**APPENDICES to**

**WORKING PAPER NO. 04-16/R**

# MATCHING AND LEARNING IN CITIES: URBAN DENSITY AND THE RATE OF INVENTION *

Gerald Carlino

Satyajit Chatterjee

Robert Hunt

Federal Reserve Bank of Philadelphia

First Draft: October 2004
This Draft: April 2005

* The working paper was previously circulated under the title "Matching and Learning in Cities: Evidence from Patent Data."

We thank Marcus Berliant and Leonard Nakamura for many useful discussions, but we alone are responsible for any remaining errors.

The views expressed here are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

# Appendix A: A Model of Matching and Innovation

## Introduction

In our paper, we argue that the inventive output of cities may be explained in part by the nature of worker interactions within firms. A formalization of this intuition can be found in the labor market search model of Berliant, Reed, and Wang (2004), hereafter BRW. In this appendix, we develop a more general version of their model and derive a number of comparative static results. We extend their model in several ways. We assume a more general production function for inventions. We allow for consumption amenities that influence the equilibrium size of cities. Finally, we characterize the effect of an exogenous increase in land area on city population and inventive output.

## The Model

Time is continuous. Workers live forever and discount consumption at the rate $r$. A measure $N$ of these workers decides to live in the city because the increase in their income, plus utility derived from the available local public goods exceeds the utility lost to the effects of congestion (including higher rents and taxes). Once located in the city, workers meet randomly and decide whether to match in order to engage in R&D that results in inventions. If the agents agree to match, the match may continue indefinitely, but all matches are subject to a constant dissolution rate $\pi$. Workers consume equal shares of the resulting output and their resulting utility is linear in the output consumed.

The output generated in a match depends on the characteristics of the workers matched together. Workers are differentiated in terms of the variety of knowledge they possess. Agents' types are distributed uniformly along a circle of unit circumference. The most productive matches occur with an intermediate degree of heterogeneity. We focus on the properties of the unique symmetric stationary equilibrium of the game.

In steady state, a measure of agents $M \in [0, N]$ will already be matched and a corresponding measure $U = N - M$ will engage in search for a new partner. In a given match, workers produce inventions at the rate $A \cdot y(x)$. $A$ is a scalar that reflects the overall productivity of the city.[1] Match-specific output depends on the Euclidean distance $x$ between the type of an agent's ideal partner and location of his actual partner.

BRW assume $y(x)$ is a linear, decreasing function of $x$. Here, we assume that *expected* output $\tilde{y}(x)$ is continuous, differentiable, and strictly log concave. This implies that output falls more and more rapidly as we continue to increase $x$.[2] To simplify the analysis, we also assume that matches between identical agents are un-productive.[3] BRW make a comparable assumption, but they show the results are qualitatively the same when

[1] In our empirical specification, $A$ reflects the utilization of inputs other than labor plus any factors external to the firm.

[2] This assumption on production technology is sufficient to satisfy the conditions for assortative matching characterized in Shimer and Smith (2000).

[3] Let $\overline{d}$ denote the Euclidean distance from the agent's own type and his ideal partner. It is sufficient to assume there exists a value $\overline{x} \le \overline{d}$ such that $y(\overline{x}) = 0$.

this assumption is violated, or if, instead, it is assumed that output is maximized by matching among identical agents.

Unmatched workers engage in search for a new partner. The arrival rate of meetings between unmatched workers follows a Poisson process. In BRW, the *meeting hazard rate* is simply $\mu = \alpha \cdot U$.[4] The exogenous parameter $\alpha$ is a measure of the efficiency of the meeting technology, which in turn may depend on the time required to travel to meeting places, the depth of the labor market, or both.

Workers cannot recognize each other's type until after they meet, and this implies that not all meetings result in matches. The probability that a meeting results in a match will depend on the selectivity of workers, which in turn depends on the opportunity cost of declining to form a match in the current meeting. In a symmetric equilibrium, agents will agree to match with any agent whose type is sufficiently close to the ideal, i.e., where $x \le \delta$, where $\delta$ is endogenously determined. Agents choose the value of $\delta$ that equates the output from the *marginal* match with the flow value of rejecting the match and waiting for the next encounter: $Ay(\delta) = rV_U$. Let $\beta(\delta, U)$ denote the endogenous *match rate*, the product of the meeting rate $\mu$ times the probability that a meeting will result in a match.

