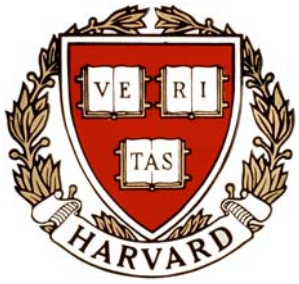

HIER

Harvard Institute of Economic Research

Discussion Paper Number 2133

What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns

by

Glenn Ellison
Edward L. Glaeser
and
William Kerr

April 2007

HARVARD UNIVERSITY
Cambridge, Massachusetts

What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns[1]

Glenn Ellison
MIT and NBER

Edward L. Glaeser
Harvard University and NBER

and

William Kerr
Harvard University

April 3, 2007

[1]email: gellison@mit.edu, glaeser@fas.harvard.edu, wkerr@hbs.edu. We thank Jim Davis, Alex Bryson, Keith Maskus, and Debbie Smeaton for comments and data assistance. The research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Boston Census Research Data Center (BRDC). Support for this research from NSF grant (ITR-0427889) is gratefully acknowledged. Research results and conclusions expressed are our own and do not necessarily reflect the views of the Census Bureau or NSF. This paper has been screened to insure that no confidential data are revealed.

Abstract

Many industries are geographically concentrated. Many mechanisms that could account for such agglomeration have been proposed. We note that these theories make different predictions about which pairs of industries should be coagglomerated. We discuss the measurement of coagglomeration and use data from the Census Bureau's Longitudinal Research Database from 1972 to 1997 to compute pairwise coagglomeration measurements for U.S. manufacturing industries. Industry attributes are used to construct measures of the relevance of each of Marshall's three theories of industry agglomeration to each industry pair: (1) agglomeration saves transport costs by proximity to input suppliers or final consumers, (2) agglomeration allows for labor market pooling, and (3) agglomeration facilitates intellectual spillovers. We assess the importance of the theories via regressions of coagglomeration indices on these measures. Data on characteristics of corresponding industries in the United Kingdom are used as instruments. We find evidence to support each mechanism. Our results suggest that input-output dependencies are the most important factor, followed by labor pooling.

1 Introduction

We know that industries are geographically concentrated.[1] We know that this concentration is too great to be explained by exogenous spatial differences in natural advantage.[2] We have an abundance of theories for this concentration.[3] But we do not which of these theories are important or even right. This paper uses patterns of coagglomeration–the tendency of different industries to locate near to each other–to assess the importance of different theories of geographic concentration.

Marshall (1920) emphasized three different types of transport costs–the costs of moving goods, people, and ideas–that could be reduced by industrial agglomeration. First, he considered transport costs for goods and argued that firms will locate near suppliers or customers to save shipping costs. Second, he developed a theory of labor market pooling in which firms located near one another can share labor. The larger labor pool created by agglomeration allows workers to move to more productive firms when there are shocks. Third, he began the theory of intellectual spillovers by arguing that in agglomerations, "the mysteries of the trade become no mystery, but are, as it were, in the air." Firms, such as those described by AnnaLee Saxenian (1994) in Silicon Valley, locate near one another to learn and to speed their rate of innovation.

Although each of these determinants certainly contributes to agglomeration in some industries, assessing their aggregate relative importance is challenging because they all predict that firms will co-locate with other firms in the same industry. One approach pioneered by David B. Audretsch and Maryann P. Feldman (1996) and Stuart S. Rosenthal and William C. Strange (2001) is to examine cross-industry variation in the degree of agglomeration, e.g. regressing the degree to which an industry is agglomerated on the importance of R&D to the industry. In this paper we propose an alternate approach: we study the agglomeration process through the lens of how industries are coagglomerated. This can potentially exploit the fact that the theories make different predictions about

[1] See P. Sargant Florence (1948), E. M. Hoover (1948), Victor Fuchs (1990), Paul Krugman (1991a), and Glenn Ellison and Edward L. Glaeser (1997).

[2] See Ellison and Glaeser (1999).

[3] See Johann Heinrich von Thünen (1826), Alfred Marshall (1920), and Krugman (1991b).

which pairs of industries will tend to coagglomerate.[4] For example, if transport costs for goods are important, then firms in an industry should be agglomerated near industries that are their customers or suppliers. If labor market pooling is important, then industries should locate near other industries that employ the same type of labor. Our approach, like those of the other papers mentioned above, uses industry characteristics as covariates. One could worry that these are endogenous and reflect the industry's geography. Our second main empirical innovation is to use characteristics of U.K. industries as instruments for the characteristics of their U.S. counterparts.

We begin in Section II with some material on the measurement of coagglomeration. We review the index of coagglomeration proposed in Ellison and Glaeser (1997), note that a simpler equivalent definition can be used when measuring pairwise coagglomeration, and further develop the economic motivation for the index as a measure of the importance of cross-industry spillovers and shared natural advantage.

Section III describes the data used to generate our coagglomeration index and presents some basic descriptive results. We base our coagglomeration measures on establishment-level data from the Census of Manufactures. This data set allows us to calculate co-agglomeration for every pairwise combination of three-digit Standard Industrial Classification (SIC) industries at the state, metropolitan area, and county levels.

Section IV reviews Marshall's three theories and discusses the covariates we will use to assess the importance of different reasons for co-location. We use input-output tables to construct proxies for the importance of transport costs for goods. We use the correlation across industries in their employment of different occupations to measure the potential gains from labor market pooling. Finally, we use data on technology flows and patent citations to construct proxies for the importance of technological spillovers. Our empirical approach is to regress the extent to which each pair of industries is co-located on the extent to which these two industries buy and sell from each other, hire the same type of workers, and share ideas. One potential concern with this approach is that the measures may not be innate characteristics of industries: firms may buy and sell from one another because they are

[4]See J. Vernon Henderson (2003) for a related approach: Henderson examines how plant-level productivity is related to the set of plants in the area.

close, not be close because they buy and sell from each other. To address this, we use British input-output tables, employment patterns, and patent citations to instrument for our American measures.

Section V presents our main empirical results. The ordinary least squares relationships support the importance of all three theories. Input/output relationships appear to be the most important determinants of co-location. Given the remarkable decline of transportation costs over the 20th century (Glaeser and Janet E. Kohlhase, 2004), it is striking that transport costs remain so important. Industries that hire the same type of workers are also quite likely to locate near one another. This effect is almost as strong as the role of supplier/customer relationship. Our proxies for intellectual spillovers have a slightly weaker, but still quite significant, impact on the tendency of industries to coagglomerate. Our instrumental variables strategy delivers similar results.

Section VI concludes. Industrial co-location patterns are far from random. Firms unsurprisingly locate near their customers and suppliers. They are almost as driven by the advantages of sharing a large labor pool, and intellectual spillovers also matter.

2 Measurement of Coagglomeration

In this section we discuss an index of coagglomeration introduced in Ellison and Glaeser (1997). We note that the index takes on a simpler form when used to measure pairwise coagglomeration and we further develop the economic motivation for the index as a measure of the importance of cross-industry spillovers and shared natural advantages.

2.1 Background

Consider a group of industries indexed by $i = 1, 2, \ldots, I$. Suppose that a geographic whole is divided into M subareas and suppose that $s_{1i}, s_{2i}, \ldots, s_{Mi}$ are the shares of industry i's employment contained in each of these areas. Let $x_1, x_2, \ldots, x_M$ be some other measure of the size of these areas, such as each area's share of population or aggregate employment. A simple measure of the *raw geographic concentration* of industry i is

$$G_i = \sum_{m=1}^{M} (s_{mi} - x_m)^2.$$

Ellison and Glaeser (1997) note that it is problematic to make cross-industry or cross-country comparisons using this measure because it will be affected by the size distribution of plants in the industry and the fineness of the available geographic data. They propose an alternate measure of agglomeration we will refer to as the EG index:

$$\gamma_i \equiv \frac{G_i/(1-\sum_m x_m^2) - H_i}{1 - H_i},$$

where H_i is the plant-level Herfindahl index of industry i.[5] They show that the EG index "controls" for differences in the plant size distribution and the fineness of the geographic breakdown, in the sense of being an unbiased estimator of a parameter reflecting the importance of natural advantages and spillovers in a simple model of location choice.

Ellison and Glaeser (1997) also propose a measure of the coagglomeration of a group of I industries. Let w_i be industry i's share of total employment in the I industries. Let $s_1, \ldots, s_M$ be the shares of the total employment in the group of I industries in each of the geographic subareas. (Note that $s_m = \sum_{i=1}^{I} w_i s_{mi}$.) Write G for the raw geographic concentration for the I-industry group: $G = \sum_{m=1}^{M}(s_m - x_m)^2$. Write H for the plant-level Herfindahl of the I-industry group: $H = \sum_i w_i^2 H_i$. The EG index of coagglomeration is

$$\gamma^c \equiv \frac{G/(1-\sum_m x_m^2) - H - \sum_{i=1}^{I} \gamma_i w_i^2 (1 - H_i)}{1 - \sum_{i=1}^{I} w_i^2}. \tag{1}$$

The index reflects excess concentration of the I-industry group relative to what would be expected if each industry were as agglomerated as it is, but the locations of the agglomerations were independent. The particular form is motivated by a proposition relating the expected value of the index to properties of the location-choice model.

Proposition 0 *Ellison and Glaeser (1997)*

In an I-industry probabilistic location choice model, suppose that the indicator variables $\{u_{km}\}$ for whether the k^{th} plant locates in area m satisfy $E(u_{km}) = x_m$ and

$$Corr(u_{km}, u_{\ell m}) = \begin{cases} \gamma_i & \text{if plants } k \text{ and } \ell \text{ both belong to industry } i \\ \gamma_0 & \text{if plants } k \text{ and } \ell \text{ belong to different industries.} \end{cases}$$

Then, $E(\gamma^c) = \gamma_0$.

[5] This is defined by $H_i = \sum_{k=1}^{N_i} z_{ki}^2$, where $k = 1, 2, \ldots, N_i$ indexes the plants in industry i and z_{ki} is the employment of plant k as a share of the total employment in industry i.

2.2 A simpler formula

The EG coagglomeration index is a measure of the average coagglomeration of industries in a group. A simpler equivalent formula can be given for the coagglomeration of two industries.

Proposition 1 *An equivalent formula for the EG coagglomeration index when* $I = 2$ *is*

$$\gamma^c = \frac{\sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}.$$

The formula makes clear that the EG coagglomeration index is closely related to the covariance of the state-industry employment shares in the two industries. The denominator rescales the simple covariance to eliminate a sensitivity to the fineness of the geographic breakdown. Note that plant-level Herfindahls do not enter into the formula: the lumpiness of plants causes an increase in the variance of the state-industry employment shares that could be mistaken for within-industry agglomeration, but does not by itself lead to a spurious increase in the cross-industry covariance. (Larger plant Herfindahls will, however, make γ^c a noisier parameter estimate.)

2.3 Explicit models of location choice

Proposition 0 is in a sense quite general: it shows that the coagglomeration index is appropriate if location decisions are made in any manner that satisfies one property. This generality, however, is obtained at the expense of losing explicit connections to the economics of location decisions and how they are influenced by natural advantages, spillovers, etc. In this section we extend the single-industry model of Ellison and Glaeser (1997) to make these connections.

We will discuss spillovers and natural advantages separately using two models with many elements in common. There are two industries indexed by $i = 1, 2$, with N_1 plants in industry 1 and N_2 plants in industry 2. Plants are indexed by $k \in K_1 \cup K_2$, with K_1 being the set of plants in industry 1 and K_2 being the set of plants in industry 2. The plants choose among M possible locations. Each plant has an exogenously fixed employment level e_k that is independent of its location choice.

2.3.1 Spillovers

We conceptualize spillovers as mechanisms that make plant k's profits a function of the other plants' location decisions. A general model of this form would be to assume that firm k's profits when locating in area m are of the form $\pi_{km} = f(m, \ell_{-k}, \epsilon_{km})$, where ℓ_{-k} is the vector of all plants' location decisions and ϵ_{km} is a random shock. A difficulty with discussing the degree of geographic concentration in such a model is that the location choice process becomes a game that can have multiple equilibria. For example, if plants k and k' receive substantial benefits from co-locating, then there may be equilibria in which the two plants co-locate in any of several areas that are fairly good for each plant, and also an equilibrium in which the plants forego the spillover benefits and locate in the areas that are most advantageous for each plant separately. (This will only be an equilibrium if plant k's most-preferred location is sufficiently unattractive to plant k' and vice-versa.) The different equilibria will typically lead to different levels of measured concentration.

Ellison and Glaeser (1997) note that the impact of equilibrium multiplicity is substantially reduced if one considers random "all-or-nothing" spillovers. To extend their analysis, define a *partition* ω of $K_1 \cup K_2$ to be a correspondence $\omega : K_1 \cup K_2 \rightrightarrows K_1 \cup K_2$ such that $k \in \omega(k)$ for all k and $k' \in \omega(k) \Rightarrow \omega(k) = \omega(k')$. Suppose that plants' location decisions are the outcome of game in which the plants choose locations in some (possibly random) exogenously specified order and plant k's profits from locating in area m are given by

$$\log(\pi_{km}) = \log(x_m) + \sum_{k' \in \omega(k)} I(\ell_{k'} \neq m)(-\infty) + \epsilon_{km}.$$

The first term on the right-side of this expression, x_m, is the measure of the size of area m we used when constructing the concentration index. Its inclusion allows the model to match real-world data in which many more plants locate in California than in Wyoming.[6] The second term reflects the impact of spillovers: the interpretation is that a spillover exists between plants k and k' if $k' \in \omega(k)$ and that when spillovers exist they are sufficiently strong so as to outweigh all other factors in the location decision process. The third term in

[6] Ellison and Glaeser (1997) note that their model has an equivalent formulation in which each potential "location" is equally profitable on average and the reason why there are many more plants in California is that California is an aggregate of a larger number of "locations".

the profit function, ϵ_{km}, is a Weibull distributed random shock that is independent across plants and locations.

