DISCUSSION PAPER SERIES

No. 3577

CITY SIZE DISTRIBUTIONS AS A CONSEQUENCE OF THE GROWTH PROCESS

Gilles Duranton

INTERNATIONAL TRADE

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP3577.asp**

ISSN 0265-8003

CITY SIZE DISTRIBUTIONS AS A CONSEQUENCE OF THE GROWTH PROCESS

Gilles Duranton, London School of Economics (LSE) and CEPR

Discussion Paper No. 3577
October 2002

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 3577

October 2002

ABSTRACT

City Size Distributions as a Consequence of the Growth Process*

The size distribution of cities in many countries follows some broadly regular patterns. Any good theory of city size distributions should (i) be able to account for this regularity, but also (ii) rely on a plausible economic mechanism and (iii) be consistent with other fundamental features of cities like the existence of agglomeration economies and crowding costs. Unlike the previous literature, the model proposed here satisfies these three requirements. It views small innovation-driven technological shocks as the main engine behind the growth and decline of cities. Cities grow or decline as they win or lose industries following new innovations. Formally, this is achieved by embedding the quality-ladder model of growth developed by Grossman and Helpman (1991) in an urban framework.

JEL Classification: O18, R11 and R12
Keywords: agglomeration economies, city-size distribution and quality-ladder models of growth

Gilles Duranton
Department of Geography &
Environment
London School of Economics
Houghton Street
LONDON
WC2A 2AE
Tel: (44 20) 7955 7604
Fax: (44 20) 7955 7412
Email: g.duranton@lse.ac.uk

* This Paper is produced as part of a CEPR Research Network on ‘The Economic Geography of Europe: Measurement, Testing and Policy Simulations’, funded by the European Commission under the Research Training Network Programme (Contract No: HPRN-CT-2000-00069). I am grateful to Yannis Ioannides, Janet Kohlhase, Henry Overman, John Parr, Diego Puga, participants at the 2001 RSAI North American meetings in Charleston, the 2002 CEPR workshop on economic geography in Villars and the 2002 ESRC seminar in regional and urban economics in Preston for helpful comments and suggestions.

Submitted 12 August 2002

NON-TECHNICAL SUMMARY

One of the most striking and least understood facts about cities regards their differences in population sizes. In the United States for instance, the Consolidated Metropolitan Statistical Area of New York has a population count above 21 million whereas Lynchburg (Virginia) has around 210,000 inhabitants and Paris (Texas) only slightly more than 21,000. The Paper proposes a novel approach to account for these very large differences in size. Small innovation-driven technological shocks at the industry level are viewed as the main engine behind the growth and decline of cities.

In developing this approach, I have been guided by the principle that any good theory of city size should meet three requirements. First, it should be able to replicate existing empirical regularities about size distributions with some accuracy. Many theories may be able to generate some unevenness with respect to city sizes. A more stringent empirical fit is thus necessary to discriminate between these theories. Second, the model should rely on a plausible economic argument and be based on sound microeconomic foundations. Unspecified or implausible economic mechanisms do not deepen our understanding of cities nor are they being helpful with respect to normative and policy questions. Third, any good theory of city size and urban development should be consistent with fundamental features of cities such as the existence of agglomeration economies and crowding costs increasing with city population. Such requirement is necessary because any theory of city size should be part of a broader body of knowledge regarding the economics of cities. The model developed below satisfies these three requirements.

According to the empirical literature, there is indeed some regularity in the size distribution of cities. The so-called Zipf's law, however, which claims that the size distribution of cities follows a Pareto distribution with power –1, must be rejected and only constitutes a poor approximation of existing distributions. Consequently, satisfying the first requirement above should not imply attempting to generate Zipf's law but rather replicating existing city size distributions. In the Paper, I focus on the US distribution, which has been extensively studied, as well as the French distribution, which exhibits features very different from its US counterpart, as a robustness test.

Turning to the second requirement, the economic mechanism I would like to highlight is the following. Anecdotal evidence suggests that technology shocks at the industry level constitute an important channel of urban growth and decline. Extreme examples range from the demise of the steel industry in Pittsburgh to the rise of internet-related industries in San Francisco / San Jose or the growth of financial services in New York. More systematic evidence shows that changes in employment in US metropolitan areas are mostly caused by local shocks at the industry level. Furthermore, the literature also

suggests that the location of production in many industries changes significantly over time. Finally, the changes in the spatial patterns of production in the US are closely related to the processes of plant creation and destruction.

To model this argument and show how innovation-driven shocks within industries can explain existing patterns of city sizes, I embed the canonical quality-ladder model of growth developed by Grossman and Helpman in an urban framework. In each industry, research firms try to invent the next step up the quality ladder in order to reap monopoly profits. Research firms may be successful in their own industry or may develop a new leading quality in another industry. Local spillovers induce research firms in an industry to locate with production in the same industry and in most industries, successful innovators need to start producing where they did their research. This implies that own-industry innovations lead to a change of monopoly but to no change of location for the industry. Cross-industry innovations by contrast imply not only a change of monopoly but also typically a change of location since the old and new monopolies are not in general located in the same city.

Rochester and New York provide two very interesting illustrations of this process. In the late 19th century, New York was the capital of the newly created photographic industry, whereas Rochester was at the forefront of technological developments in precision instruments. George Eastman, while working at improving optical instruments in Rochester, invented an emulsion-coating machine which enabled him to mass-produce photographic dry plates. His company soon took over the market for photographic films. As a consequence, Rochester replaced New York as the main centre for the industry. Rochester, 50 years later, was still the capital of the US film industry, whereas New York was that of the duplication industry. Then, Haloid Company, a firm specialized in the manufacturing of photographic papers and operating in the shadow of Eastman Kodak, came up with a new process for making copies without the need for developing. The process, called xerography, made Rochester the new capital of the duplication industry in place of New York (again) where the previous dominant technology, the varityper was produced.

In steady-state and in absence of agglomeration economies and urban crowding, this dynamic process generates city size distributions for which the Zipf's curve is concave. When calibrating the model on the US urban system, the performance is mixed. One the one hand, the simulations approximate the US distribution of city sizes better than Zipf's law. On the other hand, the simulated Zipf's curves are a bit too concave with cities at the top of the distribution being smaller than the likes of New York and Los Angeles and cities in the middle of the distribution being a bit too large.

This simple benchmark however is such that there are neither costs nor benefits to city size. To satisfy my third requirement, the model is extended to allow for urban congestion and economies of agglomeration. When agglomeration economies dominate crowding costs, the probability of innovating in a city increases more than proportionately to its size. This reduces the Zipf's exponent in the upper tail and increases it in the lower tail. Under realistic values for the trade-off between dynamic agglomeration economies and crowding costs, it is possible to generate urban systems very similar to that of the US. The French urban system can also be replicated. Hence adding realistic urban features to the model strengthens its empirical fit. The model can be extended even further and made consistent with city creation, new products development and plant relocations.

1. Introduction

One of the most striking and least understood facts about cities regards their differences in population sizes. In the United States (US) for instance, the Consolidated Metropolitan Statistical Area (CMSA) of New York has a population count above 21 million whereas Lynchburg (Virginia) has around 210,000 inhabitants and Paris (Texas) only slightly more than 21,000. This paper proposes a novel approach to account for these very large differences in size. Small innovation-driven technological shocks at the industry level are viewed as the main engine behind the growth and decline of cities.

In developing this approach, I have been guided by the principle that any good theory of city size should meet three requirements. First, it should be able to replicate existing empirical regularities about size distributions with some accuracy. Many theories may be able to generate some unevenness with respect to city sizes. A more stringent empirical fit is thus necessary to discriminate between these theories. Second, the model should rely on a plausible economic argument and be based on sound microeconomic foundations. Unspecified or implausible economic mechanisms do not deepen our understanding of cities nor are they being helpful with respect to normative and policy questions. Third, any good theory of city size and urban development should be consistent with fundamental features of cities such as the existence of agglomeration economies and crowding costs increasing with city population. Such requirement is necessary because any theory of city size should be part of a broader body of knowledge regarding the economics of cities. The model developed below satisfies these three requirements.

The empirics of city size distributions

Since Auerbach (1913) and Zipf (1949), the well-known claim is that in a large number of countries, if one ranks cities from the largest to the smallest and correlates this with their population in the following manner:

$$\log \text{Rank} = \text{Constant} - \xi \log \text{Size} \, ,$$

the estimated coefficient ξ (known as the *Zipf's exponent*) is very close to one. This statistical relation (i.e., the *Zipf's curve* being best approximated by a straight line with a slope -1) is referred to as *Zipf's law* and it corresponds to having city sizes randomly drawn from a Pareto distribution with exponent -1. This claim is however controversial. Krugman (1996) speaks of "one of the most overwhelming empirical regularities in economics" and states that Zipf's law describes the US city size distribution remarkably well, whereas Black and Henderson (1998), who also look at US cities, conclude that "this is an inappropriate framework to use". In light of the first requirement above, a deeper examination of the facts is obviously needed before going any further.

Rosen and Resnick (1980) and Soo (2002) are the most complete comparative studies on the issue. Rosen and Resnick (1980) examine the distribution of city sizes for 44 countries in 1970. Their main finding is that the average Zipf's exponent across countries is close to unity — precisely 1.13 with a standard deviation of 0.19. For all countries but one in their sample, the Zipf's exponent is between 0.8 and 1.5. Soo (2002) confirms most of these findings for an enlarged sample of 75

countries over the last 30 years. For the more recent period, he obtains a mean Zipf's exponent of 1.10 for cities. In his sample, 71 out of 75 countries have a Zipf's exponent between 0.8 and 1.5.

These cross-country differences are small compared to similar coefficients calculated on firm size distributions across industries or individual income distributions across countries (Arnold, 1983). Hence, it is fair to first conclude that there is indeed some regularity in the size distribution of cities. However the existing evidence offers only weak support for taking the Pareto distribution with exponent -1 as the underlying "true" distribution. Zipf's law is only a very rough approximation.[1] Consequently, *satisfying the first requirement above should not imply attempting to generate Zipf's law but rather replicating existing city size distributions*. Below I focus on the US distribution, which has been extensively studied, as well as the French distribution, which exhibits features very different from its US counterpart, as a robustness test.

Contradicting threads in the literature

Trying to replicate empirical distributions is very much in contrast to the existing theoretical literature. The prevailing options have been instead either to take Zipf's law very seriously or to forget completely about existing regularities in city size distribution.

The main purpose of most of the literature taking Zipf's law seriously (in urban economics just like in industrial organisation) is to propose statistical processes generating unitary Zipf's exponents regardless of their plausibility (see Sutton, 1998, for a critical review of this literature). Simon (1955) is still arguably the best model in this strand of literature. He assumes that the urban population grows over time by discrete increments. With some probability, a new lump goes to form a new city. Otherwise it is added to an existing city, with the probability that any particular city gets it proportional to its population. This mechanism generates a Pareto distribution for city sizes. The Zipf's exponent falls to one at the limit as the probability of new cities being created goes to zero.

This model is somewhat problematic with respect to the first requirement set above because it is unable to generate Zipf's exponents below one (which are observed in 12 out of the 44 countries studied by Rosen and Resnick, 1980). Besides, it can generate Zipf's law only when the urban population goes to infinity (which implies cities of infinite sizes and infinitely slow convergence). Because of its complete lack of economic content, this model also fails to pass my second requirement. It can *generate* Zipf's law in a statistical sense but is unable to *explain* it in terms of economic or social forces. Finally it does not satisfy my third requirement either since it seems inconsistent with agglomeration economies being at the root of the existence of cities.

[1] Rosen and Resnick (1980) highlight two systematic deviations from Zipf's law. First, primacy, that is the largest city being much larger than predicted by Zipf's law, is prevalent in many countries. Second, there are systematic deviations from Zipf's law in the lower tail of the distribution with either too many or too few small cities. These lower tail deviations from Zipf's law are undoubtedly at the root of many of the disagreements in the recent literature about the validity of Zipf's law. For instance, Krugman (1996) gets a Zipf's exponent of 1.00 from the 130 largest US cities, whereas Black and Henderson (1998) obtain a Zipf's exponent close to 0.8 using a much larger sample of around 300 US metropolitan areas. In further research, Dobkins and Ioannides (2000) estimate directly the counter-cumulative of city size distribution for the US. They find a Zipf's exponent above one in the upper tail of the distribution and well below one in the lower tail. These results on US cities are consistent with the non-parametric estimation of local Zipf's exponents by Ioannides and Overman (2002).

