

CHAPTER 2

Why Do Cities Exist?

*Nobody ever saw a dog make a fair and deliberate exchange
of one bone for another with another dog.*

—ADAM SMITH

Cities exist because individuals are not self-sufficient. If each of us could produce everything we consumed and didn't want much company, there would be no reason to live in dirty, noisy, crowded cities. We aren't self-sufficient, but instead specialize in a labor task—writing software, playing the accordion, performing brain surgery—and use our earnings to buy the things we don't produce ourselves. We do this because labor specialization and large-scale production allow us to produce and consume more stuff. As we'll see in this chapter, production happens in cities, so that's where most of us live and work. By living and working in cities, we achieve a higher standard of living but put up with more congestion, noise, and pollution.

To explain why cities exist, we'll start with a model that implies that they don't. In the model of backyard production, every consumer is a producer, and all production occurs in backyards (or apartment roofs). In other words there is no need for concentrated production or population. As we drop the assumptions of the backyard-production model, the new models imply that cities will develop. In other words, the short list of assumptions in the model identifies the key factors behind the development of cities.

A REGION WITHOUT CITIES—BACKYARD PRODUCTION

Consider a region that produces and consumes two products, bread and shirts. People use the raw materials from land (wool and wheat) to produce the two consumer products. The following assumptions eliminate the possibility of cities.

- **Equal productivity.** All land is equally productive in producing wheat and wool, and all workers are equally productive in producing shirts and bread.

- **Constant returns to scale in exchange.** The unit cost of exchange (the cost of executing one transaction, including transportation cost) is constant, regardless of how much is exchanged.
- **Constant returns to scale in production.** The quantity of shirts produced per hour is constant, regardless of how many shirts a worker produces. The same is true for bread production.

Together these assumptions eliminate the possibility of exchange and guarantee that each household will be self-sufficient. If a person were to specialize in bread and then trade some bread for shirts, she would incur a transaction cost equal to the time required to execute the trade. Under the assumption of equal productivity, there is no benefit from specialization because everyone is equally productive. Under the assumption of constant returns to scale, there is no benefit from producing shirts in factories because an individual is just as efficient as a shirt factory. In sum, exchange has costs without any benefits, so every household will be self-sufficient, producing everything it consumes.

1762653 2014/09/02 157.252.98.229

The absence of exchange guarantees a uniform distribution of population. If population were concentrated at some location, competition for land would bid up its price. People in the city would pay a higher price for land without any compensating benefit, so they would have an incentive to leave the city. In the locational equilibrium, the price of land would be the same at all locations, and population density would be uniform. Recall the first axiom of urban economics:



Prices adjust to ensure locational equilibrium

In this case, all sites are equally attractive, so locational equilibrium requires the same price of land at all locations.

A TRADING CITY

Now that we have a short list of assumptions under which cities don't develop, let's drop the assumptions, one by one, and see what happens. We'll start by dropping the assumption of equal productivity for all workers. Suppose households in the North are more productive in producing both bread and shirts. This could result from differences in soil conditions, climate, or worker skills. Table 2–1 shows the output per hour for the two regions. While each worker in the South can produce one shirt or one loaf per hour, workers in the North are twice as productive in producing bread and six times as productive in producing shirts.

TABLE 2–1 Comparative Advantage

	North		South	
	Bread	Shirts	Bread	Shirts
Output per hour	2	6	1	1
Opportunity cost	3 shirts	1/3 loaf	1 shirt	1 loaf

Comparative Advantage and Trade

A region has a comparative advantage in producing a particular product if it has a lower opportunity cost. For every shirt produced, the North sacrifices 1/3 loaf of bread, so that's the opportunity cost of a shirt. In the South, the opportunity cost of a shirt is one loaf. The North has a lower opportunity cost for shirts, so it has a comparative advantage in producing shirts. It is sensible for the North to specialize in shirts (and not produce any bread) because, although the North is twice as productive as the South in producing bread, the North is *six* times as productive in producing shirts.

Comparative advantage may lead to specialization and trade. Suppose the two regions are initially self-sufficient, with each household producing all the bread and shirts it consumes. Table 2–2 shows what happens if a North household switches one hour from bread to shirt production, and a South household goes the other direction, switching two hours from shirt to bread production. The first row shows the changes in production: –2 loaves and +6 shirts for North; +2 loaves and –2 shirts for South. As shown in the second and third rows, if the households exchange two loaves and four shirts, each has a gain from trade of two shirts. After specialization and exchange, each household has just as much bread as before and two additional shirts.

What about transaction costs? The transaction cost is the opportunity cost of the time required to exchange products and is equal to the amount of output that could be produced during that time. For example, a North household can produce six shirts per hour, so the opportunity cost for a 10-minute (1/6 hour) transaction is one shirt. As long as the transaction time is less than 1/3 hour (two shirts), trade is beneficial. The South household, with a lower opportunity cost, has a lower transaction cost.

Scale Economies in Exchange

The presence of specialization and trade will not necessarily cause a city to develop. The second assumption of the backyard-production model is that there are constant returns to scale in exchange. Under this assumption, an individual household is just as efficient in executing trades as a trading firm, so there is no reason to pay a firm to execute an exchange. Therefore, each North household will link up with a South household to exchange shirts and bread directly, without intermediaries.

TABLE 2–2 Specialization and Gains from Trade

	North		South	
	Bread	Shirts	Bread	Shirts
Change in production from specialization	–2	+6	+2	–2
Exchange 4 shirts for 2 loaves	+2	–4	–2	+4
Gain from trade	0	+2	0	+2

Trading firms will emerge if there are economies of scale associated with exchange and trade. Recall the fourth axiom of urban economics:



Production is subject to economies of scale

A trading firm could use indivisible inputs such as a large truck to transport output between North and South. Similarly, workers who specialize in transportation tasks will be more efficient than workers who spend most of their time producing bread or shirts. In general, because trading firms have lower transaction costs, individual households will pay trading firms to handle exchanges.

The emergence of trading firms will cause the development of a trading city. To fully exploit scale economies, trading firms will locate at places that can efficiently collect and distribute large volumes of output. The concentration of trade workers will bid up the price of land near crossroads, river junctions, and ports. The increase in the price of land will cause people to economize on land by occupying smaller residential lots. The result is a place with a relatively high population density—a city.

TRADING CITIES IN URBAN HISTORY

Our simple model of the trading city suggests that trading cities develop when comparative advantage is combined with scale economies in transport and exchange. This observation provides some important insights into the history of cities before the Industrial Revolution of the 1800s. Most of the workers in these trading cities didn't produce goods, but instead collected and distributed goods produced elsewhere, such as agricultural products from the hinterlands and handcrafted goods from various locations. Trade was a risky business, and firms in the trading city provided insurance, credit, investment opportunities, banking, and legal services.

Trading Cities in World History

Trading cities have a long history. In the third millennium B.C., Phoenicians used fast sailing ships to serve as traders for the entire Mediterranean basin, trading dye, raw materials, foodstuffs, textiles, and jewelry. They established trading cities along the Mediterranean coast in present-day Lebanon. Around 500 B.C., Athens was a thriving site for regional trade, exchanging household crafts and olive products for food and raw materials from the countryside. During the 11th and 12th centuries, Italian city-states forged agreements with the Byzantine and Islamic rulers for trade with North Africa and the East. The Europeans traded wood, iron, grain, wine, and wool cloth for medicines, dyes, linen, cotton, leather, and precious metals. This trade was the major force behind the growth of Venice, Genoa, and Pisa.

Some cities were built on coercive transfer payments rather than voluntary trade. The Athenian empire developed in the aftermath of the successful war against Persia in the fifth century B.C. After the Greek city-states repelled the Persian invasion, they formed the Delian League for joint defense and later to carry the war into

Asia Minor. By the end of the successful campaign, Athens controlled the league and transformed the voluntary contributions of member city-states into payments of tribute to Athens. The system of homage and tribute led to the Peloponnesian War between the Athenian Empire and Sparta (431 to 404 B.C.). The war ended when Athens renounced control over its empire and demolished its defensive walls.

By the third century A.D., Rome had a population exceeding 1 million. The Romans established colonial cities throughout Europe and focused on collecting the agricultural surplus while they neglected urban production activity (Hohenberg and Lees, 1985). Instead of exchanging urban goods for agricultural products, Rome used conquest and tribute to feed its population. In the fourth and fifth centuries, attacks from Germanic tribes disrupted the Roman collection system. It appears that there was little interest outside of Rome in restoring the “trade” routes, so the losses from successive attacks were cumulative. If Rome had relied to a greater extent on voluntary exchange, the colonies would have had a greater stake in maintaining the exchange network and the Western empire might have recovered from the Germanic raids.

What are the lessons from the rise and fall of Athens and Rome? Early in its history, Athens thrived under a system of voluntary trade with other areas, exchanging urban goods for food from the countryside. The Athenians eventually switched to a system of conquest and tribute, resulting in war and the decline of the city. Mumford (1961) suggests that the city of Rome should have been called “Parasitopolis” to indicate its dependence on the labors of outsiders. The decline of Rome was caused in part by the disruption of its collection system by the Germanic raids. Perhaps the lesson is that cities based on coercive transfer payments are not sustainable.

Trading Cities in American History

The history of urban America illustrates the role of transport costs and comparative advantage in trading cities (Bartlett, 1998). In the 1700s, most cities served largely as trading posts for ocean trade. On the eastern seaboard, cities collected agricultural products from their hinterlands to the west and shipped them overseas. The volume of trade was limited by the dirt roads serving the interior: Travel was always slow and, in times of rain and melting snow, slippery. The Pennsylvania Turnpike, built with stone and gravel in 1792, increased travel speeds to a steady two miles per hour, increasing the market area and trading volume of the city of Philadelphia.

Farther to the north, New York State took more drastic steps, completing the 360-mile Erie Canal in 1825. The canal linked New York City, with its natural harbor, to vast agricultural areas to the north and west, and it cut freight costs from about 20 cents per ton mile to 1.5 cents. An additional canal connecting Lake Champlain to the Hudson River extended the market area of New York City to northern New England. The vast transportation network increased the volume of trade through New York City, increasing its size. By 1850, the city had a population of half a million, about 20 times its size at the end of the American Revolution. Other cities, including competitors to the south (Baltimore and Philadelphia), responded by building canals to connect hinterlands and ports, and by 1845 there were over 3,300 miles of artificial waterways in the United States.

Comparative advantage also plays a role in urban history. Eli Whitney's cotton gin (1794) provided a means of removing the sticky seeds of green-seed cotton, which could be grown throughout the south. The total output of cotton increased by a factor of 50 over a 15-year period, with most of the output coming from inland areas far from the east coast ports. American cotton was transported along rivers to New Orleans for shipment to textile firms in New England and Europe. The increase in cotton trade caused the rapid growth of New Orleans at the mouth of the Mississippi, and the development of upriver commercial cities such as Mobile, Alabama, and Natchez, Mississippi.

Later innovations in transportation reduced transport costs and contributed to the development of trading cities. Before the introduction of the steamboat in 1807, traffic was strictly downstream: After cargo was unloaded at the terminal point, wooden boats were broken up for lumber. The steamboat allowed two-way traffic and cut river freight costs, increasing the volume of trade and the size of river cities. Later, the steam engine was used to power locomotives, and railroad freight replaced river shipping as the principal means of transporting goods. Between 1850 and 1890, the ratio of railroad freight to river freight went from 0.10 to 2.0, and the volume of railroad freight increased by a factor of 240. The shift from river to railroad caused the decline of commercial cities along rivers and the rise of cities along the vast railroad network.

A FACTORY TOWN

The third assumption of the backyard-production model is constant returns to scale in production. We'll maintain this assumption for bread production, but apply the fourth axiom of urban economics to shirt production:



Production is subject to economies of scale

A shirt factory will use indivisible inputs (machines) and allow workers to specialize in narrowly defined tasks, leading to a higher output per worker and lower average cost. Suppose a household can produce either a loaf of bread or one shirt per hour. A worker in a shirt factory is six times as productive as a home worker, so the factory worker produces six shirts per hour.

Determining Wages and Prices

We assume that workers are perfectly mobile, so the utility level of a city worker must be the same as the utility level for a rural worker. Recall the first axiom of urban economics:



Prices adjust to ensure locational equilibrium

A factory must pay its workers enough to make them indifferent between working in the factory town and in the rural area. A rural worker earns one loaf of bread per

TABLE 2–3 Cost of Factory Shirt

Labor cost per hour	3/2 loaves
Cost of indivisible inputs per hour	1/2 loaf
Total cost per hour	2 loaves
Cost per shirt with 6 shirts produced per hour	1/3 loaf

hour, so city workers must earn one loaf per hour plus an amount high enough to offset the higher cost of living in the factory town, such as higher land prices. For example, if the cost of urban living is 50 percent higher, locational indifference requires an hourly wage of $3/2$ loaves of bread. A city worker will pay $1/2$ loaf for land, leaving one loaf per hour of factory work, the same that she could earn producing bread in a rural area.

What's the price of factory shirts? The price must be high enough to cover the costs of labor and the indivisible inputs used to produce shirts. In Table 2–3 the labor cost per hour is the wage ($3/2$ loaves) and the hourly cost of indivisible inputs is $1/2$ loaf. Adding these together, the hourly cost of producing shirts is two loaves of bread. To translate this into a cost per shirt, recall that a factory worker produces six shirts per hour, so the cost per shirt is one-sixth of the cost per hour, or $2/6 = 1/3$ loaf. Therefore, for zero economic profit, the price per shirt must be $1/3$ loaf of bread.

Suppose there is a single shirt factory in the region. The factory competes with homemade shirts, and will sell shirts to any household for which the net price of factory shirts is less than the cost of a homemade shirt. The cost of a homemade shirt is the one loaf of bread that is sacrificed to produce a shirt. The net price of a factory shirt equals the price charged by the factory ($1/3$ loaf) plus the opportunity cost of travel to and from the factory to buy the shirt.

The Market Area of a Factory Town

Figure 2–1 (page 24) shows the net price of factory shirts and the market area of the shirt factory. As shown by point *f*, the net price for a consumer located just across the road from the factory (distance = 0) is the factory price, equal to $1/3 = 4/12$ loaf of bread. Other consumers bear a travel cost when they buy factory shirts, so the net price is higher. Suppose the travel time is $1/12$ hour per round-trip mile: It takes $1/12$ hour to complete a round trip of one mile in each direction. In an hour, a rural household can produce one loaf of bread, so in $1/12$ of an hour of travel, it sacrifices $1/12$ loaf. For example, at point *g* (two miles from the factory), the net price of a factory shirt is $6/12$ loaves, equal to $4/12$ paid at the factory plus $2/12$ in travel cost (forgone bread production at home).

The market area of the factory is the area over which it underprices the home production of shirts. In Figure 2–1, the horizontal line shows the opportunity cost of homemade shirts, which is one loaf of bread. The net price of factory shirts is $4/12$ at the factory and increases by $1/12$ per mile, reaching one loaf at a distance of eight miles ($4/12 + 8/12$). In other words, the factory underprices home production up

in industrial cities. In contrast to the earlier trading cities, workers in factory cities produced products rather than simply distributing products produced elsewhere.

Innovations in Manufacturing

One of the key innovations of the Industrial Revolution was Eli Whitney's system of interchangeable parts for manufacturing, developed around 1800. Under the traditional craftsman approach, the component parts of a particular product were made individually—and imprecisely. Skilled craftsmen were necessary to produce the parts and then fit them all together. Under Whitney's system, the producer made a large batch of each part, using precise machine tools to generate identical parts. The identical parts were interchangeable, so unskilled workers could be quickly trained to assemble the parts. The replacement of handcraft production with standardized production generated large scale economies, causing the development of factories and factory cities.

Whitney applied this system to the production of muskets for the army. To prove to President-elect Jefferson and other government officials that his system would work with unskilled labor, he unloaded a random collection of parts onto the floor and had the officials assemble the muskets. He got the contract to manufacture 10,000 muskets and built a factory in New Haven, Connecticut, close to a stream that he used to power the factory. His system, which became known as the American System of Manufacturing, became the standard system for mass production.

The new system of manufacturing caused the development of factory cities. New machines, made of iron instead of wood, were developed to fabricate products in large factories. Manual production by skilled artisans was replaced by mechanized production using interchangeable parts, specialized labor, and steam-powered machines. Mass production decreased the relative cost of factory goods, causing the centralization of production and employment in large industrial cities.

As an illustration of the role of scale economies in the development of cities, consider the sewing machine, which was developed in the middle of the 19th century. At the beginning of the century, about four-fifths of the clothing worn in the United States was hand-sewn in the home for members of the household, and the rest was hand-sewn by tailors. The sewing machine (patented in 1846) allowed factories to underprice home producers, and by 1890 nine-tenths of U.S. clothing was being made in factories. New cities developed around the clothing factories.

A similar story line applies to shoes. Before 1700, most shoes were produced in the home or the local village. The cost of transportation was so high that local production was efficient. Over time, transportation costs decreased, and the putting-out system was implemented in the 1700s: Shoe producers distributed raw materials to cottage workers, collected their rough output, and finished the shoes in a central shop. As new shoemaking machines were developed, the number of operations performed in the central shops increased. The McKay sewing machine (for which a patent was granted to Lyman Blake in 1858) mechanized the process of sewing the soles to the uppers. The scale economies in shoe production increased to the point that shops became genuine factories, and cities developed around the shoe factories.

Innovations in Transportation

Innovations in intercity transportation contributed to industrialization and urbanization. As we saw earlier in the chapter, the dirt roads of the 1700s were replaced by turnpikes, and the construction of canals allowed a more dense network of inland water transport. The development of the steamship allowed two-way travel on major rivers, and the railroad system increased the speed and reach of the transportation system. All of these innovations decreased the relative price of factory goods, contributing to the growth of factory cities.

Innovations in Agriculture

One of the three conditions for the development of cities is an agricultural surplus to feed city dwellers. The Industrial Revolution generated a number of innovations that increased agricultural productivity. Farmers substituted machinery for muscle power and simple tools, increasing the output per farmer. The increased agricultural productivity freed people to work in urban factories and commercial firms. Between 1800 and 1900, the share of the population living in cities increased from 6 percent to 35 percent, reflecting the decrease in the number of agricultural workers required to feed city dwellers.

Consider first the sowing side of agriculture. At the start of the 19th century, plows were fragile, awkward, and often made of wood. These inefficient plows were replaced in the 1830s by the cast-iron plow, which was produced in factories in Pittsburgh and Worcester. In the 1840s, John Deere introduced the steel plow, which was lighter, stronger, and easier to handle. Later innovations allowed the farmer to adjust the depth and angle of the plow blade, increasing productivity further.

Consider next the reaping side of agriculture. In 1831, McCormick combined several earlier innovations into a horse-drawn harvesting machine that increased the productivity of the most labor-intensive part of agriculture. Using a horse-drawn reaper, two people could harvest the same amount of grain as eight people using traditional harvesting methods.

Other innovations contributed to higher agricultural productivity. The development of agricultural science led to innovations in planting, growing, harvesting, and processing. Innovations in transportation cut transport costs and allowed each farmer to serve a wider market area. Because of rising productivity, the share of employment in agriculture decreased over the 19th and 20th centuries, from over 90 percent to less than 3 percent.

Energy Technology and Location Decisions

During the Industrial Revolution, the location pattern of factory towns reflected changes in energy technology. The first factories used waterwheels turned by waterfalls and fast-moving streams to translate moving water into mechanical motion. The power was transmitted by systems of belts and gears. Textile manufacturers

built factories along backcountry streams in New England and used waterwheels to run their machines. Some examples of waterwheel cities are Lowell, Lawrence, Holyoke, and Lewiston.

