Agglomeration, Knowledge Spillovers, and Corporate Investment
Internet Appendix

Section A of this appendix describes the O*NET data and how we aggregate the measure of knowledge intensity to the Fama–French 48 industry level. Section B provides additional figures and tables.

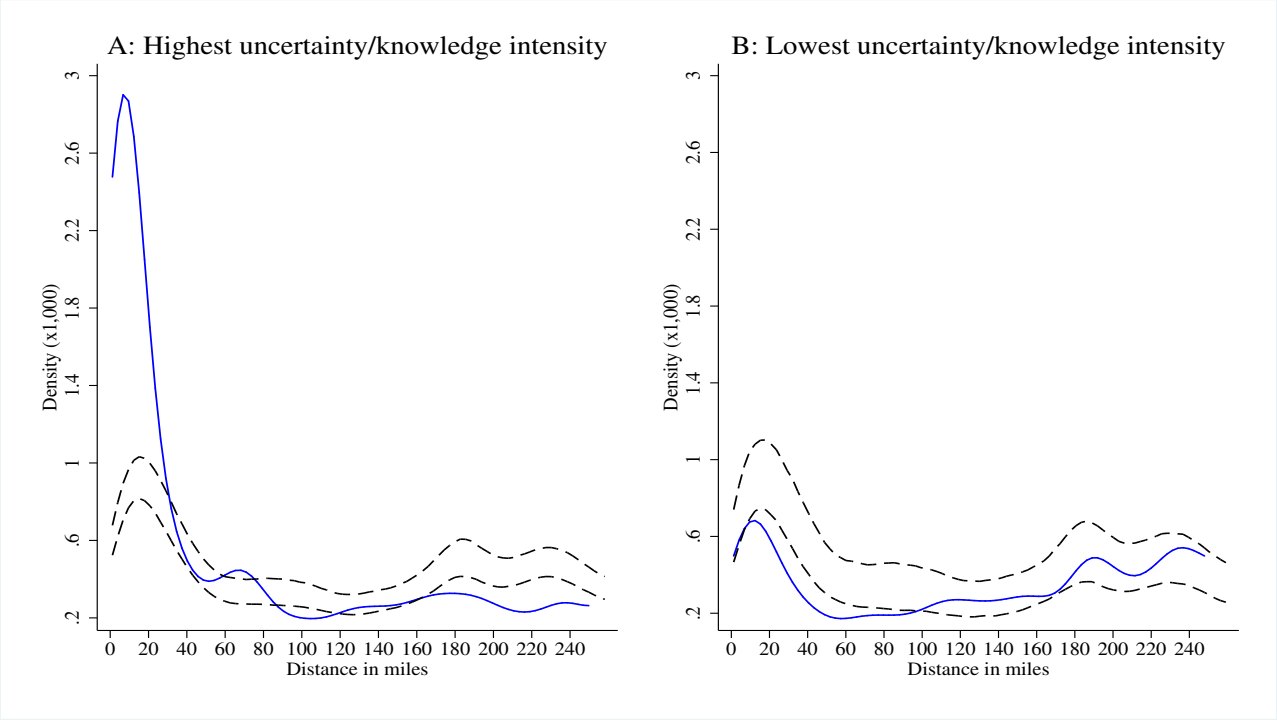
A. O*NET data and aggregation of the measure of knowledge intensity

O*NET classifies each occupation into one of five skill categories according to the degree of preparation needed. The skill levels of occupations range from *little or no preparation needed* (Job Zone 1) to *extensive preparation* (Job Zone 5). More specifically, Job Zone 1 includes occupations that may require a high school diploma or GED, little or no previous work-related skill required, and a few days to a few months of on-the-job training. Job Zone 5 includes occupations that typically require a master's degree, Ph.D., M.D., or J.D.; in other words, extensive skill, knowledge, and experience. Examples of occupations in Job Zone 1 include taxi drivers, amusement and recreation attendants, and non-farm animal caretakers, while examples from Job Zone 5 include lawyers, sports medicine physicians, surgeons, treasurers, and controllers.