Independently of this statistical approach and starting with Henderson (1974), a large economic literature on urban systems has developed. Its basic insight is the recognition of a tension between local external economies of scale and dis-economies of urban crowding. This tension implies that urban efficiency is a concave function of city size. Under mild assumptions regarding their management, cities reach their optimal size in equilibrium. Furthermore, with local external economies of scale taking place within industries, cities specialise in equilibrium (if not fully, at least in their production of tradable goods) and their sizes vary depending on their specialisations since technology differs across industries. For instance, if localisation economies are stronger for banking than for textile, banking cities are larger in equilibrium than textile cities.

This type of model provides a very fruitful way to think about cities and gives consistent explanations for many observed stylised facts about what cities are and what they do (see Duranton and Puga, 2000, for a survey of this literature). Regularities about the size distribution of cities stand here as conspicuous exceptions.[2] In conclusion this literature does well on the second and third requirements but fails in its ability to deal with city sizes, the first requirement set above.

Gabaix (1999*a*) stands as a landmark in the literature for being the first to propose a model that not only generates Zipf's law but also explains it using an economic argument. The key hypothesised mechanism is that cities grow and decline in population following exogenous idiosyncratic shocks on their amenities. The details are as follows. Overlapping generations of workers choose at birth a city where to live depending on local wages and amenities. Wages are in turn determined locally by the ratio of old to young workers as old immobile workers complement the young in the production function. In equilibrium, the utility of young workers, which is given by the product of amenities by wages, is equalised across cities. A positive amenity shock in a city leads to an influx of young workers. But this crowds out the local labour market so that any positive amenity shock is partially offset by lower wages. More precisely, amenity shocks, as they enter multiplicatively in the utility function, lead to population shocks in cities proportional to their current population. Hence, when amenity shocks are identical and independently distributed, the growth of a city is independent of its size. Gabaix (1999*a*) demonstrates that this crucial property, known as Gibrat's law, implies Zipf's law in steady-state. To see this, consider for instance that cities may double in size with probability $1/3$ or halve with probability $2/3$ so that the expected growth of a city is zero. In steady-state, for the distribution of city sizes to remain constant, there must be twice as many cities of size 1 than cities of size 2, etc. This is Zipf's law. As argued above, generating Zipf's law may not be the right thing to do but Gabaix (1999*a*) shows that his approach can also account for some of the known deviations from Zipf's law observed in the data. The first requirement above can thus be taken to be satisfied.

However, with respect to the specific channel of transmission proposed by Gabaix (1999*a*), namely young workers migrating to cities that just received positive amenity shocks, the evidence is far from supportive. Although the quality-of-life literature typically reports large cross-city

[2]With this static approach, city size is a positive function of the intensity of localisation economies. The number of cities of each type is also determined by demand parameters. There is no a priori reason why the distribution of localisation economies across industries and demand parameters should generate the observed regularities in the distribution of city sizes. Instead, the main prediction regarding the distribution of city sizes is that there should be clusters of cities of the same type and size.

differences for amenity values, these are much smaller once outliers are eliminated. According to Gyourko, Khan, and Tracy (1999), the interquartile range is around 1500 dollars for US cities. A large fraction of it can be accounted for by fiscal variables (and these should play no role in Gabaix' model). Another large fraction of it is climate, which does not frequently receive city-specific shocks. For amenities subject to non-fiscal shocks (like policing), the estimates are very low and cannot justify large migration flows. For instance, Gyourko *et al.* (1999) report that one standard deviation in crime is valued at 83 dollars. In line with this, the literature on internal migrations shows that amenities are not significant in the choice of destinations of young migrants (but play a significant role in that of older workers – the opposite of Gabaix' assumption). Instead, employment opportunities and wages come up empirically as far more important to determine migration choices (see Greenwood, 1997, for a survey). Furthermore, the causality appear to be more from population growth to worse amenities or from population decline to better amenities than from amenity improvements to population growth.[3]

Although further empirical work is certainly warranted, Gabaix (1999*a*)'s model does not seem to satisfy the second requirement set above: Amenities constitute theoretically a perfectly valid source of shocks leading to Zipf's law, but empirically the magnitude of the forces at stake appears implausibly small. Furthermore, the theory proposed by Gabaix (1999*a*) neither proposes any reason for the existence of cities in the first place nor is compatible in any direct fashion with the existence of local external returns, the third requirement set above. In Gabaix (1999*b*), the basic mechanism of Gabaix (1999*a*) is extended to account for external returns and urban crowding. The sum of the two however is assumed to follow a negative power law with exponent -1 to offset exogenous total factor productivity shocks at the urban level. This does not appear very realistic. Recent work by Córdoba (2001) explores a wide range of statistical processes also based on Gibrat's law and confirms the basic crux of Gabaix (1999*a*)'s model. He even argues in favour of rejecting the idea of local increasing returns, which has proved so fruitful to understand cities since Marshall (1890).

Preview of the argument

Rather than trying to build on a literature that does not appear very successful, my point of departure is the following. Anecdotal evidence suggests that technology shocks at the industry level constitute an important channel of urban growth and decline. Extreme examples range from the demise of the steel industry in Pittsburgh to the rise of internet-related industries in San Francisco / San Jose or the growth of financial services in New York. Further supportive case-study evidence is discussed in-depth by Brezis and Krugman (1997). Coulson (1999) and Carlino, DeFina, and Sill (2001) provide more systematic evidence and show that changes in employment in US metropolitan areas are mostly caused by local shocks at the industry level. Furthermore, Beardsell and Henderson (1999), Henderson (1999), and Dumais, Ellison, and Glaeser (2002) all suggest that

[3]The existing evidence regarding a causality running from amenities to growth is mostly intra-metropolitan – see Kahn (2000) for an analysis of differences in county growth rates in greater Los Angeles. For evidence of population decline leading to better amenities, see Kahn (1999) on Rust-Belt cities in the US. Finally, see Kahn (2001) for evidence of population growth leading to worse amenities.

the location of production in many industries changes significantly over time. Furthermore, the changes in the spatial patterns of production in the US are closely related to the processes of plant creation and destruction (Dumais *et al.*, 2002).

To model this argument and show how innovation-driven shocks within industries can explain existing patterns of city sizes, I embed Grossman and Helpman (1991*b*)'s quality-ladder model in an urban framework. In each industry, research firms try to invent the next step up the quality ladder in order to reap monopoly profits. Research firms may be successful in their own industry or may develop a new leading quality in another industry. Local spill-overs induce research firms in an industry to locate with production in the same industry and in most industries, successful innovators need to start producing where they did their research. This implies that own-industry innovations lead to a change of monopoly but to no change of location for the industry. Cross-industry innovations by contrast imply not only a change of monopoly but also typically a change of location since the old and new monopolies are not in general located in the same city.

Rochester (N.Y.) and New York provide two very interesting illustrations of this process (Jacobs, 1969). In the late 19th century, New York was the capital of the newly created photographic industry, whereas Rochester was at the forefront of technological developments in precision instruments. George Eastman, while working at improving optical instruments in Rochester, invented an emulsion-coating machine which enabled him to mass-produce photographic dry plates. His company soon took over the market for photographic films. As a consequence, Rochester replaced New York as the main centre for the industry. Rochester, 50 years later, was still the capital of the US film industry, whereas New York was that of the duplication industry. Then, Haloid Company, a firm specialised in the manufacturing of photographic papers and operating in the shadow of Eastman Kodak, came up with a new process for making copies without the need for developing. The process, called xerography, made Rochester the new capital of the duplication industry in place of New York (again) where the previous dominant technology, the varityper was produced.

In steady-state and in absence of agglomeration economies and urban crowding, this dynamic process generates city size distributions for which the Zipf's curve is concave. When calibrating the model on the US urban system, the performance is mixed. One the one hand, the simulations approximate the US distribution of city sizes better than Zipf's law. On the other hand, the simulated Zipf's curves are a bit too concave with cities at the top of the distribution being smaller than the likes of New York and Los Angeles and cities in the middle of the distribution being a bit too large.

This simple benchmark however is such that there are neither costs nor benefits to city size. To satisfy my third requirement, the model is extended to allow for urban congestion and economies of agglomeration. When agglomeration economies dominate crowding costs, the probability of innovating in a city increases more than proportionately to its size. This reduces the Zipf's exponent in the upper tail and increases it in the lower tail. Under realistic values for the trade-off between dynamic agglomeration economies and crowding costs, it is possible to generate urban systems very similar to that of the US. The French urban system can also be replicated. Hence adding realistic urban features to the model strengthens its empirical fit. The model can be extended even further and made consistent with city creation, new products development and plant relocations.

To summarise, the contributions of this paper are the following. First, it provides a simple mechanism able to replicate observed patterns of city size distributions. Second, this mechanism (innovation-driven shocks at the level of industries and cities) is arguably a strong candidate to explain the growth and decline of cities. Third, it shows that observed regularities about city size distribution are compatible with the basic building blocks of urban economics like the existence of agglomeration economies, crowding costs, etc. Importantly, these building blocks are crucial for a good empirical fit with the data. In this respect, the model can be viewed as a small step towards a unified theory of urban systems. In short, the theory presented here satisfies the three requirements set above. The last contribution of the paper is more subtle. Replicating existing patterns of city size distributions may not be as difficult as previously thought. This implies that the real test to distinguish between different mechanisms like the one highlighted here, that of Gabaix (1999*a*) or any potential alternative is not whether they can replicate observed patterns but rather their quantitative importance as sources of growth and decline for cities.[4]

The exposition proceeds in steps. Section 2 proposes a simple benchmark model. It is solved in Section 3. The model is then enriched in Section 4 where more realistic urban assumptions are introduced. Studying first a simple benchmark makes it easier to isolate the main mechanism of the model and present the main argument in a clear and simple fashion. The differences between the results of the benchmark model and those of the complete model with richer urban features also allow an easier assessment of the respective contributions of the different building blocks. Finally, the last Section contains some conclusions.

2. The benchmark model

The benchmark builds on Grossman and Helpman (1991*b*) (or Grossman and Helpman, 1991*a*, Chapter 4). Consider an economy with a large (discrete) number of industries, n, each of which produces one good which can potentially be supplied in an infinite number of qualities.[5] Quality j of good z is given by $q_j(z) = \delta^j$ with $\delta > 1$. At time $t = 0$, the quality of all goods is normalised to unity so that any good must be improved j times to reach quality j. Quality improvements find their source in research investments, which are described below.

Preferences

Consider a population of long-lived households whose mass is normalised to one and whose instantaneous utility is given by

$$u(t) \equiv \sum_{z=1}^{n} \frac{1}{n} \log \left[\sum_{j=1}^{\bar{j}(z,t)} q_j(z) d_j(z,t) \right] , \tag{1}$$

[4]Among the potential alternatives, Black and Henderson (1999) develop a model where human capital externalities lead to urban growth and a skewed distribution of city sizes. Their model generates only two classes of city sizes. Introducing some perturbations in this model could potentially lead to something more realistic.

[5]Note here a small difference with Grossman and Helpman (1991*b*). There is a discrete set of industries instead of a continuum. This prevents the law of large numbers from applying at the level of individual cities.

where $d_j(z,t)$ is the consumption of quality j of good z at time t and $\bar{j}(z,t)$ its highest available quality. For reasons made clear below, location indices can be ignored for the time being. Total expenditure is given by

$$E(t) \equiv \sum_{z=1}^{n} \sum_{j=1}^{\bar{j}(z,t)} p_j(z,t) d_j(z,t) \ , \tag{2}$$

where $p_j(z,t)$ is the price of quality j of good z at time t. The objective of consumers is to maximise the discounted sum of their future instantaneous utilities

$$U \equiv \int_0^{\infty} u(\tau) e^{-\rho\tau} \mathrm{d}\tau \ , \tag{3}$$

subject to the intertemporal budget constraint

$$\int_0^{\infty} E(\tau) e^{-R(\tau)} \mathrm{d}\tau \leq W(0) \ , \tag{4}$$

where $R(\tau)$ is the cumulative interest factor between 0 and τ and $W(0)$ is the net present value of the stream of income plus the initial asset holdings at $t = 0$.

The maximisation of the consumers' programme can be performed in two stages: First allocate instantaneous expenditure, $E(t)$, to maximise $u(t)$ and then choose the intertemporal allocation of expenditure. The maximisation of instantaneous utility (1) for any positive level of consumption expenditure implies equal shares of expenditure across industries. Then to solve for the allocation of expenditure within industries, define $J(z,t) \equiv \mathrm{Argmin}_{j \leq \bar{j}(z,t)} \left(p_j(z,t)/q_j(z) \right)$, the quality of good z for which the ratio of price to quality is the lowest. When $J(z,t)$ is unique (and it is so in equilibrium), demand in industry z is then given by

$$d_j(z,t) = \begin{cases} \frac{E(t)}{n p_j(z,t)} & \text{for } j = J(z,t), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Inserting these demands into (1) yields

$$u(t) = \frac{1}{n} \sum_{z=1}^{n} [\log E(t) - \log n + \log q_J(z,t) - \log p_J(z,t)] \ , \tag{6}$$

where $p_J(z,t)$ and $q_J(z,t)$ are the price and quality of $J(z,t)$, respectively. Equation (6) can now be used to solve the optimal consumption path whose solution is characterised by

$$\dot{E}/E = \dot{R} - \rho \ , \tag{7}$$

together with the budget constraint and a transversality condition. After normalising total expenditure $E(t)$ to n through the choice of numéraire, equation (7) implies $\dot{R} = \rho$, that is the nominal interest rate is always equal to the subjective discount rate.

