Discussion Paper 141

Institute for Empirical Macroeconomics
Federal Reserve Bank of Minneapolis
90 Hennepin Avenue
Minneapolis, Minnesota 55480-0291

November 2003

Urban Structure and Growth

Esteban Rossi-Hansberg and Mark L. J. Wright*

Stanford University

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 how the urban structure is the margin that eliminates local increasing returns to yield constant returns to scale in the aggregate, thereby implying a city size distribution that is well described by a power distribution with coefficient one: Zipf's Law. Under strong assumptions our theory produces Zipf's Law exactly. More generally, it produces the systematic deviations from Zipf's Law observed in the data, namely, the underrepresentation of small cities and the absence of very large ones. In these cases, the model identifies the standard deviation of industry productivity shocks as the key element determining dispersion in the city size distribution. We present evidence that the dispersion of city sizes is consistent with the dispersion of productivity shocks in the data.

*Rossi-Hansberg: erossi@stanford.edu. 579 Serra Mall, Stanford, CA, 94305. Ph. (650) 724-1427; Wright: mlwright@stanford.edu. 579 Serra Mall, Stanford, CA, 94305. Ph. (650) 725-9967. We thank 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 authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. INTRODUCTION

Most economic activity occurs in cities. 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. As a result, understanding aggregate economic activity requires perforce theories of *urban* economic activity. One of the key elements of such a theory is that cities emerge out of the trade-off between agglomeration effects and congestion costs. Agglomeration in cities implies urban scale effects, which create a tension between increasing returns at the city level and constant returns at the aggregate level, which are crucial for balanced growth[1]. In general, models with aggregate scale effects do not exhibit the linearity necessary for balanced growth. In particular, growth theories should explain why permanent growth is possible and why growth rates are stable, bounded and do not depend solely on population growth. In this paper we explain how urban structure eliminates local increasing returns to yield constant returns to scale in the aggregate, which is crucial for balanced growth, and how balanced growth implies a city size distribution that is well described by a power distribution with coefficient one: Zipf's Law.

The essence of our approach is to identify the urban structure as the margin that leads to constant returns to scale in the economy. In equilibrium, city sizes are determined out of the trade-off between the increasing returns implied by agglomeration and the decreasing returns caused by congestion forces. Given factor proportions and productivity levels, each city will then produce at an optimal scale and industries will behave as if using a constant returns to scale technology by varying the number of cities. In this way, introducing urban structure results in linear aggregate production functions in a world with increasing return technologies. This mechanism then has very strong implications for the size distribution of cities once we include factor accumulation and productivity shocks. In particular it delivers the striking regularity

[1]By this we are referring to the production set of the aggregate economy. In models such as Lucas [16], increasing returns at the industry level are transformed into constant returns at the aggregate by assuming a linear human capital accumulation technology.

known as Zipf's Law of cities: The rank of city sizes is proportional to the inverse of their size.

The ability of this mechanism to replicate the city size distribution depends upon the way we introduce factor accumulation and productivity shocks. In particular, its ability to produce Zipf's Law is derived from its ability to produce Gibrat's Law of cities – the mean and variance of the growth rate of a city are independent of its size – which, as shown by Gabaix [10] and extended by Cordoba [5], is both necessary and sufficient to produce an invariant distribution for city sizes that satisfies Zipf's Law. To fix ideas, consider first a simple economy in which the only factors of production are labor and human capital both growing at constant rates. With constant total factor productivity, city sizes evolve at a constant rate. Mean zero shocks to the level of total factor productivity will not affect mean city growth rates, but will affect the distribution of city sizes directly, which implies that, if shocks are permanent, the growth process of cities is scale independent. More generally, productivity shocks will affect the distribution of city sizes both directly and through their effect on factor accumulation. The bulk of the paper is devoted to a study of the interaction of these effects and their ability to produce or approximate Gibrat's Law.

In addition to establishing the remarkable robustness of Zipf's Law as a description of the size distribution of cities, the empirical literature has also stressed a number of robust deviations. One of the most notable is that, relative to Zipf's Law, small cities are underrepresented 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, in the cases where the direct effect of shocks does not exactly balance the indirect effect of shocks through factor accumulation, is able to produce these systematic deviations from Zipf's Law. In particular, the model identifies the standard deviation of industry productivity shocks as the key element determining dispersion in the size distribution of cities.

There are potentially many ways of introducing agglomeration forces and factor accumulation into an urban growth model. This paper illustrates these interactions

using a very particular specification for these forces, but we argue that the insights are much more general. In our formulation, cities are the result of the trade-off between production externalities and commuting costs, while growth can be either endogenous as a result of linear human capital accumulation or the exogenous result of technological change. Adding industry productivity shocks to this specification will then result in a distribution of city sizes where all cities in one industry will have the same size.

This paper draws from four related literatures. The first is the extensive literature on growth and, in particular, the large number of papers on endogenous growth that were spawned by the contributions of Lucas [16] and Romer [17]. In this literature, as emphasized by Jones [15], 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. The two main papers in this group are Black and Henderson [4] and Eaton and Eckstein [8], which 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. Our paper is most closely related to the contribution of Black and Henderson [4], whom we follow in using the formulation of Henderson [12] as a vehicle for introducing cities. 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 parameters.

Following the original paper of Auerbach [3], a substantial literature has arisen that investigates the empirical foundations of Zipf's Law. Rosen and Resnick [18] documented this regularity in the 1980s for a wide range of countries, while Soo [19] has updated this study using modern data and more sophisticated econometric techniques. One of the key findings of this literature is the robustness of this phenom-

enon both over time and across countries. As illustrated in Figure One for the United States, 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.

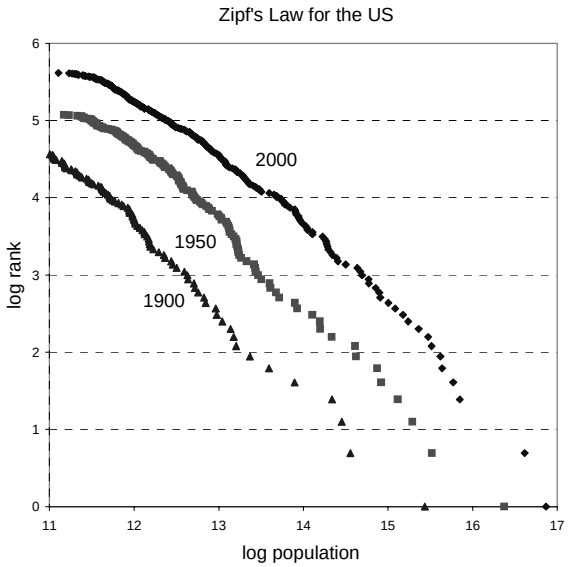

Figure One

This visual impression for U.S. data has been verified in the careful statistical work of Dobkins and Ioannides [6] and Ioannides and Overman [14]. 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. The description is not perfect. Some countries have a size distribution that is more or less even than that predicted by Zipf's Law, which is reflected in flatter or steeper plots of log-rank against log-size. There is also a broad tendency for the relationship to be slightly concave, at least once one controls for a country's capital city. 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. These papers can be distinguished by the emphasis given to the process leading to the formation of cities as opposed to the process determining the growth of cities. Gabaix [10] provided a proof that Gibrat's Law is sufficient to generate an invariant

distribution for city sizes that satisfies Zipf's Law. This result was later extended by Cordoba [5], who proved that Gibrat's Law is both necessary and sufficient. Both papers provide economic models that generate city growth that satisfies Gibrat's Law: in Gabaix [10] and [11], cities grow as labor migrates in response to city amenity shocks, while in Cordoba [5] labor is allocated across cities in response to a power distribution of taste shocks. Neither paper generates the existence of cities endogenously. In a recent study, Duranton [7] presents a quality ladder model of growth which, under very particular assumptions on the location and mobility of new firms, is capable of producing a size distribution of cities that is close to a power distribution. By contrast, in our paper, the city size distribution arises endogenously out of the growth process in a way that both eliminates scale effects in growth and approximates Gibrat's Law of city growth.

