APPENDIX

"GROWING THROUGH VINTAGES" by Fieler and Lee

CONTENTS

A. THEORY APPENDIX

A.1 Second order conditions of researchers' problem

We rewrite the problem of the researchers of nests as:

$$\max_{m,z} m z^{1/(\eta-\sigma)} \tag{A.1}$$

subject to

$$\begin{cases} (H_m m)^{1/\kappa} + (H_z z)^{1/\kappa} \leq 1 & \text{if } \kappa > 0 \\ (H_m m)^{1/\kappa} + (H_z z)^{1/\kappa} \geq 1 & \text{if } \kappa < 0 \end{cases} \tag{A.2}$$

where $H_m = f_2 H_2(t)^{-\nu_m}$ and $H_z = f_2 H_2(t)^{-\nu_z}$. Let λ be the Lagrange multiplier. We consider the two cases separately.

Case 1. $\kappa > 0$ The first and second order conditions with respect to m are:

$$\begin{aligned} z^{1/(\eta-\sigma)} - \lambda \frac{1}{\kappa} H_m^{1/\kappa} m^{1/\kappa - 1} &= 0 \\ -\lambda \frac{1-\kappa}{\kappa^2} H_m^{1/\kappa} m^{1/\kappa - 2} &< 0 \end{aligned}$$

The second order conditions hold if $\kappa \in (0,1)$.

For z, the first and second order conditions are respectively:

$$\begin{aligned} \frac{1}{(\eta-\sigma)} m z^{1/(\eta-\sigma)-1} - \frac{1}{\kappa} \lambda H_z^{1/\kappa} z^{1/\kappa - 1} &= 0 \\ \frac{1-(\eta-\sigma)}{(\eta-\sigma)^2} m z^{1/(\eta-\sigma)-2} - \lambda \frac{1-\kappa}{\kappa^2} H_z^{1/\kappa} z^{1/\kappa - 2} &< 0 \end{aligned}$$

Substituting the first order conditions, we re-write the second order conditions as:

$$\lambda \frac{1}{\kappa} H_z^{1/\kappa} z^{1/\kappa - 2} \left[\frac{1-(\eta-\sigma)}{(\eta-\sigma)} - \frac{1-\kappa}{\kappa} \right] < 0$$

The inequality holds if $\kappa < \eta - \sigma$. In the quantification of the model, $\eta - \sigma < 1$ so that the objective function (A.1) is convex in z. The condition $\kappa < \eta - \sigma$ is that the constraint has to be more convex than (A.1).

Case 2. $\kappa < 0$ The first and second order conditions with respect to m are:

$$z^{1/(\eta-\sigma)} + \lambda \frac{1}{\kappa} H_m^{1/\kappa} m^{1/\kappa - 1} = 0$$
$$\lambda \frac{1-\kappa}{\kappa^2} H_m^{1/\kappa} m^{1/\kappa - 2} < 0$$

which does not hold for any $\kappa < 0$.

A.2 Stability of balanced growth path

We check the stability of the balanced growth path with respect to shifts in research labor from creating new varieties to new nests, or vice-versa. We do not consider non-monotonic deviations from the balanced growth path, because cycles where researchers shift to and from the creation of nests may arise as in Matsuyama (1999), and Acemoglu (2009). The model is purposely fit to accommodate these cycles, which are the object of our analysis in Section 7.

The stability of balanced growth with respect to monotonic increases in $H_2(t)$ or to $H_1(t)$ is non-trivial because researchers of new nests create positive externalities on each other through ν_z. The proof below shows, however, that this positive externality accrues both to the creator of new nests and new varieties within nests, but the negative externality of H_2 decreasing the measure of nests per researcher accrues only to the creators of nests.

Note also that deviations of zero measure are not profitable. Since researchers are indifferent between creating new varieties or nests, a zero measure of researchers changing their choice doesn't affect the balanced growth path. Similarly, a deviation for a zero time interval is not profitable, because it doesn't change the expected flow profits and the measure of nests per researcher m decreases in $H_2(t)$ (so $m(t)V(\bar{z}(t),t)$ decreases in $H_2(t)$).

We then consider a shift of researchers from creating nests to creating varieties for a non-zero measure of time with a lower bound of zero, without loss of generality. That is, let $T \subset \mathbb{R}_+$ be the set of time with this shift of researchers and assume $[0,\epsilon] \subset T$ where $\epsilon > 0$. We denote the new allocations and variables with a superscript Δ. The assumed shift has $H_1^\Delta(t) > H_1(t)$ and $H_2^\Delta(t) = H_2(t) - [H_1^\Delta(t) - H_1(t)]$ for all $t \in T$. We will show that under the new equilibrium, the creation of new nests becomes more profitable than the creation of new varieties so that market forces would push researchers back to their original allocations.

With the shift, the productivity of new nests decreases for any $t \in T$:

$$\frac{z^\Delta(t)}{z(t)} = \left[\frac{H_2^\Delta(t)}{H_2(t)}\right]^{\nu_z} < 1$$

since $\nu_z > 0$ and $H_2^\Delta(t) < H_2(t)$.

We claim that there's entry into the old nests or any other nest $\tau \notin T$, as in the balanced growth path. With the deviation, the nests $\tau \in T$ have lower productivity and are fewer relative to the balanced growth path. Hence, creators of new varieties, $H_1^\Delta(t) \geq H_1(t)$, cannot be allocated disproportionately to these nests $\tau \in T$. Equalized profits $\tau \notin T$ and (12) implies that we can write

$$n^\Delta(\tau,t) = \bar{n}^\Delta(t)\left[\frac{z(\tau,t)}{Z^\Delta(t)}\right]^{1/(\eta-\sigma)} < n(\tau,t) = \bar{n}(t)\left[\frac{z(\tau,t)}{Z(t)}\right]^{1/(\eta-\sigma)}$$

for all $\tau \notin T$. The inequality arises because $\bar{n}^\Delta(t)Z^\Delta(t)^{1/(\sigma-\eta)} > \bar{n}(t)Z(t)^{1/(\sigma-\eta)}$ for all $t > 0$ because of the lower productivity and fewer nests in $\tau' \in T$ and because of the increase in $H_1(t)$. Then, $n^\Delta(\tau,t) > n(\tau,t)$ for all $\tau \notin T$ because for these nests $z^\Delta(\tau,t) = z(\tau,t)$.

At the balanced growth path, $m(t)V(\bar{z}(t),t) = v(\bar{z}(\tau),\tau,t)$ for any $t \geq 0$ and any $\tau \leq t$. Using (13) and (14), and the profit-equalization for $\tau \notin T$, we write this value differential as:

$$\begin{aligned} &m^\Delta(t)V(\bar{z}^\Delta(t),t) - v^\Delta(\bar{z}^\Delta(\tau),\tau,t) = \\ &\chi\tilde{f}_2^{-1}[H_2^\Delta(t)]^{\nu_m}\int_t^\infty n^\Delta(\tau,s)\overline{\pi}^\Delta(s)\left[1 - \frac{1-\chi}{\chi}\tilde{f}_2 e^{-\delta_1 s}\frac{1}{[H_2^\Delta(t)]^{\nu_m} n^\Delta(\tau,s)}\right]ds \end{aligned}$$

for any $t > 0$ and $\tau < t$ with $\tau \notin T$, where we use the notation $\overline{\pi}^\Delta(s)$ to be the flow profit from varieties with entry at time s. At the balanced growth path (without the superscripts Δ) the term in the integral adds to zero. But $[H_2^\Delta(t)]^{\nu_m} \leq [H_2(t)]^{\nu_m}$ for all $t \geq 0$ with strict inequality when $t \in T$, and $n^\Delta(\tau,s) > n(\tau,s)$ and all $\tau \notin T$ and $s > \tau$. Hence, $m^\Delta(t)V(\bar{z}^\Delta(t),t) > v^\Delta(\bar{z}^\Delta(\tau),\tau,t)$ which contradicts entry into $\tau \notin T$. That is, the deviation Δ cannot be profitable and market forces push researchers back to the creation of new nests.

Note that the proof doesn't require profits $\bar{\pi}^\Delta(s)$ to go down. Depending on χ, the allocation of researchers may not be welfare maximizing in the balanced growth path. So, total output and profits per variety may well increase with the deviation.