Technology

As is standard in quality-ladder models of growth, research firms in each industry compete to innovate and occupy the next step up the quality-ladder. Any successful innovator is rewarded

with a patent giving it a monopoly right over the production of this quality, $\bar{j}(z,.)$. This patent cannot be licensed and it expires only when yet another successful innovator manages to develop the following quality step. For qualities below $\bar{j}(z,t)$, there is free-entry among price-setting oligopolists. In all industries and irrespective of quality, producers need one unit of labour to produce one unit of good.

Free-entry and unitary marginal costs imply that for any non-leading quality, $j < \bar{j}(z,t)$, the price, $p_j(z,t)$, is equal to the wage rate, $w(t)$. Together with (5), this implies that any quality leader in industry z has a revenue $p_{\bar{j}}(z,t)d_{\bar{j}}(z,t) = \frac{E(t)}{n} = 1$ when $\frac{p_{\bar{j}}(z,t)}{q_{\bar{j}}(z,t)} \leq \frac{p_{\bar{j}-1}(z,t)}{q_{\bar{j}-1}(z,t)}$, that is when $p_{\bar{j}}(z,t) \leq \delta w(t)$. Prices above $\delta w(t)$ imply zero demand for the industry's leading quality. Hence, with unit elastic demand, any industry leader maximises its profits by selling its quality at the limit price $p = \delta w$. Since the assumptions about product development ensure that in every industry there is a unique quality leader, this leader is also the only active firm in the industry: $\bar{j}(z,t) = J(z,t)$. For all industry leaders, this limit pricing strategy implies a profit equal to

$$\pi = 1 - \frac{1}{\delta}\,. \tag{8}$$

There is free-entry in the race to be the next leader in each industry. A research firm k in industry z, by investing $\lambda^k(z)$ units of research labour for a time interval of length dt to work on the highest existing quality, $\bar{j}(z,t)$, succeeds in inventing the next step up the quality-ladder in this industry, $\bar{j}(z,t)+1$, with probability $\beta\lambda^k dt$. Thus, as in Grossman and Helpman (1991*b*), research firms use the state-of-the-art technology $\bar{j}(z,t)$ in an industry as a base to invent the next step up the quality-ladder in the same industry.[6] There is however a slight difference with Grossman and Helpman (1991*b*)'s framework regarding the research technology. A research effort targeted at improving industry z may be successful, not only for this particular industry (as just described), but also for any other industry because of serendipity in the research process. Scherer (1984) provides very strong empirical support regarding the pervasiveness of such cross-industry innovations.[7] More formally, a research firm k in industry z, by investing $\lambda^k(z)$ units of research labour over dt succeeds in inventing the next step up the quality-ladder in industry $z' \neq z$ with probability $\gamma\lambda^k dt$ with $\gamma < \beta$.

In total, a research firm k in industry z, which invests $\lambda^k(z)$ over dt, expects to invent the next step up the quality-ladder in industry z with probability $\beta\lambda^k(z)dt$ and in any of the other industries with probability $\gamma\lambda^k(z)dt$. After denoting $\lambda(z) \equiv \sum_k \lambda^k(z)$, the sum of all research investments made by research firms in industry z, the probability of an innovation taking place in industry z

[6] As in most endogenous growth models, innovations have both a private good dimension (patenting) as well as a public good aspect (increase of own-industry stock of knowledge). Without any cost advantage in research, industry leaders do not attempt to innovate since, in case of success, the incremental profit would be less than that of a new entrant. Thus in equilibrium, research is performed only by would-be entrants. Furthermore note that the size of research firms is indeterminate because of constant returns in this activity. This need not be a concern because employment and not industry structure is the main variable of interest here.

[7] So far, after the choice of a discrete number of industries as opposed to a continuum, this is only the second difference with Grossman and Helpman (1991*b*). None of these differences would change their results in any meaningful way. In a spatial setting, however, the possibility of inter-industry innovations is crucial to allow a research firm located in a given city to "capture" an industry previously located in another city and thus provides a reason for city size to change.

over dt is $\iota(z)dt$ where

$$\iota(z) \equiv \beta\lambda(z) + \gamma \sum_{z' \neq z} \lambda(z') \ . \tag{9}$$

For the sake of clarity, the assumptions presented here stick as closely as possible to the canonical model of Schumpeterian growth developed by Grossman and Helpman (1991*b*): A multi-industry model where firms compete and invest in research in order to reap the monopoly profits associated with the highest quality. Self-sustaining and non-explosive growth is possible since new innovations are neither more difficult nor easier than past ones. This well-known model can now be embedded in the simplest possible urban setting. With firms located in different cities, the churning of monopolies will provide the basis for population changes in cities.

Cities

Consider m cities across which final goods are freely tradable with there being many more industries than cities: $n \gg m$. Workers are freely mobile and any city can accommodate any number of workers at zero cost so that there are neither advantage nor cost to city size. This last assumption, which is clearly counterfactual, is relaxed in Section 4.

Regarding research, recall that a quality improvement in an industry requires the use of the knowledge associated with the leading quality. In turn this knowledge is available only to research firms located in the same city (or cities) as the industry leader. In other words, the public good dimension of leading technologies is restricted to be local. One may learn about leading technologies only by observing how industry leaders produce, through small-talk with workers involved in production or by being involved indirectly in the production as supplier. All this requires physical proximity. This assumption of local knowledge spill-overs has received ample empirical support: See for instance Jaffe (1989), Jaffe, Trajtenberg, and Henderson (1993) or Feldman (1994).

Turning to the location of production, assume that industries can be of two types: *First-nature* or *second-nature*. First-nature industries are immobile and each city hosts one such industry. Any successful innovator in a first-nature industry, if located in a different city, must thus relocate at no cost to implement its innovation. One may think of some natural advantages like a primary resource that tie these industries to some particular cities. For instance any improvement in coal extraction can only be implemented close to coal fields. Note that first-nature industries are only a small fraction of all industries since $m \ll n$. They provide both a "first nature" justification for the existence of cities and a way by which to identify them.

The remaining $n - m$ industries are labelled second-nature in the sense that production must take place where the last quality innovation occurred. One possible justification for this assumption is that in this type of industries the production of the highest quality depends on many workers who took part in the innovation. Although they are individually freely mobile, coordinating the relocation of these workers to a particular city may be difficult.[8] In this respect, note that in many industries where quality innovations are rather complex such as the electronics industry, the

[8] Alternatively assume that state-of-the-art knowledge is too complex to be codified and exported to another city. This would also of course prevent patenting. To remain consistent with the existence of monopolies, it is however enough to make the imitation of the highest quality costly. See Dosi (1988) for further discussion on these issues.

highest quality products are nearly always manufactured close to the research centres where they were developed (see Fujita and Ishii, 1998, for evidence regarding Japanese firms in this industry).

In summary, in first-nature industries innovators follow production, whereas in second-nature industries production follows innovators. Second-nature industries relocate across cities following successful innovators and thus provide a reason for the growth or decline of cities. In contrast first-nature industries are "anchors" which prevent cities from losing all their industries and thus falling in a zero population trap.[9] Note finally that these assumptions about industry location together with local knowledge spill-overs also imply a unique location for each industry.

3. Steady-State growth and city size distribution

Before enriching the urban side of the model and allow for the existence of agglomeration economies, I solve the benchmark case presented above.

Steady-state economic growth

In absence of cost or benefit to city size, profits are independent of location. The model can thus be solved for research and growth independently of the urban structure. Denote v the stock-market value of an industry leader. If this stock market value is the same across industries (which is the case at the symmetric equilibrium), then firm k at a cost $w\lambda^k(z)dt$ can expect to win $(\beta + (n-1)\gamma) \times \lambda^k(z) \times vdt$. Profit maximisation by research firms implies that in equilibrium

$$v = \frac{w}{\beta + (n-1)\gamma} \,. \tag{10}$$

Turning to the stock-market valuation of firms, industry leaders pay a dividend πdt over the period dt since their profits are not re-invested in research. The value of an industry leader appreciates by $\dot{v}dt$ when no research firm succeeds in inventing the next step up the quality-ladder, whereas it goes to zero in the opposite case. This loss of v occurs with probability $\iota(z)$, the aggregate probability of any research firm being successful in industry z as defined in (9). Thus, with any industry leader, the expected rate of return for a shareholder is $\frac{\pi+\dot{v}}{v} - \iota(z)$. This return is risky but can be perfectly diversified since there is always one leader with constant profit in each industry: The loss of one firm is the gain of another. Thus firms are valued so that their expected stock-market return is equal to the safe interest rate, $\dot{R}$, which is itself equal to the subjective discount rate ρ. Consequently

$$\frac{\pi + \dot{v}}{v} - \iota(z) = \rho \,. \tag{11}$$

Equations (8), (10), and (11) imply the following no-arbitrage equation

$$\frac{\dot{w}}{w} + \left(1 - \frac{1}{\delta}\right) \frac{\beta + (n-1)\gamma}{w} = \iota(z) + \rho \,. \tag{12}$$

[9]This property can be generated in many other ways. For instance, qualitatively similar results are obtained with the assumption of costly industry relocation and a small group of immobile workers in every city who pay to attract an industry when their city becomes empty. In Section 4, this assumption is relaxed when stochastic industry relocations are allowed.

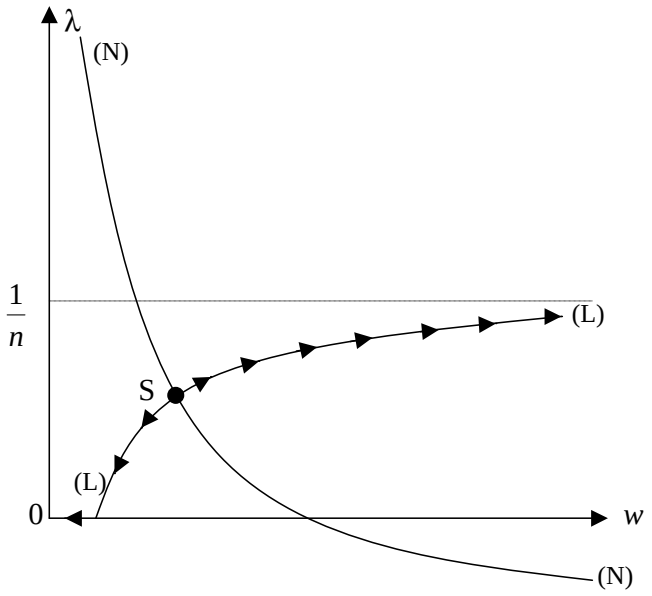

Figure 1. Determination of the equilibrium

In steady-state, equations (9)–(11) imply a symmetric research effort in all industries: $\lambda(z) = \lambda$. The aggregate probability of an innovation taking place in an industry becomes $\iota(z) = \iota = (\beta + (n-1)\gamma)\,\lambda$. Inserting this in equation (12) and re-arranging yield a differential equation guiding wages

$$\lambda = \frac{1}{\beta + (n-1)\gamma}\frac{\dot{w}}{w} + \left(1 - \frac{1}{\delta}\right)\frac{1}{w} - \frac{\rho}{\beta + (n-1)\gamma}\,. \tag{13}$$

This no-arbitrage condition is such that higher wages, which make research more expensive, have a negative effect on employment in research. By contrast, a higher efficiency of research (β or γ) raises returns to this activity and thus employment therein. An increase in the discount rate, ρ, is equivalent to an increase in interest rate. This lowers the net present value of future profits and the incentive to innovate. In turn, this implies less employment in research.

The model is closed by equating labour market demand and supply. Recall that each monopoly employs $\frac{1}{\delta w}$ units of labour and that research is the same across industries so that aggregate labour demand is equal to $n(\frac{1}{\delta w} + \lambda)$. Since aggregate labour supply is inelastic and equal to one, labour market clearing implies

$$\lambda = \frac{1}{n} - \frac{1}{\delta w}\,. \tag{14}$$

The evolution of the economy is depicted in Figure 1 where the (NN) locus is the no-arbitrage condition (13) and the (LL) locus is the labour market clearing condition (14). The economy must always lie along (LL) for the labour market to clear. For values of w below (NN), wages and research employment fall to zero. Since expenditure is constant, this implies that expected profits must rise above research costs, a contradiction with profit maximisation by research firms. A similar argument applies for values of w above (NN). The economy must thus always be in steady-state and jump at point S.

