NBER WORKING PAPER SERIES

URBAN STRUCTURE AND GROWTH

Esteban Rossi-Hansberg
Mark L.J. Wright

Working Paper 11262
http://www.nber.org/papers/w11262

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2005

We thank Vernon Henderson, Yannis Ioannides, Chad Jones, Narayana Kocherlakota, Dirk Krueger, Robert E. Lucas, Jr., and numerous seminar participants for comments, and Yannis Ioannides, Linda Dobkins and Romain Wacziarg for sharing their data. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Urban Structure and Growth
Esteban Rossi-Hansberg and Mark L.J. Wright
NBER Working Paper No. 11262
April 2005
JEL No. E0, O4, R0

ABSTRACT

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 theory of economic growth in an urban environment. We show that the urban structure is the margin that eliminates local increasing returns to yield constant returns to scale in the aggregate, which is sufficient to deliver balanced growth. In a multi-sector economy with specific factors and productivity shocks, the same mechanism leads to a city size distribution that is well described by a power distribution with coefficient one: Zipf's Law. Under certain assumptions our theory produces Zipf's Law exactly. More generally, it produces the systematic deviations from Zipf's Law observed in the data, including the under-representation of small cities and the absence of very large ones. In general, the model identifies the standard deviation of industry productivity shocks as the key parameter determining dispersion in the city size distribution. We present evidence that the relationship between the dispersion of city sizes and the variance of productivity shocks is consistent with the data.

Esteban Rossi-Hansberg
Stanford University
Department of Economics
579 Serra Mall
Stanford, CA 94305-6072
and NBER
erossi@stanford.edu

Mark L.J. Wright
Stanford University
mlwright@stanford.edu

1. INTRODUCTION

Aggregate economic activity is primarily *urban* economic activity. For example, in the United States at the turn of the millennium, 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.

In the absence of aggregate constant returns to scale, long run growth rates in income per capita either explode or tend to zero.[1] An endogenous urban structure is, however, sufficient to generate balanced growth in the presence of local increasing returns. To see this, note that the size of cities is determined by the trade-off between agglomeration effects and congestion costs. In our theory, this trade-off is affected by 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 evolution of the urban structure that produces linear aggregate production functions in a world with urban scale effects.

Theories that use this mechanism to generate balanced growth face the challenge of being consistent with a number of well-established empirical regularities about the size distribution of cities. To address these facts, we embed this mechanism in a multi-sector economy with industry specific factors and productivity shocks in which

[1]Specifically, the production set of the aggregate economy is, asymptotically, a convex cone. In both exogenous growth models, 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)).

the size distribution of cities depends on the allocation of industries across cities. The growth, birth and death of these cities in turn depends upon the evolution of productivity shocks and the way they are propagated through the accumulation of industry specific factors. We show that under two polar sets of assumptions this mechanism delivers the stylized empirical regularity known as Zipf's Law of cities: The rank of a city is inversely proportional to its size. Zipf's Law is, however, only an approximate description of the data. To address these discrepancies, we analyze the implications of our theory away from these two polar cases and show that the mechanism delivers precisely the systematic deviations from Zipf's Law observed in the data.

The main step in establishing the implications of our theory for city size distributions is to demonstrate the ability of this mechanism to produce Gibrat's Law of cities: the mean and variance of the growth rate of a city are independent of its size. In our framework, cities result out of the trade-off between commuting costs and local production externalities in human capital and labor. Industry specific externalities imply that cities specialize in an industry, and so all cities operating in an industry have the same size. The interaction between commuting costs and the urban production externality leads to a city size that varies only with changes in the average product of labor in the city (and hence industry). In response to a positive productivity shock, cities grow, and the number of cities operating in an industry falls as long as employment in the industry changes less than proportionately.

To see how the mechanism generates Gibrat's Law, first consider a simple economy in which the only factors of production are labor and human capital both growing at constant rates. In such an economy, the growth rate of the average product of labor, 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 labor 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 labor, and hence

also in the size of cities. In this case, we also obtain a scale independent growth process for cities. After establishing Gibrat's Law for these polar cases, we combine the growth, entry, and exit processes to show that, in these cases, the invariant distribution of city sizes satisfies Zipf's Law.[2]

Apart from these polar cases, productivity shocks affect the distribution of city sizes both directly and indirectly through their effect on factor accumulation. This implies a scale dependent growth process for cities. The bulk of the paper is devoted to a study of the interaction of these effects and their ability to produce a number of robust deviations from Zipf's Law observed in the data. One of the most notable is that, relative to Zipf's Law, small cities are under-represented and the largest cities are not 'large enough.' A second is that there is some systematic variation in the dispersion of city sizes across countries. We show that our theory is able to produce these robust deviations from Zipf's Law in between the two polar cases discussed above. 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 we show leads to the under-representation of small cities and the absence of very large ones. We also show that the model identifies the standard deviation of industry productivity shocks as the key element determining dispersion in the size distribution of cities across countries.

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

[2]The relationship between Gibrat's Law and Zipf's Law has been previously studied in both the physics (for example, Levy and Solomon (1996), Malcai, Biham and Solomon (1999), and Blank and Solomon (2000)) and economics (for example, Gabaix (1999a) and Cordoba (2003)) literatures. In contrast to these papers, our proof emphasizes the interaction between Gibrat's Law and the process of entry and exit in producing Zipf's Law.

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. Black and Henderson (1999) and Eaton and Eckstein (1997) 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.