The refinement of the steam engine in the second half of the 19th century made energy a transportable input. A key innovation was John McNaught's development of a compounding engine (using steam twice, at descending pressures, to drive pistons) in 1845. The steam engine could be operated anywhere, with the only constraint being the availability of coal to fuel the engine. Some energy-intensive manufacturers located near the coal mines in Pennsylvania. Others located along navigable waterways and shipped coal from the mines to their factories. In New England, textile firms shifted from backcountry waterfall sites to locations along navigable waterways. Production shifted to the Fall River–New Bedford area along the south coast of New England. The later development of the railroad gave coal users another transport option, causing the development of factories along the vast network of rail lines. In general, the steam engine widened the location options for factories.

The development of electricity changed the location patterns of factories. Electricity generators were refined in the 1860s, and the electric motor was developed in 1888. Factories replaced belt-and-gear systems driven by a central steam engine with small electric motors for individual machines. The first factory to use electric power was adjacent to a hydroelectric generating facility at Niagara Falls. Rapid improvements in the electricity transmission soon allowed factories to be hundreds of miles from hydroelectric and coal-powered generating plants. Between 1900 and 1920, the share of factory horsepower from electric motors increased from 2 percent to 33 percent.

The development of electricity made factories more footloose. A firm could tap water power without locating close to the stream and use coal without shipping the bulky fuel to the factory. In general, the development of electricity decreased the importance of energy considerations in location decisions, causing firms to base their location choices on the accessibility to other inputs and to consumers.

A SYSTEM OF FACTORY TOWNS

We can widen our horizon by looking at the entire region and consider the possibility of additional factory towns. Firms can enter the shirt industry by building shirt factories at different locations, and each firm will have a local monopoly in the area surrounding its factory. Recall the fifth axiom of urban economics:

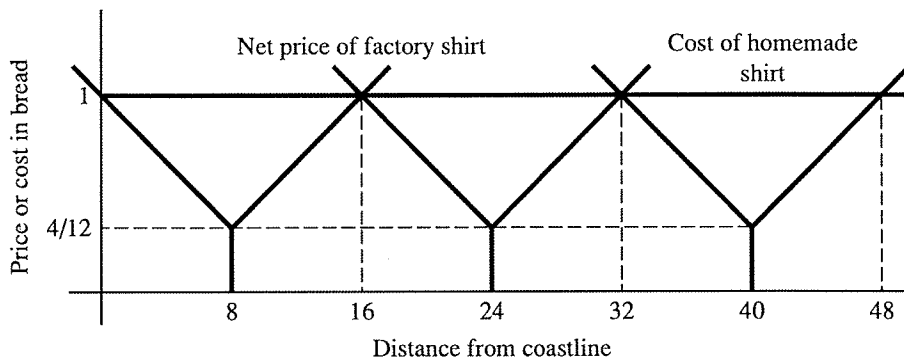


Competition generates zero economic profit

If there are no restrictions on entry, firms will continue to enter the market until economic profit is zero.

Figure 2–2 (page 28) shows the equilibrium in the region. The horizontal axis measures distance from a coastline. The rectangular region is 48 miles wide, and in equilibrium has three shirt factories, each with a market area 16 miles wide. The

FIGURE 2-2 System of Factory Towns



Each factory's market area is 16 miles wide, so a system of factory towns develops with a distance of 16 miles between towns. In this equilibrium, workers specialize, with shirt workers in towns and bread producers in rural areas between the towns.

market areas of the factories span the region: Every location in the region lies within the market area of some factory. There is complete labor specialization: Workers in factory towns produce shirts (and receive bread as wages), and workers in rural areas produce bread (and pay bread to get factory shirts).

This is an equilibrium because each firm makes zero economic profit and workers are indifferent between rural and city life:

- **Zero economic profit.** The factory price of $4/12$ loaf equals the average cost of producing shirts, including the cost of urban workers and the cost of indivisible inputs.
- **Locational indifference for workers.** The wage for factory workers is high enough to cover (1) the opportunity cost of working in factories rather than producing bread in the rural area and (2) the higher cost of urban living (land rent).

What about rural residents? For a rural resident just outside the factory towns, the net price of a factory shirt is $4/12$ loaf of bread, compared to a homemade cost of one loaf. At the other extreme, a rural household eight miles from the factory pays a net price of one loaf per shirt ($4/12 + 8/12$ in travel cost). Recall the first axiom of urban economics:



Prices adjust to ensure locational equilibrium

In this case, the price of land in rural areas will adjust to make people indifferent between locations that differ in their accessibility to the shirt factory. The shorter the distance to the factory, the lower the net price of factory shirts, and the more a household is willing to pay for land. In other words, the price of land adjusts to fully compensate for differences in accessibility.

Landowners benefit from the scale economies in production that generate the regional system of factory towns. In the rural areas, the price of land is higher at

locations close to the factory city. In the factory city itself, competition among workers for locations near the factory bids up the price of land.

RESOURCES-ORIENTED FIRMS AND PROCESSING TOWNS

Up to this point, we have ignored the cost of transporting the raw materials required to produce urban goods (shirts). We have implicitly assumed that factory workers harvest wool from wild sheep who wander by the factory at just the right time to be sheared for shirts. In the language of urban economics, we have assumed that the raw materials required for production are ubiquitous—available at all locations at the same price. This is an extreme case of a market-oriented industry, defined as an industry in which the cost of transporting output is large relative to the cost of transporting inputs. The Appendix to this chapter explores the location decisions of market-oriented firms.

Consider the opposite extreme. Suppose it is costly to transport material inputs, but output can be transported at zero cost. This is the extreme case of a materials-oriented industry, defined as an industry for which the cost of transporting material inputs is large relative to the cost of transporting output. The Appendix to this chapter explores the location decisions of materials-oriented firms. For example, the sugar content of sugar beets is roughly 15 percent, so it takes seven tons of beets to produce one ton of sugar. Beet-sugar firms locate their plants close to the beet fields to economize on transport costs (Holmes and Stevens, 2004). Beet-sugar producers will cluster in the regions of the country where weather and soil conditions are favorable for the production of sugar beets.

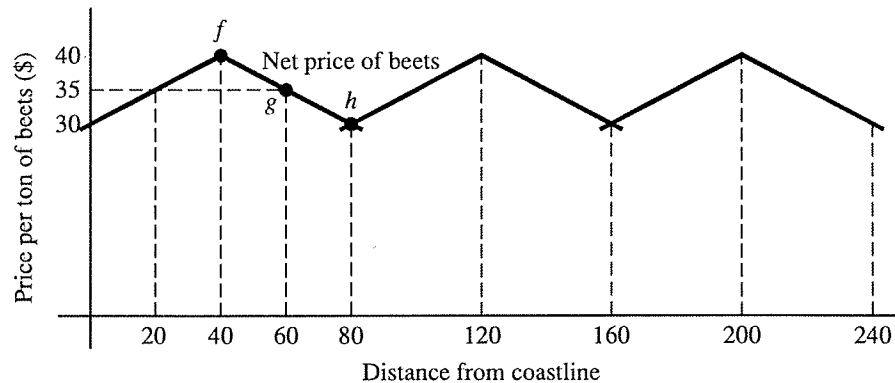
Scale Economies and Market Areas

The process of transforming sugar beets into sugar is subject to scale economies. Processors use indivisible inputs and engage in factor substitution, so the average processing costs decrease as the quantity increases. The typical sugar-beet processing plant employs 186 workers, about four times the average number of employees per plant in manufacturing.

The market area of a processing plant is determined by the net price farmers receive. The net price equals the price paid by the processor minus the cost of transporting the beets from the farm to the processing plant. In Figure 2–3, the horizontal axis measures the distance from a coastline. Consider a processing plant located 40 miles from the coastline. If the price paid by the processor is \$40, the net price is \$40 for a farmer across the road from the processor (point *f*), and drops to \$35 for a farmer 20 miles away (point *g*). Farmers naturally sell to the processing plant that generates the highest net price, so this processing plant has a market from the coastline to 80 miles inland.

Figure 2–3 shows a regional equilibrium with three processing plants, each with a market area of 80 miles. Each firm is the single buyer of sugar beets within

FIGURE 2-3 System of Processing Towns



The net price of beets (received by farmers) decreases as the distance to the processing plant increases. The market area of the typical sugar-beet processing plant is 80 miles wide, so a system of processing towns develops with a distance of 80 miles between them.

its 80-mile market area, so it has a local monopsony (as opposed to monopoly for a single seller). Recall the fifth axiom of urban economics:



Competition generates zero economic profit

If there are no restrictions on entry, firms will continue to enter the market until economic profit is zero.

System of Processing Towns

The location of sugar-beet processing plants leads to the development of a system of processing towns. The people who work in the processing plants live nearby to economize on commuting, resulting in a place with a relatively high population density: a small city or town. As Holmes and Stevens (2004) show, beet-sugar processing plants locate in the regions where beet production occurs, and they carve out input market areas within each beet-growing region.

Note the similarities of the beet-sugar industry and the shirt industry. In the shirt industry, with relatively high cost of transporting output, each firm gets a local monopoly, with all consumers patronizing the nearest factory. In the beet-sugar industry, with relatively high cost of transporting input, each firm gets a local monopsony, with all farmers selling their output to the nearest beet-sugar plant.

1762653 2014/09/02 157.252.98.229

Other Examples of Resources-Oriented Industries

The same logic applies to other resources-oriented industries (Kim, 1999). The production of leather requires hides and tannin (from tree bark) for the tanning process. The tannin content of bark is only 10 percent, so it takes a lot of bark to produce a ton of leather. In 1900, U.S. leather producers located close to forests to economize on the transport costs of tannin. In the 20th century, improvements in the extraction

process and the development of synthetic tannin reduced the orientation toward forest sites, and firms moved closer to other input sources.

The location decisions of steel producers reflected changes in the input requirements of coal and iron ore. Early in the history of the industry, a ton of steel required five tons of coal and two tons of ore, and steel production was concentrated near coal deposits. Technological innovations reduced the coal content, and steel producers were pulled toward locations that provided access to ore deposits, including sites on the Great Lakes that offered water access to ore from the Masabi Range in Minnesota. Each ton of steel required 175 tons of water, and the Great Lakes sites also provided a plentiful supply of water.

The location decisions of these and other resources-oriented industries caused the development of processing towns. Leather towns developed around tanneries and steel towns developed around steel mills. Lumber producers locate near forests, causing the development of lumber towns centered on sawmills. Ore processors locate near mineral deposits, causing the development of mining towns.

SUMMARY

Cities exist because of the benefits of centralized exchange (trading cities) and centralized production (factory and processing cities). We have focused on the market forces that generate cities. For a discussion of other possible reasons for cities, such as religion and defense and their role in the development of the first cities, see Mumford (1961) and O'Sullivan (2005). Here are the main points of this chapter:

1. A trading city develops when comparative advantage is combined with scale economies in exchange.
2. A factory city develops when there are scale economies in production.
3. The Industrial Revolution caused massive urbanization because of its innovations in agriculture, transportation, and production.
4. Changes in energy technology altered the location decisions of firms, with water power generating factories along streams, steam power generating factories along rivers and railroads, and electricity making firms more footloose.
5. Spatial competition among firms generates a market area for each firm and a system of cities.

APPLYING THE CONCEPTS

For exercises that have blanks (____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. Innovation and Trading Cities: Numbers

Consider the example shown in Tables 2–1 and 2–2. Suppose a single number changes: The shirt output per hour in the North is four shirts instead of six shirts.

To specialize, a North household switches one hour from bread to shirts, and a South household switches two hours from shirts to bread. The exchange rate is three shirts for two loaves. There are scale economies in exchange, and a transaction takes 30 minutes to execute.

- a. The gains from trade are _____ for North and _____ for South.
- b. A trading city [will/won't] develop because. . .
- c. Suppose an innovation in transportation decreases the transaction time. A trading city will develop if the transaction time is less than _____ minutes because. . .

2. Matter Transmitter and Trading City

Consider a region with two standardized products (bread and shirts) that are transported by horse-drawn wagons. There is a single trading city. Consider the effects of a new matter transmitter, an indivisible input that is economical for trading firms in the city, but not economical for an individual household. A transmitter instantly transports goods from a production site (a farm or workshop) to the trading firm, and the marginal cost of transport is zero.

- a. The labor used in transporting products from production sites to trading firms will [increase, decrease, not change] because. . .
- b. The volume of trade in the region will [increase, decrease, not change] because. . .
- c. The labor used in processing transactions (banking, accounting, insuring) will [increase, decrease, not change] because. . .
- d. The trading city will grow if _____ exceeds _____.
- e. Suppose the transmitter technology changes, and it becomes economical for an individual household. The trading city will [grow, shrink, disappear] because. . .

3. Drilling for Cities

Consider a country with two regions that are separated by a mountain range. Initially each region is self-sufficient in shirts and bread, and there are no cities. Suppose that a tunnel is drilled through the mountain, decreasing travel costs between the two regions. The tunnel will cause the development of a trading city if three following conditions are met. . .

4. Spring-Loaded Sneakers

Consider the example shown in Figure 2–1, with the “martini-glass” shape showing the net price of the factory product. Suppose all consumers switch to spring-loaded sneakers, decreasing walking time per round-trip mile from $1/12$ hours to $1/18$ hours.

- a. The slope of the net-price curve changes to _____ and the width of the market area changes from 16 miles (8 on each side) to _____ miles (_____ on each side).
- b. Using Figure 2–2 as a starting point, the spring-loaded sneakers change the number of factories in the 48-mile region from _____ to _____, each with a market area _____ miles wide.

5. Innovation and Market Areas

Consider the example shown in Figure 2–1. Depict graphically the effects of the following sequence of changes on the “martini glass” figure and the width of a factory’s market area. The changes are cumulative.

- a. An innovation in production that doubles labor productivity in factories will [shorten, lengthen, not change] the stem of the martini glass from _____ loaves to _____ loaves because. . . .
- b. The width of the market area increases from 16 miles (8 miles on each side) to _____ miles (_____ on each side) because. . . .
- c. An innovation in transportation that doubles consumer travel speeds will [decrease, increase, not change] the slope of the martini glass from _____ loaf per mile to _____ loaf per mile.
- d. The width of the market area increases from _____ miles (_____ on each side) to _____ (_____ on each side) miles because. . . .

6. Matter Transmitter in a Factory City

Consider a region with a single factory city that developed as a result of scale economies in production. Consider the effects of a new matter transmitter, which can instantly transport goods (but not people) from the factory to any consumer up to 12 miles away, with a zero marginal cost of transport. The hourly cost of the transmitter is one loaf of bread. Using Figure 2–1 as a starting point, show the effects of the matter transmitter on the market area of the factory. Is “martini glass” still an apt descriptor of the figure? If not, what is a better descriptor?

7. Singing and the Internet

Consider a region where households produce and consume two products, bread and live musical performances. All workers are equally productive at producing bread and music. The production of bread is subject to constant returns to scale, with one hour required to produce each loaf. In an hour, a single person can produce one unit of music for herself. A choral group of 20 people working for an hour (on practice and performance) can produce one unit of music for an audience of 80 people. Assume that the opportunity cost of actually listening to the music is zero. The travel cost for music consumers is $1/8$ hour per round-trip mile.

- a. The equilibrium price for choral music—the price paid by each person who listens—is _____ loaves because. . . .
- b. Use a martini-glass figure like Figure 2–1 to show the equilibrium market area of the choral group. The stem of the martini glass is _____ and the slope is _____ per mile, so the market area is _____ on each side because. . . .
- c. Suppose choral music becomes available on the Internet, and the provider charges $1/2$ loaf per song. Use a martini-glass figure like Figure 2–1 to show the new market area. The market area is _____ miles on each side because. . . .

8. Catapult in Retireland

In Retireland, no one commutes to work, and everyone consumes a single good (food), which is imported from another region and can be purchased (one meal

at a time) from the nearest vending machine. Alternatively, food can be delivered by a distant catapult, capable of flinging a meal through a food slot on the customer's roof. The price of a delivered catapult meal is \$6 and the price of a vending-machine meal is \$2. The travel cost for consumers is \$0.04 per round-trip meter (\$0.02 per meter traveled).

- a. Use a martini-glass figure like Figure 2–1 to show the equilibrium market area of the vending machine. The stem of the martini glass is _____ and the slope is _____ per meter, and the horizontal curve showing the cost of catapult meals is at _____. The market area is _____ meters on each side because. . . .
- b. Arrows up, down, or horizontal: As the distance to the nearest vending machine decreases, the price of land _____ and population density _____.

9. Performance City

Consider a region where households produce and consume musical performances, either live performances or free recorded performances transmitted over the Internet into homes. After accounting for price, the utility from a live performance is three times the utility from recorded music. There are scale economies in the production of live music, which are provided at a performance center at the center of the region.

- a. Using Figure 2–1 as an inspiration, depict graphically a method for determining the market area of the performance center. Hint: The vertical axis measures utility rather than price.
- b. Suppose the demand for live performance is income-elastic, and the demand for recorded performances is independent of income. Use a second graph to show the effect of an increase in consumer income on the market area of the performance center.

10. Cheap Beet Hauling

Consider Figure 2–3, which shows the market areas of sugar-beet processing plants. Suppose that an increase in the price of diesel fuel doubles the cost of hauling sugar beets from farms to processing plants. Modify Figure 2–3 to show the effects of higher transport costs.

- a. The slope of the net-price curve changes from _____ per mile to _____ per mile.
- b. If the number of processing plants remains at three, the net price received by the farm that is most distant from its closest processing changes from _____ per ton to _____ per ton.
- c. We would expect the number of processing plants to [increase, decrease, not change] because. . . .

11. Beer and Wine

Consider the locations of breweries and wineries.

- a. Most breweries locate close to their customers (far from their primary input sources) because. . . .
- b. Most wineries locate close to their input sources (far from their consumers) because. . . .
- c. Consider a nation that is 120 miles wide. Beer consumers are uniformly distributed throughout the nation, while grapes are uniformly distributed through

the western region of the nation. There will be two evenly spaced wineries and two evenly spaced breweries. The wineries will locate at mile _____ and mile _____, splitting the _____. The breweries will locate at mile _____ and mile _____, splitting the _____.

REFERENCES AND ADDITIONAL READING

1. Bartlett, Randall. *The Crisis of American Cities*. Armonk, NY: M.E. Sharp, 1998.
2. Combes, P., and H. Overman. "The Spatial Distribution of Economic Activities in the European Union." Chapter 64 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
3. Davis, Kingsley. "Urbanization." In *The Urban Economy*, ed. Harold Hochman. New York: W. W. Norton, 1976.
4. Ellison, Glen, and Edward Glaeser. "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?" *American Economic Review* 89 (1999), pp. 311–316.
5. Fujita, Masahisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
6. Hohenberg, Paul M., and Lynn H. Lees. *The Making of Urban Europe 1000–1950*. Cambridge, MA: Harvard University Press, 1985.
7. Holmes, T., and J. Stevens. "Spatial Distribution of Economic Activities in North America." Chapter 63 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, ed. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
8. Kim, Sukkoo. "Regions Resources, and Economic Geography: Sources of U.S. Regional Comparative Advantage, 1880–1987." *Regional Science and Urban Economics* 29 (1999), pp. 1–32.
9. Mumford, Lewis. *The City in History*. New York: Harcourt Brace Jovanovich, 1961.
10. O'Sullivan, Arthur. "The First Cities." In *A Companion to Urban Economics*, ed. Richard Arnott and Daniel McMillen. Boston: Blackwell, 2005.