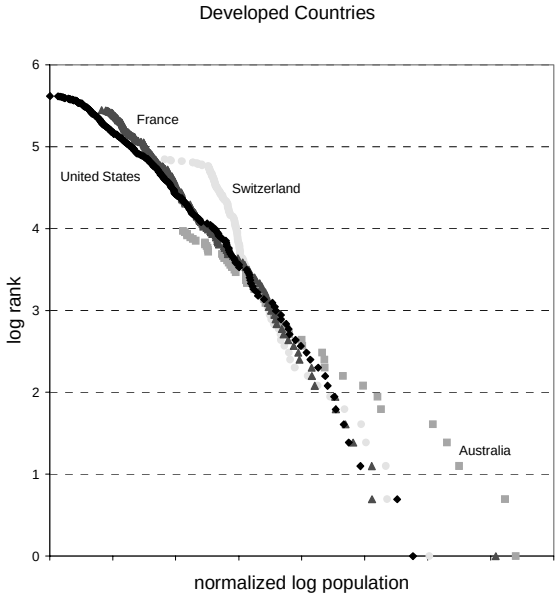

Figure Two A

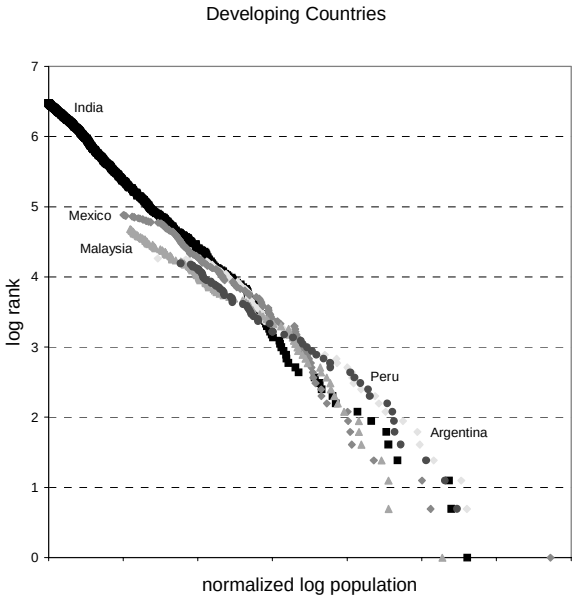

Figure Two B

The rest of the paper is organized as follows. The next section presents the model. Section 3 derives the main results of the paper, namely, the results 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 on several countries. Section 5 concludes. An appendix contains the decentralization of the allocation presented in the text, as well as proofs of all propositions in the paper.

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 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 [12] and has been used in the urban growth model of Black and Henderson [4]. 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, which is the numeraire commodity.

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 the commuting cost, 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

$$\bar{z} = \left(\frac{n}{\pi}\right)^{\frac{1}{2}},$$

and so

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

where $b = 2\pi^{-\frac{1}{2}}\tau/3$. 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

$$\frac{3b}{2}n^{\frac{1}{2}}$$

in terms of rents and commuting costs.

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 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. This is the force causing 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 need to impose an additional restriction on the technology. The original set of J industries has to allow a partition, with at least two elements of J in each component of the partition, where all elements in a component have the same technology parameters. That is, each industry has to have at least two varieties (counted in J) that are produced with exactly the same technology, but may be produced with different

amounts of human and physical capital, and receive different shocks. 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. Limiting the amount of ex-ante heterogeneity within these groups of industries is necessary for the growth process of the corresponding group of cities to satisfy Gibrat's Law in certain cases we describe in detail below. In these cases we can then aggregate all industries to show that cities in the economy satisfy Zipf's law.

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

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

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

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

Here 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 according 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 problem of choosing the optimal sizes of cities is a static problem: 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 industrywide variables and the number of cities in an industry. Namely,

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

The first order condition with respect to μ_{tj} (which we show in the appendix is necessary and sufficient) is given by

$$\frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{\frac{3}{2}} = \left(\gamma_j + \varepsilon_j\right)\frac{Y_{tj}}{\mu_{tj}}.$$

To interpret this equation, rewrite the first order condition as

$$\frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{-\frac{1}{2}} = \left(\gamma_j + \varepsilon_j\right)\frac{Y_{tj}/N_{tj}}{N_{tj}/\mu_{tj}}. \tag{8}$$

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).

Now 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 will be, in essence, a mechanism for producing permanent increases in the average product of labor in a city, while at the same time remaining consistent with the aggregate growth facts.

It is important that, in response to a productivity shock, average commuting costs do not rise by exactly the same amount as the average product of labor; if they do, the planner would find it optimal not to change the city's size. 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 will remain. In the model below we ensure that this is the case by denominating commuting costs within a city in terms of the output of that city. Other assumptions would work as well. However, one combination of assumptions that does *not* work is if commuting costs are denominated in units of time while at the same time workers supply labor inelastically and the production function is Cobb-Douglas. The reason is that, with Cobb-Douglas production, marginal and average products are proportional and hence commuting costs measured as forgone wages will rise at exactly the same rate as the average product of labor.