Proposition 2 *Consider the model of location choices with spillovers described above:*

(a) The Perfect Bayesian equilibrium outcome is essentially unique. In equilibrium, each plant k chooses the location m that maximizes $\log(x_m)+\epsilon_{km}$ if no plant with $k' \in \omega(k)$ has previously chosen a location, and the location of previously located plants with $k' \in \omega(k)$ if some such plants have previously chosen a location.

(b) If $0 \leq \gamma_0^s \leq \gamma_1^s, \gamma_2^s$ or $0 \leq \gamma_1^s, \gamma_2^s$ and $0 \leq \gamma_0 \leq \min(1/N_1, 1/N_2)$, then there exist distributions over the set of possible partitions for which

$$Prob\{k' \in \omega(k)\} = \begin{cases} \gamma_i^s & \text{if plants } k \text{ and } k' \text{ both belong to industry } i \\ \gamma_0^s & \text{if plants } k \text{ and } k' \text{ belong to different industries,} \end{cases}$$

(c) If the distribution satisfies the condition in part (b), then in any PBE of the model the agglomeration and coagglomeration indexes satisfy

$$\begin{aligned} E(\gamma_i) &= \gamma_i^s \\ E(\gamma^c) &= \gamma_0^s. \end{aligned}$$

Remarks:

1. Note that Proposition 2 shows a degree of robustness to equilibrium selection: it shows that the agglomeration index has the same expected value in any PBE of the sequential move games obtained by ordering the plants in different ways.

2. Proposition 2 also shows some robustness to the distribution of spillover benefits. Our agglomeration and coagglomeration indexes have the same expected value for any distribution over partitions satisfying the condition in part (b). The proof of the proposition describes a couple different ways to generate distributions satisfying the condition. One is very simple technically and has a four point support. Another generates coagglomeration patterns that look more reasonable by first creating clusters within each industry and then joining clusters across industries.

2.3.2 Shared natural advantage

Another mechanism that can lead to the coagglomeration of plants in two industries is the presence in some areas of "shared natural advantages" that provide benefits to firms in both industries. The natural advantages can be exogenous factors, as when a coastal location makes a state attractive both to shipbuilding plants and to oil refineries. They can also be endogenous factor advantages of the types described in each of Marshall's theories, e.g. airplane manufacturers and automobile parts manufacturers may be coagglomerated because both benefit from locating in areas with skilled machinists.

To model natural-advantage influenced location choice, we suppose that profits for a plant k that belongs to industry $i(k)$ and locates in area m are given by

$$\log(\pi_{mk}) = \log(\eta_m + \xi_{mi(k)}) + \epsilon_{mk}, \tag{2}$$

where the η_m, ξ_{mi}, and ϵ_{mk} are mutually independent random variables. The η_m can be thought of as reflecting shared natural advantages of each area m that make it attractive or unattractive to plants in both industries.[7] The ξ_{mi} reflect additional factors that are idiosyncratic to industry i. As in the previous model, we also assume that there are plant-idiosyncratic factors, ϵ_{mk}.

Proposition 3 *Suppose that profits are as in equation (2) and that each plant k chooses the location m that maximizes π_{mk}.*

(a) Suppose $0 < \gamma_1^{na} \leq \gamma_2^{na}$, and that $0 \leq \gamma_0^{na} \leq \frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}$. Write δ_{mi} for $\eta_m + \xi_{mi}$. Then, there exist distributional choices for the η_m and ξ_{mi} for which

$$E(\delta_{mi} / \sum_{m'=1}^{M} \delta_{m'i}) = x_m,$$

$$Var(\delta_{mi} / \sum_{m'=1}^{M} \delta_{m'i}) = \gamma_i^{na} x_m (1 - x_m),$$

$$Cov(\delta_{m1} / \sum_{m'=1}^{M} \delta_{m'1}, \delta_{m2} / \sum_{m'=1}^{M} \delta_{m'2}) = \gamma_0^{na} x_m (1 - x_m).$$

[7] These could include state policies as discussed in Thomas Holmes (1998).

(b) If the distributions of the η_m and the ξ_{mi} are such that the conditions in part (a) are satisfied and the ϵ_{mk} are independent Weibull random variables, then the agglomeration and coagglomeration indexes satisfy

$$\begin{aligned} E(\gamma_i) &= \gamma_i^{na} \\ E(\gamma^c) &= \gamma_0^{na}. \end{aligned}$$

Remarks:

1. As is described in more detail in the proof of Proposition 3, one specification of the shared- and industry-idiosyncratic natural advantages that can be made to satisfy the conditions in part (a) involves choosing the η_m and ξ_{mi} to be χ^2 random variables with appropriately chosen degrees of freedom. In this specification the δ_{mi} are χ^2 random variables with $2x_m(1-\gamma_i^{na})/\gamma_i^{na}$ degrees of freedom. The lowest level of coagglomeration, $E(\gamma^c)=0$, obtains when there are no shared natural advantages: if we assume that the η_m are identically zero, then the δ_{mi} are independent across industries and state-industry employments will be independent across industries. The greatest degree of coagglomeration, $E(\gamma^c)=\frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}$, obtains when we make the shared natural advantages as important as possible: if the ξ_{m2} are identically zero, then all of the natural advantages affecting industry 2 are shared natural advantages.[8]

2. Ellison and Glaeser (1997) also provide a result characterizing the expected value of the agglomeration index when both spillovers and natural advantages are present. This result does not have a clean generalization to the multi-industry case. The difficulty is that both agglomeration and coagglomeration are no longer independent of the equilibrium selection. For example, if a spillover exists between plants in separate industries, there will be more agglomeration in each industry if the plant from the more agglomerated industry chooses the joint location than if the plant from the less agglomerated industry does so.

3 Data on Coagglomeration

In this section we present some descriptive statistics on coagglomeration patterns.

[8] In this case, the η_m are distributed χ^2 with $2x_m(1-\gamma_2^{na})/\gamma_2^{na}$ degrees of freedom and the ξ_{m1} are χ^2 with $2x_m(\frac{1-\gamma_1^{na}}{\gamma_1^{na}}-\frac{1-\gamma_2^{na}}{\gamma_2^{na}})$ degrees of freedom.

We compute pairwise coagglomation measures for manufacturing industries using the confidential plant-level data from the U.S. Census Bureau's *Census of Manufactures*.[9] We examine the censuses from 1972 to 1997, each of which contains data on approximately 300,000 establishments employing about 17 million workers. We aggregate the plant-level employment data in the census up to the county-level, PMSA-level, and the state-level and compute coagglomeration metrics all three ways.[10] At the industry level we focus on the three-digit level of the 1987 Standard Industrial Classification (SIC3). The sample analyzed in this section includes 134 industries, consisting of all SIC3 manufacturing industries except Tobacco (210s), Fur (237), and Search and Navigation Equipment (381).[11]

Table 1 presents descriptive statistics of several measures of agglomeration and coagglomeration.[12] The table is divided into three panels. The top panel presents indices calculated from state-level employment data. The first row shows that the EG industrial agglomeration index remains fairly stable between 1972 and 1982, and then falls by about 10% in the following decade. The next two rows summarize trends in the pairwise coagglomeration index. The mean pairwise coagglomeration is approximately zero. This is largely by definition: our benchmark measure of a state's "size" is its share of manufacturing employment so each industry's deviations from the benchmark will be approximately uncorrelated with the average of the deviations of all other industries. The standard deviation of the coagglomeration index is more interesting, showing a decline (tighter distribution) from 1972 to 1997.

The second panel presents corresponding figures computed using PMSA-level employments. The average decline in agglomeration from 1982 to 1992 is smaller at this geographic level and agglomeration appears to have increased from 1992 to 1997. The coagglomeration distribution again shows a declining standard deviation. At the industry-pair level,

[9] Timothy Dunne, Mark Roberts, and Larry Samuelson (1989a, 1989b), Robert McGuckin and Suzanne Peck (1992), Steven Davis, John Haltiwanger, and Scott Schuh (1996), and David Autor, William Kerr, and Adriana Kugler (forthcoming) provide detailed accounts of this dataset.

[10] We use reported employment in all manufacturing establishments excluding auxiliary units as our measure x of aggregate activity in the geographic unit.

[11] These six industries are omitted due to major industry reclassifications at the plant-level in the Census of Manufacturers that are difficult to interpret.

[12] Additional details on the dataset construction are catalogued in the data appendix. A portion of these coagglomeration estimates have been released for public use by the Census Bureau and are available from the authors upon request.

the coagglomeration indices computed using the PMSA-level data have an 0.59 correlation with indices computed from state-level data.

A nice feature of the Census of Manufactures is that one can track plants over time and separate new plants from old plants. The third panel provides statistics on agglomeration and coagglomeration indices for the new "startups" in each industry.[13] The agglomeration and coagglomeration of these startups could be different from the overall pattern because they are less tied to past industrial centers than existing plants or the new establishments of existing firms (see Guy Dumais, Ellison, and Glaeser 2002) and their location choices come after the inter-industry dependencies described below are formed. These measures are naturally more noisy than those calculated through total employment due to smaller number of plants involved and the distinct sets of plants being considered in each census year. The agglomeration data show an initial decline and a later increase, particularly in the final census year. This pattern is reflected in the standard deviation of the coagglomeration index too. At the industry-pair level, the correlation between coagglomeration measures computed at the state level using all firms and those computed using new startups is 0.33.

Dumais *et al.* (2002) noted that the EG agglomeration index for an industry is highly correlated over time (even relative to the magnitude of state-industry employment changes). Table 2 indicates that coagglomeration indices are also highly correlated over time. For example, the number in the upper left cell indicates that the correlation between the 1972 and 1977 coagglomeration indices for an industry-pair is 0.953. The correlations are at least 0.936 for each five-year period. The correlation between 1972 and 1997 coagglomeration indices is still about 0.740.

Table 3 contains a list of the fifteen most coagglomerated industry pairs. Most involve textile and apparel industries, which are heavily concentrated in North Carolina, South Carolina, and Georgia. None of these coagglomerations are as strong as the within-industry agglomerations of the most agglomerated industries. For example, Ellison and Glaeser

[13]More precisely, we first compute the total employment in each state-industry attributable to plants that did not appear in the previous census and did not belong to a firm that existed in the previous census (in this or any other industry). We then compute the agglomeration and coagglomeration indices using these totals as the state-industry employments. Approximately 80% of new manufacturing plants are startups in this sense. These startups enter at smaller sizes and account for about 50% of entering establishment employment. See Kerr and Ramana Nanda (2006) for more detail regarding the differences in entry sizes and entry rates between firm births and the expansion establishments of existing firms.

(1997) find that $\gamma = 0.63$ for the fur industry (SIC 237). Many, many industry-pairs have approximately zero coagglomeration. Negative values of the index arise when pairs of industries are agglomerated in different areas. The lowest value of -0.065 obtains for the coagglomeration of the Guided Missiles and Space Vehicles (376) and Railroad Equipment (374) industries. We imagine that most strong negative coagglomerations like this are mostly due to coincidence.

Appendix Table 1 summarizes the mean 1987 coagglomeration between SIC3 pairs within SIC2 pairwise bins. The matrix confirms that SIC3 pairs within the same SIC2 category are generally positively coagglomerated. Apart from the high coagglomeration of the subindustries of the textile industry (SIC 22), none of the means are very large. This further illustrates that there is a great deal of idiosyncratic variation in coagglomeration levels across industry pairs. The remainder of this paper attempts to exploit this variation to provide insight into the relative importance of different theories of geographic concentration.

4 Why Do Firms Agglomerate? Empirical Methodology

The gains from concentration, whether in cities or geographic clusters, always ultimately come from reducing some form of transport costs. Marshall emphasized that these transport costs could be for goods, people, or ideas. Firms locate near suppliers or customers to reduce the costs of buying or selling goods. Firms concentrate to reap the advantages from a large pool of potential employees. Firms locate near one another to reduce the costs of accessing new ideas and innovations. Our primary goal is to assess which of these are relatively more important.

Interesting papers by Audretsch and Feldman (1996) and Rosenthal and Strange (2001) have addressed this question by examining cross-industry variation in the degree to which industries are agglomerated. The idea of these papers is that even though all three of Marshall's theories predict that industries will agglomerate, one might be able to tell them apart by looking at which industries are more and less agglomerated. Audretsch and Feldman examine whether industries that are more R&D intensive are more agglomerated, and Rosenthal and Strange add proxies for Marshall's other factors as well, e.g. looking at

whether agglomeration is greater in industries with highly educated workers and in which material input costs are large relative to value-added.

The motivation for our empirical approach is that there may be a great deal of additional information in coagglomeration patterns. For example, we can examine not just whether input-intensive industries are agglomerated, but whether they are located *near the industries that produce their inputs.* There is additional information about location patterns to explore because industries that are agglomerated will be coagglomerated with some industries but not with others. There is also additional potentially useful variance in the explanatory variables at the industry-pair level: each industry has supplier/customer relationships with some industries and not with others; each industry has labor needs that are similar to those of some other industries and unlike those of others; and each industry is more likely to benefit from ideas generated by some industries than others.