Appendix: Location Decisions of Transfer-Oriented Firms

This brief appendix explores the location decisions of *transfer-oriented* firms. For a transfer-oriented firm, the dominant location factor is the cost of transporting inputs and outputs. The firm chooses the location that minimizes total transport costs, defined as the sum of procurement and distribution costs. *Procurement cost* is the cost of transporting raw materials from the input source to the production

facility. *Distribution cost* is the cost of transporting the firm's output from the production facility to the output market.

The classic model of a transfer-oriented firm has four assumptions that make transportation cost the dominant location variable.

- **Single transferable output.** The firm produces a fixed quantity of a single product, which is transported from the production facility to an output market.
- **Single transferable input.** The firm may use several inputs, but only one input is transported from an input source to the firm's production facility. All other inputs are ubiquitous, meaning that they are available at all locations at the same price.
- **Fixed-factor proportions.** The firm produces its fixed quantity with fixed amounts of each input. In other words, the firm uses a single recipe to produce its good, regardless of the prices of its inputs. There is no factor substitution.
- **Fixed prices.** The firm is so small that it does not affect the prices of its inputs or its product.

Under these assumptions, the firm maximizes its profit by minimizing its transportation costs. The firm's profit equals total revenue (price times the quantity of output) less input costs and transport costs. Total revenue is the same at all locations because the firm sells a fixed quantity of output at a fixed price. Input costs are the same at all locations because the firm buys a fixed amount of each input at fixed prices. The only costs that vary across space are procurement costs (the costs of transporting the firm's transferable input) and distribution costs (the costs of transporting the firm's output). Therefore, the firm will choose the location that minimizes its total transport cost.

The firm's location choice is determined by the outcome of a tug-of-war. The firm is pulled toward its input source because the closer to the input source, the lower the firm's procurement costs. On the other side, the firm is pulled toward the market because proximity to the market reduces the firm's distribution costs.

RESOURCE-ORIENTED FIRMS

A resource-oriented firm is defined as a firm that has relatively high costs for transporting its input. Table 2A–1 shows the transport characteristics for a firm that produces baseball bats, using 10 tons of wood to produce three tons of bats. The firm is involved in a weight-losing activity: Its output is lighter than its transferable input because the firm shaves down logs to make bats.

The outcome of the locational tug-of-war is determined by the monetary weights of the firm's inputs and outputs. The monetary weight of the input equals the physical weight of the input (10 tons) times the transportation rate (\$1 per ton

TABLE 2A–1 Monetary Weights for a Resource-Oriented Firm

	Input (wood)	Output (bats)
Physical weight (tons)	10	3
Transport rate (cost per ton per mile)	\$1	\$2
Monetary weight (physical weight times rate)	\$10	\$6

per mile), or \$10 per mile. Similarly, the monetary weight of the output is three tons times \$2, or \$6 per mile. This firm is a resource-oriented firm because the monetary weight of its transferable input exceeds the monetary weight of its output. Although the unit cost of transporting output is higher (because finished bats must be packed carefully, but logs can be tossed onto a truck), the loss of weight in the production process generates a lower monetary weight for the output.

Figure 2A–1 shows the firm's transportation costs. Using x as the distance from the input source (the forest) to the production site (the factory), the firm's procurement cost equals the monetary weight of the input (the physical weight w_i times the transport cost rate t_i) times the distance between forest and factory:

$$PC = w_i \cdot t_i \cdot x$$

The slope of the procurement-cost curve is the monetary weight of the input, so PC rises by \$10 per mile, from zero at the forest to \$100 at the market 10 miles away.

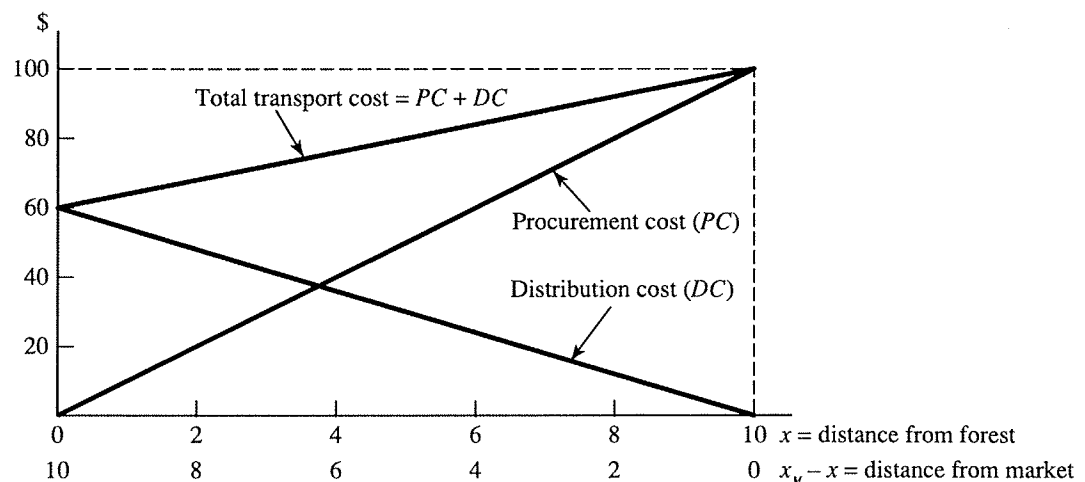
The distribution costs are computed in an analogous way. Using x_M as the fixed distance between the forest and the market (10 miles in our example), the firm's distribution cost is the monetary weight of the output (weight w_o times the transport cost rate t_o) times the distance from the factory to the market:

$$DC = w_o \cdot t_o \cdot (x_M - x)$$

The slope of the distribution-cost curve is the monetary weight of the output, so as we move from the forest toward the market, DC decreases by \$6 per mile, from \$60 at the forest (10 miles from the market) to zero at the market.

Total transport cost is the sum of procurement and distribution costs. In Figure 2A–1, total transport cost is minimized at the forest site at \$60. Starting at any location except the forest, a one-mile move toward the forest would decrease procurement cost by \$10 (the monetary weight of the wood) and increase distribution

FIGURE 2A–1 A Weight-Losing Firm Locates at Its Input Source



Total transport cost (procurement cost plus distribution cost) is minimized at the input source (the forest) because the monetary weight of the input (\$10) exceeds the monetary weight of the output (\$6).

cost by \$6 (the monetary weight of bats), for a net reduction of \$4. The firm's total transport cost is minimized at the input source because the monetary weight of the input exceeds the monetary weight of the output. Some other examples of weight-losing firms that locate close to their input sources are beet-sugar factories, onion dehydrators, and ore processors.

Some firms are resource oriented because their inputs are relatively expensive to transport. For example, a canner produces one ton of canned fruit with roughly a ton of raw fruit. The firm's input is perishable and must be transported in refrigerated trucks, while its output can be transported less expensively on regular trucks. Because the cost of shipping a ton of raw fruit exceeds the cost of shipping a ton of canned fruit, the monetary weight of the input exceeds the monetary weight of the output, and the firm locates near its input source, a fruit farm. In general, a firm's input will be more expensive to ship if it is more bulky, perishable, fragile, or hazardous than the output.

There are many examples of industries that locate close to their transportable inputs (Ellison and Glaeser, 1999). The producers of soybean and vegetable oil are concentrated in Nebraska, North Dakota, and South Dakota, close to the farms that supply soybeans and corn. Milk and cheese producers are concentrated in South Dakota, Nebraska, and Montana, close to dairy farms. Sawmills and other wood processors are concentrated in Arkansas, Montana, and Idaho, close to vast timberlands.

MARKET-ORIENTED FIRMS

A market-oriented firm is defined as a firm that has relatively high costs for transporting its output to the market. Table 2A–2 shows the transport characteristics for a bottling firm that uses one ton of sugar and three tons of water (a ubiquitous input) to produce four tons of bottled beverages. The firm is involved in a weight-gaining activity in the sense that its output is heavier than its transferable input. The monetary weight of the output exceeds the monetary weight of the input, so this market-oriented firm will locate near its market.

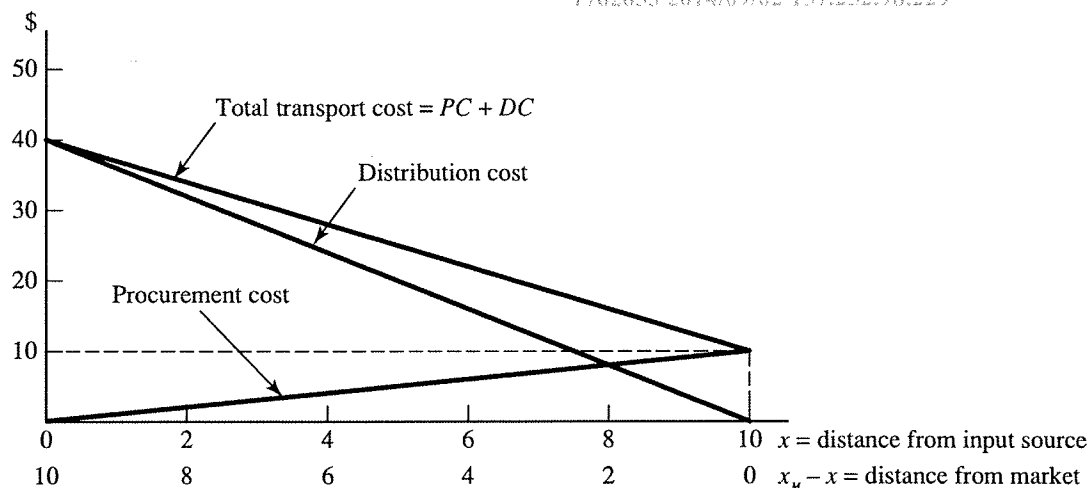
As shown in Figure 2A–2, the firm's transport cost is minimized at the market. Because the monetary weight of the output exceeds the monetary weight of the input, a one-mile move away from the market increases the distribution cost by more than it decreases procurement cost. Specifically, such a move increases distribution cost by \$4, but decreases procurement cost by only \$1, for a net loss of \$3. For this weight-gaining activity, the tug-of-war between input

TABLE 2A–2 Monetary Weights for a Market-Oriented Firm

	Input (sugar)	Output (beverages)
Physical weight (tons)	1	4
Transport rate (cost per ton per mile)	\$1	\$1
Monetary weight (physical weight times rate)	\$1	\$4

FIGURE 2A-2 A Weight-Gaining Firm Locates at Its Output Market

1762653 2014/09/02 157.252.98.229



Total transport cost is minimized at the market because the monetary weight of the output (\$4) exceeds the monetary weight of the transferable input (\$1). The weight-gaining firm locates at its output market.

source and market is won by the market because there is more physical weight on the market side.

Some firms are market oriented because their output is relatively expensive to transport. Output will be relatively costly to transport if it is bulky, perishable, fragile, or hazardous. The output of an automobile assembly firm (assembled cars) is more bulky than the inputs (e.g., rolls of wire, sheets of metal). The cost of shipping a ton of automobiles exceeds the cost of shipping a ton of component parts, so the monetary weight of the output exceeds the monetary weight of the inputs, pulling the firm toward the market. The output of a bakery is more perishable than its inputs, pulling bakeries toward consumers. A weapons producer combines harmless inputs into a lethal output, and the firm locates near its output market to avoid transporting the hazardous (or fragile) output long distances. In general, when a firm's output is relatively bulky, perishable, fragile, or hazardous, the tug-of-war will be won by the market, not because the output is heavier, but because it is more expensive to transport.

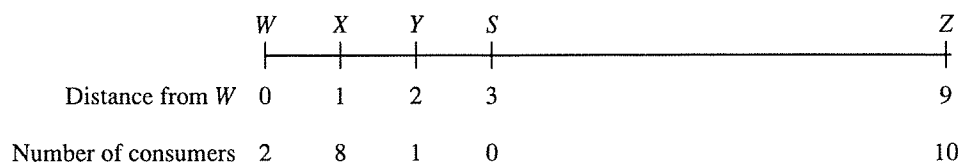
The Principle of Median Location

The classic model of the transfer-oriented firm assumes that the firm has a single input source and a single market. For more complex cases involving multiple inputs or markets, we can use *the principle of median location* to predict where a firm will locate:

The median location minimizes total travel distance

The median location splits travel destinations into two equal halves, with half the destinations in one direction and half in the other direction.

FIGURE 2A–3 The Principle of Median Location



Total delivery cost is minimized at the median location (Y), defined as the location that splits travel destinations into two equal halves.

We can illustrate this principle with the location decision of Ann, who makes and delivers pizzas to consumers along a highway. Under the following assumptions, her objective is to minimize her total delivery distance.

1. All inputs (labor, dough, toppings) are ubiquitous (available at all locations for the same price), so input transport costs are zero.
2. The price of pizzas is fixed, and each consumer along the highway demands one pizza per day.
3. Ann bears the delivery cost of \$1 per mile traveled. Each pizza delivery requires a separate trip.

Figure 2A–3 shows the distribution of consumers along the highway. Distances are measured from the western end of the highway (point W). There are two customers at point W, eight customers at point X (one mile from W), one customer at Y, and 10 customers at Z.

Ann will minimize her total delivery distance at point Y, the median location. Point Y is the median because there are 10 customers to the west (at points W and X) and 10 customers to the east (at point Z). To show the superiority of the median location, suppose Ann starts at Y, and then moves one mile east to point S. The move reduces her travel distance to customers at point Z by 10 miles, but increases her travel distance to customers to the west (at points W, X, and Y) by 11 miles. Her total delivery distance increases because she moves closer to 10 customers but farther than 11 customers. In general, any move away from the median location will increase delivery distances for the majority of consumers, so the total delivery distance increases.

It is important to note that the distance between the consumers is irrelevant to the firm's location choice. For example, if the Z consumers were located 100 miles from W instead of nine miles from W, the median location would still be point Y. Total delivery distance would still be minimized (at a higher level, of course) at point Y.

The principle of median location provides another explanation of why large cities become larger. In Figure 2A–3, suppose the locations (W, X, Y, Z) are cities, with the populations of cities W, X, and Y (in millions) indicated by the numbers listed for customers. For example, the population of city W is 2 million. In addition, suppose city Z, at the end of the line, has a population of 12 million, making it the median location. In this case, firms will minimize total delivery cost by locating in the large city, so the large city will grow. A firm that started in city Z and then moved

to the west would decrease delivery costs for a minority of its customers (11 million) while increasing delivery costs for a majority (12 million), so total delivery cost would increase. The lesson from this example is that the concentration of demand in large cities causes large cities to grow.

Transshipment Points and Port Cities

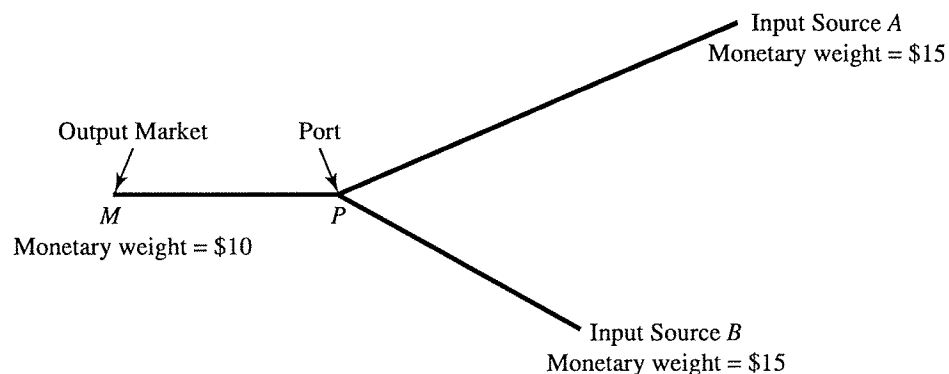
The principle of median location also explains why some industrial firms locate at transshipment points. A transshipment point is defined as a point at which a good is transferred from one transport mode to another. At a port, goods are transferred from trucks or trains to ships; at a railroad terminal, goods are transferred from trucks to the train.

Figure 2A–4 shows the location options for a sawmill. The firm harvests logs from locations *A* and *B*, processes the logs into lumber, and then sells the lumber in an overseas market at *M*. Highways connect points *A* and *B* to the port, and ships travel from the port to *M*. The sawmill is a weight-losing activity: The monetary weights of the inputs are \$15 for point *A* and \$15 for point *B*, and the monetary weight of the output is \$10.

Where will the firm locate its sawmill? Although there is no true median location, the port is the closest to a median location. If the firm starts at the port (*P*), it could move either toward one of its input sources or to its market.

- **Toward input source *A*.** A one-mile move from *P* toward point *A* will cause offsetting changes in the costs of transporting logs from the two input sources: the cost of logs from *A* would decrease, but the cost of logs from *B* would increase. At the same time, the cost of transporting output would increase by \$10. Given

FIGURE 2A–4 Median Location and Ports



The firm locates its sawmill at the port (*P*) because it is the median transport location. A one-mile move from the port toward either input source would increase output transport costs by \$10 while generating offsetting changes in the input transport costs for sources *A* and *B*. A one-mile move from the port toward the market would increase input transport costs by \$30 but decrease output transport costs by only \$10.

the offsetting changes in input transport costs and the increase in output costs, the port location is superior to locations between P and A . The same argument applies for a move from P toward B .

- **To market (M).** Unless the firm wants to operate a floating sawmill, it would not move to points between the port and the overseas market at M . It could, however, move all the way to the market. A move from P to M would decrease output transport cost by \$10 (the monetary weight of output) times the distance between M and P , and increase input transport cost by \$30 (the monetary weight of the inputs) times the distance. Therefore, the port location is superior to the market location.

Although the sawmill is a weight-losing activity, it will locate at the port, not at one of its input sources. The port location is efficient because it provides a central collection point for the firm's inputs.

There are many examples of port cities that developed as a result of the location decisions of industrial firms. Seattle started in 1880 as a sawmill town: Firms harvested trees in western Washington, processed the logs in Seattle sawmills, and then shipped the wood products to other states and countries. Baltimore was the nation's first boomtown: Flour mills processed wheat from the surrounding agricultural areas for export to the West Indies. Buffalo was the Midwestern center for flour mills, providing consumers in eastern cities with flour produced from Midwestern wheat. Wheat was shipped from Midwestern states across the Great Lakes to Buffalo, where it was processed into flour for shipment by rail to cities in the eastern United States. In contrast with Baltimore, which exported its output (flour) by ship, Buffalo imported its input (wheat) by ship.

CHAPTER 3

Why Do Firms Cluster?

People don't go there anymore. It's too crowded.

—YOGI BERRA

If two firms compete for customers in a region, will they locate close together or far apart? It is natural to imagine that the two firms will split the region into two halves, giving each firm a local monopoly. That's what happened in the theoretical models of Chapter 2, and it happens for many firms in the real world. Yet all sorts of competing firms locate close to one another, including carpet producers in Georgia and television producers in Los Angeles. Why?

This chapter explores agglomeration economies, the economic forces that cause firms to locate close to one another in clusters. The forces acting on firms in a single industry together are called localization economies, indicating that they are "local" to a particular industry. For example, firms in the software industry cluster in Silicon Valley. When agglomeration economies cross industry boundaries, they are called urbanization economies. The idea is that the presence of firms in one industry attracts firms in other industries. For example, the corporate headquarters of different industries cluster in cities. Urbanization economies lead to the development of large, diverse cities. As we'll see, localization and urbanization economies have common roots.

