Online Appendix

Income Segregation and the Rise of the Knowledge Economy

by Enrico Berkes[1] and Ruben Gaetani[2]

[1] Ohio State University, berkes.8@osu.edu.

[2] University of Toronto, ruben.gaetani@utoronto.ca.

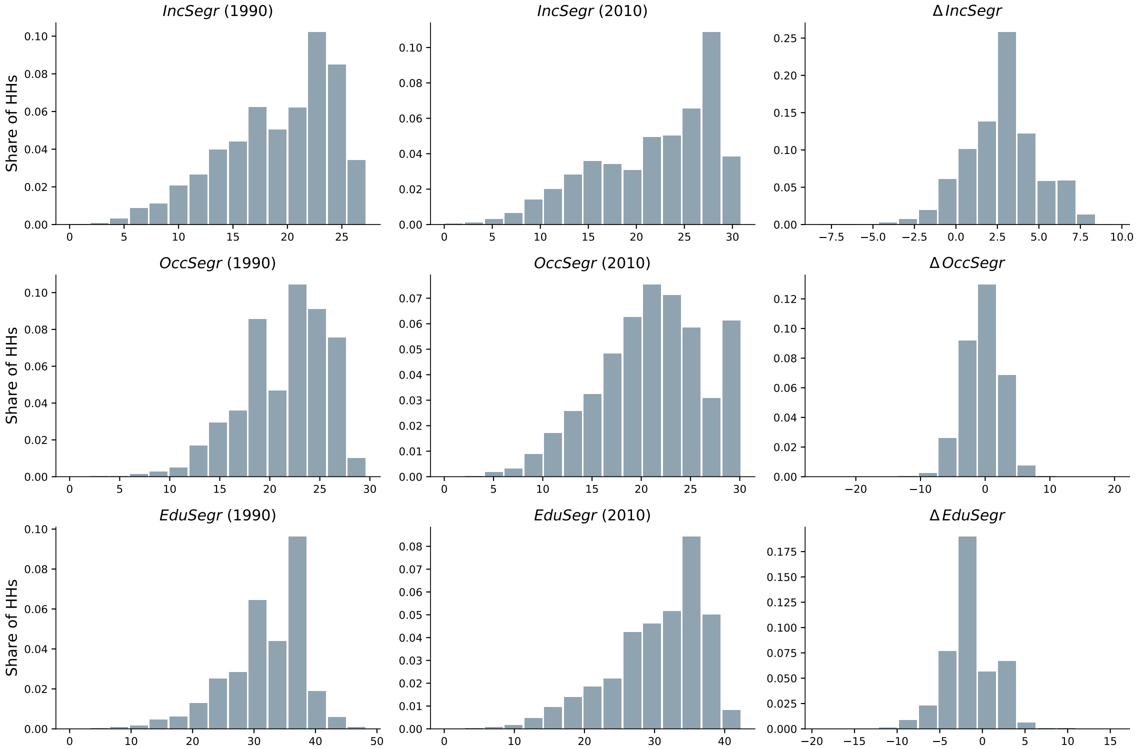
A Additional Tables and Figures

Table A.1: **Summary statistics**

	Obs	Min	Max	Mean	S.D.
$\Delta \log(1 + Patents)$	722	-2.77	3.32	0.18	0.50
Total patents (1990)	722	0	30,227	6,853	8,610
Total patents (2010)	722	0	69,719	9,220	12,716
Average income (1990)	722	18,144	59,698	39,149	8,438
Number of households (1990)	722	520	4,914,645	1,041,053	1,319,922
Number of census tracts	722	1	3031	655.4	821.1
$IncSegr$ (1990)	722	0	27.3	19.2	5.2
$IncSegr$ (2010)	722	0	30.9	21.6	6.5
$\Delta IncSegr$	722	-8.3	9.6	2.8	2.1
$OccSegr$ (1990)	722	0	29.7	21.5	4.2
$OccSegr$ (2010)	722	0	30.1	20.7	5.5
$\Delta OccSegr$	722	-25.7	20.2	-0.5	2.7
$EduSegr$ (1990)	722	0	48.3	32.0	6.4
$EduSegr$ (2010)	722	0	42.4	30.3	6.9
$\Delta EduSegr$	722	-19.2	15.6	-1.4	2.8
Share of college graduates (1990)	722	0.05	0.39	0.20	0.06
Trade exposure	722	-0.08	25.41	1.14	1.00

Notes: Summary statistics are weighted by the number of households in the corresponding year.

Figure A.1: **Economic segregation: distribution across cities**



Notes: Histograms weighted by the number of households in 1990.