The empirical strategy in this paper is to look at whether industries locate near other industries that are their suppliers or customers, near other industries that use similar labor, or near other industries that might share ideas. We do this via regressions with pairwise coagglomeration as the dependent variable and proxies for the importance of Marshall's agglomerative forces as the independent variables. Our goal is to learn not just about coagglomeration, but to learn more generally about the relative importance of goods, people, and ideas in the location decisions of manufacturing firms.

In the following subsections, we briefly discuss the three agglomeration forces and our approach to measuring them. Our empirical specification will look at the extent to which every pair of industries co-locates, so our empirical strategy requires us to construct potential explanatory variables reflecting the extent to which each pair of industries connects in goods, people, and ideas. We will do this in a single cross-section: we regress the 1987 values of the coagglomeration index on measures of industry-pair connections constructed using data from as close to 1987 as possible.[14]

[14]We did not feel that it was worthwhile to try to do this analysis in a panel setting for several reasons: we know that industry-pair coagglomeration is very highly correlated over time; we think that the industry-pair connections also do not change greatly over time; and data limitations would prevent us from calculating several of our measures at higher frequency.

4.1 Proximity to customers and suppliers: Goods

The most straightforward reason for firms to locate near one another is to reduce the costs of getting inputs or shipping goods to downstream customers. When inputs are far away from the eventual market, Marshall (1920) argued that firms will trade off the distance between customers and suppliers based on the costs of moving raw inputs and finished goods. The "new economic geography" of Masahisa Fujita, Krugman, and Anthony Venables (1999) views reducing the costs of transporting goods as the driver behind agglomeration. To assess the importance of this factor, we must assess the extent to which different industries buy and sell from one another. We use the 1987 Benchmark Input-Output Accounts of the Bureau of Economic Analysis to measure the extent that industries buy and sell from one another. The input-output tables provide commodity-level flows which we aggregate up to the three-digit SIC level. We define $Input_{i \leftarrow j}$ as the share of industry i's inputs that come from industry j. We also define $Output_{i \rightarrow j}$ as the share of industry i's outputs that are sold to industry j. These shares are calculated relative to all suppliers and customers, some of whom may be non-manufacturing industries or final consumers.

$Input_{i \leftarrow j}$ and $Output_{i \rightarrow j}$ are share variables that could go from zero to one. In fact, the highest observed value of $Input_{i \leftarrow j}$ is 0.39, which represents the share of inputs that come to Leather Tanning and Finishing (SIC 311) from Meat Products (SIC 201). The highest relative value of $Output_{i \rightarrow j}$ is 0.82, which represents the importance of output sales from Public Building and Related Furniture (SIC 253) to Motor Vehicles and Equipment (SIC 371).[15] For most industry pairs, of course, $Input_{i \leftarrow j}$ and $Output_{i \rightarrow j}$ are approximately zero.[16] To construct a single proxy for the connection in goods between a pair of industries, we define undirectional versions of the input and output variables by $Input_{ij} = \max\{Input_{i \leftarrow j}, Input_{j \leftarrow i}\}$ and $Output_{ij} = \max\{Output_{i \rightarrow j}, Output_{j \rightarrow i}\}$. We also define a combined input-output measure: $Input\text{-}Output_{ij} = \max\{Input_{ij}, Output_{ij}\}$.

One significant empirical issue is that these patterns of customers and suppliers may

[15] The large supplier share for Public Building and Related Furniture is due in part to the relatively small output of the industry. The largest absolute supplier relationship, Plastic Materials and Synthetics (282) sales to Misc. Plastic Products (308), has a relative output share of 0.32. The data appendix lists the top five dependencies for all of the metrics discussed below in both absolute and relative terms.

[16] Approximately 70 percent are less than 0.0001.

reflect rather than create geographic concentration. If an omitted variable causes two industries to locate in the same region, they may start selling to each other. To address the possibility that the vagaries of American geography are responsible for the input-output measures, we turn to U.K. input-output tables. Keith Maskus, C. Sveikauskas, and Allan Webster (1994) and Maskus and Webster (1995) use the 1989 Input-Output Balance for the United Kingdom published by the Central Statistical Office in 1992. The original table contained 102 sectors, but Maskus *et al.* (1994) aggregated those into 80 sectors that could be matched with U.S. industries. We form $UKInput_{ij}$ and $UKOutput_{ij}$ measures as described above using the U.K. input-output data and map these measures to the three-digit SIC code system. We will use these U.K. measures as instruments for the U.S. input-output relationships.

4.2 Labor market pooling: People

A second reason to coagglomerate is to take advantage of scale economies associated with a large labor pool. Marshall himself emphasized the risk-sharing properties of a large labor market. As individual firms become more or less productive, workers can shift across employers thereby maximizing productivity and reducing the variance of worker wages (see Diamond and Simon, 1990, for evidence and Krugman, 1991a, for a simple model). A variant on this theory is that agglomerations make it possible for workers to match better across firms and industries by providing a wider range of alternatives. Rotemberg and Saloner (2000) provide yet a third model of labor-market based agglomeration where firms cluster together so that workers will come and invest in human capital, knowing that they do not face ex post appropriation. A final model that emphasizes employment sharing is that new startups locate near older firms so that they can hire away their workers.

We will not be able to test between these different labor-based theories of agglomeration, but we can test whether industries that employ the same type of workers locate near one another. All of these labor market pooling hypotheses suggest that agglomeration occurs because workers are able to move across firms and industries. These cross-industry moves will only be likely if the industries use the same type of workers. Therefore we measure the extent to which different industries hire the same occupations. We start with the 1987

National Industrial-Occupation Employment Matrix (NIOEM) published by the Bureau of Labor Statistics (BLS). This matrix provides industry-level employment in 277 occupations, and we use this detail to determine for each industry the share of its employment association with each occupation, which we denote $Share_{io}$ for industry i and occupation o. We then construct a measure of the similarity of employment in industries i and j by defining $LaborCorrelation_{ij}$ to be the correlation of $Share_{io}$ and $Share_{jo}$ across occupations.

Table 4 contains summary statistics for this variable. The mean value is 0.470. The measured correlations of one arise because the industry-occupation matrix reports data for NIOEM industries, which is a coarser division than three-digit SIC industries. Motor Vehicles (371) and Motorcycles, Bicycles and Parts (375) have the most similar employment patterns (0.984) among industries with different NIOEM data.

As in the case of input-output matrices, reverse causality is a potential concern. Industries may be hiring the same type of workers because they are located in the same places and those workers happen to be there. To address this issue, we again turn to U.K. data where employment patterns should not reflect the patterns of American geography. Since the U.K. does not publish a detailed equivalent of the BLS NIOEM matrix, we constructed our own by pooling six years (2001-2006) of the U.K. Labour Force Survey (which is roughly akin to the U.S. Current Population Survey). We then developed matrices of the occupation-by-industry distribution of currently employed workers over all six surveys, which together contained 224,528 employed workers and 42,948 workers in manufacturing.

We mapped the British industry codes into the American system, but kept the occupation measures in their British format. Using this data, we calculated correlations in occupation employment shares between every two British industries just as we did for the American industries. This measure will be used as an instrument for the American labor correlation measure.

4.3 Intellectual or technology spillovers: Ideas

A final reason that firms co-locate is to speed the flow of ideas. Marshall himself emphasized the advantages that accrue to firms when workers learn skills quickly from each other in an industrial cluster. Alternatively, firms may locate near one another so that the

firm's leaders can learn from each other. Saxenian (1994) argues that this is one cause of industrial concentration in Silicon Valley. Glaeser and Matthew Kahn (2001) argue that the urbanization of high human capital industries, like finance, is evidence for the role that density plays in speeding the flow of ideas.

The potential for intellectual spillovers is harder to identify than the potential for trade in goods and for sharing a labor pool. We construct proxies using data from two different sources.

The first of these is Frederic Scherer's (1984) technology flow matrix. Scherer's matrix is designed to capture the extent to which R&D activity in one industry flows out to benefit another industry. This technology transfer occurs either through a supplier-customer relation between these two industries or through the likelihood that patented inventions obtained in one industry will find applications in the other industry. We develop two metrics, $TechIn_{i \leftarrow j}$ and $TechOut_{i \rightarrow j}$, for these technology flows that mirror $Input_{i \leftarrow j}$ and $Output_{i \rightarrow j}$ described above. These dependencies are again directional in nature and are calculated relative to total technology flows that include non-manufacturing industries and government R&D. The strongest relative technology flows are associated with Plastic Materials and Synthetics (282) and its relationships to Misc. Plastics Products (308), Tires and Inner Tubes (301), and Industrial Organic Chemicals (286).[17] Our second data source is the NBER Patent Database. Using data on patent citations for inventors residing in the U.S., we develop a measure of the extent to which technologies associated with industry i cite technologies associated with industry j, and vice versa. The measures $PatentIn_{i \leftarrow j}$ and $PatentOut_{i \rightarrow j}$ are normalized by total citations for the industries.[18]

For our regression analysis we construct undirectional measures of the intellectual

[17] Similar to the NIOEM industries, Scherer industries map to multiple SIC3s. Our regressions account for and are robust to this overlap.

[18] The NBER Patent Data File was originally compiled by Bronwyn Hall, Adam Jaffe, and Manuel Trajtenberg (2001). It contains records for all patents granted by the United States Patent and Trademark Office (USPTO) from January 1975 to December 1999. Each patent record provides information about the invention (e.g., technology classification, citations of prior art) and the inventors submitting the application (e.g., name, city). The USPTO issues patents by technology categories rather than by industries. Combining the work of Daniel Johnson (1999), Brian Silverman (1999), and Kerr (forthcoming), concordances are developed between the USPTO classification scheme and SIC3 industries (a probabilistic mapping). In practice, there is little directional difference between $PatentIn_{i \leftarrow j}$ and $PatentOut_{i \rightarrow j}$ due to the extensive number of citations within a single technology field, in which case the probabilistic citing and cited industry distributions are the same.

spillovers across an industry pair, $Tech_{ij}$ and $Patent_{ij}$, in a manner analogous to our construction of $Input\text{-}Output_{ij}$.

Many authors have used patent citations to assess intellectual spillovers, but they are obviously only an imperfect measure of intellectual spillovers.[19] As Michael Porter (1991) emphasizes, much knowledge sharing occurs between consumers and suppliers and this may be captured more by input-output relationships than by these citations. Idea sharing through the exchange of workers may be better captured by our occupational employment correlation than through patent-based metrics. As such, we see our patent citation measure as a proxy for the importance of exchanging technology rather than as a proxy for all forms of intellectual spillovers.

Again there is the concern of endogeneity of intellectual exchanges, as industries may cite each other's patents because of locational proximity. To address this issue, we use the U.K. patents in the NBER patent database to form a citations matrix based entirely on non-U.S. patents. We use the patent flow numbers across U.K. industries as an instrument for the U.S. technology flows.

5 Empirical Results

We now present our main empirical results. As described above, we examine the relationships between the coagglomeration metrics calculated from the Census Bureau data and various inter-industry dependency metrics. We first examine direct partial correlations evident in the U.S. data, and then we turn to instrumental variable regressions using U.K. data to confirm a causal interpretation. We find evidence to support all three agglomeration theories, and our results most strongly emphasize the importance of input-output and labor pooling explanations.

The core empirical specification is

$$Coagg_{ij} = \alpha + \beta_L LaborCorrelation_{ij} + \beta_{IO} InputOutput_{ij} + \beta_T Tech_{ij} + \varepsilon_{ij},$$

where $Coagg_{ij}$ is our measure of the pairwise coagglomeration between industries i and j

[19] See Zvi Griliches (1990), Jaffe, Trajtenberg, and Rebecca Henderson (1993), and Jaffe, Trajtenberg, and Michael Fogarty (2000).

in 1987. The sample contains 7381 industry pair observations: all distinct pairs from a sample of 122 industries.[20]

We perform these analyses with four different versions of the dependent variable: we calculate the coagglomeration measures using state-, PMSA-, and county-level total industry-employment data, and also with the state-level data on employment in startups.

To make it easier to assess the magnitude of each variable's importance, we normalize both the left- and right-hand side variables in all of our regressions so that they have standard deviation one.

5.1 Correlations in univariate regressions

Before proceeding to the actual regressions, Table 5 presents results from univariate regressions where coagglomeration is regressed on four different measures of the different theory: our measure of labor pool similarity (*LaborCorrelation*), our combined input-output measure (*Input-Output*), and the measures of technology flows from the Scherer matrix (*Tech*) and from patent citations (*Patent*). Each cell of the table reports a coefficient from a separate univariate regression. Each row represents a different explanatory variable. Each column corresponds to a different measure of coagglomeration. The first column uses the state-level measures. The second column uses the PMSA-level measures. The third column uses the county-level measures, and the fourth column uses the startup coagglomeration measure.

Column (1) shows that the basic relationships between the first three measures and state total employment coagglomeration are quite similar. A one standard deviation increase in the labor correlation measure is associated with a 0.18 standard deviation increase in the state-level coagglomeration measure. A one standard deviation increase in the input-output measure is association with a 0.205 standard deviation increase in the coagglomeration measure. The Scherer technology flow variable yields a 0.18 coefficient. The patent citation variable yields a somewhat lower coefficient of 0.08.

[20]The sample omits twelve industries for which we could measure coagglomeration: all Apparel industries (230s), a portion of Printing and Publishing (277-279), and Secondary Non-Ferrous Metals (334). Some exclusions are due to an inability to construct appropriate Marshallian explanatory measures and some are due to outlier concerns.

At the PMSA and county levels, the gap between the magnitudes of the effects widens. The coefficient on labor correlation is 0.106 and 0.082 in columns (2) and (3), respectively. The coefficient on input-output in the same two columns is 0.167 and 0.130. The coefficient on the Scherer technology flows is 0.148 and 0.107. Coagglomeration relationships are weaker at the metropolitan level, which can be explained if firms are drawn to counties because of other firms in neighboring counties.