B. ESTIMATION WITH PATENT DATA

B.1 Estimation with gamma function

B.1.1 Parametrization with gamma function

The special case $\psi(x) = x^\gamma$ with $\gamma > 0$ satisfies Assumption 1, delivers a closed-form solution to the balanced growth path and serves as a basis for the estimation of the model. The constants above involving integrals become:

$$\begin{aligned}
\Upsilon_2 &= \Upsilon_1 g_H(1+\nu_m)\Gamma\left(\frac{\gamma}{\eta-\sigma}+1\right)\left[g_H\left(1+\nu_m+\frac{\nu_z}{\eta-\sigma}\right)\right]^{-[\gamma/(\eta-\sigma)+1]} \\
\Upsilon_3 &= \Gamma\left(\frac{\gamma}{\eta-\sigma}+1\right)\left[\rho+g_H\left(1+\nu_m+\frac{\nu_z}{\eta-\sigma}\right)\right]^{-[\gamma/(\eta-\sigma)+1]} \\
\Upsilon_4 &= \left[\frac{g_H[1+\nu_m+\nu_z/(\eta-\sigma)]}{\rho+g_H[1+\nu_m+\nu_z/(\eta-\sigma)]}\right]^{1+\gamma/(\eta-\sigma)}
\end{aligned}$$

where Γ is the gamma function.

In this special case, the evolution of the number of varieties within a nest over time, in equation (21), has the shape of the density of a gamma distribution:

$$n(\bar{z},\tau,t) = \Upsilon_2^{-1}\bar{n}(\tau)\left(\frac{\bar{z}}{\mathbf{z}(\tau)}\right)^{1/(\eta-\sigma)} e^{-g_H[\nu_m+\nu_z/(\eta-\sigma)](t-\tau)}(t-\tau)^{\gamma/(\eta-\sigma)} \tag{B.3}$$

where the shape parameter $\gamma/(\eta-\sigma)+1$ and scale parameter $g_H[\nu_m+\nu_z/(\eta-\sigma)]$ are the same across nests, but the height of this bell-shaped curve, governed by $\Upsilon_2^{-1}\bar{n}(\tau)(\bar{z}/\mathbf{z}(\tau))^{1/(\eta-\sigma)}$, and the timing of its first appearance τ are nest-specific.

Assuming in addition $\delta = g_H[\nu_z/(\eta-\sigma)+\nu_m]$, the share of patents in class j in equation (32) becomes:

$$s_j(t) = \sum_{\omega\in\Omega_j} a(\omega)e^{-\beta_1(t-\tau(\omega))}(t-\tau(\omega))^{\beta_2}$$

where we redefine $a(\omega)$ here to be $a(\omega)\gamma/(\eta-\sigma)$ in (33), $\beta_2 = \gamma/(\eta-\sigma)-1$ and $\beta = \{\beta_1,\beta_2\}$ has two elements. The evolution of the number of patents also follows a Gamma distribution, though it's shape parameter is $\gamma/(\eta-\sigma)$ instead of $\gamma/(\eta-\sigma)+1$ in equation (B.3).

B.1.2 Estimation results with gamma distribution

From step 1, the estimated parameters are $\beta_1 = 0.39$ (s.e. 0.003) and $\beta_2 = 9.36$ (s.e. 0.14).[42] The estimated rate and shape parameters of the distribution are $\beta_1 = 0.39$ (s.e. 0.003) and $\gamma/(\eta - \sigma) = 10.36$ (s.e. 0.14) in stage 1 and $g_H(1 + \nu_m) = 0.0034$ (s.e. 0.0003) in stage 2. Together with $g_H = 0.011$, $\sigma = 1.11$, $\eta = 1.667$, these parameters imply the growth rate of real income per worker $g_w = g_H[\eta - 1 + (\eta - \sigma)\nu_m + \nu_z] = 0.20$ —an implausibly high rate. This implausibly high rate arises for the following reason. The density of the gamma distribution has a bell shaped curve only if $\gamma/(\eta - \sigma)$ is large. For smaller values, its left tail is concave, which doesn't fit the data as well. But a large γ implies a fast and lasting growth of productivity within nest. Then, the model needs very fast growth in the rest of the economy (a large obsolescence rate β_1) to capture the fast decline of patents after the booms observed in the data.

The fit of the model and other empirical implications and fit of the model are similar to the Weibull parametrization in the main text. Tables A1 presents the distribution of nests per patent class, and A2 presents the fit of the model. These tables are analogous to Tables 2 and 3, respectively, and we have kept the results in the original tables (for the Weibull distribution) for ease of reference. The number of nests per patent class is larger for the gamma parametrization. It averages 5.21 compared to 4.34 with the Weibull parametrization. This implies that each patent class has on average $10.4 = 5.21 \times 2$ to match 141 observations (one per year). Despite the larger number of parameters, the fit of the model in Table A2 is very similar in the two parametrizations.

Figure A1 plots the share of patents that experience a nest birth in each year. The slope is almost identical in the two parametrizations and the estimated growth in the number of nests $g_N = g_H(1 + \nu_m) = 0.0034$ is the same. The drop in number of births of nests after 1970 is also visible in both parametrizations. The two parametrizations exhibit similar distribution of nest productivity, captured by $\tilde{a}(\omega)$ in Figure A2. The standard deviation parameter $\theta(\eta - \sigma)$ is again nearly identical in the two parametrizations, 1.21 in the Weibull and 1.20 in the Gamma.[43]

Overall, the two parametrizations yields a similar fit with a similar number of parameters. The stark difference is the economic growth rate that leads us to rule out an unbounded productivity within nest as in $\psi(x) = x^\gamma$.

[42] We apply the formula in Section 16.4.6, equation 16-18 of Greene (2008) for estimating the asymptotic variance of the maximum likelihood estimator, where we approximate the derivatives numerically.

[43] The five smallest estimates of $\tilde{a}(\omega)$ were significant outlyiers in the Gamma parametrization and we drop them from the estimation of $\theta(\eta - \sigma)$.

TABLE A1: Distribution of number of nests per patent class

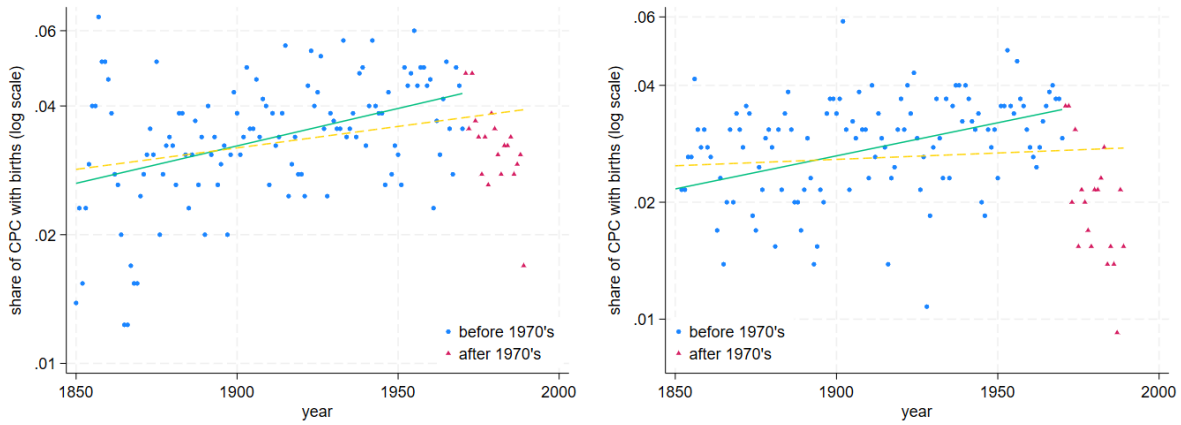
number of nests (k)	number of patent classes with k nests		cumulative percentile	
	Gamma	Weibull	Gamma	Weibull
0	0	1	0	0.15
1	18	29	2.8	4.6
2	35	60	8.2	14
3	56	98	17	29
4	106	165	33	54
5	144	143	55	76
6	139	93	77	91
7	81	44	89	97
8	56	14	98	99.5
9	12	2	99.5	99.8
≥ 10	3	1	100	100
total	**650**	**650**		

The table reports the estimated number of nests in each four-digit CPC code. The average number of nests is 5.21 in the Gamma parametrization and 4.34 in the Weibull one.