Table A.2: **Economic segregation in a selected sample of cities**

	IncSegr			*OccSegr*			*EduSegr*		
	1990	2010	Δ	1990	2010	Δ	1990	2010	Δ
Los Angeles, CA	24.0	27.2	+3.3	27.1	29.8	+2.6	35.7	38.3	+2.6
New York City, NY	27.3	30.9	+3.6	27.0	29.1	+2.1	37.5	36.0	-1.4
Chicago, IL	25.0	28.8	+3.8	25.2	26.3	+1.1	38.6	37.3	-1.3
Philadelphia, PA	22.7	29.0	+6.2	24.1	25.3	+1.1	37.4	36.4	-1.0
Newark, NJ	23.4	28.6	+5.2	23.2	25.4	+2.2	32.5	34.9	+2.4
Detroit, MI	24.3	27.5	+3.1	25.9	24.9	-1.0	39.5	36.8	-2.6
Boston, MA	19.4	25.8	+6.4	20.5	23.2	+2.7	31.7	33.6	+1.9
San Francisco, CA	23.0	28.2	+5.2	23.5	26.5	+2.9	33.1	35.9	+2.7
Baltimore, MD	22.7	29.6	+6.8	24.6	22.8	-1.8	37.4	35.1	-2.3
Houston, TX	26.6	28.2	+1.6	28.1	29.0	+0.9	40.9	40.2	-0.7
Miami, FL	21.0	28.3	+7.3	22.6	22.4	-0.1	29.5	28.8	-0.7
Bridgeport, CT	22.6	27.0	+4.4	19.7	21.9	+2.2	30.7	32.3	+1.6
Seattle, WA	17.6	23.9	+6.3	19.2	25.2	+6.0	30.2	34.1	+3.9
Pittsburgh, PA	21.3	22.2	+0.8	22.4	21.6	-0.8	34.8	32.0	-2.8
Atlanta, GA	12.2	12.3	+0.1	17.6	14.6	-3.0	27.6	19.7	-7.9

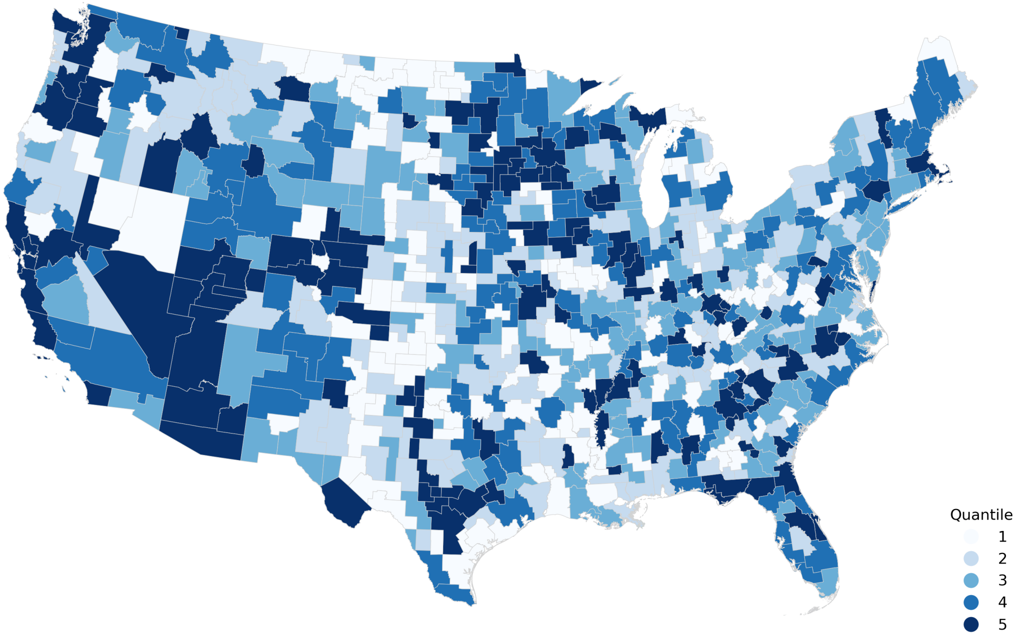
Notes: The table reports measures of economic segregation for the 15 largest commuting zones in 1990. For clarity, commuting zone names refer to the largest city only.

Table A.3: **Structure of the instrument**

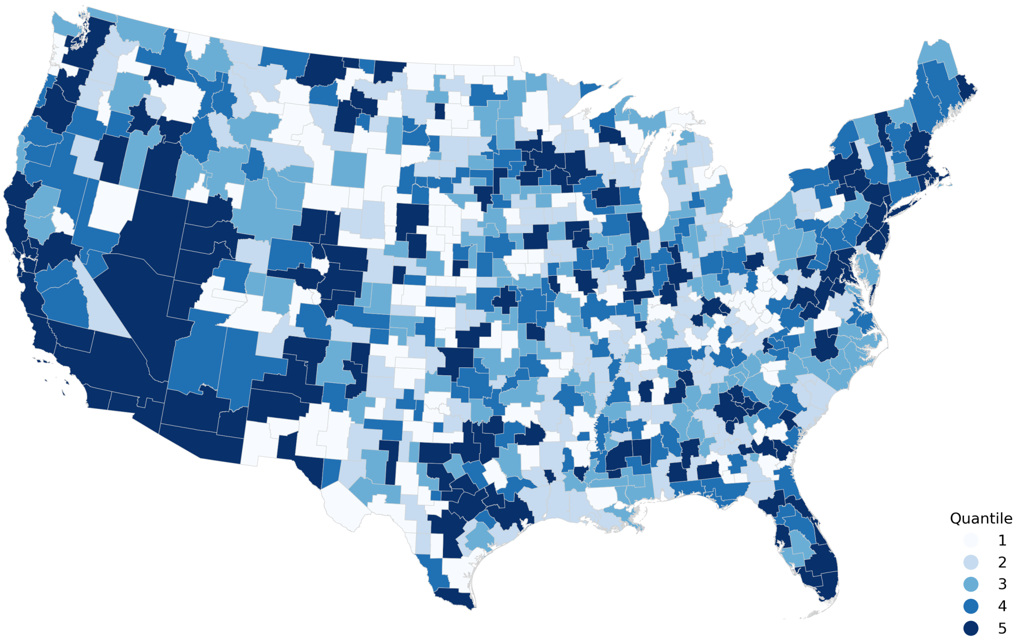
	1995	1996	...	2003	2004	$\hat{2005}$	$\hat{2006}$	...	$\hat{2013}$	$\hat{2014}$
$\hat{2005}$	d^{10}	d^9	...	d^2	d^1					
$\hat{2006}$		d^{10}	...	d^3	d^2	d^1				
...										
...										
$\hat{2013}$				d^{10}	d^9	d^8	d^7	...		
$\hat{2014}$					d^{10}	d^9	d^8	...	d^1	

