JOURNAL OF THE JAPANESE AND INTERNATIONAL ECONOMIES **10,** 399–418 (1996)
ARTICLE NO. 0023

Confronting the Mystery of Urban Hierarchy

PAUL KRUGMAN

Stanford University, Stanford, California 94305-6072

Received January 11, 1996; revised September 1, 1996

Krugman, P.—Confronting the Mystery of Urban Hierarchy

The size distribution of cities in the United States is startlingly well described by a simpler power law: the number of cities whose population exceeds S is proportional to $1/S$. This simple regularity is puzzling; even more puzzling is the fact that it has apparently remained true for at least the past century. Standard models of urban systems offer no explanation of the power law. A random growth model proposed by Herbert Simon 40 years ago is the best try to date—but while it can explain a power law, it cannot reproduce one with the right exponent. At this point we are in the frustrating postion of having a striking empirical regularity with no good theory to account for it. *J. Japan. Int. Econ.,* December 1996, **10**(4), pp. 399–418. Stanford University, Stanford, California 94305-6072

Journal of Economic Literature Classification Numbers R0, R1.

The usual complaint about economic theory is that our models are oversimplified—that they offer excessively neat views of complex, messy reality. This paper is an interim report on my efforts to understand why in one important case the reverse is true: we have complex, messy models, yet reality is startlingly neat and simple. In large part this is a report of failure—that is, while there must be a compelling explanation of the astonishing empirical regularity in question, I have not found it. At best, this paper offers some clues and hints about where the explanation might be found—as well as some news about which trails have turned out to be false.

The regularity in question is that involving the size distribution of metropolitan areas. It has been known for at least 70 years that the distribution of larger cities in the United States is surprisingly well described by a power law—that is, the number of cities with a population larger than S is approximately proportional to S^{-a}, with a quite close to 1. (See Carroll

0889-1583/96 $18.00

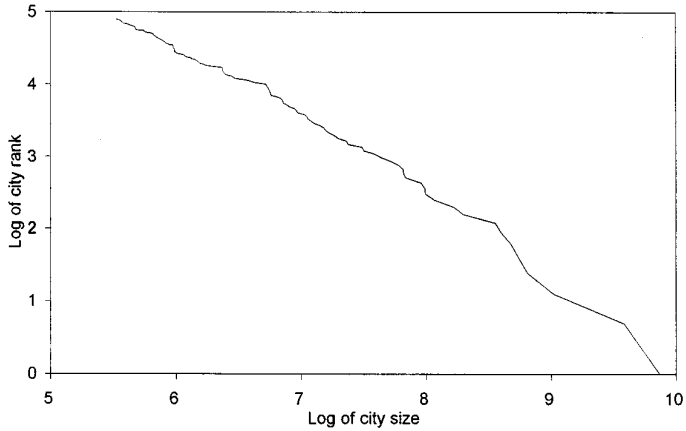

FIGURE 1

(1982) for a survey of the massive empirical literature on city size distributions). To get an idea of how well this works, consider that in 1991 there were 40 U.S. metropolitan areas with more than 1 million people, 20 with more than 2 million, and 9 with more than 4 million (Houston was just a bit too small). Figure 1 plots the log of metropolitan area size against the log of rank (i.e., New York = 1, Los Angeles = 2, etc.) for the 130 such areas listed in the *Statistical Abstract of the United States;* the near-linearity, and the approximately $-45°$ slope, are remarkable, and this visual impression is confirmed by more formal statistical analysis. Let $N(S)$ be the number of cities of population S or greater; then a log-linear regression finds

$$\ln(N) = 10.549 - 1.004 \ln(S). \\ (.010)$$

Nor is this just a fact about a single time and place. As already suggested, the distribution of city sizes in the United States has been well described by a power law with an exponent close to 1 at least for the past century. International data are more problematic, in particular because it is difficult to assemble comparably defined metropolitan areas. However, the classic study by Rosen and Resnick (1980) suggests that most national metropolitan size distributions are well described by a power law with an exponent not too far from 1—and that the exponent gets closer to 1 the more carefully the metropolitan areas are defined.

Why should this be so? Perhaps we should break the question into three parts:

1. Why is there such a wide size range of cities? An economist who did not know the data might expect cities to cluster around some typical size, reflecting a balance between the external economies that give rise to cities in the first place and the external diseconomies that make large cities so difficult to deal with. In fact, there is no typical size of city, nor any tendency to converge on any particular size.

2. Why does the size distribution obey a power law? (Strictly speaking, it is only the upper tail that obeys such a law—there are fewer small cities than one would expect from the power law.) One way to put this question, which ought to give us some kind of clue to the explanation, is to note that power laws involve a sort of principle of self-similarity. That is, if the number of cities with populations greater than S is S^{-a}, then the fraction of cities with more than a million people that have 2 million or more is the same as the fraction of cities with more than 500,000 that have 1 million or more. Surely this is significant of something—but what?