The fourth column examines the coagglomeration of startup activity in industry pairs. The coefficients in these regressions are lower than in the previous regression. Again, input-output relationships seem to be the most important.

5.2 OLS regression results

Table 6 presents OLS coefficient estimates for our core empirical specification. Each column reports coefficients from a single regression with a pairwise coagglomeration (measured using state-level data) as the dependent variable. We find a coefficient of 0.146 for labor correlation, 0.149 for the input-output measure, and 0.112 for the Scherer technology flows.

In the second column, we break input-output effects into an input measure and an output measure. Both effects are quite significant and large.

The third column excludes all industry pairs involving two industries belonging to the same two-digit SIC industry. There are both conceptual and methodological reasons for this exclusion. Conceptually, we might think that industries within the same two-digit SIC code are more likely to be driven to coagglomerate because of omitted geographic factors that drive the location patterns of such similar industries. Methodologically, some of our measures, like the technology flow measure, have variation that straddles the two-digit and three-digit levels. The coefficient estimates in this regression are slightly lower, but similar in magnitude to the base regression in the first column.

The regressions in the fourth, fifth, and sixth columns provide another robustness check. For these regressions we have redefined the input-output and technology flow variables to be means (rather than maximums) of the directional variables on which they are based. In these specifications, input-output measures are generally more important, and labor correlation is somewhat less important.

Appendix Table 2 presents regressions similar to the base regression in the first column of Table 6, but with the three alternative coagglomeration measures as the dependent variables. These substitutions yield similar results.

Two general conclusions emerge from these regressions. First, all three of Marshall's (1920) theories regarding agglomeration find support in the coagglomeration patterns. Second, the input-output relationship appears to be the most important contributor. The labor pooling hypothesis finds the second most support.

5.3 IV regression results

In this section we present our core results using the U.K. instruments. Appendix Table 3 presents the first-stage regression estimates. The t-statistics are over 15 for the relevant instruments.

Table 7 presents estimates from sixteen regressions. The regressions reported in Panel A include the input-output and labor correlation measures of industry relatedness. The odd regressions (1), (3), (5), and (7) are OLS results. One slight difference from Table 6 is that we exclude all industry-pairs involving two firms in the same two-digit industry.[21] The even regressions (2), (4), (6), and (8) are the IV results. We present results for state-, PMSA-, and county-level coagglomeration and for the coagglomeration of startups.

Comparing regressions (1) and (2) in Panel A shows that the IV specifications cause the coefficients on labor correlation and the input-output measures to rise modestly. Both coefficients remain significant.

The regressions in Panel B also include the Scherer measure of technology flows (with U.K. patent flows as the instrument in the IV specifications). The first regression is Panel B is an exact duplicate of regression (3) in Table 6, repeated to permit easy comparison between the OLS and IV results. The differences between regressions (1) and (2) are modest in magnitude, but the coefficients on input-output and technology flows become statistically insignificant. The labor correlation remains robust.

[21] We made this change for two reasons. First, the U.K. input-output tables have a relevant limitation. We explicitly exclude intra-industry flows at the SIC3 level from the U.S. input-output tables. In several cases, we are required to map the same U.K. industry to multiple SIC3 industries within an SIC2. In these cases, we are not able to distinguish flows across these SIC3 industries from intra-industry flows. In addition, some of the instruments have limited variation within two-digit industries.

The regressions in columns (3) and (4) use PMSA-level coagglomeration as the dependent variable. In both panels there is a reversal in the importance-ranking of the labor correlation and input-output measures when we move from the OLS to the IV specifications. In the OLS estimates, input-output relationships look dramatically more important than labor correlation. In the IV specification, labor correlation is both larger in magnitude and more statistically significant.

At the county level in Panel A, input-output measures are more important and significant in both the OLS and IV regressions. In Panel B, the IV measures are all statistically insignificant.

Finally, regressions (7) and (8) examine the coagglomeration of new firm births in industry pairs. In both panels the use of instrumental variables makes both the input-output and labor correlation measures much more important. In each case, these the coefficients are statistically significant in the IV specifications.

We interpret Table 7 as indicating that at least the labor correlation and input-output relationships with coagglomeration are robust to our IV approach. The IV specification generally causes coefficients to rise, but the increase is generally statistically insignificant. When we include the technology flows measure, measurement becomes more difficult. Our instrument for technology flows is highly correlated with the input-output measure, so it becomes difficult to identify separately input-output effects and technology flow effects.

6 Conclusion

Our first conclusion from our analysis of coagglomeration patterns is that there is support for the importance of all three theories of agglomeration. In the OLS specifications, all variables have statistically significant and economically meaningful effects. The IV results continue to show a robust effect for labor correlation. The input-output coefficients are similar in magnitude or rise in size, but are less significant when we also include technology flows. The technology flows measures are less robust when we use our IV measures.

Which of the theories seems to be more important? Our basic conclusion is that this work suggests all three are roughly equal in magnitude. A one standard deviation growth in labor correlation or input-output increases coagglomeration by around one seventh of a

standard deviation. In some specifications, the technology flows effect is somewhat weaker, but in others it is also close in magnitude.

It is unclear how these results would generalize to non-manufacturing industries. Services are more costly to transport since they involve face-to-face interaction and therefore we might think that input-output relationships are particularly important in that sector (Jed Kolko, 1997). Ideas may be more important in more innovative sectors, so idea flows might be more important elsewhere. But at least in manufacturing, transport costs for goods, people, and ideas all still seem to matter, and all three of Marshall's theories find vindication in the data.

References

Audretsch, David B. and Feldman, Maryann P. (1996): "R&D Spillovers and the Geography of Innovation and Production," *American Economic Review*, 86, 630–640.

Autor, David, Kerr, William and Kugler, Adriana (forthcoming): "Do Employment Protections Reduce Productivity? Evidence from U.S. States," *The Economic Journal*.

Davis, Steven, Haltiwanger, John and Schuh, Scott (1996): *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Diamond, Charles and Simon, Curtis (1990): "Industrial Specialization and the Returns to Labor," *Journal of Labor Economics*, 8, 175–201.

Dumais, Guy, Ellison, Glenn and Glaeser, Edward L. (2002): "Geographic Concentration as a Dynamic Process," *Review of Economics and Statistics*, 84, 193–204.

Dunne, Timothy, Roberts, Mark and Samuelson, Larry (1989a): "The Growth and Failure of Manufacturing Plants in the U.S.," *Quarterly Journal of Economics*, 104, 671–698.

Dunne, Timothy, Roberts, Mark and Samuelson, Larry (1989b): "Plant Turnover and Gross Employment Flows in the U.S. Manufacturing Sector," *Journal of Labor Economics*, 7, 48–71.

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

Ellison, Glenn and Glaeser, Edward L. (1999): "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?," *Papers and Proceedings, American Economic Review*, 89, 311–316.

Florence, P. Sargant (1948): *Investment, Location and Size of Plant*. London, U.K.: Cambridge University Press.

Fuchs, Victor (1962): *Changes in the Location of Manufacturing in the US Since 1929*. New Haven, CT: Yale University Press.

Fujita, Masahisa, Krugman, Paul and Venables, Anthony (1999): *The Spatial Economy: Cities, Regions and International Trade*. Cambridge, MA: MIT Press.

Glaeser, Edward L. and Kohlhase, Janet E. (2003): "Cities, Regions and the Decline of Transport Costs," *Papers in Regional Science*, 83, 197–228.

Glaeser, Edward L. and Khan, Matthew (2001): "Decentralized Employment and the Trans-

formation of the American City," NBER Working Paper 8117.

Griliches, Zvi (1990): "Patent Statistics as Economic Indicators: A Survey," *Journal of Economic Literature,* 28, 1661–1707.

Hall, Bronwyn, Jaffe, Adam and Trajtenberg, Manuel (2001): "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498.

Henderson, J. Vernon (2003): "Marshall's Scale Economies," *Journal of Urban Economics*, 53, 1–28.

Holmes, Thomas (1998): "The Effects of State Policies on the Location of Manufacturing: Evidence from State Borders," *Journal of Political Economy*, 106, 667–705.

Hoover, E.M. (1948): *The Location of Economic Activity.* New York, NY: McGraw Hill.

Jaffe, Adam, Trajtenberg, Manuel and Fogarty, Michael (2000): "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," *Papers and Proceedings, American Economic Review*, 90, 215–218.

Jaffe, Adam, Trajtenberg, Manuel and Henderson, Rebecca (1993): "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 108, 577–598.

Johnson, Daniel (1999): "150 Years of American Invention: Methodology and a First Geographic Application," Wellesley College Economics Working Paper 99-01.

Johnson, Norman L, Kotz, Samuel and Balakrishnan, N. (1994): *Continuous Univariate Distributions,* Vol. 1. New York: John Wiley & Sons.

Johnson, Norman L, Kotz, Samuel and Balakrishnan, N. (1995): *Continuous Univariate Distributions,* Vol. 2. New York: John Wiley & Sons.

Kerr, William (forthcoming): "Ethnic Scientific Communities and International Technology Diffusion," *The Review of Economics and Statistics.*

Kerr, William and Nanda, Ramana (2006): "Banking Deregulation, Financing Constraints, and Entrepreneurship," Harvard Business School Working Paper 07-033.

Kolko, Jed (1999): "Can I Get Some Service Here? Information Technology, Service Industries, and the Future of Cities," Harvard University Working Paper.

Krugman, Paul (1991a): *Geography and Trade.* Cambridge, MA: MIT Press.

Krugman, Paul (1991b): "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99, 483–499.

Marshall, Alfred (1920): *Principles of Economics.* London, U.K.: MacMillan and Co.

Maskus, Keith, Sveikauskas, C. and Webster, Allan (1994): "The Composition of the Human Capital Stock and Its Relation to International Trade: Evidence from the U.S. and Britain," *Weltwirtschaftliches Archiv*, 1994, Band 130, Heft 1.

Maskus, Keith and Webster, Allan (1995): "Factor Specialization in U.S. and U.K. Trade: Simple Departures from the Factor-Content Theory," *Swiss Journal of Economics and Statistics*, 1, 419–440.

McGuckin, Robert and Peck, Suzanne (1992): "Manufacturing Establishments Reclassified into New Industries: The Effect of Survey Design Rules," Center for Economic Studies Working Paper 92-14.

Porter, Michael E. (1990): *The Competitive Advantage of Nations.* New York, NY: The Free Press.

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

Rotemberg, Julio and Saloner, Garth (2000): "Competition and Human Capital Accumulation: A Theory of Interregional Specialization and Trade," *Regional Science and Urban Economics* 30, 373–404.

Saxenian, AnnaLee (1994): *Regional Advantage: Culture and Competition in Silicon Valley and Route 128.* Cambridge, MA: Harvard University Press.

Scherer, Frederic M. (1984): "Using Linked Patent Data and R&D Data to Measure Technology Flows," in Griliches, Zvi (ed.) *R & D, Patents and Productivity.* Chicago, IL: The University of Chicago Press.

Silverman, Brian (1999): "Technological Resources and the Direction of Corporate Diversification: Toward an Integration of the Resource-Based View and Transaction Cost Economics," *Management Science*, 45, 1109–1124.

von Thünen, Johann Heinrich (1826): *Der Isolirte Staat in Beziehung auf Landwirthschaft und Nationalökonomie, oder Untersuchungen über den Einfluss, den die Getreidepreise, der Reichtum des Bodens und die Abgaben auf den Ackerbau Ausüben.* Reprinted in English as *Von Thünen's Isolated State*, Pergamon Press, 1966.

Appendix A

PROOF OF PROPOSITION 1: Note that

$$\begin{aligned} G &= \sum_{m=1}^{M}(w_1 s_{m1} + w_2 s_{m2} - x_m)^2 = \sum_{m=1}^{M}(w_1(s_{m1} - x_m) + w_2(s_{m2} - x_m))^2 \\ &= w_1^2 \sum_{m=1}^{M}(s_{m1} - x_m)^2 + w_2^2 \sum_{m=1}^{M}(s_{m2} - x_m)^2 + 2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m) \\ &= w_1^2 G_1 + w_2^2 G_2 + 2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m), \end{aligned}$$

and that $H = w_1^2 H_1 + w_2^2 H_2$. Hence,

$$\begin{aligned} (1 - \sum_{i=1}^{2} w_i^2)\gamma^c &= G/(1 - \sum_m x_m^2) - H - \sum_{i=1}^{2} \gamma_i w_i^2 (1 - H_i) \\ &= G/(1 - \sum_m x_m^2) - (w_1^2 H_1 + w_2^2 H_2) - \sum_{i=1}^{2} \left(\frac{G_i/(1 - \sum_m x_m^2) - H_i}{1 - H_i} \right) w_i^2 (1 - H_i) \\ &= \frac{w_1^2 G_1 + w_2^2 G_2 + 2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m) - \sum_{i=1}^{2} w_i^2 G_i}{1 - \sum_{m=1}^{M} x_m^2} \\ &= \frac{2w_1 w_2 \sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}. \end{aligned}$$

The final formula results from noting that $2w_1 w_2 = 1 - \sum_{i=1}^{2} w_i^2$ when $w_1 + w_2 = 1$.

PROOF OF PROPOSITION 2: Part (a) of the theorem follows immediately from backward induction. The final plant to move must choose in this way. Given that the final plant will locate in this way, the next-to-last plant maximizes its payoff by choosing the location that maximizes $\log(x_m) + \epsilon_{km}$ if it has no spillover with a previously located plant, because it will receive full spillover benefits from the final plant (if such spillovers exist) regardless of its location choice. The qualification "essentially unique" in the proposition reflects that the maximizing choice is not unique when the maximizer of $\log(x_m) + \epsilon_{km}$ is not unique. This occurs with probability zero.