Notes: Structure of the timing of the instrument. To obtain a prediction for patenting activity in 2005 (denoted by $\hat{2005}$), the coefficients of diffusion d^τ are applied to the actual patenting between 1995 and 2004. To obtain a prediction for patenting activity in 2006, the coefficients of diffusion d^τ are applied to the actual patenting between 1996 and 2004 for $\tau = 2, ..., 10$, and to predicted patenting in 2005 for $\tau = 1$. This process continues for all years up to 2014, where a prediction is obtained by applying the coefficient of diffusion d^τ to the actual patenting in 2004 for $\tau = 10$, and to predicted patenting between 2005 and 2013 for $\tau = 1, ..., 9$.

Figure A.2: **Actual and predicted patenting growth, 1990–2010**



Notes: Quantiles of actual patenting growth, 1990–2010



Notes: Quantiles of predicted patenting growth, 1990–2010

Table A.4: **Patenting growth and economic segregation: pre-trend analysis**

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: $\Delta IncSegr$, 1980-1990						
Predicted patenting growth	0.16	-0.71	-0.72*	-0.56	-0.57	-0.32
	(0.35)	(0.44)	(0.41)	(0.41)	(0.40)	(0.41)
Share of college graduates		9.95**	0.92	0.26	1.51	-0.94
		(4.57)	(4.78)	(4.55)	(6.40)	(6.30)
Log CTs			0.67**	1.58	1.58	1.75
			(0.26)	(1.03)	(1.04)	(1.10)
Log households				-0.90	-0.85	-1.08
				(1.07)	(1.09)	(1.16)
Log average income					-0.66	-0.19
					(2.11)	(2.02)
Import exposure						-0.46***
						(0.12)
R^2	0.00	0.05	0.17	0.18	0.18	0.22
Panel B: $\Delta OccSegr$, 1980-1990						
Predicted patenting growth	0.69*	0.82*	0.82	0.52	0.53	0.54
	(0.39)	(0.48)	(0.50)	(0.49)	(0.50)	(0.49)
Share of college graduates		-1.48	0.85	2.11	1.13	1.10
		(3.05)	(4.35)	(4.11)	(6.37)	(5.92)
Log CTs			-0.17	-1.90*	-1.90*	-1.90*
			(0.23)	(1.07)	(1.06)	(1.10)
Log households				1.71	1.67	1.67
				(1.13)	(1.11)	(1.16)
Log average income					0.52	0.53
					(2.02)	(1.91)
Import exposure						-0.01
						(0.25)
R^2	0.02	0.02	0.03	0.05	0.05	0.05
Panel C: $\Delta EduSegr$, 1980-1990						
Predicted patenting growth	0.29	0.42	0.43	0.09	0.07	0.08
	(0.48)	(0.59)	(0.61)	(0.63)	(0.65)	(0.63)
Share of college graduates		-1.57	1.32	2.74	5.08	4.93
		(3.58)	(5.27)	(5.32)	(7.46)	(7.30)
Log CTs			-0.22	-2.16	-2.15	-2.14
			(0.28)	(1.36)	(1.37)	(1.38)
Log households				1.93	2.02	2.01
				(1.34)	(1.33)	(1.35)
Log average income					-1.24	-1.21
					(1.76)	(1.73)
Import exposure						-0.03
						(0.20)
R^2	0.00	0.00	0.01	0.03	0.03	0.03
# Obs.	310	310	310	310	310	310

Notes: Regressions are weighted by total number of households in 1990. Controls are at 1990 values, with the exception of import exposure (provided by Autor et al., 2013). Standard errors clustered at the state level in parenthesis. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table A.5: **Patenting growth and economic segregation: Sub-sample of cities with data on 1980 economic segregation**