The steady-state values of w and λ solve equations (13) and (14) with $\dot{w} = 0$:

$$w = \frac{1}{\frac{1}{n} + \frac{\rho}{\beta+(n-1)\gamma}} , \tag{15}$$

$$\lambda = \frac{1}{n}\left(1 - \frac{1}{\delta}\right) - \frac{\rho}{(\beta + (n-1)\gamma)\delta} . \tag{16}$$

Note that to obtain positive values for the research employment in each industry, the subjective discount rate must be low enough: $\rho \leq (\delta - 1)\frac{\beta+(n-1)\gamma}{n}$. From equation (16), it is easy to compute the expected growth rate, g, of quality adjusted production in the economy over the period dt:

$$g = \left(1 - \frac{1}{\delta}\right)\left((\delta - 1)(\gamma + \frac{\beta - \gamma}{n}) - \rho\right) . \tag{17}$$

This expected growth rate is increasing with the size of the quality improvements, δ. There is a direct effect caused by larger quality improvements and an indirect effect whereby larger improvements imply larger profit and thus a stronger incentive to do research. The growth rate also increases with β and γ, the two efficiency parameters of research. It is also obviously decreasing with the rate of time preference, ρ. More interestingly, growth decreases with the number of industries. This is a dilution effect: With more industries, a research investment has a probability of success, β, to yield a monopoly over a smaller part of the economy, $\frac{1}{n}$, and a lower probability of success, γ, over a larger part of the economy, $\frac{n-1}{n}$. Note that because of the discrete number of industries, there may not be any quality improvement in any given period.[10] Furthermore the model also predicts an uneven and ever changing distribution of qualities, $\bar{j}(.,.)$ across industries.

As in Grossman and Helpman (1991*b*) or Aghion and Howitt (1992), the equilibrium growth rate may be above or below the optimal growth rate. This is the result of two conflicting effects. There is a *consumer-surplus effect* whereby quality innovations benefit to consumers who still pay the same price and enjoy higher quality goods. This effect depresses investment and thus maintains growth below its optimal level. However there is also a *business-stealing effect* whereby successful research firms displace existing incumbents. These attempts to capture monopoly rents imply over-investment. The sum of these two effects is ambiguous.

Steady-state city size distribution

Turning to city sizes, recall that each firm with a monopoly over the highest quality in an industry is the sole producer in this industry. Furthermore, because of local spill-overs, research is geographically tied to production. Consequently, symmetry across industries and free worker mobility imply that the population of a city is $\frac{1}{n}$ times its number of monopolies. Thus, the latter quantity is a sufficient statistic to describe a city. As a shorthand, the number of active industries in a city is referred to as its size. Denote $m_i(t)$ for $i = 1,...,n$, the number of cities with i industries at time t. The distribution of city sizes is in steady-state when the expected growth of $m_i(t)$ for all $i = 1,...,n$ is zero, that is when the following steady-state equation holds:

$$E[m_i(t + dt)] - m_i(t) = 0 . \tag{18}$$

[10]The law of large numbers does not apply here because time periods can be made arbitrarily small. This does not matter with respect to the intertemporal consumer program because of log-linear utility.

To characterise the steady-state, note first that in absence of cross-industry innovation between t and $t + dt$, the urban structure is left unchanged. Changes in the size distribution of cities happen only when a second-nature industry is improved by a research firm located in a different city.

Conditional on the occurrence of a cross-industry innovation, the number of cities of size 1, m_1, increases by one unit when a second-nature industry located in a city of size 2 is successfully improved by research in an industry located in another city of size 2 or above. In steady-state, all $n(n-1)$ possible cross-industry innovations occur with the same probability. Since, there are m_2 second-nature industries in cities of size 2 which can be each improved by one of $n - m_1 - 1$ other industries, the conditional probability of this event is thus:

$$\frac{m_2}{n} \times \frac{n - m_1 - 1}{n - 1} \, . \tag{19}$$

The number of cities of size 1, m_1, declines by one unit when a second-nature industry not located in a city of size 2 is successfully improved by research in a city of size 1. Since there are $n - m - m_2$ such industries which can be captured by m_1 industries in cities of size 1, this event occurs with the conditional probability:

$$\frac{n - m - m_2}{n} \times \frac{m_1}{n - 1} \, . \tag{20}$$

In steady-state, the probability of having one more city of size 1 must equal that of having one less city of the same size. Hence, for the expected change of m_1 to be zero, (19) and (20) must be equal. This implies:

$$m_2 = \frac{n - m}{n - 1} m_1 \, . \tag{21}$$

In Appendix A, this reasoning is generalised to cities of size $i \geq 2$. The steady-state equation (18) then becomes:

$$i(n - i)m_{i+1} = [(2i - 1)n - im - 2i(i - 1)]m_i - (i - 1)(n - m - i + 2)m_{i-1} \, . \tag{22}$$

Equations (21) and (22) characterise the *deterministic* steady-state of the benchmark. With the continual occurrence of innovations and because the number of cities is not arbitrarily large, the model cannot be expected to settle at the deterministic steady-state but rather to experience small-scale fluctuations around it as cities grow and decline. Before turning to these fluctuations, I first explore the main features of the deterministic steady-state.

First, in absence of agglomeration economies and congestion costs, this steady-state is determined only by the numbers of cities and industries. This can be observed directly from equations (21) and (22). The other parameters of the model (e.g., β, γ, and ρ) only affect the growth rate of output. Hence, they may affect convergence towards the steady-state city size distribution but not the steady-state itself.

Second, the size distribution of cities in steady-state is skewed with cities randomly experiencing small positive or negative shocks driving them up or down the urban hierarchy. More precisely, the steady-state Zipf's curve is concave with a Zipf's exponent below one in the lower tail and above one in the upper tail. To understand this, recall that any series s_i with a unitary Zipf's exponent is such that $\frac{s_{i+1}}{s_i} = \frac{i-1}{i+1}$. With $n \gg m$, straightforward calculations using (21) and

(22) show that $\frac{m_{i+1}}{m_i} \geq \frac{s_{i+1}}{s_i}$ for the first terms of the series. This implies a Zipf's exponent below one in the lower tail of the distribution. Then, a simple induction argument also using (21) and (22) shows that $m_i \leq m_1(\frac{n-m}{n-1})^{i-1}$. Indeed, from (21), this inequality weakly holds for $i = 2$. Then simple calculations using (22) show that if it holds for i, it is also satisfied for $i + 1$. Hence it holds for any $i \geq 2$. The steady-state is thus bounded from above by a decreasing geometric series. Such series decreases faster than any series following a Pareto distribution and has consequently a Zipf's exponent above one in the upper tail. Hence the steady-state Zipf's curve is concave with an exponent locally below unity for smaller cities and above unity for larger cities.[11]

The economic intuition for this result is the following. Recall first that growth processes whose means and variances are independent of size (i.e. Gibrat's law) are needed to generate Zipf's law, that is a straight Zipf's curve with slope -1. Then, note that conditional on a cross-industry innovation taking place, the probability of a city with i industries gaining an industry is equal to the probability of the innovation taking place there (which is proportional to city size and equal to $\frac{i}{n}$) multiplied by the probability of the improved industry being a second-nature industry originally located in another city. This second quantity (equal to $\frac{n-m-i+1}{n-1}$) decreases with size. Hence the probability of gaining an industry is less than proportional to city size due to a negative own-size effect. By the same token, the probability of loosing an industry (equal to $\frac{i-1}{n-1} \times \frac{n-i}{n}$) is less than proportional to city size. In total, the conditional expected growth of a city (equal to $\frac{n-im}{n(n-1)}$) decreases with city size and so does its expected growth rate and its variance. Hence, the steady-state distribution generated by the benchmark is less skewed than a Pareto distribution with exponent -1 and has thus a concave Zipf's curve. The main force at work here is this own-size effect which implies that larger cities have less to win than smaller cities.

The important issue at this stage is how good an approximation this concave Zipf's curve is with respect to real city size distributions. Given the fluctuations of the model around the steady-state, this evaluation may only be carried out by means of simulations.

Baseline simulations

The size distribution of US CMSAs is used as reference. Since there is a population threshold of around 100,000 for a city to qualify as CMSA, the existing 280 US CMSAs in 2000 must be viewed as a truncation of the real distribution. To circumvent this censoring problem, consider a population of 400 cities, each of which is initially endowed with 25 industries. The latter are assumed to employ 25,000 workers each so that the total urban population is set to 250 million to match (roughly) that of the US. Consider 1,000 independent sequences of cross-industry innovations occurring randomly as in the model. In the simulations, we speak of steady-state when the mean Zipf's exponent for 1,000 simulations does not increase nor decrease by more than 0.01 as more

[11]The steady-state distribution is not log-normal either as made clear by equations (21) and (22) which define it. However as will be clear from the simulations below, the steady state is rather well approximated by a log-normal distribution.

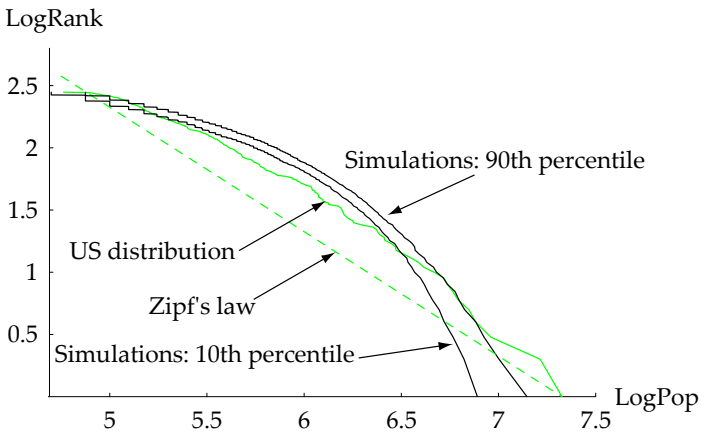

Figure 2. Baseline simulations and actual US city size distribution.
Source: US Census Bureau (2000 decennial census) and author's simulations. The US Zipf's curve is in grey. The two curves in black represent the 10^{th} and 90^{th} percentile for each rank from 1 to 280 in the simulations. 1,000 simulations of 10 million events were run. The grey dotted straight line represents the expected Zipf's curve for a Pareto distribution with exponent -1.

innovations are considered.[12]

The Zipf exponent for the 280 largest cities in each of the simulations can be computed and compared to that of the 280 US CMSAs. In steady-state, the mean Zipf's exponent is 0.74 with a standard deviation of 0.03 whereas the equivalent number for US cities in 2000 is 0.85. When looking only at the 150 biggest cities (as customary in the empirical literature), the average Zipf's exponent in the simulations is 1.10 with a standard deviation of 0.07. The corresponding figure for US CMSAs is within two standard deviations at 0.98.[13] Figure 2 plots the Zipf's curve for US CMSAs in 2000 together with the 10^{th} and 90^{th} percentile for every rank in the simulations.

The overall ability of the model to replicate the observed US patterns is mixed. On the one hand in Figure 2, the simulations predict a fairly skewed distribution of city sizes which, at first sight, looks in line with the real US distribution, which is also fairly concave (see Parr, 1976, and Vining,

[12]Such steady-state is reached after around 2 million events. This number may seem large. However, after 500,000 events, the Zipf's exponent for the 150 biggest cities is already around 1.25 and then slowly goes down to around 1.10 over the following 1.5 million events. This may well replicate the slow anticlockwise movement of the Zipf's curve observed in the US during the 20^{th} century (Parr, 1985; Black and Henderson, 1998). Furthermore, the next section argues that the development of new industries should lead to an "initial distribution" of industries across cities that is much more skewed than the initial uniform distribution assumed here for the simulations. This should make convergence much faster. Finally, it must be remembered that "churning" at the micro-level is very significant. The annual turnover rate for US plants is typically above 10% (Davis, Haltiwanger, and Schuh, 1996). At the same time, the annual US patent count is also high, typically above 600,000.

[13]Note further that the simulations are not completely neutral with respect to the number of industries n as expected from the analytic results above. Considering a greater number of industries reinforces the concavity of the Zipf's curve whereas considering fewer industries makes it straighter. An important issue here is thus how lumpy industries are. On the one hand, there are certainly more production plants in the US than the 10,000 assumed in the baseline simulations. On the other hand, remember that a central tenet of urban economics is the existence of localisation economies, i.e., local external economies within industries leading firms to cluster by industry (see Duranton and Puga, 2000, for references and a discussion of the empirical literature). Hence the use of rather big lumps (25,000 persons) could be viewed as a reduced-form for a more detailed model encompassing explicit localisation economies. Such localisation economies, although absent here, could easily be introduced by considering a more complex production structure with for instance input-output linkages with differentiated industry-specific intermediates as in Abdel-Rahman and Fujita (1990). In this respect, note that localisation economies have been shown to be no barrier to the mobility of industries. Quite the contrary, the most localised industries appear to be the most mobile in the data (Henderson, 1999).

