Krugman, Geography & Trade

1 Center and Periphery

About a year ago I more or less suddenly realized that I have spent my whole professional life as an international economist thinking and writing about economic geography, without being aware of it.

By "economic geography" I mean "the location of production in space"; that is, that branch of economics that worries about where things happen in relation to one another. It's not worth trying to define my subject more exactly than that—you'll see better what I mean once I start describing models. Most of regional economics, and some but not all of urban economics, is about economic geography in the sense I have in mind.

If you had never looked at the theory of international trade, you might have supposed that international economics would also be largely treated as a special case of economic geography, one in which borders and the actions of sovereign governments play a special role in shaping the location of production. What I will argue in these lectures is that that is how international economics ought to be done, at least part of the time. But

it is almost never the way that it is done at present. Instead, the analysis of international trade makes virtually no use of insights from economic geography or location theory. We normally model countries as dimensionless points within which factors of production can be instantly and costlessly moved from one activity to another, and even trade among countries is usually given a sort of spaceless representation in which transport costs are zero for all goods that can be traded.

There is nothing wrong with simplifying assumptions—on the contrary, it is only through strategic simplification that we can hope to make any sense of the buzzing complexity of the real world. The particular simplifying assumptions of conventional trade theory have led to an impressive and very useful intellectual construct. For some purposes it does no harm to ignore the fact that countries are not points and that some pairs of countries are much closer than others—that California is farther from New York than any place in the European Community is from any place else, or that London and Paris are much closer to each other than are New York and Chicago, or for that matter that Canada is essentially closer to the United States than it is to itself.

Yet the tendency of international economists to turn a blind eye to the fact that countries both occupy and exist in space—a tendency so deeply entrenched that we rarely even realize we are doing it—has, I would submit, had some serious costs. These lie not so much in lack of realism—all economic analysis is more or less unrealistic—as in the exclusion of important

issues and, above all, of important sources of evidence. As I hope I will be able to show, one of the best ways to understand how the international economy works is to start by looking at what happens *inside* nations. If we want to understand differences in national growth rates, a good place to start is by examining differences in regional growth; if we want to understand international specialization, a good place to start is with local specialization. The data will be better and pose fewer problems of compatibility, and the underlying economic forces will be less distorted by government policies.

The decision by international economists to ignore the fact that they are doing geography wouldn't matter so much if someone else were busy exploiting the facts and insights that can come from looking at localization and trade within countries. Unfortunately, nobody is. That is, of course, an unfair statement. There are excellent economic geographers out there, as well as urban and regional economists who worry about geographical issues. For reasons that I will discuss in a moment, however, these people are almost uniformly peripheral to the economics profession. International economics is a flagship field: no serious economics department can get by without at least one international trade expert and without offering international economics as a field for its graduate students. By contrast, regional and even urban economics are given far less priority. And economic geographers proper are almost never found in economics departments, or even talking to economists; at best they are in urban studies departments, more usually in geography departments. They may do excel-

lent work, but it does not inform or influence the economics profession.

There are good reasons why this has happened and equally good reasons why it should change. Before I begin to present my own ideas, I want to talk briefly about why international economists don't acknowledge that they are doing geography—and why they should.

Geography: Why Not and Why

The neglect of spatial issues in economics arises for the most part from one simple problem: how to think about market structure. Essentially, to say anything useful or interesting about the location of economic activity in space, it is necessary to get away from the constant-returns, perfect-competition approach that still dominates most economic analysis. As long as economists lacked the analytical tools to think rigorously about increasing returns and imperfect competition, the study of economic geography was condemned to lie outside the mainstream of the profession. Indeed, as standards of rigor in economics have risen over time, the study of location has been pushed further and further into the intellectual periphery.[1]

1. A major exception is urban economics, where there is a strong modeling tradition that informs a large body of empirical work. Henderson (1974, 1988), in particular, has developed a very persuasive framework for analyzing the evolution of an urban system and has provided extensive empirical evidence in support. I think it is fair to say, however, that international economists have largely ignored or been unaware of this body of work.

Not all students of economic geography have understood this. Much of the literature on industrial location, in particular, has ignored the issue of market structure and instead been obsessed with geometry—with the shape of market areas on an idealized landscape, or with the optimal siting of facilities given markets and resources—while paying little or no attention to the problem of modeling markets. This is, to my mind, doing things in the wrong order, worrying about the details of a secondary problem before making progress on the main issue.

Step back and ask, what is the most striking feature of the geography of economic activity? The short answer is surely *concentration*. Think of the United States: most of the population of a huge, fertile country lives along parts of two coasts and the Great Lakes; within these belts, population is further concentrated in a relative handful of densely populated urban areas. As I will document in the next lecture, these urban areas in turn are highly specialized, so that production in many industries is remarkably concentrated in space.

This geographic concentration of production is clear evidence of the pervasive influence of some kind of increasing returns. And there is the problem. Increasing returns are simply harder to model than constant or diminishing returns. If the increasing returns are purely external to firms, we can still use the tools of competitive analysis; but external economies turn out to be both analytically awkward and empirically elusive. If the

increasing returns are internal to firms, we are faced with the necessity of modeling imperfect competition.

Economics tends, understandably, to follow the line of least mathematical resistance. We like to explain the world in terms of forces that we know how to model, not in terms of those we don't. In international economics, what this meant from Ricardo until the 1980s was an almost exclusive emphasis on comparative advantage, rather than increasing returns, as an explanation for trade.[2] The point was that comparative advantage could be modeled using models that assumed constant returns and competition, which were the tools at hand. The profession simply put those aspects of international trade that could not be modeled that way on one side.

Unfortunately, the evident importance of increasing returns in economic geography is so great that this understandable impulse to focus on what we know how to deal with has led to an avoidance of the subject as a whole. After 1940, in particular, as the expected level of rigor in economic discussion steadily rose, economic geography was simply submerged.

But times have changed. During the 1970s there was a new wave of theory in industrial organization, which provided the economics profession with a menu of models of imperfect

2. For those who worry about definitions, by comparative advantage I mean the general idea that countries trade in order to take advantage of their differences. The increasing returns approach asserts instead that countries trade because there are inherent advantages to specialization, even for initially similar countries.

competition. No one of these models is totally convincing, but they make it possible to write down coherent, rigorous, and often elegant models of economies subject to increasing returns. So increasing returns are no longer something to be avoided or assumed away at all costs. The new intellectual opportunities offered by this revolution in theory have in turn transformed a series of other fields. In international economics the past decade has seen a virtually complete rethinking, with the emergence of a new view in which much trade represents arbitrary specialization based on increasing returns, rather than an effort to take advantage of exogenous differences in resources or productivity.[3] More recently, growth theorists have reintroduced the idea that sustained growth may arise from the presence of increasing returns, and old concepts like the "big push" have regained intellectual respectability.[4] And very recently some macroeconomists have suggested that increasing returns play a crucial role in business cycles.[5]

I believe that the time has come to use the same new tools to resurrect economic geography as a major field within economics. It is no longer the case that the need to model increasing returns makes a field untouchable. Instead, increasing returns are, for the moment at least, actually fashionable. And

3. See Helpman and Krugman 1985 for a survey of most of the concepts of the "new international economics."
4. See in particular Romer 1986, 1987, 1990 and Murphy, Schleifer, and Vishny 1989a.
5. See Hall 1989 and also Murphy, Schleifer, and Vishny 1989b.

so we can now admit to ourselves that space matters and try to bring geography back into economic analysis.

There are three reasons in particular why it is important to start doing economic geography. First, the location of economic activity within countries is an important subject in its own right. Certainly for a large country like the United States, the allocation of production between regions is an issue as important as international trade—and more important than many issues that occupy a much larger part of economists' time. (I have my favorite candidates, but I won't tell you what they are; I have to live with these people for the next thirty years.)

Second, the lines between international economics and regional economics are becoming blurred in some important cases. One need only mention 1992 in Europe: as Europe becomes a unified market, with free movement of capital and labor, it will make less and less sense to think of the relations between its component nations in terms of the standard paradigm of international trade. Instead the issues will be those of regional economics—and it will help if we actually have some interesting regional economics to offer when the time comes.

To my mind, however, the most important reason to look again at economic geography is the intellectual and empirical laboratory that it provides. The "new" trade, growth, and business cycle theories of the past decade have suggested to us a world view of economics that is very different from that of

most pre-1980 theory. Pervasive increasing returns and imperfect competition; multiple equilibria everywhere; an often decisive role for history, accident, and perhaps sheer self-fulfilling prophecy: these are the kind of ideas that are now becoming popular. Yet it is very difficult to produce compelling evidence from trade, growth, and business cycles that this is the way the world really works. I at least am convinced that there is a strong arbitrary, accidental component to international specialization; but not everyone agrees, and the limitations of the data make a decisive test difficult. Paul Romer is convinced that increasing returns play a large role in explaining sustained growth; but not everyone agrees, and even I am agnostic. Robert Hall thinks that increasing returns play a crucial role in business cycles (he argues that a city and a boom are essentially the same thing—one in space, one in time); not everyone agrees, and I for one find this totally implausible (but interesting!).

But when one turns to the location of production within countries, the evidence for what Nicholas Kaldor called "the irrelevance of equilibrium economics" is far more compelling. The long shadow cast by history and accident over the location of production is apparent at all scales, from the smallest to the largest—from the concentration of most U.S. manufacture of wind musical instruments in the tiny town of Elkhart, Indiana, to the fact that a third of the U.S. population still lives within the original thirteen colonies. And this clear dependence on history is the most convincing evidence available that we live

in an economy closer to Kaldor's vision of a dynamic world driven by cumulative processes than to the standard constant-returns model.