	$\Delta Segr$, 1980-1990			$\Delta Segr$, 1990-2010		
	Inc	Occ	Edu	Inc	Occ	Edu
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted patenting growth	-0.32	0.54	0.08			
	(0.41)	(0.49)	(0.63)			
Patenting growth				1.38***	2.95***	2.19***
				(0.63)	(0.61)	(0.74)
Baseline controls	✓	✓	✓	✓	✓	✓
Estimation	OLS	OLS	OLS	2SLS	2SLS	2SLS
# Obs.	310	310	310	310	310	310
R^2	0.22	0.05	0.03			
First stage F-stat				27.67	27.67	27.67

Notes: The sample only includes CZs for which data on economic segregation 1980 is available. Regressions are weighted by total number of households in 1990. Controls are at 1990 values, with the exception of import exposure (provided by Autor et al., 2013). Standard errors clustered at the state level in parenthesis. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

Table A.6: **Patenting growth and economic segregation: unweighted**

	$\Delta IncSegr$		$\Delta OccSegr$		$\Delta EduSegr$	
	(1)	(2)	(3)	(4)	(5)	(6)
Patenting growth	0.09 (0.24)	1.04* (0.61)	0.83** (0.37)	1.83*** (0.66)	0.80** (0.39)	1.29* (0.73)
Baseline controls	✓	✓	✓	✓	✓	✓
Estimation	OLS	2SLS	OLS	2SLS	OLS	2SLS
# Obs.	259	259	259	259	259	259
R^2	0.32		0.20		0.20	
First stage F-stat		25.95		25.95		25.95

Notes: Observations are restricted to CZs with 1990 number of households above 60,000. Controls are at 1990 values, with the exception of import exposure (provided by Autor et al., 2013). Standard errors clustered at the state level in parenthesis. $***p < 0.01$; $**p < 0.05$; $*p < 0.1$.

Table A.7: **Patenting growth and economic segregation: Inverse hyperbolic sine transformation**

	$\Delta IncSegr$		$\Delta OccSegr$		$\Delta EduSegr$	
	(1)	(2)	(3)	(4)	(5)	(6)
Patenting growth (arcsinh)	0.14 (0.29)	1.27** (0.60)	1.15** (0.43)	2.90*** (0.61)	1.15*** (0.41)	2.16*** (0.71)
Baseline controls	✓	✓	✓	✓	✓	✓
Estimation	OLS	2SLS	OLS	2SLS	OLS	2SLS
# Obs.	722	722	722	722	722	722
R^2	0.41		0.38		0.28	
First stage F-stat		34.73		34.73		34.73

Notes: Regressions are weighted by total number of households in 1990. Controls are at 1990 values, with the exception of import exposure (provided by Autor et al., 2013). Standard errors clustered at the state level in parenthesis. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

B Data description

B.1 Income distribution at the CT level

The NHGIS provides information on yearly household income at the CT level by dividing households into 15 income bins. The lower bounds of each income bin are: 0$, 15,000$, 20,000$, 25,000$, 30,000$, 35,000$, 40,000$, 45,000$, 50,000$, 60,000$, 75,000$, 100,000$, 125,000$, and 150,000$. In order to measure inequality and segregation, we need to approximate the income distribution. For each bracket except for the top one, we assume that all households in that bracket have income equal to the midpoint of the bracket. The top bin is unbounded, with an average that potentially varies substantially across CTs. Our measures will critically depend on the assumptions made to approximate the income distribution within this bracket. The literature has dealt with this issue by either fitting the parameters of an income distribution (usually assumed to be Pareto) or assuming that the average is a fixed percentage above the amount reported in top coded data (usually 40-50% more).[3] These two methods have been subject to several critics.[4]

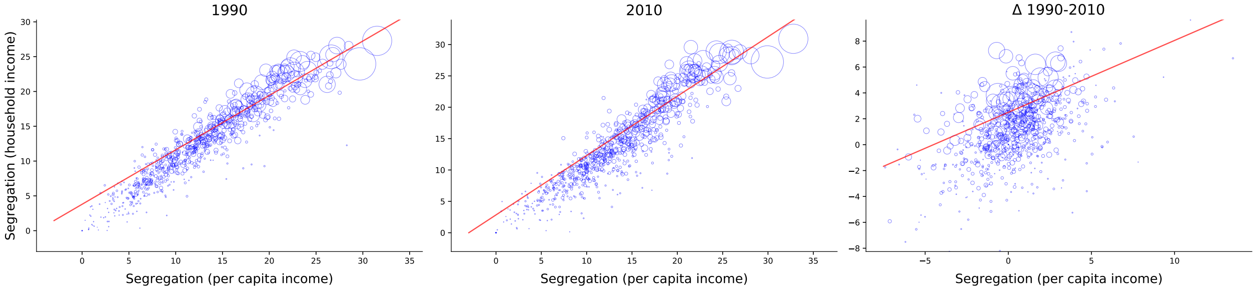
For our analysis, we design an alternative approach to assign a value to the top bin, and validate our procedure by comparing the resulting segregation index with the corresponding index we obtain by using information on average personal income, that does not require to make any assumption. First, the 5-year 2008-2012 ACS provides CT-level Gini indices using households as basic unit of analysis. For each census tract in 2010, we set the average of the top bin so that the resulting Gini matches the one reported in the ACS.[5] Second, we use the time series of individual-level Gini data at the state level computed by Frank (2009a) and provided by Frank (2009b). From there we collect estimates for the Gini index for all the states in 1990 and 2010 and calculate the percentage change. Assuming that the state trends for individual-level Gini are mirrored by the corresponding CT trends for household-level Gini, we set the average income in the top bin so that

[3] See for example Autor et al. (2008) and Lemieux (2006).