Before we explore the reasons for localization economies, it will be useful to look at some facts on industry clusters in the United States. Table 3–1 shows the facts on employment clusters for six industries. In the production of aircraft engines, the four metropolitan areas listed (Hartford, Phoenix, Cincinnati, and Indianapolis) contain nearly half of U.S. employment in the industry. Firms producing biopharmaceuticals cluster in New York, Chicago, Philadelphia, and San Francisco, while firms producing software cluster in Seattle, the San Francisco Bay Area, and Boston. Among the small metropolitan areas with relatively large clusters of software employment are Austin, Texas, and Raleigh, North Carolina. Bloomington, Indiana, has over a fifth of national employment in the production of elevators and

TABLE 3–1 Select Industrial Clusters in U.S. Metropolitan Areas, 2004

Product	Metropolitan Area	2004 Employment	Nationwide Employment (%)
Aircraft engines	Hartford, CT	15,619	22.67
	Phoenix, AZ	7,500	10.89
	Cincinnati, OH	6,957	10.10
	Indianapolis, IN	5,872	5.87
Biopharmaceutical products	New York, NY	51,604	27.21
	Chicago, IL	19,754	10.42
	Philadelphia, PA	11,383	6.00
	San Francisco, CA	10,706	5.65
Computer software	Seattle, WA	36,454	20.30
	San Francisco, CA	31,353	9.54
	San Jose, CA	29,221	8.89
	Boston, MA	23,415	7.13
Elevators and moving stairways	Bloomington, IN	1,750	20.02
	New York, NY	1,170	13.39
Financial services	New York, NY	427,296	12.97
	Chicago, IL	151,499	4.60
	Los Angeles, CA	142,337	4.32
	Boston, MA	133,342	4.05
Video production and distribution	Los Angeles, CA	161,561	44.00
	San Francisco, CA	28,394	7.73
	New York, NY	27,541	7.50

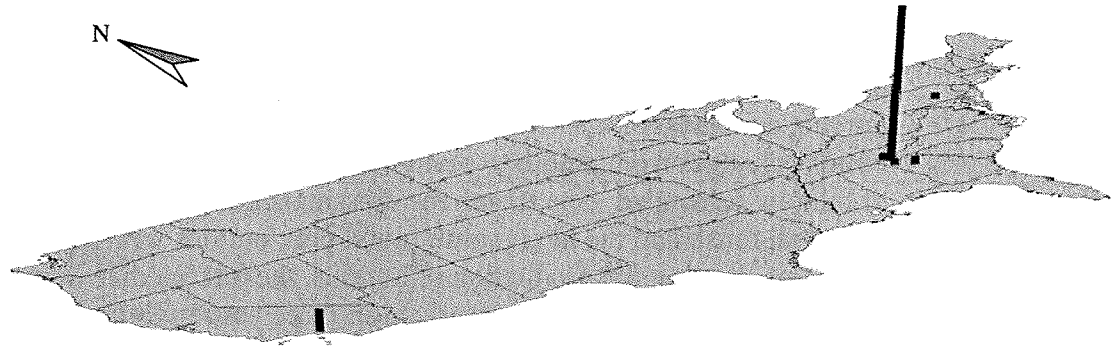
Source: Author's calculations based on data from Cluster Mapping Project, Harvard Business School.

moving stairs, while Los Angeles has over two-fifths of national employment in video production and distribution.

Of course, not all industry clusters occur because of agglomeration economies. We saw in the previous chapter that beet-sugar production facilities are concentrated in beet-growing areas. Similarly, employment in the tobacco-products industry is concentrated in tobacco-growing areas: North Carolina has about 31 percent of national employment in the industry. For the hospitality and tourism industry, two clusters occur in cities with legalized gambling (Las Vegas and Atlantic City).

Maps 3–1 and 3–2 show the locations of job clusters of two industries. Each vertical bar shows the number of jobs in a particular industry in a specific metropolitan area. In Map 3–1, the bars show the job clusters for the carpet and rug industry, which is concentrated in the area around Dalton, GA, home to almost 17,000 jobs in the industry, or 41 percent of the industry's nationwide employment. There are smaller clusters nearby in Atlanta and Chattanooga, and more distant clusters of 2,300 jobs in Los Angeles and 750 jobs in Harrisburg, PA. Map 3–2 shows the job clusters for the production of costume jewelry. The bars show employment clusters for the costume-jewelry industry, which is concentrated in Providence, RI (4,100 jobs, or 55 percent of national employment), with smaller concentrations in New York, Los Angeles, Tampa, and Dallas. For additional maps of employment clusters, visit the Web site of the book.

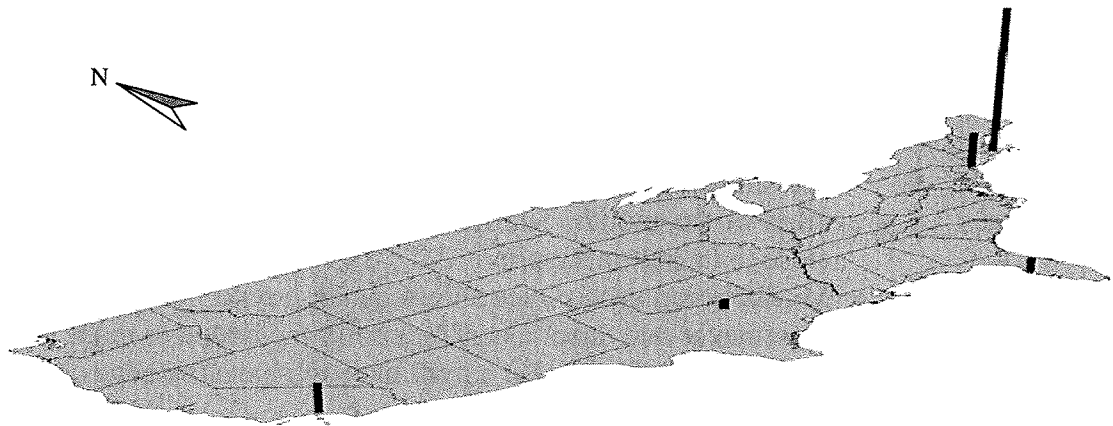
MAP 3-1 Job Clusters: Carpets and Rugs



The bars show employment in the production of carpets and rugs, with 16,790 jobs in Dalton, GA, and smaller clusters in Los Angeles; Atlanta; Chattanooga, TN; Harrisburg, PA; and Rome, GA.

1762653 2014/09/02 157.252.98.229

MAP 3-2 Job Clusters: Costume Jewelry



The bars show employment in the production of costume jewelry, with 4,100 jobs in Providence, RI, and smaller clusters in Los Angeles; New York; Tampa, FL; and Dallas, TX.

SHARING INTERMEDIATE INPUTS

Some competing firms locate close to one another to share a firm that supplies an intermediate input. The conventional list of production inputs includes labor, raw materials, and capital (machines, equipment, structures), but usually ignores intermediate inputs. An intermediate input is something one firm produces that a second firm uses as an input in its production process. For example, buttons produced by one firm are used as inputs by a dressmaking firm. The classic example of a cluster motivated by sharing an intermediate input is a cluster of dressmakers around a buttonmaker (Vernon, 1972).

Dresses and Buttons

Consider the production of high-fashion dresses. The demand for dresses is subject to the whims of fashion, so the dressmaking firms must be small and nimble, ready to respond quickly to changes in fashion. The varying demand for dresses causes varying demands for intermediate inputs such as buttons. A dressmaker's demand for buttons changes from month to month, not in the quantity demanded, but in the type of buttons demanded. One month the dressmaker might use square blue buttons with a smooth finish and the next month round pink buttons with a rough finish.

Consider next the production of buttons, the intermediate input. The production technology for buttons is summarized in three assumptions. The first is one of the axioms of urban economics:



Production is subject to economies of scale

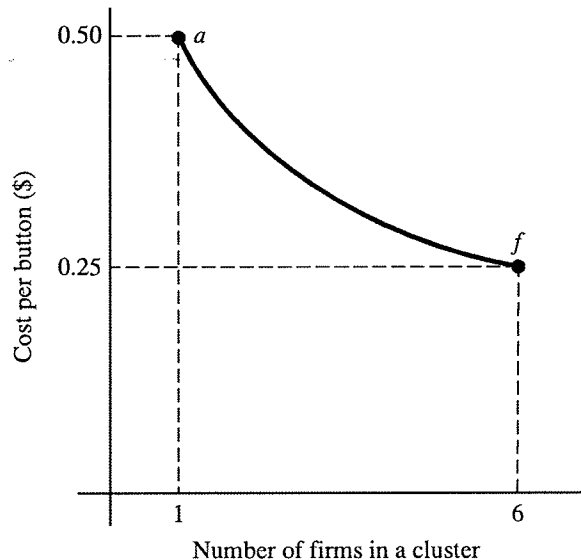
Because button producers use indivisible inputs and specialized labor, the cost per button decreases as the quantity increases. The scale economies are large relative to the button demand of an individual dressmaker, so dressmakers won't produce their own buttons but will buy them as intermediate inputs from button producers. There are two other assumptions in the dress-button model:

- **Face time.** A button for a high-fashion dress is not a standardized input that can be ordered from a catalog or a Web site, but requires interaction between dressmaker and buttonmaker to design and produce the perfect button for the dress of the month. The face time means that a dressmaker must be located close to its button supplier.
- **Modification cost.** Once a dressmaker buys a button from a buttonmaker, the dressmaker may incur a cost to modify the button to make a perfect match. For example, the dressmaker might have to shave the edges of a square button to make it a hexagon.

Figure 3–1 shows the average cost of buttons from the perspective of the dressmaker. Point *a* shows the cost for an isolated dressmaker, which has a relatively high button cost for two reasons. First, the buttonmaker produces for a single dressmaker, so output will be relatively low and the average cost (and price) of buttons will be relatively high. Second, the buttonmaker produces just one type of button (e.g., square buttons), so the dressmaker's modification costs will be relatively high. When the dress of the month calls for square buttons, modifications won't be necessary, but in all the other months, the dressmaker incurs a modification cost.

A dressmaker in a cluster has lower button costs for two reasons. First, a cluster of several dressmakers will generate sufficient button demand to allow buttonmakers to exploit scale economies, leading to lower button prices. Second, the larger total demand for buttons will allow buttonmakers to specialize in different varieties of buttons, reducing the modification costs of dressmakers. In a cluster, a dressmaker might be able to choose from buttons that are squares, hexagons, or triangles. In Figure 3–1, the average cost (and price) of buttons drops from \$0.50 for

FIGURE 3–1 Clustering and the Average Cost of Intermediate Inputs



An isolated firm has a relatively high unit cost of buttons (point *a*). As the number of dressmakers in a cluster increases, the unit cost of buttons decreases because the firms generate sufficient demand to realize scale economies in button production and can support a wider variety of buttons.

an isolated firm (point *a*) to \$0.25 for a six-firm cluster (point *f*). The lower cost provides an incentive for dressmakers to cluster to share a buttonmaker.

High-Technology Firms

The lessons from the button–dressmaker story apply to other industries. Firms producing high-technology products face rapidly changing demand for their cutting-edge products. The small, innovative firms share the suppliers of intermediate inputs, such as electronic components, and cluster to get the face time required to match components and new products. Innovative high-technology firms also share firms that provide product-testing services and locate close enough to quickly when needed tap the facilities.

Intermediate Inputs in the Movie Industry

The U.S. movie industry is concentrated in the area in and around Hollywood, CA. There are seven major studios and hundreds of independent movie producers. Both types of movie producers rely on other firms to provide all sorts of intermediate inputs, such as script writing, film processing and editing, orchestras, and set design and construction. The scale economies associated with producing these inputs are large relative to the demands of individual firms, so movie producers share the suppliers of intermediate inputs. The inputs are not standardized and require face-to-face collaboration in their design and production. The result is a cluster of movie producers and the suppliers of intermediate inputs.

The market for movie props provides an example of intermediate inputs. The objects used in film scenes include mundane items such as table lamps and chairs, special items such as castoff medical instruments and vintage cars, and signature props such as elf ear tips and Griffindor scarves. Although the major studios have their own internal props departments, most independent producers get their props from firms known as “prop houses.” In the Hollywood area, there are three clusters of prop houses. Set decorators and dressers go from one prop house to another looking for the perfect object for the set, and the prop houses cluster to facilitate this comparison shopping.

SELF-REINFORCING EFFECTS CAUSE INDUSTRY CLUSTERS

So far we have seen that clustering is beneficial because it allows firms to take advantage of agglomeration economies from input sharing, labor pooling, skills matching, and knowledge spillovers. What about the costs? In this part of the chapter, we will use an example of the clustering of movie producers to explore the costs and benefits of clustering. When agglomeration economies are strong enough to offset the cost of clustering, firms will form industry clusters, causing the development of specialized cities.

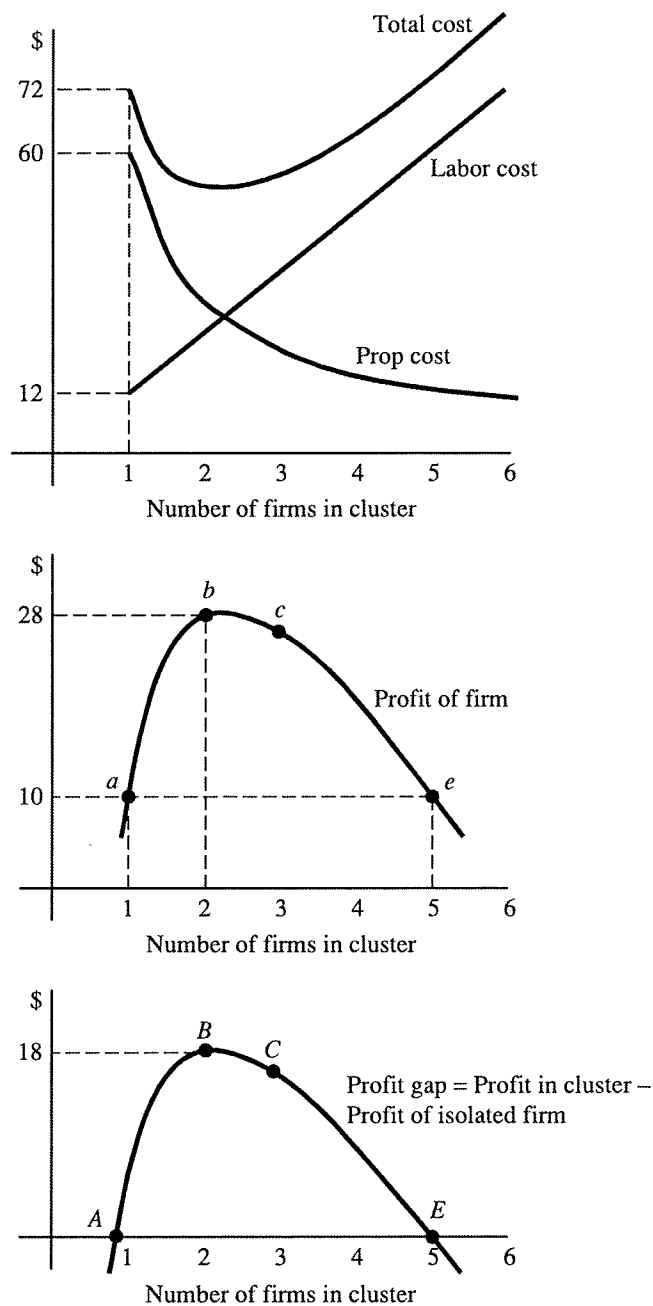
The Benefits and Costs of Clustering

Consider the location decisions of movie producers. Suppose the scale economies in providing movie props are large relative to the demand of an individual producer. As a result, movie producers won't run their own prop departments, but will instead purchase this intermediate input from prop houses. Locating in a cluster allows the producers to share prop houses and benefit from a lower price for props.

Figure 3–2 shows the trade-offs associated with clustering. In the upper panel, the negatively sloped curve shows the prop cost of the typical movie producer, which decreases as the number of producers in the movie cluster increases and the average cost—and price—of props decreases. The positively sloped curve shows the labor cost of the typical movie producer. The larger the number of firms in the cluster, the greater the competition for labor, and thus the higher the wages of movie workers and the higher the labor costs. The U-shaped curve shows the firm's total cost, equal to the sum of prop and labor costs. Going from one to two firms, the savings in prop cost dominate the increase in labor cost, and the total-cost curve reaches its minimum with two movie producers. Beyond that point, the increase in labor cost dominates, generating a positively sloped total-cost curve.

The middle panel of Figure 3–2 shows the profit of a typical movie producer for different numbers of producers in the cluster. Assume that the revenue of the typical producer is constant at \$82, and the profit equals this fixed revenue minus prop costs and labor costs. The profit of an isolated producer (a one-firm cluster), shown by point *a*, is $\$10 = \$82$ in revenue minus \$60 in prop cost and \$12 in labor

FIGURE 3-2 Self-Reinforcing Effects and Clustering



The profit gap, equal to the profit for firm in a cluster, minus the profit of an isolated firm increases, then decreases, reflecting the trade-offs from lower advertising costs and higher labor costs. The profit gap reaches zero with five firms in the cluster, the equilibrium number.

cost. As the number of firms in the cluster increases, profit increases, then decreases. The inverted U reflects decreasing, then increasing total cost. With five producers in the cluster, the profit per producer again equals the profit of the isolated producer (point *e*).

The Profit Gap and the Size of the Cluster

The lower panel of Figure 3–2 shows the gap between the profit of a movie producer in a cluster and the profit of an isolated producer (\$10). The profit gap is of course zero with a one-firm cluster (point *A*). The profit gap grows to \$18 (point *B*, with 2 firms), then shrinks. In a five-firm cluster, the profit gap is again zero (point *E*).

How many movie producers will locate in the cluster? Suppose we have a large number of movie producers, and initially each is isolated, earning a profit of \$10 (point *a* in the middle panel of Figure 3–2). Will this dispersed outcome persist? Suppose a single movie producer relocates next to another, forming a two-firm cluster. As shown in Figure 3–2, each firm in the cluster will earn \$28 (point *b*), or \$18 more than an isolated firm (point *B*). This higher profit gives the remaining isolated firms an incentive to relocate to the cluster. The third firm in the cluster will earn \$26 in the cluster compared to \$10 in isolation—a \$16 gap (point *C*). Firms will continue to join the cluster as long as the profit gap is positive, that is, as long as the cluster location is more profitable than the isolated location. In the stable equilibrium, there are five firms in the cluster (point *E*). At this point, each firm in the cluster earns \$10, the same as an isolated firm.

The agglomeration economies from sharing an intermediate input supplier generate self-reinforcing changes. Recall the second axiom of urban economics:



Self-reinforcing changes generate extreme outcomes

In this case, movie producers that compete for labor don't disperse to minimize labor costs, but instead cluster to realize agglomeration economies. In this example, rising labor costs generate diseconomies of scale that limit clustering, but the same logic applies with other diseconomies, such as rising land costs or rising transport costs for inputs and outputs.

SHARING A LABOR POOL

What do the producers of television programs and the producers of computer software have in common? Every year, dozens of new television programs are aired, and only a few are hits. In the rapidly changing software industry, hundreds of new products are introduced every year, and only a few succeed. For an individual firm in either industry, this year's new product—television program or computer program—may be wildly successful, and next year's may be a dud. In this environment of rapidly changing demand, unsuccessful firms will be firing workers at the same time that successful firms are hiring them. A cluster of firms facilitates the transfer of workers from unsuccessful firms to successful ones.

The key notion of sharing a labor pool is that the boom-bust process occurs at the level of the firm, not the industry. Suppose the total demand for output in an industry is constant over time, but the demand facing an individual firm varies from year to year. For example, the number of slots for television programs is fixed, so the success of one television firm (a hit) comes at the expense of another (a canceled

dud). Similarly, the success of one firm's encryption software comes at the expense of other firms that introduce similar products.