To aggregate the O*NET skill measures to the industry level, we create a wage-weighted average skill for each 4-digit NAICS code, using the job zone assigned to each occupation according to the O*NET database. Wage estimates come from the Bureau of Labor and Statistics (BLS) Occupational Employment Statistics (OES) database. We calculate the total industry cost of input (wage) for each occupation by multiplying its annual mean wage by the number of people employed in an industry at that occupation according to the OES. Finally, we aggregate the average skill level across all 4-digit NAICS in each Fama–French 48 industry classification.

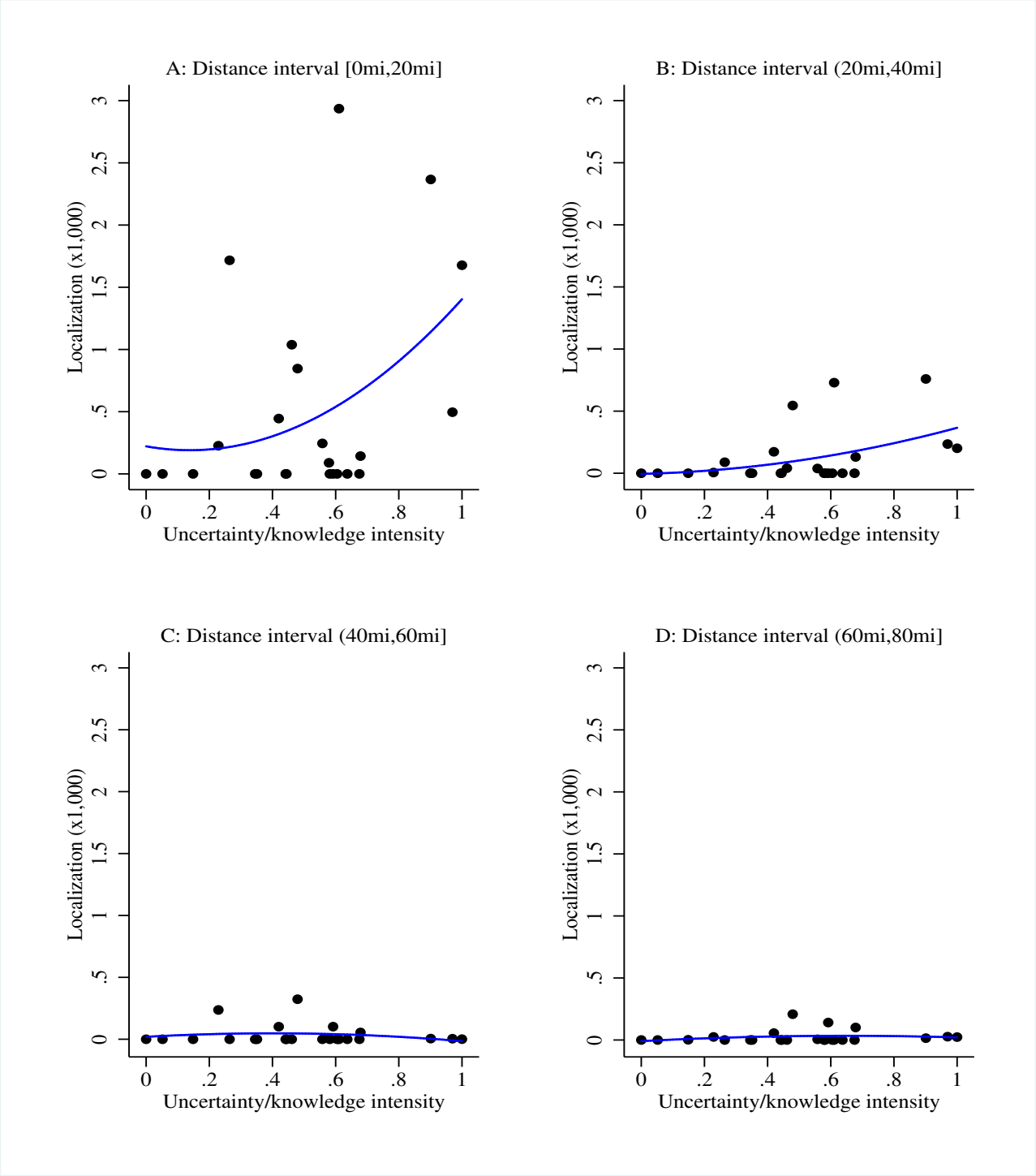
B. Additional figures and tables

Figure IA.1. Localization and industry uncertainty/knowledge intensity



This figure plots the kernel densities of the highest and lowest uncertainty/knowledge intensity (UKI Index) industries (×1,000 for scale). Industries are based on the Fama and French 48 industry classification (we exclude finance and utilities industries). Industries are ranked by the UKI index. Dashed lines represent 95% confidence intervals.

Figure IA.2. Localization and industry uncertainty/knowledge intensity by distance intervals



This figure plots the industry localization index against uncertainty/knowledge intensity (UKI index) for various distance intervals. The solid line represents a quadratic interpolation.

Table IA.1. Firm-level summary statistics

	Mean	SD	p25	p50	p75
Log(Assets)	3.23	2.76	1.60	3.44	5.09
ROA	-0.13	0.30	-0.20	0.01	0.07
Log(Size)	3.28	2.57	1.50	3.38	5.08
Market leverage	0.16	0.19	0.00	0.09	0.25
Investment	0.34	22.70	0.01	0.03	0.07
R&D	7.4	13.1	0.0	0.0	7.4

This table describes the 9,167 firms in the main sample. We define all variables in Appendix A.