Rearranging the first order condition, we also find that the optimal number of cities is given, as a function of output and employment in industry j, by

$$\mu_{tj} = \left[\frac{2\left(\gamma_j + \varepsilon_j\right)}{b}\frac{Y_{tj}}{N_{tj}}\right]^{-2} N_{tj}, \tag{9}$$

and so total commuting costs are

$$TCC_{tj} = 2\left(\gamma_j + \varepsilon_j\right) Y_{tj}. \tag{10}$$

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. To interpret this restriction, write industry output minus total commuting cost as

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

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 yields

$$C_{tj} + X_{tj} \leq F_j \hat{A}_{tj} H_{tj}^{\hat{\alpha}_j} K_{tj}^{\hat{\beta}_j} N_{tj}^{1-\hat{\alpha}_j-\hat{\beta}_j} u_{tj}^{\hat{\phi}_j} \equiv \hat{Y}_{tj} \tag{11}$$

where

$$\begin{aligned} F_j &= \left(1 - 2\left(\gamma_j + \varepsilon_j\right)\right) \left[\frac{2\left(\gamma_j + \varepsilon_j\right)}{b}\right]^{\frac{2\left(\gamma_j+\varepsilon_j\right)}{1-2\left(\gamma_j+\varepsilon_j\right)}}, \\ \hat{A}_{tj} &= A_{tj}^{\frac{1}{1-2\left(\gamma_j+\varepsilon_j\right)}}, \end{aligned}$$

and

$$\begin{aligned} \hat{\alpha}_j &= \frac{\alpha_j + \gamma_j}{1 - 2\left(\gamma_j + \varepsilon_j\right)}, \\ \hat{\beta}_j &= \frac{\beta_j}{1 - 2\left(\gamma_j + \varepsilon_j\right)}, \end{aligned}$$

and

$$\hat{\phi}_j = \frac{1 - \alpha_j - \beta_j}{1 - 2\left(\gamma_j + \varepsilon_j\right)}.$$

Since $u_{tj} \leq 1$, 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 (10) output in the industry is also a constant returns to scale function of inputs in the industry.

The constraint in (11) 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. This has implications for the way in which we view the growth process. First, it allows us to reconcile the coexistence of cities, which in turn 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, differences in the way production is organized in cities will determine the level of aggregate productivity (the magnitude of F_j in equation (11)). This suggests the possibility that differences in the pattern of urbanization are the source of differences in total factor productivity across countries[2]. In our theory, these sort of differences in productivity can be distinguished from technology levels through the fact that there is likely to be more time variation in the latter. 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 (11) 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.

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

[2] Au and Henderson [1] examines this possibility for the particular case of China.

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 diverge to infinity. However, with increasing returns at the city level, the mechanism we have introduced in this paper would yield constant returns in the aggregate and balanced growth.

After substituting for the optimal number of cities, the result is 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 (11), (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, together with expressions for the parameters of the value function, are contained in the appendix. Three basic results are immediate. The share of population working in each industry is constant and satisfies

$$N_{tj} = \frac{\left(1 - \hat{\alpha}_j - \hat{\beta}_j\right)\left(\delta D_j^K (1 - \omega_j) + (1 - \delta)\theta_j\right)}{\sum_{j=1}^{J}\left[\left(1 - \hat{\alpha}_j - \hat{\beta}_j\right)\left(\delta D_j^K (1 - \omega_j) + (1 - \delta)\theta_j\right)\right]} N_t.$$

Investment is a constant share of output net of commuting costs

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

and the fraction of time used for production is constant,

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

The original problem is not a convex dynamic optimization problem. However, since the city size problem is static, we can solve it separately and, as discussed in the text, transform the problem into a convex dynamic optimization problem. This argument, when formalized, leads to the following proposition.

Proposition 1 *There exists a unique Pareto efficient allocation for this economy.*

Decentralization

We have shown how, by solving the urban problem separately, we can convert a problem with local increasing returns to scale into a problem with aggregate constant returns to scale. This was possible because the planner internalizes the externality and therefore the Pareto optimum implies efficient city sizes. In order to explain the observed city size distributions, it is necessary to consider also competitive equilibrium allocations. We now proceed to introduce a competitive equilibrium framework for which the unique equilibrium allocation attains the optimum. As is standard in the previous literature, we use city developers that internalize the urban production externality.

We follow Henderson [12] 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 workers to the city, developers may pay each resident a transfer. They may also attract firms by subsidizing physical and human capital, although they never choose to use the former as there is no externality in physical capital. Agents derive utility out of consumption of goods that are costlessly tradable, and so they will live in the city if their income after rents, commuting costs, and transfers is larger than what they could obtain elsewhere. Firms will 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 given by a condition that is identical to the expression for the optimal size of cities arising from the social planner's problem (8). Given the size of the industry, this will, in general, mean that we must allow for the possibility of a non-integer number of cities, all of which will be identical in size within an industry. Since developers are fully internalizing the external effect, the equilibrium allocation will be efficient.

The details of the developer's problem are presented in the appendix, together with the complete statement of the competitive equilibrium, which is standard. We present the analogs of both Welfare Theorems in the next two propositions.

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

Proposition 3 *Every competitive equilibrium in this economy is Pareto efficient.*

3. ANALYTICAL RESULTS

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. As shown in the appendix, under our functional form assumptions, we are able to solve for the entire equilibrium growth path and size distribution of cities in closed form. A couple of general points are worth making. First, although the main reason for our functional form assumptions is tractability, they have some additional expository merit: the assumptions imply that the labor allocation across industries is fixed independently of productivity shocks. This means that our ability to match the size distribution of cities is being driven solely by forces operating at the city level. It also means that if we were to relax this assumption and calibrate the model to match

the size distribution of industries (which, although not obeying a rank size rule, is at least closer to it than produced by our model) we should get a city size distribution even closer to Zipf's Law.

Second, 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, 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

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

Hence, on the balanced growth path $\ln \hat{Y}_{tj} - \ln K_{tj}$ is constant. That is,

$$g_{\hat{Y}j} = g_{K_j}.$$

For human capital,

$$g_{H_j} = B_j^0 + (1-u_j^*)B_j^1.$$

For income, when $\hat{\beta}_j < 1$,

$$\begin{aligned} g_{\hat{Y}_{t+1j}} &= \ln \hat{Y}_{t+1j} - \ln \hat{Y}_{tj} \\ &= \frac{1}{1-\hat{\beta}_j}\left[\left[\ln A_{t+1j} - \ln A_{tj}\right] + \hat{\alpha}_j g_{H_j} + \left(1-\hat{\alpha}_j-\hat{\beta}_j\right) g_{N_{tj}}\right], \end{aligned}$$

so in the balanced growth path[3] (with no uncertainty),

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

Third, the distribution of city sizes is determined by a static process each period. All cities of a given type are the same size, which is given by

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

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

From this equation it is easy to see that anything that increases the level of the average product of labor will increase the average size of the city. Indeed, it is the effect of shocks on the average product of labor, both contemporaneously and in the future, that determines the growth process of a city.

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:

$$\begin{aligned}&\ln\left(\frac{N_{t+1j}}{\mu_{t+1j}}\right) - \ln\left(\frac{N_{tj}}{\mu_{tj}}\right) \\ =\ & \frac{2\hat{\alpha}_j}{1-\hat{\beta}_j}\left[g_{Hj} - g_N\right] + 2\left[\ln\left(A_{t+1j}\right) - \ln\left(A_{tj}\right)\right] \\ & + 2\left(1-\omega_j\right)\hat{\beta}_j\left[\ln(A_{tj}) - \sum_{s=1}^{\infty} \frac{\left(\omega_j + (1-\omega_j)\,\hat{\beta}_j\right)^{-s}}{\left(1 - \left(\omega_j + (1-\omega_j)\,\hat{\beta}_j\right)\right)^{-1}} \ln(A_{t-sj})\right]. \quad (12)\end{aligned}$$

Both the aggregate labor growth rate and human capital growth rate are constants, with the only stochastic part of the growth expression coming from productivity shocks today and the effects of past shocks on capital accumulation. Note that, as the economy grows, and more human capital is accumulated, the size of cities may increase or decrease indefinitely. This may result in the number of cities in the economy going to zero or infinity. Human capital accumulation implies that cities become bigger, while population growth implies that cities become smaller (since per capita human and physical capital decrease). The condition that guarantees that the

number of cities is constant over time (without uncertainty) is given by

$$g_N = \frac{2\hat{\alpha}_j}{1 - \hat{\beta}_j + 2\hat{\alpha}_j} g_{H_j},$$

which implies that population grows slower than human capital.