TABLE A2: Distribution of R-squared across CPC codes after step 2

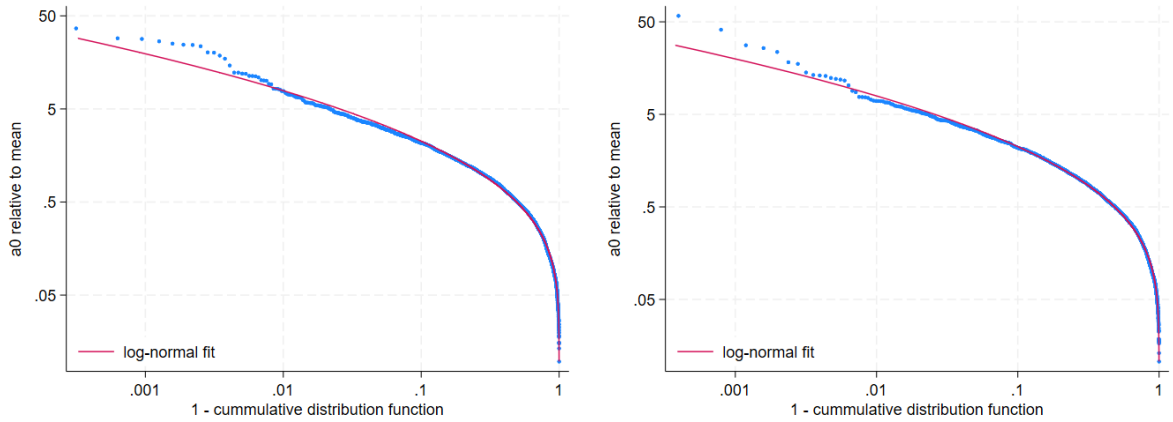
	percentiles of distribution					
	10%	25%	50%	75%	90%	**mean**
Gamma	0.46	0.64	0.79	0.89	0.95	0.75
Weibull	0.49	0.65	0.80	0.90	0.95	0.75

Step 2 separately estimates the model separately by CPC code using maximum likelihood for a given common parameter β. The table presents moments from the distribution of R-squared (not targetted) across CPC codes to illustrate the model fit. For each CPC code, there are 141 observations (number of years) and the number of parameters on average is 10.4 for the Gamma parametrization and 8.7 for the Weibull (two times the number of nests, excluding the standard error ϑ which doesn't enter the fit).



(A) Gamma (B) Weibull

FIGURE A1: Share of CPC codes with the birth of a nest per year



(A) Gamma (B) Weibull

FIGURE A2: Distribution of $\tilde{a}_0(\omega)$

B.2 Birth of nests after 1970's

Figure 6(a) plots the model's predictions for number of nests per year (divided by 650 patent classes) against time. There's a clear increase in the number of nests up until 1970 and a sharp drop afterward (observations in red triangles). This appendix argues that the decline in observed births is probably not real, but a result of births being concentrated in a few patent classes in the last few decades.

Before addressing this hypothesis, we first confirm that this result doesn't arise because our estimation procedure cannot identify the birth of nests from their left tail. We repeat the estimation procedure using data up until 1995, instead of 2015. Since the time from the birth of a nest to its peak is 25 years, 1995 is the last year for which we could reliably estimate the birth of nests born after 1970. Like in Figure 6(a), we only consider births 25 years before the beginning and end of our data (1850 to 1970). Figure A3 shows an increasing pattern of births throughout the sample period, similar to the one in Figure 6(a). The slope in the two graphs, from 1850 to 1970, is the same 0.039 (0.008).

Figure A4 plots two measures of the concentration of patents across patent classes (four-digit CPC codes) over time: The Herfindahl-Hirschman index and the share of patents in the largest ten classes. Both measures show a similar picture. Concentration levels were stable until the early 1990s and rose sharply afterward. In the model, this rise in concentration is consistent with the birth of many nests within a few categories starting in the 1970s because patents within nests don't start to grow significantly until about 16 years after birth and only peak at 25 years after birth (see 5 for the life cycle of patents within a nest).

Tables A3 and A3 list the ten CPC codes with the largest number of patents in the first and last year of our data. There's very little overlap in the two lists. In 1875, the main categories relate to agriculture and textiles, and in 2015, they relate to computer science

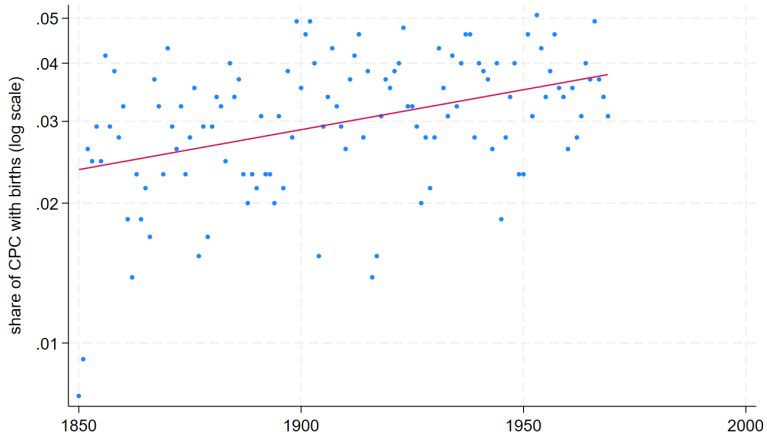


FIGURE A3: Share of CPC codes with the birth of a nest per year, estimated using data up until 1995

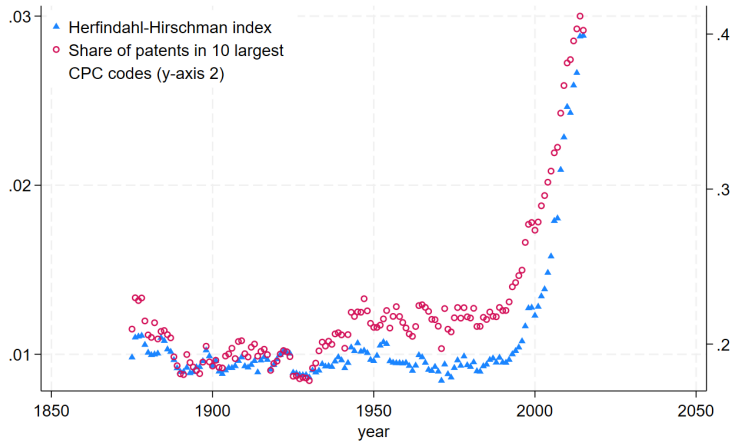


FIGURE A4: Two measures of the concentration of patents across four-digit CPC codes over time

and healthcare.

B.3 Nest birth and measures of patent influence

Figure A5 replicates Figure 11 using KPST measured at the three digit CPC code, instead of the four-digit code used in Figure 11.

In Table A5, we include information on nest productivity, measured by $\tilde{a}(\omega)$, in addition to its birth. In column (1), we regress the KPST measure on a dummy for whether the patent is within 12 years of the birth of a nest in its CPC code and on the $\log \tilde{a}(\omega)$ of the nest with the closest birth date to the patent filing date. The regression includes fixed effects for year and CPC code. Both coefficients are positive and significant. A dummy for the patent being within 12 years of the birth of a nest is associated with an increase in its KPST measure by about one-third of a standard deviation, and a one standard deviation increase in productivity $\log \tilde{a}(\omega)$ is associated with an increase in the KPST

TABLE A3: Largest 10 patent classes in **2015**

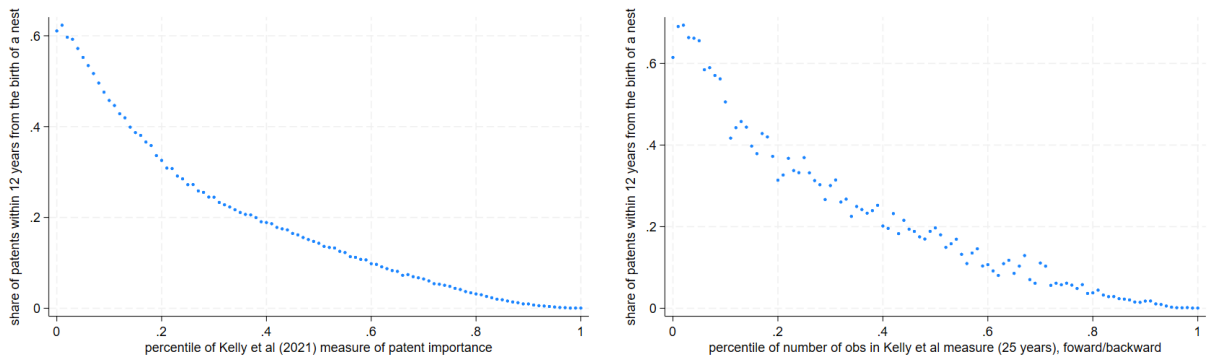
CPC code	description*	num. of patents	share of total
G06F	electric digital data processing	19,453	0.125
H04L	transmission of digital information	10,947	0.070
H01L	semiconductor devices, not in class H10	5,586	0.036
H04W	wireless communication networks	5,108	0.033
A61B	diagnosis, surgery, identification	4,986	0.032
H04N	pictorial communication (e.g., television)	4,179	0.027
Y10T	technical subj. covered by former classification (e.g., buckles, metal working)	4,179	0.027
G06Q	information and communication technology	3,490	0.022
A61K	preparations for medical, dental or toiletry purposes	2,721	0.017
A61P	specific therapeutical activity of chemical compounds or medicinal preparations	2,225	0.014

* See USPTO https://www.uspto.gov/web/patents/classification/ for the full CPC code descriptions.