What I want to do in this lecture is to offer a first illustration of the importance of economic geography, both as a field in its own right and as a way to see what kind of economy we live in. In particular, I want to show two things: that increasing returns are in fact a pervasive influence on the economy, and that these increasing returns give a decisive role to history in determining the geography of real economies.

I have already suggested that increasing returns affect economic geography at many scales. At the bottom of the scale, the location of particular industries—autos in Detroit, chips in Silicon Valley—clearly often reflects the "locking in" of transitory advantages. At an intermediate level, the existence of cities themselves is evidently an increasing returns phenomenon. At the grand level, the uneven development of whole regions (which in the United States may well be bigger than European nations) can be driven by cumulative processes that have increasing returns at their root.

In this lecture series I will pass over the question of urbanization relatively lightly. It has been better studied than the other issues I will consider (urban economics is more of an accepted field than economic geography), and it is also less relevant than the other aspects to international trade, which remains my ultimate interest. So I will focus on the small and the large:

the localization of particular industries and the differential development of huge regions. Today we look at the large, next lecture on the small.

To introduce the subject of divergent regional development, I turn to economic history to provide a particularly clear-cut example of the forces of economic geography at work. I then offer a simple model that helps make sense of that example. The example is the case of the U.S. "manufacturing belt": a relatively narrow stretch of territory within which the preponderance of U.S. manufacturing was concentrated from the mid–nineteenth century until the 1960s. The model—which is here developed only sketchily—is one in which the interaction of demand, increasing returns, and transportation costs drives a cumulative process of regional divergence.

The Case of the U.S. Manufacturing Belt

Early in this century, geographers noted that the great bulk of U.S. manufacturing was concentrated in a relatively small part of the Northeast and the eastern part of the Midwest—roughly speaking, within the approximate parallelogram Green Bay—St. Louis—Baltimore—Portland (figure 1.1). This "manufacturing belt"[6] took shape in the second half of the nineteenth

6. The term was apparently first used by DeGeer (1927). The belt is not unique, nor are the forces that established it confined to national boundaries. Industrial Canada, concentrated in part of Ontario, is essentially a part of the U.S. manufacturing belt. Continental Europe has a "manufacturing triangle" containing the Ruhr, Northern France, and Belgium that is a close cousin of the U.S. belt.

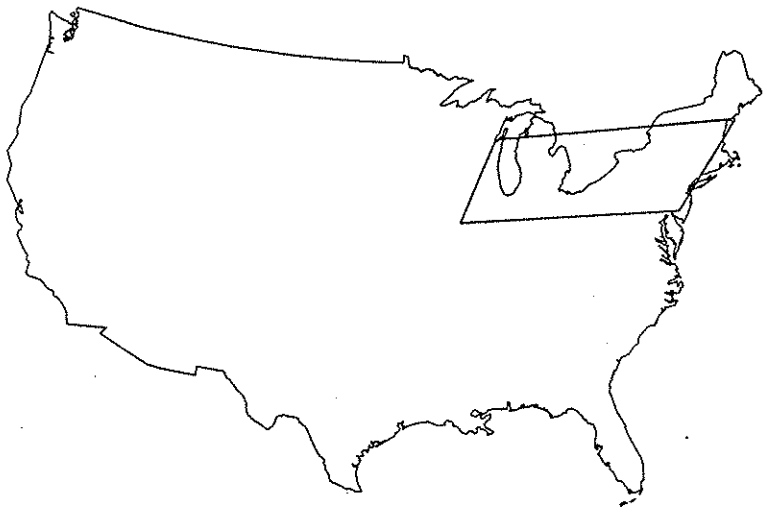

Figure 1.1

century and proved remarkably persistent. Perloff et al. (1960) estimated that as late as 1957 the manufacturing belt still contained 64 percent of U.S. manufacturing employment—only slightly reduced from its 74 percent share at the turn of the century.

Even this number understates the manufacturing dominance of this region, because during the heyday of the belt most of the manufacturing outside it consisted either of processing of primary products or of production for a very local market. That is, the manufacturing belt contained virtually all manufacturing that was "footloose," not tied to other locations either by the need to be very close to the consumer or by the need to use natural resources very close to their source.

Why did the manufacturing belt play such a dominant role for so long? It was clearly not a case of an enduring advantage in natural resources: the manufacturing belt persisted even as the center of gravity of agricultural and mineral production shifted far to the west. In 1870 the Northeast and East North Central regions—within which the emerging manufacturing belt lay—accounted for 44 percent of U.S. "resource extraction" employment (agriculture, mining, forestry, fisheries). By 1910 this share had already fallen to 27 percent; yet these regions still accounted for 70 percent of manufacturing employment. And whereas the belt's share of manufacturing employment understates its manufacturing dominance, its share of resource employment overstates its resource base. The reason is that much of the agriculture that took place within or near to the manufacturing belt was quite different from that outside it: it consisted largely of truck farming and dairying, existing less because of the suitability of the land than because of proximity to the urban centers. In other words, if the manufacturing belt had not existed, the Northeast and Great Lakes areas would have had an even smaller share of agricultural employment.

H. H. McCarty, writing during the belt's heyday, summarized the divergence between regions bluntly: "Outside the manufacturing belt, cities exist to serve the farms; inside, farms exist to serve the cities."

As for mineral resources, the manufacturing belt originally drew some of its critical raw materials from nearby coal mines

and oil wells. By the mid–twentieth century, however, the great bulk of the raw materials for manufactures were imported from other regions.

Why, then, did so much of U.S. manufacturing stay within this relatively small stretch of territory? The answer in broad terms is, of course, obvious: each individual manufacturing facility stayed within the manufacturing belt because of the advantages of being near other manufacturers. And the apparent incentive for manufacturers to cluster together explains the persistence of the manufacturing belt even after the bulk of U.S. primary production had shifted to other regions. Once the belt had been established, it was not in the interest of any individual producer to move out of it.

One may ask why this geographical concentration became established in the first place—a question about historical specifics to which I will return below. First, however, let us ask the more fundamental question: what were the forces that led manufacturers to want to cluster together? I will sketch out a simple model in which geographical concentration results from demand externalities. This surely does not capture the full story, but it is strongly suggestive of the kind of explanation that is needed.

A Model of Geographic Concentration

The basic story of geographic concentration that I will propose here relies on the interaction of increasing returns, transportation costs, and demand.[7] Given sufficiently strong economies of scale, each manufacturer wants to serve the national market from a single location. To minimize transportation costs, she chooses a location with large local demand. But local demand will be large precisely where the majority of manufacturers choose to locate. Thus there is a circularity that tends to keep a manufacturing belt in existence once it is established.[8]

Imagine a country in which there are only two possible locations of production, East and West, and two kinds of production. Agricultural goods are produced using a location-specific factor (land), and as a result the agricultural population is exogenously divided between the locations; for the moment we assume that the division is fifty-fifty.

Manufactured goods (of which there are many symmetric varieties) can be produced in either or both locations. If a given manufactured good is produced in only one location, transportation costs must be incurred to service the other market. On the other hand, if the good is to be produced in both

7. This lecture presents only a sketch of a model. It will be apparent that this sketch is sloppy about a number of issues, including, What is the market structure in manufacturing? What happens to profits, if any? and What resources are used in both fixed costs and transportation? It is possible to derive similar results in a fully specified general equilibrium monopolistic competition model; such a model is presented in appendix A. I adopt the more ad hoc approach here for ease of exposition.

8. In this model I stress the role of demand in determining the location of production of goods that are traded interregionally. An alternative approach would stress the role of increasing returns in the production of nontraded goods, as in Faini 1984. Eventually deciding between approaches will have to be an empirical matter; but for now it is a matter of taste.

locations, an additional fixed setup cost is incurred. The manufacturing labor force in each location is proportional to manufacturing production in that location. Finally, assume that the demand for each manufactured good in each location is strictly proportional to that location's population.

The basic idea can then be illustrated with a simple numerical example. Suppose that 60 percent of a country's labor force are farmers, divided equally between East and West. Suppose also that the total demand for a typical manufactured good is 10 units. Then if all manufacturing is concentrated in one location, that location will demand 7 units (3 demanded by the local farmers, 4 by the manufacturing workers), while the other demands 3; if manufacturing is evenly divided between the locations, each location will offer a local demand of 5.

To figure out what happens, we need to specify the fixed costs and transportation costs; suppose that the fixed cost of opening a plant is 4, and that the transportation cost per unit is 1. Then we have the situation shown in table 1.1. The table shows the costs to a typical firm of three locational strategies, contingent on the locational strategies of all other firms. Thus suppose that all other manufacturing is concentrated in East. Then our firm will have a local demand in East of 7 units, a local demand in West of only 3 units. If it serves the national market from a single plant in East, it will incur a fixed cost of 4 and a transport cost of 3. This is obviously less than serving the national market from a plant in West, which will have the same

Table 1.1
A manufacturing location story

Distribution of manufacturing employment	Costs of typical firm if it produces in	East	Both	West
East only	Fixed	4	8	4
	Transportation	3	0	7
	Total	7	8	11
Fifty-fifty split	Fixed	4	8	4
	Transportation	5	0	5
	Total	9	8	9
West only	Fixed	4	8	4
	Transportation	7	0	3
	Total	11	8	7

fixed cost plus a transport cost of 7; it is also less than building a plant to serve each local market, which saves the transport cost but incurs a double fixed cost of 8. In this case, then, the typical firm will choose to produce in East for a national market.

If each firm concentrates its production in East, however, then manufacturing production as a whole will be concentrated in East—which is what is assumed. So concentration of production in East is an equilibrium.