Equation (12) is the key equation of our model. From this equation we can deduce conditions under which we can guarantee Gibrat's Law for each group of cities defined by our partition of industries, that is, conditions under which the expected long run growth rate and variance do not depend on any past information and hence are independent of city size distributions in previous periods. That Gibrat's Law implies Zipf's Law follows from the results in Gabaix [10] and [11], later extended by Cordoba [5].

The first set of conditions amount to eliminating physical capital from the model. Without physical capital, productivity shocks are not propagated via capital stocks. This implies that if the *growth rate* of productivity shocks is time independent (shocks are permanent), the growth rate of cities will be time independent as well. Physical capital is eliminated if it either cannot be accumulated ($\omega_j = 1$ for all j) or is not an input in production ($\hat{\beta}_j = 0$ for all j). Under either of these conditions, we obtain that the mean long run growth rate is given by

$$E_t \left[\ln \left(\frac{N_{t+1j}}{\mu_{t+1j}} \right) - \ln \left(\frac{N_{tj}}{\mu_{tj}} \right) \right] = 2\hat{\alpha}_j \left[g_{Hj} - g_N \right]$$

and the long run variance by

$$V_t \left[\ln \left(\frac{N_{t+1j}}{\mu_{t+1j}} \right) - \ln \left(\frac{N_{tj}}{\mu_{tj}} \right) \right] = 4V_t(\ln \left(A_{t+1j} \right) - \ln \left(A_{tj} \right)),$$

both of which are obviously scale independent.

The second set of conditions amount 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 so the only remaining source of uncertainty is the contemporaneous productivity shock. If on top of this we assume that industry shocks are transitory, we obtain Gibrat's

Law. The next proposition formalizes these arguments; 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 therefore 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* $\left(\beta_j = \hat{\beta}_j = 0 \text{ or } \omega_j = 1\right)$*, 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.*

The intuition for the above result is straightforward. In order to generate 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 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. 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 accumulation of physical capital in a linear production setting[4].

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.

[4]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.

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 [9]) 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 [19]).

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 experienced a history of productivity shocks above average, it can be expected to grow slower than average in the future, while the opposite is true of small cities. To understand this we can use the expression for the long run growth rates of cities (12) to show how capital investments affect the urban size distribution. Suppose that an industry has experienced very high shocks in the past. This implies that output in that industry will be relatively high, and, since investment is a fraction of output, investment in industry specific physical capital has been high. This is expressed in equation (12) by a large value of the term

$$\sum_{s=1}^{\infty} \frac{\left(\omega_j + (1-\omega_j)\,\hat{\beta}_j\right)^{-s}}{\left(1-\left(\omega_j + (1-\omega_j)\,\hat{\beta}_j\right)\right)^{-1}} \ln(A_{t-sj}).$$

Since this term reduces the growth rate of cities, it implies that large cities will grow at a relatively lower rate than small cities (cities that have experienced low shocks and so have invested little in capital). 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 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. Since city growth is proportional to industry output growth, this implies that small cities grow faster than large cities: urban growth rates exhibit reversion to the mean. The result is that the log rank-size relationship will in general (apart from particular realizations of the shocks) be concave[5]. That is, relative to a linear relationship, there are not enough small cities and large cities are not large enough.

Proposition 5 *(Concavity) If conditions 1 and 2 in Proposition 4 are not satisfied, the growth rate for cities exhibits reversion to the mean.*

Unless the conditions of Proposition 4 are satisfied, variation in the standard deviation of productivity shocks will affect 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 *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 if international evidence of Zipf's coefficients is consistent with the evidence on the volatility of industry productivity shocks.

[5]Reversion to the mean in the productivity process can generate exogenous mean reversion in city growth rates.

4. NUMERICAL EXERCISES

This section is devoted to illustrating the solution 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 presented are asymptotic, and the long run distribution is stochastic. 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 logarithm of the transitory shocks are drawn.

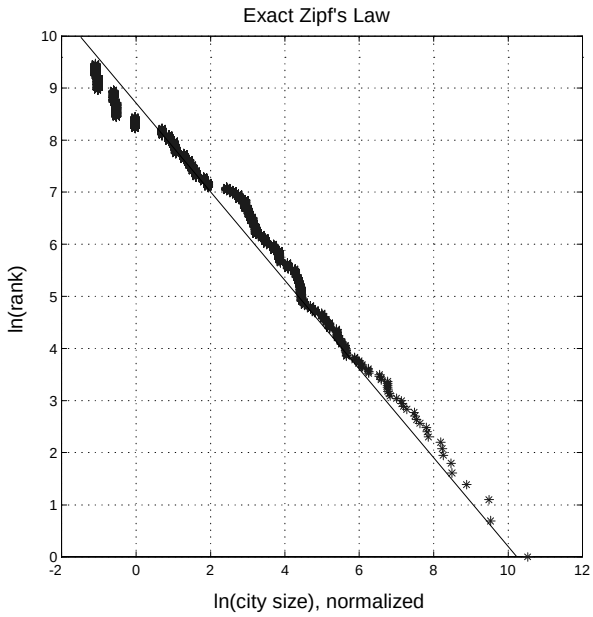

Figure Three

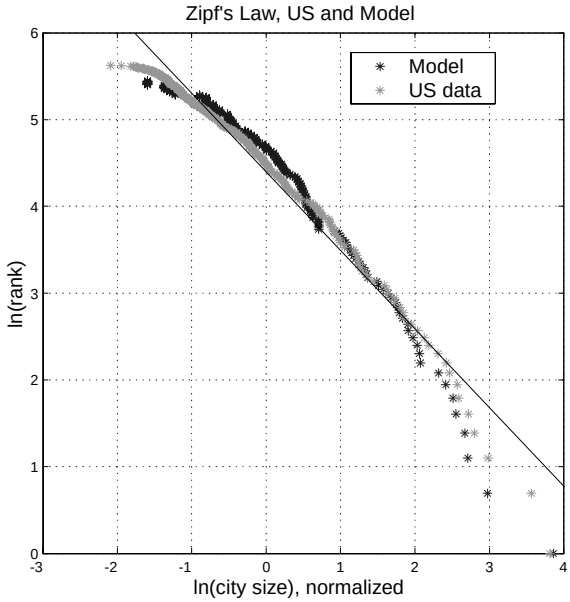

Figure Four

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. However, the model seems to be robust (not invariant) to all other parameter values. 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.