1976, for early discussions of this). Figure 2 also plots a straight line with a negative unitary slope representing a strict version of Zipf's law which takes New-York as a reference. As a first-order approximation, the model does better than Zipf's law.[14] On the other hand, closer inspection of the simulated distributions and the real US Zipf's curve reveals sizeable differences. The Zipf's curves simulated from the model are more concave than what is observed in the US creating some visible differences for the largest three cities and for cities ranking between 20 and 100. For the largest city, the predicted size in the simulations is between 15 and 65% the size of New York, with a mean slightly below 50%. For cities ranking between 50 and 60 (i.e., where the concavity of the simulated distributions is at its strongest), their simulated size is between 20 and 40% bigger than in reality.

Note that beyond their size distribution, the cities in the model share some important features with real life cities. First, the model predicts a one-to-one correlation between the size of a city and its number of industries. The relationship between size and diversity appears very clearly in the data, albeit not as strongly as predicted by the model (see Mori and Nishikimi, 2001, for strong Japanese evidence and a discussion of the literature on this point). Second, technological change is a dominant factor at the root of the rise and decline of cities. Recent studies by Coulson (1999) and Carlino *et al.* (2001) on two samples of US CMSAs show that local industry shocks explain most of the changes in metropolitan employment. Both papers use a time-series methodology to identify local industry shocks. According to Coulson (1999), the latter explain between 67 and 97% of the difference between the 36-month ahead prediction and what really occurs. Using greater frequency data and a slightly different identification method, Carlino *et al.* (2001) find that local industry shocks explain between 87 and 94% of the same forecast error over nearly 50 years. Third, Beardsell and Henderson (1999), Henderson (1999), and Dumais *et al.* (2002) all document a substantial degree of mobility of industries across US CMSAs. Finally, cities move slowly up and down the distribution as observed in the US case (Black and Henderson, 1998).

4. Adding urban features to the benchmark

The benchmark model, whose working has just been described, builds on a potentially highly relevant economic mechanism. Hence it satisfies the second requirement imposed in the Introduction. With respect to the first requirement, the ability of this benchmark to replicate the US distribution of city sizes is not perfect but nonetheless better than previous approximations building on Zipf's law. However, cities so far have been modelled in a highly simplistic manner which goes against the basic tenets of urban economics. In this respect, the objective of this section is first to demonstrate that the model can take into account fundamental features of cities like the existence of agglomeration economies and crowding costs, which have been so far ignored. The model is also extended to take into account industry relocations, the creation of new cities, and the development of new industries. Hence the benchmark model can be extended to satisfy the third requirement imposed in the Introduction regarding the consistency of the urban modelling. Interestingly, these

[14] A simple efficiency criterion is to integrate on a log plot (like that of Figure 2) the absolute value of the difference between the real US distribution and any other distribution chosen to approximate it. According to this metric, the baseline simulations do on average better than Zipf's law.

extensions also show that these more realistic urban features improve the empirical performance of the model so as to closely replicate existing urban systems.

Agglomeration economies and urban crowding

The economic literature on systems of cities and more generally urban economics typically views any city as the outcome of a trade-off between some form of economies of agglomeration and crowding costs. Crowding costs are easy to introduce. For simplicity, think of them as a labour-reducing factor. A worker in a city of population ℓ has its labour supply divided by a factor $C(\ell)$. These crowding costs are such that $C' > 0$ with $C(1)$ being normalised to unity.

Turning to agglomeration economies, they may be either static or dynamic and take place either within cities but across industries (urbanisation economies) or within cities and industries (localisation economies). Static urbanisation economies can be thought as a labour-augmenting factor in production $A(\ell)$ which increases with city population ℓ and is normalised to unity in cities of unit population.[15] As for static localisation economies, because of symmetry they apply in the same fashion to all industries within a city and thus can be confounded with static urbanisation economies.

With respect to the benchmark proposed above, these new features imply that unit labour costs are no longer uniform across cities. They depend instead on city population ℓ. The labour supply (in efficiency units) of a worker in a city of population ℓ is $A(\ell)/C(\ell)$ times that of a worker in a city of unitary population. Given the utility function (1) and free mobility, equilibrium labour income must be equalised across cities.[16] Hence unit labour costs are equal to $w\frac{C(\ell)}{A(\ell)}$ where w is the unit labour cost in a city of unit population. If for simplicity agglomeration economies are assumed to apply only to the leading technologies (i.e., non-leading technologies can be implemented with equal efficiency everywhere), the profit function (8) of an industry leader in a city of population ℓ becomes:

$$\pi(\ell) = 1 - \frac{C(\ell)}{\delta A(\ell)} \,. \tag{23}$$

In this equation, the effect of city size on profits is twofold. First, increasing city population leads to more crowding and in turn to higher costs for firms as they must compensate workers for this extra crowding. At the same time, urbanisation economies increase the efficiency of labour, which means a lower wage bill for firms. The overall effect of an increase in city population on profits is ambiguous. Note that $C(\ell) \leqslant \delta A(\ell)$ is needed for an industry leader to be willing to produce in a city.

As for dynamic urbanisation economies, the easiest and most natural way to represent them is through another labour-augmenting factor $B(\ell)$ but this time affecting research labour. More formally, a research firm k located in a city of population ℓ and hiring $\lambda^k(z)$ workers in industry z

[15] Crowding costs and agglomeration economies are directly assumed here rather than derived within the model from microeconomic foundations. The specific functional forms used below however are similar to the standard reduced form obtained from microeconomic models of urban increasing returns (see Duranton and Puga, 2003, for a review of this literature).

[16] This labour income is net of commuting and housing costs as these are already captured by $C(.)$. In absence of migration costs for households, the latter maximise instantaneous utility which in absence of trading costs imply the maximisation of their labour income $wA(\ell)/C(\ell)$. In equilibrium this quantity must be equalised across cities.

over dt expects to innovate in the same industry with probability $B(\ell) \times \beta\lambda^k(z)dt$ and with $B(\ell) \times \gamma\lambda^k(z)dt$ in any other industry with $B' > 0$ and $B(1) = 1$. Again, because of symmetry, dynamic localisation economies apply equally to all industries in the same city. Consequently, they can be confounded with dynamic urbanisation economies.

Profit maximisation by research firms in any second-nature industry z located in city $a(z)$ of population $\ell_{a(z)}$ implies that in equilibrium the value of a firm is no longer given by equation (10) but instead by

$$[\beta + \gamma(n - m - 1)]v(\ell_{a(z)}) + \gamma \sum_{z' \neq z} v(\ell_{a(z')}) = w\frac{C(\ell_{a(z)})}{B(\ell_{a(z)})} \,. \tag{24}$$

Again the effects of city size are ambiguous as they make innovations more likely through the dynamic urbanisation economies $B(.)$ but also more costly since research labour must be compensated for crowding costs in larger cities relative to smaller cities. Either may dominate. For all cities to host some research labour, no city should be such that the probability of an own-industry innovation per unit of numéraire spent is below the probability of a cross-industry innovation per unit of numéraire in another city. Formally this implies that there should be no two cities a and a' such that $\beta\frac{B(\ell_a)}{C(\ell_a)} < \gamma\frac{B(\ell_{a'})}{C(\ell_{a'})}$.

The model can then be solved as previously using a no-arbitrage condition similar to (12) and labour market clearing.[17] The main source of difficulty is that closed forms solutions are impossible to obtain even under very simple specifications for $A(.)$, $B(.)$, and $C(.)$. This is because research in an industry is no longer independent of its location and instead depends non-linearly on the whole distribution of city sizes as made clear by equation (24). However from equations (23) and (24), it is easy to see that the effects of crowding costs are to lower profits and to raise the costs of research in larger cities. Both effects depress research employment in larger cities and make smaller cities more likely innovate relative to their population. Static urbanisation economies, through (23), increase profits in larger cities and thus the incentive to innovate there. Dynamic urbanisation economies, through (24), lower the costs of doing research in larger cities. Both effects contribute to making larger cities more likely to innovate relative to their size. Hence *when agglomeration economies dominate, the distribution of city sizes is expected to become more skewed than that generated by the benchmark above, whereas it should be less skewed when crowding costs dominate*. This statement can be verified using simulations.

Simulation results

For simplicity (and to restrict the number of degrees of freedom in the simulations), crowding costs are assumed to offset static agglomeration economies, $A(.) = C(.)$, but not dynamic extern-

[17] Although a reduced-form is used for tractability, the model in its urban features is very close to Henderson (1974) with a trade-off between crowding costs and economies of agglomeration in a system of cities. The main difference is that here industries are not perfectly mobile. Production can relocate only following cross-industry and cross-city innovations. The issue of non-replicability in a static framework is discussed in-depth by Papageorgiou and Pines (2000).

alities.[18] Then, and also for simplicity, I only consider a reduced-form whereby the probability of an innovation taking place in a city with i industries in steady-state is proportional to $i \times \psi(i)$, where $\psi(i)$ is the innovativeness of a city relative to its size. This reduced-form encapsulates both dynamic agglomeration economies and crowding costs. As argued above, crowding costs alone imply $\psi(i)' < 0$ whereas dynamic agglomeration economies alone imply $\psi(i)' > 0$. The advantage of this reduced-form is that for calibration purpose equilibrium innovativeness ($\psi(i)$) can be proxied by some measures of innovative output at the city level like patent data. Data on patent applications from the US Patent and Trademark Office support the idea of a non-linear relationship between the number of patents per capita granted in cities (a proxy for $\psi(.)$) and their population. The number of patents per capita increases with population but seems to peak at around one million.[19]

To understand the mechanics of this extension, it is worth simulating first the same urban system as previously with only dynamic agglomeration economies and no crowding cost so that the probability of a city innovating is more than proportional to its size. Assume $\psi(i) = i^{\epsilon}$ with $\epsilon > 0$. In this case, the probability of a city innovating is proportional to $i^{1+\epsilon}$. The parameter ϵ captures dynamic economies of scale: the probability of innovating within a given industry increases by $1 + \epsilon$% when the size of its city increases by 1%.

The results in this case are the following. Increasing ϵ lowers the Zipf's exponent in the upper tail of the distribution. Stated differently, stronger agglomeration economies increase the skewness of the upper tail of the distribution. This is rather intuitive since the expected growth of cities as a function of their size now depends not only on the negative own-size effect described above but also on the positive effect of dynamic agglomeration economies. Higher expected growth for the largest cities makes them even larger and it comes at the expense of smaller cities. Those which suffer the most are not the smallest cities with mostly immobile first-nature industries but cities in the middle of distribution with many second-nature industries to lose. Hence, with dynamic agglomeration economies, the concavity of the Zipf's curve is reduced with respect to the benchmark case or even reversed.

More specifically, for $\epsilon = 0.02$, the Zipf's exponent is equal to 1.01 for the 150 largest cities and 0.89 for the largest 280. The corresponding numbers for US CMSAs in 2000 are respectively 0.98 and 0.85. For $\epsilon = 0.04$, the Zipf's curve is now convex rather than concave. The Zipf's exponents are equal to 0.96 for the 150 largest cities and 1.04 for the largest 280. Interestingly, the Zipf's exponents

[18] With crowding costs and static agglomeration economies, city population is not in general proportional to the number of industries. Hence, strictly speaking, only the distribution of the number of industries across cities is simulated. Nonetheless, with static agglomeration economies offsetting congestion costs, employment in final production is the same in all industries. Only research labour may differ. However research labour represents only a tiny fraction of total employment in most industries. Hence, the proportionality between number of industries and population is approximately preserved. The second effect of static agglomeration economies is on profits, which turn affect the dynamic incentives to innovate. These dynamic incentives can also by captured by the dynamic agglomeration economies, $B(.)$. Hence $A(.) = C(.)$ is only a simplifying assumption which pins down employment in industries.

[19] When regressing the log of the average number of patents per capita between 1990 and 1999 for US CMSAs on their log population and its square, log population has a positive and significant coefficient whereas the quadratic term has a negative and significant coefficient. Since the relationship between patenting per capita and population is quite flat for cities above 1 million, the maximum of the parabolic trend-line is very sensitive to the sample of cities one considers. The most innovative US city during the 1990s was Rochester (NY) with a population around 1 million. Among the largest cities, the most innovative is San Francisco with a population around 7 million.