But it is not the only equilibrium. As the rest of the table shows, if manufacturing is concentrated in West, each firm will similarly also want to concentrate its production in West. And if production is split between East and West, each firm will want

to split its production, too. So in fact all three distributions of production—all in East, all in West, and a fifty-fifty split—are equilibria in this example.

The possibility of multiple equilibria can also be seen graphically (figure 1.2). On the horizontal axis we measure the share of the manufacturing labor force employed in West, on the vertical axis the share of West in the total population. The line *MM* represents the dependence of the distribution of manufacturing on the distribution of population; the line *PP* the converse effect of manufacturing on population distribution.

Let's begin with *PP*. This line represents the relationship between manufacturing labor force employment and total population. Let π be the share of the total population engaged in manufacturing, let s_M be the share of the manufacturing labor force employed in West, and let s_N be West's share of the total population. West is home to half of the farmers, so that at

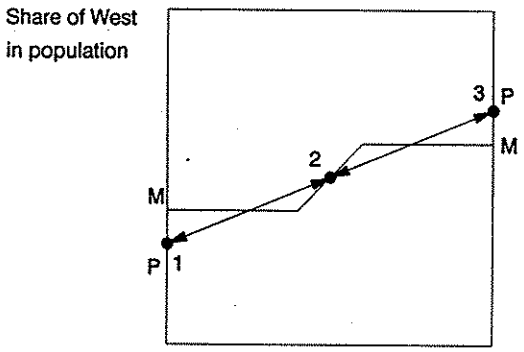

Figure 1.2

minimum it has a population share of $(1-\pi)/2$. The more manufacturing it has, the larger this share:

$$s_N = \frac{1-\pi}{2} + \pi s_M.$$

This is an upward-sloping line that is, however, flatter than a 45-degree line.

Next turn to *MM*. Suppose that West has a very small share of the population. Then it will not be worthwhile incurring the fixed costs of establishing a manufacturing facility there; it is cheaper to serve the market from facilities in East. Conversely, if West has a very large share of the population, it is not worth producing manufactures in East. If the fixed cost is not too large relative to transportation costs, a sufficiently equal division of population will lead manufacturers to produce locally for both markets. Putting these observations together, we get the illustrated shape of *MM*: no Western production for low Western population, production proportional to population for intermediate levels, no Eastern production if the West is big enough. Let x be the sales of a typical manufacturing firm, F the fixed cost of opening a branch plant, and t the transportation cost of shipping a unit of manufactures from East to West or vice versa. Then it is cheaper to service West from a plant in East than to open a Western plant as long as $s_N xt < F$; it is cheaper to service East from West if $(1-s_N)xt < F$; and it is cheaper to have a plant in each region if neither is true.

Provided that fixed costs are not too high relative to transport costs,[9] we therefore have

$$s_M = 0 \text{ if } s_N < \frac{F}{tx}$$
$$= s_N \text{ if } \frac{F}{tx} < s_N < 1 - \frac{F}{tx}$$
$$= 1 \text{ if } 1 - \frac{F}{tx} < s_N.$$

Suppose that manufacturing production adjusts gradually toward its equilibrium level. Then the dynamics are illustrated by the arrows in figure 1.2. There are three stable equilibria: manufacturing may be concentrated in either location, at 1 or 3, or it may be equally divided, at 2. Which equilibrium you get to depends on where you start: history matters.

Of course there need not be multiple equilibria. The concentration of production, if it happens, depends on a demand externality. Manufacturers want to locate where the market is largest; the market is largest where the manufacturers locate. This circularity, however, need not always be strong enough to prevail over the pull of the dispersed agricultural sector. The situation could instead look like figure 1.3: a unique, stable equilibrium with manufacturing equally divided between the two locations.

9. If $F > tx/2$, then it is always cheaper to service both markets from a single plant, even if the population is equally divided. In this case *MM* is simply a horizontal line, and the possibility of an equilibrium with equally divided manufacturing disappears.

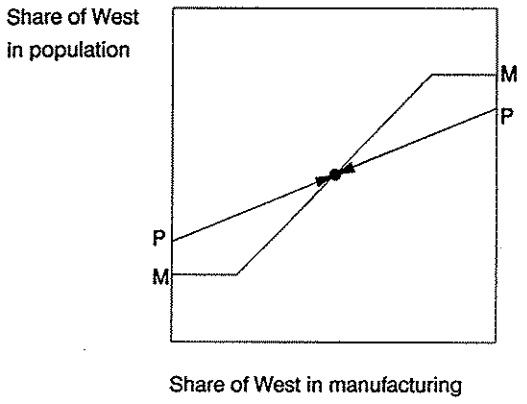

Figure 1.3

We can easily derive a necessary condition for concentration of manufacturing production in one location. With all manufacturing in East, West has a share of total population equal to only $(1-\pi)/2$. The transportation cost of serving this market from East for a typical manufacturer is therefore $tx(1-\pi)/2$. The cost of setting up a plant in West is F. So a concentration of production in East, once established, will persist as long as

$$F > \frac{1-\pi}{2} tx .$$

If this criterion is not met, history does not matter: the geography of manufactures will follow that of agriculture.

We can immediately see that a key role for history depends on three parameters: large F, i.e., sufficiently strong economies of scale; small t, i.e., sufficiently low costs of transportation; and

large π, i.e., a sufficiently large share of "footloose" production not tied down by natural resources.

We can now tell a stylized story of the emergence of the manufacturing belt.[10] In the early United States, with its primarily agricultural population, where manufacturing was marked by few scale economies and where transportation was costly, no strong geographical concentration could occur. As the country began its industrial transition, manufacturing arose in areas that contained most of the agricultural population outside the South—and the South was, for reasons having to do with its uniquely awful institutions, unsuited for manufacturing. During the second half of the nineteenth century, however, manufacturing economies of scale increased,[11] transportation costs fell, and the share of the population in nonagricultural occupations rose. The result was that the initial advantage of the manufacturing belt was locked in. Even though new land and new resources were exploited to the west, even though slavery ended, for three-quarters of a

10. This story is based on the fascinating work of David Myers (1983), who however bears no responsibility for the crudity of the representation.

11. Chandler (1990) provides a fascinationg story of the emergence of large manufacturing firms in the period between the Civil War and the 1920s—that is, during the heyday of the manufacturing belt. He shows that in one industry after another, a "first-mover" led the way by taking advantage of new technology and lower transport costs to build one or two plants of unprecedented size, serving the whole national market. While Chandler does not emphasize the point, his U.S. firms invariably established their first huge plant somewhere inside the manufacturing belt. Sometimes this choice was dictated by the availability of specific resources—for example, hydroelectric power for aluminum smelters at Niagara Falls—but access to markets seems to have played a key role in ruling out sites outside the manufacturing belt.

century the pull of the established manufactured areas was strong enough to keep the manufacturing core virtually intact.

Of course this story oversimplifies in a number of ways. On one side, it probably underemphasizes the role of certain conventional factors in giving rise to the manufacturing belt—there is a suspicious correlation between the location of heavy industry and that of coalfields, both in the United States and in Europe. On the other side, it says nothing about the sources of local specialization within the manufacturing belt—about why Detroit emerged as the automotive center, New York as the garment center, Grand Rapids as the furniture center, etc. Yet it surely captures an important aspect of what happened. And it also contains elements—increasing returns at the level of individual firms, and external economies resulting from the interaction of these firms' decisions—that will reappear as one further elaborates the story.

Before changing the subject, however, there is one particular aspect of the rise of the manufacturing belt that deserves some further elaboration. This is the role of the endogeneity of transport costs themselves.

Transport Networks and Regional Divergence

It is obvious from even a cursory reading of U.S. economic history that part of the advantage of the manufacturing belt arose from the density of the railroad network connecting the region's cities, a density that was itself a product of the region's

manufacturing dominance. This transport network effect deserves a little more attention.

Imagine for a moment a nation with not two but three locations—Center, West, and South—with equal transportation costs between any two locations. Where will a manufacturing firm locate? By analogy with our previous discussion, if one of these locations offers a sufficiently larger local market than the others, and if fixed costs are large enough relative to transport costs, the more populated location will attract a concentration of manufacturing production.

Now imagine a nation with *four* locations: East, Midwest, West, and South. But now suppose that the transportation cost between East and Midwest is much lower than that in other directions. Then in economic terms East and Midwest will in effect form a *single* location. The East-Midwest region will be a more attractive place to locate manufacturing than South or West, even if the individual markets are no bigger, because factories in either place will have better access to the combined market.

But why should transportation costs in one direction be much lower than in others? The most natural answer is that there are economies of scale in transportation itself. A railway or a highway represents indivisible investments, while the frequency of air service and the ability to use large, efficient planes depends on the volume of demand. Suppose that manufacturing production, and hence both demand and supply, is concentrated in East and Midwest. Then there will be a greater volume of transportation between these locations than on other routes. This will mean lower transport costs, which will in turn reinforce the advantage of East and Midwest as locations for production.

It is possible in principle to imagine this transportation network effect as an independent source of geographical concentration of industry—that is, to set up a model in which the local market size effect that is the driving force for our basic model is absent. In practice, of course, the two effects work together. The U.S. manufacturing belt was characterized not only by a denser population but also by a better transport network than any other part of the country, and thus offered much better market access to manufacturers.

Further Thoughts

The case of the U.S. manufacturing belt is of substantial interest in its own right. The rise and persistence of that belt is an important yet much neglected aspect of U.S. economic history. More important than its immediate significance, however, is what the history of manufacturing location says about the nature of our economy in general. And what it says is that increasing returns and cumulative processes are pervasive and give an often decisive role to historical accident.