[4] Critics of the former approach have argued that if the underlying distribution is far from the assumed one, a researcher would obtain better results by taking the bin averages. Critics of the latter have pointed to the fact that the assumption of the average income for the last bin is arbitrary. Different methods to deal with binned income data have been reviewed by von Hippel et al. (2016).

[5] Note that in 3,609 out of 98,032 CTs (3.7%) there is no value that allows us to exactly match the Gini reported in the ACS. This might be due to measurement errors or the approximation that all the households earn the average of the income braket. In this case, our algorithm diverges, either assigning values that are too low (i.e., smaller than 150,000$ which is the lower bound of the top bin) or too high (i.e., bigger than 1,000,000$). When this happens we assign to the CTs in question a default value of 200,000$ which is in line with the 1.4-rule. We experimented with different default values and the main results are robust. Another 908 CTs (or 0.9%) appear in the income data but not in the Gini data. In that case, we try to match the 2010 national Gini (0.48).

Figure B.3: **Income segregation: household VS per capita income**



Notes: Scatter plots of the unconditional correlations between income segregation computed using household income and per capita income in 1990 (left panel), 2010 (middle panel), and 1990–2010 difference (right panel). Circles and regression lines are weighted by the total number of households in 1990.

the percentage change in the Gini index is equal to the one in Frank (2009a).[6]

To further validate our procedure, in Figure B.3 we show scatter plots of income segregation in 1990 (left panel), 2010 (middle panel), and 1990–2010 change (right panel), using the household income distribution approximated using the procedure described above, and the same measure computed using per capita income at the CT level, which does not require to make arbitrary assumptions on the distribution of income within brackets. The correlation between the two variables is equal to 93% in 1990 and 91% in 2010. The correlation between the 1990–2010 change in the two variables is also remarkably high (44%).

B.2 Other data sources

Residents by occupation

The distribution of residents by occupation at the CT level is constructed as follows. First, from the NHGIS we obtain information on the CT-level distribution of residents according to a coarse definition of occupations, comprising 13 occupations in 1990 and 25 occupations in 2010. Then, using IPUMS, we construct a city-specific crosswalk that maps the coarse definition of occupation into the fine one (386 occupations in 1990 and 454 in 2010). To this end, we exploit the city-specific frequency of each fine occupation code in each coarse category. We categorize occupations into two classes:[7] knowledge intensive and non-knowledge intensive. Knowledge intensive occupations are defined according to Florida (2017) definition of "creative class": "*The creative class is made up of*

[6] We are not able to match 20,966 (or 21%) of the 1990 CTs with the 2010 data. In this case, we assume that their Gini is the same as the national one in 1990 (0.43). As we did in 2010, when the algorithm diverges or estimates an implausible value, we assign to the top bin a default value of 200,000$.

[7] This categorization is available as part of the replication package of this paper.

workers in occupations spanning computer science and mathematics; architecture and engineering, the life, physical, and social sciences; the arts, design, music, entertainment, sports, and media; management, business, and finance; and law, health care, education, and training." (p. 217).

Workers by occupation

We assign workers to workplaces using the National Establishment Time Series (NETS). This data set contains information about employment for the near universe of establishments between 1990 and 2010, as well as their location and NAICS code. The latitude and longitude is provided at 5 geographical levels (namely block face, block group, census tract centroid, ZIP code centroid or street level). We allocate workers to each census tract according to the following procedure. First, we assign to a census tract those establishments whose geographical coordinates are provided at a block face, block group or census tract centroid level. Second, we assign the workers of each establishment geo-located at ZIP code level based on the area of the census tracts it contains.[8] We discard all those establishments whose coordinates are missing, more aggregated than the ZIP level, or reported as ZIP codes that do not appear in the NHGIS files (e.g., P.O. boxes). The final data set includes about 10.6 million establishments in 1990 and about 30.6 million establishments in 2010. This procedure gives us an estimate of workers per NAICS at a census tract level.

