I. Introduction

Agglomeration economies are the benefits that come when firms and people locate near one another together in cities and industrial clusters. These benefits all ultimately come from transport costs savings: the only real difference between a nearby firm and one across the continent is that it is easier to connect with a neighbor. Of course, transportation costs must be interpreted broadly and they include the difficulties in exchanging goods, people and ideas. The connection between agglomeration economies and transport costs would seem to suggest that agglomerations should become less important as transportation and communication costs fallen. Yet a central paradox of our time is that in cities, industrial agglomerations remain remarkably vital despite ever easier movement of goods and knowledge across space.

Declining transport costs has facilitated trade between China, India and the rest of the world, but within those countries, development has centered in urban areas. Across the world, urbanization continues to increase and the United Nations reports that by the end of 2008, one half of the world will live in cities.[1] Indeed, mega-cities have become the gateways between those developing countries and the developed world. Within the richer nations of the West, many cities, like New York and London, have experienced remarkable comebacks since the dire days of the 1970s. Wages, population and especially housing prices in many dense centers have experienced robust growth. Indices of industrial agglomeration show only a slight decrease in concentration over the last 30 years (Dumais, Ellison and Glaeser, 2002). If transport costs are so low, then why has the urge to agglomerate remained so strong?

This volume collects eleven essays on the economics of agglomeration. They cover far-ranging topics from the productivity of hospitals to the location of fast food joints, yet they are all joined by a common goal of seeking to understand why economic activity

[1] http://www.un.org/esa/population/meetings/EGM_PopDist/EGM_PopDist_Report.pdf.

clusters together. Making sense of this clustering is the crucial step in understanding the present and future economics of place. All of these essays approach agglomeration economies from different angles, but taken together, the volume is meant to provide a sample of cutting edge work on the economics of agglomeration.

Measuring Agglomeration— Prices, Wages, Quantities

Urban economists infer urban success from high local wages, robust real estate prices, and growth in the number of people within an area. If a place is doing well, then employers should be willing to pay more for workers in that area, people should be willing to pay more for access to that place and more people should move to that area. The first three essays in the volume separately consider these three different measures of local economic well-being.

Over the last 45 years, the spatial equilibrium has been the primary tool for urban and regional economists trying to make sense of cities. The logic of the spatial equilibrium is that since people can move freely within a nation, they must be indifferent between different locales. This indifference implies that high wages must be offset by high prices or low amenities, otherwise people would flock to high wages areas. High housing prices reflect high wages or high amenities or both.

But the spatial equilibrium concept only gives us one-half of the labor market equilibrium that determines area wages. The other half is labor demand, the willingness of firms to pay for their workers. So while high wages must reflect something bad about an area, like high prices or poor amenities, high wages must also reflect something good about an area that makes firms willing to tolerate a high cost of labor. Firms wouldn't continue to locate in New York City or the San Francisco Bay region unless those areas were productive enough to offset the cost of expensive workers.

Neoclassical economics tells us that wages reflect the marginal product of labor. In a standard Cobb-Douglas formulation of the producer's problem, where most capital is

mobile, the high marginal product of labor in a given area must either reflect a high productivity level or an abundance of non-traded capital inputs to production. Wages, therefore, can be interpreted as telling us about the core determinants of urban productivity and high wages in an area are usually interpreted as meaning that the area is unusually productive.

One of the facts that supports the existence of agglomeration economies is the strong relationship between density and high wages. This fact is mirrored in the strong relationship between area density and per capita Gross Metropolitan Product shown in Figure 1-1. This fact is quite statistically robust, but the causal chain in the relationship is difficult to infer. Does the density-productivity relationship mean that the dense places become more productive or that productive places attract more people? The need to tease out the direction of causality in this relationship motivates the first essay in this volume, on agglomeration in France, written by Pierre-Philippe Combes, Gilles Duranton, Laurent Gobillon and Sebastion Roux.