It is also interesting that the story of the manufacturing belt reaches back to the mid–nineteenth century. It is common to

argue, as Brian Arthur has, that external economies and cumulative processes have become more important in recent decades because of the growing importance of technology. The geographical concentration of manufacturing in the United States took shape, however, long before the dawn of the information age. So it is not simply true that our economy is not now well described by the conventional constant-returns model. It never was.

The Process of Change

The circular relationship in which the location of demand determines the location of production, and vice versa, can be a deeply conservative force, tending to lock into place any established center-periphery pattern. In the case of the U.S. manufacturing belt, the geographical structure of production that happened to exist at the point at which industrialization, factory production, and the railroad came into force remained essentially intact for the next century.

Nothing, however, is forever. Indeed, one of the most interesting things about the type of model sketched in this lecture is what is says about the process of economic change. What I want to do at this point is illustrate two ideas in particular that are suggested by the center-periphery model. First is that while the geographical structure of production may be stable for long periods of time, when it does change it may change rapidly. In fact, a gradual change in underlying conditions can

at times lead to explosive, or more accurately, catastrophic change. Second, change when it comes may be influenced strongly not only by objective conditions but also by expectations—expectations that may be self-fulfilling.

The Logic of Sudden Change

To see how change in the geography of production can sometimes take place abruptly, suppose that instead of being divided equally between the locations, the agricultural labor force is unevenly split, with West initially having a smaller population. The hypothetical position is shown in figure 1.4. *PP* is the initial relationship between manufacturing employment and total population. Although a possible equilibrium exists at 2 in which West would produce manufactures, we suppose that owing to a head start for East we are instead at 1, with no manufacturing production in West.

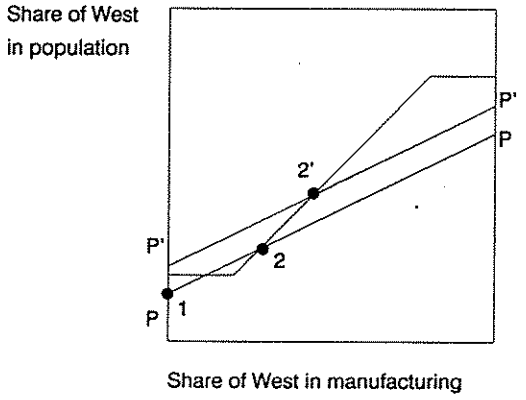

Figure 1.4

Now suppose that there is a gradual reallocation of the agricultural labor force from East to West. This will shift *PP* upward, toward *P'P'*. It is immediately obvious that at a certain point East's dominance of manufactures will collapse. When the Western population reaches a critical mass, it becomes worthwhile for manufacturers to produce there; as manufacturing production in West increases, the population rises further, stimulating still further increases in manufacturing production. A small increase in the agricultural base may therefore set in motion a cumulative process of import substitution and growth, leading eventually to an equilibrium at a point like 2'.

This scenario may not be entirely hypothetical. Paul Rhode (1988) has pointed out that late nineteenth-century California was a resource-based economy with limited manufacturing, largely because the local market was too small to support much industry. He suggests that the discovery of oil around the turn of the century raised California to critical mass, starting it on a process of explosive growth (and in particular causing the rapid emergence of Los Angeles as a manufacturing center).

The point is that the same kind of model that helps explain why history matters also suggests that when change does come, it will often be sudden. And we may also note that changes in regional fortunes will be difficult to predict: in the hypothetical history illustrated in figure 1.4 one would see a

sudden acceleration in West's growth without any obvious reason.

History versus Expectations

Now that I have described the logic of sudden change, let me raise a problem with that logic—a problem that will no doubt already have occurred to those readers with a background in modern macroeconomics, with its emphasis on rationality of expectations. Suppose that the distribution of agricultural population were in fact evolving in the way illustrated in figure 1.4. Wouldn't manufacturing workers and/or firms realize that a sudden increase in West's population was in prospect? And wouldn't they begin to move into West in anticipation of that increase, thereby smoothing out the process of change?

The answer is yes, if they were sufficiently well informed. In practice I have my doubts—here as elsewhere, the assumption of rational expectations seems to presume a degree of information and sophistication that is unreasonable. That is not a controversy that I want to get too far into; I just want to argue that the kind of static expectations that implicitly underlies the dynamics in figure 1.4 retains a useful place in analysis.

But even if I am skeptical about the literal relevance of the assumption of rational expectations, now that we have raised the issue of the role of expectations in regional development,

we should pursue it. For if you think about it a bit, you realize that the kind of circular process that I have argued leads to regional differentiation can also lead to self-fulfilling prophecies.

Imagine again our two-region nation; this time assume for simplicity that the economies of scale are sufficiently large relative to transport costs that there are only two long-run equilibria, with manufacturing concentrated either entirely in East or entirely in West. Suppose, however, that workers cannot all move at once; that there is some kind of adjustment cost that limits the rate at which manufacturing can shift. Thus a worker who chooses to locate in one region or the other is stuck with that choice, at least for a while.

It is immediately apparent that in this case workers will be concerned with more than their current wage—they will base their decisions to move on something like the present value of future wages. But the real wage rates in each region at any point in time depend on the distribution of manufacturing workers; so this means that each worker's current location decision depends on her expectations about the future decisions of other workers.

The possibility of self-fulfilling prophecy now becomes apparent. Suppose that East and West have equal numbers of farmers, and that East initially has somewhat more manufacturing, so that by virtue of its superior forward and backward

linkages East offers a higher real manufacturing wage. Then one might expect to see migration of manufacturing from West to East. But suppose that for some reason the public is convinced that West will be the destination of migration, not East, and that as a result real wages in West will eventually exceed those in East. This belief will induce seemingly perverse migration from the region with higher real wages to that with lower—and this migration will indeed eventually reverse the real wage differential! And if this reversal takes place sufficiently quickly, the worker who migrates from East to West will find that she actually made the right decision. Thus the belief that West is the land of opportunity turns out to be a self-fulfilling prophecy. If everyone had instead had faith in the East, of course, East would have gotten the industry.

When can self-fulfilling prophecy outweigh initial advantage? Several factors clearly matter. First, the rate at which workers and firms can move must be rapid enough relative to the rate at which future wage differentials are discounted that the future advantage of one region can matter more than the current advantage of another. Second, increasing returns must be strong enough that an expected future shift in population distribution moves the real wage differential quickly. Finally, the starting position must not be too unequal: if enough manufacturing is concentrated in one region, this initial advantage may be too much for even the most optimistic expectations about the other region to overcome.

It is possible to formalize the problem of self-fulfilling expectations rather neatly; such a formalization is presented in appendix B. What the formalization tells us is that there may be a range of initial distributions of manufacturing workers from which either region can end up with the manufacturing concentration, depending on expectations. Whether such a range exists, and the size of the range if it does exist, depend crucially on the speed of adjustment; only if the adjustment is slow can we be sure that initial advantage cumulates over time, instead of potentially being overruled by self-fulfilling expectations.

So much for the logic. To what, if anything, does this story correspond in reality? The answer is that I am not sure. In the case of the U.S. manufacturing belt, history clearly determined what happened. Perhaps there could have been a self-fulfilling belief in the industrial future of, say, the Great Plains that could have outweighed the historical advantage of the traditional manufacturing locations; but there wasn't. (And I doubt it.)

At a smaller scale, however, the case for self-fulfillment is better. Certainly a prominent part of the tradition of local economic development in the United States has been boosterism—the sometimes ludicrous efforts by local businessmen and chambers of commerce to convince footloose individuals and firms of the virtues of their state or town, in the belief that if they can draw a critical mass into the local economy, it will become self-sustaining. Some of this boosterism involved concrete incentives, sort of proto-industrial policy; we'll see an example in the story of Akron and the rubber industry in the next lecture. But often it was simply an attempt to create optimism about the locale. The analysis sketched out here suggests that in principle, at least, boosterism may make perfectly good sense.

There is also the possibility of reverse boosterism: if for some reasons businesses and workers become pessimistic about a region's prospects, this pessimism can become self-justifying. It is difficult for me to avoid speculating that something like this may be happening in my own home state. As you may know, two years ago the governor of Massachusetts ran for president, in part on the impressive economic record of his state. He was humiliatingly defeated by George Bush—and the Massachusetts economy itself went into a tailspin. Was this just a coincidence, or did the psychological impact of the campaign, and the political civil wars that followed within the state, create a self-fulfilling downward spiral? (And will the state's economy continue to implode?) I don't know the answer, but such seemingly fanciful ideas don't seem as silly to me as they might to a more conventionally minded economist.

Where We Stand

In this lecture I have tried to argue for the acceptance of economic geography as a major field within economics, on a par with or even in some sense encompassing the field of

international trade. Since we all know that economic argumentation succeeds at least as much on its aesthetics as its empirical support, I have tried to make my case with the cutest model of geography that I have been able to come up with: one that shows how a core-periphery structure can emerge endogenously on a nationwide scale. And I have argued that something like this actually happened in the United States between the Civil War and the First World War.

The phenomenon of concentration in economic geography takes place at many scales, however. While the emergence of huge metropolitan belts may be the most dramatic, for international affairs the forces that lead to localization of particular industries, usually but not always within those belts, are possibly of even more interest. So in the next lecture I will move from the large to the small: from core-periphery to localization.

2 Localization

In 1895 a teenaged girl named Catherine Evans, living in the small Georgia city of Dalton, made a bedspread as a wedding gift. It was an unusual bedspread for the time, in that it was tufted; the craft of tufting or candlewicking had been common in the eighteenth and early nineteenth centuries but had fallen into disuse by that time. As a direct consequence of that wedding gift, Dalton emerged after World War II as the preeminent carpet manufacturing center of the United States. Six of the top twenty U.S. carpet manufacturing firms are located in Dalton; all but one of the rest are located nearby, and the carpet industry in Dalton and vicinity employs nineteen thousand workers.