Part (b) states that we can choose a distribution over partitions that satisfies

$$\text{Prob}\{k' \in \omega(k)\} = \begin{cases} \gamma_i^s & \text{if plants } k \text{ and } k' \text{ both belong to industry } i \\ \gamma_0^s & \text{if plants } k \text{ and } k' \text{ belong to different industries,} \end{cases}$$

if either of two hypotheses holds.

The first hypothesis is that $0 \leq \gamma_0^s \leq \gamma_1^s, \gamma_2^s$. In this case, a four-point distribution suffices. Let ω_0 be the fully disjoint partition: $\omega_0(k) = \{k\}$ for all k. Let ω_i be the partition in which all plants in industry i are in a single cluster with the remaining plants disjoint: $\omega_i(k) = K_i$ if $k \in K_i$ and $\omega_i(k) = \{k\}$ otherwise. Let ω_{12} be the partition with all plants in a single cluster: $\omega_{12}(k) = K_1 \cup K_2$ for all k. The distribution that places probability γ_0^s on ω_{12}, probability $\gamma_i^s - \gamma_0^s$ on ω_i, and the remaining probability on ω_0 has the desired property.

The second hypothesis is that $0 \leq \gamma_1^s, \gamma_2^s$ and $0 \leq \gamma_0 \leq \min(1/N_1, 1/N_2)$. In this case, it is simplest to describe the construction of a distribution on the set of partitions on $K_1 \cup K_2$ as a two-step process. Let p_1 be a probability distribution over partitions of K_1 that satisfies $p_1(\{\omega | k' \in \omega(k)\}) = \gamma_1^s$ for all $k, k' \in K_1$. This can be done easily by putting probability γ_1^s on the partition with all plants in a single cluster and the remaining probability on a disjoint partition, and can also be done in many other ways if γ_1^s is not too large. Similarly, let p_2 be a distribution over partitions of K_2 that satisfies $p_2(\{\omega | k' \in \omega(k)\}) = \gamma_2^s$ for all $k, k' \in K_2$. To choose a partition of $K_1 \cup K_2$, first draw partitions ω_1 of K_1 and ω_2 of K_2 according to p_1 and p_2. Let C_i be the set of clusters in partition i: $C_i = \{S \subset K_i | \omega_1(k) = S \text{ for some } k \in K_1\}$. Assuming WLOG that $|C_1| < |C_2|$, let f be a one-to-one function from C_1 to C_2 chosen uniformly from the set of all such functions. Then, define a partition ω on $K_1 \cup K_2$ by setting $\omega(k) = \omega_1(k)$ with probability $1 - |C_2|\gamma_0$ and $\omega(k) = \omega_1(k) \cup f(\omega_1(k))$ with probability $|C_2|\gamma_0$ for $k \in K_1$, and defining $\omega(k) = \omega_2(k)$ if $k \in K_2$ and k has not previously been defined as belonging to some $\omega(k)$ with $k \in K_1$. (The randomization in this definition is perfectly correlated across k and k' if $k' \in \omega_1(k)$ and can have any correlation if k and k' are not in the same cluster.) It is straightforward that a partition created this way has the desired property.

Part (c) is a corollary of Proposition 0. Let u_{km} be an indicator for plant k locating in area m. A standard property of the logit model is that $\text{Prob}\{u_{km} = 1\} = x_m / \sum_{m'} x_{m'} = x_m$. The locations of plants k and k' are identical if $k' \in \omega(k)$ and independent otherwise, so

$$E(u_{km} u_{k'm} | \omega) = \begin{cases} x_m & \text{if } k' \in \omega(k) \\ x_m^2 & \text{otherwise.} \end{cases}$$

The unconditional expectation is $E(u_{km} u_{k'm}) = x_m^2 + \text{Prob}\{k' \in \omega(k)\}(x_m - x_m^2)$. Using this we calculate

$$\begin{aligned} \text{Corr}(x_{km}, x_{k'm}) &= \frac{E(u_{km}u_{k'm}) - E(u_{km})E(u_{k'm})}{\sqrt{\text{Var}(u_{km})\text{Var}(u_{k'm})}} \\ &= \frac{x_m^2 + \text{Prob}\{k' \in \omega(k)\}(x_m - x_m^2) - x_m^2}{\sqrt{x_m(1-x_m)x_m(1-x_m)}} \\ &= \text{Prob}\{k' \in \omega(k)\} \end{aligned}$$

Hence, the hypothesis of Proposition 0 is satisfied whenever the condition on the distribution over partitions in part (b) of Proposition 2 holds.

Proof of Proposition 3: Suppose that the η_m and ξ_{mi} are independent χ^2 random variables with $\frac{1-\gamma_2^{na}}{\gamma_2^{na}} 2c_m x_m$ and $\frac{1-\gamma_i^{na}}{\gamma_i^{na}} 2x_m - \frac{1-\gamma_2^{na}}{\gamma_2^{na}} 2c_m x_m$ degrees of freedom, respectively, for some constants $c_m \in [0, 1]$. The additive property of χ^2 random variables implies that δ_{mi} is a χ^2 random variable with $\frac{1-\gamma_i^{na}}{\gamma_i^{na}} 2x_m$ degrees of freedom. Note that δ_{mi} and $\delta_{m'i}$ are independent if $m \neq m'$. A standard result on Chi-square distributions implies that $\delta_{mi} / \sum_{m'=1}^{M} \delta_{m'i}$ has a Beta distribution with parameters $\frac{1-\gamma_i^{na}}{\gamma_i^{na}} x_m$ and $\frac{1-\gamma_i^{na}}{\gamma_i^{na}}(1 - x_m)$.[22]

[22] See Chapter 25 of Johnson, Kotz and Balakrishnan (1995).

A Beta distribution with parameters θ_1 and θ_2 has expectation $\theta_1/(\theta_1+\theta_2)$ and variance $\frac{\theta_1\theta_2}{(\theta_1+\theta_2)^2(\theta_1+\theta_2+1)}$. Using these formulas gives

$$E\left(\frac{\delta_{mi}}{\sum_{m'=1}^{M}\delta_{m'i}}\right)=\frac{\frac{1-\gamma_i^{na}}{\gamma_i^{na}}x_m}{\frac{1-\gamma_i^{na}}{\gamma_i^{na}}}=x_m$$

$$\text{Var}\left(\frac{\delta_{mi}}{\sum_{m'=1}^{M}\delta_{m'i}}\right)=\frac{\left(\frac{1-\gamma_i^{na}}{\gamma_i^{na}}\right)^2 x_m(1-x_m)}{\left(\frac{1-\gamma_i^{na}}{\gamma_i^{na}}\right)^2\frac{1}{\gamma_i^{na}}}=\gamma_i^{na}x_m(1-x_m).$$

This shows that the distributions have two of the three desired properties given in part (a) of the Proposition.

To complete the proof of part (a) it suffices to show that the third property,

$$\text{Cov}(\delta_{m1}/\sum_{m'=1}^{M}\delta_{m'1},\delta_{m2}/\sum_{m'=1}^{M}\delta_{m'2})=\gamma_0^{na}x_m(1-x_m)$$

holds for some choice of $c_m\in[0,1]$. The covariance is a continuous function of c_m. When $c_m=0$, the covariance is zero. Hence, by the intermediate value theorem we can complete the proof by showing that the covariance is equal to $\frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}x_m(1-x_m)$ when $c_m=1$.

When $c_m=1$ the covariance can be written as

$$\begin{aligned}\text{Cov}\left(\frac{\delta_{m1}}{\sum_{m'=1}^{M}\delta_{m'1}},\frac{\delta_{m2}}{\sum_{m'=1}^{M}\delta_{m'2}}\right) &= \text{Cov}\left(\frac{\eta_m}{\sum_{m'=1}^{M}\eta_{m'}},\frac{\eta_m+\xi_{m2}}{\sum_{m'=1}^{M}\eta_{m'}+\xi_{m'2}}\right),\\ &= \text{Cov}\left(\frac{Y_0}{Y_0+Y_0'},\frac{Y_0+Y_1}{Y_0+Y_0'+Y_1+Y_2}\right),\end{aligned}$$

where $Y_0=\eta_m$, $Y_0'=\sum_{m'\neq m}\eta_{m'}$, $Y_1=\xi_{m2}$, and $Y_2=\sum_{m'\neq m}\xi_{m'2}$. Note that Y_0, Y_0', Y_1 and Y_2 are mutually independent Chi-square random variables. By rewriting the last term on the right side of the above expression as $\frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}+\frac{Y_1}{Y_1+Y_2}\frac{Y_1+Y_2}{Y_0+Y_0'+Y_1+Y_2}$ we find that it is equal to

$$\begin{aligned}&\text{Cov}\left(\frac{Y_0}{Y_0+Y_0'},\frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}+\frac{Y_1}{Y_1+Y_2}\frac{Y_1+Y_2}{Y_0+Y_0'+Y_1+Y_2}\right)\\ &\quad=\text{Cov}\left(\frac{Y_0}{Y_0+Y_0'},\frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}\right)+\text{Cov}\left(\frac{Y_0}{Y_0+Y_0'},\frac{Y_1}{Y_1+Y_2}\frac{Y_1+Y_2}{Y_0+Y_0'+Y_1+Y_2}\right)\end{aligned}$$

Another standard property of Chi-square (and Gamma) random variables is that $\frac{Y_0}{Y_0+Y_0'}$ and Y_0+Y_0' are independent.[23] This immediately implies that the second covariance in the line above is zero. It also implies that $\frac{Y_0}{Y_0+Y_0'}$ and $\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}$ are independent. This implies

$$\text{Cov}\left(\frac{Y_0}{Y_0+Y_0'},\frac{Y_0}{Y_0+Y_0'}\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}\right)=\text{Var}\left(\frac{Y_0}{Y_0+Y_0'}\right)E\left(\frac{Y_0+Y_0'}{Y_0+Y_0'+Y_1+Y_2}\right).$$

[23] See Chapter 17 of Johnson, Kotz and Balakrishnan (1994).

Plugging in the appropriate degrees of freedom into the formulas for the mean and variance of Beta-distributed random variables we find that this is equal to

$$\gamma_2^{na} x_m(1-x_m)\frac{\frac{1-\gamma_2^{na}}{\gamma_2^{na}}}{\frac{1-\gamma_1^{na}}{\gamma_1^{na}}} = x_m(1-x_m)\frac{1-\gamma_2^{na}}{1-\gamma_1^{na}}\gamma_1^{na}.$$

Table 1: Levels of Geographic Agglomeration and Coagglomeration 1972-1997

	1972	1977	1982	1987	1992	1997
			A. State-Level Total Employment			
EG Agglomeration Index γ Mean	0.0398	0.0399	0.0392	0.0368	0.0351	0.0342
EG Coagglomeration Index γc Mean	0.0003	0.0003	0.0002	0.0004	0.0003	0.0003
EG Coagglomeration Index γc SD	0.0150	0.0139	0.0140	0.0133	0.0129	0.0124
			B. PMSA-Level Total Employment			
EG Agglomeration Index γ Mean	0.0298	0.0292	0.0286	0.0285	0.0271	0.0299
EG Coagglomeration Index γc Mean	0.0003	0.0003	0.0002	0.0003	0.0002	0.0002
EG Coagglomeration Index γc SD	0.0086	0.0075	0.0069	0.0061	0.0054	0.0060
			C. State-Level Employment in Firm Births			
EG Agglomeration Index γ Mean	0.0290	0.0022	0.0121	0.0107	0.0158	0.0285
EG Coagglomeration Index γc Mean	0.0001	0.0003	0.0003	0.0005	0.0004	0.0003
EG Coagglomeration Index γc SD	0.0193	0.0172	0.0177	0.0150	0.0187	0.0181

Notes: Measures of industrial agglomeration and coagglomeration calculated from the Census of Manufacturers. Estimates include all manufacturing SIC3 industries, except those listed in the text, for 134 observations per year.

Table 2: Correlation of EG Coagglomeration Index

	1972	1977	1982	1987	1992
1977	0.953				
1982	0.891	0.944			
1987	0.841	0.889	0.936		
1992	0.791	0.840	0.895	0.959	
1997	0.740	0.789	0.832	0.890	0.941

Notes: See Table 1. EG Coagglomeration Index measured through state total employments for each industry.

Table 3: Highest 1987 Pairwise Coagglomerations

Rank	Industry 1	Industry 2	Coaggl.
1	Broadwoven Mills, Cotton (221)	Yarn and Thread Mills (228)	0.207
2	Knitting Mills (225)	Yarn and Thread Mills (228)	0.187
3	Broadwoven Mills, Fiber (222)	Textile Finishing (226)	0.178
4	Broadwoven Mills, Cotton (221)	Broadwoven Mills, Fiber (222)	0.171
5	Broadwoven Mills, Fiber (222)	Yarn and Thread Mills (228)	0.164
6	Handbags (317)	Photographic Equipment (386)	0.155
7	Broadwoven Mills, Wool (223)	Carpets and Rugs (227)	0.149
8	Carpets and Rugs (227)	Yarn and Thread Mills (228)	0.142
9	Photographic Equipment (386)	Jewelry, Silverware, Plated Ware (391)	0.139
10	Textile Finishing (226)	Yarn and Thread Mills (228)	0.138
11	Broadwoven Mills, Cotton (221)	Textile Finishing (226)	0.137
12	Broadwoven Mills, Cotton (221)	Carpets and Rugs (227)	0.137
13	Broadwoven Mills, Cotton (221)	Knitting Mills (225)	0.136
14	Carpets and Rugs (227)	Pulp Mills (261)	0.110
15	Jewelry, Silverware, Plated Ware (391)	Costume Jewelry and Notions (396)	0.107

Notes: See Table 1. EG Coagglomeration Index measured through state total employments for each industry.

