Clusters of Knowledge: R&D Proximity and the Spillover Effect

BY GERALD A. CARLINO AND JAKE K. CARR

The United States is home to some of the most innovative companies in the world, such as Apple, Facebook, and Google, to name a few. Inventive activity depends on research and development, and R&D depends on, among other things, the exchange of ideas among individuals. People's physical proximity is a key ingredient in the innovation process. Steve Jobs understood this when he helped to design the layout of Pixar Animation Studios. The original plan called for three buildings, with separate offices for animators, scientists, and executives. Jobs instead opted for a single building with a vast atrium at its core. To ensure that animators, scientists, and executives frequently interacted and exchanged ideas, Jobs moved the mailboxes, the cafeteria, and the meeting rooms to the center of the building.

There is nothing really new in the recognition that face-to-face contact among individuals is one key to innovation. Mervin Kelly, who for a time ran AT&T's legendary Bell Labs, was, according to a *New York Times* article, "convinced that physical proximity was everything."[1] According to the article, Kelly personally helped to design a building that opened in 1941 "where everyone would interact with one another." Hallways were designed to be so long that when walking a hall's length one would encounter "a number of acquaintances, problems, diversions and ideas. A physicist on his way to lunch in the cafeteria was like a magnet rolling past iron filings." Within this unique culture, Bell Labs' employees developed some of the most important inventions of the 20th century, including the transistor, the laser, and the solar cell.

Most American companies are small in size, and they obviously lack the resources of companies such as Apple, Facebook, and Google. Does their small size deprive these firms of the benefits of knowledge spillovers — the continuing exchange of ideas among individuals and firms — that physical proximity provides? The answer appears to be no. There is an exceptionally high spatial concentration of individual R&D labs in the Northeast corridor, around the Great Lakes, in Southern California, and in California's Bay Area. The high geographic concentration of R&D labs creates an environment similar to that found at Bell Labs, in which ideas move quickly from person to person and from lab to lab.[2] This exchange of ideas underlies the creation of new goods and new ways of producing existing goods.

In this article, we will discuss a recent study that we coauthored with Robert Hunt and Tony Smith. That

[1] Jon Gertner, "True Innovation," *New York Times*, February 25, 2012.

Gerald A. Carlino is a senior economic advisor and economist specializing in regional analysis at the Federal Reserve Bank of Philadelphia. **Jake K. Carr** is a former economic analyst in the Research Department of the Federal Reserve Bank of Philadelphia. The views expressed in this article are not necessarily those of the Federal Reserve. This article and other Philadelphia Fed research and reports are available at www.philadelphiafed.org/research-and-data/publications.

[2] Knowledge spillovers are the *unintended* transmission of knowledge that occurs among individuals and organizations. For example, as pointed out by AnnaLee Saxenian, although there is intense competition in California's Silicon Valley, a remarkable level of knowledge spillovers occurs.

study has two main goals. First, our study introduces a more accurate way to measure the extent of the spatial concentration of R&D activity. This new approach allows us to document the spatial concentration of more than 1,000 R&D labs in the Northeast corridor of the U.S. An important finding that emerged from this approach is that the clustering of labs is by far most significantly concentrated at very small spatial scales, such as distances of about one-quarter of a mile, with significant clustering attenuating rapidly during the first half-mile. The rapid attenuation of significant clustering is consistent with the view that knowledge spillovers are highly localized.

We also observe a secondary node of significant clustering at a scale of about 40 miles. This secondary node of clustering is interesting because its spatial scale is roughly the same as that of the local labor market. That is, firms will draw most of their workers and most residents will commute to jobs within 40 miles. Hence, this scale is consistent with the view that the efficiency gains and cost savings at the labor market level (e.g., better matching of workers' skills to the needs of labs) are important for innovative activity.

A second goal of our study is to provide evidence on the extent to which knowledge spillovers are geographically localized within the R&D clusters we identify. Data on patent citations have been used to track knowledge spillovers. Patents contain detailed geographic information about the inventors as well as citations to prior patents on which the inventions were built. If knowledge spillovers are localized within the clusters that we identify, then citations of patents generated within a cluster should come disproportionately from within the same cluster as previous patents. We find that citations are a little over four times more likely to come from the same cluster as earlier patents than one would expect based on the preexisting concentration of technologically related activities.

LEARNING IN CLUSTERS

An enormous increase in the material well-being of individuals has been achieved over the past 200 to 300 years. We not only have more of the same goods and services but also a variety of new goods and services — such as the personal computer, the Internet, and cellular phones — whose specific characteristics could not have been imagined just 50 years ago. It took an accumulation of knowledge to design and build these goods and services and bring them to market. Inventions or innovations do not happen in a vacuum but instead are created by individuals working together to solve common problems. Often, new knowledge is tacit knowledge, that is, knowledge that is highly contextual and difficult or even impossible to codify or electronically transmit.

Beginning with Alfred Marshall, economists have studied the benefits that individuals and firms gain from locating near one another, in what are referred to as *agglomeration economies*. Knowledge spillovers, an important aspect of agglomeration economies, have proved hard to empirically verify. The empirical evidence on knowledge spillovers is rather sparse. What the limited research suggests is that the transmission of knowledge rapidly diminishes the farther one gets from the source of that knowledge. Looking at innovative activity, Adam Jaffe, Manuel Trajtenberg, and Rebecca Henderson and, more recently, Ajay Agrawal, Devesh Kapur, and John McHale find that nearby inventors are much more likely to cite each other's inventions in their patents, suggesting that knowledge spillovers are indeed localized. Mohammad Arzaghi and Vernon Henderson look at the location pattern of firms in the advertising industry in Manhattan. They show that for an ad agency, knowledge spillovers and the benefits of networking with nearby agencies are extensive, but the benefits dissipate quickly with distance from other ad agencies and are gone after roughly one-half mile.