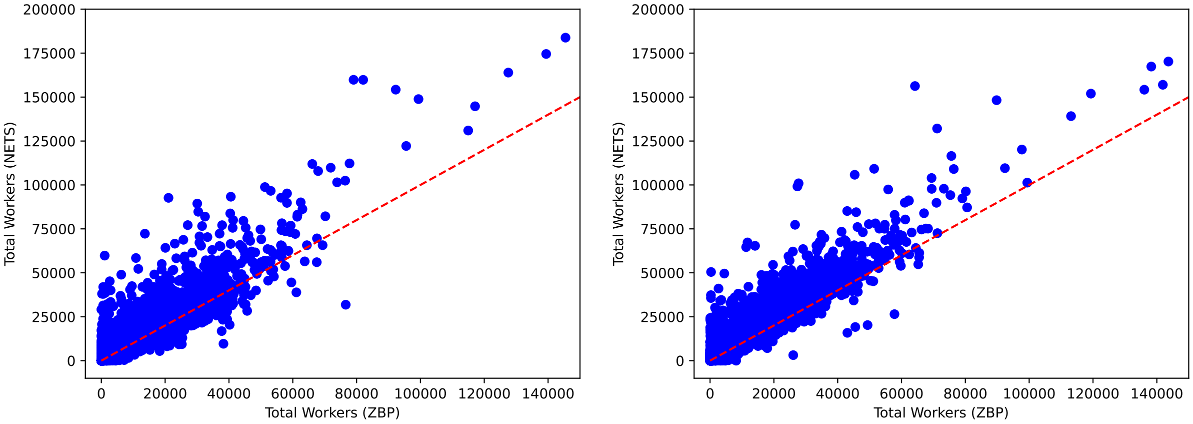
Since the NETS is a relatively new data set in the literature and there might be some concerns related to its validity, before assigning each NAICS to a distribution of occupations, we compare our employment estimates with the distribution of workers obtained from the ZIP Code Business Patterns (ZBP, U.S. Census Bureau, 2017).[9] We aggregate the employment data obtained from the NETS at the ZIP code level and we then check whether they systematically differ in the two data sets. Note that we do expect them to somewhat differ for various reasons. For example, the ZBP does not consider workers that are employed by the public sector. Therefore, the number of workers in ZIP codes that contain public universities or government buildings is likely to be significantly lower in the ZBP.[10] Figure B.4 shows the correlation between total workers by ZIP code estimated using the ZBP (horizontal axis) and the NETS (vertical axis) in 1994 (left panel)

[8] For example, if a certain ZIP code contains two census tracts that cover 40% and 60% of its area, respectively, we assign 40% of the employment of an establishment assigned to that ZIP code to the first census tract and 60% to the second one.

[9] The ZBP provides employment count by establishment in 7 bins, with the top bin including all establishments with employment above 1,000. We assign to each establishment the number of workers corresponding to the midpoint of each bin, and 1,500 workers for establishments in the top bin.

[10] Some other NAICS codes, for example agriculture, are excluded from the ZBP and the sampling frame differs in the two data sets.

Figure B.4: **Total workers: NETS VS ZBP**



Notes: Correlation between total workers by ZIP code as reported in the ZBP (horizontal axis) and in the NETS (vertical axis) in 1994 (left) and 2010 (right). The dashed red line is the 45-degree line.

and 2010 (right panel).[11] As we expected, the NETS systematically reports more workers than the ZBP, although the two measures are very close. Interestingly and in line with our prior expectations, the difference between the two employment estimates is highest in ZIP codes that contain public universities or government buildings. For example, the three largest differences in 1994 come from ZIP codes 90012, 43215, and 77002 (92,662 vs. 21,060; 159,815 vs. 82,058; and 159,847 vs. 79,047, respectively). ZIP code 90012 contains the Los Angeles City Hall as well as other government buildings (e.g., the California Department of Transportation's offices), ZIP code 43215 contains the Ohio State house, and ZIP code 77002 contains the Houston City Administration. In 1994, the NETS reports an estimate of 33,410 workers for ZIP code 43210 (UC Berkeley), whereas the ZBP of only 2,924.

We use the NETS data in conjunction with the Occupational Employment and Wage Statistics (OEWS, Bureau of Labor Statistics, 2017) to get an estimate of the occupational distribution of workers in each census tract. The OEWS reports the percentage of workers active in a certain occupation for each NAICS (SIC90 for 1990) code.[12] Similarly to what we did for residents, occupations are then assigned to either the knowledge intensive or the non-knowledge intensive category according to the procedure describe above.

[11] We used 1994 instead of 1990, since this is the first year for which the ZIP Code Business Patterns was made available.

[12] Since in the 1990s only certain industry codes were reported in each year, we build the crosswalk for 1990 using OEWS data from 1990 to 1993. Also, since the data are provided for SIC (instead of NAICS) codes, we first use a crosswalk from NAICS to SIC (NAICS Association, 2012) and we then use the appropriate distributions reported in the OEWS.

Local consumption amenities

The establishment count for the three categories of consumption amenities used in Section 4.3 is also computed from NETS. The "Restaurants" category includes establishments from NAICS 722511 ("Full-Service Restaurants"). The "Food Shops" category includes establishments from all the NAICS in 4452 ("Specialty Food Stores"). The "Fitness Centers" category includes establishments from NAICS 71394 ("Fitness and Recreational Sports Centers").

