APPENDIX

A.1 Variable Definitions

In this section, we detail the construction of relevant variables for our analysis. All state-level variables have analogous definitions at the county-level for our border-county analysis.

- *Top Corporate Marginal Tax Rate (MTR)* - The additional tax burden accruing to a firm in the top tax bracket in state s for an additional one dollar of revenue if all of its operations were in s. In firm-level regressions (Table 16), we assign firms the average corporate tax in states in which the firm operates an R&D lab, weighted by the share of labs in that state.

- *90th Percentile Income Marginal Tax Rate (MTR)* - The additional tax burden accruing to an individual at the 90th percentile of the national income distribution for an additional one dollar of earnings. Calculated using the tax calculator by Bakija (2017).

- *90th Percentile Income Average Tax Rate (ATR)* - The total tax burden for an individual at the 90th percentile of the national income distribution divided by that individual's total income. Calculated using the tax calculator by Bakija (2017).

- *Median Income Marginal Tax Rate (MTR)* - The additional tax burden accruing to an individual at the 50th percentile of the national income distribution for an additional one dollar of earnings. Calculated using the tax calculator by Bakija (2017).

- *Median Income Average Tax Rate (ATR)* - The total tax burden for an individual at the 50th percentile of the national income distribution divided by that individual's total income. Calculated using the tax calculator by Bakija (2017).

- *Inventor productivity* - An inventor's productivity in year t is defined to be the number of eventually-granted patents that the inventor has applied for as of year $t-1$. In robustness table A7, inventor i's productivity in year t is defined to be the total number of citations ever received by patents applied for by i through year t. An inventor is said to be "high productivity" in year t if he/she is in the top 10% of the national inventor productivity distribution in year t. In robustness table A4, an inventor is said to be high productivity if he/she is in the top 5% of the national productivity distribution in year t. In robustness table A6, an inventor is said to be high productivity if he/she is *ever* in the top 10% of the national productivity distribution in a single year. Finally, robustness table A8 allows an inventor to be high productivity if he/she is in the top 10% of the productivity distribution, of middle productivity if he/she is between the 75th and 90th percentile of the productivity distribution, and low productivity otherwise.

- *Effective Tax Rates* - An inventor's effective marginal (average) tax rate is defined to be the marginal (average) tax rate faced by the 90^{th} percentile earner in the national income distribution if the inventor is high productivity, and the marginal (average) tax rate faced by a median earner if the inventor is low productivity. In appendix table A8, middle productivity inventors have an effective tax rate equal to the tax rate faced by an individual earning at the 75^{th} percentile of the national income distribution. In all regressions, we use lagged effective tax rates as independent variables. Thus an inventor living in state s will face an effective tax rate for innovation output in year t which is the effective tax rate the inventor would have faced in year $t-1$ given his/her $t-1$ productivity level and the tax laws in place in year $t-1$.

- *Log Patents* - The natural logarithm of the number of eventually-granted patents applied for in state s in year t. Similarly, in firm regressions (Table 16), Log Patents refers to the natural logarithm of the number of successful patent applications for firm j in year t.

- *Log Citations* - The natural logarithm of the number of citations ever received by eventually-granted patents which were applied for in state s in year t. Similarly, in firm regressions (Table 16), Log Citations refers to the natural logarithm of the number of citations ever received by eventually-granted patents which were applied for by firm j in year t. Citation counts adjusted according to the algorithm of Hall et al. (2001), detailed for our data in Akcigit et al. (2017) Appendix B.1.

- *Log Inventors* - The natural logarithm of number of inventors in state s in year t as implied by the Lai et al. (2014) algorithm applied to our dataset. A detailed description of this algorithm is provided in Appendix OA.1.

- *Log Superstars* - The natural logarithm of the number of inventors in state s in year t who are in the top 5% of the national inventor productivity distribution.

- *Corporate Patent* - A corporate patent is one which is assigned to a corporation after being granted.

- *Share Assigned* - The share of patents in state s in year t which are assigned to a corporation.

- *Log Patents (3-year)* - The log of the number of eventually-granted patents applied for by inventor i years t through $t+2$.

- *Log Citations (3-year)* - The log of the number of citations ever received by eventually-granted patents which were applied for by inventor i years t through $t+2$. Citation counts adjusted according to the algorithm of Hall et al. (2001), detailed for our data in Akcigit et al. (2017) Appendix B.1.

- *Has Patent (3-year)* - An indicator variable, equal to 100 (for legibility) if the inventor has at least one successful patent application between years t and $t+2$. Inventors are included in the regression sample for the period between their first ever successful patent application, and their last ever successful patent application.

- *Has 10+ Cites (3-year)* - An indicator variable, equal to 100 (for legibility) if the inventor's patents, applied for between years t and $t+2$, ever receive at least 10 citations in total between them. Inventors are included in the regression sample for the period between their first ever successful patent application, and their last ever successful patent application. Patent citation counts adjusted according to the algorithm of Hall et al. (2001), detailed for our data in Akcigit et al. (2017) Appendix B.1.