TABLE A4: Largest 10 patent classes in **1875**

CPC code	description*	num. of patents	share of total
Y10T	technical subj. covered by former classification (e.g., buckles, metal working)	773	0.064
A01B	soil working in agriculture or forestry	290	0.024
B65D	containers for storage (e.g., bags, boxes, cans)	288	0.024
D06F	laundering, drying, ironing of textile articles	213	0.018
A01D	harvesting, mowing	210	0.017
A47C	chairs, sofas, bed	190	0.016
A47B	tables, desks, office furniture, cabinets, drawers	179	0.015
Y10S	technical subj. in former USPC cross ref. art collections	138	0.011
E06B	closures for openings in buildings, vehicles, fences (e.g., doors, windows, gates)	131	0.011
A01K	animal husbandry; aciculture, apiculture, pisciculture	130	0.011

* See USPTO https://www.uspto.gov/web/patents/classification/ for the full CPC code descriptions.



(A) KPST measure (B) Number of observations

FIGURE A5: Most influential patents at CPC-3 digit level and birth of nests

Note: The figure repeats Figure 11 but uses three-digit CPC codes in the comparison groups $\mathcal{F}$ and $\mathcal{B}$ when constructing the Kelly, Papanikolaou, Seru, and Taddy (2021) measure of influence. The rest of the figure is constructed in the same way. The x-axis in (a) is the ranking of KPST measure within four-digit CPC code grouped into percentile bins. In (b), the x-axis is the ranking (also in bins) of the ratio of the number of observations used to construct the KPST measure, $\mathcal{F}_j/\mathcal{B}_j$. There's an average of about 3500 patents per CPC code. On the y-axes is the share of patents in each percentile bin that is within the first 12 years of the birth of a nest or in the year of the birth minus one. Of all patents, 0.21 are within this interval of a birth.

measure by 0.13 standard deviations.

In Column (2) we add an interaction of the two independent variables from column (1). The interaction is positive suggesting that the KPST measure is particularly high when the patent is within the first 12 years of a productive nest. Column (3) replaces the dummy variable in (1) with the absolute value of the time lag between the patent filing date and the birth of its closest nest. Qualitatively, the results are unchanged: The shorter is the distance to a birth, the higher is the patent influence.

Columns (4) through (7) replaces the dependent variable in the baseline (1) with different measures of patent influence. Columns (4) and (5) change the set $\mathcal{F}_j$ to include patents filed 10 years after j or in all years after j respectively, as opposed to 25 years in the baseline. The coefficient and the R-squared is much lower in column (4) than (1) consistent with our estimation where it takes 25 years for a breakthrough patent to reach its peak influence. When we use all years in column (5), the sign of the coefficient on $\log \tilde{a}(\omega)$ becomes negative. Using all years may confound a patent's true influence with its use of vocabulary that later became common in the field.

The dependent variable in column (6) is the number of citations. The coefficients are still positive and statistically significant, but the R-squared drops significantly. In column (7) when we include in sets $\mathcal{F}_j$ and $\mathcal{B}_j$ patents in the same three-digit CPC code as j, not only in its four-digit code, the results are very similar to column (1), in magnitude, statistical significance and R-squared.

TABLE A5: Patent influence and nest characteristics from estimation

	(1) baseline 25 years in $\mathcal{F}$	(2) interaction	(3) dist. to nest	(4) 10 years in $\mathcal{F}$	(5) all years in $\mathcal{F}$	(6) # citations	(7) CPC-3 digit
Patent within 12 years of the birth of a nest	0.289*** (0.022)	0.200*** (0.028)		0.061*** (0.015)	0.264*** (0.022)	0.041*** (0.013)	0.222*** (0.0179)
Height of closest nest $\log\tilde{a}(\omega)$	0.085*** (0.032)	0.056* (0.0300)	0.107*** (0.0357)	0.071*** (0.020)	-0.162*** (0.056)	0.068*** (0.025)	0.0498*** (0.0179)
Interaction		0.144*** (0.0217)					
log(distance to closest nest)			-0.0027*** (0.00071)				
Observations	2,305,457	2,305,457	2,305,457	2,304,646	2,305,549	2,330,265	2,327,097
R-squared	0.651	0.658	0.642	0.390	0.762	0.426	0.765
Fixed effects							
four-digit CPC code	yes	yes	yes	yes	yes	yes	yes
year	yes	yes	yes	yes	yes	yes	yes
share of patents within 12 years of a nest	0.214	0.214		0.214	0.214	0.214	0.214
standard deviation of...							
dependent variable	0.954	0.954	0.954	0.665	1.30	1.36	0.884
$\log\tilde{a}(\omega)$	1.51						
interaction term		0.685					
distance to closest nest (logs)			18.73				

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered by four-digit CPC code. All regressions include fixed effects for four-digit CPC codes and for year. The number of CPC codes and year are respectively 646 and 166. The dependent variable varies across columns. In columns (1), (5) and (6), it is the KPST measure of influence as described above with five years of patents in set $\mathcal{B}_j$ and 25 years in $\mathcal{F}_j$. We change $\mathcal{F}_j$ to span ten years after the filing of j in column (2) or all years after j in column (3). In column (4) the dependent variable is the number of citations of a patent. The comparison sets $\mathcal{B}_j$ and $\mathcal{F}_j$ in columns (1)-(6) contain only patents in the same four-digit CPC code as j. In column (7) we expand this set to include patents in the same three-digit CPC code as j. Consider the set of birth dates and nest productivity $\tau(\omega), \tilde{a}(\omega)$ estimated in 5.2 above for the same CPC code as a patent j. The independent variable "patent within 12 years of the birth of a nest" is a dummy for whether patent j was issued in the time interval $[\tau(\omega)-1, \tau(\omega)+12]$ and variable $\log\tilde{a}(\omega)$ corresponds to the productivity parameter of the nest whose birth $\tau(\omega)$ is closest to j. Columns (5) modifies the baseline column (1) by introducing an interaction term between the two independent variables. In column (6), we substitute the dummy for whether the patent is within 12 years of the birth of nest with the absolute value of the distance to the closest nest.

C. SIMULATIONS OUT OF THE GROWTH PATH

This appendix provides the technical details for Section 7. In Appendix C.1, we derive some of the hat algebra expressions in the main text, and in Appendix C.2 we detail the simulation procedure.

C.1 Auxiliary derivations for model in changes

Equation (43). Aggregating across all varieties (4) implies

$$[\bar{N}(t)\dot{\bar{n}}(t)] = -\delta\bar{N}(t)\bar{n}(t) + \frac{H_1(t)}{f_1} \tag{C.4}$$

Rearranging this same expression in the bgp and setting the growth rate of $\bar{N}^*(t)\bar{n}^*(t)$ to g_H, we have

$$\bar{N}^*(t)\bar{n}^*(t) = \frac{H_1^*(t)}{f_1(\delta + g_H)}. \tag{C.5}$$

Then

$$\begin{aligned}
\frac{d}{dt}\left[\widehat{\bar{N}}(t)\widehat{\bar{n}}(t)\right] &= \widehat{\bar{N}}(t)\widehat{\bar{n}}(t)\left(\frac{\bar{N}(t)\dot{\bar{n}}(t)}{\bar{N}(t)\bar{n}(t)} - \frac{\bar{N}^*(t)\dot{\bar{n}}^*(t)}{\bar{N}^*(t)\bar{n}^*(t)}\right) \\
&= -(\delta + g_H)\widehat{\bar{N}}(t)\widehat{\bar{n}}(t) + \frac{H_1(t)}{f_1\bar{N}^*(t)\bar{n}^*(t)} \\
&= (\delta + g_H)\left[\widehat{H}_1(t) - \widehat{\bar{N}}(t)\widehat{\bar{n}}(t)\right]
\end{aligned}$$

where the second line uses (C.4) and the last line uses (C.5).

Value of research From (24) and (25), real wages and aggregate profits in changes are:

$$\widehat{w}(t) = \widehat{Z}(t)\widehat{\bar{N}}(t)^{\eta-1}\widehat{\bar{n}}(t)^{\sigma-1} \tag{C.6}$$

$$\widehat{\bar{\pi}}(t) = \frac{\widehat{w}(t)}{\widehat{\bar{N}}(t)\widehat{\bar{n}}(t)} \tag{C.7}$$

The stochastic interest rate is:

$$r(t) = \rho + g + \frac{\dot{\widehat{w}}(t)}{\widehat{w}(t)}. \tag{C.8}$$

To derive the value of new varieties in (44), we substitute (13) and using (21), which holds with entry in all periods and nests, and we use $e^{-\int_t^s \dot{\widehat{w}}(s')/\widehat{w}(s')ds'} = \widehat{w}(t)/\widehat{w}(s)$ and

$\widehat{\tilde{\pi}}(s)/\widehat{w}(s) = [\widehat{\tilde{N}}(s)\widehat{\tilde{n}}(s)]^{-1}$. For the value of nests, we use these same two expressions together with (14) and the optimal number of varieties per nest in (21), which also holds with entry into all nests.