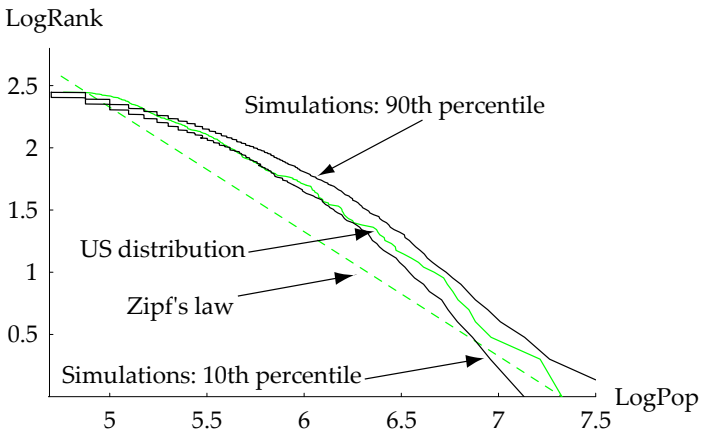

Figure 3. Augmented simulations and actual US city size distribution.
Source: US Census Bureau (2000 decennial census) and author's simulations. The US Zipf's curve is in grey. The two curves in black represent the 10^{th} and 90^{th} percentile for each rank from 1 to 280 in the simulations. 1,000 simulations of 4 million events were run with $\psi(i) = i^{0.04-0.035i/(i+1)}$. The grey dotted straight line represents the expected Zipf's curve for a Pareto distribution with exponent -1.

are not very sensitive to ϵ provided it remains below 0.06.This corresponds to net scale economies of around 6% which is in the upper bound of what this literature suggests (see Henderson, 1999, for a discussion of the empirical literature). The only counter-factual prediction of these simulations is that for ϵ even as low as 0.01, the largest city is already much larger than the second largest city by a factor of ten or more.[20] This highlights the need to consider crowding costs along with agglomeration economies.

I now consider some simulations with both crowding costs and dynamic agglomeration economies. The general form is $\psi(i) = i^{\epsilon(i)}$. Figure 3 plots the 10^{th} and 90^{th} percentile for every rank between 1 and 280 in the simulations for $\epsilon(i) = 0.04 - 0.035i/(i+1)$ together with the Zipf's curve for US CMSAs in 2000. The Figure also plots the Zipf's curve for an urban system following Zipf's law from New-York downwards. This specification assumes a constant degree of increasing returns of 4% in innovations. Crowding costs are rising with city size but they never fully offset dynamic agglomeration economies. The shape of the locus representing the number of innovations as a function of city size is very similar to that of patenting in US cities as a function of their size with this specification.

With these parameter values, the mean Zipf's exponent for the first 150 cities is equal to that of US CMSAs at 0.98. The bands constructed from the 10^{th} and 90^{th} percentile contain the US Zipf's curve almost entirely. It is only in the very lower tail of the distribution that these bands are hit due to their stepwise nature. The simulations capture the US city size distribution much better than Zipf's law (the dotted line). It is also interesting to note that the US urban system is widely acknowledged to be close to its steady-state (Black and Henderson, 1998; Ioannides and Overman,

[20]Such ratio is very large for the US but not uncommon in the world (Soo, 2002). The intuition for such primacy is quite simple. Consider the case of the two largest cities being initially of equal size. When ones gains a small advantage after a few lucky draws, it becomes relatively better at innovating and thus draws new industries. These new industries reinforce the strength of its innovation advantage, etc. This cumulative causation mechanism stops only when the growth of the primate city is limited by the own-size effect described above.

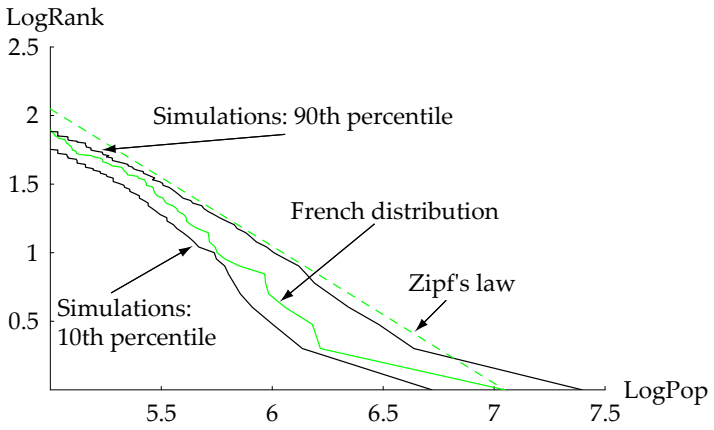

Figure 4. Augmented simulations and actual French city size distribution.
Source: INSEE (Urban areas data for 2000) and author's simulations. The French Zipf's curve is in grey. The two curves in black represent the 10^{th} and 90^{th} percentile for each rank from 1 to 80 in the simulations. 1,000 simulations of 4 million events were run with $\psi(i) = i^{0.037-0.03i/(i+1)}$. The grey dotted straight line represents the expected Zipf's curve for a Pareto distribution with exponent -1.

2002). Consequently if the model is able to replicate the US steady-state, it must also replicate the US dynamic of relative urban growth.

Turning to the robustness of these simulations, a few comments are in order. First, the simulations do not show any knife-edge property. Very similar results are obtained when changing the negative term which captures congestion to 0.036 or 0.034 instead of 0.035. It is also possible to replicate well the US Zipf's curve with either stronger or weaker agglomeration economies. For instance, very good results are also obtained with $\epsilon(i) = 0.02 - 0.016i/(i+1)$ or $\epsilon(i) = 0.06 - 0.055i/(i+1)$ (implying dynamic agglomeration economies of 2 and 6%). To replicate the US city size distribution, the crucial element is to have $\epsilon(i)$ being a decreasing and convex function which takes a small positive values (around 0.5%) for cities with population between 5 and 25 million.

Second, it is worth turning to another empirical distribution. France was chosen because the concept of French urban area matches rather closely that of US CMSA. At the same time, the French distribution is fairly different from that of the US. Firstly, cities are fewer: there are only 80 French urban areas with a population above 100,000. Secondly, the French Zipf's coefficient is much higher than in the US at 1.06. Thirdly the French urban system has a dominant city, Paris, about seven times as large as the second largest city, Lyons. As a consequence of these last two features, the Zipf's curve is convex rather than concave in its upper tail. The simulations are the same before except that I consider a population of 200 cities with initially 20 industries each. For consistency with the total French urban population, each industry is assumed to employ 12,000 workers. Figure 4 plots the 10^{th} and 90^{th} percentile for every rank between 1 and 80 in the simulations for $\epsilon(i) = 0.037 - 0.03i/(i+1)$ together with the Zipf's curve for the French urban areas in 2000. Compared to the simulations for the US above, urban increasing returns are weaker.

So is urban congestion.[21] The Figure also plots the Zipf's curve for an urban system following Zipf's law from Paris downwards. As made clear by the figure, the simulations do better than Zipf's law and can successfully replicate the French distribution.

Third, it is also possible to reproduce very closely an imaginary urban system following Zipf's law. For instance with the simulated "US" system, it suffices to make $\epsilon(i)$ more convex than in the simulations reported above. When $\epsilon(i)$ is more convex, cities in the middle of the distribution tend to be relatively less innovative than these in the upper tail. Hence they must be smaller in steady state and this makes the Zipf's curve straighter in the middle of the distribution.

Firm relocations

The results above show that it is possible to relax the assumption of no cost nor benefit to city size made in the benchmark model. In this sub-section, the relaxation of another restrictive set of assumptions is discussed. Recall that in the benchmark new innovations should either immediately locate to a particular city (first-nature industry) or remain permanently in the city where the innovation was produced (second-nature industry). It appears more realistic to assume instead that all industries are such that their leader may relocate at times.

The empirical literature suggests that the real benefits of large cities may not lie in lower production costs but in a higher propensity to innovate. In the model, this implies that congestion costs may dominate static agglomeration economies after city size reaches a fairly low population threshold, whereas dynamic agglomeration economies possibly always dominate congestion or at least dominate them until a much higher population threshold is reached. With respect to the notations, this implies $C(\ell) > A(\ell)$ for $\ell > \underline{\ell}$ and $B(\ell) > C(\ell)$ for $\overline{\ell} > \ell$ with $\overline{\ell} > \underline{\ell}$. Under these realistic conditions, successful innovators in large cities would like to relocate their production to a smaller city where the costs of production are lower.

Such relocation may be costly and following the argument articulated by Duranton and Puga (2001), it might be feasible only after a successful search for the best way to produce has been conducted.[22] In larger cities, searching may be easier and the intensity of the search is also greater since the incentive to relocate is stronger. Consequently the probability of an industry leader being able to relocate to a smaller city at any point in time should increase with city size. As shown by the data in Duranton and Puga (2001), the amount of plant relocations is not negligible and relocating plants usually go from large diversified cities to smaller and more specialised cities better suited to their needs.

[21] Given that it is possible to replicate the French and US distribution with different functional forms, it is best to be cautious about drawing conclusions regarding the relative values of agglomeration economies and congestion costs in France and the US.

[22] In Duranton and Puga (2001), firms search for their ideal production process by sampling successively different types of production processes which in turn require different sets on intermediate inputs. After every trial, the firm learns if the process it just tried was its ideal production process or not. Since firms relocate only after hitting their ideal production process, this implies a strong stochastic element in the relocation process. This learning process is fully consistent with the assumptions of the benchmark model provided that any product innovation (i.e., a better quality) must be followed by a process innovation (i.e., a stochastic search) with no relocation being feasible before the latter takes place.

In the benchmark, first-nature industries were needed to prevent cities from becoming empty and remaining so forever. Stochastic relocations with a bias from large cities to small cities also prevent this since empty cities end up receiving relocating industries. Hence this relocation mechanism makes it possible to relax the assumptions about industries being either first- or second-nature and be more realistic in this respect. The other advantage of relocations (as assumed here) regards the balance between agglomeration economies and crowding costs. As the simulations above make clear, it is only when the dynamic agglomeration economies net of crowding costs are small that realistic distributions of city sizes can be generated. If dynamic agglomeration economies substantially outweigh crowding costs, the largest city becomes unrealistically large. Relocations from large to small cities may allow for more flexibility in this respect since firms may be leaving the largest cities when they get too large and thus prevent too much primacy from occurring. However this extra realism in the assumptions and this greater flexibility with respect to parameter values would necessitate a much more complex model. A fully-fledged model in this direction is beyond the scope of this paper and is left for future work.

New industries, new cities and uneven initial conditions

For simplicity, the benchmark model also took the number of cities and industries as given. However new products and industries keep being developed. Dobkins and Ioannides (2000) also document significant entry of new cities in the US urban system over the 20th century. As shown here, it is possible to relax this exogeneity of cities and industries but doing so requires making significant changes to the basic framework. The model outlined in what follows no longer builds on the quality-ladder framework but instead on the standard horizontal proliferation framework developed by Romer (1990) and Grossman and Helpman (1991*a*, Chapter 3). It receives a full treatment in Appendix B.

First, the Cobb-Douglas instantaneous utility function (1) must be transformed into a CES to be consistent with the arrival of new industries since with Cobb-Douglas utility, there is no demand for new goods. For simplicity only one quality level is considered for each good. These changes lead to a different pricing strategy by monopolies (mark-up instead of limit pricing) but leave the dynamic optimisation of consumers unchanged.

To allow for new industries, the research process must also be amended slightly. As in the benchmark, there is a monopoly protected by a patent in each industry. Competitive research firms can freely use each existing patent as a line of research. But, this time, it is with the goal of developing new industries rather than improving existing ones. Individually, these research firms face constant returns. However, there are decreasing returns to research in each line of research. Such decreasing returns in research are easily justified by a duplication argument. When there are neither costs nor benefits to city size, this assumption implies a symmetric research investment across lines of research (i.e., industries) in equilibrium. Finally, all existing patents form a stock of knowledge which also enters as a factor of production for new patents. This is a standard assumption in the literature, which allows for self-sustaining growth to take place.[23]

[23]Recall that self-sustaining growth requires the number of new patents to be proportional to the existing number of patents.

As in the benchmark, research is geographically tied to production in each industry because of local spill-overs. New industries can be either first- or second-nature. In first-nature industries, each patent holder must go to a specific location. Only at this location can the patent be exploited. Assuming that first-nature locations are always different, each new patent in a first-nature industry leads to the creation of a new city. In second-nature industries, each patent can only be implemented by its innovator where it was developed.

Symmetry in the utility function and decreasing returns in each line of research imply equal levels of output and profit across all industries. Then the research process implies an equal level of research employment across industries. Hence in equilibrium all industries are symmetric with respect to output, profit and research employment. This implies that the probability of a new second-nature industry being developed in any city is proportional to its size. Consequently, whenever an innovation takes place, it leads to the creation of a new city with some probability α (the probability of a first-nature industry being created) or it is added to an existing city with the probability that any particular city gets it is proportional to its population.