**Equilibrium**

BRW prove the existence of a unique, symmetric, stationary equilibrium of the game for two situations: (1) a city of fixed population and (2) an "open" city where population is endogenously determined. We explore the latter equilibrium, taking into account our assumption about the properties of expected output. Note that uniqueness follows from the fact that agents are distributed uniformly over the type space.[5]

We begin by computing the steady-state value functions holding population constant and assuming that agents choose symmetric strategies. The flow value of being matched to an agent whose type is of distance *x* from one's own is:

$$r\hat{V}_m(x, U, \delta) = Ay(x) + \pi\left[V_U(U, \delta) - \hat{V}_m\right],$$

which consists of the instantaneous rate of inventions generated by the current match plus the capital loss associated with the cessation of the match, discounted by the arrival rate of the exogenous separation shock.

The agent does not know the exact type of his next partner, so $V_u$ is calculated as an expected value taking into account the agent's selectivity $\delta$. The next match does not

[4] Note that the meeting technology exhibits increasing returns. BRW argue that the findings of empirical studies are consistent with this assumption, but they also point out that their results do not depend on this assumption.

[5] The closed city version of BRW can be mapped to a special case of the model in Burdett and Wright (1998), where agents' payoffs from prospective matches are symmetric. Proposition 2 of that paper shows that if the distribution over types is log-concave, there is only one steady-state equilibrium. Proposition 4 shows this holds even where the matching function exhibits increasing returns. Uniqueness is preserved in the open city version of the model because congestion costs are monotonically increasing in population.

arrive immediately after the current one is dissolved, so $V_u$ also reflects the expected amount of time before meeting a suitable partner, and this depends on both $U$ and $\delta$. More explicitly:

(A.1) $$rV_U(U,\delta) = \beta(\delta,U)\left[V_m(U,\delta) - V_U(U,\delta)\right],$$

which is simply the discounted present value of the capital gain resulting from the next match. Note that $V_m$ is derived by integrating $\hat{V}_m$ over $x \in [0,\delta]$, holding constant both $U$ and $\delta$. Thus

(A.2) $$rV_m(U,\delta) = A\tilde{y}(\delta) + \pi\left[V_U(U,\delta) - V_m(U,\delta)\right],$$

where $\tilde{y}(\delta)$ is the expected value of output from matches when agents' selectivity is $\delta$:

$$\tilde{y}(\delta) = \int_0^\delta y(x)\,f(x)dx \Big/ \left[1 - F(\delta)\right] = \frac{1}{\delta}\cdot\int_0^\delta y(x)\,f(x)dx,$$

where $f(x)$ and $1 - F(\delta)$ are the pdf and cdf, respectively, of the distribution of worker locations. The equality with the right hand side of the expression follows from our assumption that workers are distributed uniformly over a circle of unit circumference.

Substituting for the value functions in (A.1) and (A.2), we find that

(A.3) $$rV_u(U,\delta) = \frac{\beta A\tilde{y}(\delta)}{r + \pi + \beta}.$$

**First Order Conditions.** Agents that are currently matched enjoy the proceeds of the match until it is dissolved. Unmatched agents are engaged in search and must decide whether to match with the un-matched agents they meet. They choose $\delta$ to maximize $rV_U(U,\delta)$. The associated first order condition is

(A.4) $$[r + \pi + \beta]\beta\frac{\partial\tilde{y}}{\partial\delta} + [r + \pi]\tilde{y}\frac{\partial\beta}{\partial\delta} = 0.$$

Holding constant $U$ and the selectivity of other agents, becoming less selective (raising $\delta$) increases the arrival rate of the next partner $(\partial\beta/\partial\delta > 0)$ but reduces the expected income generated in the next match $(\partial\tilde{y}/\partial\delta < 0)$.

Because agents' types are distributed uniformly over the unit circle and we assume that matches of identical agents are unproductive, $\beta(\delta,U) = 4\delta\mu(U)$. This allows us to rewrite the first order condition:

(A.5) $$\frac{\beta}{r + \pi + \beta} = \frac{y(\delta)}{\tilde{y}(\delta)} \equiv 1 - \eta,$$

where $y(\delta)$ represents the output generated by a marginal match and $\eta$ is the elasticity of *average* output with respect to increases in $\delta$.

**Lemma 1**: There exists an interior solution to the worker's optimization problem characterized by the equality in (A.5).