Commuting times

Commuting times between each pair of CTs are calculated using driving times between the centroids of each census tract. Because of the high number of possible combinations we were unable to use commercial routing services (e.g., Google Maps) and we relied on the Open Source Routing Machine (OSRM).[13] The advanatage of using the OSRM is that it is possible to run it locally. This allows us to send queries without limits and in parallel. In particular, it was possible to collect data on commuting times for each pair of neighborhoods withing each city (for a total of almost 19.4 million pairs) in just few hours. The disadvantage is that the OSRM does not contain any data on traffic which might underestimate the actual commuting times/costs faced by workers, particularly during rush hour.[14]

[13] http://project-osrm.org/

[14] Note that, because of the lack of traffic data, commuting times are undirected, that is the time necessary to go from A to B and from B to A is the same. The commuting matrices are therefore symmetric and overall contain more than 38.8 million values.

C Citation Network: Demand Pull or Supply Push?

In Section 3.2.1, we developed an instrument for local patenting activity that exploits a predetermined network of knowledge flows. Our instrument is valid as long as these knowledge links are determined by factors that are orthogonal to the local future economic activity. A possible concern that would invalidate our identification strategy is that the channels captured through the network of citations reflect demand instead of supply links. This would be problematic for the validity of the model, since demand links are likely to be informative about the state of the local economy. To fix ideas, suppose that an IT firm in San Jose supplies innovation to a car manufacturer in Detroit under commission. In this case, our network would record a strong link from San Jose to Detroit, but the associated knowledge flows would violate the orthogonality assumption, since demand from Detroit is likely to be correlated with unobservable factors in Detroit.

The structure of our network and the long time series of patents data can be used to test for the presence of demand-driven links. Formally, we proceed in three steps. First, we use the knowledge network defined in Section 3.2.1 and the observed patenting activity in the period 1985-1994 to get a forward estimate of the patenting activity between 1995 and 2004. Second, we reverse the network and use the patents filed between 2005-2014 to get an upstream estimate of the patenting activity we expect to observe in the period 1995-2004 if the citations were capturing demand links. The reversed network closely mirrors the one defined in (5), but exploits citations received instead of citations given:

$$o^{\tau}_{(b,\mu)\to(c,\nu)} = \begin{cases} \dfrac{\sum\limits_{p\in\mathcal{P}(b,\mu)} \widetilde{ShareCit}^{\tau}_{p\to(c,\nu)}}{\widetilde{TotPat}^{\tau}_{c,\nu}} & b\neq c \\ 0 & b=c \end{cases} \qquad \text{for } \tau\in\{1,\dots,10\}\,, \tag{C.1}$$

where $\widetilde{ShareCit}^{\tau}_{p\to(c,\nu)}$ denotes the share of citations *received* by patent p from patents from class ν and commuting zone c filed τ years after p, and $\widetilde{TotPat}^{\tau}_{c,\nu}$ denotes all the potential destination patents in (c,ν) at diffusion lag τ. The coefficient $o^{\tau}_{(b,\mu)\to(c,\nu)}$ represents the number of patents of class μ in commuting zone b that we expect to observe upstream if τ years later we observe one patent of class ν in commuting zone c downstream. We then compare the two models (the one based on citations given and the one based on citations received) to see which one offers the most accurate description of the innovation process. To do this, we follow Acemoglu et al. (2016) and regress the actual 1995-2004 patenting activity on the patenting activity predicted by the two

Table C.8: **Predicting patents with "supply" and "demand" links**

	$\log(1 + P^{actual}_{95-04,\,c})$
$\log(1 + \hat{P}^{up}_{95-04,\,c})$	0.771***
	(0.105)
$\log(1 + \hat{P}^{down}_{95-04,\,c})$	-0.153***
	(0.045)
$\log(1 + P^{actual}_{85-94,\,c})$	0.370***
	(0.101)
R^2	0.984
# Obs.	722

Notes: $\hat{P}^{up}_{95-04,\,c}$ is total 1995-2004 patents predicted with the "supply" links of Equation (5). $\hat{P}^{down}_{95-04,\,c}$ is total 1995-2004 patents predicted with the "demand" links of Equation (C.1). Standard errors clustered at the state level in parenthesis. $***p < 0.01$; $**p < 0.05$; $*p < 0.1$.

procedures, and controlling for actual patenting in 1985-1994:

$$\log(1 + P^{actual}_{95-04,\,c}) = \alpha + \beta \log(1 + \hat{P}^{up}_{95-04,\,c}) + \gamma \log(1 + \hat{P}^{down}_{95-04,\,c}) + \delta \log(1 + P^{actual}_{85-94,\,c}) + \epsilon_c,$$

where $\hat{P}^{up}_{95-04,\,c}$ is patenting activity predicted by the model in Equation (5), whereas $\hat{P}^{down}_{95-04,\,c}$ is the patenting activity predicted by the model in Equation (C.1). The results, reported in Table C.8, show that only "supply" links ($\hat{P}^{up}_{95-04,\,c}$) have a strong predictive power, while "demand" links ($\hat{P}^{down}_{95-04,\,c}$) have a small negative impact on actual patenting. The sign and magnitude of the estimates are consistent with the ones obtained by Acemoglu et al. (2016), who consider an analogous setting but use the network of citations to predict patenting growth across technological fields and do not consider its geographical dimension.