An important finding that emerged from our new approach is that the clustering of labs is by far most significantly concentrated at very small spatial scales, such as about one-quarter of a mile.

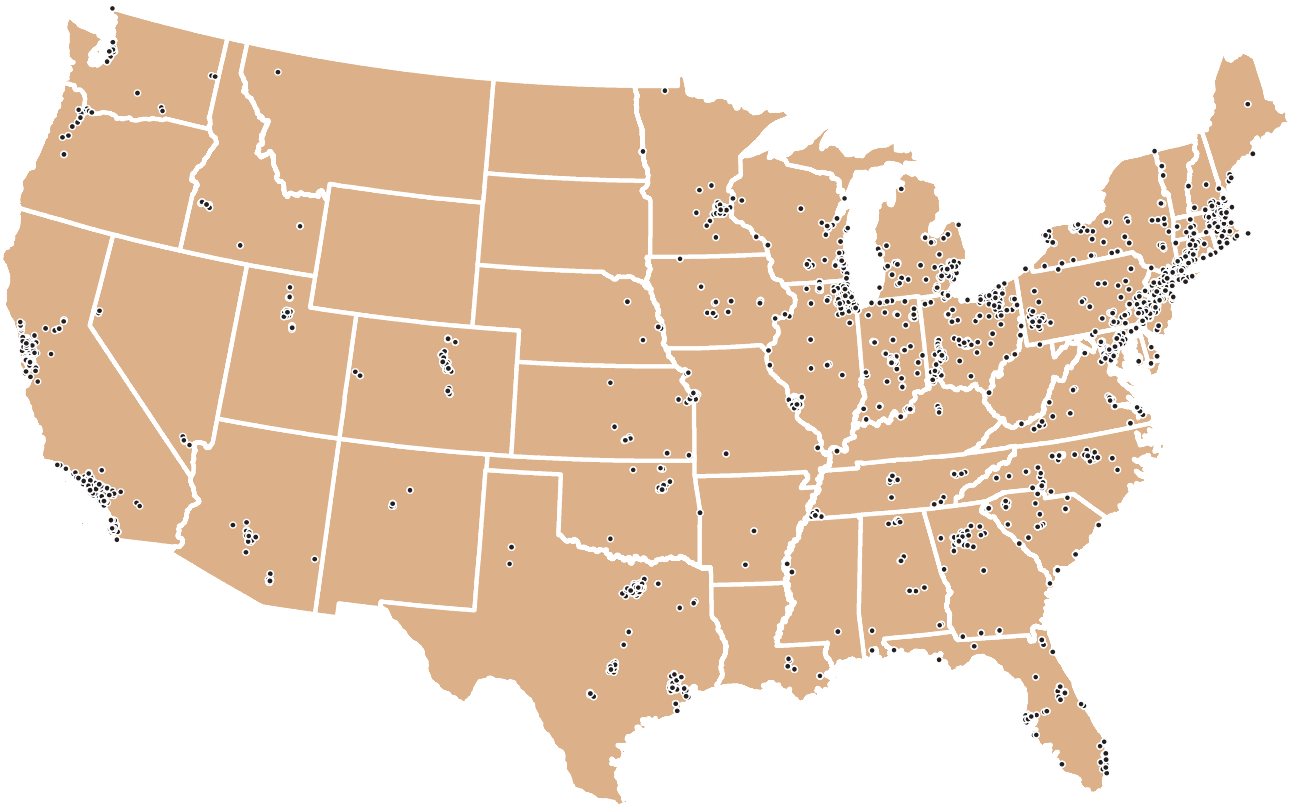
More than most economic activity, innovative activity such as R&D depends on knowledge spillovers. R&D labs will have an incentive to locate near one another if knowledge spillovers tend to dissipate rapidly with increasing distance from the source of that knowledge.

A map of the spatial distribution of R&D labs reveals a striking clustering of R&D activity (Figure 1). In places that have little R&D activity, each dot on the map represents the location of a single R&D lab. For example, there is only one lab in Montana, represented by the single dot. In counties with a dense clustering of labs, the dots tend to sit on top of one another, representing a concentration of labs.

FIGURE 1

Location of R&D Labs

Each dot on the map represents the location of a single R&D lab in 1998. In areas with dense clusters of labs, the dots tend to sit on top of one another.



Sources: Directory of American Research and Technology and authors' calculations

A prominent feature of the map is the high concentration of R&D activity in the Northeast corridor, stretching from northern Virginia to Massachusetts. There are other concentrations, such as the cluster around the Great Lakes and the concentration of labs in California's Bay Area and in Southern California.

The high geographic concentration of R&D labs creates an environment in which ideas move quickly from person to person and from lab to lab. Locations that are dense in R&D activity encourage knowledge spillovers, thus facilitating the exchange of ideas that underlie the creation of new goods and new ways of producing existing goods. The tendency for innovative activity to cluster raises a number of interesting and important questions. How strong is the tendency for R&D labs to cluster? Where in space do these labs cluster, and what are the geographic sizes of these clusters? How rapidly does the mutual attraction among labs attenuate with distance? Providing answers to these questions is an important objective of our study with Hunt and Smith.