Their paper looks at the connection between density and both wages and total factor productivity in France. They start by confirming the existence of a strong, robust relationship between density and both wages and productivity in France. This fact parallels the well known density-productivity relationship in the U.S. (Ciccone and Hall, 1996). They then consider two challenges to interpreting this fact as evidence for agglomeration economies. One possibility is that dense places are more productive because they attract more skilled workers. Glaeser and Mare (2001) find little evidence that this is the case in U.S. cities, but the selection of the skilled into cities seems to be much stronger in France. They use an individual fixed effects approach and find that allowing for individual fixed effects reduces the estimated impact of density on wages by about one-third.

Their second contribution is to use a wide range of historical and geological instruments for current density levels. Population patterns in France are remarkably permanent. The density of districts in France today is highly correlated with density 170 years ago and

with basic features of the soil. Their instrumental variables estimates are generally quite close to estimates found using ordinary least squares. As long as we believe that these instruments are not independently correlated with productivity today, then this provides evidence for strong agglomeration economies. If readers doubt that this orthogonality condition holds, then their results, at least, provide a striking set of facts about the correlation between geology and prosperity.

Real estate prices provide a second means of assessing the success of an area. One sign that agglomeration has been well over the last twenty-five years is that housing prices have risen more dramatically in dense metropolitan areas. Figure 1-2 shows the xx percent correlation between density in 1980 and price growth between 1980 and 2006 (calculated using the Office of Federal Housing Enterprise Oversight repeat sales index). The spatial equilibrium framework suggests that this fact can either mean that dense places have become more pleasant over time or that dense places have become more productive.

But the growth in housing prices has not been uniformly experienced across all metropolitan areas. Some places, like San Francisco and New York City, have been christened "Superstar Cities," by the authors of the second essay in this volume, Joseph Gyourko, Christopher Mayer and Todd Sinai. Their essay documents the extraordinary price growth of a small set of urban areas, which has continued decade-by-decade since 1940 and then tries to understand the causes for price growth in these areas.

Broadly speaking, high prices in a region can reflect economic vitality that pushes up wages, consumer amenities that increases the willingness to pay to live in an area, or rigid housing supply. Gyourko, Mayer and Sinai argue that rising prices in superstar cities cannot be completely explained by rising productivity levels in those areas. They argue instead that these places have high amenities and restrictions on housing supply. Rising levels of inequality in the country as a whole have led the wealthiest Americans to be willing to pay more and more to live in high amenity areas of the country.

The growth of population or employment provides a third means of measuring local success. If housing supply is neither perfectly elastic nor perfectly inelastic, then a boom in local productivity will increase both wages and population in an area. In places with more elastic housing, area-level success should show up primarily in the form of larger population levels not in higher wages or higher housing prices.

The concentration of people and industries has long been seen by economists as evidence for the existence of agglomeration economies. After all, why would so many people suffer the inconvenience of crowding into the island of Manhattan if there weren't also advantages from being close to so much economic activity. However, there is a debate about interpreting the concentration of people and firms, just as there is about interpreting the connection between density and productivity. People and firms might be clustering because of some innate advantage possessed by a particular spot of earth, not just because of agglomeration economies. Indeed, in the 19th century, some of Manhattan's mass appeal may well have reflected the natural advantages bestowed by its remarkable port.

Today, it is harder to believe that industrial and urban clusters reflect natural advantage rather than agglomeration economies. The statistical work that has tried to assess the importance of natural advantage to geographic concentration finds that only about one-quarter of industrial concentration can be explained by observable sources of natural advantage (Ellison and Glaeser, 1999). But all of the work measuring the clustering of population has tended to measure agglomerations based on political boundaries. These political boundaries are often drawn around existing agglomerations, and this creates an inherent bias in using political borders.

If political boundaries are drawn in a way that reflects existing population patterns, then we might think that we observe agglomerations of activity even when there was no innate tendency for clustering. Even a random distribution of population across space will be lumpy. While some measures of industrial concentration correct for that lumpiness (Ellison and Glaeser, 1997), standard corrections for lumpiness can do little if the

geographic units are drawn around the lumps. In many cases, the statistical properties of spatial areas would be far easier to understand if geographic areas were defined by a fixed grid, rather than political boundaries.