C.2 Simulation procedure

Let t_0 be the initial period of the simulation, and T the final period. We simulate the economy in itervals of one year. The expressions in the main text that refer to a cross section remain unchanged. For expressions that involve integrals, we rewrite them here in discrete time.

For each guess of $\widehat{H}_1(t)$, we perform the following steps.

1. $\widehat{H}_2(t)$, $\widehat{\tilde{H}}_2(t)$Calculate $\widehat{H}_2(t)$ in equation (38). For $\widehat{\tilde{H}}_2(t)$, equation (39) becomes:

$$\widehat{\tilde{H}}_2(t) = \left(1 - e^{-(\phi+g_H)}\right)\left[\sum_{s=t_0}^{t} e^{-(\phi+g_H)(t-s)}\widehat{H}_2(s)\right] + e^{-(\phi+g_H)(t-t_0+1)}. \tag{C.9}$$

2. **Random nests** We draw the number of nests $N(t)$ in each interval $[t, t+\Delta)$ as a Poisson with parameter $\Delta K(t)$, where $K(t)$ is in (40). For each of these nests, we draw a reference productivity $\hat{\tilde{z}}(\omega)$ from a log normal with variance parameter θ^2 and mean parameter $\log \hat{\mathbf{z}}(t)$ given by (41).

   These random draws gives as a set $\tilde{\Omega}(t)$ with $|\tilde{\Omega}(t)| = N(t)$ and productivity $\hat{\tilde{z}}(\omega)$ for each $\omega \in \tilde{\Omega}(t)$.

3. $\widehat{\tilde{N}}$, $\widehat{Z}$, **and** $\widehat{\tilde{N}\tilde{n}}$**.** The total number of nests at time t is $\bar{N}(t) = \bar{N}^*(t_0 - \Delta) + \sum_{s\in\mathcal{T}(t)} N(s)$, where $\bar{N}^*(t) = N^*(t)/g_N$ where $N^*(t) = J\lambda^*(1970)e^{(t-1970)g_N}$ as in the main text. In deviations $\widehat{\tilde{N}}(t) = \bar{N}(t)/\bar{N}^*(t)$.

   The productivity of the representative nest is:

$$Z(t) = \mathbf{z}^*(t_0)\bar{N}(t)^{\sigma-\eta}\left\{A_1(t) + \sum_{s=t_0}^{t-1}\left[\psi(t-s)^{1/(\eta-\sigma)}\sum_{\omega\in\tilde{\Omega}(s)}\bar{z}(\omega)^{1/(\eta-\sigma)}\right]\right\}^{\eta-\sigma} \tag{C.10}$$

   where we take $\bar{z}(\omega) = \hat{\tilde{z}}(\omega)e^{g_z(t-t_0)}$, and we exclude the nests created at time t from the sum because they have zero productivity and $A_1(t)$ is the component of the nests created before t_0:

$$A_1(t) = \Upsilon_1 N^*(t_0)\sum_{s=-S}^{t_0-1} e^{(s-t_0)[g_N + g_z/(\eta-\sigma)]}\psi(t-s)^{1/(\eta-\sigma)}$$

   We take $S = 5{,}000$, a sufficiently large number so that $e^{(S-t_0)[g_N+g_z/(\eta-\sigma)]} \approx 0$. We

build the bgp Z^* from (C.10):

$$Z^*(t) = \mathbf{z}^*(t_0)\bar{N}^*(t)^{\sigma-\eta}\left\{A_1(t) + \Upsilon_1 \sum_{s=t_0}^{t-1} N^*(t)\left[e^{g_z(t-t_0)}\psi(t-s)\right]^{1/(\eta-\sigma)}\right\}^{\eta-\sigma}.$$

Note that in simulating $Z(t)$ and $Z^*(t)$ we can set $\mathbf{z}^*(t_0)$ to any value since below we only use the ratio $\hat{Z} = Z(t)/Z^*(t)$.

Starting with $\widehat{\bar{N}(t_0)\bar{n}(t_0)} = 1$, we construct the sequence of deviations in the average number of varieties per nest using:[44]

$$\widehat{\bar{N}(t)\bar{n}(t)} = \left(\frac{1-\delta}{1+g_H}\right)\widehat{\bar{N}(t-1)\bar{n}(t-1)} + \left[1 - \left(\frac{1-\delta}{1+g_H}\right)\right]\widehat{H}_1(t). \tag{C.11}$$

4. **Value of innovation.** With the perfect foresight assumption, we approximate the value of innovation in (44) and (C.12) as:

$$\frac{\hat{v}(t)}{\hat{w}(t)} = \left[\frac{1-e^{-(\delta+g_H+\rho)}}{e^{-(\delta+g_H+\rho)}}\right]\left\{\sum_{s=t+1}^{T} e^{-(\delta+g_H+\rho)(s-t)}\left[\widehat{\bar{N}(s)\bar{n}(s)}\right]^{-1}\right\} + e^{-(\delta+g_H+\rho)(T+1-t)}$$

(C.12)

$$\frac{\hat{V}(1,t)}{\hat{w}(t)} = \Upsilon_6 \sum_{s=t+1}^{S} e^{-[\rho+g_H+g_H\nu_z/(\eta-\sigma)](s-t)}\psi(s-t)\left[\widehat{\bar{N}}(s)\widehat{Z}(s)^{1/(\eta-\sigma)}\right]^{-1}$$

$$\Upsilon_6^{-1} = \sum_{s=1}^{S} e^{-[\rho+g_H+g_H\nu_z/(\eta-\sigma)](s)}\psi(s).$$

where S is a large number (5000) and we set all $\widehat{\bar{N}}(s) = \widehat{Z}(s) = 1$ for $s > T$ in (C.12).

5. **Equilibrium in research** The research market is in equilibrium if

$$\frac{\hat{v}(t)}{\hat{w}(t)} = \widehat{\bar{H}}_2(t)^{\nu_m+\nu_z/(\eta-\sigma)}\frac{\hat{V}(1,t)}{\hat{w}(t)}$$

for all t.

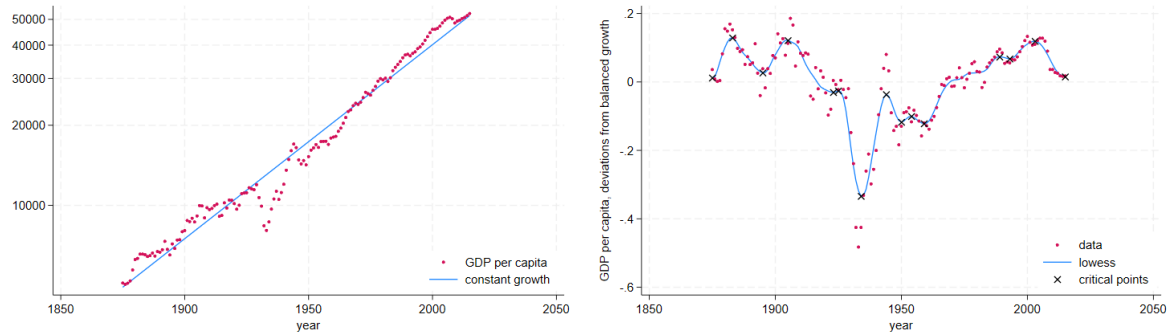
[44] The evolution of total varieties in discrete time (Δ intervals) is:

$$\bar{N}(t)\bar{n}(t) = (1-\delta)^{\Delta}\bar{N}(t-\Delta)\bar{n}(t-\Delta) + \Delta\frac{H_1(t)}{f_1}$$

Evaluating this expression in the bgp where $\bar{N}(t)\bar{n}(t)$ grows at a rate g_H, we get

$$\Delta\frac{H_1^*(t)}{f_1} = \left[1 - \left(\frac{1-\delta}{1+g_H}\right)^{\Delta}\right]\bar{N}^*(t)\bar{n}^*(t)$$

Substituting back into the first equation, dividing by $\bar{N}^*(t)\bar{n}^*(t)$ and rearranging gives us (C.11).



(A) Original data & trend (B) Detrended and with local smoothing

FIGURE A6: GDP per capita in USA, 1875-2015

We iterate over guesses of $\hat{H}_1(t)$. For each guess, we follow steps 1-4 above until step 5 holds.[45]

C.3 USA GDP Data from Maddison Project

Typical analyses of business cycles are concerned with small short-term deviations from GDP trend concerning monetary policy. Our analysis longer real fluctuations as in Comin and Gertler (2006). For an order of magnitude of economic fluctuations in the data, Figure A6 illustrates the data on real income per capita data in the USA from the Maddison Project.[46] Panel (A) plots the raw data and the constant growth line. Panel (B) plots the detrended data in red dots together with a LOWESS line (bandwidth 0.1). The crosses represent the points of local minima and maxima. The standard deviation in the LOWESS series is 0.105, and the average time between critical points is ten years. Compared to Table 5, cycles in the data are deeper and shorter.