Proof: Existence follows from the fact that $\tilde{y}(\delta)$, $\tilde{y}(\delta)$ and $\beta$ are continuous and differentiable. If we take the limit as $\delta \to 0$, holding *U* constant, the right-hand side of (A.5) converges to 1 while the left-hand side converges to 0. Conversely, if we take the limit as $\delta \to \bar{\delta}$, the right-hand side of (A.5) converges to 0, while the left-hand side converges to a value that is strictly greater than zero. Finally, note that the second order condition is simply $[r+\pi+\beta]\beta y'(\delta)/\delta < 0$.■

The left hand side of (A.5) reflects agents' discount rate and the rate of turnover between the states of being matched and unmatched. The first order condition implies that when the match rate $\beta$ is relatively large, the ratio of marginal to average output is close to one. The following lemma shows that if $\tilde{y}(x)$ is log-concave, relatively large values of $\beta$ are associated with relatively small values of $\delta$. This establishes the intuition that, in equilibrium, a higher arrival rate of matches is associated with more selectivity (lower values of $\delta$) even though $\partial\beta/\partial\delta|_U > 0$.

**Lemma 2**: If $\tilde{y}(x)$ is strictly log-concave, $\eta$ is strictly increasing in *x*.

Proof: This follows directly from the definition of $\tilde{y}(\delta)$, which is the expected value of output, conditional on the level of selectivity $\delta$. Taking first and second derivatives of $Ln(\tilde{y}(\delta))$, we have

$$Ln(\tilde{y}(\delta))' = \frac{1}{\delta}\left[\frac{y(\delta)}{\tilde{y}(\delta)} - 1\right] = \frac{-\eta}{\delta} < 0; \qquad Ln(\tilde{y}(\delta))'' = \frac{1}{\delta^2}\left[\eta - (1-\eta)(\varepsilon-\eta)\right];$$

where $\varepsilon \equiv -\delta y'(\delta)/y(\delta)$ is the elasticity of *marginal* output with respect to agents' selectivity. Thus strict log concavity of expected output implies that

$$\text{(A.6)} \qquad \frac{\eta}{1-\eta} < \varepsilon - \eta.$$

A necessary characteristic of the production function, then, is that that marginal output falls proportionately more rapidly than average output as $\delta$ rises. In addition, the first order condition implies that $\eta \le 1$.[6] But these are sufficient conditions for our proof as

$$\frac{\partial\eta}{\partial\delta}\cdot\delta = \frac{\partial\left(1 - y(\delta)/\tilde{y}(\delta)\right)}{\partial\delta} = \frac{-y(\delta)}{\tilde{y}(\delta)}\cdot\left\{\frac{\delta y'(\delta)}{y(\delta)} - \left[\frac{y(\delta)}{\tilde{y}(\delta)} - 1\right]\right\} = (1-\eta)(\varepsilon-\eta). \blacksquare$$

In steady state the instantaneous measure of agents finding matches is just equal to the measure of agents whose matches are terminated: $\beta(U,\delta)U = \pi M$. Substituting for $\beta$ and $M$, we can solve a quadratic equation in *U*. One root is positive:

[6] We know there are values of $\delta$ where this condition holds because $\lim_{\delta\to 0}\tilde{y}(\delta) = y(0)$.

$U = \pi(\lambda - 1)/8\alpha\delta$ where $\lambda = \sqrt{1 + 16\alpha\delta N/\pi} \geq 1$. The corresponding match rate is $\beta(U,\delta) = \pi(\lambda - 1)/2$.

**Equilibrium Population Mass.** To close the model, we formalize the worker's participation constraint:

(A.7) $$rV_u(U,\delta) + C \geq w \cdot N,$$

where agents take $U$ and the selectivity of other agents as given.[7] $C$ represents the utility derived from the available public goods, net of any lump sum taxes used to finance them. We assume $C \geq 0$. The last term, $w \cdot N$, represents the loss of utility that results from congestion, including the rise in costs that result as scarce local inputs are bid up or depreciate more rapidly.

Taken together, (A.3) and (A.5) imply that $rV_u(U,\delta) = Ay(\delta)$. In other words, the flow value of being an un-matched agent is just equal to the output generated by the marginal match. In equilibrium (A.7) is satisfied with equality, so $N = \left(Ay(\delta) + C\right)/w$. This in turn implies that

$$\beta(U,\delta) = \frac{\pi}{2}\left(\sqrt{1 + 16\alpha\delta\left(Ay(\delta) + C\right)/\pi w} - 1\right).$$

**Comparative Static Results for Agents' Selectivity and Population**

We can substitute the preceding expression for $\beta$ into our first order condition and derive comparative static results based on a single equation.

**Proposition 1**: $\delta$ is decreasing in $\alpha$, $A$, and $C$, but increasing in $w$ and $\pi$.

**Corollary**: $N$ is increasing in $\alpha$, $A$, and $C$, but decreasing in $w$ and $\pi$.

In words, agents become more selective in response to an increase in the efficiency of the matching technology, the average productivity of matches, or local public goods. Agents become less selective in response to an increase in congestion costs or the separation rate.