- *Has Corporate Patent (3-yr)* - An indicator variable, equal to 100 (for legibility) if the inventor successfully applies for at least one patent, which is assigned to a corporation, between years t and $t+2$. Inventors are included in the regression sample for the period between their first ever successful patent application, and their last ever successful patent application.

- *Corporate Inventor* - An inventor is said to be a corporate inventor if he/she is granted at least one corporate patent in his/her career.

- *# of Research Workers* - The number of research workers employed by the firm as stated on the National Research Council (NRC) Surveys of *Industrial Research Laboratories of the United States* (IRLUS).

- *Agglomeration* - The number of patents, in thousands, applied for by inventors $j \neq i$ who share inventor i's modal patent class in year t in state s.

- *Mover* - An inventor is said to be a mover if he/she applies for patents in at least two states over the sample period. Analagously, non-movers are those inventors who only apply for patents in one state over the entire course of their career.

- *Home State* - The state in which an inventor first applies for a patent.

- *Assignee Has Patent in Destination* - An indicator variable equal to one if an inventor i's firm has at least one patent applied for in year t by an inventor $j \neq i$ in destination state s.

- *Inventor Tenure/Experience* - an inventor's tenure is the number of years that have passed since the inventor's first successful patent application.

A.2 Additional Tables and Figures

Table A1: Macro Effects of Taxes: State Level Regressions, excluding Federal Taxes

	Log	Log	Log	Log	Citations/	Share
	Patents	Citations	Inventors	Superstars	Patent	Assigned
	(1)	(2)	(3)	(4)	(5)	(6)
90th Pctile Income MTR	-0.026***	-0.021***	-0.027***	-0.028***	0.188***	-0.018
	(0.002)	(0.003)	(0.002)	(0.004)	(0.042)	(0.065)
Top Corporate MTR	-0.053***	-0.056***	-0.045***	-0.077***	-0.231***	-0.799***
	(0.008)	(0.009)	(0.007)	(0.011)	(0.056)	(0.142)
Median Income MTR	-0.042***	-0.041***	-0.042***	-0.056***	0.087	-0.085
	(0.004)	(0.004)	(0.004)	(0.006)	(0.053)	(0.090)
Top Corporate MTR	-0.055***	-0.056***	-0.048***	-0.077***	-0.176***	-0.790***
	(0.008)	(0.009)	(0.007)	(0.011)	(0.048)	(0.135)
90th Pctile Income ATR	-0.047***	-0.040***	-0.046***	-0.060***	0.232***	-0.011
	(0.003)	(0.004)	(0.003)	(0.006)	(0.056)	(0.093)
Top Corporate MTR	-0.053***	-0.055***	-0.046***	-0.074***	-0.210***	-0.803***
	(0.008)	(0.009)	(0.007)	(0.011)	(0.051)	(0.139)
Median Income ATR	-0.095***	-0.102***	-0.086***	-0.144***	-0.541***	-0.691***
	(0.008)	(0.010)	(0.007)	(0.010)	(0.129)	(0.143)
Top Corporate MTR	-0.054***	-0.054***	-0.047***	-0.074***	-0.108**	-0.739***
	(0.008)	(0.008)	(0.007)	(0.010)	(0.043)	(0.125)
Observations	2867	2867	2867	2661	2867	2867
Mean of Dep. Var.	7.18	9.87	7.31	4.37	17.68	71.74
S.D. of Dep. Var.	1.31	1.59	1.33	1.60	12.48	14.01