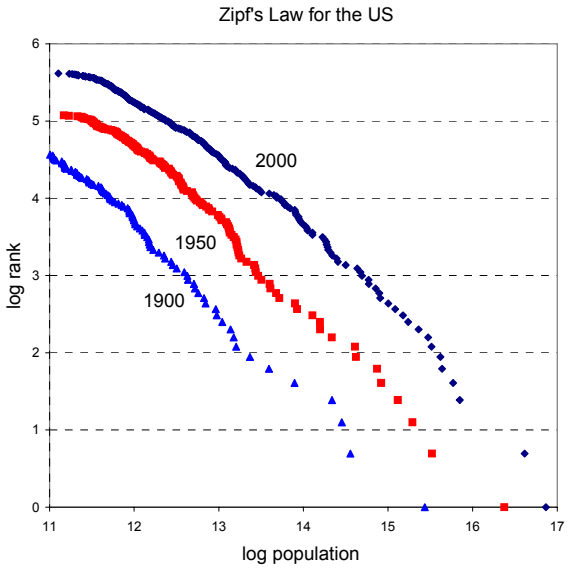

Figure One

Following the original paper of Auerbach (1913), a substantial literature has arisen that investigates the empirical foundations of Zipf's Law. A number of authors, including Rosen and Resnick (1980), Dobkins and Ioannides (2000), Ioannides and Overman (2001), and Soo (2003) have documented the robustness of this phenomenon

both over time and across countries. This is illustrated in Figure One for the United States, where Zipf's Law appears to be as good a description of the size distribution of cities at the turn of the Twenty-First century as it was at the turn of the Twentieth. Further, as illustrated in Figures Two 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.

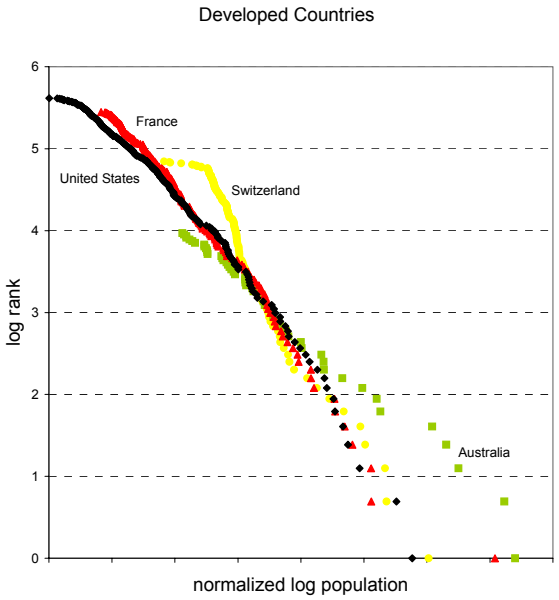

Figure Two A

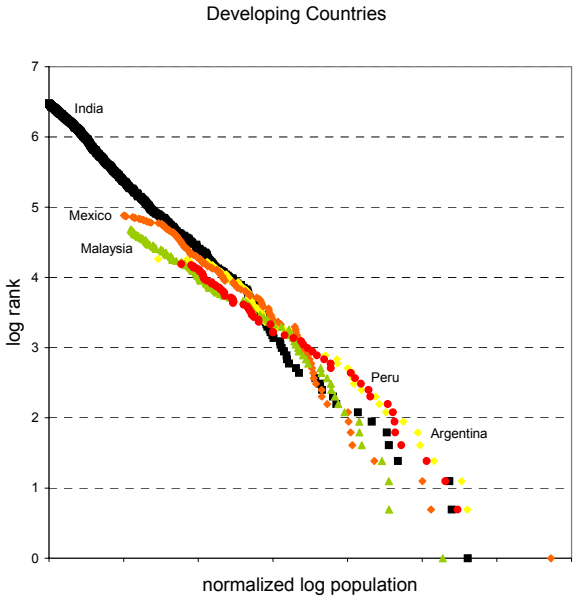

Figure Two B

Much recent empirical work on the size distribution of cities (surveyed in Gabaix and Ioannides (2003)) has emphasized the fact that there are systematic deviations from Zipf's Law. One of the most robust is the under-representation of small cities and the absence of very large ones, which is illustrated in Figures One and Two as a broad tendency for the relationship to be slightly concave, at least once one controls for a country's capital city (see also Eeckhout (2004)). A second, as shown in Figure Two, is that 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. These deviations from Zipf's Law are precisely the ones emphasized in the discussion above.

Finally, this paper is related to a number of proposed explanations of Zipf's Law. A large number of papers, including most notably Champernowne (1953), Kalecki (1945), Levy and Solomon (1996), Malcai, Biham and Solomon (1999), Gabaix (1999a), Blank and Solomon (2000) and Cordoba (2003), have studied statistical processes that generate Zipf's Law. In economics, Gabaix (1999a,b), Cordoba (2003) and Eeckhout (2004) have presented models in which cities grow as labor migrates as the direct response to city amenity, taste or productivity shocks. Neither paper generates the existence of cities endogenously, and in all three the city growth process directly inherits the behavior of the exogenously specified shocks. In a recent study, Duranton (2002) presents a quality ladder model of growth which, under certain assumptions on the location and mobility of new firms, produces a size distribution of cities that matches some aspects of the data. In contrast, our paper focuses upon the relationship between factor accumulation, productivity shocks and the urban structure. Importantly, it is endogenous city formation that both eliminates scale effects in growth and provides an alternative 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. An 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 labor 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 assume log-linear preferences and Cobb-Douglas production functions so

that both the growth path and the city size distribution can be solved in closed form.

Cities

Our approach to modeling 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 central business district (CBD). It is assumed that every agent that works at the CBD must reside in the area surrounding the city. Locations closer to the CBD are more desirable because they involve a shorter commute to work. Specifically, we assume that the cost of commuting is linear in the distance travelled, and we let τ be the cost per mile of commuting in terms of the output of the city.