Table IA.2. Industry Summary Statistics

Industry	Annualized volatility	Skill	Rank (volatility)	Rank (skill)	UKI index	Number of firms	Firm pairs within 20mi	Fraction within 20mi
Electronic equipment	0.339	3.508	1	4	1	558	10,878	7.7
Measuring and control equipment	0.322	3.556	2	2	0.970	188	695	4.1
Computers	0.271	3.763	6	1	0.901	221	2,150	9.5
Automobiles	0.320	2.734	4	17	0.679	136	258	3.2
Steel	0.320	2.718	3	18	0.675	119	125	2.0
Machinery	0.262	3.071	7	10	0.637	288	633	1.7
Oil	0.215	3.368	13	6	0.611	521	12,493	11.8
Personal Services	0.230	3.229	11	9	0.605	126	132	1.9
Electrical equipment	0.255	2.997	9	11	0.592	140	124	1.4
Construction	0.271	2.836	5	16	0.581	110	149	3.0
Healthcare	0.202	3.372	15	5	0.579	187	461	3.1
Telecommunications	0.200	3.328	16	7	0.558	418	2,484	3.3
Pharmaceuticals	0.147	3.524	22	3	0.479	786	12,457	4.8
Entertainment	0.260	2.580	8	20	0.461	218	1,232	6.2
Chemicals	0.211	2.920	14	13	0.444	194	368	2.2
Construction materials	0.239	2.688	10	19	0.442	148	145	1.4
Medical equipment	0.154	3.299	21	8	0.420	385	2,463	3.8
Transportation	0.185	2.853	18	14	0.351	237	527	2.2
Wholesale	0.168	2.969	20	12	0.345	374	1,339	2.2
Clothing	0.229	2.255	12	23	0.264	105	491	9.3
Household consumer goods	0.141	2.844	24	15	0.229	155	368	3.6
Retail	0.173	2.367	19	21	0.149	477	2,099	2.1
Food	0.142	2.329	23	22	0.052	150	225	2.3
Meals, restaurants, and hotels	0.188	1.812	17	24	0	208	337	1.8

This table ranks the various industries based on industry-level uncertainty and knowledge intensity, as well as on measures of firm localization. To construct the uncertainty/knowledge-intensity index (UKI index), we add standardized values of industry-level volatility and worker skill. Then, the resulting values are normalized so that the UKI index ranges from 0 to 1.

