defined as having experienced a sequence of past shocks whose weighted average is $E(\ln A)$, then the expected growth rate of the city is 0. By contrast, if the past shocks have a weighted average greater than (less than) $E(\ln A)$, then the expected growth rates are negative (positive).

We next turn to the proof of existence of an invariant distribution. The first step is to bound the space of city sizes. For this, first note that city sizes depend on output per capita, and given A_{tj}, Y_{tj}/N_{tj} is bounded only if human capital and labour do not grow permanently. So normalize city size by the amount of human capital in the industry and population, or alternatively, let $g_{H_j}+g_N=0$. Then, in steady state the growth rate of output per capita in all industries is 0. Denote by $y_j^{SS}(A)$ steady-state output per capita in industry j given a sequence of productivity shocks all equal to A. Then city sizes are such that there exists a $t^*(\varepsilon)$ such that for $t>t^*(\varepsilon)$,

$$\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\in[c_{\tilde{N}}+\ln y_j^{SS}(\underline{A}_j)-\varepsilon,c_{\tilde{N}}+\ln y_j^{SS}(\overline{A}_j)+\varepsilon]\equiv LN,$$

where $c_{\tilde{N}}$ is some constant that depends on the parameters of the model. For what follows, without loss of generality assume that $t>t^*(\varepsilon)$. To simplify notation let $N_{tj}/\mu_{tj}=\tilde{N}_{tj}$. Since the argument can be made industry by industry we also drop the industry subindex.

Define the function $g(\cdot)$ using the evolution of $\tilde{N}$ as

$$g(\tilde{N},A^t,A_{t+1})\equiv\ln\tilde{N}+2[\ln(A_{t+1})-\ln(A_t)]$$

$$+2(1-\omega)\hat{\beta}\left[\ln(A_t)-\sum_{s=1}^{\infty}\frac{(\omega+(1-\omega)\hat{\beta})^{s-1}}{(1-(\omega+(1-\omega)\hat{\beta}))^{-1}}\ln(A_{t-s})\right]. \tag{16}$$

This lies in the compact set LN defined above. Let ϕ be the probability measure over A. Then, the probability of a transition from a point $\tilde{N}$ to a set S is given by

$$Q(\tilde{N},S)=\phi(A:g(\tilde{N},A^t,A)\in S),$$

where A^t denotes the sequence of productivity levels up to t. For any function $f:LN\to\mathbb{R}$ define the operator T by

$$(Tf)(\tilde{N})=\int_{LN}f(\tilde{N}')Q(\tilde{N},d\tilde{N}')=\int_{\underline{A}}^{\overline{A}}f(g(\tilde{N},A^{t+1}))d\phi^{t+1}(A^{t+1}).$$

where $\phi^{t+1}(A^{t+1})$ denotes the probability measure of a sequence A^{t+1}. Define also the operator T^*, which maps the probability of being in a set S next period given the current distribution, say λ, as

$$(T^*\lambda)(S)=\int_{LN}Q(\tilde{N},S)\lambda(d\tilde{N}).$$

Since the set LN is compact, we are able to use Theorem 12.12 in Stokey, Lucas and Prescott (1989) to prove that there exists a unique invariant distribution, if we can show that the transition probability function Q satisfies the Feller property, is monotone, and satisfies the mixing condition.

To see that it satisfies the Feller property, note that the function g is continuous in $\ln \tilde{N}$, and $\ln A^{t+1}$. Since g is continuous and bounded, if f is continuous and bounded, $f(g(\cdot))$ will be continuous and bounded and therefore so is Tf. Hence T maps the space of bounded continuous functions into itself, $T : C(\bar{S}) \to C(\bar{S})$. To see that it is monotone, we need to prove that if $f : LN \to \mathbb{R}$ is a non-decreasing function, then so is Tf. But this follows from the fact that the g is non-decreasing in $\tilde{N}$. Hence $f(g(\tilde{N}, A^{t+1}))$ is non-decreasing in $\tilde{N}$ and therefore so is Tf.