3. Why is the exponent (close to) 1? In general, the persistence of an apparently constant parameter across vast stretches of time and space is puzzling. American cities in 1890 were largely built around steam-powered manufacturing and railroads, with a workforce that commuted by trolleys; in 1990 they were service centers whose workforce commuted by private automobile. Why should the tension between external economies and diseconomies have produced similar results? Beyond this, 1 is a special number—as we will see below, it is the exponent at which a power law becomes degenerate, and at which the expected population in cities whose population exceeds S would (if we ignored integer constraints) be infinite. It is hard to escape the feeling that there must be something deeply significant about the fact that the size distribution of cities is right at the ''singularity'' at which the very distribution that works so well ceases to make sense, or alternatively at the point at which the ''quantum'' nature of urban places—no fractional cities allowed—becomes crucial to save us from absurdity. But what does it mean?

What I will argue in the remainder of this paper is that the fact that there are three questions rather than just one makes it impossible to be comfortable with the present state of our understanding. The standard urban system models, whether based on the classic work of Henderson (1974) or the more recent efforts of Fujita *et al.* (1994), offer seemingly persuasive answers to question 1—why are cities so unequal in size?—but they fail to predict a power law. The ingenious analysis of Simon (1955) suggests a mechanism that could produce a power law—albeit only by sacrificing much of the economic content of the urban systems literature—but runs into severe difficulties when one tries to get it to produce the particular, edge-of-degeneracy power law that we actually see in the data. I will offer a sketch of a different way that we might explain a power law,

one that could be *consistent* with an exponent of 1—but this approach still does not offer a compelling reason why the exponent should be near 1 across a wide range of times and places.

The remainder of this paper is in five parts. It begins with reviews of the existing economic models of urban systems: the classic aspatial approach of Henderson (1974) and the recent work by Fujita *et al.* on spatial urban hierarchies. The main message of these sections is that while these models offer considerable insight into why cities of different sizes may coexist, they offer no hint of why the size distribution should obey any sort of power law, let alone one with $a = 1$. The third part turns to the Simon (1955) model of random growth, which does at least have the virtue of offering a rationale for the existence of power laws. I show, however, that Simon's model has two defects: aside from lacking much economic motivation, it also runs into trouble when it is called on to explain power laws with a close to 1. The fourth part of the paper offers a sketch of a new approach, which abandons the traditional assumption of a homogeneous landscape and instead considers a landscape with randomly varying transport costs. I argue that, rather oddly, seemingly complicating the model in this way may be necessary in order to explain the mysterious simplicity of the actual data. A final section summarizes the state of play.

1. The Urban System Model

There is an extensive literature on urban systems—that is, on the reasons an economy generates a particular distribution of city sizes and roles. For the most part, we can classify this literature as drawing on two basic ideas: the spatial models of urban hierarchy that derive from the ''central place theory'' of Christaller (1933) and Lösch (1940), and the aspatial urban system model that was introduced by Henderson (1974). In this section I briefly review the Henderson-type model; the next section turns to spatial urban hierarchies.

The basic idea of Henderson's analysis is extremely simple: there is a tension between external economies associated with geographic concentration of industry within a city, on one side, and diseconomies associated with large cities on the other. The net effect of this tension is that the relationship between the size of a city and the utility of a representative resident is an inverted U, like the one shown in Fig. 2.

It might seem obvious that if this is the tradeoff between city size and welfare, all cities will be of the optimum size, as indicated by point O. This is in fact Henderson's assertion; but it is, as he recognizes, not quite that easy. The way in which he argues that cities will in fact tend to be of optimal

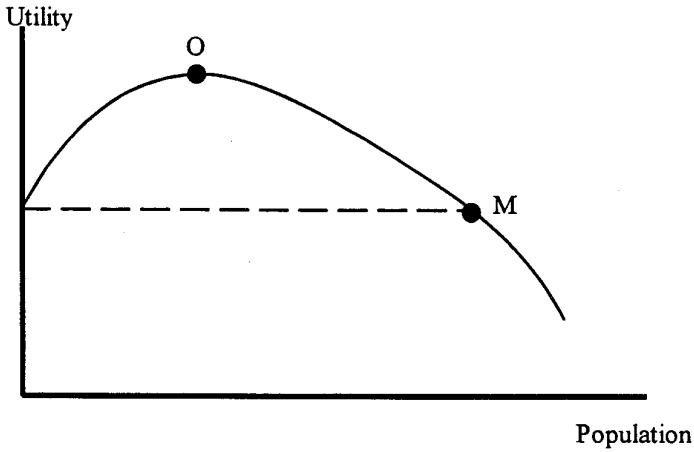

FIGURE 2

size and the way in which he alters the model to get multiple sizes of cities are what makes his work distinctive.

Suppose for a moment that there were too few cities, and thus that the typical city was too large, i.e., lay somewhere along the arc OM.[1] Then it is straightforward to see that no individual resident would have any incentive to move to a new location: any existing city would still yield a higher level of welfare than moving in isolation to a new location. This seems to imply the possibility both of substantially suboptimal city sizes and of multiple equilibria in the size distribution as well as location of cities. What Henderson argues, however, is that reality is simplified through the forward-looking behavior of large agents: any situation with too few cities would offer a profit opportunity. Anyone who could organize a "city corporation" that moves a number of people to a new city of optimal size would be able to profit (perhaps through land prices). It turns out that developers of often startling size play a significant role in urban growth in the United States. So Henderson argues that the actual city sizes are, to a first approximation, optimal.