Table IA.3. Location and uncertainty/knowledge intensity

	log(distance)		Within 20mi		20-40mi	40-60mi	60-80mi
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
UKI index$_{FF48}$	-0.0503***	-0.1363***	0.0434***	0.0541***	0.0047***	0.0007	-0.0023***
	(0.0143)	(0.0228)	(0.0031)	(0.0041)	(0.0017)	(0.0010)	(0.0008)
log(Sales)$_{Firm1}$		-0.0171***		0.0014***	0.0004*	-0.0003***	-0.0004***
		(0.0025)		(0.0003)	(0.0002)	(0.0001)	(0.0001)
log(Sales)$_{Firm2}$		-0.0172***		0.0014***	0.0004***	-0.0003***	-0.0004***
		(0.0008)		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Number of firms$_{FF48}$		0.0120***		-0.0013***	-0.0005***	-0.0001***	0.0000
		(0.0008)		(0.0001)	(0.0001)	(0.0000)	(0.0000)
R^2	0.0029	0.0038	0.0016	0.0035	0.0008	0.0006	0.0001
N	3,730,680	3,730,680	3,730,680	3,730,680	3,567,790	3,491,016	3,457,266

This table reports estimates for the relation between industry peers' locations and uncertainty/knowledge intensity. The dependent variable in Columns (1)–(2) is the natural log of the distance (in miles) between industry peers' HQ locations. The dependent variables in Columns (3)–(7) are indicators of whether these firms are located within 20 miles, between 20 miles and 40 miles, between 40 miles and 60 miles, or between 60 miles and 80 miles, respectively. Distances are calculated from geographic coordinates for corporate HQ ZIP codes. The regressions in Columns (2) and (4)–(7) include controls for the log of sales of both firms, as well as the number of firms in the same Fama–French 48 industry classification. Standard errors are clustered at the Fama–French 48 industry level and are reported in parentheses, below the coefficient estimates. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table IA.4. Industry Citation Networks

	$Citation_{ij}$	
	(1)	(2)
Edges	-5.275***	-5.274***
	(0.016)	(0.016)
Same city (20mi)	-0.145 **	0.107*
	(0.073)	(0.060)
Same FF-48	0.010	0.015
	(0.020)	(0.020)
Uncertainty × Same city (20mi)	0.460***	
	(0.087)	
Knowledge capital × Same city (20mi)		0.092
		(0.080)
Uncertainty	0.070***	0.069***
	(0.003)	(0.003)
Knowledge capital	0.039***	0.040***
	(0.002)	(0.002)
Organizational capital	-0.011***	-0.011***
	(0.002)	(0.002)
Classes per patent	0.473***	0.472***
	(0.015)	(0.015)
AIC	747570	747616
Firm-level controls	Yes	Yes

This table presents exponential random graph model (ERGM) estimates of firms' citation relationships. The dependent variable in all models is a binary variable indicating whether a firm cites another firm's patents. The coefficients represent the contribution of industry and firm covariates on the conditional log-odds that two firms cite each other's patents. The intercept estimate (*Edges*) indicates the unconditional probability that a firm cite's another firm's patents. We define all variables in Appendix A. The ERGM is estimated via MCMC maximum likelihood. The standard errors (reported in parentheses) are calculated using the standard deviations of the posterior distribution of the corresponding parameter estimates. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table IA.5. Agglomeration, UKI, and Investment Expenditures