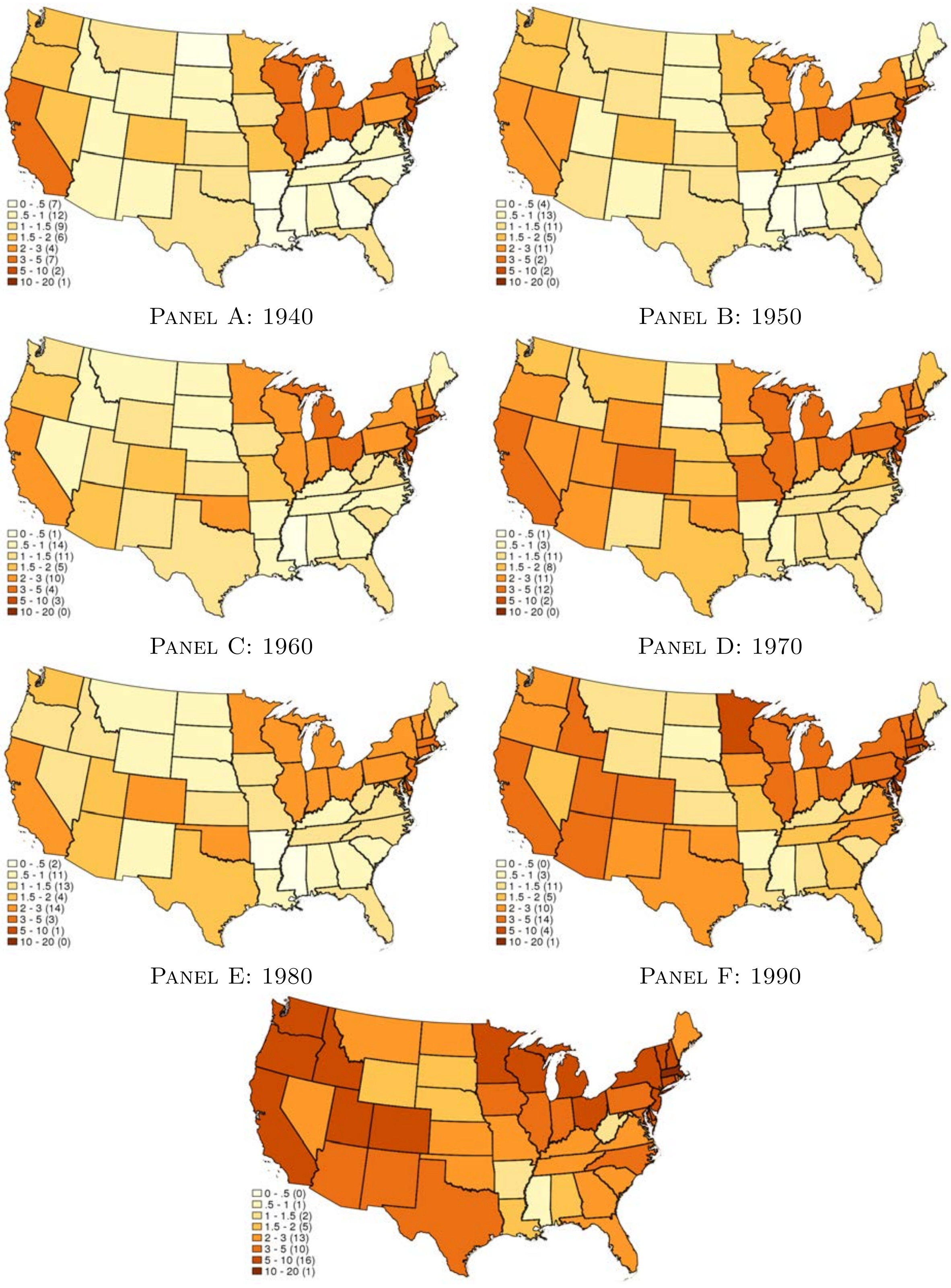
Notes: The period covered is 1940-2000. White heteroskedasticity robust standard errors clustered at year level reported in parentheses. All regressions include controls for lagged population density, real GDP per capita, and R&D tax credits, as well as state and year fixed effects. Tax rates measured in percentage points and lagged by 1 year. Only state tax rates included in tax measures.

Table A2: The Macro Effect of Taxation: Estimates from State Level Regressions, excluding California

	Log Patents (1)	Log Citations (2)	Log Inventors (3)	Log Superstars (4)	Citations/ Patent (5)	Share Assigned (6)
90th Pctile Income MTR	-0.055***	-0.052***	-0.053***	-0.063***	-0.001	-0.390***
	(0.005)	(0.006)	(0.005)	(0.008)	(0.058)	(0.072)
Top Corporate MTR	-0.056***	-0.050***	-0.045***	-0.083***	0.156**	-1.007***
	(0.006)	(0.007)	(0.006)	(0.010)	(0.066)	(0.146)
Median Income MTR	-0.076***	-0.083***	-0.073***	-0.097***	-0.247***	-0.337***
	(0.005)	(0.006)	(0.005)	(0.006)	(0.080)	(0.086)
Top Corporate MTR	-0.050***	-0.041***	-0.039***	-0.073***	0.233***	-1.026***
	(0.007)	(0.008)	(0.006)	(0.011)	(0.072)	(0.156)
90th Pctile Income ATR	-0.107***	-0.111***	-0.102***	-0.133***	-0.248**	-0.453***
	(0.006)	(0.008)	(0.005)	(0.009)	(0.101)	(0.105)
Top Corporate MTR	-0.041***	-0.033***	-0.031***	-0.064***	0.230***	-0.995***
	(0.006)	(0.007)	(0.006)	(0.010)	(0.078)	(0.155)
Median Income ATR	-0.099***	-0.105***	-0.090***	-0.150***	-0.435***	-0.585***
	(0.007)	(0.010)	(0.007)	(0.010)	(0.126)	(0.144)
Top Corporate MTR	-0.057***	-0.049***	-0.047***	-0.080***	0.227***	-1.035***
	(0.007)	(0.007)	(0.006)	(0.011)	(0.073)	(0.155)
Observations	2806	2806	2806	2600	2806	2806
Mean of Dep. Var.	6.99	9.64	7.12	4.19	16.86	71.97
S.D. of Dep. Var.	1.24	1.48	1.25	1.56	11.51	14.32

Notes: White heteroskedasticity robust standard errors clustered at year level reported in parentheses. All regressions include controls for lagged population density, real GDP per capita, and R&D tax credits, as well as state and year fixed effects. Tax rates measured in percentage points and lagged by 1 year. Regressions weighted by state-year level population counts.