Table 4: Descriptive Statistics for 1987 Pairwise Coagglomeration Regressions

	Mean	Standard Deviation	Minimum	Maximum
	A. Pairwise Coagglomeration Measures			
State Total Empl. Pairwise Coaggl.	0.000	0.013	-0.065	0.207
PMSA Total Empl. Pairwise Coaggl.	0.000	0.006	-0.025	0.119
County Total Empl. Pairwise Coaggl.	0.000	0.003	-0.018	0.080
State Birth Empl. Pairwise Coaggl.	0.000	0.015	-0.082	0.259
	B. Pairwise Labor Similarities Index			
Labor Correlation	0.470	0.226	-0.046	1.000
	C. Pairwise Input-Output Relationship Indices			
Input-Output Maximum	0.007	0.029	0.000	0.823
Input-Output Mean	0.002	0.010	0.000	0.240
Input Maximum	0.005	0.019	0.000	0.392
Input Mean	0.002	0.010	0.000	0.196
Output Maximum	0.005	0.026	0.000	0.823
Output Mean	0.002	0.013	0.000	0.411
	D. Pairwise Technology Relationship Indices			
Scherer R&D Tech Maximum	0.005	0.026	0.000	0.625
Scherer R&D Tech Mean	0.002	0.010	0.000	0.263
Patent Citation Tech Maximum	0.015	0.025	0.000	0.400
Patent Citation Tech Mean	0.007	0.014	0.000	0.203

Notes: Descriptive statistics for 1987. All pairwise combinations of manufacturing SIC3 industries are included, except those listed in the text, for 7381 observations. Coagglomeration measures are calculated from the 1987 Census of Manufacturers. Labor Correlation indices are calculated from the BLS National Industry-Occupation Employment Matrix for 1987. Input-Output relationships are calculated from the BEA Benchmark Input-Output Matrix for 1987. Technology Flows are calculated from the Scherer (1984) R&D tables for the 1970s and from the NBER Patent Citation Database for 1975-1997. See the data appendix for further details.

Table 5: OLS Univariate Specifications for 1987 Pairwise Coagglomeration

Each row and column reports a separate estimation with single regressor	Dependent Variable is EG Coagglomeration Index			
	State Total Employment Coagglomeration	PMSA Total Employment Coagglomeration	County Total Employment Coagglomeration	State Firm Birth Employment Coagglomeration
	(1)	(2)	(3)	(4)
Labor Correlation	0.180	0.106	0.082	0.077
	(0.011)	(0.012)	(0.012)	(0.012)
Input-Output	0.205	0.167	0.130	0.112
	(0.011)	(0.011)	(0.012)	(0.012)
Technology Flows Scherer R&D	0.180	0.148	0.107	0.089
	(0.011)	(0.012)	(0.012)	(0.012)
Technology Flows Patent Citations	0.081	0.100	0.085	0.068
	(0.012)	(0.012)	(0.012)	(0.012)

Notes: Each cell reports a separate regression of pairwise EG Coagglomeration Index on a determinant of industrial co-location. Coagglomeration measures are calculated from the 1987 Census of Manufacturers using the employments listed in the column headers. All pairwise combinations of manufacturing SIC3 industries are included, except those listed in the text, for 7381 observations. Labor Correlation indices are calculated from the BLS National Industry-Occupation Employment Matrix for 1987. Input-Output relationships are calculated from the BEA Benchmark Input-Output Matrix for 1987. Technology Flows are calculated from the Scherer (1984) R&D tables for the 1970s and from the NBER Patent Citation Database for 1975-1997. Maximum values for the pairwise combination are employed. All variables are transformed to (mean 0, standard deviation 1) for interpretation. Regressions are unweighted. Standard errors are in parentheses.

Table 6: OLS Multivariate Specifications for 1987 Pairwise Coagglomeration

Dependent variable is EG Coaggl. Index calculated with state total emp.	Pairwise Maximum Regressions			Pairwise Mean Regressions		
	Base Estimation	Separate Input & Output	Exclude Pairs in Same SIC2	Base Estimation	Separate Input & Output	Exclude Pairs in Same SIC2
	(1)	(2)	(3)	(4)	(5)	(6)
Labor Correlation	0.146	0.142	0.110	0.135	0.134	0.108
	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.012)
Input-Output	0.149		0.108	0.185		0.117
	(0.012)		(0.012)	(0.012)		(0.012)
Input		0.109			0.116	
		(0.014)			(0.014)	
Output		0.095			0.098	
		(0.013)			(0.013)	
Technology Flows Scherer R&D	0.112	0.096	0.050	0.125	0.121	0.032
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Observations	7381	7381	7000	7381	7381	7000

Notes: See Table 5. Regression of pairwise EG Coagglomeration Index on determinants of industrial co-location. Coagglomeration measures are calculated from the 1987 Census of Manufacturers using state total employments for each industry. Columns 3 and 6 exclude SIC3 pairwise combinations within the same SIC2. Appendix Table 2 repeats Column 1 with alternative coagglomeration metrics.

Table 7: OLS and IV Multivariate Specifications for 1987 Pairwise Coagglomeration

	Dependent Variable is EG Coagglomeration Index							
	State Total Empl. Coaggl. OLS	State Total Empl. Coaggl. IV	PMSA Total Empl. Coaggl. OLS	PMSA Total Empl. Coaggl. IV	County Total Empl. Coaggl. OLS	County Total Empl. Coaggl. IV	State Birth Empl. Coaggl. OLS	State Birth Empl. Coaggl. IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A. OLS and IV Multivariate Specifications with Labor and Input-Output Only							
Labor Correlation	0.108	0.140	0.033	0.151	0.029	0.047	0.042	0.187
	(0.012)	(0.056)	(0.012)	(0.057)	(0.012)	(0.056)	(0.012)	(0.056)
Input-Output	0.121	0.149	0.096	0.078	0.075	0.103	0.051	0.152
	(0.012)	(0.045)	(0.012)	(0.045)	(0.012)	(0.045)	(0.012)	(0.043)
	B. OLS and IV Multivariate Specifications with Scherer Technology Metric							
Labor Correlation	0.110	0.120	0.035	0.136	0.030	0.028	0.042	0.254
	(0.012)	(0.059)	(0.012)	(0.060)	(0.012)	(0.060)	(0.012)	(0.066)
Input-Output	0.108	0.095	0.085	0.039	0.068	0.051	0.047	0.341
	(0.012)	(0.121)	(0.012)	(0.123)	(0.012)	(0.123)	(0.012)	(0.136)
Technology Flows	0.050	0.104	0.041	0.076	0.026	0.099	0.015	-0.359
Scherer R&D	(0.012)	(0.181)	(0.012)	(0.183)	(0.012)	(0.183)	(0.012)	(0.204)

Notes: See Tables 5 and 6. OLS and IV Regression of pairwise EG Coagglomeration Index on determinants of industrial co-location. Appendix Table 3 documents the first-stage coefficients.

App. Table 1: Inter-Industry 1987 Pairwise Coagglomeration Averages

	20	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
20	0.002																		
22	-0.003	0.102																	
23	0.000	0.021	0.012																
24	0.003	0.012	0.003	0.013															
25	-0.001	0.016	0.002	0.002	0.000														
26	0.001	0.012	0.004	0.006	-0.001	0.005													
27	0.001	-0.001	0.001	-0.003	-0.004	0.000	0.004												
28	0.001	0.004	0.001	0.000	-0.002	0.002	0.001	0.007											
29	0.004	-0.018	-0.003	0.000	-0.006	-0.002	0.001	0.008	0.013										
30	-0.001	0.003	-0.003	-0.001	0.001	0.000	-0.001	0.001	-0.001	0.002									
31	0.000	-0.005	0.006	-0.001	-0.003	0.005	0.006	0.001	-0.003	-0.003	0.019								
32	0.001	0.001	0.002	0.000	-0.002	0.000	0.000	0.003	0.006	0.001	-0.001	0.003							
33	-0.001	-0.012	-0.006	-0.002	-0.004	0.001	-0.001	0.001	0.002	0.004	-0.003	0.004	0.010						
34	-0.001	-0.014	-0.007	-0.004	-0.002	-0.002	0.000	-0.001	0.001	0.002	-0.003	0.000	0.005	0.004					
35	0.000	-0.011	-0.006	-0.003	-0.002	-0.001	0.001	-0.001	0.000	0.001	-0.001	-0.001	0.003	0.004	0.001				
36	0.001	-0.007	-0.001	-0.002	0.000	-0.003	0.001	-0.002	0.000	0.000	0.002	-0.001	-0.001	0.000	0.000	0.000			
37	-0.001	-0.017	-0.008	-0.001	0.001	-0.004	-0.002	-0.004	0.000	-0.002	-0.008	-0.002	0.004	0.004	0.001	-0.001	-0.004		
38	-0.002	-0.010	0.005	-0.005	-0.003	-0.002	0.006	-0.003	-0.005	-0.005	0.009	-0.003	-0.005	-0.002	0.000	0.002	-0.004	0.008	
39	-0.001	-0.007	0.005	-0.004	-0.004	0.000	0.005	-0.001	-0.003	-0.003	0.010	-0.002	-0.002	-0.001	0.000	0.003	-0.006	0.012	0.014

Notes: Table entries are the weighted-average pairwise SIC3 coagglomerations within the pairwise SIC2 cell. EG Coagglomeration Index measured through state total employments for each industry.

App. Table 2: OLS Multivariate Specifications for 1987 Pairwise Coagglomeration

	Dependent Variable is EG Coagglomeration Index			
	State Total Employment Coagglomeration	PMSA Total Employment Coagglomeration	County Total Employment Coagglomeration	State Firm Birth Employment Coagglomeration
	(1)	(2)	(3)	(4)
Labor Correlation	0.146	0.078	0.060	0.060
	(0.011)	(0.012)	(0.012)	(0.012)
Input-Output	0.149	0.125	0.101	0.086
	(0.012)	(0.012)	(0.012)	(0.013)
Technology Flows	0.112	0.098	0.067	0.054
Scherer R&D	(0.012)	(0.012)	(0.012)	(0.012)

Notes: See Table 6. Column 1 repeats the first column of Table 6 with coagglomeration measured through state total employments for each industry. Columns 2-4 substitute alternative metrics of coagglomeration.

App. Table 3: Univariate and Multivariate First-Stage Specifications for UK IV of Determinants of Co-Locations

Dependent variable is the explanatory regressor listed in the column header	Univariate First-Stage Specifications			Multivariate First-Stages without Technology		Multivariate First-Stages with Technology		
	Labor Correlation	Input-Output	Technology Scherer	Labor Correlation	Input-Output	Labor Correlation	Input-Output	Technology Scherer
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
UK Labor IV	0.278			**0.262**	0.095	**0.247**	0.041	0.048
	(0.011)			**(0.012)**	(0.011)	**(0.012)**	(0.012)	(0.012)
UK Input-Output IV		0.345		0.070	**0.323**	0.064	**0.300**	0.159
		(0.011)		(0.012)	**(0.011)**	(0.012)	**(0.011)**	(0.012)
UK Technology Flows IV			0.237			0.054	0.202	**0.195**
			(0.012)			(0.012)	(0.012)	**(0.012)**

Notes: First-stage regressions of U.S. pairwise determinants of industrial co-location on similarly constructed U.K. instruments. All pairwise combinations of manufacturing SIC3 industries are included, except those listed in the text, for 7000 observations. The decline in observations from Table 5 is due to the exclusion of pairwise combinations within the same SIC2. Variable constructions are described in the data appendix. Maximum values for the pairwise combination are employed. All variables are transformed to (mean 0, standard deviation 1) for interpretation. Regressions are unweighted. Standard errors are in parentheses.

Data Appendix to Ellison, Glaeser and Kerr (2007)

Glenn Ellison
MIT

Edward Glaeser
Harvard University

William Kerr*
Harvard Business School

April 2007

Abstract

This appendix provides additional details about the data employed in Ellison, Glaeser and Kerr (2007).

*Comments are appreciated and can be sent to wkerr@hbs.edu. We are grateful to Jim Davis, Alex Bryson, Keith Maskus, and Debbie Smeaton for data assistance. The research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Boston Census Research Data Center (BRDC). Support for this research from NSF grant (ITR-0427889) is gratefully acknowledged. Research results and conclusions expressed are our own and do not necessarily reflect the views of the Census Bureau or NSF. This paper has been screened to insure that no confidential data are revealed.

1 Overview

This note provides more detail about the coagglomeration dataset developed for Ellison, Glaeser, and Kerr (2007). It first outlines the Census Bureau data employed for the construction of the Ellison and Glaeser (1997) coagglomeration metrics. It then outlines the development of the U.S.-based metrics of Marshall's agglomeration theories. The note closes with the design of the U.K.-based instrumental variables.