MEASURING CLUSTERING OF ECONOMIC ACTIVITY

Although R&D labs tend to be spatially concentrated, a similar pattern of geographic concentration would be found for either population or employment. Thus, studies that look at the concentration of R&D labs need to control for the general tendency for economic activity and population to cluster spatially. In a 1996 study,

David Audretsch and Maryann Feldman introduced the "locational Gini coefficient" to show that innovative activity at the state level tends to be considerably more concentrated than is manufacturing employment and that industries that stress R&D activity also tend to be more spatially concentrated.[3]

Glenn Ellison and Edward Glaeser have identified a potential problem with the Audretsch and Feldman study. They argue that an industry may appear to be spatially concentrated if that industry consists of a few large firms. In this instance, the industry would be classified as industrially concentrated but not necessarily spatially concentrated. Ellison and Glaeser developed an alternative measure of spatial concentration — called the EG index — that controls both for the overall concentration of economic activity and for the industrial organization of the industry. Typically, the EG index has been used to gauge the geographic concentration of various manufacturing industries with fixed spatial boundaries, such as states, metropolitan areas, and counties.[4]

The EG index suffers from a number of important aggregation issues that result from using fixed spatial boundaries. For example, when calculating EG indexes at the county level, researchers will not take into account any activity that crosses county borders. As a result, measures of spatial concentration will be underestimated for counties. For example, Philadelphia County shares a border with Montgomery County. One stretch of City Avenue divides these two counties. Economic activity on the Philadelphia side of City Avenue is allocated to Philadelphia County, while activity on the Montgomery County side is assigned to that county. But this partition of economic activity is artificial, since this activity is really part of the same cluster. As a result, concentration will be underestimated for both counties. To avoid problems associated with fixed spatial boundaries, authors of several recent studies have used geocoded data to identify the exact location of establishments. These studies base their approach on the actual distance between establishments and are, therefore, not bound by a fixed geographical classification.[5]

MEASURING THE CLUSTERING OF R&D LABS

In our study, we used 1998 data from the Directory of American Research and Technology to electronically code the R&D labs' addresses and other information. Since the directory lists the complete address for each establishment, we were able to assign a geographic identifier (using geocoding techniques) to more than 3,100 R&D labs in the U.S. in 1998. We limited our analysis to 1,035 R&D labs in the 10 states (Connecticut, Delaware, Maryland, Massachusetts, New Hampshire, New York, New Jersey, Pennsylvania, Rhode Island, and Virginia) and the District of Columbia that make up the Northeast corridor of the United States.

A key question we need to determine is whether an observed spatial collection of labs in this corridor is somehow unusual; that is, is it different from what we would expect based on the spatial concentration of manufacturing employment? We used manufacturing employment instead of manufacturing firms as our benchmark.[6] In our study, we start with a "global" measure of concentration that is based on the observed concentration of R&D labs at various distances, ranging from a quarter-mile to 100 miles. For example, suppose we want to calculate the average number of labs that are located within a quarter-mile radius of

[3] A locational Gini coefficient shows how similar (or dissimilar) the location pattern of employment (or innovative activity, in Audretsch and Feldman's case) in a particular manufacturing industry is to the location pattern of overall manufacturing employment. The larger the value found for the locational Gini, the more concentrated is employment (or innovative activity) in a particular industry relative to overall manufacturing employment. See the *Business Review* article by Kristy Buzard and Gerald Carlino for a discussion of the construction of the locational Gini coefficient. The study by Audretsch and Feldman looked at the spatial concentration of innovative activity by industry. Their analysis, which is at the state level, uses 1982 census data provided by the United States Small Business Administration. They construct a data set on innovations by state and industry that is culled from information on new product announcements in over 100 scientific and trade journals.

[4] For examples of studies that use the EG index, see the studies by Ellison and Glaeser; Stuart Rosenthal and William Strange; and Glenn Ellison, Edward Glaeser, and William Kerr. See the *Business Review* article by Buzard and Carlino for a discussion of the EG index.

[5] Another problem is that authors of studies based on the EG index often provide only indexes of localization, without any indication of the statistical significance of their results. Without such statistical analyses, it is unclear whether the concentrations found differ from concentrations that would have been found if the locations of economic activity were randomly chosen. See the article by Gilles Duranton and Henry Overman for a discussion of statistical issues with the EG index.

[6] The concentration of R&D establishments is measured relative to a baseline of economic activity as reflected by the amount of manufacturing employment in the Zip code, as reported in the 1998 vintage of Zip Code Business Patterns. Since one of our objectives is to describe the localization of total R&D labs, manufacturing employment represents a good benchmark because most R&D labs are owned by manufacturing firms. We elected to use manufacturing employment as our benchmark rather than the number of manufacturing establishments in a Zip code, since past studies (such as the study by Audretsch and Feldman) use manufacturing employment as their benchmark. When we look at the clustering of R&D labs in specific industries relative to the location of all R&D labs in our data set, we find that the patterns of clustering in specific industries are highly similar to the overall clustering of labs that we found when we used manufacturing employment as the benchmark.

one another. We start by choosing one of the labs and drawing a ring with a quarter-mile radius around that lab. We then count the number of *other* labs in that quarter-mile ring and enter that number in a spreadsheet. Next, we move to another lab and draw a quarter-mile ring around it; then we count the number of other labs in its quarter-mile ring and enter that number in the spreadsheet. We repeat this procedure for all of the 1,035 labs in the corridor. Finally, we can compute the global measure of concentration at the level of a quarter-mile by averaging the 1,035 entries in the spreadsheet. This gives us the *average* number of labs that are located within a quarter-mile of one another.

We computed the global measures of the concentration of R&D labs for distances ranging from a quarter-mile to 100 miles. Finally, R&D clusters for a given distance, such as a quarter-mile, are identified as "significant" only when they contain more R&D labs than would be expected at that distance based on manufacturing employment (see Appendix: *Measuring Concentration Based on K-Functions*). We show that for every distance we considered, the spatial concentration of R&D labs is much more pronounced than it is for manufacturing employment. As we have noted, physical proximity is a key ingredient in order for firms and individuals to maximize the benefits from knowledge spillovers. This suggests that we should expect to see evidence that the benefits from such spillovers decline rapidly with increasing distance among the labs. More important, we find that the concentration of labs is most significant when labs are located within a quarter-mile radius of one another and that the significance of clustering of labs relative to manufacturing falls off rapidly as the distance among labs increases. The rapid attenuation of significant clustering at small spatial scales is consistent with the view that knowledge spillovers are highly localized.