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 rents on the city edge equal to zero. Therefore, in a city of radius $\bar{z}$, rents at a distance z from the center must be given by

$$R(z) = \tau\left(\bar{z} - z\right).$$

Hence, total rents in a city of radius $\bar{z}$ are given by

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

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

$$TR = \frac{\pi\tau}{3}\left(\frac{n}{\pi}\right)^{\frac{3}{2}} = \frac{b}{2}n^{\frac{3}{2}},$$

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

$$TCC = \int_0^{\bar{z}} 2\pi z\tau z dz = bn^{\frac{3}{2}},$$

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

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

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

where $\tilde{A}_{tj}$ is the total factor productivity of an urban firm (given that good j is produced in that city), 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 labor 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 finite order Markov and is distributed according to a density function with finite moments.

We divide the original set of J industries into groups. 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. 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 fulfill different purposes, but are produced with the same ex-ante technology. We use the homogeneity of technology within a group to establish the conditions under which Zipf's Law holds for each of these groups. We then aggregate across groups to obtain Zipf's Law for the entire economy.

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_{j0}) and human (H_{j0}) capital.

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

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

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

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.

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 labor 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. Hence we can think of the household distributing N_j of its members to each industry j subject to

$$\sum_j N_{tj} \leq N_t,$$

in each period.

Workers spend time producing new human capital according to

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

where B_j^0 and B_j^1 are positive 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.

Efficient allocations

All Pareto efficient allocations are the solution of the following *Social Planning Problem*: Choose state contingent sequences $\left\{C_{tj}, X_{tj}, N_{tj}, \mu_{tj}, u_{tj}, K_{tj}, H_{tj}\right\}_{t=0,j=1}^{\infty,J}$ so as to maximize

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

subject to, for all t and j,

$$C_{tj} + X_{tj} + b\tilde{N}_{tj}^{\frac{3}{2}}\mu_{tj} \leq A_{tj}\tilde{K}_{tj}^{\beta_j}\tilde{H}_{tj}^{\alpha_j+\gamma_j}\tilde{N}_{tj}^{1-\alpha_j-\beta_j+\varepsilon_j}u_{tj}^{1-\alpha_j-\beta_j}\mu_{tj}, \tag{2}$$

$$N_t = \sum_{j=1}^{J} N_{tj} = \sum_{j=1}^{J} \mu_{tj} \tilde{N}_{tj}, \tag{3}$$

$$K_{tj} = \mu_{tj} \tilde{K}_{tj}, \tag{4}$$

$$H_{tj} = \mu_{tj} \tilde{H}_{tj}, \tag{5}$$

$$K_{t+1j} = K_{tj}^{\omega_j} X_{tj}^{1-\omega_j}, \tag{6}$$

$$H_{t+1j} = H_{tj} \left[B_j^0 + (1 - u_{tj}) B_j^1 \right]. \tag{7}$$

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, where μ_{tj} denotes the number of cities in industry j at time t.

The original problem is not a convex dynamic optimization problem. However, since the city size problem is static, we can solve it separately and transform the problem into a convex dynamic optimization problem. This allows us to prove the existence of a unique Pareto efficient allocation below.

Decentralization

In order to explain the observed city size distributions, it is necessary to consider also competitive equilibrium allocations. It is easy to introduce a competitive equilibrium framework for which the unique equilibrium allocation attains the solution of the Social Planning Problem. As is standard in the previous literature, we use city developers that internalize the urban production externality.

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. Property developers 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 (although they never choose to subsidize physical capital as there is no externality in physical capital). 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 larger than what they could obtain elsewhere. Firms produce

in the city as long as profits are nonnegative. Free entry implies that developers earn zero profits in equilibrium. 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 labor, 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. Some examples of policies that may achieve this goal in practice are subsidies to firms that employ high skilled workers, or the provision of local public goods preferred by highly educated agents (e.g. fine arts)[3].

The next two propositions establish uniqueness of the Pareto efficient allocation, and the analogs of both Welfare Theorems. Apart from the developers problem, the proofs of these propositions are standard. The details of the developers problem are presented in the appendix. The full decentralization and a detailed proof of these propositions can be found in Rossi-Hansberg and Wright (2003).

Proposition 1 *Every competitive equilibrium in this economy is Pareto efficient.*

Proposition 2 *There exists a competitive equilibrium that attains the unique Pareto efficient allocation.*

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.

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

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.

Aggregate Constant Returns

The problem of choosing the optimal sizes of cities is static: 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. Toward 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,

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

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,

$$\mu_{tj} = \left[\frac{2\left(\gamma_j+\varepsilon_j\right)}{b}\frac{Y_{tj}}{N_{tj}}\right]^{-2} N_{tj}, \tag{8}$$

and so total commuting costs satisfy

$$TCC_{tj} = 2\left(\gamma_j+\varepsilon_j\right)Y_{tj}. \tag{9}$$

Notice that we need to impose

$$\gamma_j + \varepsilon_j < \frac{1}{2},$$

since otherwise total commuting costs would be larger than total output in the industry (this assumption also guarantees that the first order condition is necessary and sufficient). To interpret this restriction, write industry output minus total commuting cost as

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

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

$$C_{tj} + X_{tj} \leq F_j \hat{A}_{tj} H_{tj}^{\hat{\alpha}_j} K_{tj}^{\hat{\beta}_j} N_{tj}^{1-\hat{\alpha}_j-\hat{\beta}_j} u_{tj}^{\hat{\phi}_j} \equiv \hat{Y}_{tj}, \tag{10}$$

where

$$\begin{aligned} F_j &= \left(1 - 2\left(\gamma_j + \varepsilon_j\right)\right) \left[\frac{2\left(\gamma_j + \varepsilon_j\right)}{b}\right]^{\frac{2\left(\gamma_j+\varepsilon_j\right)}{1-2\left(\gamma_j+\varepsilon_j\right)}}, \\ \hat{A}_{tj} &= A_{tj}^{\frac{1}{1-2\left(\gamma_j+\varepsilon_j\right)}},\ \hat{\alpha}_j = \frac{\alpha_j + \gamma_j}{1 - 2\left(\gamma_j + \varepsilon_j\right)}, \\ \hat{\beta}_j &= \frac{\beta_j}{1 - 2\left(\gamma_j + \varepsilon_j\right)},\ \text{and}\ \hat{\phi}_j = \frac{1 - \alpha_j - \beta_j}{1 - 2\left(\gamma_j + \varepsilon_j\right)}. \end{aligned}$$

Since, under our assumptions $u_{tj} \leq 1$ 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 (9) output in the industry is also a constant returns to scale function of inputs in the industry.