I'll come back to Catherine Evans and her story later in this lecture. For now let me simply assert that aside from being particularly charming, the carpet story is actually a fairly typical one. To a remarkable extent, manufacturing industries within the United States are highly localized; and when one tries to understand the reasons for that localization, one finds that it can be traced back to some seemingly trivial historical accident.

I'll tell some stories later. First, let's develop an analytical structure.

Sources of Industry Localization

The observation of high localization of industries is, of course, not a new one. Indeed, it was such a striking feature of the process of industrialization that it attracted a great deal of attention in the later nineteenth century, with a fascinating monograph on the subject contained in the 1900 U.S. Census. The literature on industry localization is much too extensive to cite; notable examples over the years include Hoover 1948, Lichtenberg 1960, and very recently, Porter 1990. There is also considerable overlap between the subject of industry localization and urban economics; theory, anecdotal evidence, and solid empirical work can be found in such works as Bairoch 1988, Jacobs 1969 and 1984, and Henderson 1988.

Let us, however, go back to the source. It was Alfred Marshall who presented the classic economic analysis of the phenomenon. (Actually, it was the observation of industry localization that underlay Marshall's concept of external economies, which makes the modern neglect of the subject even more surprising.)

Marshall (1920) identified three distinct reasons for localization. First, by concentrating a number of firms in an industry in the same place, an industrial center allows a pooled market for workers with specialized skills; this pooled market benefits both workers and firms:

> [A] localized industry gains a great advantage from the fact that it offers a constant market for skill. Employers are apt to resort to any place where they are likely to find a good choice of workers with the special skill which they require; while men seeking employment naturally go to places where there are many employers who need such skill as theirs and where therefore it is likely to find a good market. The owner of an isolated factory, even if he has good access to a plentiful supply of general labor, is often put to great shifts for want of some special skilled labor; and a skilled workman, when thrown out of employment in it, has no easy refuge.

Second, an industrial center allows provision of nontraded inputs specific to an industry in greater variety and at lower cost:

> [S]ubsidiary trades grow up in the neighborhood, supplying it with implements and materials, organizing its traffic, and in many ways conducing to the economy of its material...the economic use of expensive machinery can sometimes be attained in a very high degree in a district in which there is a large aggregate production of the same kind, even though no individual capital employed in the trade be very large. For subsidiary industries devoting themselves each to one small branch of the process of production, and working it for a great many of their neighbors, are able to keep in constant use machinery of the most highly specialized character, and to make it pay its expenses. . .

Finally, because information flows locally more easily than over greater distances, an industrial center generates what we would now call technological spillovers:

> The mysteries of the trade become no mystery; but are as it were in the air. . . . Good work is rightly appreciated, inventions and improvements in machinery, in processes and the general organization of the business have their merits promptly discussed: if one man starts a new idea, it is taken up by others and combined with

suggestions of their own; and thus it becomes the source of further new ideas.

(On the whole, I like Marshall's turn of phrase better than the modern one!)

Cut through the archaism of Marshall's language and his lack of formalism, and you will see that he had a pretty sophisticated model in mind. He missed a few tricks that I will try to point out, but on the whole the main purpose of this part of the lecture is to rephrase Marshall in a drier, less felicitous style, and thereby bring it up to date.

Let us therefore consider in turn each of the Marshallian reasons for localization.

Labor Market Pooling

Imagine for a moment that there is some industry that consists of just two firms, each of which can produce in either of only two locations. These firms both use the same distinctive kind of skilled labor. For whatever reason, however, the firms' demands for labor are not perfectly correlated. For example, they may produce differentiated products that face uncertain demand; or they may be subject to firm-specific production shocks. Whatever the reason, the labor demand of the firms is both uncertain and imperfectly correlated.

To make matters more concrete, suppose that each firm may experience either "good times," in which it would like to hire 125 specialized workers at the going wage, or "bad times," in

which it would like to hire only 75. We also suppose that there are 200 of these workers in total, so that average demand for labor equals supply. (In this example I take the wage rate for the specialized labor as given, so that there may be either excess demand or excess supply for labor. If you like, imagine a wage bargaining process that sets the wage at an expected market-clearing level before the shocks to labor demand are revealed—not too unrealistic an assumption. This assumption is not, however, essential. I will show in a moment that even if the wage rate is completely flexible and the labor market clears, the basic story remains the same.)

Now we may ask: will firms and workers be better off if the two firms choose different locations—each thus forming a company town with a local labor force of 100—or if the two firms choose the same location, with a pooled labor force of 200 that can work at either firm?

You may immediately ask about the possibility of exploitation: wouldn't each firm prefer to have a captive local labor force? I'll come back to that soon and show that it doesn't work the way you may think. For now, just put it aside and assume that the wage is set at an expected market-clearing level. Then it should immediately be apparent that it is in the interest of both firms and workers that everyone be in the same place.

First, consider the situation from firms' point of view. If each has its own town, with a labor force of 100, then it will be unable to take advantage of its good fortune when labor demand is high: during good times, there will be an unfillable

excess demand of 25 workers. If the two firms are in the same place, however, then at least occasionally one firm's good times will coincide with the other firm's bad times, and there will be additional workers available.

Next consider the situation from workers' point of view. If they live in a company town, then the firm's bad times are their bad times too: whenever the firm has low labor demand, 25 workers will be laid off. If the firms are in the same place, then at least sometimes one firm's bad times will be offset by the other firm's good times, and the average rate of unemployment will correspondingly be lower.

This is a pretty trivial example. Yet I think it is useful as a way to clear up some points that are often misunderstood. First, the example clarifies the nature of the gains from labor market pooling. Because of the word "pooling" some people may be tempted to assume that the incentive to create a pooled labor market is something like portfolio diversification, that is, that it has something to do with risk aversion on the part of workers. No doubt minimizing risk is also an issue, but I didn't mention it in the example. So even if workers were entirely risk-neutral, there would be an efficiency gain from creating a localized industry with a pooled labor market.

Second, if you think about it, the example shows that uncertainty alone won't generate localization. You need increasing returns as well. The key point is that in order to make a pooled labor market advantageous, I needed to assume that each firm had to choose one location or the other, not both. If each firm

could produce in both locations, or for that matter if each firm could be split into two identical firms, one in each place, then the full "portfolio" of firms and workers could be replicated in each location, and the motivation for localization would be gone. But the most natural justification for the assumption that firms do not locate in both places is that there are sufficient economies of scale to militate for a single production site.

So it is the *interaction* of increasing returns and uncertainty that makes sense of Marshall's labor pooling argument for localization. But as I have described it so far, it is only an argument for the advantage of concentrated production, not a description of the process that might bring about such concentration. Can we provide such a description?

Consider figure 2.1 (which bears more than a coincidental resemblance to some of the figures from the last lecture). I envision an industry in which there are a fixed number of firms

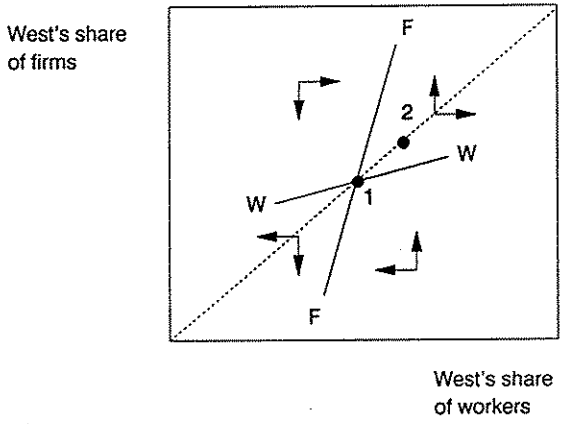

Figure 2.1

and a limited supply of specialized workers. Each firm and each worker must choose one of two locations, East or West. On the horizontal axis I show West's share of the labor force, on the vertical axis its share of the number of firms.

Which location will firms and workers prefer? The curves *FF* and *WW* show which distributions of firms and workers will leave the typical firm and worker respectively indifferent between the two locations. Both curves cross the 45-degree line at point 1, in the middle of the box: if everything is the same, everything is the same.

For any given labor force, firms would rather face less competition for the available workers. Thus an increase in the share of firms in West will make that location less attractive for the typical firm, unless this increase is offset by an increase in the labor force as well. So the set of combinations of share of firms and share of labor force that leaves firms indifferent between the locations, shown by *FF*, is upward sloping.

Workers would prefer to share the demand of any given number of firms with as few other workers as possible; an increase in the labor force in West will therefore make it a less attractive location for workers, unless offset by an increase in the number of firms. So the set of combinations of share of firms and share of labor force that leaves workers indifferent between East and West, shown by *WW*, is also upward sloping.

As drawn, *FF* is steeper than *WW*. To see why, consider point 2. At that point, the ratio of firms to workers is the same in West and East; but West has more of both. This means that West will offer better labor market pooling, and hence that at point 2 both firms and workers will prefer West to East. So point 2 is below *FF* (West is preferred by firms) and above *WW* (West is preferred by workers). This is only possible if at point 1 *FF* is steeper than *WW*.

Now ask what happens if we are off these schedules of equal attractiveness. Presumably firms move to the more attractive location, as do workers. The resulting dynamics are illustrated by the arrows in figure 2.1. There are three equilibria; but the one in the middle is knife-edge unstable. So we will converge to concentration of both firms and workers either in East or West, depending on initial conditions.[1]

This is a rough formalization of Marshall's labor pooling argument. It has many loose ends, of course; let me try to tie up a few of them.