Panel A: 20-mile Concentric Geographic Areas

	CapEx			R&D		
	(1)	(2)	(3)	(4)	(5)	(6)
Industry Inv. within 20mi	0.0052***	0.0077***	0.0006	0.0021***	0.0018***	-0.0237***
	(0.0012)	(0.0016)	(0.0038)	(0.0004)	(0.0004)	(0.0033)
Industry Inv. outside 20mi	0.0034***	0.0054***	0.0054***	-0.0006***	-0.0007***	-0.0006***
	(0.0002)	(0.0008)	(0.0008)	(0.0001)	(0.0001)	(0.0001)
Industry Inv. within 20mi ×UKI index			0.0112**			0.0404***
			(0.0046)			(0.0054)
Year FE	yes	yes	yes	yes	yes	yes
Firm FE	no	yes	yes	no	yes	yes
Firm controls	yes	yes	yes	yes	yes	yes
R^2	0.0282	0.6657	0.6657	0.0059	0.6732	0.6739
N	92,014	92,014	92,014	92,014	92,014	92,014

Panel B: 40-mile Concentric Geographic Areas

	CapEx			R&D		
	(1)	(2)	(3)	(4)	(5)	(6)
Industry Inv. within 40mi	0.0062***	0.0084***	0.0032	0.0018***	0.0016***	-0.0166***
	(0.0011)	(0.0013)	(0.0033)	(0.0003)	(0.0003)	(0.0024)
Industry Inv. outside 40mi	0.0031***	0.0052***	0.0052***	-0.0006***	-0.0007***	-0.0006***
	(0.0002)	(0.0008)	(0.0008)	(0.0001)	(0.0001)	(0.0001)
Industry Inv. within 40mi ×UKI index			0.0083*			0.0290***
			(0.0043)			(0.0038)
Year FE	yes	yes	yes	yes	yes	yes
Firm FE	no	yes	yes	no	yes	yes
Firm controls	yes	yes	yes	yes	yes	yes
R^2	0.0279	0.6625	0.6625	0.0058	0.6741	0.6745
N	92,014	92,014	92,014	92,014	92,014	92,014

This table reports estimates for the relation between investment comovement, regional proximity, and uncertainty/knowledge intensity. Estimated regressions are of the form:

$$\begin{aligned} Investment_{j,t}^{i,a} &= \delta + \beta_1 Investment_{p,t}^{i,a} + \gamma Investment_{p,t}^{i,a} \times UKI_{index} \\ &\quad + \beta_2 Investment_{p,t}^{i,-a} + \beta_3 Controls_t^{i,a} + \epsilon_{j,t}^{i,a}, \end{aligned}$$

where $Investment_{j,t}^{i,a}$ represents the investment of firm j in industry i and area a during year t. The dependent variable in Columns (1)–(3) ((4)–(6)) is the natural log of capital expenditures (R&D). The independent variable *Industry Inv within 20mi (40mi)* is an equally weighted portfolio (p) of firms within firm $j's$ industry (i) and its 20-mile (40-mile) area (a). Similarly, *Industry Inv outside 20mi (40mi)* is the equally weighted portfolio of firms within firm $j's$ industry (i) but located outside its area (a). Firm j is excluded from the calculation of each portfolio. The area a is defined to be the 20-mile (Panel A) and the 40-mile (Panel B) concentric circle that surrounds the centroid of firm $j's$ HQ ZIP code. Columns (3) and (6) include the interaction term ($Investment_{p,t}^{i,a} \times UKI_{index}$). All columns include year fixed effects, and Columns (2)-(3) and (5)-(6) include firm fixed effects. Firm controls account for Tobin's Q, the natural log of total assets, Z-score, cash flow, cash holdings, and tangibility. Standard errors are clustered at the Fama–French 48 industry level and are reported in parentheses, below the coefficient estimates. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table IA.6. Firm Performance, Industry Clusters, and the UKI Index