The constraint in (10) 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 3 *(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 labor 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 (10)) 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[4]. As productivity shocks are likely to be more frequent than changes in these patterns, one could in principle decompose their effect on total factor productivity empirically. To clarify this last point, suppose that cities are organized at a suboptimal size, either too large or too small, captured by a parameter $\kappa_j \neq 1$, such that

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

Then, output net of commuting costs would be given by equation (10) with a modified F_j given by

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

which, as can be easily checked, has a global optimum at $\kappa_j = 1$. Hence, by organizing cities inefficiently (too small *or* too large), the economy would produce with lower total factor productivity. In what follows we set $\kappa_j = 1$, since it does not affect any of the urban or growth implications of the model.

City Sizes

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

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

[4] Au and Henderson (2002) examines this possibility for the particular case of China.

That is, the planner increases the number of people in the city until the change in commuting costs per person for current residents (left hand side) is equal to the change in earnings per person for current residents (right hand side).

From this equation it is easy to see that anything that increases the level of the average product of labor 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 labor 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 labor. If commuting costs were to rise by less, or even more, than the average product of labor, the basic result that productivity shocks are translated into fluctuations in city size remains. However, one combination of assumptions that does not work is if commuting costs are denominated *purely* in units of time, *and* workers supply labor 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 labor. More generally, any combination of time and material cost 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.

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 (1) subject to (10), (3), (6), and (7). The value function of the planner has the form

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

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 are suppressed.[5] Three basic results are immediate. 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

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

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[6] (with no uncertainty),

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

[5] The details are contained in Rossi-Hansberg and Wright (2003).

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

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 $g_{\hat{Y}_j} = g_N$.

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. In particular, the growth rate of a city in industry j is given by

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

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

$$
\begin{aligned}
& \ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) \\
= \quad & \frac{2\hat{\alpha}_j}{1-\hat{\beta}_j}\left[g_{Hj} - g_N\right] + 2\left[\ln\left(A_{t+1j}\right) - \ln\left(A_{tj}\right)\right] \\
& +2\left(1-\omega_j\right)\hat{\beta}_j\left[\ln(A_{tj}) - \sum_{s=1}^{\infty}\frac{\left(\omega_j + \left(1-\omega_j\right)\hat{\beta}_j\right)^{s-1}}{\left(1-\left(\omega_j + \left(1-\omega_j\right)\hat{\beta}_j\right)\right)^{-1}}\ln(A_{t-sj})\right]. \quad (11)
\end{aligned}
$$

Equation (11) 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. We can then show that Gibrat's Law implies Zipf's Law in our framework once we modify the results on the convergence of the growth process in Levy and Solomon (1996) and Malcai, Biham and Solomon (1999) to allow for the entry and exit of cities.

In order to generate Gibrat's Law, and Zipf's Law as an invariant distribution, we need the growth processes at the city level to be independent of scale. As labor is perfectly mobile across cities and industries, this in turn requires that the marginal product of labor 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 labor, so that the growth rate of the marginal product is independent of scale[7]. These assumptions eliminate the third term in equation (11) 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 labor through the linear accumulation of physical capital. This

[7]This is essentially the mechanism at work in Gabaix (1999a,b), Cordoba (2003) and Eeckhout (2004) for an exogenous number of cities. Gabaix (1999a) and Cordoba (2003) 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 one. 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 differ in all three cases.

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.[8]

The next proposition formalizes these arguments and proves the link between Gibrat's Law and Zipf's Law in the model. 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 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 1999a). All proofs are relegated to the appendix.

Proposition 4 *(Exact Gibrat's Law and Zipf's Law) The growth process of city sizes satisfies Gibrat's Law, and the invariant distribution for city sizes satisfies Zipf's Law, if and only if one of the following two conditions is satisfied:*

1. *(No physical capital) There is no physical capital* $\left(\hat{\beta}_j = 0 \text{ or } \omega_j = 1\right)$*, and productivity shocks are permanent.*

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

Scale Dependence

Obviously, the conditions outlined in Proposition 4 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

[8]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 Zipf's Law exactly for any parameter set.

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 some variation in cross country estimates of Zipf's coefficients, with this variation positively correlated with per capita income: richer countries have a more even city size distribution (Soo (2003)).

In the next two Propositions we argue that, in general, the model produces 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 one. Eeckhout (2004) emphasizes exactly this feature of the data.

Proposition 5 *(Concavity) If conditions 1 and 2 in Proposition 4 are not satisfied, the growth rate of cities exhibits reversion to the mean. If productivity levels are bounded for all industries (so that there exist uniform bounds such that $A_{tj} \in \left[\underline{A}_j, \overline{A}_j\right]$ for all t, j), then there exists a unique invariant distribution of city sizes with thinner tails than a Pareto distribution with coefficient one.*

Unless the conditions of Proposition 4 are satisfied, variation in the standard de-

viation of productivity shocks affects the distribution of city sizes. Intuitively, given capital stocks, a larger standard deviation of shocks implies a larger standard deviation of city sizes and a larger standard deviation of investments, which in turn implies a more dispersed distribution of capital stocks. This would explain the positive correlation between Zipf's coefficients and income, documented in Soo (2003), *if* high income countries experience less volatile shocks. We formalize this intuition in the following proposition.