Figure A1: Inventors per Capita Over Time



Panel A: 1940

Panel B: 1950

Panel C: 1960

Panel D: 1970

Panel E: 1980

Panel F: 1990

Panel G: 2000

Table A3: Macro Effects of Taxes: Excluding Movers (OLS)

	Log	Log	Log	Citations/	Share
	Patents	Citations	Inventor	Patent	Assigned
	(1)	(2)	(3)	(4)	(5)
90th Pctile Income MTR	-0.042***	-0.043***	-0.041***	-0.108**	-0.433***
	(0.005)	(0.005)	(0.004)	(0.051)	(0.074)
Top Corporate MTR	-0.061***	-0.063***	-0.050***	-0.157**	-1.091***
	(0.006)	(0.008)	(0.006)	(0.061)	(0.148)
Median Income MTR	-0.045***	-0.044***	-0.045***	0.038	-0.195**
	(0.004)	(0.005)	(0.004)	(0.047)	(0.083)
Top Corporate MTR	-0.062***	-0.064***	-0.050***	-0.207***	-1.183***
	(0.008)	(0.009)	(0.007)	(0.068)	(0.162)
90th Pctile Income ATR	-0.064***	-0.060***	-0.062***	0.079	-0.321***
	(0.004)	(0.005)	(0.004)	(0.053)	(0.094)
Top Corporate MTR	-0.056***	-0.059***	-0.044***	-0.220***	-1.144***
	(0.007)	(0.008)	(0.006)	(0.064)	(0.161)
Median Income ATR	-0.095***	-0.103***	-0.088***	-0.554***	-0.905***
	(0.007)	(0.010)	(0.007)	(0.126)	(0.141)
Top Corporate MTR	-0.060***	-0.061***	-0.049***	-0.102	-1.092***
	(0.007)	(0.008)	(0.006)	(0.061)	(0.151)
Observations	2867	2867	2867	2867	2867
Mean of Dep. Var.	6.90	9.56	7.11	16.85	68.40
S.D. of Dep. Var.	1.30	1.57	1.32	11.31	14.66

Notes: White heteroskedasticity robust standard errors clustered at year level reported in parentheses. All regressions include controls for lagged population density, real GDP per capita, and R&D tax credits, as well as state and year fixed effects. Tax rates measured in percentage points and lagged by 1 year. Regressions weighted by state-year level population counts.

Table A4: Micro Regressions: High Productivity Cutoff at Top 5%

Dependent Variable:	Has Patent (3-year)	Has 10+ Cites (3-year)	Log Patents (3-year)	Log Citations (3-year)	Has Corporate Patent (3-yr)
	(1)	(2)	(3)	(4)	(5)
Effective MTR	-0.625***	-0.624***	-0.013***	-0.017***	-0.674***
	(0.092)	(0.103)	(0.003)	(0.003)	(0.081)
Top Corporate MTR	-0.064	-0.076	-0.002	-0.003	-0.047
	(0.050)	(0.058)	(0.001)	(0.002)	(0.047)
State FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Effective MTR	-0.618***	-0.600***	-0.011***	-0.012***	-0.656***
	(0.104)	(0.117)	(0.004)	(0.004)	(0.092)
State × Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Observations	5964243	5964243	4550168	4396954	5964243
Mean of Dep. Var.	76.291	45.069	0.442	2.758	61.399
S.D. of Dep. Var.	42.530	49.756	0.664	1.453	48.683

Notes: Standard errors clustered at year level reported in parentheses. Only state-years undergoing progressive spells included. All tax rates on percentage point scale, and lagged by one year. Effective taxes defined as the marginal tax rate faced by the 90^{th} percentile earner in state s in year t for high productivity inventors, and the marginal tax rate rate faced by the median earner for low productivity inventors. Inventor productivity defined as being in the top 5% of dynamic patent counts. Regressions with state and year fixed effects include controls for lagged real state GDP per capita, population density, and a quadratic in inventor tenure. All regressions include controls for inventor productivity, and a local agglomeration force, measured as the number of patents applied for in the inventor's modal class in state s in year $t-1$ by other residents in the state.

Table A5: Micro Regressions: Only Including States with Progressive Tax Spells

Dependent Variable:	Has Patent (3-year)	Has 10+ Cites (3-year)	Log Patents (3-year)	Log Citations (3-year)	Has Corporate Patent (3-yr)
	(1)	(2)	(3)	(4)	(5)
Effective MTR	-0.390***	-0.453***	-0.010***	-0.014***	-0.501***
	(0.086)	(0.100)	(0.003)	(0.003)	(0.068)
Top Corporate MTR	-0.023	0.064	-0.004	0.001	0.020
	(0.117)	(0.114)	(0.003)	(0.004)	(0.120)
State FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Effective MTR	-0.425***	-0.449***	-0.008***	-0.011***	-0.509***
	(0.090)	(0.102)	(0.003)	(0.003)	(0.069)
State × Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Observations	2759975	2759975	2095175	2027602	2759975
Mean of Dep. Var.	75.913	45.713	0.447	2.814	59.769
S.D. of Dep. Var.	42.761	49.816	0.672	1.484	49.036