Proof: For simplicity, we present the comparative static results in terms of elasticities. We begin with the result for $\alpha$, the arrival rate of meetings per unmatched worker:

$$\frac{\partial\delta}{\partial\alpha} \cdot \frac{\alpha}{\delta} = -\eta\left(\frac{\lambda + 1}{2\lambda}\right) \Big/ \left\{\eta\left(\frac{\lambda + 1}{2\lambda}\right)(1 - \phi) + (\varepsilon - \eta)\right\} < 0,$$

where $\phi = -Ay'(\delta)\delta/\left(Ay(\delta) + C\right)$ is the elasticity of population with respect to changes in $\delta$. We do not wish to rule out the possibility that $\phi > 1$, so we must verify that the denominator in the preceding expression is indeed positive. We begin by noting that $\varepsilon \geq \phi$ because we have assumed that $C \geq 0$. It is then sufficient to verify that

[7] The net benefit of living outside the city is normalized to zero.

$$\left(\frac{\lambda+1}{2\lambda}\right)\eta\left(1-\varepsilon\right)+\left(\varepsilon-\eta\right)>0.$$

Dividing by $\eta$ and rearranging terms we have

$$\frac{\varepsilon}{\eta}\left(1-\left(\frac{\lambda+1}{2\lambda}\right)\eta\right)-\left(1-\left(\frac{\lambda+1}{2\lambda}\right)\right)>0,$$

which is positive if $\tilde{y}(\delta)$ is strictly log concave.

Next, we consider the effect of increasing *w*, the coefficient on population that reflects utility lost to congestion:

$$\frac{\partial\delta}{\partial w}\cdot\frac{w}{\delta}=\frac{\partial\delta}{\partial\pi}\cdot\frac{\pi}{\delta}=\eta\left(\frac{\lambda+1}{2\lambda}\right)\Big/\left\{\eta\left(\frac{\lambda+1}{2\lambda}\right)\left(1-\phi\right)+\left(\varepsilon-\eta\right)\right\}>0.$$

This is also the expression for the elasticity of $\delta$ with respect to changes in the separation rate $\pi$. Now we consider an increase in *A*, the overall productivity of matches:

$$\frac{\partial\delta}{\partial A}\cdot\frac{A}{\delta}=-\eta\left(\frac{\lambda+1}{2\lambda}\right)\cdot\left(\frac{Ay(\delta)}{Ay(\delta)+C}\right)\Big/\left\{\eta\left(\frac{\lambda+1}{2\lambda}\right)\left(1-\phi\right)+\left(\varepsilon-\eta\right)\right\}<0,$$

where the extra term in the numerator is the elasticity of population with respect to *A*.

Finally, we consider the effect of an increase in public goods available in the city:

$$\frac{\partial\delta}{\partial C}\cdot\frac{C}{\delta}=-\eta\left(\frac{\lambda+1}{2\lambda}\right)\cdot\left(\frac{C}{Ay(\delta)+C}\right)\Big/\left\{\eta\left(\frac{\lambda+1}{2\lambda}\right)\left(1-\phi\right)+\left(\varepsilon-\eta\right)\right\}<0,$$

where the extra term in the numerator is the elasticity of population with respect to *C*.

The participation constraint (A.7) can be rewritten as $Ay(\delta)+C=wN$. Thus equilibrium population is inversely related to $\delta$, which proves the corollary. ■

**Comparative Static Results for Per Capita Output and Population Mass**

Here we present the main results characterized in our companion working paper:

**Proposition 2**: Per capita output of inventions is increasing in $\alpha$, *A*, and *C*, but decreasing in *w* and $\pi$.

**Corollary**: Per capita output of inventions is higher in cities with a larger population mass.

Proof: Per capita output is the product of the share of the population that is matched and the average productivity of matches:

$$\Phi\equiv\frac{M}{N}\cdot A\tilde{y}(\delta)=\frac{\beta}{\pi+\beta}\cdot A\tilde{y}(\delta)=\left(\frac{\lambda-1}{\lambda+1}\right)\cdot A\tilde{y}(\delta).$$

Again we present comparative static results in terms of elasticities, which take the form:

$$\frac{d\Phi}{dz}\cdot\frac{z}{\Phi\lambda}=\frac{z}{\lambda\left(\lambda^2-1\right)}\cdot\frac{\partial\lambda^2}{\partial z}+\left(1-\phi\right)\cdot\frac{\partial\delta}{\partial z}\cdot\frac{z}{\delta}-\eta\lambda\cdot\frac{\partial\delta}{\partial z}\cdot\frac{z}{\delta}.$$