Proposition 6 *If conditions 1 and 2 in Proposition 4 are not satisfied, the standard deviation of city sizes increases with the standard deviation of industry shocks.*

Proposition 6 points to the standard deviation 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

This section illustrates the characterization of the urban structure presented in the previous section. Summarizing, we obtain Zipf's Law exactly if we either eliminate capital or make capital accumulation linear; in all other cases the log rank-size relationship is concave and the absolute value of the slope is negatively related to the variance of industry shocks. All the results we have presented are asymptotic; for any particular realization of the stochastic process there may be random deviations from Zipf's Law. This is illustrated in Figure Three, where we simulate the model for 100 identical industries for the case of $\omega_j = 1$ for all $j = 1, ..., J$ and permanent shocks (Case 1 of Proposition 4). Along a given sample path, Zipf's Law holds exactly, apart from stochastic deviations.

The next step is to illustrate the deviations of Zipf's Law obtained in our model when we move away from the assumptions in Proposition 4. Figure Four presents U.S. data in 2002 for MSAs, together with a numerical simulation of the model with

transitory shocks. We let the model run for 10,000 periods so that the distribution of city sizes is not changing significantly through time.

As one can see in Figure Four, the model does very well – arguably better than Zipf's Law – in matching the U.S. data. In particular, and as expected given Proposition 5, the curve is slightly concave, as in the data. That is, large cities are too small, and there are not enough small cities. Both simulations above have been computed for the particular set of parameter values collected in the following table:

$\alpha = \beta = \phi$	B	$\gamma = \varepsilon$	ω	δ	τ	g_N	m	sd
$1/3$	0.2	0.01	.9	.95	10	1.02	0	0.5

,

where m and sd are the mean and standard deviation of the normal distribution from which the logarithms of the transitory shocks are drawn.

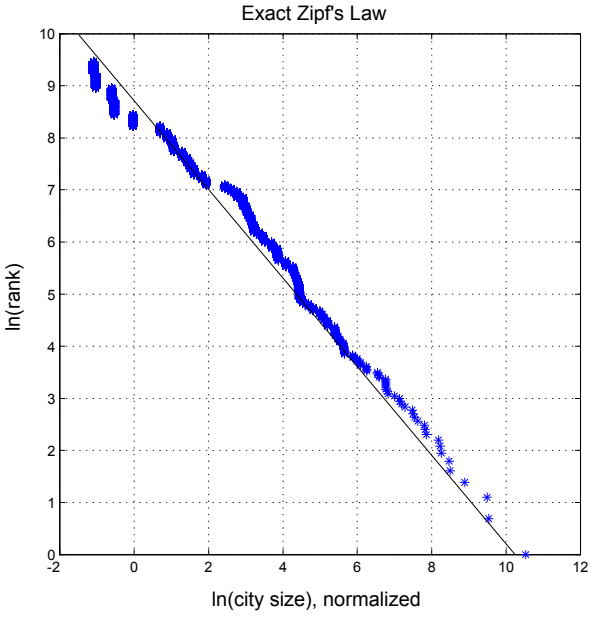

Figure Three

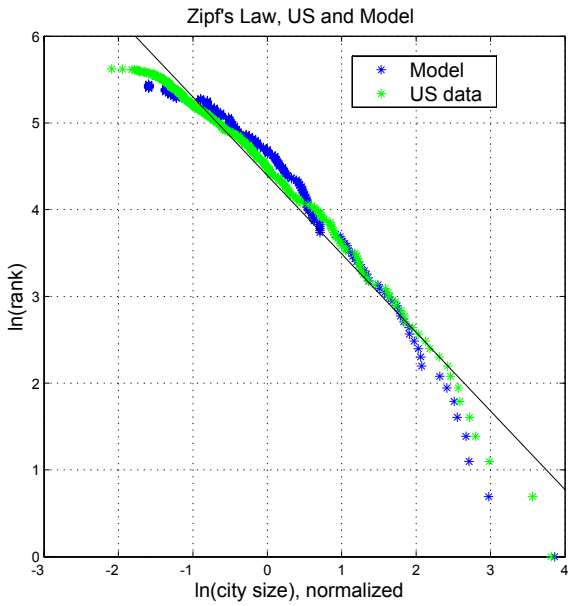

Figure Four

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. Proposition 6 tells us that one key parameter is the standard deviation of industry shocks. Otherwise, the model seems to be robust (not invariant) to all other parameter values[9]. This justifies our focus on the standard deviations: 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 standard deviation of temporary shocks in Figure Five.

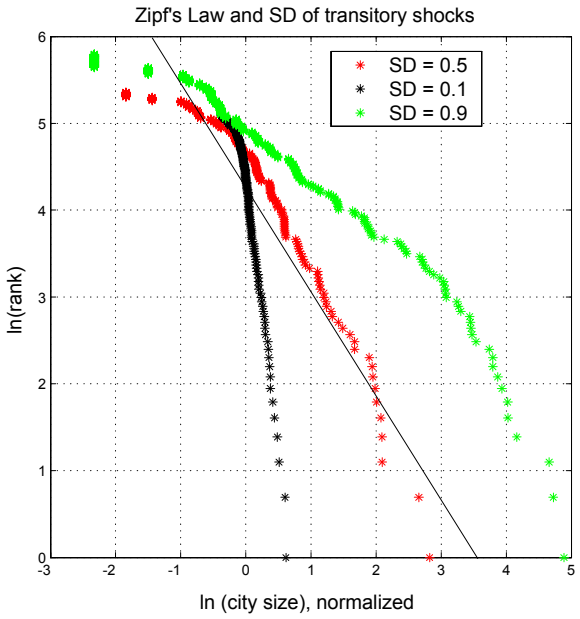

Figure Five

The figure starts with a standard deviation of 0.5, which implies a Zipf's coefficient close to 1. If we increase *sd* to 0.9, the absolute value of the slope of the curve decreases. That is, the dispersion of the city size distribution increases. The opposite happens if we reduce *sd* substantially, say to 0.1. Soo (2003) finds that the coefficients

[9]Except the discount factor, δ, that is related to the standard deviation, *sd*, via the period length, which is calibrated to one year.