Notes: Standard errors clustered at year level reported in parentheses. Only state-years undergoing progressive spells included. All tax rates on percentage point scale, and lagged by one year. Effective taxes defined as the marginal tax rate faced by the 90^{th} percentile earner in state s in year t for high productivity inventors, and the marginal tax rate rate faced by the median earner for low productivity inventors. Inventor productivity defined as being in the top 10% of dynamic patent counts. Regressions with state and year fixed effects include controls for lagged real state GDP per capita, population density, and a quadratic in inventor tenure. All regressions include controls for inventor productivity, and a local agglomeration force, measured as the number of patents applied for in the inventor's modal class in state s in year $t-1$ by other residents in the state.

Table A6: Micro Regression Coefficients, Using Static Measure of Inventor productivity

Dependent Variable:	Has Patent (3-year) (1)	Has 10+ Cites (3-year) (2)	Log Patents (3-year) (3)	Log Citations (3-year) (4)	Has Corporate Patent (3-yr) (5)
Effective MTR	-0.835***	-0.532***	-0.012***	-0.013***	-0.559***
	(0.153)	(0.101)	(0.003)	(0.004)	(0.098)
Top Corporate MTR	-0.131	-0.109	-0.001	-0.001	-0.095
	(0.114)	(0.102)	(0.002)	(0.003)	(0.096)
State FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Effective MTR	-0.856***	-0.473***	-0.010**	-0.010	-0.512***
	(0.172)	(0.127)	(0.004)	(0.006)	(0.116)
State × Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Observations	5960430	5960430	4548136	4394979	5960430
Mean of Dep. Var.	76.306	45.078	0.442	2.758	61.407
S.D. of Dep. Var.	42.521	49.757	0.664	1.454	48.681

Notes: Standard errors clustered at year level reported in parentheses. All mainland states, excluding Louisiana, included for the period 1940-2000. All tax rates on percentage point scale, and lagged by one year. Inventor productivity defined as ever being in the top 10% of dynamic patent counts. Regressions with state and year fixed effects include controls for lagged real state GDP per capita, population density, and a quadratic in inventor tenure. All regressions control for local agglomeration forces, measured as the number of patents applied for in the inventor's modal class in state s in year $t-1$ by other residents in the state.

Table A7: Micro Regression Coefficients: Productivity Measure: Dynamic Citation Counts

Dependent Variable:	Has Patent (3-year)	Has 10+ Cites (3-year)	Log Patents (3-year)	Log Citations (3-year)	Has Corporate Patent (3-yr)
	(1)	(2)	(3)	(4)	(5)
Effective MTR	-0.467***	-0.458***	-0.009***	-0.015***	-0.515***
	(0.097)	(0.121)	(0.003)	(0.003)	(0.076)
Top Corporate MTR	-0.243**	-0.144	-0.003*	-0.001	-0.131
	(0.101)	(0.100)	(0.002)	(0.002)	(0.091)
State FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Effective MTR	-0.425***	-0.394***	-0.007**	-0.012***	-0.456***
	(0.101)	(0.131)	(0.003)	(0.003)	(0.079)
State × Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Observations	5960430	5960430	4548136	4394979	5960430
Mean of Dep. Var.	76.306	45.078	0.442	2.758	61.407
S.D. of Dep. Var.	42.521	49.757	0.664	1.454	48.681

Notes: Standard errors clustered at year level reported in parentheses. All mainland states, excluding Louisiana, included for the period 1940-2000. All tax rates on percentage point scale, and lagged by one year. Effective taxes defined as the marginal tax rate faced by the 90^{th} percentile earner in state s in year t for high productivity inventors, and the marginal tax rate rate faced by the median earner for low productivity inventors. Inventor productivity defined as being in the top 10% of dynamic citation counts. Regressions with state and year fixed effects include controls for lagged real state GDP per capita, population density, and a quadratic in inventor tenure. All regressions include controls for inventor productivity, and a local agglomeration force, measured as the number of patents applied for in the inventor's modal class in state s in year $t-1$ by other residents in the state.

Table A8: Micro Regression Coefficients: Three Productivity Cutoffs

Dependent Variable:	Has Patent (3-year)	Has 10+ Cites (3-year)	Log Patents (3-year)	Log Citations (3-year)	Has Corporate Patent (3-yr)
	(1)	(2)	(3)	(4)	(5)
Effective MTR	-0.629***	-0.563***	-0.011***	-0.016***	-0.587***
	(0.102)	(0.072)	(0.001)	(0.001)	(0.072)
Top Corporate MTR	-0.202**	-0.104	-0.002*	-0.002	-0.059
	(0.090)	(0.074)	(0.001)	(0.002)	(0.078)
State FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Effective MTR	-0.791***	-0.661***	-0.012***	-0.017***	-0.682***
	(0.130)	(0.090)	(0.002)	(0.002)	(0.089)
State × Year FE	Y	Y	Y	Y	Y
Inventor FE	Y	Y	Y	Y	Y
Observations	3940182	3940182	2541638	2465129	3940182
Mean of Dep. Var.	64.644	41.771	0.614	2.972	54.645
S.D. of Dep. Var.	47.807	49.318	0.742	1.505	49.784