In this part of the chapter, we develop a formal model of labor pooling. The total demand at the industry level is constant, but the demand for each firm varies from year to year. For each firm, there are two possibilities—high demand or low demand—and each outcome is equally likely. As we'll see, there is an incentive for firms in such an industry to cluster to share a pool of workers.

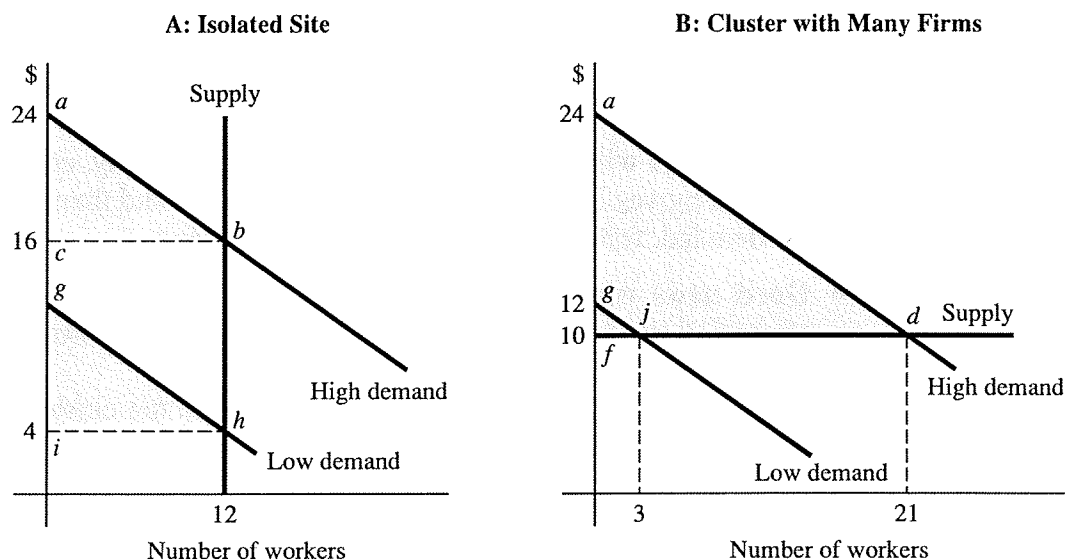
The Isolated Firm

Consider first the situation for an isolated firm outside any industry cluster. The isolated firm doesn't face any competition for labor within its town, and to simplify matters, we assume that labor supply in the isolated site is perfectly inelastic, fixed at 12 workers. This means that wages will rise and fall with the demand for the firm's product.

When the demand for the firm's product is high, so is the firm's demand for labor. In Panel A of Figure 3–3, the high-demand equilibrium is shown by the intersection of the upper demand curve and the vertical supply curve at point *b*, generating a wage of \$16. When demand for the firm's product is low, so is its demand for labor and the equilibrium wage (\$4 at point *h*). To summarize, the isolated firm hires the same number of workers during high and low demand but pays a lower wage when demand is low.

1762653 2014/09/02 157.252.98.229

FIGURE 3–3 Clustering to Share a Labor Pool



In an isolated site, the firm faces a perfectly inelastic supply of labor (12 workers). The firm hires the same number of workers during high demand and low demand but pays a higher wage during high demand.

In a cluster, the firm faces a perfectly elastic supply of labor, and the wage is fixed at \$10. The firm hires 21 workers during high demand but only three workers during low demand.

Locating in a Cluster

The key difference between an isolated site and a cluster concerns the competition for labor and the variability of wages. Workers in the cluster can choose from a large number of firms. For every successful firm hiring workers, there is an unsuccessful firm firing them. Therefore, the total demand for labor in the cluster is constant, and so is the equilibrium wage.

Workers are mobile between the isolated site and the cluster, and in equilibrium they will be indifferent between the two locations. Recall the second axiom of urban economics:



Prices adjust to generate locational equilibrium

At the isolated site, the wage is uncertain, being either \$16 during high demand or \$4 during low demand. The two outcomes are equally likely, so the expected wage (the sum of the probabilities times the wages) is \$10:

$$\text{Expected wage} = \frac{1}{2} \cdot \$16 + \frac{1}{2} \cdot \$4 = \$10$$

To make workers indifferent between the two sites, the certain (constant) wage in the cluster must be \$10.

Panel B in Figure 3–3 shows the outcomes in the cluster. An individual firm can hire as many workers as it wants at the market wage. The typical firm hires 21 workers when demand is high (point *d*), but only three workers when demand is low (point *j*). When the demand for a firm's product goes from high to low, the firm fires 18 workers at the same time that another firm in the cluster is hiring 18 workers as its demand goes from low to high.

Expected Profits Are Higher in the Cluster

Expected profits will be higher in the cluster. To see why, consider what happens when a firm moves from the isolated site to a cluster and then experiences one year of high demand, followed by one year of low demand.

- **Good news when demand is high.** The move to the cluster cuts the wage (from \$16 to \$10) and allows the firm to hire more workers (21 instead of 12), generating higher profit in the cluster.
- **Bad news when demand is low.** The move to the cluster increases the wage (from \$4 to \$10), generating lower profit in the cluster.

Which is larger, the good news with high demand, or the bad news with low demand?

The good news will dominate the bad news because a firm in the cluster responds to changes in the demand for its product. When demand is high, the firm takes advantage of the lower wages (a \$6 gap) in the cluster by hiring more workers (21). When demand is low, a firm in the cluster cushions the blow of low demand by hiring fewer workers (only three). Because the firm changes its workforce

when the demand for its product changes, the good news will be large relative to the bad news, and profit will be higher in the cluster.

Another way to show that profit is higher in the cluster is to compute the expected profits at the two sites. As shown in Section 3 of “Tools of Microeconomics” (the Appendix to the book), the labor-demand curve shows the marginal benefit of labor, the value of output produced by the marginal worker. A firm’s profit from hiring a worker equals the difference between the worker’s marginal benefit and the wage, and a firm’s profit from its entire workforce is shown by the gap between the labor demand curve and the horizontal wage line. In Panel A of Figure 3–3, triangle *abc* shows the profit for an isolated firm when demand is high (\$48), and triangle *ghi* shows the profit when demand is low (\$48). In Panel B, the profit with high demand is shown by triangle *adf* (\$147), and the profit with low demand is shown by triangle *gjf* (\$3). So if the two outcomes are equally likely, the expected profit in the cluster is \$75 (the average of \$147 and \$3), compared to \$48 in the isolated site.

Labor Pooling in the Movie Industry

The U.S. movie industry, concentrated in the area in and around Hollywood, CA, provides an example of the benefits of labor pooling. One segment of the labor market includes workers involved in the craft and technical side of the industry. These workers move periodically from one producer to another as projects come and go, and rely on an “economy of favors,” building and maintaining personal relationships to keep informed about potential jobs and ease the moves from one firm to another. The same phenomenon occurs for creative workers (actors, directors, writers) as they move between firms to work on different projects. In a cluster of movie producers, firms draw from a common labor pool, facilitating the flow of workers between firms.

There are a number of mechanisms that facilitate coordination in the local labor market and improve the flow of workers between firms. Intermediaries such as agents, casting directors, and talent managers match labor demanders and suppliers. Dozens of worker associations, including the Production Assistants Association and the Stuntmen’s Association, provide useful information and training programs to their members. Colleges and universities in the area have professional programs that train students in the production of film and television. These coordinating mechanisms improve the efficiency of the labor market and help maintain the competitive advantages of the movie-industry cluster.

LABOR MATCHING

In a typical economic model of a labor market, we assume that workers and firms are matched perfectly. Each firm can hire workers who have precisely the skills the firm requires. In the real world, things are not so tidy. Workers and firms are not always perfectly matched, and mismatches require costly worker training. As we’ll see, a large city can improve the matching of workers and firms in the untidy real world, decreasing training costs and increasing productivity.

equals its total revenue of \$200 minus the sum of its marketing and labor costs. There are two location options: an isolated site ($n = 1$) or a cluster with up to five corporations.

- a. Use a graph like Figure 3–4 to show the profit gap (profit in cluster – profit in isolation) for one through five corporations.
- b. If initially all corporations are isolated and then one joins another to form a two-corporation cluster, other firms [will, won't] have an incentive to join the cluster because. . . .
- c. In the long-run equilibrium, there will be a cluster of _____ corporations, each of which will earn a profit of _____, differing from the profit of an isolated site by _____.

8. Agglomeration Economies and Auto Row

Chapter 1 uses Auto Row as an example of self-reinforcing changes that lead to extreme outcomes. Consider a city with three isolated automobile dealers, each of which has three buyers per day. The profit per car sold is \$1,000. A two-dealer cluster would get six times as many buyers (18), and a three-dealer cluster would get 12 times as many buyers (36).

- a. Use a graph like Figure 3–4 to show the profit gap (the profit for a firm in a cluster minus the profit for an isolated firm) for one, two, and three dealers.
- b. If initially all dealers are isolated and then one joins another to form a two-dealer cluster, other firms [will, won't] have an incentive to join the cluster because. . . .

9. Personal and Pet Grooming in the Minimall

Suppose that personal grooming and pet grooming are complementary products. Betty Beehive could move her beauty shop from an isolated location to a minimall that also contains Peter's pet-grooming shop. If she moves, she will attract some of Peter's customers and her present profit will increase by \$180. Her current rent is \$100, compared to \$300 in the minimall.

- a. Betty [will, won't] make the move because. . . .
- b. Betty's presence in the minimall would increase Peter's profit by \$100. If you were the manager of the minimall, with the power to set the rent of each tenant, what would you do?

10. Diversify the Economy?

According to the conventional wisdom concerning urban economic development, a city should develop a diverse economy with a large number of industries. Evaluate the merits of this advice in light of the empirical evidence concerning the magnitudes of localization and urbanization economies.

REFERENCES AND ADDITIONAL READING

1. Aarland, K., J. C. Davis, J. Henderson, and Y. Ono, *Spatial Organization of Firms: The Decision to Split Production and Administration*. Providence, RI: Brown University Press, 2003.
2. Audretsch, David, and Maryann Feldman. "Knowledge Spillovers and the Geography of Innovation." Chapter 61 in *Handbook of Regional and Urban*

- Economics 4: Cities and Geography*, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
3. Carlton, D.W. "The Location and Employment Choices of New Firms." *Review of Economics and Statistics* 65 (1983), pp. 440–49.
 4. Dumais, Guy, Glen Ellison, and Edward Glaeser. "Geographic Concentration as a Dynamic Process." *Review of Economics and Statistics* 84 (2002), pp. 193–204.
 5. Duranton, Gilles, and Diego Puga. "Micro-foundations of Urban Agglomeration Economies." Chapter 48 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004. 1762653 2014/09/02 157.252.98.229
 6. Duranton, Gilles, and Diego Puga. "From Sectoral to Functional Specialization." *Journal of Urban Economics* 57 (2005), pp. 343–70.
 7. Fujita, Masahisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
 8. Glaeser, Edward. "Learning in Cities." *Journal of Urban Economics* 46 (1999), pp. 254–277.
 9. Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer. "Growth in Cities." *Journal of Political Economy* 100 (1992), pp. 1126–52.
 10. Hanson, Gordon. "Scale Economies and the Geographic Concentration of Industry." *Journal of Economic Geography* 1 (2001), pp. 255–76.
 11. Harvard Business School, Cluster Mapping Project. <http://data.isc.hbs.edu/isc/>.
 12. Head, K., J. Ries, and D. Swenson. "Agglomeration Benefits and Location Choice." *Journal of International Economics* 38 (1995), pp. 223–48.
 13. Helsley, R., and W. Strange. "Matching and Agglomeration Economies in a System of Cities." *Regional Science and Urban Economics* 20 (1990), pp. 189–212.
 14. Henderson, J. V. "Efficiency of Resource Usage and City Size." *Journal of Urban Economics* 19 (1986), pp. 47–90.
 15. Henderson, J. V., Kuncoro, A., and M. Turner. "Industrial Development and Cities." *Journal of Political Economy* 103 (1995), pp. 1067–81.
 16. Jacobs, Jane. *The Economy of Cities*. New York: Random House, 1969.
 17. Marshall, Alfred. *Principles of Economics*. London: Macmillan, 1920, p. 352.
 18. Mun, Seil, and Bruce G. Hutchinson. "Empirical Analysis of Office Rent and Agglomeration Economies: A Case Study of Toronto." *Journal of Regional Science* 35 (1995), pp. 437–55.
 19. Rosenthal Stuart and William Strange. "The Micro-Empirics of Agglomeration." Chapter 1 in *A Companion to Urban Economics*, eds. R. Arnott and D. McMillen. London: Blackwell, 2005.
 20. Rosenthal, S. S., and W. C. Strange (2001). "The Determinants of Agglomeration," *Journal of Urban Economics* 50 (2001), pp. 191–229.
 21. Rosenthal, S. S., and W. C. Strange. "Geography, Industrial Organization, and Agglomeration," *Review of Economics and Statistics* 85 (May 2003), pp. 377–93.

plant births, with the largest effect on industries that employ college graduates. Their results suggest that knowledge spillovers are important in determining the locations of firms in idea-oriented industries. Rosenthal and Strange (2001) show that the most innovative industries are more likely to form clusters. They also show that knowledge spillovers are highly localized, petering out over a distance of a few miles.

There is also evidence that knowledge spillovers are more important for industries with small, competitive firms. A recent study compared two clusters of the electronics industry, California's Silicon Valley and Route 128 near Boston (Saxenian, 1994). Knowledge spillovers are more important in Silicon Valley because its network of specialized companies generates an atmosphere of collaboration, experimentation, and shared knowledge. In contrast, the firms in the Route-128 cluster are less interdependent so there are fewer knowledge spillovers.

EVIDENCE OF LOCALIZATION ECONOMIES

A large volume of economics literature examines the magnitude of localization economies. In searching for evidence of localization economies, researchers focus on the effects of industry concentration on (1) worker productivity, (2) the number of new production plants (plant births), and (3) growth in industry employment. If there are localization economies, we expect industry clusters to generate higher productivity, more births, and more rapid employment growth.

Consider first the effect of concentration on worker productivity. Henderson (1986) estimates the elasticity of output per worker with respect to industry output, defined as the percentage change in output per worker divided by the percentage change in industry output. For the electrical machinery industry, the elasticity is 0.05, meaning that a 10 percent increase in the output of the industry increases output per worker by 0.50 percent. The elasticity estimates for other U.S. industries range from 0.02 for the pulp and paper industry to 0.11 for the petroleum industry.

Mun and Hutchison (1995) use data from Toronto to estimate agglomeration economies in the office sector. They estimate a productivity elasticity of 0.27, suggesting that localization economies are more powerful in the office sector than in the manufacturing sector. The productivity effects are larger for growth in central locations and are localized.

Consider next the implications of industry concentrations for the location of new production facilities. Carlton (1983) examines the location choices of firms in three industries: plastics products, electronic transmitting equipment, and electronic components. His estimated elasticity of firm births with respect to industry output is 0.43: A 10 percent increase in industry output increases the number of births by 4.3 percent. More recently, Head, Reis, and Swenson (1995) show that Japanese corporations locate their new plants close to other Japanese plants in the same industry. Rosenthal and Strange (2003) show that firm births are more numerous in locations close to concentrations of employment in the same industry.

Consider next the effects of industry concentration on employment growth. Henderson, Kuncoro, and Turner (1995) show that growth in mature industries is

more rapid in areas that start with large concentrations of the industry. Rosenthal and Strange (2003) compute this localization effect for six industries, including computer software. A zip code area that starts out with 1,000 more software jobs than another zip code area experiences a larger increase in software employment—about 12 more jobs. On average, the localization effect peters out at a rate of about 50 percent per mile. The rapid attenuation of the localization economies explains the local in “localization economies.”

URBANIZATION ECONOMIES

So far in this chapter, we have considered agglomeration economies experienced within a particular industry, also known as localization economies. These localization economies generate clusters of firms producing the same product. In contrast, urbanization economies—defined as agglomeration economies that cross industry boundaries—cause firms of different industries to locate close to one another. The result is the development of large, diverse cities. The four agglomeration economies that generate localization economies also generate urbanization economies.

Sharing, Pooling, and Matching

Consider first the notion of input sharing. Although some intermediate inputs such as buttons are specific to an industry, others are shared by firms in different industries. For example, most industries use business services such as banking, accounting, building maintenance, and insurance. Similarly, firms in different industries share hotels and firms providing transportation services. In addition, firms share public infrastructure such as highways, transit systems, ports, and universities. By sharing these intermediate inputs, firms in larger cities pay lower prices and tap a wider variety of inputs.

Another source of agglomeration economies is labor pooling. Recall that labor pooling is beneficial when the product and labor demand per firm varies while total industry demand remains constant. A cluster of firms in the same industry facilitates the movement of workers from firing firms to hiring firms. Labor pooling generates urbanization economies when demand varies across industries, with some industries expanding while others decline.

Consider next the benefits of labor matching. Recall that an increase in a city's workforce increases the density of worker skills, reducing the mismatches between workers' skills and firms' skill requirements. Because some skill requirements are common to multiple industries, the benefits of labor matching cross industry boundaries. For example, firms in many industries require computer programmers, and firms in these industries benefit from producing in a city with a high density of programmers.

Corporate Headquarters and Functional Specialization

Corporations locate their headquarters in cities to exploit urbanization economies. Corporate executives and managers perform a variety of tasks—developing marketing

TABLE 3–3 Increase in Functional Specialization of Metropolitan Areas

Population	Percentage Gap between Metropolitan Ratio of Management to Production Workers and the National Ratio		
	1950	1970	1990
5–20 million	+10.2	+22.1	+39.0
1.5 to 5 million	+0.30	+11.0	+25.7
75,000 to 250,000	–2.1	–7.9	–20.7
67,000 to 75,000	–4.0	–31.7	–49.5

Source: Gilles Duranton and Diego Puga. “From Sectoral to Functional Specialization,” *Journal of Urban Economics* 57 (2005), pp. 343–70.

campaigns, picking locations for new plants, and fending off lawsuits—and draw on other firms to accomplish these tasks. Corporate expenditures on outsourced legal, accounting, and advertising services are equivalent to about two-thirds of their wage bill (Aarland, Davis, Henderson, Ono, 2003).

Corporations cluster to share firms that provide business services. For example, given the large economies of scale in producing advertising campaigns, corporations cluster to share advertising firms, and they get specialized marketing campaigns at a lower cost. Similarly, corporations are attracted by the large concentrations of firms providing financial and business services in midtown Manhattan, the Loop in Chicago, and the financial district of San Francisco.

In the last several decades, there has been a fundamental shift in the specialization of cities. Large cities have become increasingly specialized in managerial functions, while smaller cities have become more specialized in production. Duranton and Puga (2005) compute the ratio of managerial workers to production workers for the nation as a whole and for different metropolitan areas. Table 3–3 shows the percentage differences between the national ratio and the metropolitan ratio for metropolitan areas of different sizes. For example, in the largest metropolitan areas in 1950, the metropolitan ratio was 10.2 percent higher than the national ratio, indicating a slight specialization in managerial functions. At the other extreme, for the smallest areas, the metropolitan ratio was 4.0 percent lower than the national ratio, indicating slight specialization in production.

Over the 40-year period shown in Table 3–3, there was a dramatic change in specialization. By 1990, the ratio for largest cities was 39 percent higher than the national ratio, indicating substantial specialization in managerial functions. At the other extreme, the ratio for the smallest cities was nearly 50 percent lower than the national ratio, indicating a high degree of specialization in production. These changes in specialization were caused by decreases in the cost of managing production facilities from afar. Firms are better equipped to operate multiplant firms from headquarters in large cities where agglomeration economies generate lower production costs. The most important cost reductions have come from innovations in telecommunications, in particular the development of duplicators (photocopiers, fax machines, and e-mail) that have facilitated the rapid transmission of information and reduced the cost of coordination.