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 (that is, smaller industry shocks), we view the response of the model to changes in *sd* as identifying the source of the differences in Zipf's coefficients observed in the data.

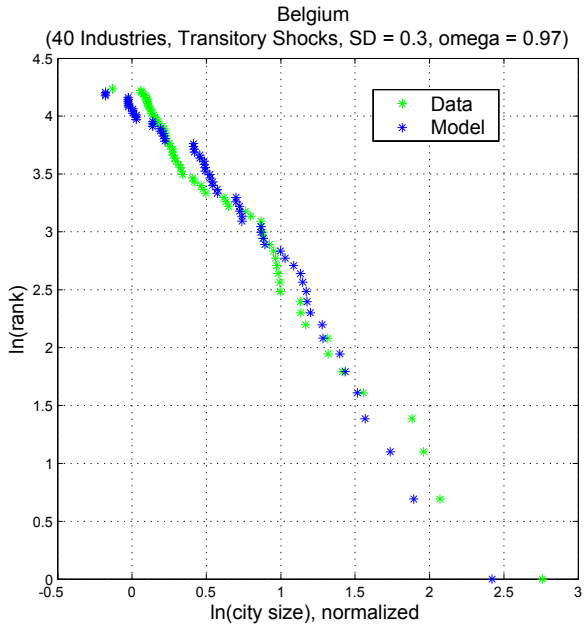

Figure Six

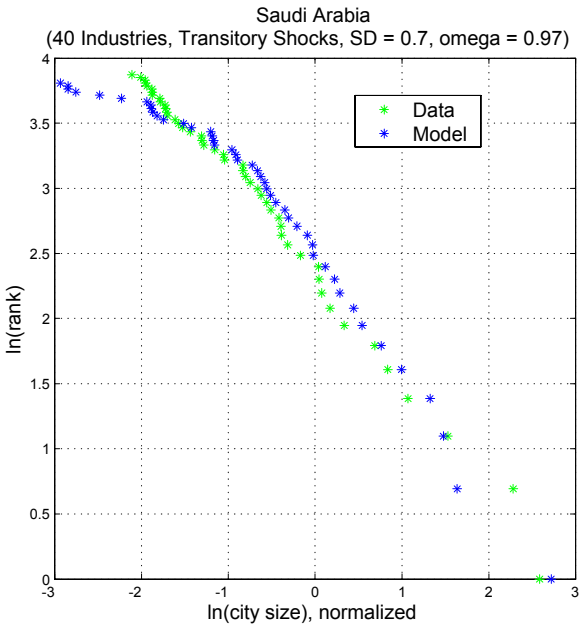

Figure Seven

International evidence on urban structures implies bounds on observed Zipf's Law coefficients. These bounds, in turn, imply bounds on admissible industry productivity shocks. In the rest of this section we compare available evidence on this relationship. Toward this, we first select two countries that exhibit city size distributions that are either extremely concentrated or extremely dispersed. The rank size relationship in Belgium is very steep with a Zipf's coefficient of 1.59. The standard deviation of transitory shocks that yields a city size distribution consistent with the Belgian data is 0.31. The data and the simulation are presented in Figure Six.

We perform the same exercise for a country 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. Figure Seven shows the simulation and Saudi Arabia's data[10]. The standard deviation used in the numerical simulation is $sd = 0.73$.

These two extreme cases give us a range of standard deviations that would imply city size distributions consistent with what we observe in the data. The next question is whether this range is in line with measures of productivity shocks by industry. The model gives us a method to map observed Zipf's coefficients into standard deviations of productivity shocks, given industry heterogeneity. As we have done so far, we want to gauge the performance of the model without relying on particular forms of industry heterogeneity that would help our theory, but obscure the main mechanisms in play. Hence, we assume identical industries and solve for the standard deviation that produces Zipf's coefficients consistent with the ones in the data. This produces bounds on standard deviations that we compare with the evidence on productivity shocks in the data. Horvath (2000) measures the standard deviation and persistence of industry shocks in the United States for 36 industries[11].

It is important to stress that this comparison puts a heavy burden on our theory. To clarify, consider a situation where all of the standard deviations 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 standard deviations. That is, there is no

[10] 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 coeficient of 0.728, while Kuwait, with 28 cities, has a Zipf's coefficent of 1.720. Using these countries would only improve the performance of the model in the comparisons that follow.

[11] As the United States is the world's largest economy, we will take this data to represent the universe of possible productivity shock processes. In order to compare Horvath's estimates with our range of standard deviations, we first need to map the standard deviations of persistent shocks into standard deviations of transitory shocks.

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 standard deviations were inside the implied range, it would be evidence against our theory.

Table One presents these estimates and the percentage of industries in Horvath's study that lie inside the interval of standard deviations implied by the international city size data. Perhaps surprisingly, given the nature of the exercise, half of the industries have standard deviations that lie within these bounds.