Notes: Standard errors clustered at year level reported in parentheses. All mainland states, excluding Louisiana, included for the period 1940-2000. All tax rates on percentage point scale, and lagged by one year. Effective taxes defined as the marginal tax rate faced by the 90^{th} percentile earner in state s in year t for high productivity inventors, the rate faced by the 75^{th} percentile earner for mid-productivity inventors, and the marginal tax rate rate faced by the median earner for low productivity inventors. Inventors are said to be high, or middle productivity if they are above the 10^{th}, or 25^{th} percentiles of dynamic patent counts, and low productivity otherwise. Regressions with state and year fixed effects include controls for lagged real state GDP per capita, population density, and a quadratic in inventor tenure. All regressions include controls for inventor productivity, and a local agglomeration force, measured as the number of patents applied for in the inventor's modal class in state s in year $t-1$ by other residents in the state.

ONLINE APPENDIX – NOT FOR PUBLICATION

for "The Effects of Taxes on Innovation: Evidence from Historical U.S. Patent Data"

by Ufuk Akcigit, John Grigsby, Tom Nicholas, and Stefanie Stantcheva

OA.1 Disambiguation Algorithm

We employ the algorithm of Lai et al. (2014) to disambiguate inventors in our historical patent data.[19] The goal of disambiguation is to determine if two patent-inventor level records were produced by the same inventor. A problem of this sort may be distilled into a clustering problem well-suited to standard machine learning algorithms: given a training dataset and a set of features – such as inventor name, location, technology class, assignee, and coauthor networks – we wish to group records together into profiles which indicate that the two records were produced by the same inventor. The goal is to assign probabilities of an inventor match based on the characteristics of every pair of observations. The central idea is that two records coming from two very similar names (not necessarily identical: "John A Smith" vs "John Adam Smith" for instance) working in similar subject areas, working for the same company in roughly the same geographic location, are likely to be the same person.

Such a machine learning approach has three central benefits relative to other more rudimentary approaches, such as treating each individual name as a separate inventor, or hand-matching innovators' records to one another. First, the Lai et al. approach permits minor name typos or data entry errors, without incorrectly decoupling these inventors. Second, it provides probabilistic matches based on more information than name and location, which helps disambiguate between common names – a John Smith working in software is likely different to a John Smith with patents in bootmaking. Finally, the algorithm does not impose any functional forms on the relationship between a pair's set of attributes and the probability that those pairs belong to the same inventor.

Of course, this machine learning approach is imperfect and will struggle to correctly match inventors who drastically change their names or have exceptional careers. For instance, if an inventor named Jane Smith changes her name after marrying a man with surname Robertson, the algorithm will struggle to adapt, as names are the most distinguishing piece of information amongst records. Similarly, if a software engineer living in California and working for Apple decides to change his career and move to Montana to open a new shoe factory, the algorithm is likely to suggest that these are two separate inventors, rather than one inventor with such an uncommon career trajectory.

The clustering exercise is subject to two principal challenges. First, one must produce a suitable

[19] The code and associated files for the original disambiguation may be downloaded from https://github.com/funginstitute/downloads; accessed October 13, 2016.

training dataset from which to glean the probability that two patent records with a similarity profile of x belong to the same inventor. Here, one may follow two approaches. One could submit a hand-curated dataset of known matches to the disambiguation algorithm to determine the likelihood of a match. However, the construction of these datasets are often subject to bias if, for example, researchers are more likely to include better-known inventors. An alternative approach, and the one followed by Lai et al., is to allow the algorithm to produce its own training dataset based on features in the data. For example, a training dataset of known matches could be constructed by examining individuals with matching rare names.

Our baseline approach lies somewhere in between these two strategies. We use the matches of Lai et al. to form the basis of our training dataset. We draw twenty million pairs of records belonging to different inventors according to Lai et al. to complete our training dataset. Using this as a training dataset relies on two principal assumptions: first, we assume that the Lai et al. disambiguation correctly identifies inventors, and second we assume that the sets of features that were predictive of inventor clustering are stable over time, so that the same rules for determining matches in the modern sample of Lai et al. will apply to our historical sample. We choose this approach in order to best match the state-of-the-art disambiguation of inventors in the modern data.[20]

The second major challenge to the disambiguation exercise is computational. Ideally, one would compare every pair of records in our data, and build a similarity profile for each. However, with over 12 million unique patent-inventor records in our dataset, one would have to compare over 144 trillion record pairs in order to compare each record to each other, which is computationally infeasible. To circumvent this challenge, we follow Lai et al. in disambiguating successively larger blocks. We first group records into blocks of possible matches, based on the first characters of an inventor's name. Then we compare all records within a block to one another, but never compare across blocks. After disambiguating a set of narrow blocks, we expand the size of the block, for example by considering all record pairs that match the first three letters of an inventor's name, rather than the first five letters. By iteratively allowing progressively larger blocks, and assuming clusters within prior blocking rounds were successfully disambiguated, we greatly reduce the computational burden of the disambiguation.