Finally, to show that it satisfies the mixing condition, we need to show that there exists $c \in LN$, T and $\eta > 0$ such that

$$Q^T(c_{\tilde{N}} + \ln y_j^{SS}(\underline{A}_j) - \varepsilon, [c, c_{\tilde{N}} + \ln y_j^{SS}(\overline{A}_j) + \varepsilon]) \geq \eta,$$

and

$$Q^T(c_{\tilde{N}} + \ln y_j^{SS}(\overline{A}_j) + \varepsilon, [c_{\tilde{N}} + \ln y_j^{SS}(\underline{A}_j), c]) \geq \eta,$$

where the superindex T denotes the number of steps. Let $c = c_{\tilde{N}} + \ln y_j^{SS}((\underline{A}_j + \overline{A}_j)/2) + \varepsilon$. As g is continuous and increasing in A, there exists a sequence of identical shocks $A' > (\underline{A}_j + \overline{A}_j)/2$ of length T', large enough, such that $g^{T'}(\tilde{N}, A^t, A'^{t+T'}) > c$. This, since we know that $g^T(\tilde{N}, A^t, A'^{t+T'})$ converges to $c_{\tilde{N}} + \ln y_j^{SS}(A') + \varepsilon > c$ when $T \longrightarrow \infty$. Let $\eta' = \phi^{T'}(A') > 0$. Similarly there exists a sequence of shocks of $A'' < (\underline{A}_j + \overline{A}_j)/2$ of length T'' such that $g^{T''}(\tilde{N}, A^t, A''^{t+T''}) < c$. Let $\eta'' = \phi^{T''}(A'')$. Then if $\eta = \min\{\eta', \eta''\}$ and $T = \max\{T', T''\}$. Then c, T, and η guarantee that the mixing condition holds. Theorem 12.12 in Stokey *et al.* (1989) then guarantees that there exists a unique invariant distribution and that the iterates of T^* converge weakly to that invariant distribution.

The final step is to prove that the invariant distribution has thinner tails than a Pareto distribution with coefficient 1. But this is immediate from Proposition 3 and the fact that outside those cases the city growth process exhibits reversion to the mean (Proposition 4). Hence, in these cases the operator T^* maps the Pareto distribution with coefficient 1 into distributions with thinner tails, and so its fixed point must have thinner tails too. ‖

Proof of Proposition 5. If conditions 1 and 2 in Proposition 3 are not satisfied, the variance of the log of city sizes is given by

$$V_0\left[\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\right] = o(V_0[\ln(A_{tj})] + \hat{\beta}_j^2 V_0[\ln(K_{tj})]),$$

were o is a constant that depends on the parameters of the model, and

$$V_0[\ln K_{tj}] = V_0\left[\sum_{T=1}^{t}(\omega_j + (1-\omega_j)\hat{\beta}_j)^{t-T}(1-\omega_j)\ln(A_{T-1j})\right].$$

If shocks are i.i.d. with variance v, we obtain

$$V_0[\ln K_{tj}] = v\left[\sum_{T=1}^{t}(\omega_j + (1-\omega_j)\hat{\beta}_j)^{t-T}(1-\omega_j)\right]^2,$$

or as $t \to \infty$, $V_0[\ln K_{tj}] = v/(1+\hat{\beta}_j)^2$, so that the variance of the long-run city size distribution is given by

$$V_0\left[\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\right] = ov\left[1 + \frac{\hat{\beta}_j^2}{(1+\hat{\beta}_j)^2}\right],$$

which is increasing in v, thereby proving the result.

If shocks are not i.i.d., a higher unconditional variance implies that $V_0[\ln K_{tj}]$ is larger, since $(\omega_j + (1-\omega_j)\hat{\beta}_j)^{t-T}$ is positive for every $1 > \omega_j > 0$ and $1 > \hat{\beta}_j > 0$. Higher unconditional variance implies that $V_0[\ln(A_{tj})]$ is larger for every t, and so the variance of city sizes increases. ‖

Acknowledgements. We thank the editor, two anonymous referees, Vernon Henderson, Yannis Ioannides, Chad Jones, Narayana Kocherlakota, Dirk Krueger, Robert E. Lucas, Jr., and numerous seminar participants for comments, and Yannis Ioannides and Linda Dobkins for sharing their data.