2 US Coagglomeration Metrics

Our estimates of industrial coagglomeration patterns are developed through confidential data housed by the U.S. Census Bureau. The Census of Manufacturers is conducted every five years (those ending with 2 or 7) and surveys the universe of manufacturing plants operating in the U.S. With appropriate clearance, researchers can analyze the microdata of these Censuses, which is essential for estimating coagglomeration levels of detailed industries as public reports suppress values that risk disclosing the operating details of individual firms. Moreover, as the microdata for plants can be linked longitudinally across Censuses, we can compare the coagglomeration of existing establishment with that of new entrants. We focus on the six Census of Manufacturers conducted from 1972 to 1997, providing approximately 300k establishment observations employing 17 million workers in each census year.[1]

Following Proposition 1, the pairwise coagglomeration between industry pair 1 and 2 can be analyzed with the simple formula

$$\gamma^c = \frac{\sum_{m=1}^{M}(s_{m1} - x_m)(s_{m2} - x_m)}{1 - \sum_{m=1}^{M} x_m^2},$$

where M indexes geographic regions. $s_{1i}, s_{2i}, \ldots, s_{Mi}$ are the shares of industry i's employment contained in each of these areas. $x_1, x_2, \ldots, x_M$ are some other measure of the size of these areas, such as each area's share of population or aggregate employment. We model x_m in this paper through the mean employment share in the region across manufacturing industries.[2]

We operationalize this coagglomeration measure among manufacturing industries using the three-digit level of the 1987 Standard Industrial Classification (SIC3). Our primary measure of the economic activity in an industry j in a given geographic area m is the total employment in all manufacturing establishments excluding auxiliary units. The s_{mj} measure is then the share of the industry j's employment in region m. Throughout the paper, we simultaneously report coagglomeration metrics calculated at the state-level (including the District of Columbia), the

[1] The 2002 Census of Manufacturers recently became available. It employs the NAICS industry codes, however, that make it difficult to compare to earlier years. Ellison et al. (2006) discusses the calculation of coagglomeration measures under the NAICS framework. Dunne *et al.* (1989), McGuckin and Peck (1992), Davis *et al.* (1996), Kerr and Nanda (2006), and Autor *et al.* (2007) provide detailed accounts of the Census Bureau data.

[2] While the Ellison and Glaeser (1997) formula allows for the x_m to vary across industries, the equivalency formula in Proposition 1 requires that they be the same.

PMSA-level, and the county-level. These variants only adjust the M demarcations on which s and x are calculated.

The regression sample includes the pairwise combinations of 122 SIC3 industries. Tobacco (210s), Fur (237), and Search and Navigation Equipment (381) are excluded throughout the paper due to major industry reclassifications at the plant-level in the Census of Manufacturers that are difficult to interpret. In the empirical estimations in Section 5, the remainder of Apparel (230s), a portion of Printing and Publishing (277-279), and Secondary Non-Ferrous Metals (334) are also excluded due to either an inability to construct appropriate Marshallian explanatory matrices or outlier concerns in the explanatory data. Finally, we exclude same-industry pairs for a total of 7381 unique pairwise industry combinations per Census of Manufacturers.

3 US Coagglomeration Determinants

We use industry attributes to design coagglomeration-oriented metrics that mirror each of Marshall's three theories of industry agglomeration: (1) labor market pooling, (2) proximity to input suppliers or industrial customers to save on transportation costs, and (3) intellectual or technology spillovers. Data Appendix Table 1 documents the summary statistics for these metrics, and Data Appendix Table 2 lists the extreme pairwise values. A condensed version of this section appears in the main text.

3.1 Labor market pooling

One of Marshall's theories of industrial location is that firms locate near one another to shield workers from the vicissitudes of firm-specific shocks. Workers are willing to accept lower wages in locations where other firms stand by ready to hire them (see Diamond and Simon (1990) for evidence and Krugman (1991) for a formalization). Rotemberg and Saloner (2000) present an alternative theory in which workers gain not because of insurance from shocks, but because multiple firms protect workers against ex post appropriation of investments in human capital. Both theories predict that plants that use the same type of workers will locate near one another.

To test the labor pooling theory, we construct a metric of the similarity in the occupational labor requirements for pairwise industries. We build from the 1987 National Industry-Occupation Employment Matrix (NIOEM) published by the Bureau of Labor Statistics (BLS). The NIOEM provides industry-level employments (at the national level) in 277 occupations. We convert the occupational employment counts into occupational percentages for each industry and map the BLS industries to the SIC3 framework. 52 of the 185 broadly-defined BLS industries are within manufacturing. Each SIC3 industry is assumed to possess the same occupational composition of employment as that of the NIOEM industry to which it belongs.[3]

[3]The BLS has recently released a 1983-1998 longitudinal version of the NIOEM. Users should note that the occupations employed in the standardized version differ slightly from those in the 1987 NIOEM we employ. Metrics calculated from the new panel are very close to those used in this paper.

Our metric of labor similarity, $LaborCorrelation_{ij}$, is a vector correlation of occupational percentages between two industries. $LaborCorrelation_{ij}$ averages 0.47 across the pairs of manufacturing industries, with a range of -0.05 to 1.00. The least correlated industry pair is Logging (241) and Aircrafts and Parts (372) at -0.046. The perfect correlation maximum value reflects that some NIOEM industries map to two or more SIC3 industries; the empirical specifications in the main paper account for this multiplicity. The most correlated industry pair, not by construction, is Motor Vehicles and Equipment (371) and Motorcycles, Bicycles, and Parts (375) at 0.984. Finally, note that the labor pooling metrics are symmetrical for a pairwise industry combination i,j: $LaborCorrelation_{ij} = LaborCorrelation_{ji}$. This is not generally the case for the next two factors discussed, where directional flows are evident.

3.2 The presence of suppliers and customers

Marshall (1920) also argues that transportation costs should induce plants to locate close to their inputs, close to their customers, or most likely at some point optimally trading off distance between inputs and customers. To test this theory, we construct metrics of the importance of customer or supplier relationships for pairwise industries. We build our metrics from the 1987 Benchmark Input-Output Accounts published by the Bureau of Economic Analysis. The "Use of Commodities by Industries" table provides commodity-level make and use for flows for very detailed industries at the national level, which we aggregate to the SIC3 framework. While some commodities can partly be produced by other industries than the one associated with these commodities, we ignore this distinction and therefore interpret the numbers from the table as providing an estimate of how much of an industry's production is used as an input to other industries.

We define $Input_{i \leftarrow j}$ as the share of industry i's inputs that come from industry j, and $Output_{i \rightarrow j}$ as the share of industry i's outputs that go to industry j. These measures run from 0 (no input or output purchasing relationship exists) to 1 (full dependency on the paired industry). These shares are calculated relative to all input-output flows, including those to non-manufacturing industries or to final consumers.

The strongest relative customer or input dependency is Leather Tanning and Finishing's (311) purchases from Meat Products (201) at 0.39. The highest absolute customer dependency (with a relative share of 23%) is Misc. Plastics Products (308) purchases from Plastic Materials and Synthetics (282). The strongest relative output or supplier dependency is Public Building and Related Furniture's sales to Motor Vehicles and Equipment (371) at 82%. The highest absolute supplier dependency (with a relative share of 32%) is Plastic Materials and Synthetics (282) sales to Misc. Plastics Products (308). Approximately 70% of pairwise combinations have an input-output dependency less than 0.01%.

This construction results in four potential metrics for a pairwise industry i,j combination: $Input_{i \leftarrow j}$, $Input_{j \leftarrow i}$, $Output_{i \rightarrow j}$, and $Output_{j \rightarrow i}$. Unlike the labor pooling metrics, customer and

supplier flows are not symmetrical ($Input_{i \leftarrow j} \neq Input_{j \rightarrow i}$). Moreover, the flows between the plastics industries highlights how differences in industry size and the importance of flows to or from non-manufacturing industries and final consumers result in asymmetries between pairwise customer and supplier dependencies ($Input_{i \leftarrow j} \neq Output_{j \rightarrow i}$). To operationalize these metrics for the pairwise coagglomeration regressions, we take either the maximum or the mean of the *Input* and *Output* relationships for the pairwise i,j combination. We also examine jointly the input-output role by calculating means and maximums across all four metrics.

3.3 Intellectual or technology spillovers

Firms may also locate where they are likely to learn from other firms. This learning can take the form of workers learning skills from one another (as Marshall argued) or industrial innovators copying each other (as Saxenian (1994) reports for Silicon Valley). Firms will group near one another either because of the gains from continued presence or because the idea leading to the opening of a new establishment came from an existing concentration of employment in nearby plants. To test this third theory, we develop two metrics of intellectual spillovers that focus specifically on the sourcing of technological innovations. The primary metric is derived from technology flow matrices developed by Scherer (1984); the second metric is derived from patent citations.

Of Marshall's three theories, intellectual spillovers are the most difficult to quantify and to assess empirically. We first note that our metrics focus only technology spillovers. Other intellectual or information spillovers may exist between industries that are not captured by our design, although technology sourcing is a very important form of knowledge sharing for the manufacturing sector. Second, the discussion below highlights that technology flows are not mutually exclusive to Marshall's first two theories. Technologies embodied in products and machinery can be transferred directly through input-output exchanges. Likewise, industries that share similar labor pools may also be industries between which there is a greater possibility for intellectual spillovers. Our empirical exercises attempt to isolate technology spillovers by joint testing with these other two factors, but it is important to note that intellectual spillovers do occur within these channels too.

3.3.1 Scherer Technology Flows

Scherer (1984) develops a technology flow matrix that estimates the extent to which R&D activity in one industry flows out to benefit another industry. This technology transfer occurs either through a supplier-customer relationship between these two industries or through the likelihood that patented inventions obtained in one industry will find applications in the other industry. We develop two metrics, $TechIn_{i \leftarrow j}$ and $TechOut_{i \rightarrow j}$, for these technology flows that mirror *Input* and *Output* described above. These dependencies are again directional in nature and are calculated relative to total technology flows that include non-manufacturing industries.

The strongest relative technology flows are associated with Plastic Materials and Synthetics (282) and its relationships to Misc. Plastics Products (308), Tires and Inner Tubes (301) and Industrial Organic Chemicals (286).

The raw technology flows are taken from Table 20.1 of Scherer (1984). Each entry in that table is a dollar amount of 1974 R&D spending in a given industry that is estimated to flow out to benefit another industry. We converted the 38 manufacturing industries reported by Scherer (1984) to the SIC3 framework by apportioning entries through total value of shipments (obtained from the 1987 Census of Manufactures). For instance, if T^*_{mn} is the entry in Scherer's table corresponding to the dollar flow of benefits from industry m to industry n, and j (resp., i) is a three-digit industry that is part of industry group m (resp., n) and accounts for a fraction w_j (resp, w_i) of all shipments in that industry group, then $T_{ji} = w_i w_j \, T^*_{mn}$.

3.3.2 Patent Citation Flows

The NBER Patent Data File was originally compiled by Hall *et al.* (2001). This dataset offers detailed records for all patents granted by the United States Patent and Trademark Office (USPTO) from January 1975 to December 1999. Each patent record provides information about the invention (e.g., technology classification, citations of prior art) and the inventors submitting the application (e.g., name, city). Patent citation patterns can be informative about technology diffusion and knowledge exchanges. Griliches (1990) and Jaffe *et al.* (2000) further discuss employing patent citations in this context.

We construct our second knowledge spillovers metric through the patent citations. We restrict the citations data to be citations where both the citing and cited patents are a) applied for after 1975 and b) filed within the U.S. This sample includes 4,467,625 citations. These citations are first collapsed into a citation matrix using the USPTO technology categories, over 400 in number. Combining the work of Johnson (1999), Silverman (1999) and Kerr (forthcoming), concordances are developed between the USPTO classification scheme and SIC3 industries (a probabilistic mapping).

The resulting metrics estimate the extent to which technologies associated with industry i cite technologies associated with industry j, and vice versa. These $PatIn_{ij}$ and $PatOut_{ij}$ are normalized by total citations for the industries. In practice, there is little directional difference between $PatIn_{ij}$ and $PatOut_{ij}$ due to the extensive number of citations within a single technology field, in which case the probabilistic citing and cited industry distributions are the same. These patent-based metrics have the advantage of covering the 1975-2000 period, but inventor-to-inventor communication patterns represent a subset of the technology flows Scherer (1984) attempts to encompass.

We primarily use the patent citations data to construct the U.K. instrument for technology flows in the U.S. As further noted below, using the same technology-to-industry concordance structurally relates the U.S. and U.K. citation matrices. Thus, it is better to use the U.K.

citation matrices with the Scherer (1984) technology flows.

4 UK Instrumental Variables

The above U.S. metrics are useful for examining correlations in the data regarding the determinants of coagglomeration. A clear interpretation of the results, however, is limited by concerns of reverse causality. Take our observed importance of customer relationships as an example. Our exposition suggests firms are choosing their geographic locations to be near their customers in order to minimize transportation costs. An alternative explanation of the findings, however, is that these firm locations are determined by other factors (e.g., historical accidents). After these locations are determined, firms choose to sell to nearby industries. These sales are subsequently reflected in the BEA Benchmark Input-Output Accounts, leading to our observed correlations.

To recover a causal assessment, we develop instruments for our explanatory variables from equivalent data in the U.K. Their sources and construction mirror those described for the U.S. and are described below. The identifying assumption is that the observed input-output, labor pooling, and technology sourcing relationships among industries in the U.K. are correlated with the natural inter-industry dependencies but are orthogonal to any endogenous industry interdependencies present in the U.S. data that arise from reverse causality. The instruments may have applications in other contexts too.