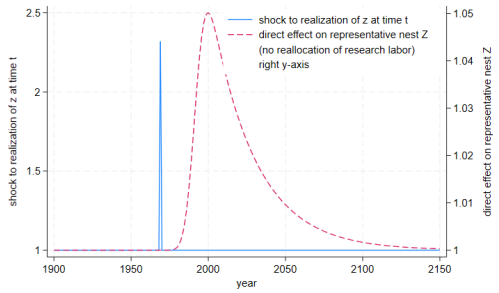
C.4 Impulse responses

Figure A7 presents the results of the shock studied in Section 7.2 with alternative values of ϕ and H_1^*/H^*. Here, we take as a baseline, the case with $H_1^*/H = 0.5$, so deviations in research are equal in magnitude for varieties and nests ($\hat{H}_1 - 1 = -(\hat{H}_2 - 1)$), and $\widehat{\bar{H}}_2 = 1$, so the reallocation does not change individual researcher productivity. This case, $\widehat{\bar{H}}_2 = 1$ arises if (7) is redefined as

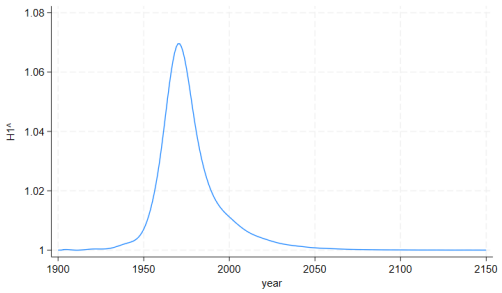
$$\bar{H}_2(t) = \int_{-\infty}^{t} e^{\phi(s-t)} H(s) ds$$

[45] In particular, we minimize the sum of squared deviations from equilibrium $\left[\hat{v}(t)\widehat{\bar{H}}_2(t)^{-\nu_m-\nu_z/(\eta-\sigma)}/\hat{V}(1,t) - 1\right]^2$. Instead of all 341 elements of $\hat{H}_1(t)$ for $t = 1775,...,2115$, we parameterize $\hat{H}_1$ using cubic B-splines with knots placed every 10 years, thereby reducing the number of parameters to 37. This approximation yields deviations from equilibrium that are almost always less than 1 percent and on average less than 0.3 percent.

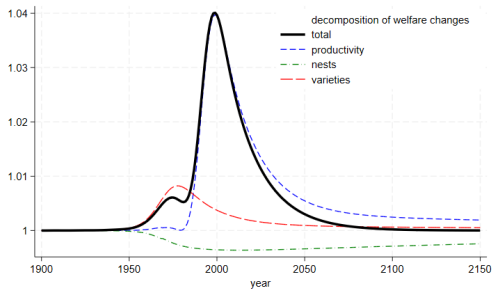
[46] Downloaded from https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020 on February 13, 2026.



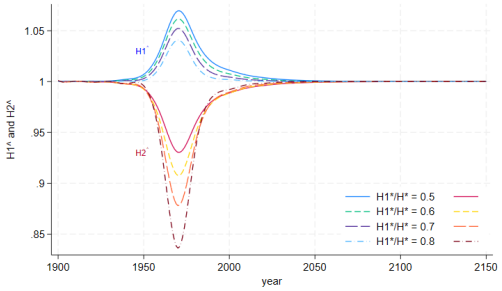
(A) Shock (all cases)



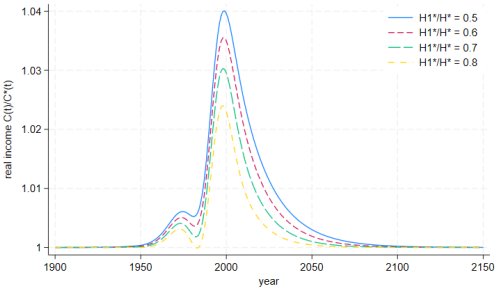
(B) Baseline: Research reallocation



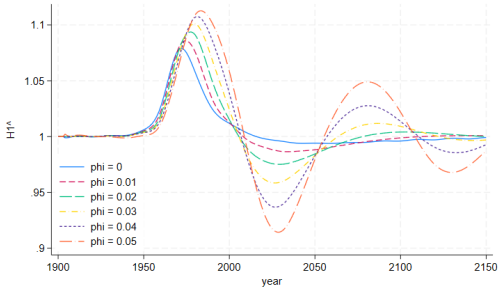
(C) Baseline: Welfare



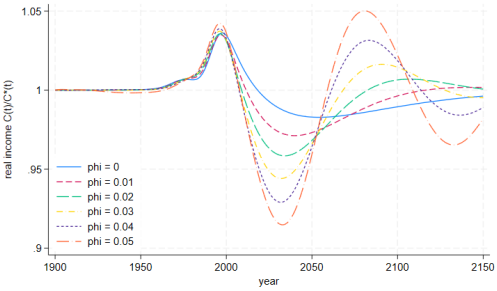
(D) Variation in H_1^*/H: Research



(E) Variation in H_1^*/H: Welfare



(F) Variation in ϕ's: Research



(G) Variation in ϕ's: Welfare

FIGURE A7: Impulse response functions to a technology shock.

The spillover in research comes from all researchers, of nests and varieties, and hence doesn't change with the reallocation of research away from balanced growth. It's straightforward to confirm that the bgp doesn't change with this re-definition except that equation (19) $N(t) = H_2(t)H(t)^{\nu_m}(g_H + \phi)^{-\nu_m}\tilde{f}_2^{-1}$.[47] Since we identify the level of $N(t)$, not its components H_2 or f_2, the two models cannot be separately identified.

This baseline is in Figure A7(B) and (C). The anticipated rise in $Z(t)$ lowers the value of nests in (C.12), so researchers shift from nests toward varieties (panel B). The variety boom raises welfare modestly and early; the offsetting decline in nests is larger and more persistent—nests do not die and are less substitutable—so the net income gain peaks at 0.04, below the 0.05 direct effect (panel C). The reallocation is stronger when nest researchers are a smaller share of the total: as H_1^*/H rises, $\hat{H}_2$ deviates by more, nests fall by more, and the income gain shrinks (panels D and E).

Allowing knowledge spillovers among nest researchers ($\phi > 0$, panels F and G) makes the boom self-reversing. A larger ϕ amplifies the initial reallocation, because the fall in $\widehat{\widehat{H}}_2$ lowers the efficiency of inventing nests in (6). For small ϕ (0 or 0.01) research returns gradually to the BGP, but the protracted decline in nest-researcher productivity generates an endogenous downturn after the boom ($\widehat{C}(t) < 1$ after $t \approx 2015$ in panel G). For large ϕ, a single shock sets off several cycles: lower nest-researcher productivity reduces competition across nests ($NZ^{\eta-\sigma}$ in (C.12)), pulling research back toward nests between $t \approx 2005$ and 2060 and then toward varieties again; at $\phi = 0.05$ the downturn and the second cycle exceed the original boom.

Mapped to history, the model associates these computing breakthroughs with the productivity boom of the late 1990s and early 2000s in every case and—when there are positive externalities among nest researchers (basic research, panels F and G)—with the productivity slowdown after 2008, though the model's downturn begins somewhat later than in the data.[48]

D. Census of Manufactures

Historical data from the manufacturing census provide information value added, number of workers, and number of establishments per industry and geographic unit.

[47] If the allocation of research doesn't affect the spillovers, the social planner's allocation becomes:

$$\frac{H_1}{H_2} = \frac{\sigma - 1}{\eta - \sigma}$$

[48] The reversals in real income in Figure 12(G) are primarily driven by productivity, and the slowdown starts in 2008 when $\phi = 0.05, 0.04, 0.03$, and between 2009 and 2019 in the other cases. Syverson (2017) surveys the literature. Jorgenson, Ho, and Stiroh (2008), Byrne, Oliner, and Sichel (2013), and Fernald (2015) also associate both the productivity boom (1995–2004) and slowdown (2005-) to information technology. Similar to Gordon (2016), the slowdown in Figure 12(G) arises from a protracted period of fewer breakthroughs with high productivity potential.

These Censuses were conducted in 1870, 1880, 1900, 1909, 1919, 1929, 1939. The geographic unit was county in 1870, 1880, 1929, and city in 1880, 1900, 1909, 1919, 1929, 1939.[49] The industry classification to the 1950 classification used by the Integrated Public Use Microdata Series (IPUMS).

We compare the aggregate, industry level-data in the two years with data both at the city and county levels, 1880 and 1929. We find that relative to the city data, the county data have systematically higher figures for total value added, number of workers and number of establishments by industry, presumably because rural areas were excluded from city censuses. So, when presenting changes below, we use censuses with the geographic unit for each decade.