	ROA_{In}–ROA_{Out}	ROA		
	(1)	(2)	(3)	(4)
UKI index	0.0615**			
	(0.0271)			
1(*In Cluster*)		0.0110***	0.0181***	0.0257
		(0.0011)	(0.0028)	(0.0239)
1(*In Cluster*) × UKI index			0.0129***	0.0636*
			(0.0046)	(0.0384)
Year FE	No	Yes	Yes	Yes
Industry FE	No	Yes	Yes	No
Industry×year FE	No	No	No	Yes
R^2	0.1896	0.0009	0.0011	0.0075
N	24	92,014	92,014	92,014

This table presents estimates for relations of return on Assets (ROA) between firms inside and outside of industry clusters. We define firms as being in an industry cluster if the number of industry peers located within 20 miles is above the industry median. Column (1) reports results for the average differences in ROA between firms inside and outside of industry clusters, for each of the 24 Fama and French industries in our sample (i.e., regression results for the analog of Panel B of Figure 2). Columns (2)–(4) repeat the analysis at the firm level. Column (2) considers differences in ROA for firms inside and outside of industry clusters. Column (3) considers the interaction of being inside of a cluster with the UKI index. Column (4) repeats the specification of Column (3), but with industry×year-fixed effects. The specifications in Columns (2)–(4) only include dummy variables on the right-hand side, which explains the low R^2. Standard errors are clustered at the Fama–French 48 industry level (in Columns (2)–(4)) and are reported in parentheses below the coefficient estimates. No adjustments are made to the standard errors in Column (1), which is estimated using only 24 observations. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table IA.7. Geography Networks

	HQ locations				*State mentions*	
	(1)	(2)	(3)	(4)	(5)	(6)
Edges	−3.507***	−3.526***	−3.631***	−3.698***	−3.492***	−3.482***
	(0.003)	(0.006)	(0.012)	(0.014)	(0.003)	(0.006)
UKI high	0.007	0.002	0.003		0.043***	0.031***
	(0.007)	(0.007)	(0.007)		(0.004)	(0.004)
Same FF-48	0.0002	−0.001	−0.003	0.021	−0.121***	−0.083***
	(0.015)	(0.015)	(0.015)	(0.009)	(0.007)	(0.007)
Same FF-48 × UKI high		0.046***	0.041***		0.042***	0.047***
		(0.006)	(0.007)		(0.003)	(0.003)
Organizational capital			0.009***	0.009***		0.002***
			(0.002)	(0.002)		(0.001)
| Δ Organizational capital |			0.031***	0.024***		0.027***
			(0.002)	(0.002)		(0.001)
Classes per patent			0.153***	0.156***		0.090***
			(0.017)	(0.017)		(0.009)
| Δ Classes per patent |			0.124***	0.135***		0.018
			(0.023)	(0.023)		(0.012)
Knowledge capital				0.019***		0.029***
				(0.002)		(0.001)
Knowledge intensity				0.006***		
				(0.002)		
Uncertainty				0.001***		
				(0.0001)		
AIC	1369243	1369228	1368812	1368644	5182782	5176680
Firm-level Controls	No	No	Yes	Yes	No	Yes

This table repeats the analysis in Table 1 after excluding from the sample firms in the oil, gas, and coal industries. The table presents exponential random graph model (ERGM) estimates of firm headquarters location. The dependent variable in all models is a binary variable indicating whether two firms are geographic neighbors. In models (1)-(4), the geography network is based on firms' distances from the headquarters, and in columns (5)-(6), it is based on the inverse of the Mahalanobis distances between firms' vector of state-mentions in their 10-Ks (State mentions data come from Garcia and Norli (2012)). The coefficients represent the contribution of industry and firm covariates on the conditional log-odds that two firms are located within 20 miles while holding all other ties fixed. The intercept estimate (*Edges*) indicates the unconditional probability that a firm is located within 20 miles of a random firm that is added to the network. We define all variables in Appendix A. The ERGM is estimated via MCMC maximum likelihood. The standard errors (reported in parentheses) are calculated using the standard deviations of the posterior distribution of the corresponding parameter estimates. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.