Knowledge Spillovers

The essential feature of knowledge spillovers is that physical proximity facilitates the exchange of knowledge between people, leading to new ideas. The ideas lead to new products as well as new ways to produce old products. Some knowledge spillovers occur within an industry, but knowledge spillovers often cross industry boundaries. A city that produces a wide variety of products is fertile ground for applying ideas refined in the design and production of one product to new products.

The evidence for the knowledge spillovers in large cities comes from data on new product innovations (Audretsch and Feldman, 2004). The bulk of patents for new products and production technology is issued to people in large cities. For example, in 1860 the incidence of patents (the number of patents per capita) in the largest 35 U.S. cities was over four times the national average. More recently, the most innovative metropolitan areas in the United States are San Francisco (8.9 innovations per 100,000 population), Boston (8.7), New York (4.2), and Philadelphia (3.6).

The evidence from the paper trail of patents provides additional evidence of knowledge spillovers in large cities. On each patent is a list of related patents—earlier patents that provide a foundation of knowledge for the new patent. A disproportionate number of these citations involve patents from the same metropolitan area. Specifically, the number of local citations (of patents from the same metropolitan area) is five times the number that would occur in the absence of knowledge spillovers.

Following the paper trail of innovation reveals two other features of knowledge spillovers. First, citations to a particular patent are highly localized in the first year following the patent, but the local nature of knowledge spillovers declines over time as knowledge is diffused geographically. Second, the fertility of patents (the number of subsequent citations) varies across the originating institutions: University patents are the most fertile, followed by corporate patents and government patents.

Evidence of Urbanization Economies

There have been many studies of urbanization economies. The general conclusion is that the elasticity of productivity with respect to population is in the range 0.03 to 0.08 (Rosenthal and Strange, 2004). In other words, a doubling of population increases output per worker by between 3 percent and 8 percent. Two studies (Glaeser, Kallal, Scheinkman, and Schleifer, 1992, and Henderson, Kuncoro, and Turner, 1995) suggest that diversity promotes employment growth, especially in new and innovative industries. Hanson (2001) concludes that long-run industry growth is higher in cities with a wider variety of industries, suggesting that diversity promotes growth.

OTHER BENEFITS OF URBAN SIZE

The urbanization economies discussed so far—input sharing, labor pooling, skills matching, and knowledge spillovers—generate higher productivity and lower production costs. In this part of the chapter, we'll consider three other advantages

associated with a larger urban economy: better employment opportunities for families, a better learning environment for workers, and better social opportunities.

These advantages of size increase the relative attractiveness of large cities and increase the supply of labor to big cities. How does that contribute to the clustering of firms in cities? Recall the first axiom of urban economics:



Prices adjust to generate locational equilibrium

An increase in the relative attractiveness of a big city decreases the wage that workers are willing to accept to live and work in the city, generating lower production costs for firms. This is similar to the Dullsville versus Coolsville example in Chapter 1: A city that has superior opportunities for family employment, learning, and social interactions has lower wages, everything else being equal.

Joint Labor Supply

Most families have two workers, but are tied to a single residential location. In other words, families must confront the problem of joint labor supply. If the skills of the two workers are suited to different industries, the family will be attracted to locations with a mix of industries. Therefore, the joint supply of labor encourages firms in different industries to cluster. The role of cities in resolving the issue of joint labor supply has a long history. In the 1800s, mining and metal-processing firms (employing men) located close to textile firms (employing women), and each industry benefited from the presence of the other. More recently, “power couples” (defined as a pair of college graduates) are concentrated in large cities, where they are more likely to find good employment matches for both workers.

Learning Opportunities

Another benefit of urban size comes from the greater learning opportunities in cities. Human capital is defined as the knowledge and skills acquired by workers in formal education, work experience, and social interaction. Human capital can be increased through learning by imitation, that is, observing other workers and imitating the most productive workers. A larger city provides a wider variety of role models for workers so it attracts workers looking for learning opportunities.

The evidence for urban learning comes from data on the wages earned by workers who migrate to cities (Glaeser, 1999). Wages are higher in cities, reflecting the higher productivity of urban workers. But when a worker migrates from a rural area, she doesn't earn the higher urban wage immediately. Instead she experiences rising wages over time as learning increases her productivity. When a worker leaves the city, her wage does not drop back to the wage she earned prior to coming to the city. Instead, the higher productivity resulting from urban learning leads to a higher wage outside the city. In other words, the benefits of urban learning translate into higher wages everywhere.

Social Opportunities

A third benefit of city size comes from social interactions. An implicit assumption of the backyard-production model in Chapter 2 is that people do not value social interaction. Of course, people enjoy interacting with one another, and a larger city provides more opportunities for social interactions.

To think about the social dimension of cities, recall the labor-matching model. Suppose we replace labor skills with social interests: People have different hobbies, conversational topics, and social activities. In addition, suppose we replace firms seeking good skills matches with people seeking a network of friends with similar interests. In a model of social-interest matching, a larger city will generate better interest matches, with each network (like each firm) achieving a tighter range of social interests. Some people live in cities to take advantage of better opportunities for social-interest matching.

To illustrate the notion of social benefits of large cities, suppose you want to form a book club to discuss your favorite book, *Giles Goat Boy* (by John Barth). In a small town, you may be the only person who has read the book. In contrast, thousands of people in the typical large city have read the book and perhaps a dozen will be eager to discuss the masterpiece. A quick Internet search reveals that larger cities have more book clubs on a wider variety of topics, consistent with the notion that bigger cities provide better social matches.

SUMMARY

Firms cluster to exploit agglomeration economies, including localization economies at the industry level and urbanization economies at the city level. Here are the main points of the chapter:

1. Firms may cluster to share a supplier of an intermediate input if the input is subject to relatively large scale economies and requires face time for its design and production.
2. Firms may cluster to share a labor pool if the variation in product demand is greater at the firm level than at the industry level.
3. Larger cities provide better skill matches, leading to higher productivity and wages.
4. People and firms are attracted to cities because they facilitate knowledge spillovers, learning, and social opportunities.
5. Agglomeration economies cause self-reinforcing changes in location: The movement of one firm to a city increases the incentive for other firms to move to the city.

APPLYING THE CONCEPTS

For exercises that have blanks (____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as

necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. Attention Kmart Shoppers

Most of the dresses sold in the United States are produced in large factories that are dispersed, not concentrated. Reconcile this fact with the text discussion of localization economies that cause dressmakers to cluster.

2. Labor Pooling: What's Fixed and Variable?

Consider the model of labor pooling, with each firm locating either in an isolated site or in a cluster with other firms. Fill the blanks with "fixed" or "variable."

- In an isolated site, the wage is _____ and the firm's workforce is _____ because. . . .
- In a cluster, the wage is _____ and the firm's workforce is _____ because. . . .
- Illustrate with two graphs, one for the isolated site and one for the cluster.

3. Trade-offs with Clustering for Labor Pooling

Consider the model of labor pooling, with each firm locating either in an isolated site or in a cluster with other firms. Suppose that good times (high demand) and bad times (low demand) are equally likely. The table shows wages and workforces in different times and locations.

	Isolated		Cluster	
	Wage	Workforce	Wage	Workforce
Good times (high demand)	\$40	50	\$30	60
Bad times (low demand)	\$20	50	\$30	40

- Use a figure like Figure 3–2 to illustrate the situation.
- During good times, the benefit of being in the cluster as opposed to being isolated is _____, computed as. . . .
- During bad times, the cost of being in the cluster as opposed to being isolated is _____, computed as. . . .
- The benefit exceeds the cost because a firm in a cluster. . . .

4. Mr. Mullet's Carnival

Mr. Mullet runs a traveling carnival that hires local workers in each city it visits. The demand for carnival activities is uncertain, with low or high demand equally likely in any given city. At the end of the year, Mr. Mullet reviews his financial records and discovers some puzzling differences between his experiences in small and large cities.

- He always paid the same wage in large cities (\$9), but paid different wages in small cities (\$6 or \$12).
 - He always hired the same quantity of labor in small cities (20 workers), but different quantities in big cities (10 or 30 workers).
- Using Figure 3–2 as a model, illustrate with two graphs, one for the typical small city and one for the typical big city. Assume that the demand curves

for labor are linear and parallel, with vertical intercepts of \$18 (high demand) and \$12 (low demand).

- b. In the typical big city with high demand, profit is _____ computed as. . . .
- c. In the typical big city with low demand, profit is _____ computed as. . . .
- d. In the typical small city with high demand, profit is _____ computed as. . . .
- e. In the typical small city with low demand, profit is _____ computed as. . . .
- f. The expected profit is _____ in a big city, compared to _____ in a small city.

5. Number of Workers and Net Wages

Using Table 3–2 as a starting point, suppose the gross wage is \$36 and the unit training cost is \$48. Complete the following table.

Number of Workers	Skills Gap	Training Cost	Net Wage
4	_____	_____	_____
8	_____	_____	_____
24	_____	_____	_____

6. Models on the Color Wheel

Consider the model-management industry, with firms that supply human models for advertisements. Workers (models) vary in skin tone along the color wheel, which can be divided into 12 colors. Firms enter the market with a specific skin tone requirement for their models. If a model's skin tone does not match the firm's tone requirement, the model incurs a makeup cost to close the gap, with a cost of \$3 for each unit of color shift. For example, to go from color #2 to color #4, the cost is \$6. Given the scale economies in model management, each firm manages three models. The gross wage is \$20.

- a. Smallville has six models, equally spaced on the color wheel at 12:00, 2:00, 4:00, and so on. There will be _____ firms in the city, with firm A at 12:00 and the other firm or firms at _____.
- b. Illustrate with a graph like Figure 3–3.
- c. For the typical firm, the average mismatch is _____ skin tones and the average makeup cost is \$ _____. The average net wage after makeup costs is _____, computed as. . . .
- d. Bigburg has twice as many models as Smallville. It will have _____ firms. Its average makeup costs will be [lower, higher] and its average net wage will be [higher, lower].
- e. Complete the following table.

Number of Models	Skills Gap	Makeup Cost	Net Wage
6	_____	_____	_____
12	_____	_____	_____

7. Advertising and Corporate Clusters

Consider corporations that use advertising firms to develop marketing campaigns. Each corporation buys one campaign per year, and the cost per campaign is $\$120/n$, where n = the number of corporations in the cluster (and campaigns per year). The cost of labor per firm is $\$30 \cdot n$. A corporation's profit

equals its total revenue of \$200 minus the sum of its marketing and labor costs. There are two location options: an isolated site ($n = 1$) or a cluster with up to five corporations.

- a. Use a graph like Figure 3–4 to show the profit gap (profit in cluster – profit in isolation) for one through five corporations.
- b. If initially all corporations are isolated and then one joins another to form a two-corporation cluster, other firms [will, won't] have an incentive to join the cluster because. . . .
- c. In the long-run equilibrium, there will be a cluster of _____ corporations, each of which will earn a profit of _____, differing from the profit of an isolated site by _____.

8. Agglomeration Economies and Auto Row

Chapter 1 uses Auto Row as an example of self-reinforcing changes that lead to extreme outcomes. Consider a city with three isolated automobile dealers, each of which has three buyers per day. The profit per car sold is \$1,000. A two-dealer cluster would get six times as many buyers (18), and a three-dealer cluster would get 12 times as many buyers (36).

- a. Use a graph like Figure 3–4 to show the profit gap (the profit for a firm in a cluster minus the profit for an isolated firm) for one, two, and three dealers.
- b. If initially all dealers are isolated and then one joins another to form a two-dealer cluster, other firms [will, won't] have an incentive to join the cluster because. . . .

9. Personal and Pet Grooming in the Minimall

Suppose that personal grooming and pet grooming are complementary products. Betty Beehive could move her beauty shop from an isolated location to a minimall that also contains Peter's pet-grooming shop. If she moves, she will attract some of Peter's customers and her present profit will increase by \$180. Her current rent is \$100, compared to \$300 in the minimall.

- a. Betty [will, won't] make the move because. . . .
- b. Betty's presence in the minimall would increase Peter's profit by \$100. If you were the manager of the minimall, with the power to set the rent of each tenant, what would you do?

10. Diversify the Economy?

According to the conventional wisdom concerning urban economic development, a city should develop a diverse economy with a large number of industries. Evaluate the merits of this advice in light of the empirical evidence concerning the magnitudes of localization and urbanization economies.

REFERENCES AND ADDITIONAL READING

1. Aarland, K., J. C. Davis, J. Henderson, and Y. Ono, *Spatial Organization of Firms: The Decision to Split Production and Administration*. Providence, RI: Brown University Press, 2003.
2. Audretsch, David, and Maryann Feldman. "Knowledge Spillovers and the Geography of Innovation." Chapter 61 in *Handbook of Regional and Urban*

- Economics 4: Cities and Geography*, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
3. Carlton, D.W. "The Location and Employment Choices of New Firms." *Review of Economics and Statistics* 65 (1983), pp. 440–49.
 4. Dumais, Guy, Glen Ellison, and Edward Glaeser. "Geographic Concentration as a Dynamic Process." *Review of Economics and Statistics* 84 (2002), pp. 193–204.
 5. Duranton, Gilles, and Diego Puga. "Micro-foundations of Urban Agglomeration Economies." Chapter 48 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004. 1762653 2014/09/02 157.252.98.229
 6. Duranton, Gilles, and Diego Puga. "From Sectoral to Functional Specialization." *Journal of Urban Economics* 57 (2005), pp. 343–70.
 7. Fujita, Mashisa, and Jacques-Francois Thisse. *Economics of Agglomeration*. Cambridge: Cambridge University Press, 2002.
 8. Glaeser, Edward. "Learning in Cities." *Journal of Urban Economics* 46 (1999), pp. 254–277.
 9. Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer. "Growth in Cities." *Journal of Political Economy* 100 (1992), pp. 1126–52.
 10. Hanson, Gordon. "Scale Economies and the Geographic Concentration of Industry." *Journal of Economic Geography* 1 (2001), pp. 255–76.
 11. Harvard Business School, Cluster Mapping Project. <http://data.isc.hbs.edu/isc/>.
 12. Head, K., J. Ries, and D. Swenson. "Agglomeration Benefits and Location Choice." *Journal of International Economics* 38 (1995), pp. 223–48.
 13. Helsley, R., and W. Strange. "Matching and Agglomeration Economies in a System of Cities." *Regional Science and Urban Economics* 20 (1990), pp. 189–212.
 14. Henderson, J. V. "Efficiency of Resource Usage and City Size." *Journal of Urban Economics* 19 (1986), pp. 47–90.
 15. Henderson, J. V., Kuncoro, A., and M. Turner. "Industrial Development and Cities." *Journal of Political Economy* 103 (1995), pp. 1067–81.
 16. Jacobs, Jane. *The Economy of Cities*. New York: Random House, 1969.
 17. Marshall, Alfred. *Principles of Economics*. London: Macmillan, 1920, p. 352.
 18. Mun, Seil, and Bruce G. Hutchinson. "Empirical Analysis of Office Rent and Agglomeration Economies: A Case Study of Toronto." *Journal of Regional Science* 35 (1995), pp. 437–55.
 19. Rosenthal Stuart and William Strange. "The Micro-Empirics of Agglomeration." Chapter 1 in *A Companion to Urban Economics*, eds. R. Arnott and D. McMillen. London: Blackwell, 2005.
 20. Rosenthal, S. S., and W. C. Strange (2001). "The Determinants of Agglomeration," *Journal of Urban Economics* 50 (2001), pp. 191–229.
 21. Rosenthal, S. S., and W. C. Strange. "Geography, Industrial Organization, and Agglomeration," *Review of Economics and Statistics* 85 (May 2003), pp. 377–93.

22. Rosenthal, Stuart, and William Strange. "Evidence on the Nature and Sources of Agglomeration Economies." Chapter 49 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. Vernon Henderson and Jacques-Francois Thisse. Amsterdam: Elsevier, 2004.
23. Saxenian, Annalee. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press, 1994.
24. Scott, Allen J. *On Hollywood: The Place, the Industry*. Princeton New Jersey: Princeton University Press, 2005.
25. Vernon, Raymond. "External Economies." In *Readings in Urban Economics*, eds. M. Edel and J. Rothenberg. New York: Macmillan, 1972.

CHAPTER 4

City Size

There is no need to worry about mere size. Sir Isaac Newton was very much smaller than a hippopotamus, but we do not on that account value him less.

—BERTRAND RUSSELL

New York, the largest urban area in the United States, has a population of more than 18 million, while the smallest urban area (Andrews, Texas) has a population of about 13,000. As shown in Table 4–1 (page 69), there are a few very large cities, a moderate number of medium-size cities, and many small cities. In this chapter, we'll explore the economic forces responsible for the development of cities of different size. We'll also explore why cities differ in their economic scope—from highly specialized cities to diverse cities.

UTILITY AND CITY SIZE

The previous chapter explained how agglomeration economies cause firms to cluster. As we'll see in this chapter, these agglomeration economies increase productivity and wages, so workers in larger cities earn higher wages. As a city grows, the benefits of higher wages are at least partly offset by several undesirable features of larger cities, including longer commuting times, greater density, and more congestion and pollution. Given these trade-offs, the key question is:

How does an increase in city size (population) affect the utility of the typical worker?

Benefits and Costs of Bigger Cities

We are interested in the relationship between city size and the utility level of the typical worker. Consider a city where production occurs at a single point, and workers commute from a residential area to the production center. We'll start with the benefits of a larger city, then turn to the costs.

TABLE 4–1 Size Distribution of Urban Areas, 2000

Population Range	Number of Urban Areas
Greater than 10 million	2
5 million to 10 million	4
1 million to 5 million	43
100,000 to 1 million	324
Less than 100,000	549

TABLE 4–2 Utility and City Size

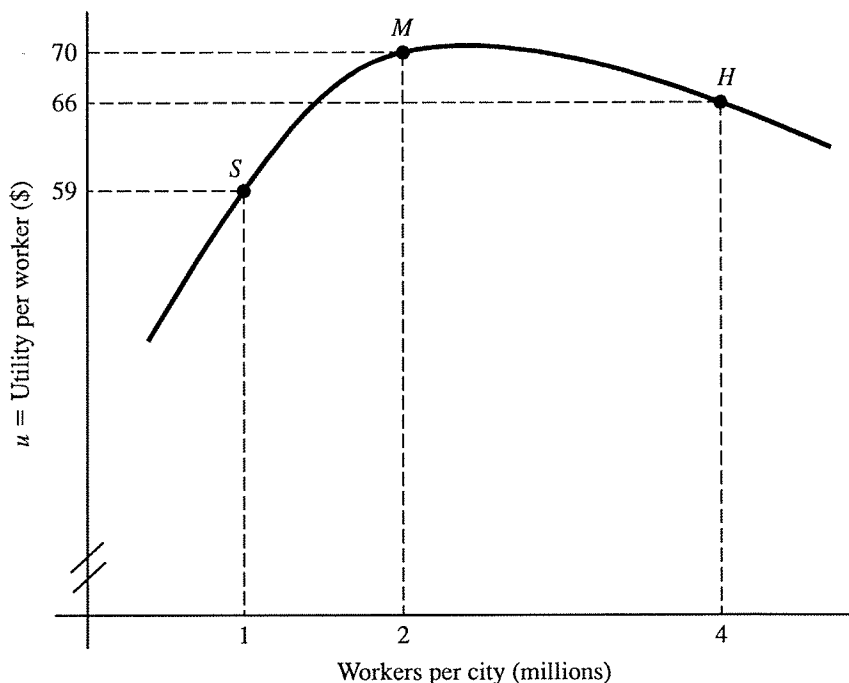
Workforce (millions)	Wage	Labor Income	Commuting Cost	Utility
1	\$ 8	\$64	\$ 5	\$59
2	\$10	\$80	\$10	\$70
4	\$11	\$88	\$22	\$66

As we saw in the previous chapter, agglomeration economies—from input sharing, labor pooling, skills matching, and knowledge spillovers—increase labor productivity. In a competitive labor market, competition among firms ensures that wages reflect labor productivity so wages are higher in larger cities. Table 4–2 provides a simple example of the relationship between wages and city size. As shown in the second column, the wage increases at a decreasing rate, reflecting the assumption that agglomeration economies diminish as the city grows: Labor productivity increases with the size of the workforce, but at a diminishing rate.