We also found evidence of a secondary node of statistically significant clustering at a distance of about 40 miles. This scale is roughly comparable to that of a local labor market, suggesting that such markets may provide additional spillovers that improve the efficiency of labs. One way dense locations improve efficiency is through the better quality of matches among labs and workers that occurs in large and dense labor markets. Workers and labs in larger, denser labor markets can be much more selective in their matches because the opportunity costs (the lost wages or profits when the worker or firm has not made a successful match) of waiting for a prospective partner are lower. That is because even though workers and labs are more selective, on average they form better matches and tend to match more quickly. As a result, the average output from matches (such as new ideas that lead to innovation) is higher, and a higher share of the workforce and labs is engaged in productive matches. Another possibility is that labs in larger and denser locations may share critical inputs into the production process. For example, Robert Helsley and William Strange argue that the necessary inputs into the process of innovation are more plentiful and more readily available in an area with a dense network of input suppliers. The dense network of input suppliers facilitates innovation by making it cheaper to bring new ideas to fruition.

R&D clusters for a given distance, such as a quarter-mile, are identified as "significant" only when they contain more R&D labs than would be expected at that distance based on manufacturing employment.

PLOTTING THE CLUSTERING OF R&D LABS

The discussion to this point has revealed at what distances the clustering of labs is most significant, but it does not tell us where this clustering takes place. Therefore, we use a second approach, referred to as a "local" measure of clustering, to identify specific geographic areas within the corridor with high concentrations of R&D labs. Thus, a novel feature of our study is the use of a local measure of clustering to identify specific R&D clusters as well as the labs that belong to them. This approach allows us to show on a map the exact locations where the clustering of labs is occurring. For example, suppose we want to know how many other labs are located within a half-mile radius of a given lab. To find this, as we did for the global measure of clustering, we draw a circle with a radius of a half-mile around a particular lab and count the number of other labs that fall within that half-mile circle. Before, to get the global measure of clustering, we computed the average number of other labs across all 1,035 labs at a half-mile distance. To get the local measure of clustering, we are interested in the number of other labs in the individual clusters themselves. The local measures of clustering focus on the size and locations of specific R&D clusters.

Once again, we are confronted with the issue of whether the count of the labs in each of these half-mile circles is greater than would be expected based on the spatial concentration of

manufacturing employment. Figure 2 shows the strength of the clustering of labs relative to manufacturing employment for labs located south of Central Park in New York City. The 11 black dots indicate that the data strongly support the concentration of labs relative to the concentration of manufacturing employment, while the grey dots indicate somewhat less support.

To identify a half-mile cluster in New York City, we start by drawing rings with a half-mile radius around each of the 11 black dots shown in Figure 2. Figure 3 shows the pattern resulting from the construction of these half-mile rings. Notice that these rings tend to overlap one another, indicating a mutual influence among these labs. Next, we take the union of these rings to form the "half-mile" cluster in New York City (Figure 4). An important thing to note about this half-mile cluster is that its actual geographic distance is greater than a half-mile.

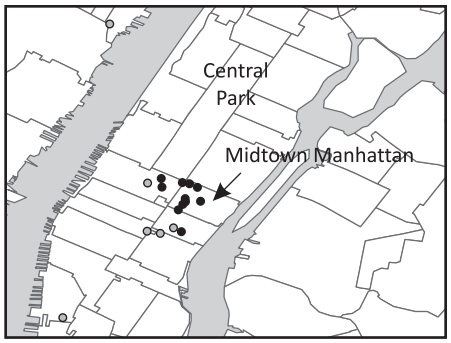
Figure 5 shows the locations of the four half-mile clusters we identified in the Boston area. The largest (both spatially and by number of labs) is found in Cambridge, MA, shown roughly at the center of the map. We also found two half-mile buffer clusters located along Route 128 and one such cluster located along Route 495.

We repeated the procedure used to create half-mile clusters, but this time we constructed one-mile rings around each of the 1,035 labs. We identified eight one-mile clusters in the Boston area, which are shown in Figure 6. Notice that all four half-mile clusters are each contained within a unique one-mile cluster. Next, we followed the same procedure to first create a five-mile cluster (of which there are two in Boston) and then a 10-mile cluster (of which there is one in Boston). Figure 7 shows the two five-mile clusters (solid black line) and the 10-mile cluster (dotted black line).

There are 187 R&D labs within

FIGURE 2

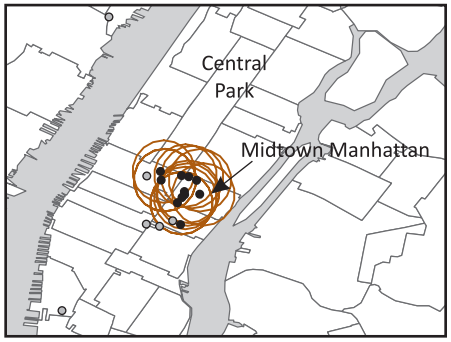
R&D Labs in New York City



Each dot represents the location of a single R&D lab. The black dots strongly indicate a local cluster of labs relative to manufacturing employment. The grey dots indicate a less significant concentration of labs relative to manufacturing employment.

FIGURE 3

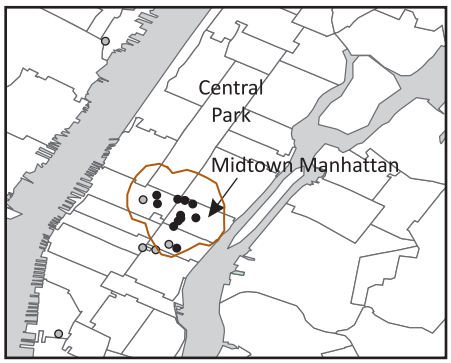
Constructing Half-Mile Buffer Rings



This half-mile cluster in New York City was created by constructing rings with a half-mile radius around each black dot. These rings tend to overlap one another, indicating a mutual influence among these labs.