But then why are cities of such different sizes? Here the argument runs as follows: external economies tend to be specific to particular industries; but diseconomies tend to depend on the overall size of a city, whatever it produces. This asymmetry has two consequences. First, because there are diseconomies to city size, it makes no sense to put industries without mutual

[1] It is straightforward to see that a situation in which there are too many cities, and thus where the typical city is too small, is unstable: some of the cities will simply collapse.

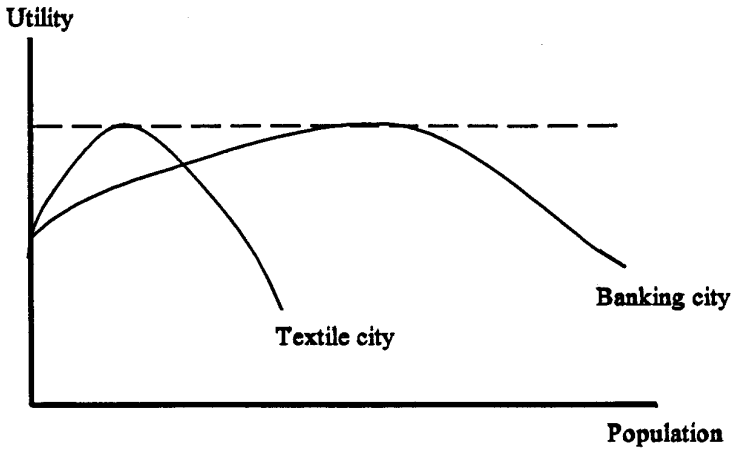

FIGURE 3

spillovers in the same city—if steel production and publishing generate few mutual external economies, steel mills and publishing houses should be in different cities, where they do not generate congestion and high land rents for each other. So each city should be specialized (at least in its "export" industries) in one or a few industries that create external economies. Second, the extent of these external economies may vary greatly across industries: a textile city may have little reason to include more than a handful of mills, while a banking center might do best if it contains practically all of a nation's financial business. So the optimal size of city will depend on its role.

The last step in Henderson's analysis is to argue that relative prices will adjust so that the welfare of a representative resident in cities of whatever type are the same. The end picture will look like Fig. 3: for each type of city there will be an optimum size. At the optimum size each will yield the same utility, but that size will vary depending on the type of city.

It's an impressively concise and clean analysis. It makes the sizes of cities an economic variable that depends on forces which can, in principle, be measured (and Henderson's model has given rise to extensive empirical work); it also helps explain why the actual distribution of cities contains a wide size range that shows no signs of collapsing.

It is hard to see, however, why this model should generate anything that looks like a power law. Suppose that the optimum banking city has 4 million people, the optimum high-tech manufacturing city 2 million, and the optimum low-tech city 1 million; why should the ratio of banking to high-tech cities required by the economy be the same as the ratio of high-tech to low-tech? Since the Henderson model generates a size distribution

out of a tension between external economies and diseconomies—both of which presumably depend on the technologies of production, communication, transportation, and so on—one would surely predict from this model that the size distribution would change over time, rather than show the mysterious stability it exhibits in practice.

2. Central-Place Theory

When geographers as opposed to urban economists discuss city sizes, they usually begin with the ''central place theory'' of Christaller (1933) and Lösch (1940). Like urban system theory, central place theory envisages urban places as emerging because of economies of scale in the provision of manufactured goods or services. What prevents all of these activities from becoming concentrated in a single location is the existence of activities, such as farming, that must make use of dispersed resources. This creates a tension between economies of scale and transport costs, with an optimum or equilibrium (the original central place theorists were blurry about the difference) number and size of cities. If all urban activities involved the same degree of scale economies and transport costs, one would expect a regular lattice of equal-sized urban places (which, on a homogeneous plain, would have Lösch's famous hexagonal market areas). Because activities differ in scale effects and transportation costs, however, one expects to see urban areas of different types; in particular, Christaller argued, one expects to find a hierarchical structure.

In its original formulation, central place theory falls far short of being what an economist would call a model; it is indeed generally unclear whether one is looking at supposed equilibria of a decentralized market process, the solution to a planning problem, or simply a plausible but ad hoc classification scheme. Recently, however, Fujita and colleagues have made use of ''new economic geography'' type models to show that Christaller-type hierarchies can indeed emerge from a decentralized process. In particular, if one imagines a dynamic process in which the population gradually increases, leading to a moving agricultural frontier and the occasional formation of new cities, one can generate an emergent hierarchy of central places.

To do this analysis right is surprisingly hard. It is possible, however, to suggest the character of the results using a heuristic approach.

A Heuristic Model of Urban Location

We envisage an agricultural population of size $2D$ spread evenly along a line, choosing units so that one farmer lives on each unit of line; this makes the line $2D$ long, and we measure it from the center, so that it

extends from $-D$ to D. We assume that each farmer demands enough manufactures to directly support μ manufacturing workers—and also that each manufacturing worker also directly supports μ manufacturing workers, so that the total manufacturing employment is $2D/(1 - \mu)$.[2]

Manufacturing is next assumed to consist of N firms, each producing a distinct product which gets a fraction $1/N$ of the total demand for manufactures. We assume that N is large enough that each firm can ignore the impact of its own workers on the geographical distribution of demand. Each firm is free to choose both the number and location of plants. It incurs a constant production cost c for each unit it produces; since this is independent of locational decisions, it will be ignored. It also incurs a transportation cost t for each unit shipped one unit of distance, and a fixed cost F for each plant it chooses to operate. The firm is simply assumed to choose the number and location of plants so as to minimize the sum of fixed and transportation costs.[3]

This is clearly a peculiar model, which is hard to justify in terms of consumers' budget constraints, let alone any coherent description of market structure. Its only justification is that it allows us to focus in a very simple way on the interdependence among firms' location choices, and on the dynamics of city formation.