This problem is particularly severe when thinking about this distribution of city sizes across, generally described by Zipf's Law. If larger cities are allowed to encompass more geographic area, then the distribution of city sizes reflects both density and the arbitrary boundaries that adjust to fit that density. If areas below a certain size are not considered cities at all, then the distribution of city sizes will be truncated below a certain population level.

The third essay in this volume, by Thomas Holmes and Sanghoon Lee, presents a new take on the measurement of spatial concentration. Instead of using political boundaries, Holmes and Lee lay down a grid of six mile-by-six mile squares. These squares then become their "cities," geographic areas that are truly random. While they focus on using their grid approach to revisit the topic of Zipf's law, this type of approach could be valuable in many other settings. For example, it would be useful to measure industrial concentration using their 36 mile squares instead of counties or to look at population growth regressions using their natural geographic areas, instead of counties or political cities.

Holmes and Lee have a number of striking findings. Cities and metropolitan areas follow a Zipf distribution, where there is always a greater density of smaller cities. However, the left tail of the distribution of squares looks much more bell-shaped and normal. For example, there are about twice as many squares with two people than there are with one person. In low density areas, the political definitions of units seem to be driving the received wisdom about the size distribution of cities.

In high density areas, Holmes and Lee find a kink in the distribution of population around 50,000 people. Above that point, the number of really populous places falls much more radically than Zipf's law suggests. While Zipf's law suggests that the coefficient

between rank and population size is one, they find a coefficient of two among their high density squares, which means that rank rises more quickly than population.

Gabaix (1999) connects Zipf's Law with Gibrat's Law. He finds that if places grow proportionally, then the distribution of place populations should follow Zipf's law. Since Holmes and Lee find that their squares do not follow Zipf's law, we shouldn't be surprised that they also find that Gibrat's Law fails for their 36 mile squares. They find that growth rates are much lower among places that start with more people, which perhaps explains the absence of ultra-high population areas. Their results can be taken to suggest that some form of congestion sets in at ultra-high densities.

The Sources of Agglomeration: The Costs of Moving People

Understanding agglomeration economies requires us to move beyond measuring the overall extent of agglomeration as revealed by housing prices, productivity and population concentration. We must also understand the exact mechanisms that make it more productive to cluster. While all agglomeration economies can be understood as consequences of reduced transport costs, the nature of the agglomeration economy will depend on what transport costs are being reduced. For example, the classic Krugman (1991) model of agglomeration emphasized agglomeration benefits that come from reducing the costs of moving goods over space. When an input supplier locates next to a final goods producer, these firms become more productive by saving the costs of shipping the input.

None of the essays in this volume focus on agglomeration economies that come from reducing the costs of moving goods over space, perhaps because researchers have reached a consensus that such agglomeration economies are now relatively second order. A century or more ago, when shipping goods was expensive, cities like Chicago and New York formed around ports and rail yards. Over the 20th century, the cost of moving goods declined enormously and few modern agglomerations seemed built on the easy movement of physical output. Today, the bulk of urban growth, at least in the U.S.,

appears to be in far flung places that seem to have little advantage in the shipment of goods. There is some evidence that manufacturing firms still cluster near suppliers and customers, but even this clustering seems relatively weak (Dumais, Ellison and Glaeser, 1997).

While the costs of moving goods may have declined dramatically, the cost of moving people is still high. After all, time is a major input into human travel and the value of time continues to rise as people become more productive. Even if changes in transportation technology make it possible to locate goods production anywhere in the world, there will still be an advantage from clusters that minimize the costs of people moving across space. This book has three essays that look at different types of agglomeration economies that come from reducing the costs of moving people.

Henry Overman and Diego Puga's essay examines labor market pooling— an idea whose pedigree stretches back to Alfred Marshall. The basic concept is that if there are many employers within an area then workers can change employers without changing residences. Job hopping creates advantages if workers don't know where they will be most productive, or if the productivity of different firms changes over time. Labor market pooling allows labor to be more efficiently allocated following productivity shocks, because workers can leave firms that have become less productive and move to employers that have become relatively more productive.