Table One

Distribution of Zipf's coefficients	Min	Max	
$[Min, Max]$	0.7287	1.7190	
$[10\%, 90\%]$	0.8590	1.3820	
$[20\%, 80\%]$	0.9207	1.2704	
Implied bounds on the *sd* of industry shocks	Min	Max	% of Horvath's industries inside the *sd* range
$[Min, Max]$	0.3080	0.7300	50
$[10\%, 90\%]$	0.3850	0.6200	25
$[20\%, 80\%]$	0.4200	0.5750	19

Similarly, we can use the evidence on the standard deviations of industry shocks to construct bounds on Zipf's coefficients. In contrast with the previous exercise, the fact that countries have diversified industrial structures implies that this exercise will produce only loose bounds on the range of Zipf's coefficients that we should observe in the data. Not surprisingly, as shown in Table Two, the Zipf's coefficient of every country in our data set is inside the interval implied by the industry data. This

remains true even if we focus only on those industries at the center of the distribution of standard deviations.

Countries produce in a variety of industries and so the ability of the model to explain the relationship between the urban structure and the variance of productivity shocks lies between the bounds implied by these two exercises. This allows us to conclude that the theory is performing well for most industries and countries. It is also clear that in order to derive tighter bounds we would need to take a stand on industry heterogeneity. This would require disaggregated data on industrial structure for a wide set of countries. To the best of our knowledge, these data are not available beyond a small sample of developed economies, and so we leave this empirical exercise for future research.

Table Two

Distribution of sd of industry shocks in the US	Min	Max	
$[Min, Max]$	0.0844	3.6816	
$[10\%, 90\%]$	0.1423	1.1727	
$[20\%, 80\%]$	0.2421	0.6936	
Implied bounds on Zipf's coefficients	Min	Max	% of countries inside the Zipf's coefficient range
$[Min, Max]$	0.1444	6.2389	100
$[10\%, 90\%]$	0.4535	3.6862	100
$[20\%, 80\%]$	0.7675	2.1933	97

5. CONCLUSIONS

We have proposed an urban growth theory that emphasizes the role of the accumulation of specific factor 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's evolution through the birth, growth and death of cities, that 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.

One of the advantages of the simple specification we adopted above is that it allowed us to identify analytically the standard deviation 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.

Our theory also 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 city 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, would not affect the main results of our theory, but would translate into lower observed levels of total factor productivity. The varying ability of local governments in different 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.

REFERENCES

[1] Au, C. and V. Henderson (2002). "How Migration Restrictions Limit Agglomeration and Productivity in China." Unpublished paper, Brown University.

[2] Auerbach, F. (1913). "Das Gesetz der Bevölkerungskonzentration." *Petermanns Geographische Mitteilungen*, 59:74-76.

[3] Black, D. and V. Henderson (1999). "A Theory of Urban Growth." *Journal of Political Economy,* 107(2): 252-284.

[4] Blank, A. and S. Solomon (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.

[5] Champernowne, D. G. (1953). "A Model of Income Distribution." *Economic Journal*, 63 (250): 318-351.

[6] Cordoba, J. (2003). "On the Distribution of City Sizes." Unpublished paper, Rice University.

[7] Dobkins, L. H. and Y. M. Ioannides (2000). "Spatial Interactions Among U.S. Cities: 1900-1990." Unpublished paper, Tufts University.

[8] Duranton, G. (2002). "City Size Distribution as a Consequence of the Growth Process." Unpublished paper, London School of Economics.

[9] Eaton, J. and Z. Eckstein (1997). "Cities and Growth: Theory and Evidence from France and Japan." *Regional Science and Urban Economics*, 27: 443-474.

[10] Eeckhout, J. (2004). "Gibrat's Law for (all) Cities." *American Economic Review*, forthcoming.

[11] Gabaix, X. and Y. Ioannides (2003). "The Evolution of City Size Distributions." In J. V. Henderson and J. F. Thisse (eds.) *Handbook of Economic Geography*, North-Holland, Amsterdam.

[12] Gabaix, X. (1999a). "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics*, 739-767.

[13] Gabaix, X. (1999b). "Zipf's Law and the Growth of Cities." *American Economic Review Papers and Proceedings,* 89(2): 129-132.

[14] Henderson, V. (1974). "The Sizes and Types of Cities." *American Economic Review*, 64: 640-656.

[15] Horvath, N. (2000). "Sectoral Shocks and Aggregate Fluctuations." *Journal of Monetary Economics*, February, 69-106.

[16] Ioannides, Y. M. and H. G. Overman (2001). "Zipf's Law for Cities: An Empirical Examination." Unpublished paper, Tufts University.

[17] Jones, C. I. (1999). "Growth: With and Without Scale Effects." *American Economic Review Papers and Proceedings*, 89 (2): 139-144.

[18] Kalecki, M. (1945). "On the Gibrat Distribution." *Econometrica*, 13 (2): 161-170.

[19] Levy, M. and S. Solomon. (1996). "Power Laws are Logarithmic Boltzmann Laws," *International Journal of Modern Physics C*, 7(4): 595-600.

[20] Lucas, R. E., Jr. (1988). "On the Mechanics of Economic Development." *Journal of Monetary Economics,* 22(1): 3-42.

[21] Malcai, O., O. Biham and S. Solomon. (1999). "Power-Law Distributions and Levy-Stable Intermittent Fluctuations in Stochastic Systems of many Autocatalytic Elements," *Physical Review E*, 60(2): 1299-1303.

[22] Romer, P. (1990). "Endogenous Technological Change." *Journal of Political Economy*, 98: S71-S102.

[23] Rosen K. and M. Resnick (1980). "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy." *Journal of Urban Economics*, 8(2): 156-186.

[24] Rossi-Hansberg, E. and M. L. J. Wright (2003). "Urban Structure and Growth." *SIEPR Discussion Paper* 02-40.

[25] Rossi-Hansberg, E. and M. L. J. Wright (2004). "Firm Size Dynamics in the Aggregate Economy." Unpublished paper, Stanford University.

[26] Soo, K. T. (2003). "Zipf's Law for Cities: A Cross Country Investigation." Unpublished paper, London School of Economics.