To simplify matters, let's assume that the only cost of population growth is longer commuting time. Suppose commuting time comes at the expense of leisure time, and we can place a dollar figure on the value of leisure time lost to commuting. As shown in the fourth column of Table 4–2, the cost of commuting increases with city size. A doubling of the workforce from 1 million to 2 million doubles the commuting cost per worker from \$5 per day to \$10. Doubling the workforce again more than doubles commuting cost.

The last column of Table 4–2 shows the utility level of the typical worker. For now, we'll define utility as income minus the value of leisure time lost from commuting. We assume that each person works an eight-hour day, so income is eight times the wage. Moving from a city of 1 million to a city of 2 million, the increase in labor income (from higher productivity) is large relative to the increase in commuting cost, so utility increases from \$59 to \$70. In Figure 4–1, this is shown as a move from point *S* to point *M* on the utility curve. In other words, agglomeration economies are stronger than the diseconomies of scale associated with commuting, so utility increases. In contrast, an increase in the workforce from 2 to 4 million decreases utility because agglomeration economies are weaker than the diseconomies from commuting. In Figure 4–1 the city's utility curve reaches \$70 at a workforce of 2 million.

FIGURE 4-1 Utility and City Size



An increase in city size increases wages because of agglomeration economies and increases commuting costs. As long as agglomeration economies are stronger, utility increases with city size. When agglomeration economies are weaker than the diseconomies from commuting, utility decreases as city size increases.

Locational Equilibrium, Land Rent, and Utility within a City

So far we have ignored the location decisions of workers within a particular city. Consider a city with a workforce of 2 million (shown by point *M* in Figure 4-1), where workers commute from different residential locations to the city's production center. Workers differ in their commuting costs. Recall the first axiom of urban economics:



Prices adjust to achieve locational equilibrium

In this case, the price of residential land will adjust to make workers indifferent among all residential locations.

Table 4-3 (page 71) shows how differences in commuting generate differences in land rent. Let's compare two workers, one who lives near the production center and thus incurs no commuting cost, and a second who has a commuting cost of \$10. If the worker at the center has a rent of \$25, the rent for the other worker must be \$10 lower, or \$15. The same logic suggests that a worker 10 miles from the center will pay only \$5 for land. Workers are indifferent between the three residential locations because the differences in commuting cost are exactly offset by differences in land rent.

TABLE 4–3 Commuting, Land Rent, and Utility within a City

A Commute Distance	B Commute Cost	C Land Rent Paid	D Labor Income	E Rental Income	F Utility
0	0	\$25	\$80	\$15	\$70
5 miles	\$10	\$15	\$80	\$15	\$70
10 miles	\$20	\$ 5	\$80	\$15	\$70

Who gets the rent from land? Suppose that workers own land, and to keep things simple, assume that land rent is shared equally among the city's workers. The average rent is \$15 (the rent paid by the worker living five miles from the center), and as shown in column E of Table 4–3, each worker earns \$15 of rental income to supplement labor income. For the average worker (five miles from the center), land rent paid equals the rental income received. In contrast, for the worker living at the center, the rent paid exceeds rental income, while the opposite is true for the worker living 10 miles from the center.

The last column of Table 4–3 shows the utility for workers at different locations in the city. We can define utility as

$$\text{Utility} = \text{Labor income} + \text{rental income} - \text{commute cost} - \text{rent paid}$$

Each worker earns the same \$95 of total income (labor plus rental income). For locational equilibrium, the differences in commuting costs are exactly offset by differences in land rent paid, so workers living at different locations reach the same utility level, \$70.

A SYSTEM OF CITIES

We can use the utility curve to explore how a region's workforce can be distributed among its cities. The issue is whether the region will have a large number of small cities, a small number of large cities, or something between the two extremes. Consider a region with a total urban workforce of 6 million and three possible configurations:

- Six cities {A, B, C, D, E, F}, each with a workforce of 1 million.
- Three cities {D, E, F}, each with a workforce of 2 million.
- Two cities {E, F}, each with a workforce of 3 million.

Cities Are Not Too Small

We can use Figure 4–2 to explore the feasibility of the alternative configurations. Consider first the six-city outcome, with each city housing a workforce of 1 million. As shown by point S, utility in each of the six cities is \$59. Is this a stable equilibrium, or will workers have an incentive to move from one city to another?

FIGURE 4-2 Cities May Be Too Large But Not Too Small



The utility curve reaches its maximum with 2 million workers in a city (point *M*), so a region with 6 million workers will maximize utility with three cities, each with 2 million workers. The outcome with six small cities (point *S*) is unstable because the utility curve is positively sloped. The outcome with two big cities (point *L*) is stable because the utility curve is negatively sloped.

To demonstrate the instability of the six-city outcome, imagine that a group of workers moved from city A to city D. The workforce of city D will grow, causing the city to move upward along the positively sloped portion of its utility curve, leading to a higher utility level, for example, \$60. At the same time, the workforce of city A will shrink, moving the city downward along its utility curve to a lower utility level, for example, \$58. In other words, the movement of workers from city A to city D opens a utility gap of \$2, which will encourage more workers to move from A to D.

Because the utility curve is positively sloped near point *S*, migration is self-reinforcing. The more workers who migrate, the larger the utility gap between the two cities and the greater the incentive to relocate. For example, if the workforce of city D grows to 1.2 million and the workforce of city A shrinks to 0.8 million, utility in city D will be about \$61, compared to about \$57 in city A. Recall the second axiom of urban economics:



Self-reinforcing effects generate extreme outcomes

The extreme outcome is that everyone will relocate from city A to city D, so city A will disappear.

The same logic of self-reinforcing migration applies to other cities in the region. Suppose that cities A, B, and C shrink and eventually disappear, while D, E, and F

grow, each eventually doubling in size. In this case, we reach point *M* in Figure 4–2, with each of the region's three cities housing a workforce of 2 million and reaching a utility level of \$70. The three-city outcome happens to be the optimum, where utility is maximized.

Cities May Be Too Large

What happens if we start out with a small number of large cities? Suppose we start with two large cities, each with a workforce of 3 million. In Figure 4–2, the starting point in each city is shown by point *L*. Each city (*E*, *F*) has a workforce of 3 million and reaches a utility level of \$68. For all workers in the region, utility is lower than the maximum level. Is the two-city outcome a stable equilibrium?

To show why this is a stable equilibrium, consider the effects of migration from city *E* to city *F*. The workforce of city *F* will grow, so the city will move downward along the negatively sloped portion of the utility curve (from point *L* toward point *z*) and reach a lower utility level (e.g., \$67). At the same time, the workforce of city *E* will shrink, moving the city upward along the utility curve to a higher utility level (e.g., \$69). In other words, the migration of workers opens a utility gap of \$2, but utility is higher in the smaller city, not the larger one.

In this case, migration is self-correcting, not self-reinforcing, and will be reversed. The migrant workers who migrated reach a lower utility level in their new city, so they'll regret the move and perhaps return to their original city. Alternatively, other workers in the larger city now have lower utility than is available in the smaller city, giving them an incentive to move, effectively swapping places with the original migrants. In either case, migration will be reversed, restoring the original workforces and utility levels in the two cities.

Why is the situation with small cities unstable, while the situation with large cities is stable? The reason is shown by the utility curve. With small cities, the utility curve is positively sloped because agglomeration economies are stronger than the diseconomies of scale resulting from commuting. The utility of the migrants increases because their new city is larger and more efficient, and the utility of people left behind decreases because they now live in a smaller city. In contrast, when we start with large cities, the utility curve is negatively sloped because agglomeration economies are weaker than commuting diseconomies. The utility of migrants decreases because their new city is larger but less efficient, and the utility of workers left behind increases because they now live in a smaller, more efficient city.

The general lesson from this discussion is that cities tend to be too large rather than too small. The too-small outcome occurs when a region has a city on the positively sloped portion of the utility curve. This triggers self-reinforcing migration that eliminates some of the small cities and causes the others to grow. These self-reinforcing changes do not occur when cities are too large, so cities that are inefficiently large persist.

SPECIALIZED AND DIVERSE CITIES

Do cities specialize in a narrow set of economic activities, or do they generalize, producing a diverse mix of products? It turns out that the “or” in this question is misplaced: The typical region contains a wide variety of cities, from highly specialized cities to highly diverse ones (Henderson, 1988). The specialized cities develop because of localization economies, while the diverse cities develop because of urbanization economies.

In fact, specialized and diverse cities are actually complementary, serving different roles in a market economy. Many firms start their lives in a diverse city and eventually relocate to a specialized city. Diverse cities foster new ideas and experimentation so they serve as laboratories for innovative firms. Once a firm settles on a product design and production process, production is likely to be more efficient in a specialized city that fully exploits localization economies. In other words, diverse cities foster innovation, while specialized cities facilitate efficient production.

A Model of Laboratory Cities

We can use a model developed by Duranton and Puga (2001) to explore the role of cities in innovation and production. Consider a firm that is looking for the ideal production process for a new product. By experimenting with different processes, the firm will eventually find the ideal one. At that point, the firm will switch to mass production and start earning a profit. Where should the firm experiment—in a diverse city or a specialized city?

Consider first a firm that experiments in a diverse city until it discovers the ideal process and then relocates to a specialized city. An experiment entails producing a prototype of the firm’s new product with a particular production process. Suppose there are six potential production processes, and the experimenting firm can observe other firms using these processes in the diverse city, and then imitate these other firms to produce a prototype. In addition, suppose that, on average, it takes the firm three years to discover the ideal process. Once the entrepreneur discovers the ideal process, the firm will move to a specialized city and start making a profit.

The alternative scenario is to experiment—search for the ideal process—in the region’s specialized cities. As usual, there are some trade-offs:

- **Good news: lower prototype cost.** The cost of producing a given prototype will be lower in a specialized city because each city has the specialized inputs for that production process.
- **Bad news: higher moving cost.** The search for the ideal process requires the firm to move from one specialized city to another. On average, a firm adopting this strategy will have three moves, compared to a single move for the firm that experiments in the diverse city. If moving costs are large relative to the savings in prototype costs, profit will be lower when the firm experiments in specialized cities.

1762653 2014/09/02 157.252.98.229

This model of laboratory cities shows the roles of diverse and specialized cities in the product cycle. A diverse city has a rich variety of products and production processes, providing fertile ground for new ideas about how to produce new products. Once a firm finds its ideal production process, the benefits of being in a diverse environment diminish, so the firm relocates to a specialized city, where localization economies generate lower production cost.

Example: The Radio Industry in New York

Vernon (1972) identifies the radio industry in New York as the classic example of an industry that developed in a diversified city:

In the 1920s that industry had all the earmarks of an activity whose establishments were heavily dependent on external [agglomeration] economies, speed, and personal contact. Its technology was unsettled and changing rapidly; its production methods were untried; its market was uncertain. Accordingly, at that stage, producers were typically small in size, numerous, agile, nervous, heavily reliant upon subcontractors and suppliers. Mortality in the industry was high. In those circumstances, the attraction of an urban area like the New York Metropolitan Region was especially strong.

New York was attractive because it provided a wide variety of intermediate inputs and a large and diverse workforce. The area also provided production knowledge—embodied in a wide variety of production processes—that proved useful in developing a production process for the radio.

Vernon explains why the radio industry eventually left the New York metropolitan area:

A decade or two later, however, the technology of the industry had settled down. Production methods were standardized and sets were being turned out in long runs. Now, the critical competitive questions had become transport and labor costs, rather than product design. The small firm faded from the picture and large assembly plants appeared at lower-range locations more centrally placed for national markets.

When a product reaches its mature stage, with a settled design and established production process, producers have less to gain from diversified cities, and they can relocate to places with lower production costs because of localization economies, lower wages, or lower land rent.

Evidence of Laboratory Cities

Duranton and Puga (2001) provide evidence for the notion that diversified cities serve as laboratories for firms in innovative industries. Using data from firms in France, they show that among firms that change locations, over 7 in 10 relocated from a diverse city to a specialized city. As firms mature, they relocate from diverse cities with urbanization economies to specialized cities with localization economies.

The most innovative industries have the highest frequency of relocations from diverse to specialized cities. For example, the frequency is 93 percent for research and development, 88 percent for pharmaceuticals and cosmetics, and 82 percent for information technology. The frequency is high for other industries, including

business services, printing and publishing, aerospace equipment, and electronic equipment. In contrast, the frequency of movement to specialized cities is relatively low for less innovative sectors such as furniture, food, beverages, clothing, and leather.

Much of the recent work on urbanization economies focuses on how city diversity affects plant births and employment growth. Rosenthal and Strange (2004) summarize the most recent studies and conclude that diversity encourages both births and growth in employment, especially in high-technology industries.

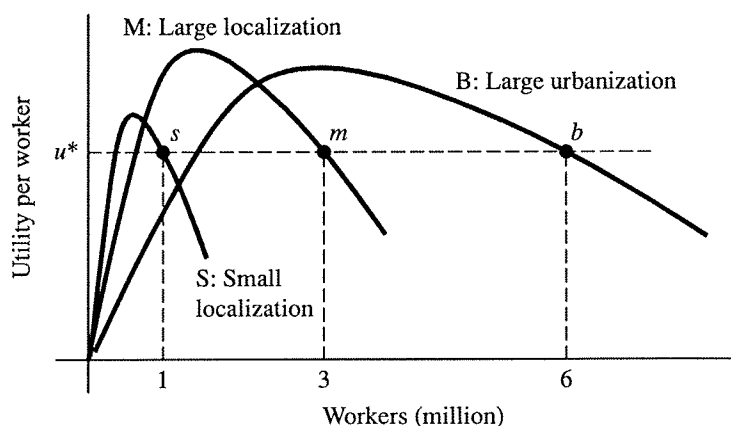
DIFFERENCES IN CITY SIZE

Table 4–1 shows the wide range of city sizes in the United States. Other countries have a similar pattern of differences in city sizes. In this part of the chapter, we'll explore the roles of localization and urbanization economies in determining city size. We'll also explore the role of consumer goods.

Differences in Localization and Urbanization Economies

Figure 4–3 shows utility curves for three types of cities in a regional economy. The curve on the left applies to an industry for which localization economies are exhausted with a relatively small workforce. In this case, the diseconomies of commuting quickly overwhelm agglomeration economies, so the optimum city size is relatively small. The middle utility curve is for a specialized city with larger localization economies and thus a larger optimum size. Finally, the utility curve on the right is for a city that experiences large urbanization economies, generating a large optimum size.

FIGURE 4–3 Differences in City Sizes from Differences in Agglomeration Economies



City S has small localization economies and its optimum population is smaller than that of city M with its large localization economies. City B has large urbanization economies and a large population. The set of points $\{s, m, b\}$ shows a possible equilibrium, with all residents achieving the utility level u^* and populations of 1 million (city S), 3 million (city M), and 6 million (city B) adding up to the regional population of 10 million.

Locational equilibrium requires that workers in the region are indifferent between the three cities, meaning that workers in the three cities must achieve the same utility level. Suppose the region has a total of 10 million workers. In Figure 4–3, points s , m , and b show one possible equilibrium. The utility level, u^* is the same in all three cities, and the city workforces (1 million in the small city, 3 million in the medium-size city, and 6 million in the big city) add up to the regional workforce (10 million). As we saw earlier in the chapter, this is a stable equilibrium because each city is on the negatively sloped portion of its utility curve.

Local Goods and City Size

So far our discussion of city size has focused on employment in industries subject to localization and urbanization economies. In other words, we have ignored the consumer side of the urban economy. We can distinguish between employment in industries that export their output to people outside the city and those that sell their products locally to residents of the city. For example, most of the cars produced in Detroit are sold to people in other cities, while most of the donuts produced in Detroit are sold to city residents, as are most haircuts and groceries. Total employment in a city is the sum of export employment and local employment.

Some local products are available in all cities, large and small. If the per-capita demand for a product is large relative to the scale economies associated with producing it, even a small city will generate sufficient demand to support at least one firm. For example, it takes just a few thousand people to support a barber, so even a small city will have at least one barber. Similarly, a pizzeria can be supported by a few thousand people, so even a small city will have pizzerias and pizza workers. Of course, a larger city has more hair to cut and more people to feed, so it will have more barbers and pizza tossers. In fact, we expect the number of barbers and pizza tossers to increase proportionately with city size.

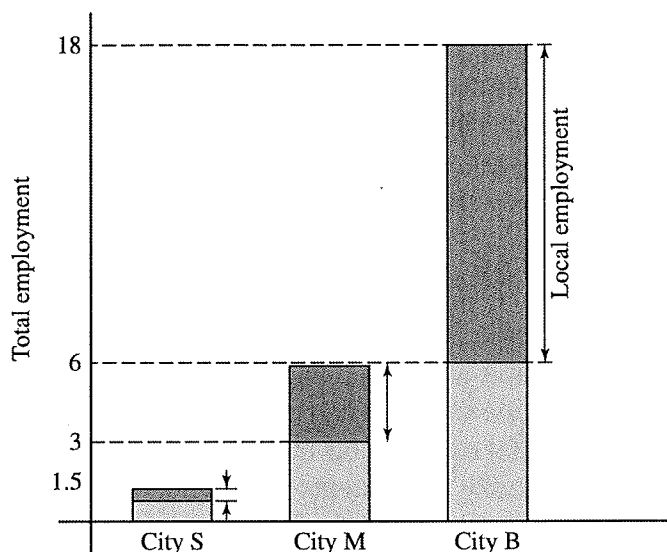
Some local products are available only in large cities. If the per-capita demand for a product is low relative to the scale economies in production, it will take a large city to generate enough demand to support a firm. For example, the per-capita demand for opera is relatively small, so it may take a million people to support an opera company. As a result, we will find opera companies in large cities but not small ones. Similarly, the per-capita demand for brain surgery is low relative to the scale economies in production, so brain surgeons operate in large cities.

Larger cities have a wider variety of consumer products. In a large city, consumers can get everything available in small cities (pizzas and haircuts) as well as products not available in small cities (opera and brain surgery). In fact, people in small cities travel to large cities to buy products that are not available locally. In contrast, consumers in large cities can buy almost anything they want locally so there is little reason to travel to smaller cities.

Local Employment Amplifies Size Differences

Figure 4–4 (page 78) shows the implications of local employment for the size of cities with different levels of export employment. Suppose that in a city with export

FIGURE 4-4 The Introduction of Local Goods Amplifies Differences in City Size



Introducing local consumer amplifies the differences in population arising from differences in export employment. The population of the small city increases by half, while the population of the medium city doubles and the population of the large city triples.