Krugman (1991) provided a simple and elegant model of labor market pooling that illustrates its basic mechanism. Overman and Puga's model extends the Krugman-model to multiple sectors and multiple locations. This extension is important because it generates predictions about which types of firms will co-locate with one another, and what types of co-location will generate the biggest benefits. A key result is that the agglomeration benefits are biggest when the sectors have shocks that are heterogeneous so that their shocks are particularly uncorrelated. This result, of course, requires that the sectors are still similar enough so that workers can move across them.

To test this implication empirically, Overman and Puga look across sectors within the U.K. They calculate a measure of the benefits of labor market pooling by estimating the extent to which different plants within a sector seem to have idiosyncratic employment shocks. Presumably, workers can always move across plants within an industrial sector, and sectors with more plant-level employment variation would seem to be sectors with more shocks to plant level productivity. They find that sectors with more plant-level employment shocks are more geographically concentrated. While one can reasonably worry whether greater geographic concentration within a sector is partially responsible for greater plant level variation in employment, that reverse causality should also be seen as a prediction of a labor market pooling model. This paper is one of the few papers that attempt to test this important, century-old idea.

The next two papers examine a simpler agglomeration mechanism that still stems from the benefits that come from reducing transport costs for people. Service industries can almost be defined as sectors which require person-to-person delivery. While this statement may be too strong, there is no doubt that services involve a lot more face-to-face contact than manufacturing. As a result, when service industries cluster near customers they reduce the travel costs either for their customers or for their workers. The continuing importance of transport costs for people may explain why services have remained urbanized, even as manufacturing has fled to lower density settings.

Jed Kolko's paper provides a sweeping view of agglomeration and urbanization in the service sector. Services are less agglomerated, but more urbanized than manufacturing. City streets are a good setting for services because they enable service providers to readily link with large numbers of their diverse customers. The higher transport costs involved in face-to-face delivery ties services to dense urban areas. Across services, Kolko finds a positive relationship between urbanization and concentration. The services that are most likely to benefit from connections to diverse urban populations are also most likely to concentrate. Perhaps these are the sectors with the highest transport costs.

Across service industries, human capital strongly predicts urbanization. As the Glaeser and Ponzetto paper in this volume emphasizes, cities seem to be particularly important for the transmission of ideas. Selling services directly to consumers also predicts location in big cities, while intensive use of natural resources is negatively associated with urbanization. The use of specialized occupations is positively associated with both urbanization and agglomeration, perhaps because the benefits of labor market pooling are higher for such specialized workers who cannot readily just take up another task.

Kolko also studies co-agglomeration—the tendency of industries to co-locate with other industries. He finds a strong tendency of service industries to locate near their suppliers and customers. This result contrasts with the much weaker links between customers and suppliers found in manufacturing (Dumais, Ellison and Glaeser, 1997). Since the costs of delivering services are much higher than the costs of delivering goods, it is reassuring that location patterns seem aimed at reducing those costs.

Waldfogel's essay continues the examination of the impact of transport costs, but he focuses on retail establishments. Since these establishments require visits by customers, we would expect them to be located near those customers. Waldfogel finds a strong pattern where retail establishment sectors locate near demographic groups that regularly buy from that sector. Stores catering to the well-educated locate near the well-educated. While the basic effect may be unsurprising, the measured magnitude of the tendency to locate near likely buyers is remarkably strong.

Waldfogel then suggests that the locational tendency of retail shops then provides an added benefit to demographic clustering. If a family is more likely to have access to stores that meet its needs if it locates near similar families, then this provides a good reason for neighborhood homogeneity. This mechanism is a consumption-related agglomeration effect, where locating near similar people increases one's ability to shop efficiently.