Interdependence of Location Choices

The interdependence of location choices can be illustrated by considering the following situation: all firms have only one plant, and all of these plants are concentrated at a single "urban" location U. Let us now ask whether an individual firm has an incentive to relocate its plant away from U.

It is obvious that if U is at location 0, that is, in the dead center of the agricultural area, there will be no reason for individual firms to locate elsewhere. But what if it is off-center? In this model we can see easily that for at least some range of locations firms will still want to stay in the city, wherever it is—a point that will shortly turn out to give a clue to the process of city formation.

Consider the transportation costs of a firm that chooses a location L, possibly different from that of the city U. The situation is illustrated in Fig. 4. We may think of the firm as shipping its good to three different markets: the $D + L$ farmers to its left, at an average distance of $(D + L)/2$; the $D - L$ farmers to its right, at an average distance of $(D - L)/2$; and the

[2] Obviously this involves pulling a bit of a fast one—what about the price elasticity of demand? However, there is worse to come.

[3] If firms are not simply concerned with minimizing costs, it becomes necessary to analyze their location decisions in terms of a "market potential" function that measures the relative profitability of different locations. See Fujita *et al.* (1994).

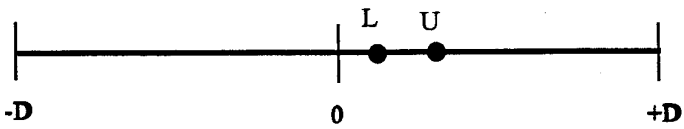

FIGURE 4

$2D/(1 - \mu)$ urban consumers, at the distance $|U - L|$. Thus the firm's overall transportation costs will be

$$T = \frac{t\mu}{N}\left[\frac{1}{2}(D + L)^2 + \frac{1}{2}(D - L)^2 + 2\frac{D}{1 - \mu}|L - U|\right]. \tag{1}$$

A planner who could choose to locate all firms simultaneously in the appropriate places would clearly set $L = U = 0$. Suppose, however, that for some reason a city has formed at an off-center location $U \neq 0$. Then it may still be in the best interests of each individual firm to locate at that city. Figure 5 illustrates how a firm's transport costs would depend on its location for a particular example of an off-center city ($D = 1$, $t = 0.1$, $U = 0.2$, $\mu = .2$). The $|U - L|$ term creates a "kink" at the city location, which may make it the minimum transport cost location for each individual firm even though it is the "wrong" location from the point of view of a central planner.

This example illustrates in a very simple way how a city, once formed,

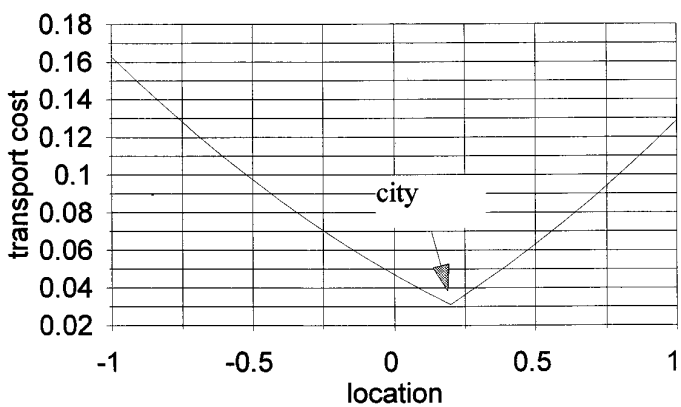

FIGURE 5

tends to have its location "locked in." We can now exploit this insight to examine how a growing urban system might give birth to new cities.

New Cities

Let us now, following Fujita *et al.,* consider how, when, and where new cities would arise if we took the model just developed and allowed it to experience gradual population growth.

Suppose that initially all manufacturing is concentrated at $U = 0$. We can now ask what happens as D increases—that is, the population grows and the agricultural frontier shifts out with it.

At any point individual firms have the option of opening new plants. Suppose that a firm were to consider opening a new plant to the right of U, say at the location L'. What would the optimal location for that new plant be?

Once the plant was opened, customers would be served from whatever plant was nearest to them. That means that, of customers to the right of $U = 0$, those from 0 to $L'/2$ would continue to be served from the old plant; those from $L'/2$ to D would be served from the new plant. It is easy to show that the transport costs of serving these consumers would be

$$T = \frac{t\mu}{2N}\left[\frac{1}{2}(L')^2 + (D - L')^2\right] \tag{2}$$

so that the optimal location for the new plant is at $2D/3$ and the transport costs are $t\mu D^2/6N$. The cost of serving these same customers from the original plant is $t\mu D^2/2N$. So it is worth introducing a new plant, at the location $L' = 2D/3$, as soon as

$$\frac{t\mu D^2}{3N} > F, \tag{3}$$

implying that new cities will appear as soon as D reaches the critical value

$$D^* = \sqrt{3FN/t\mu}. \tag{4}$$

Once the cities have come into being, however, they will be locked in place by the same logic that locked the original city into place, and we can then repeat the analysis as the population grows. When the agricultural frontier has extended D^* beyond the existing cities, a new pair will pop into existence, and so on. Thus there will be a typical distance D^* between cities, depending positively on fixed costs and negatively on transport costs, and of course a typical size of cities as well.