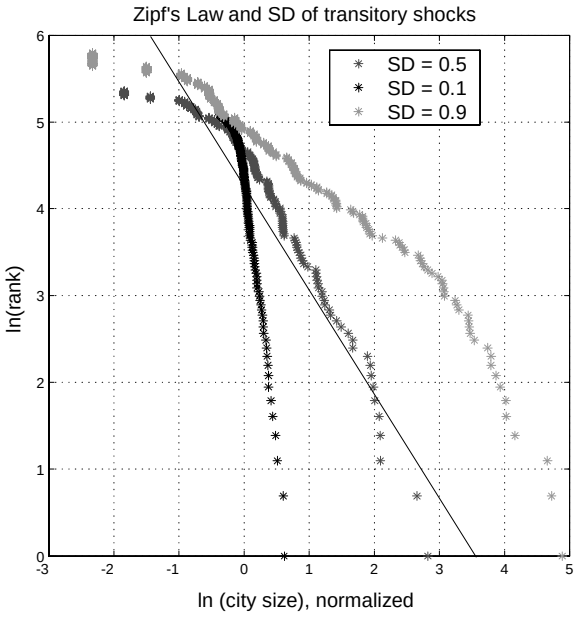

Figure Five

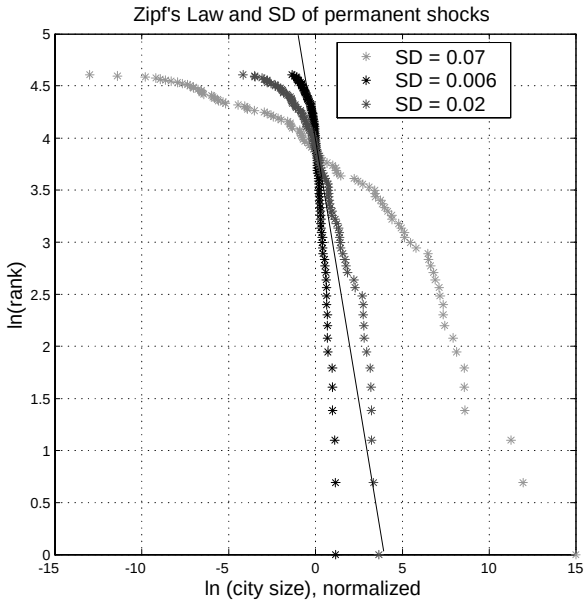

Figure Six

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, cities become less equal. The opposite happens if we reduce sd substantially, say to 0.1. Cities become more similar. Soo [19] 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 (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.

As we have mentioned, we can use permanent, instead of transitory, shocks in the model. This implies that in order to have city size distributions for which the coefficient of the Pareto distribution are close to one, we are constrained to using much lower standard deviations of shocks. Figure Six illustrates the effect of changes in the standard deviation of permanent shocks for $sd = 0.006, 0.02$ and 0.07.

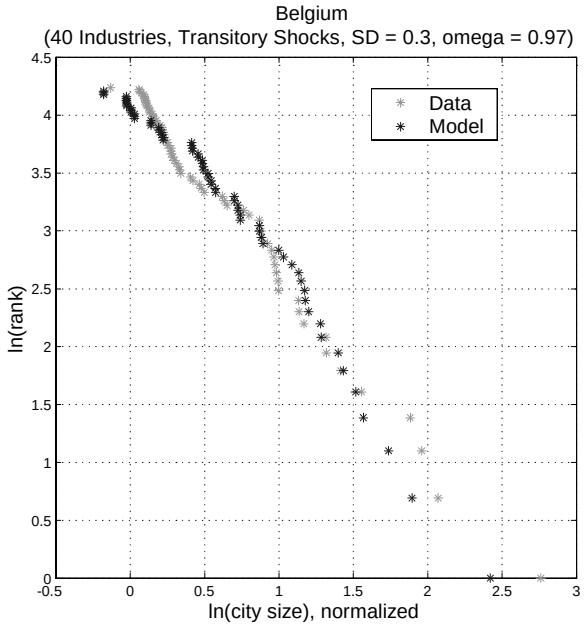

Figure Seven

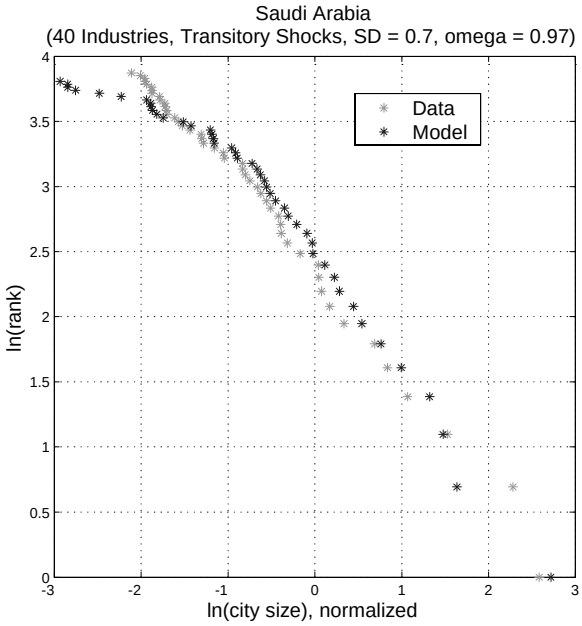

Figure Eight

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 0.3. The data and the simulation are presented in Figure Seven.

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 Eight shows the simulation and Saudi Arabia's data[6]. The standard deviation used in the numerical simulation is $sd = 0.7$.

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 will produce bounds on standard deviations that we will then compare with evidence on productivity shocks in the data. Horvath [13] measures the standard deviation and persistence of industry shocks in the United States for 36 industries[7].

It is important to stress that this comparison puts a heavy burden on our theory. To illustrate this, consider a situation where all of the standard deviations of productivity

[6]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.

[7]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.

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. Therefore, we know that 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, fully half of the industries have standard deviations that lie within these bounds[8].

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

[8]The estimates in the table were computed using city data from 73 countries. Data of agglomerations are only available for 26 mostly developed economies. Using agglomeration data, the corresponding number is 33%.

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.

Reality surely 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 United States	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 presented an urban growth theory in which cities arise endogenously out of a trade-off between agglomeration forces and congestion costs. Our theory is capable of reproducing several basic growth and urban facts. The urban structure itself leads to a reconciliation between the increasing returns at the local level that are necessary for agglomeration and constant returns at the aggregate level that are necessary for balanced growth. This has two additional implications for growth theory. First, tests for the presence of increasing returns should be conducted at the urban, and not the aggregate, level. Second, differences in the urban organization of economic activity may explain some part of the observed differences in total factor productivity across countries.

We find that the organization of economic activity in cities combined with productivity shocks and factor accumulation produce strong implications for the size distribution of cities. In particular, under special assumptions, the model predicts an exact version of Zipf's Law, while more generally the model can be used to explain some of the robust empirical deviations from Zipf's Law, including the underrepresentation of small cities and the fact that the largest cities are not large enough.