Additional Thoughts on Labor Pooling

The biggest loose end in the story just told—a loose end that is flapping free in Marshall as well—is in the description of wage determination. In fact, this is a double issue (a split loose end?). First, how dependent is the argument on the assumption of sticky wages and a non-clearing labor market? Second, what

1. Obviously there is an issue of self-fulfilling prophecy, of expectations as opposed to history, in models of localization as well as in the more macro model of the previous lecture. I will take it as understood that this issue is lurking in the background, and proceed with ad hoc dynamics that ignore the expectational problem.

about the issue of exploitation—the advantage to firms of a monopsony position in a company town?

A special example can help demonstrate that even with flexible wages it is advantageous to have a pooled labor force. Return to our two-firm example, but now suppose that each firm has a downward-sloping demand for labor rather than a simple number of workers that it wants. Let wages be flexible, and ignore the potential for monopsony. Then in the case where the firms locate in different places there will no longer be unemployment or excess demand for labor; instead, the local wage rate will fluctuate.

If the firms formed a pooled labor market, however, the wage rate would fluctuate less. Risk-averse workers would like this. But beyond this, firms would find that their profits were higher. To see this, take the extreme case in which the labor demands of the two firms are perfectly negatively correlated, so that the wage rate in the pooled case does not fluctuate at all. The average wage rate would be the same[2]; but this does not leave the firms indifferent.

Consider figure 2.2. It shows the "good times" and "bad times" labor demands of one of the two firms. If the firm is isolated, it pays a high good-times wage and a low bad-times wage, while always employing the same number of people; if it is part of a pooled market, it pays a constant wage that is the average of the two. But it now employs more workers in good

2. Actually, to leave the average wage rate unchanged we must assume both linearity of the labor demand schedules and additivity of the shocks; but the basic point goes through regardless, as shown in appendix C.

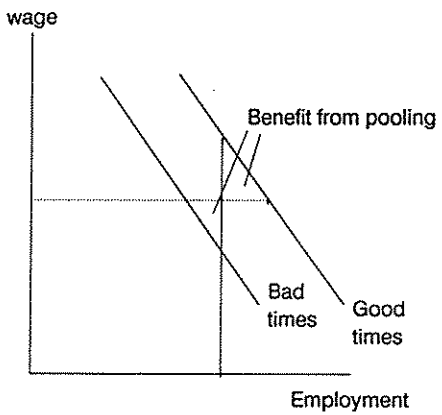

Figure 2.2

times, fewer in bad. As a result, the gain to the firm from the lower wage it pays in good times is greater than the loss it suffers from the higher wage it pays in bad times. In the figure, the net benefit is measured by the sum of the areas of the two indicated triangles.

This is an extreme example, but it makes a more general point. The gains from labor pooling do not rest in any essential way on a failure of labor markets to clear.

What about monopsony power? Won't firms prefer to have a captive labor force that they can exploit? Yes, other things equal. But other things won't be equal—and in fact the monopsony issue actually acts as a further reason for localization.[3]

3. This point was first made by Julio Rotemberg and Garth Saloner (1990). The style of their model is very different from what I present here, but essentially I am just offering a variant of their analysis.

To see why, now let us eliminate the uncertainty motive for pooling, and suppose that firms have known labor demand schedules. They will ordinarily, however, be off these schedules, because they will restrict their hiring in an effort to keep wages down. The extent of this restriction will depend on the degree of competition among the firms, which will presumably depend on how many firms there are in each location.[4] So an equal proportional increase in the number of firms and workers will benefit workers, hurt firms. So in figure 2.3, which has the same format as figure 2.1, both *FF* and *WW* are shown as flatter than the 45-degree line.

But I show *FF* as steeper than *WW*. That is, an increase in West's share of the labor force, offset by a large enough increase in its share of firms to leave workers indifferent between East and West, will make West a more attractive location for firms.

To see why, compare points 1 and 2 in the figure. At point 1, which is the middle of the box, both wage rates and the profits of a typical firm are the same in East and West. At point 2, where West has a larger labor force offset by a larger number of firms, wage rates are still the same— so workers are still indifferent. However, firms are more profitable in West, even though they pay the same wage, because they engage in less restriction of hiring in order to keep the wage down. Because

4. Formulating the nature of competition in the labor market is tricky. Rotemberg and Saloner assume Bertrand competition, so that the wage goes from the reservation wage to the competitive level when the number of firms goes from one to two. This seems a little extreme; as appendix C points out, however, it is difficult to come up with an equally elegant alternative. For the purposes of this exercise I will simply assume that exploitation goes down as the number of firms rises.

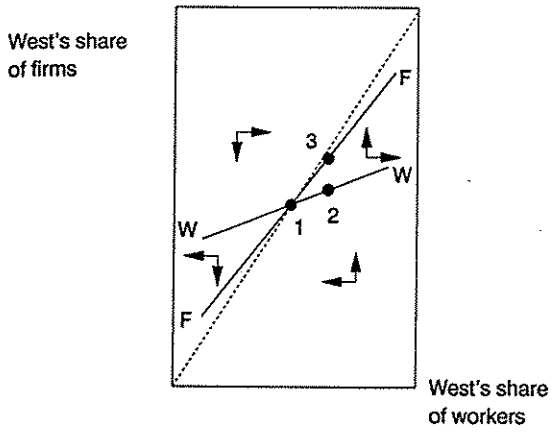

Figure 2.3

the number of firms has not risen in proportion to the number of workers, each firm has larger employment; since workers are paid less than their marginal product, this means higher profits. So in order to leave firms indifferent between West and East, it would be necessary to offset the increase in West's labor force by a larger increase in the number of firms, for example at point 3. Thus we know that *FF*, while flatter than the 45-degree line, is steeper than *WW*.

Once again the dynamics are illustrated by the arrows; and we see that the equilibrium in the middle is unstable, and the industry ends up concentrating either in East or in West. But in this example we have assumed away uncertainty, so that there are no pooling gains from concentrating the industry. Instead, the concentration of the industry is driven entirely by the issue of market power in the labor market. The intuitive argument that says that firms prefer company towns, because of the monopsony power this gives them, is exactly wrong.

Why does it turn out this way? One way to think about it is the following: there is a tug-of-war between firms, who prefer a less competitive labor market and hence production in both locations, and workers, who prefer a more competitive market and hence concentration in one location. Workers win this tug-of-war because a more competitive labor market is also more efficient; given noncollusive choice of location by firms, this efficiency gain decisively tips the scales toward the concentrated solution.

An alternative, and perhaps deeper, way to view the issue is in terms of credibility. Firms would like to convince workers that they will *not* try to exploit their monopsony power, so that they can attract workers to their production location.[5] But the only credible way to do this is to have enough firms in the location that there is an assurance of competition for workers. The commonsense idea that firms would like to have a company town in which workers could be exploited is right; but the point is that workers will shun such towns if they can, so that firms will end up finding it more profitable to locate in agglomerated centers that are not company towns.[6]

5. In Rotemberg and Saloner's formulation workers are not mobile, but immobile workers must choose whether or not to invest in industry-specific human capital. The basic principle is the same.

6. Of course company towns do exist. Impressionistically, they seem to happen for one of two reasons. First, there may be specific natural advantages that cause individual factories to be located on scattered sites, as was the case for water-driven New England textile towns. Second, the economies of scale in an industry may be so large that a single firm dominates the industry and agglomerates its plants in order to achieve pooling, like Eastman-Kodak in Rochester or Boeing in Seattle.

So it turns out that the loose ends that Marshall left in his labor market story do not invalidate it. Allowing for the possibility of flexible wage rates, while eroding some of the simplicity of the labor pooling story, does not make it go away; allowing for the possibility that firms may try to exploit monopsony power actually makes the story stronger.

Intermediate Inputs

Marshall's second reason for agglomeration, the availability of specialized inputs and services, seems straightforward enough. A localized industry can support more specialized local suppliers, which in turn makes that industry more efficient and reinforces the localization.

There are two points, however, that could perhaps use some clarification. First, the intermediate inputs story, like the labor pooling story, depends crucially on at least some degree of economies of scale. If there were no economies of scale in the production of intermediate inputs, then even a small-scale center of production could replicate a large one in miniature and still achieve the same level of efficiency. It is only the presence of increasing returns that makes a large center of production able to have more efficient and more diverse suppliers than a small one.

Second, the intermediate inputs story does not depend on some asymmetry in transportation costs between intermediate and final goods. Weber-type stories of transport cost minimization may suggest that localized industrial complexes

will emerge only if it is more costly to transport intermediate inputs than final goods. This is an impression that may have been reinforced by some models in trade, unfortunately including the work of Elhanan Helpman and myself (1985). In our models of international trade we contrasted the case where intermediate goods are tradable but final goods are not—which gave rise in effect to external economies at the level of the world rather than the individual country— with the reverse, which led to the formation of national industrial complexes. The reason for stressing these two extreme cases was convenience: it is easier to model a good that is either perfectly tradable or perfectly nontradable than one that can be traded, but only at a cost. Unfortunately, this approach can convey the impression that localization due to the clustering of suppliers occurs only in the special case in which transportation costs for intermediate goods are particularly high.

This is a misleading impression. In fact, localization will tend to occur unless the costs of transporting intermediates are particularly *low* compared with those of transporting final goods. And a general reduction in transport costs, of both intermediates and final goods, will ordinarily tend to encourage localization rather than discourage it.