FIGURE 4

Half-Mile Cluster in New York City

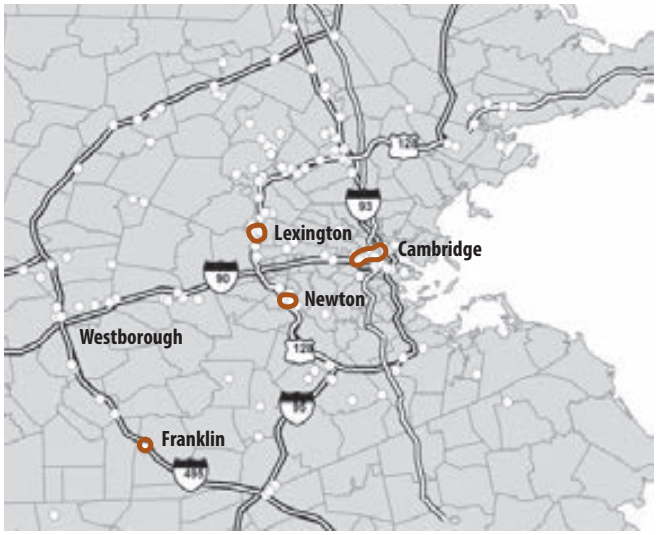


To identify New York City's half-mile cluster, we drew a line around the perimeter of the rings in Figure 3. It is important to note, however, that the actual geographic distance of this cluster is greater than a half-mile.

Sources: Directory of American Research and Technology and authors' calculations

FIGURE 5

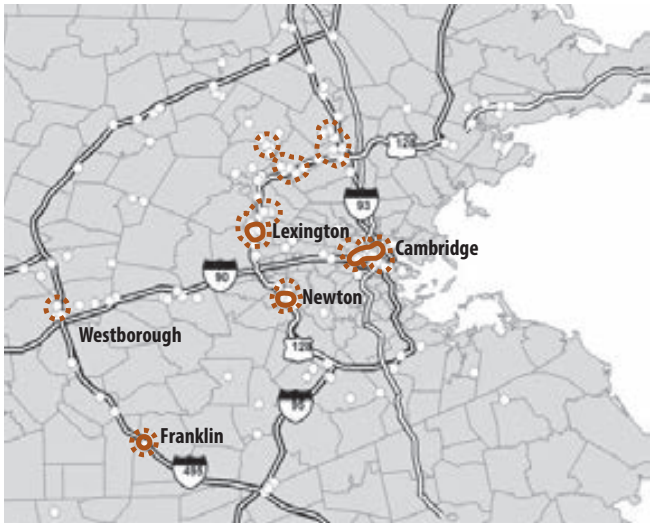
Half-Mile Clusters in Boston



This figure shows four half-mile clusters of labs in Boston, the largest of which is in Cambridge at the junction of Route 90 and Route 93.

FIGURE 6

One-Mile Clusters in Boston



Eight one-mile clusters of labs in Boston are indicated by dotted brown rings. Notice that all four half-mile clusters, which are indicated by solid brown rings, are situated within one-mile clusters.

Boston's single 10-mile cluster. Most of these labs conduct R&D in five industries: computer programming and data processing, drugs, lab apparatus and analytical equipment, communications equipment, and electronic equipment. The largest five-mile cluster, which is shown in Figure 7, contains 108 labs, which account for 58 percent of all labs in the larger 10-mile cluster. At the one-mile scale, Boston has eight clusters, six of which are centered in the largest five-mile cluster. The largest of these one-mile clusters contains 30 labs, half of which conduct research on drugs.

Figure 8 shows the clusters of R&D labs we identified in the Philadelphia region, where there are a total of 49 labs. The city of Philadelphia is shown by the darker grey area east of the center of the figure. The dotted black ring depicts Philadelphia's 10-mile cluster. Of the 49 labs in this broad cluster, 16 conduct research on drugs, and another 16 perform research in the plastics materials and synthetic resins industry. The Philadelphia region contains two five-mile clusters, shown by the solid black boundaries in Figure 8. The most prominent subcluster is centered in the King of Prussia area, directly west of the city of Philadelphia, and contains 30 labs, of which 40 percent conduct research on drugs. Within this subcluster, there is a much tighter concentration of labs (indicated by the dotted brown ring in Figure 8) located near Routes 76 and 276.

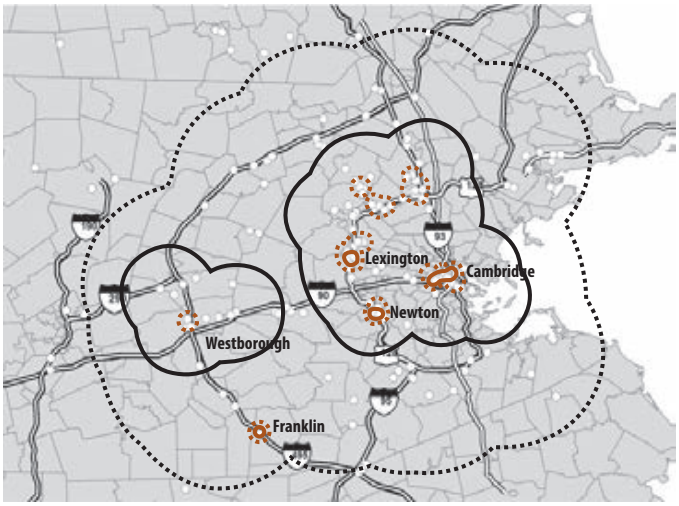
The second subcluster is centered in the city of Wilmington, DE, where about 25 percent of the labs are also engaged in research on drugs, but most (almost 60 percent) are conducting research on plastics materials and synthetic resins.