REFERENCES

AU, C. and HENDERSON, V. (2006), "How Migration Restrictions Limit Agglomeration and Productivity in China", *Journal of Economic Development*, **80** (2), 350–388.
AUERBACH, F. (1913), "Das Gesetz der Bev ölkerungskonzentration", *Petermanns Geographische Mitteilungen*, **59**, 74–76.
BLACK, D. and HENDERSON, V. (1999), "A Theory of Urban Growth", *Journal of Political Economy*, **107** (2), 252–284.
BLANK, A. and SOLOMON, S. (2000), "Power Laws in Cities Population, Financial Markets and Internet Sites (Scaling in Systems with a Variable Number of Components)", *Physica A*, **287** (1–2), 279–288.
CHAMPERNOWNE, D. G. (1953), "A Model of Income Distribution", *Economic Journal*, **63** (250), 318–351.
CORDOBA, J. (2004), "On the Distribution of City Sizes" (Working Paper, Rice University).
DOBKINS, L. H. and IOANNIDES, Y. M. (2001), "Spatial Interactions among U.S. Cities: 1900–1990", *Regional Science and Urban Economics*, **31** (6), 701–731.
DURANTON, G. (2007), "Urban Evolutions: The Fast, the Slow, and the Still", *The American Economic Review* (forthcoming).
EATON, J. and ECKSTEIN, Z. (1997), "Cities and Growth: Theory and Evidence from France and Japan", *Regional Science and Urban Economics*, **27**, 443–474.
EECKHOUT, J. (2004), "Gibrat's Law for (all) Cities", *American Economic Review*, **94** (5), 1429–1451.
GABAIX, X. (1999*a*), "Zipf's Law for Cities: An Explanation", *Quarterly Journal of Economics*, **114** (3), 739–767.
GABAIX, X. (1999*b*), "Zipf's Law and the Growth of Cities", *American Economic Review Papers and Proceedings*, **89** (2), 129–132.
GABAIX, X. and IOANNIDES, Y. (2003), "The Evolution of City Size Distributions", in J. V. Henderson and J. F. Thisse (eds.) *Handbook of Economic Geography* (Amsterdam: North-Holland) 2341–2380.
HENDERSON, V. (1974), "The Sizes and Types of Cities", *American Economic Review*, **64** (4), 640–656.
HENDERSON, V. and WANG, H. (2005), "Urbanization and City Growth" (Working Paper, Brown University).
HERCOWITZ, Z. and SAMPSON, M. (1991), "Output Growth, the Real Wage, and Employment Fluctuations", *American Economic Review*, **81** (5), 1215–1237.
HORVATH, N. (2000), "Sectoral Shocks and Aggregate Fluctuations", *Journal of Monetary Economics*, **45** (1), 69–106.
IOANNIDES, Y. M. and OVERMAN, H. G. (2003), "Zipf's Law for Cities: An Empirical Examination", *Regional Science and Urban Economics*, **33** (2), 127–137.
JONES, C. I. (1999), "Growth: With and Without Scale Effects", *American Economic Review Papers and Proceedings*, **89** (2), 139–144.
KALECKI, M. (1945), "On the Gibrat Distribution", *Econometrica*, **13** (2), 161–170.
LEVY, M. and SOLOMON, S. (1996), "Power Laws are Logarithmic Boltzmann Laws", *International Journal of Modern Physics C*, **7** (4), 595–600.
LUCAS, R. E., Jr. (1988), "On the Mechanics of Economic Development", *Journal of Monetary Economics*, **22** (1), 3–42.
LUCAS, R. E. and PRESCOTT, E. C. (1971), "Investment Under Uncertainty", *Econometrica*, **39** (3), 659–681.
MALCAI, O., BIHAM, O. and SOLOMON, S. (1999), "Power-Law Distributions and Levy-Stable Intermittent Fluctuations in Stochastic Systems of Many Autocatalytic Elements", *Physical Review E*, **60** (2), 1299–1303.
ROMER, P. (1990), "Endogenous Technological Change", *Journal of Political Economy*, **98** (5), S71–S102.
ROSEN, K. and RESNICK, M. (1980), "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy", *Journal of Urban Economics*, **8** (2), 156–186.
SOO, K. T. (2005), "Zipf's Law for Cities: A Cross Country Investigation", *Regional Science and Urban Economics*, **35** (3), 239–263.
STOKEY, N., LUCAS, R. E., Jr. and PRESCOTT, E. (1989) *Recursive Methods in Economic Dynamics* (Cambridge, MA: Harvard University Press).