The Sources of Agglomeration: Knowledge Spillovers

Many recent papers on agglomeration economies have followed Marshall and Jane Jacobs and emphasized the role that cities can play in speeding the flow of ideas. The core idea at the center of information-based agglomeration economies is that all of our knowledge builds on things that we learn from people around us. The central premise is that the presence of knowledgeable neighbors enables an apprentice steelworker to learn his craft, but it also makes a biotechnology researcher more innovative. The interaction of smart people in urban areas both enhances the development of person-specific human capital and increases the rate at which new ideas are formed.

Katherine Baicker and Amitabh Chandra look at the diffusion of high quality health care in hospitals. They argue that there are a number of low cost procedures that significantly improve health outcomes and that those procedures should be used universally. When hospitals fail to use these procedures, Baicker and Chandra argue that the hospital is being less productive. One significant contribution of this paper is to show the diversity in this productivity measure across space. In many cases, the hospitals that have high quality, using their metric, are not the same hospitals that spend more per patient.

Baicker and Chandra illustrate the remarkable heterogeneity across metropolitan areas in hospital productivity, which seems comparable to the diversity in productivity overall. However, in the case of Baicker and Chandra's measures, higher productivity doesn't require any more physical capital, just enough human capital to use these low cost, high value procedures. They find that areas with more non-government doctors and a higher overall skill base are more likely to deliver higher quality health care, which again supports the view that local human capital matters for productivity.

They also specifically test a learning model by regressing the quality of a hospital on the lagged quality of that hospital's geographic neighbors and the hospitals own lagged quality level. Hospitals that are surrounded by higher quality hospitals tend to improve in quality. One interpretation of these results is that doctors in one hospital learn how to

practice better medicine by interacting with doctors in nearby hospitals. If this is the case, then the flow of ideas across people in metropolitan areas is actually saving lives.

William Kerr's essay looks at intellectual connections among inventors. His paper shows that the American patents are increasingly being given to inventors with non-European last names. Patents are also increasingly geographically concentrated. Kerr connects these two facts and shows that the increasing geographic concentration of inventive activity is associated with the tendency of ethnic inventors to cluster in a few metropolitan areas. This clustering of ethnic inventors can explain a significant amount of the increased clustering of patents.

Why do ethnic inventors cluster in a small number of geographic areas? One possibility is that these inventors are intellectually linked and geographic proximity allows those links to flourish. An alternative explanation is that different immigrant groups cluster in different cities to explain consumption-related advantages, such as access to religious organizations or relevant consumer goods or just to friends with a similar background. Hopefully future work will sort out the different explanations of the remarkable concentration of ethnic inventors.

The essay by Rosenthal and Strange offers a third approach to invention and entrepreneurship in urban areas. Almost 50 years ago, Ben Chinitz (1961) argued that one of the reasons why New York was more dynamic than Pittsburgh was that New York had abundant small enterprises, while Pittsburgh was concentrated in a few large businesses. Abundant small enterprises facilitated a culture of entrepreneurship, because those smaller firms needed independent input providers who could also provide inputs for other start-ups. Likewise, more small firms might mean more independent customers and these could provide a ready market for start-ups. If small firms are less able to protect their ideas, then new innovations might spread more easily in places with lots of little employers.

Rosenthal and Strange find that the amount of new establishment formation in an area is tightly linked to the number of small firms. Employment in big firms doesn't predict these start-ups. Employment in small firms does. Their research is done at the Census tract level, so they are looking at very small geographic areas and within these areas, there seems to be a strong tendency of new firms to locate where there are already many small establishments.

The penultimate paper in the volume is more of a theoretical paper on the interaction between intellectual spillovers and communication and transportation costs. I began this introduction with the seeming paradox that many cities are more vital then ever despite the fact that declining transportation and communication costs would seem to be making proximity obsolete. The paper by Glaeser and Ponzetto tries to make sense of those two facts.

The model assumes that there are three sectors in the economy: an innovation sector that produces new ideas, a manufacturing sector that makes goods and a sector that directly uses natural resources (like farming). All three sectors receive advantages from urban areas and all three sectors use land. The sectors are ordered so that the innovation sector receives the biggest benefits from urban location, because of idea spillovers and the natural resource sector gets the least out of being in a sector. The natural resource sector however uses the most land and the innovative sector needs the least. This ordering means that the innovative sector is always urbanized and under some conditions, it is the manufacturing sector that will be on the margin between urban and non-urban locations.