Our starting point is the historical inventor data digitized by Akcigit et al. (2017), combined with the patent data of Lai et al. (2014) available on the Harvard Dataverse Network (HDN).[21] We first manually clean inventor names and location to correct for obvious typos. The most common correction is to remove prefixes and suffixes, such as "DR," "JR," and "SR." In addition, we standardize names to be all capital letters, and consider a person's first name to be the first word of their name. Finally, we consider only the first patent class listed on a patent document to be

[20] In early versions of the paper, we experimented with allowing the algorithm to find its own training sets, and found qualitatively similar headline results.

[21] Accessed from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/15705 on February 13, 2017.

that patent's primary classification.

To compare records, we construct a similarity profile for every pair of records to be compared. A similarity profile x is a vector of similarity scores for the active attributes in the disambiguation. Specifically, a similarity profile is encoded as follows:

- First and Last names
    1. If one of the two records is missing the name
    2. If there is no clear misspelling or abbreviation employed, and the strings do not exactly match
    3. If there is a misspelling (defined as either missing 1 or 2 characters somewhere, or switching the place of a few characters)
    4. If exact match or, in the case of first names, if one string appears to be an abbreviation of the other in that it has the first 3 characters the same (e.g. "ROB" and "ROBERT")
- Middle Names
    0. If have different middle names
    1. If one of the two records have missing middle name
    2. If both records have missing middle name
    3. If one record has a full middle name (e.g. "WILLIAM") and the other has just the middle initial which matches the full middle name (e.g. "W").
    4. If exactly the same name
- Location
    1. If over 50 miles apart
    2. If under 50 miles apart
    3. If under 25 miles apart
    4. If under 10 miles apart
    5. If under 1 mile apart
- Patent Classes
    0. If different strings
    1. If exactly the same string
- Assignees
    5. If the Jaro-Winkler string distance between assignee names is at least 0.9

4. If JW distance > 0.8
3. If JW distance > 0.7
2. If one of the two records has a missing assignee
1. Otherwise

- Coauthors

  1. If coauthors exactly the same (coauthors entered as <First Initial> . <Last Name> and separated by comma in the variable)
  0. Otherwise

- Country

  0. If different country
  1. If the same non-US country
  2. If the same US country

Next, one may construct, for every observed similarity profile, the probability that this profile belongs to the same inventor or not, by comparing the frequency with which it occurs in the training dataset. Specifically, defining $\mathcal{M}$ to be the set of matched inventor pairs in the training dataset, and $\mathcal{N}$ to be the set of non-matched inventor pairs in the training dataset, one may use Bayes' rule to write the probability of a match as

$$P(\mathcal{M}|x) = \frac{P(x|\mathcal{M})P(\mathcal{M})}{P(x|\mathcal{M})P(\mathcal{M}) + P(x|\mathcal{N})\,(1 - P(\mathcal{M}))}$$

where $P(\mathcal{M})$ is the prior probability of a match, which we follow Lai et al. in setting as proportional to the ratio of the number of within-cluster pairs (i.e. disambiguated inventors from prior blocking rounds) in a block to the total number of pairs in that block.[22] For numerical reasons, it is more convenient to work with the posterior *odds* of a match, defined as

$$\frac{P(\mathcal{M}|x)}{1 - P(\mathcal{M}|x)} = \frac{P(x|\mathcal{M})}{P(x|\mathcal{N})} \cdot \frac{P(\mathcal{M})}{1 - P(\mathcal{M})}$$

In particular, we calculate the likelihood ratio, $r(x)$, for every observed similarity profile x. This likelihood ratio is defined as

$$r(x) = \frac{P(x|\mathcal{M})}{P(x|\mathcal{N})} \tag{OA1}$$

This can be determined directly from the training dataset by comparing the number of records with similarity profile x that belong in the matched training dataset (i.e. come from the same

[22] The discrete nature of the similarity profile space described above makes the computation of this match probability much simpler.

inventor), to the number of records with similarity profile x that belong in the unmatched training dataset (i.e. come from different inventors).[23] Once we have the likelihood ratios calculated, we invert them to calculate the probability that two records originated from the same inventor:

$$P(\mathcal{M}|x) = \frac{1}{1 + \frac{1-P(\mathcal{M})}{P(\mathcal{M})}\frac{1}{r(x)}} \tag{OA2}$$

We say that two records originated from the same inventor if this posterior probability of a match is at least 0.99.[24]

Our blocking routine proceeds as follows:[25]

Round 1. Block based on exact first and last name. Compare records based on middle name and patent location.

Round 2. Block based on exact first and last name. Compare records based on middle name, coauthor network, patent class, and assignee name.

Round 3. Block based on first five characters of first name, and exact last name. Compare records based on middle name, coauthor network, patent class, and assignee name.

Round 4. Block based on first three characters of first name, and exact last name. Compare records based on middle name, coauthor network, patent class, and assignee name.