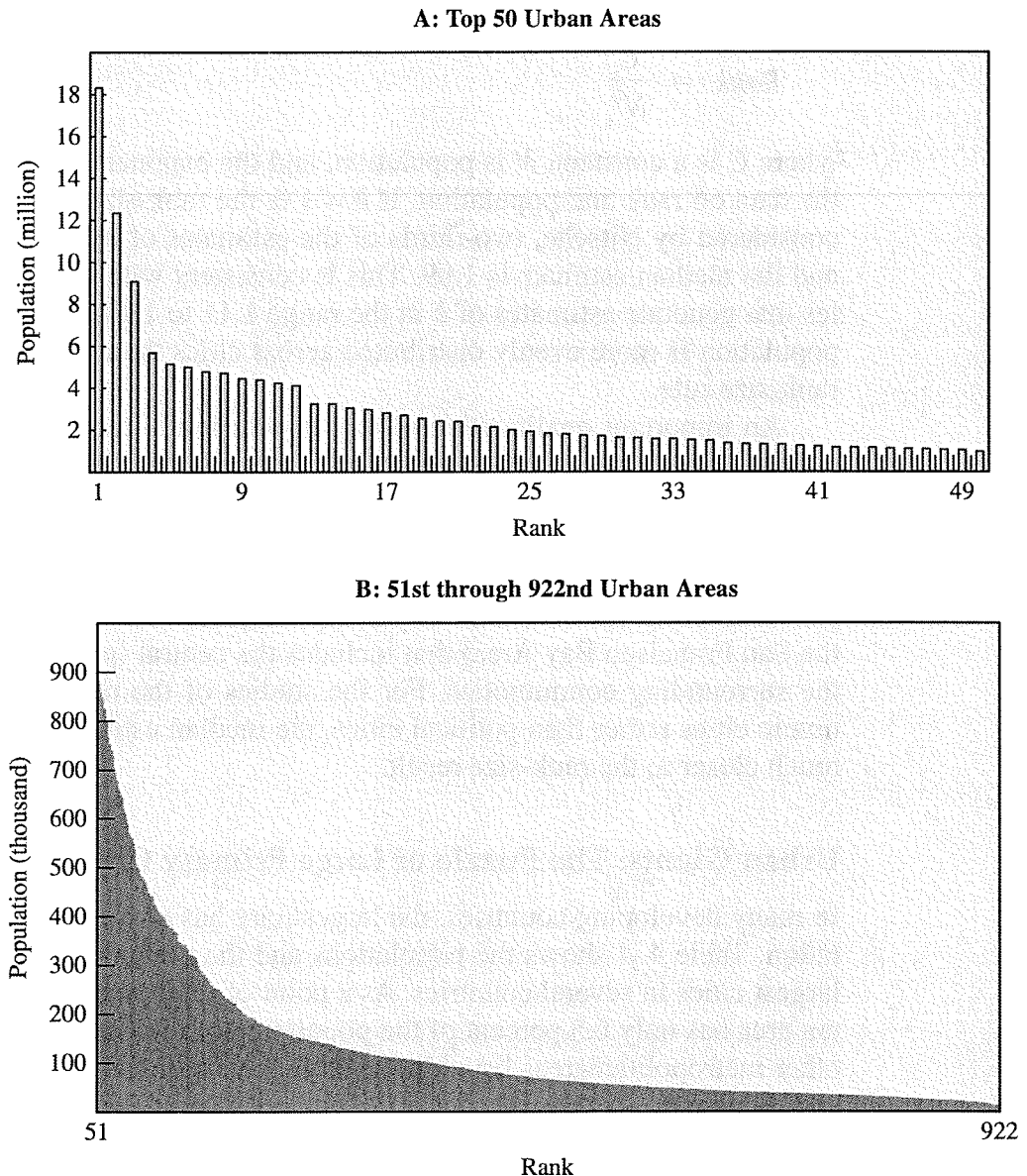
employment of 1 million, each export job supports half a job in local industry. If so, total employment in the city is 1.5 times its export employment, or 1.5 million. We know that a larger city can support a wider range of consumer goods. Suppose that in a city with export employment of 3 million, each export job supports one local job. As shown in Figure 4-4, total employment in such a city is 6 million (3 million export jobs plus 3 million local jobs). Finally, suppose that in a city with 6 million export jobs, each export job supports two local jobs. In this case, total employment is 18 million (6 million export jobs plus 12 million local jobs).

As shown in Figure 4-4, local employment amplifies differences in population. Total employment increases by half in the small city (1 to 1.5 million), while it doubles in the medium-sized city (from 3 to 6 million) and triples in the large city (from 6 to 18 million). After incorporating local employment, the largest city has 12 times the total employment of the small city, up from six times before considering local employment. This occurs because the large city has a larger consumer base and can support a wider variety of products.

THE SIZE DISTRIBUTION OF CITIES

Figure 4-5 shows the size distribution of cities split into two graphs, one for the top 50 urban areas and a second for the rest of the urban areas. The figure is drawn with the new census definitions of urban areas, with metropolitan areas (urban areas with populations exceeding 50,000) and micropolitan areas (smaller urban areas).

FIGURE 4–5 Size Distribution of U.S. Urban Areas, 2000



The Rank-Size Rule

Geographers and economists have estimated the relationship between city rank and size. One possibility is that the relationship follows the rank-size rule:

Rank times population is constant across cities

In other words, if the largest city (rank 1) has a population of 24 million, the second largest city will have a population of 12 million ($12 \cdot 2 = 24$), while the third largest will have a population of 8 million ($8 \cdot 3 = 24$), and so on.

Nitsche (2005) analyzes the results of 29 studies of the rank-size relationship, with data from countries around the world. The hypothesized relationship is

$$\text{Rank} = \frac{C}{N^b}$$

where C is a constant, N is population, and the exponent b is to be estimated from the data on rank and population. If $b = 1.0$, the rank-size rule holds. In the studies considered by Nitsche, two-thirds of the estimates of b are between 0.80 to 1.20, and the median estimate is 1.09. This is consistent with earlier cross-country studies that generate estimates of b in the range 1.11 to 1.13. In other words, the urban population is more evenly distributed across cities than would be predicted by the rank-size rule.

An important qualification of this conclusion is that many of the studies use political definitions of cities rather than the economic definition. A political city is defined by boundaries that separate political jurisdictions. In contrast, the economic definition of a city ignores political boundaries and includes in a city's population all the people who are economically involved in a particular urban economy. In practical terms, the economic city is typically defined as a metropolitan area (e.g., the San Francisco Bay Area) that includes the central (political) city along with all the surrounding communities. For the studies of the rank-size rule that use economic cities rather than political cities, the median estimate for b is 1.02, which is much closer to the rank-size result.

Urban Giants: The Puzzle of Large Primary Cities

In many developing countries, the largest city has a relatively large share of population. Table 4–4 shows the populations and the national population shares for the largest cities in several countries. As a point of reference, the New York metropolitan area has only 6.5 percent of the population of the United States. In contrast, the other metropolitan areas listed in the table have national shares between 11 percent and 39 percent.

Economists have developed several models to explain the large concentrations of population in primary cities. One theory is based on the idea that large economies of scale in trade encourage the development of a single large trading city rather than

TABLE 4–4 Population of Largest Cities as Share of National Population

Metropolitan Area	Population	Share of National Population
Tokyo	19,037,361	15.76%
Mexico City	16,465,487	20.97
São Paulo	15,538,682	11.46
Buenos Aires	10,759,291	35.47
Santiago, Chile	4,227,049	34.87
Montevideo, Uruguay	1,157,450	39.36

several smaller ones. For example, the substantial investment in a port facility encourages nations to designate a primary port city. English trade grew rapidly between 1520 and 1670, a result of military victories against the Spanish, improved shipping technology, and the expansion of markets in Asia and the Americas. The population of London increased during this period from 55,000 to 475,000, and its share of England's population increased from 2 percent to 10 percent. In Argentina, between 1887 and 1914 exports quadrupled, and the population of Buenos Aires—the principal trading city—increased by 1.1 million (265 percent).

Transportation infrastructure plays a role in urban concentration. In many developing countries, a disproportionate share of investment in roads and telecommunication facilities occurs in and around the capital city. The relatively low investment in infrastructure outside the capital area generates high transport costs within the country and encourages the development of large primary cities.

What is the role of politics in the development of large primary cities? Ades and Glaeser (1995) suggest that nations run by dictators have larger primary cities than democracies. One way for a dictator to stay in power is to take resources from the hinterland (areas outside the capital city) and transfer these resources to the people who are most likely to overthrow him—people in the capital city. If a dictator pays off local agitators, the capital city will grow as some people migrate to the city to get the payoffs and others migrate for jobs in local industry that are supported by the bribes. Based on a study of cities in 85 countries, Ades and Glaeser conclude that capital cities in countries with dictatorships are 45 percent larger than capital cities in other countries.

The experience of Rome illustrates the role of redistribution policies on urban concentration. In the period 130–50 B.C., the population of Rome increased from 375,000 to 1 million, making Rome more than twice as large as any city up to that point in history. Military successes during this period extended the empire into Gaul and the eastern provinces of Asia, providing a large hinterland from which to extract resources. The rulers responded to political unrest in Rome by distributing free grain to the residents and staging the infamous—and very expensive—Roman circuses.

SUMMARY

This chapter explains why cities come in different sizes. Here are the main points of the chapter:

1. The utility curve shows the trade-offs from an increase in population: Agglomeration economies increase productivity and wages, but diseconomies of scale from increased commuting costs reduce utility.
2. Cities are unlikely to be too small because such an outcome is not a stable equilibrium: Migration is self-reinforcing because the growing city becomes more productive while the shrinking city becomes less productive.
3. Cities may be too large because such an outcome is a stable equilibrium: Migration is self-correcting because the shrinking city becomes more productive while the growing city becomes less productive.

4. Local employment amplifies differences in workforces and population across cities.
5. A diverse city fosters experimentation and leads to innovations in product design and production.
6. The rank-size rule provides a rough approximation to the size distribution of cities.

APPLYING THE CONCEPTS

For exercises that have blanks (____), fill each blank with a single word or number. For exercises with ellipses (. . .), complete the statement with as many words as necessary. For exercises with words in square brackets ([increase, decrease]), circle one of the words.

1. Migration to Portland

In the last several months, the number of cars driving in Portland with Washington State license plates has increased dramatically. Suppose this reflects migration from Seattle to Portland. Use two utility curves, one for Portland and a second for Seattle, to represent this migration. Assume that both cities are on the negatively sloped portions of their utility curves.

- a. Label the current position for Portland (population 2 million) with "P" and the current position for Seattle (population 3 million) with "S." Use arrows to indicate the direction of movement along each utility curve.
- b. In Portland, migration [decreases, increases] utility because _____ of scale from _____ dominate _____ of scale from _____.
- c. Label the long-run equilibrium point for Portland as "Q" and the long-run equilibrium point for Seattle as "T."

2. Heli-Segways for Workers

Consider a region with 6 million workers and two cities (A and B), each with an initial workforce of 3 million (the utility-maximizing workforce). Suppose the heli-segway (a flying personal transporter) replaces the automobile, cutting commuting cost and increasing the utility-maximizing urban workforce per city to 5 million. The positively sloped part of the new utility curve is steeper than the negatively sloped part.

- a. Show the effect of the heli-segway on the urban utility curve in the typical city.
- b. A workforce of 3 million [is, isn't] a stable equilibrium because. . .
- c. Using the new utility curves, show the changes in the workforces of two cities, one that grows and one that shrinks. Label the initial point *i* and indicate the directions of the changes with arrows.
- d. The new equilibrium number of cities is _____ and each city has a population of _____.

3. A Free Circus and City Size

Consider a region with a fixed population of 2 million. The urban utility curve reaches the maximum utility with 1 million people, and initially there are two

cities, R and S, each with 1 million people. The positively sloped portion of the urban utility curve is steeper than the negatively sloped portion. Suppose a dictator in city R starts to provide free circuses, financed by coercive transfer payments from people outside the region.

- Show the effect of the free circuses on the urban utility curve in city R. Assume that the utility-maximizing population doesn't change.
- People will migrate from city _____ to city _____ because. . .
- Arrows up or down: Migration _____ utility in city R and _____ utility in city S.
- Show the new equilibrium distribution of population between the two cities. The population in city R is _____ and the population in city S is _____

4. Formation of New Cities

Consider a region with a workforce of 12 million in a single city. The urban utility curve reaches its maximum with 3 million workers and includes the following combinations (W = workers; U = utility in \$):

W	1	2	3	4	5	6	7	8	9	10	11	12
U	32	56	70	65	60	55	50	45	40	35	30	25

Suppose the government establishes a new city with 1 million workers, leaving 11 million workers in the old city. Assume that the number of cities remains at two.

- Immediately following the establishment of the new city, the utility in the small new city is _____ and the utility in the large old city is _____
- On the utility curve, mark the position of the new city with "N" and the position of the old city (immediately following the formation of the new city) with "D." Use arrows to indicate the direction of movement for each city.
- In the long-run equilibrium, the workforce of the new city = _____ with utility = _____; the workforce of the old city = _____ with utility = _____

5. New Cities Boom or Bust?

Consider a region with a workforce of 12 million in a single city. The urban utility curve reaches its maximum with 3 million workers and includes the following combinations (W = workers; U = utility in \$):

W	1	2	3	4	5	6	7	8	9	10	11	12
U	23	33	43	48	52	55	50	45	40	35	30	25

Suppose the government establishes a new city with 1 million workers, leaving 11 million workers in the old city. Assume that the number of cities remains at two.

- Immediately following the establishment of the new city, the utility in the small new city is _____ and the utility in the large old city is _____, so the new city will [shrink, grow] and the old city will [shrink, grow].
- In the long-run equilibrium, the workforce in the old city = _____.
- A new city will be viable in the long run if the starting workforce is at least _____ because. . .

6. Rural versus Urban Development

In the initial equilibrium, a nation's workforce of 24 million is divided equally between a city and a rural area. The initial common utility level is \$50. The urban utility curve has the conventional hump shape, with a peak at 8 million workers. In the population range 8–12 million, the slope of the utility curve is $-\$2$ per million workers. The rural utility curve is subject to mild increasing returns to scale. In the population range 12–16 million, the slope is $\$1$ per million workers. Suppose the nation invests in rural infrastructure, increasing rural productivity and shifting the rural utility curve upward by $\$3$.

- Depict graphically the effects of this infrastructure investment on the two sectors.
- People will migrate from _____ to _____ because. . . .
- Migration [increases, decreases] utility in the rural area and [increases, decreases] utility in the city, with a [larger, smaller] change in the [city, rural area].
- Show the new equilibrium distribution of population between the rural area and the city. In the new equilibrium, the population of the city is _____; the population of the rural area is _____; the common utility level is _____.

7. Size Distribution Example

Consider a region with three industries subject to localization economies, each of which employs 300 workers. There is a single cluster for industry L (all firms in the industry locate in a single city labeled "L"), three clusters for industry M (firms are divided equally between three cities, labeled "M1," "M2," "M3"), and five clusters for industry S (firms are divided equally between five cities, labeled "S1," "S2," . . . , "S5"). There are no urbanization economies. In a city with 300 export jobs, there are three local jobs for each export job; in a city with 100 export jobs, there are two local jobs for each export job; in a city with 60 export jobs, there is one local job for each export job.

- Complete the following table.

Export Workers	Local Workers	Total
City S	_____	_____
City M	_____	_____
City L	_____	_____

- Compared to the smallest city, the largest city has _____ times the export employment _____ but times the total employment. The difference in total employment is relatively large because. . . .

8. Specialized Services in Large Cities

Complete the statement: Large urban areas provide specialized cultural, legal, medical, financial, and other services that are not available in small urban areas because these specialized services are characterized by. . . .

9. Boring Capital Cities

Some people claim that state capitals (Sacramento, CA; Salem, OR; Olympia, WA) are boring in the sense that they have a smaller variety of goods and

services than other cities of equal size. Check a map and then provide an explanation why these cities could be boring.

10. One City Size

Consider a region with two export products (gloves and socks) and two local goods (tattoos and manicures). The production of each export good is subject to localization economies, so each city specializes in one export good. According to Mr. Wizard, "If my two assumptions (one for export products and one for local goods) are correct, all the cities in the region will be the same size." Assume that Mr. Wizard's logic is correct. List his assumptions and explain why together they imply the region's cities will be the same size.

REFERENCES AND ADDITIONAL READING

1. Ades, Alberto F., and Edward L. Glaeser. "Trade and Circuses: Explaining Urban Giants." *Quarterly Journal of Economics* (1995), pp. 195–227.
2. Abdel-Rahman, H., and A. Anas. "Theories of Systems of Cities." Chapter 52 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
3. Audretsch, D., and M. Feldman. "Knowledge Spillovers and the Geography of Innovation," Chapter 61 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
4. Duranton, G., and D. Puga. "Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products." *American Economic Review* 91 (5), 1454–77 (2001).
5. Gabaix, X., and Y. Ioannides. "Evolution of City Size Distributions," Chapter 53 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
6. Glaeser, E. "Learning in Cities." *Journal of Urban Economics* 46 (1999), pp. 254–77.
7. Henderson, V. *Urban Development: Theory, Fact, and Illusion*. Oxford: Oxford University Press, 1988.
8. Nitsche, V. "Zipf Zipped." *Journal of Urban Economics* 57 (2005), pp. 86–100.
9. Rosenthal, S., and W. Strange, "Evidence on the Nature and Sources of Agglomeration Economies," Chapter 49 in *Handbook of Regional and Urban Economics 4: Cities and Geography*, eds. V. Henderson and J. F. Thisse. Amsterdam: Elsevier, 2004.
10. U.S. Department of Transportation. *Summary of Travel Trends: 2001 National Household Travel Survey*. Washington D.C., December 2004.
11. Vernon, Raymond. "External Economies." In *Readings in Urban Economics*, ed. M. Edel and J. Rothenberg. New York: Macmillan, 1972.

Appendix: CENTRAL PLACE THEORY

Central place theory, which was developed by Christaller (translated in 1966) and refined by Losch (translated in 1954), shows how the location patterns of different industries combine to form a regional system of cities, with a small number of large cities and a large number of small cities.

The starting point for central place theory is the observation that the market areas of individual firms vary across industries. If the economies of scale in production are large relative to per-capita demand for the product, each firm will need a relatively large area to sell a large enough quantity to exhaust its scale economies. For example, the per-capita demand for brain surgery is low, and the specialized equipment used in brain surgery generates large economies of scale. As a result, a brain-surgery center draws patients from a large area and locates in a large city. In contrast, the per-capita demand for haircuts is high relative to the scale economies in haircutting, so the market area of a barber shop is small and even small towns have at least one barber shop.

Consider a region with three consumer products: books, pizzas, and jewelry. The region has the following characteristics:

1. **Population density.** The initial distribution of population is uniform. The total population of the region is 80,000.
2. **No shopping externalities.** Shopping externalities normally occur with complementary goods (one-stop shopping) and imperfect substitutes (comparison shopping). The simple central place model assumes that there are no shopping externalities.
3. **Ubiquitous inputs.** All inputs are available at all locations at the same prices.
4. **Uniform demand.** For each product, per-capita demand is the same throughout the region.
5. **Number of stores.** The three goods have different per-capita demands and scale economies:
 - a. **Jewelry.** Scale economies are large relative to per-capita demand. Every jewelry store requires a population of 80,000, so a single jeweler will serve the entire region.
 - b. **Books.** Scale economies are moderate relative to per-capita demand. Each bookstore requires a population of 20,000, so there will be four bookstores in the region.
 - c. **Pizza.** Scale economies are small relative to per-capita demand. Every pizza parlor requires a population of 5,000, so there will be 16 pizza parlors in the region.

In the central place model, firms base their location decisions exclusively on access to their consumers. Because production costs are the same at all locations