The first two terms reflect the change in the share of agents matched at any given time $\left(M/N\right)$. These reflect changes in the intensive margin (holding *N* constant, does *M* rise or fall?) and the extensive margin (holding constant $\left(M/N\right)$, does *N* rise or fall?). The last term reflects the change in expected output per match, characterized in the previous proposition. Rearranging the expression, we have

$$\frac{d\Phi}{dz}\cdot\frac{z}{\Phi\lambda}=\frac{z}{\lambda\left(\lambda^2-1\right)}\cdot\frac{\partial\lambda^2}{\partial z}+\left\{\left(1-\phi\right)-\eta\lambda\right\}\cdot\frac{\partial\delta}{\partial z}\cdot\frac{z}{\delta}.$$

Consider first an increase in the arrival rate of meetings per unmatched worker $\alpha$:

$$\frac{d\Phi}{d\alpha}\cdot\frac{\alpha}{\Phi\lambda}=1+\left\{\left(1-\phi\right)-\eta\lambda\right\}\cdot\frac{\partial\delta}{\partial\alpha}\cdot\frac{\alpha}{\delta}=\frac{\left(\frac{\lambda+1}{2}\right)\eta^2+\left(\varepsilon-\eta\right)}{\left(\frac{\lambda+1}{2\lambda}\right)\eta\left(1-\phi\right)+\left(\varepsilon-\eta\right)}>0.$$

Next consider an increase in the overall productivity of innovations *A*:

$$\frac{d\Phi}{dA}\cdot\frac{A}{\Phi\lambda}=\frac{Ay(\delta)}{Ay(\delta)+C}+\left\{\left(1-\phi\right)-\eta\lambda\right\}\cdot\frac{\partial\delta}{\partial A}\cdot\frac{A}{\delta}=\frac{\left(\frac{Ay(\delta)}{Ay(\delta)+C}\right)\left\{\left(\frac{\lambda+1}{2}\right)\eta^2+\left(\varepsilon-\eta\right)\right\}}{\left(\frac{\lambda+1}{2\lambda}\right)\eta\left(1-\phi\right)+\left(\varepsilon-\eta\right)}>0.$$

Consider the effect of an increase in public goods *C* available in the city:

$$\frac{d\Phi}{dC}\cdot\frac{C}{\Phi\lambda}=\frac{C}{Ay(\delta)+C}+\left\{\left(1-\phi\right)-\eta\lambda\right\}\cdot\frac{\partial\delta}{\partial C}\cdot\frac{C}{\delta}=\frac{\left(\frac{C}{Ay(\delta)+C}\right)\left\{\left(\frac{\lambda+1}{2}\right)\eta^2+\left(\varepsilon-\eta\right)\right\}}{\left(\frac{\lambda+1}{2\lambda}\right)\eta\left(1-\phi\right)+\left(\varepsilon-\eta\right)}>0.$$

Finally, consider the effect of an increase in the utility lost due to congestion as population increases *w* (The effect of an increase in the separation rate $\pi$ takes on exactly the same expression):

$$\frac{d\Phi}{dw}\cdot\frac{w}{\Phi\lambda}=-1+\left\{\left(1-\phi\right)-\eta\lambda\right\}\cdot\frac{\partial\delta}{\partial w}\cdot\frac{w}{\delta}=\frac{-\left\{\left(\frac{\lambda+1}{2}\right)\eta^2+\left(\varepsilon-\eta\right)\right\}}{\left(\frac{\lambda+1}{2\lambda}\right)\eta\left(1-\phi\right)+\left(\varepsilon-\eta\right)}<0.$$

The corollary then follows from the previous corollary. ■

**Comparative Static Results for Changes in Land Area**

In BRW population varies, while the physical size of the city is fixed In other words, in terms of the model, changes in population and density are the same. Thus, in our empirical work it is important for us to examine the effects of both city size (the number of workers) and density (the number of workers per square mile).

But we also wish to consider the possibility that exogenous increases in land area might reduce density and therefore per capita output. This will occur if an insufficient number of workers respond to the reduction in congestion by moving into the city. To explore this possibility, we assume that the efficiency of the meeting rate parameter $\alpha$ and congestion coefficient $w$ are decreasing in city land area $L$. We show the following:

**Proposition 3**: Increasing $L$ decreases (increases) $\delta$, and $\Phi$ if $\alpha$ is more (less) sensitive to changes in $L$ than is $w$.