THE EFFECTS OF KNOWLEDGE SPILLOVERS

Innovation is important because it can directly affect a nation's productivity growth and the economic welfare of society through the introduction of new or improved goods and lower prices. In addition to these direct benefits, as we have argued in this article, the innovative activity of one person can also influence the innovative activity of others through knowledge spillovers. Paul Krugman has argued, however, that knowledge spillovers are impossible to measure empirically because they "are invisible; they leave no paper trail by which they may be measured and tracked." However, as Jaffe and coauthors have noted, "Knowledge flows

FIGURE 7

Ten-Mile Cluster in Boston



This figure shows the two five-mile clusters of labs in Boston (solid black lines) and the single 10-mile cluster (dotted black line). Notice that all four half-mile clusters (solid brown) identified for Boston are situated within one-mile clusters (dotted brown). Similarly, most of the one-mile clusters lay within the two five-mile clusters, and the two five-mile clusters are contained within the 10-mile cluster.

FIGURE 8

Ten-Mile Cluster in Philadelphia

In the Philadelphia region, we identified a single one-mile cluster that is located west of the city (the city is shown in dark grey) approximately in the King of Prussia, PA, area. The Philadelphia region has two five-mile clusters (solid black lines) and one 10-mile cluster (dotted black line). The second five-mile cluster is centered in the city of Wilmington, DE.

do sometimes leave a paper trail in the form of patent citations to prior art."

Jaffe and coauthors pioneered a method for studying the geographic extent of knowledge spillovers using patent citations. Every patent contains the names, hometowns, and Zip codes of the inventors named in the patent. A patent can be assigned to a location by using the Zip code of one of its inventors (usually the first person named). Patent citations are similar to citations received by academic articles in that patent citations reference prior technology or prior art on which the citing patent builds. Therefore, Jaffe and coauthors hold that patent citations are a useful proxy for measuring knowledge flows among inventors. If knowledge spillovers are localized within a given metropolitan area, then citations to patents within a given metropolitan area should disproportionately come from other inventors who are located within that metropolitan area.

However, Jaffe and coauthors point out that just because we observe a geographic clustering of technologically related activities, such as the clustering of the semiconductor industry in Silicon Valley, this clustering is not necessarily evidence of knowledge spillovers among these related activities. There are other sources of agglomeration economies in metropolitan areas, such as better matching and sharing, that could explain the spatial clustering of activities in the semiconductor industry. Jaffe and coauthors deal with the spatial clustering of related activities by constructing a set of control patents designed to match the existing geographic concentration of technologically related activities. To test for localized knowledge spillovers, Jaffe and coauthors construct three patent samples. The first sample consists of a set of originating patents. The second sample consists of a set of patents that cite one of the originating patents (referred to as citing patents). The final sample consists of a control patent chosen to match each of the citing patents. To qualify as a control patent, the patent must be as similar as possible (in terms of being in the same technology class and having an appli-

cation date as close as possible) to the matched citing patent, but the control patent must not cite the matched originating patent. Jaffe and coauthors compute two geographic matching frequencies: one between the citing patents and the originating patents and one between the control patents and the originating patents. Their test for the localization of knowledge spillovers is whether the citation matching frequency for a given geographic definition (states and metropolitan areas) is significantly greater than that associated with the control matching frequency. Jaffe and coauthors find that patent citations are two times more likely to come from the same state and about six times more likely to come from the same metropolitan area as earlier patents than one would expect based on the control patents.

In our study, we adopt Jaffe and coauthors' methodology to look for evidence of localized knowledge spillovers, except that we use the boundaries determined by the nine five-mile clusters identified in our research instead of using state and metropolitan area boundaries.[7] State boundaries are politically determined, rather than economically justified, and states are too big to adequately capture knowledge spillovers, which are highly localized. In addition, the boundaries of metropolitan areas are determined by labor market flows; therefore, they are not well suited for analysis of spillovers among individuals engaged in innovative activity. Instead, we use the boundaries determined by our nine five-mile clusters as our basic geography, since these boundaries are determined by interrelationships among the R&D labs and more accurately reflect the appropriate boundaries in which knowledge spillovers are most likely to occur.

The patent citation counts that we use are constructed from the NBER Patent Citations Database. Patents are assigned to locations according to the Zip code of the first inventor named on the patent.[8] There were 9,105 patents applied for in the nine five-mile buffer clusters we identified in our study during the period 1996–1997. After removing self-citations, these originating patents received 90,159 forward citations during the period 1996–2006.[9] But we were able to find control patents for only about 55,000 of the citing patents. This limits our analysis to those citing patents for which we have controls.[10] We find that, on average, a patent that falls within one of our five-mile clusters is 4.3 times more likely to cite an earlier patent in the same five-mile cluster compared with a control patent (a finding that is highly statistically significant). Despite the fact that knowledge spillovers are not directly observable, they do leave a paper trail in the form of patent citations. We find that these paper trails provide evidence consistent with the geographic concentration of knowledge spillovers.

CONCLUSION

In this article, we summarize the findings from our study that uses distance-based measures to analyze the spatial concentration of over 1,000 R&D labs in the Northeast corridor of the United States. Rather than using a fixed spatial scale, such as counties and metropolitan areas, we attempt to describe the spatial concentration of R&D labs more precisely by considering the spatial structure at different scales. We find that the clustering of labs is by far most significant at very small spatial scales, such as distances of about one-quarter of a mile, with significance attenuating rapidly during the first half-mile. The rapid attenuation of significant clustering at small spatial scales is consistent with the view that knowledge spillovers are highly localized.

We introduce a novel way to identify the location of clusters and number of labs in these clusters. For example, this approach identified a number of clusters of R&D labs in the Boston, New York–Northern New Jersey, Philadelphia–Wilmington, and Washington, D.C., areas. We also found that each of these clusters has distinct characteristics, especially in terms of the mix of industries the R&D labs serve.

Using patent data, we are able to provide evidence that knowledge spillovers are highly localized within the clusters of R&D labs that we identify. We find that patent citations are a little over four times more likely to come from the same cluster as earlier patents than one would expect based on the preexisting geographic concentration of technologically related activities.