4.1 Labor market pooling

The U.K. does not publish a detailed equivalent of the BLS' National Industry-Occupation Employment Matrix. To construct a similar matrix for the U.K., we pooled six years of the U.K. Labour Force Survey (LFS), akin to the U.S. Current Population Survey. We then developed matrices of the occupation-by-industry distribution of currently employed workers by summing over the survey. The included surveys are March-May 2001, June-August 2002, September-November 2003, December 2004-February 2005, and April-June 2006. This pooled dataset contains 224,528 employed workers out of 520,952 respondents; 42,948 work in manufacturing. We maintained the occupation codes Soc2km (353 classifications) and Sc2kmmn (84 classifications) at their detailed level for estimating labor similarities. We mapped the industry code Indm92m (461 classifications, 265 in manufacturing) into the SIC3 system.[4]

4.2 The presence of suppliers and customers

The input-output matrices are taken from Maskus *et al.* (1994) and Maskus and Webster (1995). These researchers began with the 1989 Input-Output Balance for the United Kingdom,

[4]We employ a later period than our typical 1987 date to increase the available LFS sample size and questionnaire detail. The period starts after occupation classifications changed in 2001. The staggered surveys avoid double counting as one-fifth of the LFS' respondents rotate out each quarter. From 2005, the data collection periods shift from (mar-may, jun-aug, sep-nov, dec-feb) to (jan-mar, apr-jun, jul-sep, oct-dec).

published by the Central Statistical Office, London, in 1992. The original table contained 102 sectors; Maskus *et al.* (1994) aggregated the table into 80 sectors that formed the least common denominator with the U.S. tables they were also employing. These tables again include flows out of the manufacturing sector that are used for normalizations. We mapped the 80 Maskus *et al.* sectors that corresponded to the SIC3 system. The empirical analysis in the text accounts for this multiplicity.

4.3 Intellectual or technology spillovers

The U.K. technology flows matrices are calculated through the NBER patent citations data. The construction mirrors the U.S. citation development documented above, except that we limit the raw sample to those citations where both the citing and cited patents are filed from the U.K. 28,134 citations from 1975-1999 are used for these metrics. It is important to note that U.K. citations are converted from the USPC classification system to the SIC3 framework using the same technology concordances as used for converting the U.S. data. Using the U.K. citations as an instrument for spillovers measured through the U.S. citations will deliver overstated first-stages by construction. The U.K. citations are better suited as an instruments for the Scherer (1984) technology flow matrices.[5]

[5]The core element of the USPC-to-industry concordances comes from Canadian data that jointly classified patents into technologies and industries. Thus combining the U.K. citations with the industry concordances is still excludable for an instrument of U.S. technology flows. By themselves, the U.K. citations can serve as instruments for U.S. citations when using just the USPC codes; it is the industry conversion that introduces the common structural forms.

References

[1] Autor, David, Kerr, William and Kugler, Adriana (forthcoming): "Do Employment Protections Reduce Productivity? Evidence from U.S. States," *The Economic Journal*.

[2] Davis, Steven, Haltiwanger, John and Schuh, Scott (1996): *Job Creation and Destruction*. Cambridge, MA: MIT Press.

[3] Diamond, Charles and Simon, Curtis (1990): "Industrial Specialization and the Returns to Labor," *Journal of Labor Economics*, 8, 175–201.

[4] Dumais, Guy, Ellison, Glenn and Glaeser, Edward L. (2002): "Geographic Concentration as a Dynamic Process," *Review of Economics and Statistics*, 84, 193–204.

[5] Dunne, Timothy, Roberts, Mark and Samuelson, Larry (1989): "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries," *The RAND Journal of Economics*, 19, 495–515.

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

[7] Ellison, Glenn, Glaeser, Edward L., and Kerr, William (2007): "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," Working Paper.

[8] Ellison, Glenn, Glaeser, Edward L., and Kerr, William (2006): "The Impact of the SIC-NAICS Conversion on Industrial Organization Metrics: Evidence Building from Establishment Data," Census Bureau Technical Paper.

[9] Griliches, Zvi (1990): "Patent Statistics as Economic Indicators: A Survey," *Journal of Economic Literature,* 28, 1661–1707.

[10] Hall, Bronwyn, Jaffe, Adam and Trajtenberg, Manuel (2001): "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498.

[11] Jaffe, Adam, Trajtenberg, Manuel and Fogarty, Michael (2000): "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," *Papers and Proceedings, American Economic Review*, 90, 215–218.

[12] Johnson, Daniel (1999): "150 Years of American Invention: Methodology and a First Geographic Application," Wellesley College Economics Working Paper 99-01.

[13] Kerr, William (forthcoming): "Ethnic Scientific Communities and International Technology Diffusion," *The Review of Economics and Statistics*.

[14] Kerr, William and Nanda, Ramana (2006): "Banking Deregulations, Financing Constraints, and Entrepreneurship," HBS Working Paper 07-033.

[15] Krugman, Paul (1991): *Geography and Trade*. Cambridge, MA: MIT Press.

[16] Marshall, Alfred (1920): *Principles of Economics*. London, UK: MacMillan and Co.

[17] Maskus, Keith, Sveikauskas, C. and Webster, Allan (1994): "The Composition of the Human Capital Stock and Its Relation to International Trade: Evidence from the U.S. and Britain," *Weltwirtschaftliches Archiv*, 1994, Band 130, Heft 1.

[18] Maskus, Keith and Webster, Allan (1995): "Factor Specialization in U.S. and U.K. Trade: Simple Departures from the Factor-Content Theory," *Swiss Journal of Economics and Statistics*, 1.

[19] McGuckin, Robert and Peck, Suzanne (1992): "Manufacturing Establishments Reclassified into New Industries: The Effect of Survey Design Rules," Center for Economic Studies (U.S. Bureau of the Census).

[20] Rotemberg, Julio and Saloner, Garth (2000): "Competition and Human Capital Accumulation: A Theory of Interregional Specialization and Trade," *Regional Science and Urban Economics*, 30, 373–404.

[21] Saxenian, AnnaLee (1994): *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

[22] Scherer, Frederic M. (1984): "Using Linked Patent Data and R&D Data to Measure Technology Flows," in Griliches, Zvi (ed.) *R & D, Patents and Productivity*. Chicago, IL: The University of Chicago Press.

[23] Silverman, Brian (1999): "Technological Resources and the Direction of Corporate Diversification: Toward an Integration of the Resource-Based View and Transaction Cost Economics," *Management Science*, 45, 1109–1124.

Data App. Table 1: 1987 Descriptive Statistics

	Mean	Standard Deviation	Minimum	Maximum
	A. Pairwise Coagglomeration Measures			
State Total Empl. Pairwise Coaggl.	0.000	0.013	-0.065	0.207
PMSA Total Empl. Pairwise Coaggl.	0.000	0.006	-0.025	0.119
County Total Empl. Pairwise Coaggl.	0.000	0.003	-0.018	0.080
State Birth Empl. Pairwise Coaggl.	0.000	0.015	-0.082	0.259
	B. Pairwise Labor Similarities Index			
Labor Correlation	0.470	0.226	-0.046	1.000
	C. Pairwise Input-Output Relationship Indices			
Input-Output Maximum	0.007	0.029	0.000	0.823
Input-Output Mean	0.002	0.010	0.000	0.240
Input Maximum	0.005	0.019	0.000	0.392
Input Mean	0.002	0.010	0.000	0.196
Output Maximum	0.005	0.026	0.000	0.823
Output Mean	0.002	0.013	0.000	0.411
	D. Pairwise Technology Relationship Indices			
Scherer R&D Tech Maximum	0.005	0.026	0.000	0.625
Scherer R&D Tech Mean	0.002	0.010	0.000	0.263
Patent Citation Tech Maximum	0.015	0.025	0.000	0.400
Patent Citation Tech Mean	0.007	0.014	0.000	0.203

Notes: Descriptive statistics for 1987. All pairwise combinations of manufacturing SIC3 industries are included, except those listed in the text, for 7381 observations. Coagglomeration measures are calculated from the 1987 Census of Manufacturers. Labor Correlation indices are calculated from the BLS National Industry-Occupation Employment Matrix for 1987. Input-Output relationships are calculated from the BEA Benchmark Input-Output Matrix for 1987. Technology Flows are calculated from the Scherer (1984) R&D tables for the 1970s and from the NBER Patent Citation Database for 1975-1997.

Data App. Table 2A: Highest Labor Correlation Metrics

Industry 1	Industry 2	Labor Cor.
Motor Vehicles and Equipment (371)	Railroad Equipment (374)	0.984
Motor Vehicles and Equipment (371)	Motorcycles, Bicycles, and Parts (375)	0.984
Motor Vehicles and Equipment (371)	Miscellaneous Transportation Equipment (379)	0.984
Musical Instruments (393)	Toys and Sporting Goods (394)	0.979
Toys and Sporting Goods (394)	Pens, Pencils, Office & Art Suppliers (395)	0.979

Data App. Table 2B: Lowest Labor Correlation Metrics

Industry 1	Industry 2	Labor Cor.
Logging (241)	Aircrafts and Parts (372)	-0.046
Logging (241)	Engines and Turbines (351)	-0.029
Logging (241)	Motor Vehicles and Equipment (371)	-0.029
Logging (241)	Guided Missiles, Space Vehicles, Parts (376)	-0.029
Logging (241)	Metalworking Machinery (354)	-0.021

Data App. Table 2C: Highest Relative Customer Dependencies Metrics

Using Industry	Source Industry	Input Vol.	Input Share
Leather Tanning and Finishing (311)	Meat Products (201)	872	0.392
Sawmills and Planing Mills (242)	Logging (241)	6811	0.360
Leather Gloves and Mittens (315)	Leather Tanning and Finishing (311)	58	0.345
Yarn and Thread Mills (228)	Plastics Materials and Synthetics (282)	2154	0.309
Wood Containers (244)	Sawmills and Planing Mills (242)	548	0.271

Data App. Table 2D: Highest Absolute Customer Dependencies Metrics

Using Industry	Source Industry	Input Vol.	Input Share
Misc. Plastics Products (308)	Plastics Materials and Synthetics (282)	13,999	0.229
Motor Vehicles and Equipment (371)	Metal Forgings and Stampings (346)	11,378	0.055
Plastics Materials and Synthetics (282)	Industrial Organic Chemicals (286)	9903	0.243
Fabricated Structural Metal Products (344)	Blast Furnace and Basic Steel Products (331)	7607	0.196
Metal Forgings and Stampings (346)	Blast Furnace and Basic Steel Products (331)	7011	0.249

Data App. Table 2E: Highest Relative Supplier Dependencies Metrics

Source Industry	Using Industry	Output Vol.	Output Share
Public Building and Related Furniture (253)	Motor Vehicles and Equipment (371)	1681	0.823
Cement, Hydraulic (324)	Concrete, Gypsum, and Plaster Products (327)	3380	0.819
Primary Nonferrous Metals (333)	Nonferrous Rolling and Drawing (335)	5750	0.504
Metal Cans and Shipping Containers (341)	Beverages (208)	5768	0.491
Logging (241)	Sawmills and Planing Mills (242)	6811	0.440

Data App. Table 2F: Highest Absolute Supplier Dependencies Metrics

Source Industry	Using Industry	Output Vol.	Output Share
Plastics Materials and Synthetics (282)	Misc. Plastics Products (308)	13,999	0.322
Metal Forgings and Stampings (346)	Motor Vehicles and Equipment (371)	11,378	0.401
Industrial Organic Chemicals (286)	Plastics Materials and Synthetics (282)	9903	0.179
Blast Furnace and Basic Steel Products (331)	Fabricated Structural Metal Products (344)	7607	0.153
Blast Furnace and Basic Steel Products (331)	Metal Forgings and Stampings (346)	7011	0.141

Data App. Table 2G: Highest Relative Technology Input Dependencies Metrics

Using Industry	Source Industry	Input Vol.	Input Share
Misc. Plastics Products (308)	Plastics Materials and Synthetics (282)	104	0.217
Rubber and Plastics Footwear (302)	Plastics Materials and Synthetics (282)	1	0.200
Tires and Inner Tubes (301)	Plastics Materials and Synthetics (282)	48	0.165
Fabricated Rubber Products (306)	Plastics Materials and Synthetics (282)	11	0.131
Hose, Belting, Gaskets, and Packing (305)	Plastics Materials and Synthetics (282)	5	0.116

Data App. Table 2H: Highest Absolute Technology Input Dependencies Metrics

Using Industry	Source Industry	Input Vol.	Input Share
Misc. Plastics Products (308)	Plastics Materials and Synthetics (282)	104	0.217
Tires and Inner Tubes (301)	Plastics Materials and Synthetics (282)	48	0.165
Plastics Materials and Synthetics (282)	Industrial Organic Chemicals (286)	24	0.040
Aircrafts and Parts (372)	Computers and Office Equipment (357)	21	0.039
Petroleum Refining (291)	Computers and Office Equipment (357)	19	0.043

Data App. Table 2I: Highest Relative Technology Supplier Dependencies Metrics

Source Industry	Using Industry	Output Vol.	Output Share
Plastics Materials and Synthetics (282)	Misc. Plastics Products (308)	104	0.172
Textile Finishing (226)	Misc. Plastics Products (308)	2	0.146
Ordnance and Accessories (348)	Guided Missiles, Space Vehicles, Parts (376)	3	0.133
Broadwoven Mills, Fiber (222)	Misc. Plastics Products (308)	2	0.086
Industrial Organic Chemicals (286)	Plastics Materials and Synthetics (282)	24	0.081

Data App. Table 2J: Highest Absolute Technology Supplier Dependencies Metrics

Source Industry	Using Industry	Output Vol.	Output Share
Plastics Materials and Synthetics (282)	Misc. Plastics Products (308)	104	0.172
Plastics Materials and Synthetics (282)	Tires and Inner Tubes (301)	48	0.080
Industrial Organic Chemicals (286)	Plastics Materials and Synthetics (282)	24	0.081
Computers and Office Equipment (357)	Aircrafts and Parts (372)	21	0.018
Computers and Office Equipment (357)	Petroleum Refining (291)	19	0.017