We aggregate data to the industry level, merge across years and merge with patent data to create an unbalanced panel of 47 industries and seven years 1870, 1880, 1900, 1909, 1919, 1929, 1939. Of these industries, 46 appear in all years and industry code 377 "aircraft and parts" appears only in 1929 and 1939.

D.1 A description of the Census data

Table A6 compares the auto-correlation between the value added in an industry and its lagged value, with the correlation between the number of patents in an industry and its lagged value. The autocorrelation of patent data for different time lags is similar to those of Figure 8.[50] For the patent data, the number of observations are much larger because the industry is more disaggregated and there are more years of data. As a result, the correlation is precisely estimated. In the Census data, measurement error may lead to an attenuation bias decreasing the estimated autocorrelation. Still, the auto-correlation in both value added and in number of patents declines with the time lag, and the pace of the decline is similar.

D.2 Establishments in the model and data

Table A8 decomposes the growth and level of value added as

$$\text{value added} = \#\text{ establishments} \times \frac{\text{value added}}{\#\text{ establishments}}$$

Growth in the number of establishments accounts for almost 60 percent of the variation in value added growth across sectors, and establishments account for almost 80 percent of the variation of value added across sectors in a cross section. We do not explicitly model establishments.

[49] We also have data from the 1860 data but we don't use them since we use patent data from 1875-2015.

[50] The numbers are slightly different because in Figure 8 we separately calculate the autocorrelation for each pair $t, t-$ lag and plotted the mean. Table A6 reports the correlation in the pooled data using only data for the turn of decades.

TABLE A6: Autocorrelation of value added and number of patents

	log(value added) in Census data			log(# patents)		
	autocorr.	std. err.	# obs.	autocorr.	std. err.	# obs.
time lag	(1)	(2)	(3)	(4)	(5)	(6)
10	0.71	0.05	241	0.92	0.004	8039
20	0.64	0.06	194	0.84	0.006	7400
30	0.62	0.06	192	0.76	0.008	6757
40	0.55	0.07	146	0.67	0.009	6116
50	0.47	0.09	98	0.59	0.011	5476
60	0.34	0.10	96	0.51	0.012	4849
70	0.26	0.14	48	0.43	0.014	4227

Note: Column (1) presents the correlation between the vector of value added (in logs) of a Census and the Census of year = t - time lag. Column (2) presents the standard error of this correlation, and column (3) the number of observations used to calculate it. When the Census time lag is 9 years, we round the lag to ten years. The number of observations with 70 year time difference is larger than the 46 industries cited in the because the table uses Census data before merging them with patent data. Columns (4)-(6) presents the same moments for the vector of number of patents (in logs) across decades of patent data.

To link the model to establishment data, assume that an establishment may produce multiple varieties within a nest, but cannot produce varieties from multiple nests. For a nest with n varieties at time t, the number of varieties per establishment is Poisson with parameter n^α where $\alpha \in (0,1)$. Then, the expected number of varieties per establishment is

$$\frac{n^\alpha}{1-e^{-n^\alpha}} \tag{D.13}$$

and the number of establishments is:

$$n^{1-\alpha}\left[1-e^{-n^\alpha}\right] \tag{D.14}$$

Both are strictly increasing in n, and the number of establishments is always smaller than the number of varieties n. With $\alpha = 0.77$, the model exactly matches the decomposition of value added in Panel A of Table A8, where establishments account for 0.595 of value added growth and establishment size accounts for the remaining 0.405. Out of sample, the model predicts that variation in the number of establishments per sector accounts for 0.66 of the variation in levels of value added. In the data, this number is 0.786. That is, in both the data and the model, establishments account for a larger share of variation in the level than the growth rate of value added, but in the data, the difference is larger.

Table A7 decomposes growth in employment in CBP data as:

$$\#\text{ workers} = \#\text{ establishments} \times \frac{\#\text{ workers}}{\#\text{ establishments}}$$

TABLE A7: Decomposition of employment in CBP data

Panel A: Growth regressions		
	Δ log # establishments	Δ log(workers/estab)
Δ log(# establishments)	0.553***	0.447***
	(0.059)	(0.059)
Observations	12,503	12,503
R-squared	0.473	0.452
Industry FE	291	291
Year FE	52	52
Panel B: Level regressions		
	log # establishments	log(workers/estab)
log(# establishments)	0.814***	0.186***
	(0.005)	(0.005)
Observations	12,735	12,735
R-squared	0.965	0.908
Industry FE	291	291
Year FE	52	52

Note: *** p<0.01, ** p<0.05, * p<0.1

The decomposition yields strikingly similar results to the decomposition of data on value added from the Census of Manufactures, in equation (D.13) and A8.

D.3 Value added per worker in Census

The CBP data only report number of workers, number of establishments by industry-location pair. The Census data report in addition the total value added. In this section, we drop the assumption that value added per worker is equal across sectors.

We start by reporting in Table A9 the following decomposition of the value added in an industry-year, into the following components:

$$\text{value added} = \#\text{ establishments} \times \frac{\#\text{ workers}}{\#\text{ establishments}} \times \frac{\text{value added}}{\#\text{ workers}}$$

Panel A presents the growth regressions, all variables are in log changes, $\log(x_t/x_{t-1})$. In panel B, presents the level regressions, $\log(x_t)$. Growth in number of establishments account for almost 60% of value added growth, whereas the other two components, number of workers per establishment and value added per worker, account for about 20% of growth each.

In the main text, we do not count research value as part of a sector value added, and take the output of a sector to be proportional to the number of varieties. Here, to capture differences in value added per worker across industries, we modify the model to take the output of creators of varieties as part of value added in an industry. The BEA

TABLE A8: Decomposition of value added I

Panel A: Growth regressions	Δ log # establishments	Δ log(VA/establishement)
Δ log(value added)	0.595***	0.405***
	(0.0466)	(0.0466)
Observations	291	291
R-squared	0.689	0.386
Industry FE	51	51
Year FE	6	6
Panel B: Level regressions	log # establishments	log(VA/establishement)
log(value added)	0.786***	0.214***
	(0.0416)	(0.0416)
Observations	346	346
R-squared	0.852	0.699
Industry FE	51	51
Year FE	7	7

Note: *** p<0.01, ** p<0.05, * p<0.1

TABLE A9: Decomposition of value added in the Census of Manufactures

Panel A: Growth regressions			
	log # establishments	log(workers/estab)	log(VA/worker)
log(value added)	0.595***	0.221***	0.184***
	(0.0466)	(0.0424)	(0.0244)
Observations	291	291	291
R-squared	0.689	0.311	0.478
Industry FE	51	51	51
Year FE	6	6	6
Panel B: Level regressions			
	log # establishments	log(workers/estab)	log(VA/worker)
log(value added)	0.786***	0.127***	0.0870***
	(0.0416)	(0.0386)	(0.0186)
Observations	346	346	346
R-squared	0.852	0.566	0.816
Industry FE	51	51	51
Year FE	7	7	7

Note: *** p<0.01, ** p<0.05, * p<0.1

accounts for intellectual capital as part of value added since 2008. Before, spending on R&D was accounted for as part of firms' costs.

With intellectual capital, the per period budget in (8) becomes:

$$\dot{B}(t) = rB(t) + Y(t) - \bar{C}(t) - I(t)$$

where $I(t)$ is investment in intellectual capital. In equilibrium, condition (ii) has the additional condition $I(t) = H(t)w_H(t)$, and investors are indifferent at the margin between consumption, investing in bonds or *in intellectual capital*.

To map this modified model to the data, we derive the value added and number of workers in a nest, where value added includes the value of research directed to create varieties in the nest and its workers $h(\bar{z},\tau,t)$.

Since all varieties get the same profits, the number of production workers in a nest is proportional to the number of varieties in a nest:

$$\ell(\bar{z},\tau,t) = \frac{n(\bar{z},\tau,t)}{\bar{N}(t)\bar{n}(t)}L(t).$$

The total value added in a nest $(\bar{z},\tau)$ at time t is the payments to production workers, $w\ell(\bar{z},\tau,t)$ plus profits $(\sigma-1)w\ell(\bar{z},\tau,t)$ plus the output of its researchers, $w_H h_1(\bar{z},\tau,t)$:[51]

$$\begin{aligned} VA(\bar{z},\tau,t) =& n(\bar{z},\tau,t)\frac{w(t)L(t)}{\bar{N}(t)\bar{n}(t)}\times \\ & \left\{\sigma + \frac{(1-\chi)(\sigma-1)}{(\rho+\delta+g_H)}\left[\delta - \nu_m g_H - \frac{\nu_z g_H}{\eta-\sigma} + \frac{1}{\eta-\sigma}\frac{\dot{\psi}(t-\tau)}{\psi(t-\tau)}\right]\right\} \end{aligned} \tag{D.15}$$

The total number of workers in the nest is[52]

$$\mathcal{L}(\bar{z},\tau,t) = \frac{L(t)n(\bar{z},\tau,t)}{\bar{N}(t)\bar{n}(t)}\left\{1 + \Upsilon_6\frac{H(t)}{L(t)}\left[\delta - \nu_m g_H - \frac{\nu_z g_H}{\eta-\sigma} + \frac{1}{\eta-\sigma}\frac{\dot{\psi}(t-\tau)}{\psi(t-\tau)}\right]\right\} \tag{D.16}$$

where

$$\Upsilon_6 = \frac{\Upsilon_4(1-\chi)}{\chi\rho + (\delta+g_H)[\chi+\Upsilon_4(1-\chi)]}.$$

[51] Substituting (28) and (25) into (28), we write:

$$w_H = \frac{(1-\chi)(\sigma-1)}{f_1(\rho+\delta+g_H)}\frac{w(t)L(t)}{\bar{N}(t)\bar{n}(t)}.$$

Multiplying by $h_1(\bar{z},\tau,t)$ in (22), we get the research term in value added.