Proof: We derive the elasticity of per capita output with respect to changes in land area in the same manner as we did in Propositions 2 and 3:

$$\frac{d\Phi}{dL}\cdot\frac{L}{\Phi}=\frac{L}{\lambda}\left(\frac{\alpha'(L)}{\alpha(L)}-\frac{w'(L)}{w(L)}\right)\left\{2+\frac{(\lambda+1)(1-\eta)\eta^2}{\left(\frac{\lambda+1}{\lambda}\right)\eta(1-\phi)+2(\varepsilon-\eta)}\right\}.\blacksquare$$

In the model, increasing land area has two effects on workers' incentive to migrate to the city. The first effect is a reduction in the costs associated with congestion which, all else equal, induces additional migration. The second is a reduction in expected income that results from waiting longer for the next meeting with an unmatched worker. All else equal, this would discourage workers from entering the city.

It turns out that, in the absence of a fully specified model for the supply of developed land, we cannot sign the effect a priori. If the reduction in congestion costs is sufficiently large relative to the decline in the efficiency of meeting technology, both density and per capita output rises. But if the reduction in congestion costs is relatively small, density and per capita output will fall.

## APPENDIX B: Data Construction

**Patent Data**

Our patent data are derived from data sets furnished by the Technology Forecasting and Assessment Branch of the U.S. Patent and Trademark Office. These include the US Patent Inventor File 1977-99: All Grants; 75-76 Utility Patents Only and the PATSIC99 file. We are grateful to Jim Hirabayashi of USPTO for his assistance with these data.

We assemble the home address information of the first inventor named on each patent from the inventor's file. We began with 1,198,376 patents granted to inventors with an American address over the years 1975-1999. We assigned a state-county FIPS code for each patent by matching address information against the 1998 vintage of NIST's FIPS55 place names data set. We obtained unique ZIP code-county and place name-county

matches for 81 percent of the patents. For another 8.5 percent of the patents, we could not initially identify a unique county for the inventor but we could identify a unique metropolitan area. Another 6.6 percent of patents were geographically located via matches against the location of an R&D lab of the firm owning the patent and manual searches.

In all 1,155,133 patents were matched to a county or an MA, while 43,243 were not. Of those patents that were successfully placed in an MSA, 581,001 were granted between 1990 and 1999 (inclusive); 46,647 were located outside an MSA. That leaves 534,354 patents in MAs that were used to construct the patent variable.

Patent classifications are obtained by matching patents in the inventor file to patents in the firm file. The patent class dummies (medical, chemical, computer, electrical, and mechanical) are constructed using the primary patent classes suggested in Appendix 1 of Hall, Jaffe, and Trajtenberg (2001). The industry mix of patents is obtained by matching patent numbers with those included in the NBER Patent Citations Data File, obtaining the CUSIP of the firm initially assigned the patent and obtaining the SIC of that firm from the 1999 vintage of Compustat. The high-tech patent share is derived using the set of R&D intensive three-digit SIC industries identified in Office of Technology Policy (2001).

**Our Definition of Metropolitan Areas**

Our MA definitions are based primarily on MSA definitions defined by the Office of Management and Budget in 1983 (http://www.census.gov/population/estimates/metro-city/83mfips.txt). Several adjustments are made:

- The six MSAs in Puerto Rico are removed.
- New England County Metropolitan Areas (NECMAs) are used as our MAs in New England.
- The Bureau of Economic Analysis (BEA) uses its own set of county-equivalent codes to tabulate data for independent cities and their surrounding counties together. For all data from the Regional Economic Information System (REIS), data for our MAs are built up including these independent cities.
- Nine Consolidated Metropolitan Statistical Areas (CMSAs) are employed instead of their 25 component MSAs: Chicago, IL-IN-WI; Cincinnati-Hamilton, OH-KY-IN; Cleveland-Akron-Lorain, OH; Dallas-Fort Worth, TX; Houston-Galveston-Brazoria, TX; Kansas City, KS-MO; Portland-Vancouver, OR-WA; Seattle-Tacoma, WA; and St. Louis- East St. Louis- Alton, IL-MO.
- Seven ad hoc metropolitan areas were also created: Denver-Boulder-Greeley (Boulder-Longmont, CO PMSA; Denver, CO PMSA; & Greeley, CO MSA); Greenville-Anderson (Anderson, SC MSA and Greenville-Spartanburg, SC MSA); Los Angeles-Anaheim (Anaheim-Santa Ana, CA PMSA and Los Angeles-Long Beach, CA PMSA); Midland-Odessa (Midland, TX MSA and Odessa, TX MSA); New York-Northern New Jersey (Bergen-Passaic, Jersey City, Middlesex-Somerset-Hunterdon, Monmouth-Ocean, Nassau-Suffolk, New York, Newark, Orange County); San Francisco-Oakland (Oakland, CA PMSA and San Francisco, CA PMSA); and Sarasota-Bradenton (Bradenton, FL MSA and Sarasota, FL MSA).