Urban Hierarchies

So far this model generates only one size of city. But now suppose that there are many different industries, differing in their fixed and/or transport costs. The analysis then becomes considerably more complicated. However, it is fairly easy to convince oneself of the following:

1. New cities will form first for industries with high transport costs or low fixed costs.
2. Lower transport cost or higher fixed cost industries will concentrate in fewer cities.
3. Because of the way that existing transport costs cause a ''kink'' in the transport cost function, when low transport/high fixed cost industries open new plants they will often do so in existing small cities. (If they do not, the smaller cities will frequently be ''pulled'' into the new, larger city.)
4. The result will be the formation of a hierarchy of central places with a number of levels: small cities serving only the local agricultural population and themselves, larger cities that serve a market area that includes smaller cities, and so on.
5. For a range of parameters (i.e., not a set of measure zero), this will be a strict hierarchy: each large city will have an integer number of satellite cities, and so on. (This is an example of the ''phase-locking'' discussed in Krugman (1996).)

In short, this model, which is intended as a heuristic version of the far more careful but also more difficult work of Fujita *et al.,* offers a way to see how the urban hierarchies of central place theory can be seen as the outcome of a dynamic market process rather than as a static planning framework.

Unfortunately, just as in the case of the Henderson-type urban systems literature, it is difficult to see why the size distribution of firms should obey a power law, let alone one with the specific slope of -1.

3. Random Urban Growth

The urban system models just described, although they rely on some rudimentary dynamics to limit the range of possible outcomes, are essentially tales about static tradeoffs—the tradeoff between external economies and diseconomies in the Henderson model and the tradeoff between economies of scale and distance in central place theory as formalized by Fujita *et al.* There is, however, an alternative tradition, due mainly to Herbert Simon, which views the existence of a wide size range of cities (or for that matter of business firms) as evidence that there really aren't any tradeoffs—that size is more or less irrelevant. Simon argued that it is

precisely because size is irrelevant that a process of random growth can produce a huge range of sizes whose upper tail is well described by a power law.

Simon's original exposition of a random-growth model (Simon, 1955; Ijiri and Simon, 1977) has had surprisingly little impact on economic thinking. This may perhaps be because it is nihilistic about the economics, giving us little more to say, but it is also in part because the exposition is peculiarly dense. In recent work (Krugman, 1996) I have developed a more user-friendly way to explain Simon's model; here I offer an even more streamlined version, which serves to highlight both the insightfulness and the weakness of the model.

Simon's Model

As a starting point, it is useful to consider an alternative statement of the power law on urban sizes. We know that the upper tail is well described by a relationship of the form $N = kS^{-a}$, where N is the number of cities with populations greater than S. We may therefore also say that the *density* of city sizes is $n = -akS^{-a-1}$. Finally, in what turns out to be the most useful statement, we may say that the *elasticity* of the density of cities with respect to size is $-a - 1$:

$$\frac{dn}{dS}\frac{S}{n} = -a - 1. \tag{5}$$

We can now turn to Simon's urban growth model. Simon envisaged a process in which the urban population grows over time by discrete increments—call them "lumps"—and let the population at any point in time, measured in lumps, be P. Where does a new lump go when it arrives? Simon supposes that with some probability π it goes off to a previously unpopulated location, i.e., creates a new small city. With probability $1 - \pi$ it attaches itself to an existing city, with the probability that any particular city gets the next lump proportional to its population.

This is an extremely nihilistic and simplistic model. It supposes that there are neither advantages nor disadvantages to city size: a city is simply a clump of lumps, whose expected growth rate is independent of size. If you like, you may think of a lump as an industry; in this case, Simon's model says that industries are equally likely to give birth to other industries in the same city regardless of the city's size.

There would be little reason to take such a model seriously, except for one thing: the size distribution of cities does follow a power law, and Simon's model both predicts this result and gives at least a hint why the size distribution might have remained stable despite huge changes in technology and economic structure.

To analyze the model, we provisionally assume that over time the urban size distribution converges to a steady state. That is, the ratio of the number of cities of size S, n_S, to the population tends toward a constant. The ratio n_S/P can change for three reasons: a city of size $S-1$ may expand by one lump, which increases n_S; a city of size S may expand by one lump, which reduces n_S; or the overall population may increase, which reduces n_S/P. If we write the expected change in n_S/P when P increases, and are carefully sloppy about the discrete nature of the change in P, we find

$$\frac{Ed(n_S/P)}{dP} = \frac{1}{P^2}[(1-\pi)n_{S-1}(S-1) - (1-\pi)n_S S - n_S]. \tag{6}$$