One of the features of the model is that it was especially tractable as a result of special functional form assumptions. We were able to solve for the entire growth path of the economy, and the entire urban structure, in closed form. A potentially important extension of this paper is to check the robustness of our results to different specifications of the economy. For example, at the moment the assumption of logarithmic preferences implies that the labor allocation across industries is fixed, and in all of the experiments conducted in the paper we have assumed that this is equal across industries. However, in the data, the size distribution of industry employment levels is already much closer to Zipf's Law. What are the assumptions on preferences that would yield a distribution of industry sizes closer to the one observed in the data? Other extensions include using different types of agglomeration effects, combining

productivity shocks with taste shocks, or adding amenities and nontraded goods or services to cities (for example, as in Gabaix [10] and Cordoba [5]).

An extension that deserves special consideration is to allow for different specifications of land ownership and city formation. The current specification, which follows the contributions of Henderson [12] and Black and Henderson [4], implies that resources are allocated efficiently across cities. The basic results will continue to hold in an environment with suboptimal cities as long as the deviation from optimality is roughly proportional: industries will still act as though they face constant returns to scale, expanding the number of cities at a suboptimal size. Moreover, as long as the equilibrium city size responds to variations in factor proportions, the same mechanisms will lead to a tendency toward Gibrat's Law of city growth.

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 influencing the ability of the mechanism to match features of the data. An empirical analysis of this parameter, and how it differs across countries, is certain to be an important part of any systematic empirical evaluation of our theory.

Finally, it is worth pointing out that Zipf's Law is also a strikingly good description of the size distribution of firms (see Axtell [2]). As it stands, our theory assumes internal constant returns to scale at the firm level, and hence the size distribution of firms is indeterminate. A natural question is whether the same processes we described could be used at the firm level. Specifically, assume that there are increasing returns in production, but that the firms must bear a 'managerial cost' that is increasing in the number of employees of the firm and is denominated in terms of the firm's own output. Suppose also that the firm accumulates its own factors and faces stochastic firm productivity shocks. Then a simple relabelling of terms would make the model of the paper also a model of the firm: instead of choosing the number of cities, the firm would choose the number of plants to operate. The firms as a whole would then behave as though they had constant returns to scale in the aggregate even though there were increasing returns at the plant level. Moreover, this would allow us to

imbed this model of the firm within our existing model of city formation in which there are external economies at the city level. Whether these elements can all be combined in a version of the above framework is the subject of future research.

REFERENCES

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

[2] Axtell, R. L. (2001). "Zipf Distribution of US Firm Sizes." *Science*, 293:1818-1820.

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

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

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

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

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

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

[9] 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.

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

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

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

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

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

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

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

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

[18] 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.

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

APPENDIX

Solution of Social Planner's Problem

Our first task is to solve the planning problem:

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

subject to for, all t and j,

$$K_{t+1j} = K_{tj}^{\omega_j} X_{tj}^{1-\omega_j},$$
$$H_{t+1j} = H_{tj}\left[B_j^0 + (1-u_{tj})B_j^1\right],$$
$$F_j A_{tj} H_{tj}^{\hat{\alpha}_j} K_{tj}^{\hat{\beta}_j} N_{tj}^{1-\hat{\alpha}_j-\hat{\beta}_j} u_{tj}^{\hat{\phi}_j} = C_{tj} + X_{tj},$$

and

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

To solve this problem, we can verify that the value function of the problem takes the form

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

This leads to

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

which implies that

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

We can use this result to obtain an expression for u_{tj},

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

and N_{tj}^*,

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

where

$$D_j^K = \frac{(1-\delta)\theta_j\hat{\beta}_j}{1-\delta\omega_j-\delta\left(1-\omega_j\right)\hat{\beta}_j},$$

and

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

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

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

For physical capital

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

where

$$x_j = \frac{\delta D_j^K\left(1-\omega_j\right)}{\delta D_j^K\left(1-\omega_j\right)+(1-\delta)\theta_j}.$$

Of course,

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

so

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

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 $N_t=\left(g_N\right)^t N_0$, we obtain

$$\begin{aligned}&\lim_{t\to\infty}\left[\ln K_{tj}-\ln K_{t-1j}\right]\\ =\;&\left(1-\omega_j\right)\lim_{t\to\infty}\left[\ln(A_{t-1j})-\sum_{T=1}^{t}\frac{\left(\omega_j+\left(1-\omega_j\right)\hat{\beta}_j\right)^{t-1-T}}{\left(1-\left(\omega_j+\left(1-\omega_j\right)\hat{\beta}_j\right)\right)^{-1}}\ln(A_{T-1j})\right]\\ &+\frac{1}{1-\hat{\beta}_j}\left[\left(1-\hat{\alpha}_j-\hat{\beta}_j\right)g_N+\hat{\alpha}_j\ln\left(B_j^0+(1-u_j^*)B_j^1\right)\right].\end{aligned}$$

The size of the city is given by

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

so

$$\begin{aligned}
\ln\left(\frac{N_{tj}}{\mu_{tj}}\right) &= 2\left[\ln\left(\frac{2\left(\varepsilon_j+\gamma_j\right)}{bn_j}\right) + \ln\left(Y_{tj}\right) - \ln(N_t)\right] \\
&= 2\left[\ln\left(\frac{n_j F_j 2\left(\varepsilon_j+\gamma_j\right)}{bn_j^{\hat{\alpha}_j+\hat{\beta}_j}(1-2\left(\varepsilon_j+\gamma_j\right))}\right) + \ln\left(A_{tj}\right) + \hat{\alpha}_j \ln\left(H_{tj}\right)\right. \\
&\quad \left. + \hat{\beta}_j \ln\left(K_{tj}\right) - \left(\hat{\alpha}_j + \hat{\beta}_j\right)\ln(N_t) + \hat{\phi}_j \ln\left(u_j^*\right)\right].
\end{aligned}$$

Hence,

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

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

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

Imposing constant population growth,

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

Equilibrium Allocation

Firms.—

The problem of the firm is to hire labor and human and physical capital to maximize profits given prices for these inputs and taking as given the total amount of labor input in the city (and hence the size of the externality term) and factor prices. 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 written in terms of some numeraire commodity, the firm's optimization problem yields

$$\begin{aligned} W_{tj}/P_{tj} &= \left(1-\alpha_j-\beta_j\right) Y_{tj}/\left(u_{tj}N_{tj}\right), \\ \left(1-\tau_{tj}^k\right) R_{tj}/P_{tj} &= \beta_j Y_{tj}/K_{tj}, \\ \left(1-\tau_{tj}^h\right) S_{tj}/P_{tj} &= \alpha_j Y_{tj}/H_{tj}, \end{aligned}$$

where τ_{tj}^k and τ_{tj}^h are subsidies paid by city developers to attract firms to a particular city.

As noted above in our discussion of the property developer's problem, all cities producing good j will be the same size. Note that in our framework, all cities producing good j are identical, so that if there are μ_{tj} cities producing good j, the amounts of labor and human capital in any one city are given by H_{tj}/μ_{tj} and N_{tj}/μ_{tj}.

Households.—

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

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

where B_j^0 and B_j^1 are some positive constants.