We model an increase in communication and transportation costs as improving the productivity in the non-urban area, relative to the city, in all three sectors. This has the effect of moving the manufacturing sector out of the city and also making the manufacturing sector more productive. As manufacturing becomes more productive, the returns to ideas increases and this increases the size of the innovative sector in the city. In one version of the model, improvements in transportation and communication costs

cause the decline of cities that specialize in manufacturing, like Detroit, and the rise of cities that specialize in innovation, like New York.

This model does appear to fit some of the recent facts about urban change. In the 1960s, almost all cities specialized in manufacturing. The ability to produce goods more cheaply outside of cities caused almost all of those places to do poorly in the 1970s. However, since then, cities with abundant human capital that specialized in innovation have done exceedingly well. In many cases, these places are coming up with new ideas that will then be produced in low cost areas throughout the world. This paper suggests that globalization seems likely to be good for cities that continue to specialize in the production of innovation, but it will continue to mean decline for manufacturing areas.

A Congestion Cost: Pollution and Cities

Density is not without its costs. Not only is land more expensive in urban areas, but congestion, pollution and social problems often accompany the crowding of people into cities. The last essay in this volume, by Matthew Kahn, reviews these costs of urban density and their trends over time.

Kahn presents an intensive look at commute times by distance to the city center. He distinguishes between big and small metropolitan areas and he compares 1980 and 2000. In most metropolitan areas, commute times rise monotonically with distance to the city center, but in the largest metropolitan areas (New York, Los Angeles, Chicago) there is a non-monotonic relationship between distance to the city center and commute times. In those largest areas, people who are far from the city center aren't commuting downtown at all. In all areas at all distances from the city center, commute times have been rising. Higher levels of congestion mean that the speed of travel has slowed significantly. Those speeds are slowest in big metropolitan areas, and this congestion is one of the big costs of living in a large metropolitan area.

While commuting costs are rising, the pollution problems of big cities appear to have been falling over the last 25 years. Kahn links this decline to the exodus of manufacturing from big cities, but cleaner big city air also reflects the rise of catalytic converters and lower levels of car emissions. Crime rates have also been falling in big cities over the past 12 years. While big cities bore the brunt of the national crime increase between 1960 and 1975, big cities have also seen the biggest drops in crime rates since their peak in the early 1990s. One possible explanation for this phenomenon is that big cities, like New York, have experienced the greatest improvements in policing quality.

Overall, Kahn's paper suggests a mixed picture. Congestion is getting worse, but pollution and crime are getting better. One possible interpretation of these facts is that new technologies, whether used by automobile manufacturers or police departments, have been more effective in pollution and crime than in reducing congestion.

This volume is meant to give a sample of the exciting work that is being done to understand the mysteries of agglomeration. Big cities are more productive, for many reasons, but they also have their costs. Indeed, if they didn't, then everyone would live in one. These essays are by no means the last word on agglomeration economies, but they do illustrate the wide range of exciting work that is being done by economists in this area.

References

Chinitz, Benjamin. "Contrasts in Agglomeration: New York and Pittsburgh." *American Economic Review,* vol. 51, no. 2 (1961): 279-289.

Ciccone, Antonio and Robert E. Hall. "Productivity and the Density of Economic Activity." *American Economic Review,* vol. 86, no. 1 (1996): 54-70.

Dumais, Guy, Glenn Ellison and Edward Glaeser. "Geographic Concentration as a Dynamic Process," NBER Working Paper # 6270 (1997).

Ellison, Glenn and Edward Glaeser, "The geographic concentration of industry: Does natural advantage explain agglomeration?" *American Economic Review*, vol. 89, no. 2 (1999): 311-316.

Krugman, Paul. *Geography and Trade*. Cambridge: MIT Press, 1991.

Lucas, Robert E. "On the Mechanics of Economic Development." *Journal of Monetary Economics* vol. 22, no. 1 (1988): 3-42.