If the city size distribution is to approach a steady state, however, in the long run this expected change must be zero, giving us a relationship in steady state between the number of cities of sizes S and $S-1$:

$$\frac{n_S}{n_{S-1}} = \frac{(1-\pi)(S-1)}{(1-\pi)S+1}. \tag{7}$$

This may be rewritten as

$$\frac{n_S - n_{S-1}}{n_{S-1}} = \frac{\pi - 2}{(1-\pi)S+1}. \tag{8}$$

We now focus only on the upper tail of the distribution, for which S is large. In this case, it will be possible to approximate the discrete distribution of city sizes by a smooth distribution $n(S)$, with

$$\frac{n_S - n_{S-1}}{n_{S-1}} \cong \frac{dn/dS}{n} \cong \frac{\pi - 2}{(1-\pi)S+1}. \tag{9}$$

We can then derive the elasticity of n with respect to S,

$$\frac{dn}{dS}\frac{S}{N} = \frac{\pi - 2}{1 - \pi + 1/S} \cong \frac{\pi - 2}{1-\pi}, \tag{10}$$

which by (5) tells us that the upper tail of the city size distribution will be characterized by a power law with exponent $a = 1/(1-\pi)$.

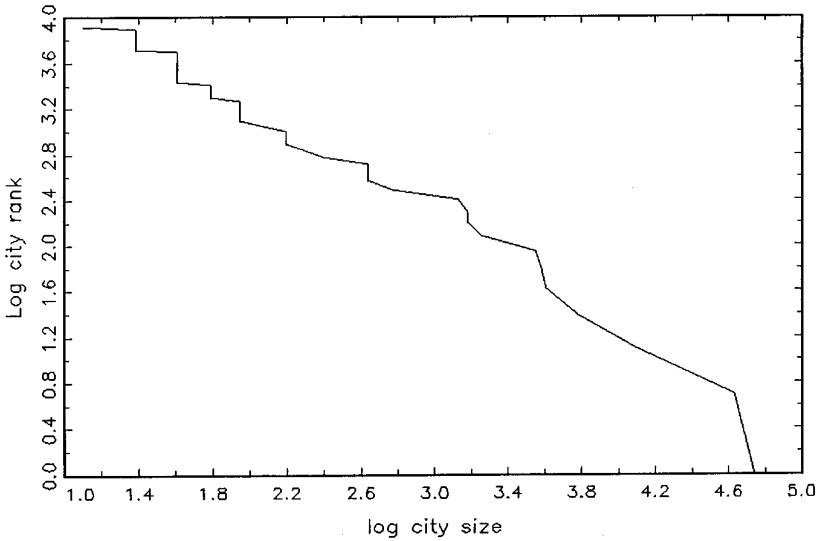

FIGURE 6

Does this really work? Yes, it does. Figure 6 shows the results of a simulation run in which π was set equal to 0.2 and in which I started with 10 seeds of one lump each and allowed the population to grow by a factor of 100. The rank–size relationship for the top 50 cities is shown; it is reasonably log-linear, with a slope not too far from the predicted value.

If your concern is with the seeming universality of a power law on city sizes, with an exponent that is stable across time and space, Simon's model represents a big improvement over the "economistic" models, for at least three reasons:

1. It predicts a power law, whereas the urban system and central place models do not.
2. The parameter that determines the exponent on the power law is the probability of forming a new city, which seems less obviously something that must have changed drastically over the past century than variables like economies of scale or urban commuting costs.
3. The mysterious exponent of 1, which seems so hard to justify, has a natural interpretation here: it is what you get when increments to urban population usually attach themselves to existing cities rather than forming new cities.

So Simon's model seems to get us much closter to the grail of understanding the power law on city sizes. One might object to its lack of economic

content. However, even if one is willing to let that slide, there is a further problem.

The Degeneracy Problem

In the derivation of the power law result in Simon's model, the crucial first step was the assumption that the urban size distribution tends to approach a steady state. Yet this can never be exactly true: there is no upper bound on the size of the largest city, which will therefore always tend to rise. The reason the model nonetheless works is that the largest city tends in the long run to have an ever smaller share of the total population, and therefore an ever smaller share of the increment in population goes to making that biggest city bigger. Suppose that the distribution really followed a power law throughout (it really does so only for the upper tail, but this will be good enough), and let $S_{\min}$ be the size of the minimum possible city. Then it is straightforward to show that the share of the population in cities larger than any given size S will be $(S/S_{\min})^{1-a}$. As long as $a > 1$, then—which will be true as long as $\pi > 0$—the share of the growth in population going to make the biggest cities bigger will eventually become negligible, justifying the steady-state assumption.

Unfortunately, the data tell us that a is extremely close to 1, implying π essentially zero. Under these conditions Simon's process does not produce a power law, as one might expect, because with $a = 1$ a power law would predict an infinite urban population! Nor can we evade the problem by assuming that π is only close to zero, but not exactly zero. Intuition suggests, and simulations confirm, that when π is very small it requires a very large increase in the urban population to produce a smooth power law. (Notice that in Fig. 6, with π still an unacceptably large 0.2 and the urban population increasing by a factor of 100, the relationship is still not nearly as smooth as in the actual data.) Yet the power law on U.S. city sizes, with a very close to 1, has prevailed at least since 1890.