We used these definitions for certain cities rather than the underlying MSAs because two or more cities shared a common border and we could not always assign some patents (a few thousand) to an MSA with certainty (see Patent Data above). If we used the underlying MSAs, we would either have to discard these patents or take the chance that the error rate in assigning patents to cities might vary across cities. The methodology used in USPTO (2000) assigns equal shares of patents to counties with common place names when a patent cannot be matched to a unique county. This too might imply a higher error rate in assigning patents when MSAs are close to each other.

We compared our MSA patent counts to those reported in USPTO (2000) and found them to be extremely close except for a few instances. In some cases two or more MSAs were in close proximity (e.g., Dallas and Fort Worth). In a few others, the place name of the inventor's address was common to more than one county, regardless of distance. The PTO algorithm divided those patents equally across those counties.

Our definition of metropolitan areas or the manner of allocating patents to MSAs does not significantly influence our results. In an earlier version of this paper (Carlino, Chatterjee, and Hunt 2001) we estimated the relationship between patent intensity and employment density using a data set of 296 MSAs and PMSAs as defined by OMB in 1983 and using patent counts built up from the data contained in USPTO (2000). The results were qualitatively the same as those reported here, although the estimated coefficients were somewhat larger.

**Missing Data**

Four MSAs are dropped in the analysis because of missing data. One MSA (Enid, OK) does not have a corresponding urbanized area. Owing to disclosure limitations, a Herfindahl index of industry employment shares cannot be calculated for Atlantic City, NJ, and Tallahassee, FL. In addition, manufacturing employment is not available in 1989 for Columbus, GA-AL.

**Geographic Variables**

Urbanized Area land area for every county was obtained from Table 34 of Census Bureau (1993a). We also obtained comparable measures for urbanized areas defined in 1980 and 1970 from Census Bureau (1979, 1984).

There are a few instances where a county includes land area associated with more than one urbanized area. For example, portions of Bucks County, PA, are associated with the Philadelphia and Trenton urbanized areas. To be consistent with our other county-based measures, we attribute this land area (and associated employment) to the MA associated with that county (e.g., Philadelphia).

The weather and topography instrumental variables are based on the USDA's Natural Amenity Scale project (http://www.ers.usda.gov/data/naturalamenities). The data are reported at county level, which we aggregate, based on county land area, to MSAs. We use the variables indicating, for the years 1941-1970, mean hours of sunlight and temperature in January, mean temperature in July, the percent of land area covered by surface water, and five dummy variables for the presence of a particular type of geography (plains, tablelands, open hills and mountains, hills and mountains, and plains with hills and mountains) built up from a finer gradation in the ERS data. We also

included as an instrument the amount of MA area covered by water n 1990. Those data were obtained from a county level tabulation reported in the Census Bureaus Gazeteer (http://www.census.gov/geo/www/gazetteer/gazette.html).

**Economic and Demographic Variables**

Our primary employment data are county-level values reported in the BEA's REIS database and aggregated to the MA level. Those data are derived from the BLS Covered Employment and Wages Program (ES-202), which represents the average annual number of full- and part-time jobs held by all workers who are covered by unemployment insurance. Industry breakdowns are based on 1987 SIC definitions.

Our measures of residency-based employment in 1990 were obtained from the Census Bureau web site (www.americanfactfinder.com). Those data are derived from the STF3 (5 percent sample) tape. A separate count was obtained for every county that includes an urbanized area. These are aggregated to MAs in the same manner as our other variables.

Our counts of employment for professional specialty occupations are derived from the 1990 census (STF3). This is a residency-based count, but here we include all residents in the counties making up an MA, not just residents in the urbanized area portion of those counties. The occupation codes for this category of jobs are codes 043-202 in the 1990 Census Occupation Classification System.

Our counts of scientists and engineers are residency-based measures of workers in these occupations living in urbanized areas as reported in Table 34 of Census Bureau (1993c). We aggregated these counts to be consistent with our MA definitions.

The shares of MSA land area, population, and employment contained in urbanized areas discussed on pages 13-14 are derived from Tables 48, 50-51 of Census Bureau (1993a) and Table 33 of Census Bureau (1993b) and Census Bureau (1993c).