APPENDIX

Competitive Equilibrium: Developers Problem

City 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. 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 P_{tj}, W_{tj}, R_{tj}, and S_{tj} be the price of output j and the rental rates of labor, physical capital, and human capital respectively written in terms of some numeraire commodity. Then the problem of a city 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

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

subject to

$$\begin{aligned} \left(1 - \tau^k_{tj}\right) R_{tj}/P_{tj} &= \beta_j Y_{tj}/K_{tj}, \\ \left(1 - \tau^h_{tj}\right) S_{tj}/P_{tj} &= \alpha_j Y_{tj}/H_{tj}, \\ \frac{W_{tj}}{P_{tj}} - T_{tj} &= \left(1 - \alpha_j - \beta_j\right) \frac{Y_{tj}}{N_{tj}} \\ I_{tj} &= \frac{W_{tj}}{P_{tj}} - \frac{3b}{2} \left(\frac{N_{tj}}{\mu_{tj}} \right)^{\frac{1}{2}}. \end{aligned}$$

Competition from other developers ensures that in equilibrium profits are zero, so

$$T_{tj} = \frac{b}{2} \left(\frac{N_{tj}}{\mu_{tj}} \right)^{\frac{1}{2}} - \tau^k_{tj} \frac{R_{tj}}{P_{tj}} \frac{K_{tj}}{N_{tj}} - \tau^h_{tj} \frac{S_{tj}}{P_{tj}} \frac{H_{tj}}{N_{tj}}.$$

Proofs of Propositions

Proposition 4 *(Exact Gibrat's Law and Zipf's Law) The growth process of city sizes satisfies Gibrat's Law, and the invariant distribution for city sizes satisfies Zipf's Law, if and only if one of the following two conditions is satisfied:*

1. *(No physical capital) There is no physical capital* $\left(\hat{\beta}_j = 0 \text{ or } \omega_j = 1\right)$*, and productivity shocks are permanent.*

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

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

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

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

In the second case, we have

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

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

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

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 that this implies an invariant distribution that satisfies Zipf's Law, start with the process

$$\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) = \xi_j.$$

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 (1999a) among others). We study the invariant distribution that results as the lower bound tends to zero. 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

$$\hat{q}\left(\hat{s}\right) = e^{\hat{s}} q\left(e^{\hat{s}}\right).$$

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

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

where $q^{\xi}\left(\xi\right)$ denotes the probability of the growth rate taking the value ξ, and $\hat{q}\left(\hat{s}, t\right)$ denotes the distribution of $\hat{s}$ at time t. Standard results (see for example Levy and Solomon (1996) and Malcai, Biham and Solomon (1999)) then imply that the only asymptotic stationary solution of the master equation is of the form

$$\hat{q}\left(\hat{s}\right) = M e^{-\eta \hat{s}},$$

for some M and η to be determined. This implies that

$$q\left(s\right) = M \frac{1}{s^{1+\eta}}.$$

Using the normalization

$$\int_{f}^{G} s q\left(s\right) ds = 1,$$

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

$$\int_{f}^{G} q\left(s\right) ds = 1,$$

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 zero, η converges to zero, and we get

$$q\left(s\right) = M\frac{1}{s}.$$

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

$$q^{\text{City}}\left(\varsigma\right) = \hat{M}\frac{1}{\varsigma^2},$$

for some $\hat{M}$ finite. The cumulative distribution function is then given by

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

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 $Q_i^{\text{City}}\left(\varsigma > \bar{\varsigma}\right) = \hat{M}_i / \bar{\varsigma}$, 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

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

which is a statement of Zipf's Law for the economy. ■

Proposition 5 *(Concavity) If conditions 1 and 2 in Proposition 4 are not satisfied, the growth rate of cities exhibits reversion to the mean. If productivity levels are bounded for all industries (so that there exist uniform bounds such that $A_{tj} \in \left[\underline{A}_j, \overline{A}_j\right]$ for all t, j), then there exists a unique invariant distribution of city sizes with thinner tails than a Pareto distribution with coefficient one.*

Proof. We have that city growth rates are given by

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

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

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

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

$$\begin{aligned}& \hat{\beta}_j\left(1 - \left(\omega_j + \left(1-\omega_j\right)\hat{\beta}_j\right)\right)\left(1-\omega_j\right)\sum_{T=1}^{t}\left(\omega_j + \left(1-\omega_j\right)\hat{\beta}_j\right)^{t-T} \\ = \; & \hat{\beta}_j\left(1 - \left(\omega_j + \left(1-\omega_j\right)\hat{\beta}_j\right)^{t-1}\right)\left(1-\omega_j\right).\end{aligned}$$

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

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 zero. 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).

To prove existence of a unique invariant distribution with thinner tails than a Pareto distribution with coefficient one, we rely on the results in Propositions 4 and 5 of Rossi-Hansberg and Wright (2004). ■

Proposition 6 *If conditions 1 and 2 in Proposition 4 are not satisfied, the standard deviation of city sizes increases with the standard deviation of industry shocks.*

Proof. If conditions 1 and 2 in Proposition 4 are not satisfied, the variance of the log of city sizes is given by

$$V_0\left[\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\right] = 4V_0\left[\ln\left(A_{tj}\right)\right] + 4\hat{\beta}_j^2 V_0\left[\ln\left(K_{tj}\right)\right]$$

and

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

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

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

or as $t \to \infty$,

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

so that the variance of the long run city size distribution is given by

$$V_0\left[\ln\left(\frac{N_{tj}}{\mu_{tj}}\right)\right] = 4v\left[1 + \frac{\hat{\beta}_j^2}{\left(1+\hat{\beta}_j\right)^2}\right],$$

which is increasing in v, thereby proving the result.

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