Clearly, households will allocate their labor and human and physical capital services to the cities with the highest wages and rental rates, so that in an equilibrium these must be equal across all cities producing a given good. If we let W_{tj}, R_{tj}, and S_{tj} denote state contingent sequences of wages and rental rates in each industry and P_{tj} denote the sequence of state contingent output prices, then the household's problem is to maximize

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

subject to sequences of flow budget constraints

$$\begin{aligned}
&\sum_{j=1}^{J} P_{tj}\left[C_{tj}+X_{tj}+\left[ACC_{tj}+AR_{tj}\right]N_{tj}\right] \\
\leq\ &\sum_{j=1}^{J}\left[W_{tj}N_{tj}u_{tj}+R_{tj}K_{tj}+S_{tj}H_{j}+P_{tj}T_{tj}N_{tj}\right],
\end{aligned}$$

where ACC_{tj} and AR_{tj} represent average commuting costs and average rents. The laws of motion for human and physical capital

$$\begin{aligned}
K_{t+1j} &= K_{tj}^{\omega_j}X_{tj}^{1-\omega_j}, \\
H_{t+1j} &= H_{tj}\left[B_{j}^{0}+(1-u_{tj})B_{j}^{1}\right],
\end{aligned}$$

and the constraint on labor allocation

$$\sum_{j} N_{tj} \leq N_{t}.$$

Note that the prices P_j, W_j, R_j, and S_j all depend on the economywide state variables $\bar{H}_j, \bar{K}_j$ and $\bar{A}_j$. The state vector for each household also includes the household's stocks of human and physical capital, H_j and K_j.

City Developers.—

City developers aim to maximize rents net of transfers offered to households and subsidies paid to firms in order to attract them to the city. That is, city developers choose factor inputs in the city $N_{tj}/\mu_{tj}, K_{tj}/\mu_{tj}$ and H_{tj}/μ_{tj}, transfers to households T_{tj} and subsidies to physical and human capital, $\tau_{tj}^{k}, \tau_{tj}^{h}$, to maximize

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

subject to

$$\begin{aligned}
\left(1-\tau_{tj}^{k}\right)R_{tj}/P_{tj} &= \beta_{j}Y_{tj}/K_{tj}, \\
\left(1-\tau_{tj}^{h}\right)S_{tj}/P_{tj} &= \alpha_{j}Y_{tj}/H_{tj}, \\
I_{tj} &= \left(1-\alpha_{j}-\beta_{j}\right)\frac{Y_{tj}}{N_{tj}} + T_{tj} - \frac{3b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{\frac{1}{2}}.
\end{aligned}$$

Competition from other developers ensures that profits are zero, so

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

Equilibrium.—
We are now in a position to define a competitive equilibrium for this economy.

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:*

$$\begin{aligned} C_{tj} + X_{tj} + bN_{tj}^{\frac{3}{2}}\mu_{tj}^{-\frac{1}{2}} &= Y_{tj}, \\ \sum_{j=1}^{J} N_{tj} &= N_t. \end{aligned}$$

Proofs of Propositions

Proposition 1 *There exists a unique Pareto efficient allocation for this economy.*

Proof. 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

$$A_{tj}K_{tj}^{\beta_j}H_{tj}^{\alpha_j+\gamma_j}N_{tj}^{1-\alpha_j-\beta_j+\varepsilon_j}u_{tj}^{1-\alpha_j-\beta_j}\mu_{tj}^{-\varepsilon_j-\gamma_j} - bN_{tj}^{\frac{3}{2}}\mu_{tj}^{-\frac{1}{2}}. \tag{13}$$

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

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

Then it is easy to see that

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

under our assumption that $\varepsilon_j + \gamma_j < 1/2$, is strictly increasing in μ_{tj}, equals zero 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 (13) is negative, while for all other μ it is strictly positive. Moreover, in the limit as μ goes to infinity, the expression in (13) goes to zero. Hence, as the expression is continuous in μ, it possesses a maximum on $[\mu^*, +\infty)$, which from the first order necessary condition satisfies

$$TCC_{tj} \equiv bN_{tj}^{\frac{3}{2}}\mu_{tj}^{-\frac{1}{2}} = 2\left(\varepsilon_j + \gamma_j\right)Y_{tj}.$$

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:

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

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. ■

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

Proposition 3 *Every competitive equilibrium in this economy is Pareto efficient.*

Proof. Let us start with the solution of the SPP. We know that this solution is the unique allocation satisfying the first order condition of the SPP. That problem is to choose

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

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

$$K_{t+1j} = K_{tj}^{\omega_j}X_{tj}^{1-\omega_j},$$
$$H_{t+1j} = H_{tj}\left[B_j^0 + (1-u_{tj})B_j^1\right],$$
$$N_t = \sum_{j=1}^{J} N_{tj} = \sum_{j=1}^{J} \mu_{tj}\tilde{N}_{tj}.$$

If we let the multipliers on these constraints be denoted respectively by λ_{tj}^{SP}, $\gamma_{Ktj}^{SP}, \gamma_{Htj}^{SP}$ and γ_{Nt}^{SP}, the first order conditions are

$$\begin{aligned}
(1-\delta)\,\delta^t N_t\theta_j\frac{1}{C_{tj}} &= \lambda_{tj}^{SP} \\
\gamma_{Ktj}^{SP}\left(1-\omega_j\right)K_{tj}^{\omega_j}X_{tj}^{-\omega_j} &= \lambda_{tj}^{SP} \\
\lambda_{tj}^{SP}\left(1-\alpha_j-\beta_j\right)\frac{Y_{tj}}{u_{tj}} &= \gamma_{Htj}^{SP}B_j^1H_{tj} \\
\lambda_{tj}^{SP}\left[\left(1-\alpha_j-\beta_j+\varepsilon_j\right)\frac{Y_{tj}}{N_{tj}} - \frac{3b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{\frac{1}{2}}\right] &= \gamma_{Nt}^{SP} \\
\lambda_{tj}^{SP}\left[\frac{b}{2}\left(\frac{N_{tj}}{\mu_{tj}}\right)^{\frac{3}{2}} - \left(\varepsilon_j+\gamma_j\right)\frac{Y_{tj}}{\mu_{tj}}\right] &= 0 \\
E_t\left\{\lambda_{t+1j}^{SP}\beta_j\frac{Y_{t+1j}}{K_{t+1j}} + \gamma_{Kt+1j}^{SP}\omega_jK_{t+1}^{\omega_j-1}X_{t+1j}^{1-\omega_j}\right\} &= \gamma_{Ktj}^{SP} \\
E_t\left\{\lambda_{t+1j}^{SP}\left(\alpha_j+\gamma_j\right)\frac{Y_{t+1j}}{H_{t+1j}} + \gamma_{Ht+1j}^{SP}\left[B_j^0+\left(1-u_{t+1j}\right)B_j^1\right]\right\} &= \gamma_{Htj}^{SP}.
\end{aligned}$$