D Stability of citation network

In this Section, we perform a comparison between the citation network in the early sample and its counterpart in the late sample to verify that the channels of knowledge diffusion inferred from the citations patterns are, at least to some extent, stable over time. We do this in three steps. First, we build the network of citations and compute the coefficients of diffusion separately for the two samples (1975-1994 and 1995-2014). For each $\tau = 1, ..., 10$, we take the difference of the two adjacency matrices and calculate its Frobenius norm as follows:

$$real_\tau = \left\| \mathbf{D}_\tau^{75-94} - \mathbf{D}_\tau^{95-14} \right\|_2 .$$

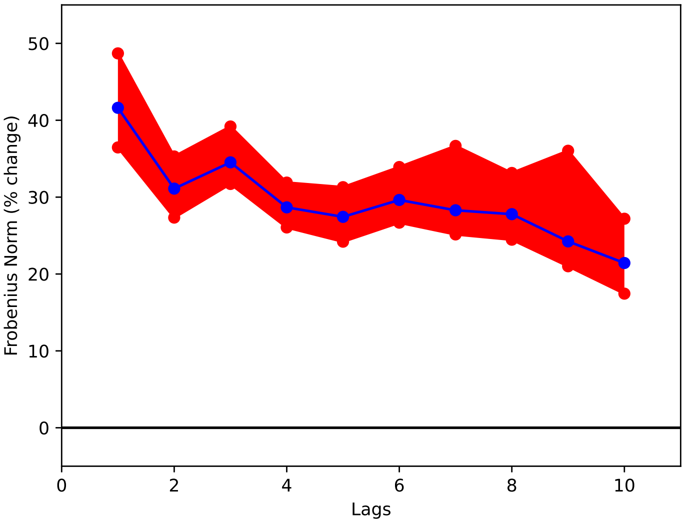
Second, for each year between 1975 and 2014, we reshuffle all the patents filed in that year under the constraint that after the reshuffling each commuting zone is assigned the same amount of patents as in the real dataset. We repeat the same exercise performed in the first step for this new sample of patents and calculate

$$random_\tau = \left\| \tilde{\mathbf{D}}_\tau^{75-94} - \tilde{\mathbf{D}}_\tau^{95-14} \right\|_2 ,$$

where $\tilde{\mathbf{D}}_\tau^{75-94}$ and $\tilde{\mathbf{D}}_\tau^{95-14}$ are the citation networks built using the reshuffled patents.

Finally, we calculate the percentage difference between $real_\tau$ and $random_\tau$ for each $\tau = 1, ..., 10$. This number captures the distance between the two actual networks (75-94 and 95-14) compared to two networks that, while maintaining the same structure and properties of the original ones, are by construction uninformative of each other. A positive value indicates that the networks built using the actual data are more similar than the reshuffled ones. Figure D.5 plots the percentage difference for all the values of τ together with the 95% confidence interval we obtained by repeating this procedure 50 times. The difference of the random networks is more than 40% larger than the one obtained with the actual networks for the first lag and it declines for larger lags. The decline implies that the more years pass after a new idea is generated the less the citation patterns are distinguishable from links that are distributed across cities at random. This result is intuitive. With time a new technology becomes widespread and is embedded in patents produced in areas that do not have any direct link with the origin city-class pair.

Figure D.5: **Proximity of 75-94 and 95-14 random relative to actual citation networks**



Notes: Percentage difference of the Frobenius norms $random_\tau$ and $real_\tau$ for $\tau = 1, ..., 10$.

References

Acemoglu, D., U. Akcigit, and W. R. Kerr (2016): "Innovation network," *Proceedings of the National Academy of Sciences*, 113, 11483–11488.

Autor, D., D. Dorn, and G. H. Hanson (2013): "Replication data for: The China Syndrome: Local Labor Market Effects of Import Competition in the United States," American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E112670V1.

Autor, D. H., L. F. Katz, and M. S. Kearney (2008): "Trends in US wage inequality: Revising the revisionists," *The Review of economics and statistics*, 90, 300–323.

Bureau of Labor Statistics (2017): "Occupational Employment and Wage Statistics (1988–2010) [tables]," https://www.bls.gov/oes/.

Florida, R. (2017): *The new urban crisis: How our cities are increasing inequality, deepening segregation, and failing the middle class-and what we can do about it*, Basic Books New York.

Frank, M. W. (2009a): "Inequality and growth in the United States: Evidence from a new state-level panel of income inequality measures," *Economic Inquiry*, 47, 55–68.

——— (2009b): "U.S. State-Level Income Inequality Data," https://www.shsu.edu/eco_mwf/inequality.html. Accessed December 2016.

Lemieux, T. (2006): "Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?" *American Economic Review*, 96, 461–498.

NAICS Association (2012): "NAICS to SIC Crosswalk," https://www.naics.com/naics-to-sic-crosswalk-2/. Accessed November 2015.

U.S. Census Bureau (2017): "ZIP Codes Business Patterns (1994–2010) [tables]," https://www.census.gov/data/developers/data-sets/cbp-nonemp-zbp/zbp-api.html.

von Hippel, P. T., S. V. Scarpino, and I. Holas (2016): "Robust estimation of inequality from binned incomes," *Sociological Methodology*, 46, 212–251.