To see why, it is useful to consider a tricky sort of model in which intermediates and final goods are the same thing. Imagine a group of products each one of which is demanded both as a final good and as an input into all of the others. For example, suppose that the typical product in the group has total sales of 10, but that 4 of these sales are to manufacturers of other products in the group. And correspondingly we must

suppose that to produce these 10 units requires 4 units of intermediate inputs, which again are drawn from the same industry. Notice that by making each good both a final and an intermediate, I have a fortiori imposed symmetry among intermediates and final goods in terms of their tradability.

Now suppose that there are two possible locations of production, each of which is also the location of half the final demand, that is, 3 units of each product. Where will a firm want to locate? The answer obviously depends on the decisions of other firms. If everyone else is in East, then 7 of the 10 units of total demand will be in East (3 final plus 4 intermediate); this will provide a firm with an incentive to locate its own production in East as well. The incentive will be reinforced by the fact that all of the firm's supplies of intermediate goods will come from East and will therefore be cheaper there. Thus there will be both backward and forward linkages that provide an incentive to concentrate production. Of course there will also be an incentive to move closer to final demand that will pull the other way.

This should be sounding familiar: it sounds an awful lot like the core-periphery model sketched out in the first lecture and derived formally in appendix A. And indeed it is possible to construct a model of intermediate goods and industry localization that is formally exactly analogous to the core-periphery model. This is helpful, because we already know something about that model. In particular, we know that the prospects for formation of a core-periphery pattern depend negatively on transportation costs, positively on the share of "foot-

loose" demand, and positively on the importance of economies of scale. The same should be true here, if these variables are suitably reinterpreted. In particular, the role played by the share of manufactures in the core-periphery model is here taken instead by the share of the industry's output that is used as an intermediate good rather than directly for final demand. Given this, we see that lower transport costs make industrial localization more likely, even if the cost of transporting intermediates falls along with that of transporting final goods.

It is also interesting to note that the same historical forces that in the last lecture I argued gave rise in the nineteenth century to the emergence of geographical concentration at a macro level—falling transportation costs, industrialization, and growing economies of scale—also should have led to increased localization of industry within the manufacturing belt. The only difference is that instead of an increased share of manufactures per se in demand, the necessary change is an increased share of manufactures that are used as inputs into other manufactures. But this did of course happen, as the "industrial ecology" of the U.S. economy became more diverse and complex. So it is not surprising that localization of industry, the emergence of sharply distinct industrial characters for particular cities or districts, became a striking characteristic of the American economy by the end of the century.

Technological Spillovers

I have saved for last the reason for localization that many economists would put first—namely, the more or less pure externality that results from knowledge spillovers between

nearby firms. The emphasis on high technology in much policy discussion and the fame of such clusters as California's Silicon Valley and Boston's Route 128 have made technological externalities the most obvious thing to mention. Furthermore, economists with a conventional background still have a hankering to preserve perfect competition in their models; purely technological externalities do this.

Yet I have chosen to put pure technological external economies last, not first, for several reasons. First, it is an empirical fact that many of the industries that are highly localized within the United States now or were highly localized in the past are nothing like high technology sectors. Silicon Valley is famous; but equally remarkable concentrations may be found of carpet producers around Dalton, Georgia; of jewelry producers around Providence, Rhode Island; of financial services in New York; and historically, of such industries as shoes in Massachusetts or rubber in Akron. Evidently forces for localization other than those involving high technology are quite strong.

Second, as a matter of principle I think we should try to focus first on the kinds of external economies that can be modeled other than by assumption. Labor pooling or intermediate goods supply are things that in principle one could examine directly and predict given a knowledge of the technology of the industry. And on the other side, the concreteness of these forces places constraints on what we can assume. Knowledge flows, by contrast, are invisible; they leave no paper trail by which they may be measured and tracked, and there is nothing to prevent the theorist from assuming anything about them

that she likes. A sociologist might be able to help with survey methods; but I would like to get as far as possible with drab, down-to-earth economic analysis before turning to the other social sciences.

Finally, high technology is fashionable, and I think we are all obliged to make a deliberate effort to fight against fashionable ideas. It is all too easy to fall into a kind of facile "megatrends" style of thought in which the wonders of the new are cited and easy assumptions are made that everything is different now. Of course the world has changed—but it was a pretty remarkable place even before the coming of large-scale integrated circuits, and even high technology industries respond to old-fashioned economic forces.

So while I am sure that true technological spillovers play an important role in the localization of some industries, one should not assume that this is the typical reason—even in the high technology industries themselves.

Some Empirical Evidence

I have alluded to some facts about the degree of localization of industry within the United States. But what sort of facts do we actually have? One kind of evidence, which I regard as very important in spite of its lack of rigor, consists of case studies. Nothing is better at suggesting the kind of model that we ought to use than a collection of examples, particularly with some historical depth, of how particular industries come to be

in particular places. And the stories are often entertaining too. Before turning to stories, however, I want to discuss some very preliminary statistical work that I have undertaken.

The objective of this work was to answer two questions. First, how localized is the "typical" U.S. industry? Are familiar examples like that of the auto concentration around Detroit or high technology in Silicon Valley normal, or are they outliers?

Second, what sorts of industry are highly localized? Are localized industries typically high technology sectors (which would lend support to a technological spillover model), or are they more prosaic? Are they industries that use highly skilled labor or more general-purpose workers?

The technique I have tried to get at these questions is to construct "locational Gini coefficients" for as many U.S. manufacturing industries as possible. A locational Gini curve for an industry is constructed as follows. First, for each of the locational units in our sample we calculate both the share of total national manufacturing employment and the share of national employment in the industry. Then we rank the units by the ratio of these two numbers. Finally, we run down the ranking, keeping a cumulative total of both the sum of total employment share and the sum of employment share in the industry.

Suppose, for example, that there were three regions. Region 1 has 20 percent of total manufacturing employment, but 50 percent of employment in the widget industry. Region 2 has 40 percent of total manufacturing employment, and 40 percent of

widget employment. Region 3 has the remaining 40 percent of manufacturing employment, and the remaining 10 percent of widget employment. Then the resulting curve would be the one illustrated in figure 2.4. Clearly, the more the geographical distribution of the widget industry matches that of overall manufacturing, the closer this curve will lie to the 45-degree line. So there is an obvious index of localization: the area between the curve and the 45-degree line. An industry that was not localized at all, but simply spread out in proportion to overall employment, would have an index of 0; one that is concentrated almost entirely in a region with small overall employment would have an index close to 0.5.

In implementing these indices, I (or rather my research assistant) used data on U.S. three-digit industries, classified by states. This is only a preliminary step; although it is illuminating, there are a number of problems that blur the results.

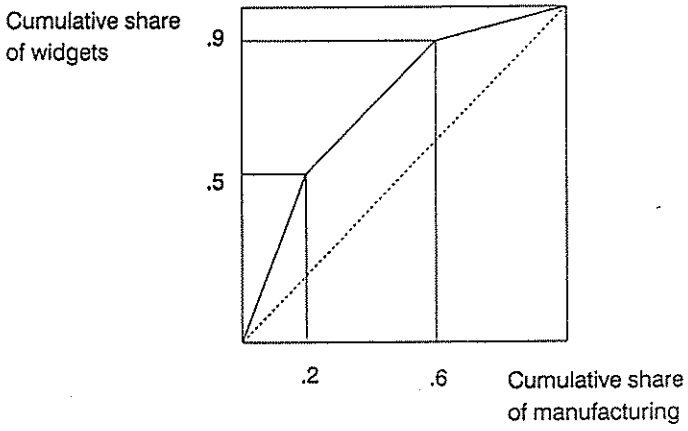

Figure 2.4

First, there are some important holes in the data. The U.S. Bureau of the Census does not reveal information that would compromise the assurance of confidentiality; perhaps surprisingly, this affects data even at the level of large states and big industries. For example, data is withheld on the aircraft industry in Washington (so as not to reveal too much about Boeing) and about the photographic equipment industry in New York (so as to protect Kodak). These are major examples of industry localization, indeed classic ones. Unfortunately, they must simply be dropped from the statistical analysis.

Second, the definition of an industry is problematic. Three-digit industrial classifications are not useless, but they are far from ideal. Old industries, like costume jewelry, rate a three-digit category in spite of very modest employment, while both Silicon Valley and Route 128 are buried inside the huge category of "electronic components." Ideally we would disaggregate and then reaggregate in order to get more meaningful comparisons—but I haven't.

Finally, states aren't really the right geographical units. First of all, they are of very unequal population, which biases industry comparisons: an industry that is completely concentrated around a small town in California will not generate as high a locational Gini as a comparably concentrated industry in Rhode Island. Second, economic regions do not respect state boundaries. The textile complex that shows up very strongly in the data occupies the Piedmont area of the Southeast; state comparisons do not really reveal the compactness (in both the lay and the mathematical senses) of this industry's location.

In spite of all these limitations, we can learn a fair amount from even this preliminary calculation. The results for 106 three-digit industries are reported in appendix D, which also shows total employment in the industry and employment in the three leading states.

The first impression that emerges from the results is that many industries are indeed highly concentrated geographically. The automotive industry offers a useful benchmark. It is a famously localized industry. Although there has been some dispersal since the heyday of Motown, half of the employment is still in the traditional automotive district of southern Michigan and neighboring regions of Indiana and Ohio. So we might expect motor vehicles to be an exceptional industry. But it isn't. It is just slightly above the median. The point is not that automotive production is not highly localized—it is. But so are a lot of other industries.

The other feature of the data that stands out is that the most highly concentrated industries by this admittedly crude calculation are not cutting-edge, high technology sectors. Indeed, the thing that leaps out from the table is not the localization of high technology industries but the cluster of textile-related industries, all of them in more or less the same place: the Piedmont area of the Carolinas and Georgia. Half of the top twenty industries according to my Gini ranking are Piedmont textile sectors. (As we will see in a moment, there is considerable localization of particular industries in this group within the Piedmont; but leave that aside for a moment.)