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

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

subject to sequences of flow budget constraints

$$\begin{aligned}
&\sum_{j=1}^{J} P_{tj}\left[C_{tj}+X_{tj}+\left\{ACC_{tj}+AR_{tj}\right\}N_{tj}\right] \\
\leq\ &\sum_{j=1}^{J}\left[W_{tj}N_{tj}u_{tj}+R_{tj}K_{tj}+S_{tj}H_{j}+P_{tj}T_{tj}N_{tj}\right],
\end{aligned}$$

the laws of motion for human and physical capital

$$\begin{aligned}
K_{t+1j} &= K_{tj}^{\omega_j}X_{tj}^{1-\omega_j}, \\
H_{t+1j} &= H_{tj}\left[B_{j}^{0}+(1-u_{tj})B_{j}^{1}\right],
\end{aligned}$$

and the constraint on labor allocation

$$\sum_{j} N_{tj} \leq N_{t}.$$

Letting λ_t^{HH} be the multipliers on budget constraints, γ_{Ktj}^{HH} and γ_{Htj}^{HH} be those on physical and human capital accumulation, and γ_{Nt}^{HH} be that on labor supply, the first order conditions of the household are

$$(1-\delta)\,\delta^{t}N_{t}\theta_{j}\frac{1}{C_{tj}}=\lambda_{t}^{HH}P_{tj}$$

$$\gamma_{Ktj}^{HH}\left(1-\omega_{j}\right)K_{tj}^{\omega_j}X_{tj}^{-\omega_j}=\lambda_{t}^{HH}P_{tj}$$

$$\lambda_{t}^{HH}W_{tj}N_{tj}=\gamma_{Htj}^{HH}B_{j}^{1}H_{tj}$$

$$\lambda_{t}^{HH}\left\{P_{tj}\left[T_{tj}-ACC_{tj}-AR_{tj}\right]+W_{tj}u_{tj}\right\}=\gamma_{Nt}^{HH}$$

$$E_{t}\left\{\lambda_{t+1}^{HH}R_{t+1j}+\gamma_{Kt+1j}^{HH}\omega_{j}K_{t+1j}^{\omega_{j-1}}X_{t+1j}^{1-\omega_j}\right\}=\gamma_{Ktj}^{HH}$$

$$E_{t}\left\{\lambda_{t+1}^{HH}S_{t+1j}+\gamma_{Ht+1j}^{HH}\left[B_{j}^{0}+\left(1-u_{t+1j}\right)B_{j}^{1}\right]\right\}=\gamma_{Htj}^{HH}.$$

2. Firms optimize:

$$\begin{aligned}
W_{tj}/P_{tj} &= \left(1-\alpha_{j}-\beta_{j}\right)Y_{tj}/N_{tj}, \\
\left(1-\tau_{tj}^{k}\right)R_{tj}/P_{tj} &= \beta_{j}Y_{tj}/K_{tj}, \\
\left(1-\tau_{tj}^{h}\right)S_{tj}/P_{tj} &= \alpha_{j}Y_{tj}/H_{tj}.
\end{aligned}$$

3. Developer choices and free entry:

The relevant first order conditions from the developer's problem after some rearranging can be expressed as

$$\begin{aligned} \tau_{tj}^{k}\frac{R_{tj}}{P_{tj}} &= 0, \\ \tau_{tj}^{h}\frac{S_{tj}}{P_{tj}} &= \gamma_j\frac{Y_{tj}}{H_{tj}}, \\ T_{tj} &= \varepsilon_j\frac{Y_{tj}}{N_{tj}}. \end{aligned}$$

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

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

Substituting the last first order condition, we obtain

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

which is exactly the first order condition of the social planner's problem with respect to μ_{tj}. Using the second first order condition and the fact that firms choose human capital optimally, we know that

$$\tau_{tj}^{h} = \frac{\gamma_j}{\alpha_j + \gamma_j}.$$

4. Markets clear:

$$\begin{aligned} C_{tj} + X_{tj} + bN_{tj}^{\frac{3}{2}}\mu_{tj}^{-\frac{1}{2}} &= Y_{tj}, \\ \sum_{j=1}^{J} N_{tj} &= N_t. \end{aligned}$$

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 (that is, it is sufficient to establish the existence of the appropriate set of Lagrange 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

$$\lambda_{tj}^{SP} = \lambda_{t}^{HH}P_{tj}.$$

Looking at first order conditions in investment, we get

$$\frac{\lambda_{tj}^{SP}}{\gamma_{Ktj}^{SP}} = \frac{\lambda_{t}^{HH} P_{tj}}{\gamma_{Ktj}^{HH}},$$

which, using the first equivalence, implies

$$\gamma_{Ktj}^{SP} = \gamma_{Ktj}^{HH}.$$

Looking at the first order condition in u_{tj} we get from the household's equation

$$B_j^1 H_{tj} = \frac{\lambda_t^{HH}}{\gamma_{Htj}^{HH}} W_{tj} N_{tj}.$$

Substituting for W_{tj} and rearranging, this implies

$$\gamma_{Htj}^{SP} = \gamma_{Htj}^{HH}.$$

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 SP problem 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

$$\frac{W_{tj}}{P_{tj}} u_{tj} - ACC_{tj} - \gamma_j \frac{Y_{tj}}{N_{tj}} = \frac{\gamma_{Nt}^{HH}}{P_{tj} \lambda_t^{HH}}.$$

Substituting for real wages, we get

$$\left(1 - \alpha_j - \beta_j - \gamma_j\right) \frac{Y_{tj}}{N_{tj}} - b \left(\frac{N_{tj}}{\mu_{tj}}\right)^{\frac{1}{2}} = \frac{\gamma_{Nt}^{HH}}{P_{tj} \lambda_t^{HH}}.$$

Substituting the results from the city developer's problem, we obtain

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

This latter equation is the same as the first order condition for N_{tj} from the social planner's problem under the equivalence

$$\frac{\gamma_{Nt}^{HH}}{P_{tj} \lambda_t^{HH}} = \frac{\gamma_{Nt}^{SP}}{\lambda_{tj}^{SP}}.$$

■

Proposition 4 *(Exact Gibrat's Law and Zipf's Law) The growth process of city sizes satisfies Gibrat's Law, and therefore 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* $\left(\beta_j = \hat{\beta}_j = 0 \text{ or } \omega_j = 1\right)$*, 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

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

But this is independent of city size.

To show that this implies an invariant distribution that satisfies Zipf's Law, we can apply the results of Gabaix [10] and Cordoba [5]. Under our restriction of ex-ante industry heterogeneity, we can do this group by group. The results then follow from Propositions 1 and 2 in Gabaix [10], which show that an invariant distribution satisfying Zipf's Law is the result of the limit of the processes above augmented with a reflecting barrier as this barrier goes to zero. ■

Proposition 5 *(Concavity) If conditions 1 and 2 in Proposition 4 are not satisfied, the growth rate for cities exhibits reversion to the mean.*

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 $E\left(\ln A\right)$, then the expected growth rate of the city is zero. By contrast, if the past shocks have a weighted average greater than (less than) $E\left(\ln A\right)$, then the expected growth rates are negative (positive). ■

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. ■