Why doesn't the power law with an exponent of 1 pose a problem in the real data? Because of an integer constraint. A continuation of the power law for the United States would predict 0.5 cities with twice the population of New York, 0.25 cities with 4 times the population, and so on, with an implied infinite population; but because fractional cities are impossible, this is not a real problem. This observation should send chills down the spine of anyone who knows something about the history of physics: it was the need to impose an integer constraint to avoid predicting infinite blackbody radiation that led to the discovery of the quantum nature of energy. It is deeply suggestive that the exponent of the power law on city sizes should be precisely at the point at which the indivisible nature of cities is necessary for the distribution to make sense. But what it suggests is still a mystery.

4. Percolation Models

What follows is closer to a suggestion for a model than an actual model; as indicated at the beginning of this paper, this is a progress report (or perhaps a lack of progress report) rather than a set of answers.

The Simon model suggests strongly that the explanation of the power law on city sizes should be sought in some kind of random process. Simon's own proposed process, however, does not seem to be compatible with the fact that a power law with an exponent very close to 1 emerged so quickly in the United States. There does not seem to have been time enough to get this result.

It is a familiar proposition in some physical contexts, however, that random growth need not proceed in time—it may proceed in space instead. If this sounds obscure, consider the classic example of "percolation theory": a porous rock in which any two holes are connected with probability p. Obviously the expected size of connected areas depends on p; there is a critical value at which this expected size becomes infinite. But it also turns out that when p is close to but slightly below this critical value, the upper tail of the size distribution of connected regions follows a power law. One can think of a sort of hypothetical growth process by which these regions are explored—start with a hole, explore all the connected holes, then explore the holes connected to them, and so on; but one need not suppose that the creation of these regions takes place over time—they are there from the beginning.

How can this be relevant to urban size distributions? It may be necessary to drop a ground rule that has implicitly been followed by nearly all attempts to model the sizes of cities: the assumption that the landscape is homogeneous, so that differences in city size have nothing to do with differences in the inherent advantages of their locations. In practice, of course, we know that special advantages such as rivers and ports have played a considerable role in the rise of great cities (see Fujita and Mori (1995) for an analysis of this phenomenon in terms of the effect of inhomogeneities in the landscape on market potential). A possible route to resolving the puzzle of a power law on city sizes is to make an assumption that is precisely the opposite of the standard one: namely, that the landscape is strongly inhomogeneous, and that this inhomogeneity is the source of the size distribution of cities.

It is obvious that differences in the quality of city sites will generate inequality in city sizes. But why should this inequality take the form of a power law? The answer is that under some circumstances the variation in the landscape may usefully be regarded as random, and that this random variation could produce a power law.

Consider the following example. Suppose that a country can be repre-

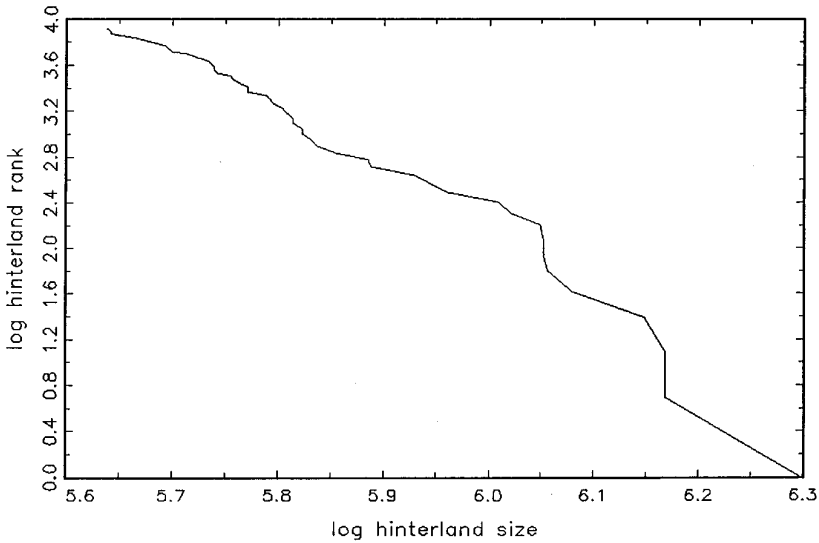

FIGURE 7

sented by a large number of locations laid out in a checkerboard fashion. The transport cost between any two adjoining locations is random, say between 0 and 1. Let the first row of the checkerboard represent the "coast," and suppose that goods from the interior are shipped to the coast by the lowest cost route (to make life simple, let us assume that no backtracking is allowed). Then each location on the coast will have a natural "hinterland," consisting of all the locations in the interior which can most cheaply reach the sea via that "port." If the interior is at all deep, these hinterlands may be of very unequal size: the hinterland of a successful port may fan out in a rough triangle, eventually taking in a large area.

Not surprisingly, the upper tail of the size distribution of hinterlands tends to follow a power law. Figure 7 shows the results of a simulation in which the country consisted of 40 rows of 1000 locations each. It plots the logarithm of the size of the top 50 hinterlands against the logarithm of their rank. The fit to a power law is quite good, and notice that this power law need not evolve over time: it is a purely static feature of the landscape.