With respect to the number of patents, this model is thus equivalent to Simon (1955) and the preceding assumptions can be viewed as microeconomic foundations for it. Then it is immediate that the distribution of the stock of patents across cities follows a power law with exponent $1/(1-\alpha)$ where α is the probability of any new industry being first-nature. Again see the Appendix for a complete derivation. Because of symmetry, the population in a city is proportional to its number of patents. Consequently, the size distribution of cities also follows a power law with exponent $1/(1-\alpha)$. Hence, this extension of the model can account for the creation of both new cities and new industries.[24]

5. Concluding comments

This paper started from the principle that any good theory of city size distribution should satisfy three requirements: (1) ability to replicate observed patterns, (2) reliance on a plausible economic argument, and (3) consistency with the idea that cities result from a trade-off between some benefits from agglomeration and crowding costs. These three requirement can be satisfied by embedding the Grossman and Helpman (1991*a*)'s quality-ladder model of growth in an urban framework. Interestingly, the paper also shows that mimicking existing city size distributions is not very difficult. Acknowledging Gabaix (1999*a*)'s results, it is my contention that several economic mechanisms could satisfy this first requirement. As a consequence, the empirical challenge is no longer to focus on the exact shape of the distribution of city sizes but instead to evaluate what the real drivers of urban growth and decline are.

[24]However, product proliferation cannot be considered alone to explain and replicate existing city size distributions. The reason is twofold. First, this process cannot generate Zipf's exponents below unity. Second, to generate Zipf's exponent close to unity, an arbitrarily large number of new industries is needed. Nonetheless, this extension can play an important accessory part in the main argument. Indeed, if one thinks of the creation of industries as taking place before quality improvement are made possible, this extension should be viewed as a way to generate skewed "initial conditions" in the quality-ladder model developed earlier.

References

Abdel-Rahman, Hesham M. and Masahisa Fujita. 1990. Product variety, Marshallian externalities, and city sizes. *Journal of Regional Science* 30(2):165–183.

Aghion, Philippe and Peter Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60(2):323–351.

Arnold, Berry C. 1983. *Pareto Distributions*. Fairland, MD: International Cooperative Publishing House.

Auerbach, F. 1913. Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen* 59:73–76.

Beardsell, Mark and J. Vernon Henderson. 1999. Spatial evolution of the computer industry in the USA. *European Economic Review* 43(2):431–456.

Black, Duncan and J. Vernon Henderson. 1998. Urban evolution in the US. Working Paper 98–21, Brown University.

Black, Duncan and J. Vernon Henderson. 1999. A theory of urban growth. *Journal of Political Economy* 107(2):252–284.

Brezis, Elise S. and Paul R. Krugman. 1997. Technology and the life cycle of cities. *Journal of Economic Growth* 2(4):369–383.

Carlino, Gerald A., Robert H. DeFina, and Keith Sill. 2001. Sectoral shocks and metropolitan employment growth. *Journal of Urban Economics* 50(3):396–417.

Coulson, Edward N. 1999. Sectoral sources of metropolitan growth. *Regional Science and Urban Economics* 29(6):723–43.

Córdoba, Juan-Carlos. 2001. Zipf's law: a case against scale economies? Processed, University of Rochester.

Davis, Steven J., John C. Haltiwanger, and Scott Schuh. 1996. *Job Creation and Destruction*. Cambridge (Mass.): MIT Press.

Dobkins, Linda Harris and Yannis M. Ioannides. 2000. Dynamic evolution of the size distribution of US cities. In Jean-Marie Huriot and Jacques-François Thisse (eds.) *Economics of Cities: Theoretical Perspectives*. Cambridge: Cambridge University Press, 217–260.

Dosi, Giovanni. 1988. Sources, procedures and microeconomic effects of innovation. *Journal of Economic Literature* 26(3):1120–1171.

Dumais, Guy, Glenn Ellison, and Edward L. Glaeser. 2002. Geographic concentration as a dynamic process. *Review of Economics and Statistics* 84(2):193–204.

Duranton, Gilles and Diego Puga. 2000. Diversity and specialisation in cities: Why, where and when does it matter? *Urban Studies* 37(3):533–555.

Duranton, Gilles and Diego Puga. 2001. Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review* 91(5):1454–1477.

Duranton, Gilles and Diego Puga. 2003. Micro-economic foundations of urban increasing returns. In J. Vernon Henderson and Jacques Thisse (eds.) *Handbook of Regional and Urban Economics vol. 4*. North-Holland: Elsevier Science. Forthcoming.

Feldman, Maryann P. 1994. *The geography of innovation*. Dordrecht and London: Kluwer Academic.

Fujita, Masahisa and Ryoichi Ishii. 1998. Global location behavior and organizational dynamics of Japanese electronics firms and their impact on regional economies. In Alfred D. Chandler Jr., Peter Hagström, and Örjan Sölvell (eds.) *The Dynamic Firm: The Role of Technology, Strategy, Organization and Regions*. Oxford: Oxford University Press, 343–383.

Gabaix, Xavier. 1999*a*. Zipf's law for cities: an explanation. *Quarterly Journal of Economics* 114(3):739–767.

Gabaix, Xavier. 1999*b*. Zipf's law and the growth of cities. *American Economic Review (Papers and Proceedings)* 89(2):129–132.

Greenwood, Michael J. 1997. Internal migrations in developed countries. In Mark R. Rosenzweig and Oded Stark (eds.) *Handbook of Population and Family Economics vol. IB*. North-Holland: Elsevier Science, 647–720.

Grossman, Gene and Elhanan Helpman. 1991*a*. *Innovation and Growth in the World Economy*. Cambridge, MA: MIT Press.

Grossman, Gene M. and Elhanan Helpman. 1991*b*. Quality ladders in the theory of growth. *Review of Economic Studies* 58(1):43–61.

Gyourko, Joseph, Matthew E. Khan, and Joseph Tracy. 1999. Quality of life and environmental comparisons. In Edwin S. Mills and Paul Cheshire (eds.) *Handbook of Regional and Urban Economics vol. III*. North-Holland: Elsevier Science, 1413–1454.

Henderson, J. Vernon. 1974. The sizes and types of cities. *American Economic Review* 64(4):640–656.

Henderson, J. Vernon. 1999. Marshall's economies. Working Paper 7358, National Bureau of Economic Research.

Ioannides, Yannis M. and Henry G. Overman. 2002. Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics* (forthcoming).

Jacobs, Jane. 1969. *The Economy of Cities*. New York: Random House.

Jaffe, Adam B. 1989. Real effects of academic research. *American Economic Review* 79(5):957–970.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108(3):577–598.

Kahn, Matthew E. 1999. The silver lining of Rust Belt manufacturing decline. *Journal of Urban Economics* 46(3):360–376.

Kahn, Matthew E. 2000. Smog reduction's impact on California county growth. *Journal of Regional Science* 40(3):565–582.

Kahn, Matthew E. 2001. City quality-of-life dynamics: measuring the costs of growth. *Journal of Real Estate Finance and Economics* 22(2/3):339–352.

Krugman, Paul. 1996. Confronting the mystery of urban hierarchy. *Journal of the Japanese and International Economies* 10(4):1120–1171.

Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan.

Mori, Tomoya and Koji Nishikimi. 2001. Self-organisation in the spatial economy: Size, location and specialization of cities. Processed, Kyoto University.

Papageorgiou, Yorgos Y. and David Pines. 2000. Externalities, indivisibility, nonreplicability, and agglomeration. *Journal of Urban Economics* 48(3):509–535.

Parr, John B. 1976. A class of deviations from rank-size regularity: Three interpretations. *Regional Studies* 10(3):285–292.

Parr, John B. 1985. A note on the size distribution of cities over time. *Journal of Urban Economics* 18(2):199–212.

Romer, Paul M. 1990. Endogenous technical change. *Journal of Political Economy* 98(5(2)):S71–S102.

Rosen, Kenneth and Mitchell Resnick. 1980. The size distribution of cities: An examination of the pareto law and primacy. *Journal of Urban Economics* 8(2):165–186.

Scherer, Frederic M. 1984. Using linked patent and R&D data to measure interindustry technology flows. In Zvi Griliches (ed.) *R&D, Patents, and Productivity*. Chicago: University of Chicago Press.

Simon, Herbert. 1955. On a class of skew distribution functions. *Biometrika* 42(2):425–440.

Soo, Kwok Tong. 2002. Zipf's law for cities: A cross country investigation. Processed, London School of Economics.

Sutton, John. 1998. *Technology and market structure: theory and history*. Cambridge, Mass.: MIT Press.

Vining, Daniel R. 1976. Autocorrelated growth rates and the pareto law: A further analysis. *Journal of Political Economy* 82(2):369–380.

Zipf, George K. 1949. *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, Mass.: Addison Wesley.

Appendix A. Steady-state city size distribution in the benchmark

The number m_i of cities of size $i \geq 2$ grows by two units when a second-nature industry located in a city of size $i+1$ is successfully improved by a research firm located in a city of size $i-1$. Since there are $n(n-1)$ possible cross-industry innovations, im_{i+1} second-nature industries in cities of size $i+1$ and $(i-1)m_{i-1}$ potential sources of innovations in cities of size $i-1$, this event occurs with the probability:

$$\frac{im_{i+1}}{n} \times \frac{(i-1)m_{i-1}}{n-1} , \tag{A 1}$$

conditional on a cross-industry innovation taking place. Then, m_i grows by one unit when an industry relocates from a city of size $i+1$ to a (different) city of any size except i and $i-1$. Since there are im_{i+1} industries which can be captured this way by $n-i-im_i-(i-1)m_{i-1}$ industries, this event takes place with the conditional probability:

$$\frac{im_{i+1}}{n} \times \frac{n-i-im_i-(i-1)m_{i-1}}{n-1} . \tag{A 2}$$

Finally, m_i also increases by one unit when research in a city of size $i-1$ successfully improves a second-nature industry located in a different city of any size but i and $i+1$. Since there are $(i-1)m_{i-1}$ industries in cities of size $i-1$ which can capture one of $n-m-(i-2)-(i-1)m_i-im_{i+1}$ industries, this event takes place with the conditional probability:

$$\frac{n-m-(i-2)-(i-1)m_i-im_{i+1}}{n} \times \frac{(i-1)m_{i-1}}{n-1} . \tag{A 3}$$

The number m_i of cities of size $i \geq 2$ decreases by two units when a second-nature industry located in a city of size i is successfully improved by a research firm located in another city of size i. Since there are im_i industries which can capture one of $(i-1)(m_i-1)$ industries, the conditional probability of this event is:

$$\frac{(i-1)(m_i-1)}{n} \times \frac{im_i}{n-1} . \tag{A 4}$$

Then, m_i declines by one unit when an industry relocates from a city of size i to a city of any size except i and $i-1$. With $(i-1)m_i$ industries which can be captured this way by $n-im_i-(i-1)m_{i-1}$, this event takes place with the conditional probability:

$$\frac{(i-1)m_i}{n} \times \frac{n-im_i-(i-1)m_{i-1}}{n-1} . \tag{A 5}$$

Finally, m_i also decreases by one unit when research in a city of size i successfully improves a second-nature industry located in a city of any size but i and $i+1$. Since there are im_i industries in cities of size i which can capture one of $n-m-(i-1)m_i-im_{i+1}$ industries, this event takes place with the conditional probability:

$$\frac{n-m-(i-1)m_i-im_{i+1}}{n} \times \frac{im_i}{n-1} . \tag{A 6}$$

Using (A 1)–(A 6), the steady-state condition (18) for $i \geq 2$ yields after simplification

$$i(n-i)m_{i+1} = [(2i-1)n-im-2i(i-1)]m_i-(i-1)(n-m-i+2)m_{i-1} . \tag{A 7}$$

Appendix B. Industry proliferation in an urban setting

The model in this Appendix builds on the standard endogenous growth framework with expending product variety developed by Romer (1990) and Grossman and Helpman (1991*a*, Chapter 3), whereas the model presented in the main text builds on Grossman and Helpman (1991*a*, Chapter 4). These two models are the two canonical models of modern growth theory.

Assumptions

Consider again a population normalised to one of long-lived households. Their instantaneous utility is now given by

$$u(t) \equiv \log \left(\sum_{z=1}^{n(t)} d(z,t)^{1-1/\sigma} \right)^{\frac{\sigma}{\sigma-1}} , \tag{B 1}$$

where $d(z,t)$ is again the consumption of the good from industry z at time t, $n(t)$ is the number of available goods, and and σ (> 1) is the elasticity of substitution between goods. In absence of quality differentiation, there is a one-to-one mapping between goods and industries. Total instantaneous consumption expenditure is given by $E(t) \equiv \sum_{z=1}^{n(t)} p(z,t)d(z,t)$ where $p(z,t)$ is the price of good z at time t. The objective of consumers is still to maximise the discounted sum of their future instantaneous utilities subject to an intertemporal budget constraint similar to (4). In comparison with the benchmark presented in Section 2, the main change so far is to have a CES instantaneous utility instead of a Cobb-Douglas. This change is necessary to accommodate the arrival of an unbounded number of new goods and industries since with Cobb-Douglas utility functions, there is positive demand only for a finite number of goods.