[7] We identified two five-mile clusters in Boston (Figure 7), three such clusters in New York, two in Philadelphia (Figure 8), and two in Washington, D.C. In this article, we present only the findings averaged across the nine clusters. See our working paper for details on the individual clusters.

[8] The patent and citation data we use from the National Bureau of Economic Research (NBER) Patent Data Project provide the name, town, and Zip code of the principal (or first named) inventor on each patent. As is standard when assigning patents to areas, we assign patents to our clusters using the Zip code of the first inventor named on the patent. Knowledge spillovers can occur among individuals who meet because they are part of either local technical or social networks. For example, AnnaLee Saxenian describes how Walker's Wagon Wheel bar in Mountain View, CA, became a popular place for engineers who lived in Silicon Valley to exchange ideas.

[9] Since self-citations may not result from knowledge spillovers, we excluded not only inventor self-citations but also citing patents owned by the same organizations as the originating patent.

[10] There was an insufficient number of control patents to confidently conduct the analysis for the one-mile or half-mile clusters.

Appendix: Measuring Concentration Based on K-Functions

The Global K-Function

A popular measure of concentration is Ripley's *K*-function, which we use to test for clustering at differing distances:

$$\hat{K}^{O}(d) = \frac{1}{n}\sum_{i=1}^{n} C_i(d)$$

where $C_i(d)$ is the count of additional labs within distance d from lab (location) i and n is the total number of locations in the study (n = 1,035 in our study). To see how this works, set d equal to one mile. Take the first lab and draw a one-mile circle around that lab. Count the number of other labs in that one-mile circle and enter the resulting count of other labs into a spreadsheet. Go to the next lab and construct a one-mile circle around that lab. Count the number of other labs in that one-mile circle and enter the resulting number into the spreadsheet. Repeat these steps for all 1,035 labs. Sum over the 1,035 observations and divide by 1,035 labs. This is the average value of concentration of labs at a distance of one mile, denoted by $\hat{K}^{O}(1)$. We calculate the average observed value of concentration, beginning at a quarter-mile and increasing at quarter-mile increments below one mile and at one-mile increments from one mile to 100 miles.

The key question of interest is whether the overall pattern of R&D locations in the 10 states and the District of Columbia exhibits more clustering than would be expected from the spatial concentration of manufacturing in those areas. To address this question statistically, our null hypothesis is that R&D locations are determined entirely by the distribution of manufacturing employment.

We use a two-step procedure for generating counterfactual observations that are used to test the null hypothesis. In the simulations, we randomly allocated labs to Zip codes based on a probability proportional to manufacturing employment in that Zip code so that Zip codes containing a large share of employment are more likely to be assigned labs. For each distance, we compute a simulated distribution of labs. We compared the observed value for their *K*-functions (the $\hat{K}^{O}(d)$) with values obtained from a simulated distribution of R&D labs. If the observed value for the *K*-function for a given distance is large relative to the simulated distribution, this is taken as evidence of significant clustering of labs relative to manufacturing employment. *P*-values can be computed as:

$$P(d) = \left\{ \frac{\text{The number of simulated values at distance } d \text{ that are at least as large as the observed value}}{\text{Number of simulation performed}} \right\}$$

For example, if we performed 1,000 simulations and there are 10 simulated values at least as large as $\hat{K}^{O}(d)$, then there is only a one-in-a-hundred chance of observing a value at least as large as $\hat{K}^{O}(d)$. In this example, there is significant clustering of R&D locations at the 0.01 level of statistical significance at spatial scale d. However, we found that the clustering of labs is so strong relative to manufacturing employment that the estimated *p*-values were uniformly 0.001 for all the distances we considered. We obtained sharper discrimination by calculating the *z*-scores for each observed estimate, $\hat{K}^{O}(d)$, as given by

$$z(d) = \frac{\hat{K}^{0}(d) - \bar{K}_d}{s_d} \quad , \quad d = \{0.25, 0.5, 0.75, 1, 2, ..., 99, 100\}$$

where $\bar{K}_d$ and s_d are the corresponding sample means and standard deviations for the $N+1$ sample *K*-values. These *z*-scores are shown along the vertical axis in the figure, while the horizontal axis shows distances among R&D labs. The higher the *z*-score for a given distance, the more spatially concentrated the R&D labs are at that distance relative to manufacturing employment. Notice that the highest *z*-score we found, which is more than 30 standard deviations away from the mean, occurs at the shortest distance among labs we considered (one-quarter of a mile) and declines rapidly with distance up to a distance of about five miles. The rapid decline in *z*-scores (significance of clustering of R&D labs) at short distances is consistent with the view that knowledge spillovers are highly geographically localized. Notice that the lowest *z*-score obtained, which occurs at a distance of about five miles, is still more than 7 standard deviations away from the mean, indicating that R&D labs are significantly more concentrated than manufacturing employment over all the distances we considered. We also observe a secondary mode of significance at a scale of about 40 miles, which is roughly associated with metropolitan areas.

The Local *K*-Function

Basically, the local version of Ripley's *K*-function for a lab at a given location is simply the count of all additional labs within distance *d* of the given lab. In terms of the notation, the local *K*-function, $\hat{K}_i$, at location i is given for each distance, *d*, by,