The Herfindahl-Hirschman Index (HHI) is calculated for 1989 and includes employment for ten industrial sectors. These include Construction, Manufacturing, Transportation and Public Utilities, Wholesale Trade, Retail Trade, Finance, Insurance, and Real Estate, Services, Civilian Federal Government, State and Local Government, and a category, other, that consists primarily of employment in the military, agriculture, and mining. The HHI is calculated at the MA level. Wherever county data are missing for 1989, either data from the previous or following year are used or MSA-level data are substituted when appropriate.

The percentage of population with a college degree or more education is derived from the 1990 Census American Fact Finder. Because these are MSA-level data, it was necessary to create a weighted average of component MSAs for the special CMSAs we created; this was done using 1990 Census Bureau mid-year population estimates aggregated from county level to MSA. The population data were also used to calculate our measure of patents per capita.

Our instrumental variables include a measure of the number of restaurants and museums and crimes in 1989. These are derived from county-level data, as reported in County Business Patterns, and aggregated to the MA level. They are converted to intensities by dividing by population in 1989.

**Research and Development Variables**

The amount of academic R&D in each MSA is derived from the National Science Foundation's (NSF) "Academic Science and Engineering: R&D Expenditures," as archived in the WebCaspar search engine at the NSF web site. Expenditures are averaged over the period 1987-89 and are normalized by the total fall enrollment as reported in the Integrated Postsecondary Education Data System assembled by the National Center for Education Statistics and archived on WebCaspar. S&E expenditures reported for a number of university systems were also allocated to particular campuses using advanced S&E degrees granted by those campuses (details are found in an extended data appendix available from the authors). Expenditures are built up from counties according to our MA definitions.

Data on the location and resources of the Federally Funded Research and Development Centers were provided to us by Ronald Meeks of the National Science Foundation.

Data on private R&D facilities were extracted from the 1989 edition of the *Bowker Directory of American Research and Technology*. They are matched to the SIC of the parent company using the 1999 vintage of Compustat.

**Trade Secrecy Index**

We assigned the industry-specific effectiveness rating in Table 1 of Cohen, Nelson, and Walsh (2000) to two-digit or three-digit SIC industries. These ratings are a categorical response to a question that asks for the proportion of product innovations for which trade secrets are effective in preserving the resulting profits. For each MA, we compute a weighted average of the industry ratings using as weights the shares of all private R&D facilities in the MA contained in those industries.

## References

Berliant, Marcus, Robert R. Reed, III and Ping Wang. 2004. "Knowledge Exchange, Matching, and Agglomeration," mimeo, Economics Department, Washington University, St. Louis.

Bagnoli, Mark, and Ted Bergstrom. 1988. "Log Concavity and Its Applications," mimeo, University of California at Santa Barbara.

Burdett, Kenneth and Randall Wright. 1998. "Two-Sided Search with Nontransferable Utility," *Review of Economic Dynamics*. Vol. 1, pp. 220-245.

Carlino, Gerald, Satyajit Chatterjee, and Robert M. Hunt, 2001, "Knowledge Spillovers and the New Economy of Cities," Federal Reserve Bank of Philadelphia Working Paper No. 01-14.

Carlino, Gerald, Satyajit Chatterjee, and Robert M. Hunt. 2005. "Matching & Learning in Cities: Urban Density and the Rate Of Invention," Federal Reserve Bank of Philadelphia Working Paper No. 04-16/R.

Cohen, Wesley. M., Richard Nelson, and John P. Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (OR NOT)," NBER Working Paper No. 7552.

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

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools," NBER Working Paper No. 8498.

Office of Technology Policy. 2001. *The Dynamics of Technology-based Economic Development: State Science and Technology Indicators, 2nd Ed*. Washington: U.S. Department of Commerce.

Shimer, Robert, and Lones Smith. 2000. "Assortative Matching and Search," *Econometrica*, Vol. 68, pp. 343-369.

U.S. Census Bureau. 1979. *Population and Land Area of Urbanized Areas, for the United States: 1970 and 1960.* Census of Population Supplementary Report PC (S1)-108.

U.S. Census Bureau. 1984. *Population and Land Area of Urbanized Areas for the United States and Puerto Rico: 1980 and 1970.* Census of Population Supplementary Report PC80-S1-14.

U.S. Census Bureau. 1993a. *Census of Population and Housing, Population Housing Unit Counts, United States* (CPH-2-1).

U.S. Census Bureau. 1993b. Census *of Population, Social and Economic Characteristics, Metropolitan Areas* (CP-2-1B)

U.S. Census Bureau. 1993c. *Census of Population, Social and Economic Characteristics, Urbanized Areas* (CP-2-1C).

U.S. Patent and Trademark Office, 2000, *United States Patent Grants by State, County, and Metropolitan Area*. Washington: U.S. Patent and Trademark Office, Office for Patent and Trademark Information.