Of course, if we think of the size of a city as simply being proportional to the size of its hinterland, the exponent is much too large—that is, the large cities are far too equal in size. One might think of a number of reasons why the real exponent might be smaller. One is economies of scale in transportation: a port with a large hinterland might tend to be better served by roads, railways, etc., helping it to expand its hinterland still further. To

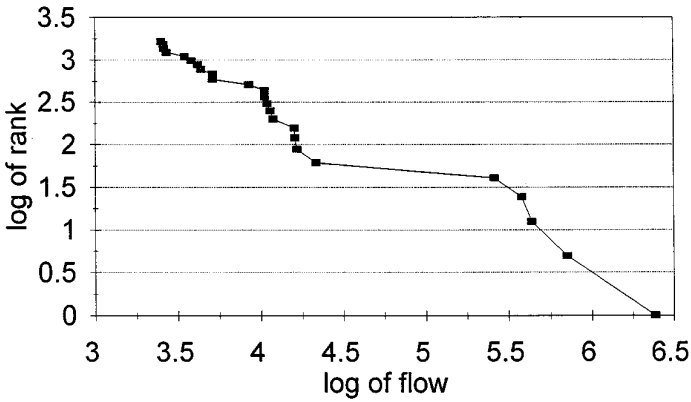

FIGURE 8

those with a vague acquaintance with geology, this story is reminiscent of the idea of "stream capture," the way in which larger streams that dig their beds deeper are sometimes able to appropriate former tributaries of smaller rivers. For that matter, we might note that the variation in the natural hinterlands of ports implied by this model may be too small, precisely because natural transportation routes tend to be correlated across space—think of the Mississippi River or the Hudson–Mohawk system.

It is impossible to resist throwing in a natural example at this point. It turns out that the size distribution of large rivers is reasonably well described by a power law. Figure 8 plots the log of the flow through the 25 largest rivers in the United States against the log of their rank. The relationship is roughly linear; the regression coefficient is -0.949.

The suggestion here, then, is that some kind of percolation process across an uneven natural landscape might be an alternative to random growth over time in explaining the power law on city sizes.[4] The virtue of such a story, aside from the way it brings back the traditional geographical concerns over rivers, harbors, and so on, is that it does not pose the same problems of timing that the random growth story does. Unfortunately, it remains equally unclear why the exponent should be so close to 1. (Perhaps the river size distribution, which also has an exponent oddly close to 1, offers some kind of clue.)

[4] We might also raise yet another possibility: the "landscape" in which cities exist is not a physical one, but one defined in some other space, such as that of technological characteristics, I.e., a city with industry A is then able also to get industries B and C, which lead to D, E, and F, and so on. It may seem bizarre to suggest that such an abstract notion could produce a regularity in city sizes—but the regularity is there, and needs some kind of explanation!

5. Summary

This paper posed three questions about the urban system. Of these, we seem to have reasonably good answers to the first, we have a glimmering of an answer to the second, and we are quite in the dark about the third.

The first question is why city sizes are persistently unequal, why there is no typical size of a city. Economic analysis has made considerable progress on this issue: the classic Henderson-type urban systems model has now been complemented by an analysis that makes sense of the geographers' concept of a central-place hierarchy. If we did not know about the mysterious regularities in the size distribution, these models might well seem quite adequate.

The second question is why the sizes of large cities are so well described by a power law. Simon's unjustly neglected random-growth model offers a potential explanation. I suggest here that there may be an alternative "percolation" approach that stresses natural differences in the advantages of different city sites, particularly in terms of access to hinterlands. Power laws are pervasive in natural systems, so it should perhaps not be too surprising that we find one for cities as well—but it is hard to see how to reconcile the types of model that might explain a power law with the urban system models that are otherwise so persuasive.

Finally, the power law on city sizes is a disturbing one: the exponent is very close to one. This is more or less baffling, while also being intriguing. A power law with an exponent of one is degenerate—it implies a population of infinite size, or would except for the "quantum" requirement that excludes fractional cities. Because that quantum requirement is ignored in Simon's model, that model breaks down for exponents close to one. "Percolation" models do not have the same problem, but it is still hard to see why 1 should consistently be the exponent.

The failure of existing models to explain a striking empirical regularity (one of the most overwhelming empirical regularities in economics!) indicates that despite considerable recent progress in the modeling of urban systems, we are still missing something extremely important. Suggestions are welcome.

References

Carroll, G. (1982). National city-size distributions: What do we know after 67 years of research?, *Prog. Human Geography,* **6,** 1–43.

Christaller, W. (1933). "Central Places in Southern Germany." Fischer Jena; English translation by C. W. Baskin, Prentice Hall, London, 1966.

Fujita, M., and Mori, T. (1995). Why are most great cities port cities? Transport nodes and spatial economic development, mimeo.

FUJITA, M., KRUGMAN, P., AND MORI, T. (1994). On the evolution of hierarchical urban systems, mimeo.

HENDERSON, J. V. (1974), The sizes and types of cities, *Amer. Econ. Rev.* **64,** 640–656.

IJIRI, Y., AND SIMON, H. (1977). "Skew Distributions and the Sizes of Business Firms." North-Holland, Amsterdam.

KRUGMAN, P. (1996), "The Self-Organizing Economy." Blackwell, Cambridge, MA.

LÖSCH, A. (1940), "The Economics of Location." Fischer, Jena; English translation, Yale Univ. Press, New Haven, 1954.

ROSEN, K., AND RESNICK, M. (1980). The size distribution of cities: an examination of the Pareto Law and primacy, *J. Urban Econ.* **8,** 165–186.

SIMON, H. (1955), On a class of skew distribution functions, *Biometrika.*