$$\hat{K}_i(d) = C_i(d)$$

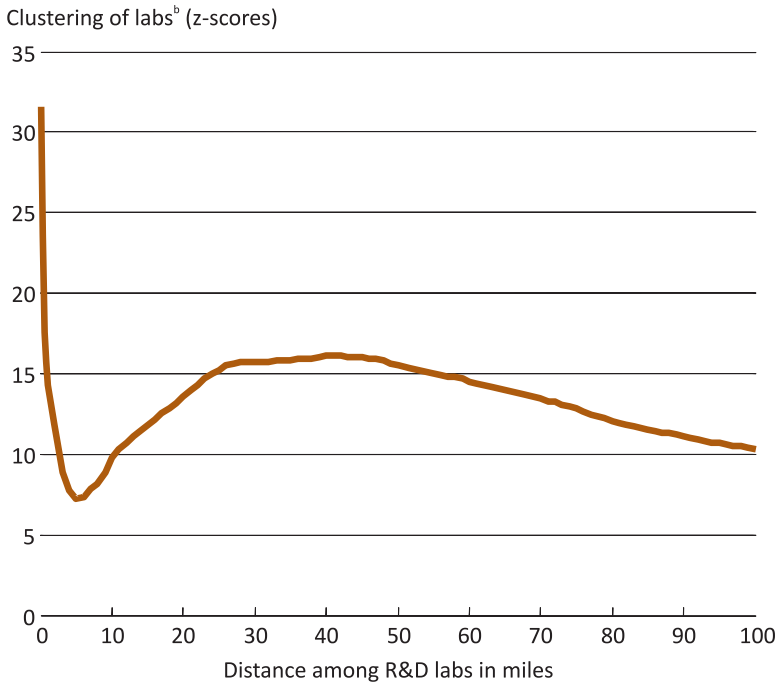
We use the same null hypothesis employed in the global *K*-function analysis that R&D labs are distributed in a manner proportional to the distribution of manufacturing employment. The only substantive difference from the procedure used in global *K*-function analysis is that the actual point associated with location *i* is held fixed when computing the simulated values for the local *K*-function. That is, for a given distance, holding the location of the lab fixed, we compute a simulated distribution of labs at that point. We compared the observed value for their *K*-functions (the $\hat{K}_i(d)$) with values obtained from a simulated distribution of R&D labs. If the observed value for the *K*-function at a given point is large relative to the simulated distribution, this is taken as evidence of significant clustering of labs relative to manufacturing employment at that location. The set of radial distances (in miles) used for the local tests was $D = \{0.5, 0.75, 1, 2, 5, 10, 11, 12.., 100\}$.

In our global analysis, the *p*-values were essentially the same for nearly all spatial scales. That is not the case for the local analysis. It is not surprising to find that many isolated R&D locations exhibit no local clustering whatsoever; therefore, wide variations in significance levels are possible at any given spatial scale. Thus, *p*-values are used in the local *K*-function analysis.

An attractive feature of these local tests is that the resulting *p*-values for each point i in the observed pattern can be *mapped*. This allows us to check visually for *regions* of significant clustering. In particular, groupings of very low *p*-values serve to indicate not only the location but also the *approximate size* of possible clusters.

Because we conduct tests for local clustering over many locations and spatial scales, we need to address two aspects of the "multiple testing" problem. First, suppose that there is, in fact, no local clustering of labs. In our simulations, we would nonetheless expect to find that 5 percent of the observed values for the local *K*-functions for a given distance are statistically significant at the 5 percent level of significance. Therefore, when many such tests are conducted (1,035 tests for each distance considered), we are likely to find some degree of significant clustering using standard testing procedures. The incidence of this type of "false positive" findings is mitigated by reducing the threshold level of significance (the *p*-value) deemed to be "significant." That is, we can minimize the incidence of false positives due to the multiple testing problem by focusing on labs with very high levels of statistical significance (*p*-values of 0.001 or lower). We refer to these as core points — the black dots in Figure 2 in the article.[a] A second condition of a core point is that there must be at least four other labs at a given distance. This condition is imposed to exclude isolated labs that happen to be in areas with little or no manufacturing employment.

Clustering of Labs Attenuates Rapidly with Distance



[a] The grey dots in Figure 2 are associated with p-values no greater than 0.005.

[b] *Z-scores are shown along the vertical axis, while the horizontal axis shows distances among R&D labs. The higher the z-score for a given distance, the more spatially concentrated the R&D labs are at that distance relative to manufacturing employment. For example, a z-score of 10, occurring at a distance of about two miles, indicates that the concentration of labs at that distance is 10 standard deviations away from the mean at that distance, indicating that labs are significantly more concentrated at that distance relative to manufacturing employment.

REFERENCES

Agrawal, Ajay, Devesh Kapur, and John McHale. "How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data," *Journal of Urban Economics*, 64:2 (2008), pp. 258-69.

Arzaghi, Mohammad, and J. Vernon Henderson. "Networking Off Madison Avenue," unpublished manuscript (2005).

Audretsch, David B., and Maryann P. Feldman. "R&D Spillovers and the Geography of Innovation and Production," *American Economic Review*, 86 (1996), pp. 630-40.

Buzard, Kristy, and Gerald A. Carlino. "The Geography of Research and Development Activity in the U.S.," Federal Reserve Bank of Philadelphia *Business Review* (Third Quarter 2008), pp. 1-10.

Carlino, Gerald A., Jake K. Carr, Robert M. Hunt, and Tony E. Smith. "The Agglomeration of R&D Labs," Federal Reserve Bank of Philadelphia Working Paper 12-22 (September 2012).

Directory of American Research and Technology, 23rd ed. New York: R.R. Bowker, 1999.

Duranton, Gilles, and Henry G. Overman. "Testing for Localization Using Micro-Geographic Data," *Review of Economic Studies*, 72 (2005), pp. 1077-1106.

Ellison, Glenn, and Edward L. Glaeser. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, 105 (1997), pp. 889-927.

Ellison, Glenn, Edward L. Glaeser, and William Kerr. "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," Discussion Paper 2133, Harvard Institute of Economic Research (April 2007).

Helsley, Robert W., and William C. Strange. "Innovation and Input Sharing," *Journal of Urban Economics*, 51:1 (2002), pp. 25-45.

Jaffe, Adam, M. M. Trajtenberg, and R. Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 108 (1993), pp. 577-98.

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

Rosenthal, Stuart, and William C. Strange. "The Determinants of Agglomeration," *Journal of Urban Economics*, 50 (2001), pp. 191-229.

Saxenian, AnnaLee. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*, 2nd ed. Cambridge, MA: Harvard University Press, 1996.