As previously, the consumer's maximisation problem can be solved in two stages: first allocate instantaneous expenditure, $E(t)$, across goods to maximise $u(t)$ and then choose the intertemporal allocation of expenditure. The maximisation of instantaneous utility (B 1) for any given level of expenditure implies the following instantaneous demand

$$d(z,t) = \frac{E(t)p(z,t)^{-\sigma}}{\sum_{z'=1}^{n} p(z',t)^{1-\sigma}} . \tag{B 2}$$

After defining the aggregate price index $P \equiv \left(\sum_{z=1}^{n} p(z)^{1-\sigma} \right)^{\frac{1}{1-\sigma}}$ and inserting (B 2) into (B 1), intertemporal utility becomes

$$U = \int_0^{\infty} [\log E(\tau) - \log P(\tau)] e^{-\rho\tau} \mathrm{d}\tau . \tag{B 3}$$

Equation (B 3) can now be used to solve the optimal consumption path whose solution is given again by equation (7) together with the budget constraint and a transversality condition. After normalising total expenditure $E(t)$ to unity through the choice of numéraire, the nominal interest rate is again equal to the discount rate, $\dot{R} = \rho$.

Regarding technology, there are still two activities: the manufacturing of goods in existing industries and the development of new industries, each of which is developed by a single successful innovator protected by an infinitely-lived patent. The number of industries is large so that each

monopoly is of negligible size and takes the aggregate price index P as given. Manufacturing one unit of a good still requires one unit of labour. Facing the demand function (B 2), the manufacturer of good z maximises its instantaneous profits by charging $p(z) = \frac{\sigma}{\sigma-1}w$, that is a constant mark-up over marginal cost. Since all manufacturers behave in the same way, the equilibrium aggregate price index is $P = \frac{\sigma}{\sigma-1}w$ and this pricing strategy implies instantaneous profits equal to

$$\pi(z) = \frac{1}{\sigma n} . \tag{B 4}$$

To allow for new goods/industries, the micro-economic foundations of the research process must be amended slightly. Each patent now intervenes in three different instances in the production process. First, it still serves as the basis to manufacture a given good for consumption. The rent associated with this can still be fully appropriated by its innovator. Second, each patent still constitutes a line of research which can be used by research firms to innovate. Third and unlike previously, each patent is also part of the general stock of knowledge and, as such, it is a pure public good. The general stock of knowledge enters as a factor in the development of new goods together with research labour. Such general stock of knowledge is necessary to make growth self-sustainable with an ever expending number of industries. Recall that self-sustaining growth requires the expected number of new patents to be proportional to the number of existing industries. After the change in the utility function, this constitutes the second difference with the benchmark in Section 2.[25]

Competitive research firms face constant returns to scale. However, in each line of research, there are aggregate decreasing returns. The justification is that, although every line of research has the same potential to generate new ideas regardless of how fruitful it has been in the past, an increase in research labour on a given line of research leads to some duplication. This duplication of the research effort is viewed as a negative congestion externality that is not internalised by research firms. More formally, any research firm k working on good z and investing $\lambda^k(z)$ units of research labour for a time interval of length dt succeeds in inventing a new good with probability $b(\lambda(z),n)\lambda^k(z)dt$ where $\lambda(z)$ is the total research labour working on z. Because of duplication, the individual hazard function $b(n,\lambda(z))$ decreases with $\lambda(z)$: $\frac{\partial b(n,\lambda(z))}{\partial \lambda(z)} < 0$. It also increases with the total stock of knowledge, which is measured by the number of goods n: $\frac{\partial b(n,\lambda(z))}{\partial n} > 0$. To allow for self-sustaining and non-explosive growth, the aggregate hazard function for any line of research z is assumed to take the following functional form:

$$B(\lambda(z),n) \equiv b(\lambda(z),n)\lambda(z) \equiv \beta(\lambda(z)n)^{1-\phi} , \tag{B 5}$$

where β is again the efficiency of the innovation process and $\phi \in (0,1)$ is the intensity of congestion in research. Finally aggregating across lines of research yields the instantaneous probability of an innovation taking place in the economy

$$\iota = \sum_{z=1}^{n} B(\lambda(z),n) . \tag{B 6}$$

[25]This is however a standard feature of growth models with expanding varieties. The main difference with respect to these latter models is that they do not need horizontal differentiation to play a role in product development. This assumption is important here since, together with local spill-overs, it pins down the location of research which would otherwise be indeterminate.

This idea of decreasing returns for each line of research is a third departure from the benchmark derived in Section 2 (and it also constitutes a minor departure from the canonical expanding variety framework). Constant returns to innovation in all lines of research would make the distribution of research across industries irrelevant. Decreasing returns instead pin down the location of research.[26]

Turning to cities, their number $m(t)$ can increase over time. Workers are still freely mobile and final goods freely tradable across cities. Initially there are more goods than cities and each city hosts the production of at least one good. Again, there are neither advantage nor cost to city size and each new patent leads to a new industry which may be first-nature with probability α or second-nature with probability $1-\alpha$. With first-nature industries, each patent holder must go to a specific location. Only at this location can the patent be exploited. Assuming that first-nature locations are always different, each new patent in a first-nature industry implies the creation of a new city. In second-nature industries, a patent can only be implemented by its innovator where it was developed. As in the benchmark model of Section 2, local spill-overs make research firms locate where the line of research they work on is implemented.

Steady-State growth and city size distribution

From equation (B 4), producers all make the same profit in equilibrium so that the present value of the uncertain profit stream is the same across industries: $v(z,t) = v(t)$ for all z. From equation (B 4), if an innovation takes place between t and $t+dt$, profits are scaled down by a factor $\frac{n(t)}{n(t)+1}$. In steady-state, this implies that the value of any manufacturer is also multiplied by $\frac{n(t)}{n(t)+1}$. Hence, research firm k when investing $\lambda^k(z)$ units of research over dt at a cost $w\lambda^k(z)dt$ can expect to win $b(\lambda(z),n) \times \lambda^k(z) \times \frac{n}{n+1} v dt$. Profit maximisation by research firms implies that in equilibrium

$$w = b(\lambda(z),n)\frac{n}{n+1}v. \tag{B 7}$$

Inserting equation (B 5) into (B 7) and re-arranging implies:

$$v = \frac{n+1}{\beta n^{2-\phi}}[\lambda(z)]^{\phi} w \ . \tag{B 8}$$

This equation implies that in equilibrium, the same quantity of research labour must be used in each line of research: $\lambda(z) = \lambda = \frac{\Lambda}{n}$ where Λ is total research labour.

Labour market clearing together with equation (B 2) and symmetry in manufacturing implies $\Lambda + \frac{1}{p} = 1$. Since $p = \frac{\sigma}{\sigma-1} w$, this expression implies $w = \frac{\sigma-1}{\sigma}\frac{1}{1-\Lambda}$. Inserting this into equation (B 8) and using symmetry across industries implies a first key equation relating the value of manufacturers to research employment:

$$v = \frac{\sigma-1}{\sigma}\frac{n+1}{\beta n^2}\frac{\Lambda^{\phi}}{1-\Lambda} \ . \tag{B 9}$$

Turning to the stock-market valuation of firms, manufacturers pay a dividend πdt over dt. The value of a manufacturer appreciates by $\dot{v}dt$ when no research firm succeeds in developing a new

[26] In the benchmark, such indetermination is avoided despite constant returns because of an advantage in own-industry innovations. This advantage is absent here since cross-industry innovations are ruled out.

patent, whereas it decreases by a factor $\frac{n}{n+1}$ in the opposite case. This loss occurs with probability ι, the aggregate probability of any research firm being successful as defined in equation (B 6). Thus, with any manufacturer, the expected rate of return for a shareholder is $\pi + \dot{v} - \iota\frac{\pi}{n+1}$. This return is risky but can be perfectly diversified since by equation (B 4), aggregate profit is constant and equal to $\frac{1}{\sigma}$. Consequently, manufacturers are valued so that their stock-market return is equal to the safe interest rate, $\dot{R}$, which is itself equal to the subjective discount rate ρ. Hence the absence of arbitrage implies a second key equation relating the value of manufacturers to research employment:

$$\pi + \dot{v} - \iota\frac{\pi}{n+1} = \rho v \; . \tag{B 10}$$

Then note that in steady-state the absence of arbitrage opportunity for investors also implies that the value of firms must remain constant between t and $t + dt$ if no new patent is developed. Inserting this together with (B 4), (B 6), and symmetry into equation (B 10) implies

$$v = \frac{1}{\rho\sigma n}\left(1 - \beta\frac{n}{n+1}\Lambda^{1-\phi}\right) \; . \tag{B 11}$$

The steady-state values of Λ and v solve equations (B 9) and (B 11). Λ is given by:

$$\rho(\sigma-1)\frac{n+1}{n}\Lambda^{\phi} - \beta\left(1 - \beta\frac{n}{n+1}\Lambda^{1-\phi}\right)(1-\Lambda) = 0 \; . \tag{B 12}$$

By inspection of equation (B 12), Λ is unique and interior. With rational investors, the no-arbitrage condition (B 11) is always satisfied. If Λ is larger (resp. lower) than determined by (B 9), the value of manufacturers goes down (resp. up) by (B 11) which decreases (resp. increases) the demand for research labour. Hence, this steady-state is stable.

The comparative statics of the implicit equation (B 12) is straightforward. It indicates that aggregate research labour (and hence the rate of innovation) decreases with ρ because of discounting. It also decreases with σ since a higher elasticity of substitution across goods reduces profits for manufacturers. The effect of the efficiency of innovation, β, is ambiguous. On the one hand, a higher β makes research more efficient and thus reinforces the incentive to invest. On the other hand, a higher rate of innovation depreciates the value of innovations. Equation (B 12) shows that the first effect dominates when β is small whereas the second effect dominates when it is large. The effect of ϕ, a measure of the decreasing returns in research, is also ambiguous for similar reasons. An increase in ϕ implies more strongly decreasing returns in research and thus a lower incentive to invest. At the same time, more congestion also decreases the rate of innovation which raises the value of existing patents.

Regarding welfare, there are four sources of inefficiency in this model. First, research firms do not take into account the surplus accruing to consumers when a greater number of products is available. Nor do they take into account the negative effect of new patents on existing profits. With CES preferences these two distortions exactly offset each other. The third distortion stems from research firms not internalising the effect of their innovations upon future innovations. Such intertemporal spill-overs imply that too little research labour is employed in equilibrium. The magnitude of this inefficiency can be shown to rise with σ (Grossman and Helpman, 1991*a*, Chapter 3). The last inefficiency is specific to this paper: Research firms can expect to get their average

and not their marginal expected returns since they do not internalise the congestion externality in research. This leads to over-investment whose magnitude increases with ϕ, the intensity of congestion. Overall the outcome is ambiguous. There may be too much or too little research in equilibrium depending on the relative values of ϕ and σ.

Because of local knowledge spill-overs and symmetry across industries, it is immediate that the population of a city where i patents are implemented is $\frac{i}{n}$: The population of a city is exactly proportional to the number of goods it manufactures. Symmetry in research labour implies that, conditional on an innovation taking place and leading to a second-nature industry, the probability that any particular city gets it is proportional to its population.

With respect to the number of patents, this model is thus equivalent to Simon (1955). Consequently, the size distribution of cities follows a power law with exponent $1/(1-\alpha)$.

A short proof of this result is as follows.[27] In steady-state, ratio of the number of cities with i patents m_i to the total number of patents n must be constant. This ratio can change for three reasons. A city with $i-1$ patents may gain one leading m_i to increase by one unit. A city with i patents may gain one leading m_i to decrease by one unit. Finally a city of any size but $i-1$ and i can gain one patent which decreases the ratio $\frac{m_i}{n}$. This implies the following steady-state condition:

$$E\left(\frac{\Delta(m_i/n)}{\Delta n}\right) = \frac{(1-\alpha)(i-1)m_{i-1} - (1-\alpha)im_i - m_i}{n^2} = 0\,. \tag{B 13}$$

This immediately yields $m_i = \frac{(1-\alpha)(i-1)}{(1-\alpha)i+1} m_{i-1}$. Substituting into the corresponding expressions for m_{i+1}, m_{i+2}, etc, shows directly that m_i in the upper tails is approximately a power law with exponent $1/(1-\alpha)$. Since the population of a city is proportional to its number of patents, the distribution of city sizes also follows a Pareto distribution with the same exponent.

[27] Simon (1955), Krugman (1996), and Gabaix (1999*a*) provide a more complete proof of this.