There is probably a bias in the way this table was constructed against finding the localization of high-tech industries. For one thing, two high-tech industries that are famously geographically localized had to be excluded from the table because of withheld data: aircraft, dominated by Boeing's huge Seattle facilities, and photographic equipment, with Kodak's Rochester complex. These are, of course, industries that are concentrated in ownership as well as in geography.

More important, the fact that the classification scheme is so antiquated means that quite small traditional industries still rate their own three-digit codes, while advanced sectors are buried in meaningless aggregates. Silicon Valley and Route 128 are real enough, but you just can't find them in the statistics.

So I don't want to be misinterpreted. This evidence does not show that high technology industries are not localized. What it shows is simply that low technology industries are also localized. Whatever drives industries to concentrate in one place, it is not solely a matter of technological spillovers.

This is about as far as I want to take this statistical exercise in this lecture; I'll come back to some U.S. regional data in the next lecture. But now I want to get to the fun part and tell stories.

Case Studies: Some Historical Examples

One of the Piedmont textile industries that I found to be highly concentrated geographically was the carpet industry. My

research assistant did some follow-up work on this industry and uncovered a classic case of the role of historical accident and cumulative processes in generating localization. So let's go back to Catherine Evans and her bedspread.

As I mentioned at the beginning of this lecture, in 1895 the teenaged Miss Evans made a bedspread as a gift. The recipients and their neighbors were delighted with the gift, and over the next few years Miss Evans made a number of tufted items, discovering in 1900 a trick of locking the tufts into the backing. She now began to sell the bedspreads, and she and her friends and neighbors launched a local handicraft industry that began selling items well beyond the immediate vicinity.

The handicraft industry became semimechanized in the 1920s, as tufting was used to satisfy the surging demand for chenille sweaters; but production continued to be done by individual households.

Immediately after World War II, however, a machine was developed for producing tufted carpets. Until that time, machine-made carpets had been woven. Tufting proved to be far cheaper. And guess where one could find people who knew about tufting and were quick to see the potential? In the late 1940s and early 1950s, many small carpet firms sprang up in and around Dalton, together with a cluster of supporting firms providing backings, dyeing, etc. Existing carpet manufacturers initially clung to weaving; they eventually either went out of business, driven out by the upstarts from Dalton, or moved their own operations from traditional sites in the Northeast to

Dalton. And so the little Georgia city emerged as America's carpet capital.

It's a lovely story; it's also a very typical one. The whole process of industrialization within the United States was marked by similar stories of small accidents leading to the establishment of one or two persistent centers of production. Anyone who thinks that Silicon Valley is a distinctly modern sort of creation should look at the fascinating monograph, "The Localization of Industries," contained in the 1900 U.S. Census. The monograph identifies fifteen highly localized industries in 1900, including: collars and cuffs, localized in Troy, New York; leather gloves, localized in the two neighboring New York towns of Gloversville (sic) and Johnstown; shoes, in several cities in the northeastern part of Massachusetts; silk goods, in Paterson, New Jersey; jewelry, in and around Providence, Rhode Island; and agricultural machinery, in Chicago.

In each of these cases there is a story similar to, if not quite as charming, as that of Catherine Evans. An accident led to the establishment of the industry in a particular location, and thereafter cumulative processes took over. The Massachusetts shoe industry owed its start to the Welsh cobbler John Adams Dagyr, who set up shop in 1750; the dominance of the Providence jewelry industry (which still makes it onto our list of highly localized industries!) began when a local man invented "filled" gold in 1794; the reign of Troy as the detachable collar and cuff center (alas, fashion!) was inaugurated by a Methodist minister in the 1820s.

What is important to the economist here is, of course, not the initial accident but the nature of the cumulative process that allowed such accidents to have such large and long-lasting effects. What the historical record shows us are two things. First, such cumulative processes are pervasive; Silicon Valley is not at all unique, either in time or space, but is simply a glitzy version of a traditional phenomenon. And second, Marshall's first two reasons for localization, labor pooling and the supply of specialized inputs, play a large role even when pure technological externalities seem unlikely to be important.

Has the basis of localization shifted over time? The authors of the monograph in the twelfth census thought that it would. They noted that many of the highly localized industries were ones that relied heavily on skilled handwork, and speculated that the arbitrary nature of localization would tend to be eroded because "The use of machinery has . . . tended to lessen the importance of a specially skilled labor supply. In proportion as an industry becomes automatic, its localization becomes independent of its supply of special labor."

To some extent they were right. First, it is doubtless true that each manufacturing industry, as it matures, tends to become less dependent on the pooled labor market, specialized inputs, and information spillovers that sustain localization. Consider, for example, the U.S. tire industry. Before 1930 this industry was spectacularly localized in the city of Akron, to which financial incentives from the chamber of commerce (industrial policy!) had lured one Benjamin Franklin Goodrich. As Detroit boomed, the rubber center in Akron came to have more than

a hundred firms, generating some of the highest wages in the United States and attracting migrants from all over (including my grandfather). Time was not, however, on Akron's side. As tire production became standardized, it could be delocalized and moved closer to the market. The market itself spread out over time as the auto industry began to establish assembly plants around the United States.

As it turned out, the end was rather sudden. The Depression dealt a devastating blow to Akron's economy and apparently destroyed the critical mass of rubber firms there: when the country emerged, the role of Akron as a rubber center was gone. No major producers of tires are now located in the one-time tire capital of the world.

But while the Akron story shows that localization within an industry tends to fade away, there are always new industries. Detroit fades, but Silicon Valley rises. Indeed, surely there is a kind of product cycle, in which emergent new industries initially flourish in localized industrial districts, then disperse as they mature.

High Technology Clusters

In the last generation, the familiar examples of localization have changed. One rarely now hears about Motown, or Iron City, or the garment district (although they still exist in somewhat attenuated form). Instead it is all high tech: Silicon Valley, Route 128, Research Triangle. How do these new clusters compare with the older localization?

The first thing to say is that on the whole the stories of their founding are less romantic. In general the new high technology clusters were the product less of intrepid individuals than of visionary bureaucrats (if that is not an oxymoron). But otherwise the stories look rather similar.

Silicon Valley was created largely through the initiative of Fred Terman, the vice-president of Stanford University. Through his initiative the university provided an initial stake for Hewlett-Packard, which became the nucleus of the Valley. It also established the famous research park on university land, on which first Hewlett-Packard, then many other firms, began operations. There was a noticeable cumulative process operating through the university itself: the revenues from the research park helped to finance Stanford's ascent to world-class status in science and engineering, and the university's rise helped make Silicon Valley an attractive place for high-tech business.

Route 128 was created, in a more diffuse way, through the initiative of MIT's president, Karl Compton, who encouraged MIT faculty to become entrepreneurs and helped mobilize private venture capital.

North Carolina's Research Triangle, finally, was created through state support of a research park, in direct emulation of Silicon Valley and Route 128.

Perhaps the most important thing to emphasize in these high technology stories is the importance of non-high-technology factors in the agglomerative process. Both in Silicon Valley

and around Route 128 a key advantage is the existence of a pool of people with certain skills. In the Boston area, for example, growth companies in the software field can be reasonably sure of being able to find people with esoteric knowledge in a variety of sub-subdisciplines. At the same time, the Boston area has been a good place for people to invest in acquiring these skills, or for those with those skills to live: if a start-up goes bust, as many do, you can find another job without having to relocate. This is just the labor pooling story; the fact that the skill involves high technology, rather than shoemaking or tufting, may be of secondary importance.

An anecdote: when involved with the MIT productivity commission, which eventually produced the best-selling book *Made in America*, I became very unpopular with the engineers for suggesting that some very non-technological-seeming sectors were in economic terms not so different from high tech. For example, consider Milan's fashion industry: a cluster of firms that rely on a highly specialized labor force (designers, models, stitchers, etc.), on specialized suppliers (fabrics, dyes, makeup, etc.), and on early access to information (what's in, what's out). I argued that economically Milan and Route 128 are similar creatures. The engineers thought I was frivolous.

Services

In the late twentieth century the great bulk of our labor force makes services rather than goods. Many of these services are nontradable and simply follow the geographical distribution of the goods-producing population—fast-food outlets, day-

care providers, divorce lawyers surely have locational Ginis pretty close to zero. Some services, however, especially in the financial sector, can be traded. Hartford is an insurance city; Chicago the center of futures trading; Los Angeles the entertainment capital; and so on.

The most spectacular examples of localization in today's world are, in fact, based on services rather than manufacturing. Tokyo and London are not essentially manufacturers. Even Silicon Valley and Route 128 are in effect more nearly centers that supply services to manufacturing than actual physical production sites. And arguably technology is moving in a direction that will promote more localization of services. Transportation of goods has not gotten much cheaper in the past eighty years: the epochal innovations were railroads and steamboats, with everything since representing only modest improvements. But the ability to transmit *information* has grown spectacularly, with telecommunication, computers, fiber optics, etc.

Some trends are visible. One of my students has been looking at the data on the growing concentration of wealth in Southeastern England; he finds that it is the service industries that are concentrating there, while manufacturing is actually shifting the other way.

The important point is that the logic of localization remains similar. Catherine Evanses—that is, small accidental events—start a cumulative process in which the presence of a large number of firms and workers acts as an incentive for still more

firms and workers to congregate at a particular location. The resulting pattern may be determined by underlying resources and technology at some very aggregative level; but at ground level there is a striking role for history and accident.