Finally, we subset our data to only consider US inventors. As was indeed the case in our time period, the most productive inventors are Kia Silverbrook, Shunpei Yamazaki, George Lyon, Donald Weder, and Melvin De Groote. We refer the reader to Lai et al. (2014) for additional statistics on the performance of the algorithm on modern data.

OA.2 Assigning Inventors to States

Our patent data provides information on the residence address of the patent's inventors. However, we do not observe the residence of all inventors on a patent in the historical period. Specifically, we observe an inventor's state if either 1) they are the first inventor on the patent, or 2) the patent is contained in the Harvard Dataverse Network (HDN) data. In order to run our inventor-level

[23] To account for small sample bias in rare similarity profiles, we follow Lai et al. in applying a Laplace correction to these likelihood ratio values.

[24] In the early stages of our analysis, we experimented with match thresholds of 0.98 and 0.95 to determine whether records originated from the same inventor. After examining the data by hand, we determined that this was too low, as common names such as Robert Smith were often spuriously considered the most prolific inventors in the data. This problem largely vanished with the threshold of 0.99.

[25] We experimented with additional rounds of blocking, as well as with allowing for inexact surname matches in the blocking routine. Manual checks of the data revealed that this routine minimized errors with common names, and correctly matched the most productive inventors as listed by outside data sources.

regressions, we must assign each inventor to a particular home state. In this section, we detail our approach to doing so.

For all non-primary authors on historical patents, we impute a location using the following algorithm:

1. We assign all HDN and first author inventors to the state listed in the data
2. If an inventor is an HDN or first author inventor on one patent in a given year, but not on another patent, we assign that inventor to his first-author state. If he is first author in multiple states in that year, we assign him to the state listed on the patent if that state matches one of his first author states; otherwise we proceed to step 3 below (using alternative years)
3. We replace the inventor's state with the preceding years state if state information is still missing.
4. We replace the inventor's state with the following years state if state information is still missing.
5. If the inventor-patent record is still missing state information, but the inventor has multiple first-author states listed in that year, then we pick a random first-author state for that inventor-patent.
6. If all else fails, we assign the state of the first-author on the patent.

An additional challenge arises from the fact that a number of inventors have patents granted in multiple states in the same year. There may be many causes for multiple unique states within a given year for an inventor. The most common causes of these multi-state inventors are:

- An inventor may live in state A until midway through a particular year, and then move to state B. They file a patent application both in state A before moving and in state B after moving. They never file a patent in state B before moving, and never file a patent in state A after moving.
- Inventors may have multiple home addresses. As a result, they consistently file in both state A and state B in multiple years. For example, inventors may spend half of the year in Chicago, IL, and half of the year in Milwaukee, WI, and thus frequently have patents in both of these states in a given year.
- Inventors have multiple coauthors, who live in different states and who alternate in terms of who is the first listed author. For instance, Harvey Clayton Rentschler lives in Pittsburgh, PA, but frequently coauthors with J. Marden, who lives in Orange, NJ. Every time they coauthor a patent, the location is listed as Orange, NJ, but every time Harvey Rentschler sole authors a patent, his location appears to be Pittsburgh. These situations are particularly common

among assigned patents, and seem to account for all individuals living in an exceptionally high number of states. Indeed, everyone who shows up in 7 or more states has a coauthor on their patents, while the share of those with a coauthor is 92.8% for those with multiple states, compared with just 66.3% for those in one state[26]

- Possible disambiguation errors: two individuals may have very similar names, work in similar classes, and live just across a state border from one another (so are close in latitude-longitude). As a result these two separate inventors may be classified as the same person by the disambiguator. This would inflate the number of states an individual lives in.

To address this concern, we assign multi-state inventors a home state using the following algorithm:

1. Each year, assign an inventor to the modal state in which we observe him/her operating as a sole author.

2. If the inventor does not have any sole authorships in that particular year, check if he/she has sole authorships in the preceding or subsequent year. If the preceding and subsequent year both have sole authorships in the same modal location, then assign the inventor to that state. This smoothes over off years for inventors and removes spurious migration.

3. If we still do not have a location for the inventor, then we assign them to the modal location we observe them in in the given year, regardless of whether the patent was sole authored or coauthored.

4. If the inventor has two modal states (e.g. has 2 patents in both Illinois and Wisconsin in the given year), then choose a random choice of those states and assign the inventor to that state.

OA.3 Historical Corporate Tax Data

We collected the corporate tax rates from a large variety of sources. We have built a documentation available at https://scholar.harvard.edu/stantcheva/publications that shows all the sources for each year and state. We only collected direct taxes and net income franchise taxes. We also collect surtaxes or surcharges, as well as additional temporary taxes imposed on top of the main rates. They are sometimes imposed as a percentage of regular tax liabilities and sometimes as a rate to add to the main rate. We record them as rates to add to the main rate with applicable thresholds. We have not collected minimum taxes (they are very low and probably not applicable to the companies in our sample) and alternative minimum taxes.

[26] This is partially mechanical as these invetnors are also more productive so have more chances to appear in multiple states.