[52] To write $\bar{N}\bar{n}/L$ as a function of primitives of the model, we use (27) and (30):

$$\frac{\bar{N}(t)\bar{n}(t)}{L(t)} = \frac{H_1(t)}{f_1(\delta+g_H)L(t)} = \frac{\Upsilon_4(1-\chi)}{f_1\{\chi\rho+(\delta+g_H)[\chi+\Upsilon_4(1-\chi)]\}}\frac{H(t)}{L(t)}.$$

Dividing (D.15) by (D.16), note that the value added per worker in a nest depends on the nest's age $t-\tau$ but not on its productivity $\bar{z}$ in $n(\bar{z},\tau,t)$.

Preliminaries We group the 650 patent classes into sectors, where the number of patent classes per sector m follows the following random process: $Pr(m=1)=0.2$, $Pr(m=2)=0.2$ and $Pr(m=\bar{m})=0.6/8$ for $\bar{m}=3,...,10$. We get 116 sectors from this grouping, each characterized by a collection of nests.

Combining equations (21) and (33), we write the number of varieties in nest $\omega=(\bar{z},\tau)$ as:

$$n(\omega,t) \propto a^C(\omega)e^{-g_H[\nu_m+\nu_z/(\eta-\sigma)](t-\tau)}\psi(t-\tau)^{1/(\eta-\sigma)} \tag{D.17}$$

where

$$a^C(\omega)=a(\omega)e^{g_H\tau}$$

where $a(\omega)$ and τ.

The term in square brackets in (D.15) and (D.16) is:

$$\left[\delta-\nu_m g_H-\frac{\nu_z g_H}{\eta-\sigma}+\frac{1}{\eta-\sigma}\frac{\dot{\psi}(t-\tau)}{\psi(t-\tau)}\right]=\delta+g_H-\beta_1+\frac{1}{\eta-\sigma}\frac{\dot{\psi}(t-\tau)}{\psi(t-\tau)}$$

Optimization Without loss of generality, we normalize $w(0)=1$ and $L(0)/[\bar{n}(0)\bar{N}(0)=1$. The levels of these variables don't matter since they enter log-linearly in VA and $\mathcal{L}$. The level of number of varieties in a nest in (D.17) matters because the number of establishments in a nest in (D.14) is a non-linear function. There are three parameters in the calculation of VA, $\mathcal{L}$ and number of establishments:

$$\varsigma_1=\frac{(1-\chi)(\sigma-1)}{(\rho+\delta+g_H)}$$
$$\varsigma_2=\Upsilon_6\frac{H(t)}{L(t)} \qquad \text{for all } t$$

which appear in (D.15) and (D.16). From (D.17), we calculate

$$n(\omega,t)=\varsigma_3 a^C(\omega)e^{-g_H(\nu_m+\nu_z/(\eta-\sigma))(t-\tau)}\psi(t-\tau)^{1/(\eta-\sigma)} \tag{D.18}$$

and $\varsigma_3=\alpha$. We pick the discount rate $\rho=0.05$.

For each guess of the parameters and each nest we calculate the value added in (D.15), number of workers in (D.16) and number of establishments in (D.14) using (D.18). We add across nests within sectors to get the sector-level variables, and calculate VA, number of workers per establishment, and value added per worker. We choose ς and α to minimize the squared distance between the data and the model for the components

of the growth decomposition in A9, panel A.

Table A10 displays the results. When we don't restrict the value of ς_1, the model perfectly matches the data with $\varsigma_1 = 7.9$ and $\varsigma_2 = 0$. But this value of ς_1 implies $\chi < 0$. When we set $\chi = 0$ and $H/L = 0$ so that $\varsigma_1 = 0.89$ and $\varsigma_2 = 0$, value added per worker accounts for only 2.9 percent of value added growth, much smaller than the 18 percent in the data. This parametrization puts an upper bound on the role of value added in growth. Increasing χ or H/L both decrease the role of value added per worker.[53]

Panel C shows the model's predictions for the decomposition in levels. Again, the model estimates the role of value added per worker. In the model and the data, the number of establishments account for a larger share of value added in levels than in growth, and the model is quantitatively not far from the data, irrespectively of whether or not we restrict χ.

TABLE A10: Decomposition of value added: model and data

Panel A: Estimated parameters	no restrictions on ς_1	restrict $\chi \in [0,1]$ in ς_1
ς_1	7.86	0.89
ς_2	0	0
α	0.85	0.85
implied χ	-7.86	0

Panel B Targeted moments: Decomposition of VA growth			
		Model	
	Data	unrestricted	$\chi \in [0,1]$
Number of establishments	0.595	0.595	0.673
Workers per establishment	0.221	0.221	0.299
Value added per worker	0.184	0.184	0.029

Panel C Untargeted moments: Decomposition of VA in levels			
		Model	
	Data	unrestricted	$\chi \in [0,1]$
number of establishments	0.786	0.737	0.766
workers per establishment	0.127	0.222	0.228
value added per worker	0.087	0.042	0.006

[53] Decreasing the discount rate to $\rho = 0.01$ increases the share of value added per worker to 0.038, still far from the data.

TABLE A11: Outcome growth and number of patents

Dependent variable: log change in the column outcome (yearly average)								
	Census of Mfg.		CBP					
	Value added		Employment		Payroll		Establishments	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log (average number of patents per year)	0.034**		0.015***		0.019***		0.017***	
	(0.013)		(0.004)		(0.006)		(0.004)	
log change in number of patents (yearly avg.)		0.230		0.031		0.010		0.059
		(0.165)		(0.041)		(0.065)		(0.045)
log outcome previous Census	−0.082***	−0.075***	−0.049***	−0.047***	−0.066***	−0.063***	−0.062***	−0.061***
	(0.0070)	(0.0069)	(0.0027)	(0.0027)	(0.0044)	(0.0043)	(0.0032)	(0.0032)
Observations	267	264	2,048	2,044	1,461	1,458	2,052	2,048
R-squared	0.644	0.621	0.407	0.402	0.350	0.345	0.387	0.381
Industry FE	47	47	253	253	252	252	253	253
Year FE	6	6	10	10	7	7	10	10

Note: Standard errors in parentheses. Each pair of columns reports, for a given outcome, the patent-level specification (odd columns) and the patent-growth specification (even columns), with industry and year fixed effects absorbed (`areg`). Columns (1)–(2) use Census of Manufacture data (dependent variable: log change in value added); columns (3)–(8) use County Business Patterns (CBP) data. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

E. County Business Patterns Data Harmonization

E.1 Industry concordances

To construct a consistent panel from the CBP files spanning 1946 to 2016,[54] we harmonize establishment counts and employment recorded under multiple historical industry classification systems into three target classifications: the U.S. Census 1950 Industry Classification (IND1950), the U.S. Census 2000 Industry Classification (IND2000), and the 2012 North American Industry Classification System (NAICS2012). Because CBP uses different classifications across this span — historical SIC and ICC1942 codes through 1997, and modern NAICS codes from 1998 onward — harmonization requires two complementary methodologies. For the early CBP files (1946–1996), we directly map each source classification to each target via an LLM-assisted three-stage pipeline (Section E.1.1 below). For the modern CBP files (1997–2016), we adopt a chain approach in which crosswalks from each NAICS vintage to the IND classifications are assembled through a sequence of pairwise concordances anchored at NAICS2012 as a hub (Section E.1.2 below). The two-pipeline structure is illustrated in Figure A8. The online appendix (Sections E.1.2 and E.1.2) documents both pipelines in full; here we describe the methodology.

[54] The harmonized CBP panel extends through 2021, but the analysis in the main paper uses CBP data only through 2016. This is because our patent-level analysis uses a forward five-year citation window, requiring patent data through 2021 to construct citation measures for patents filed in 2016.

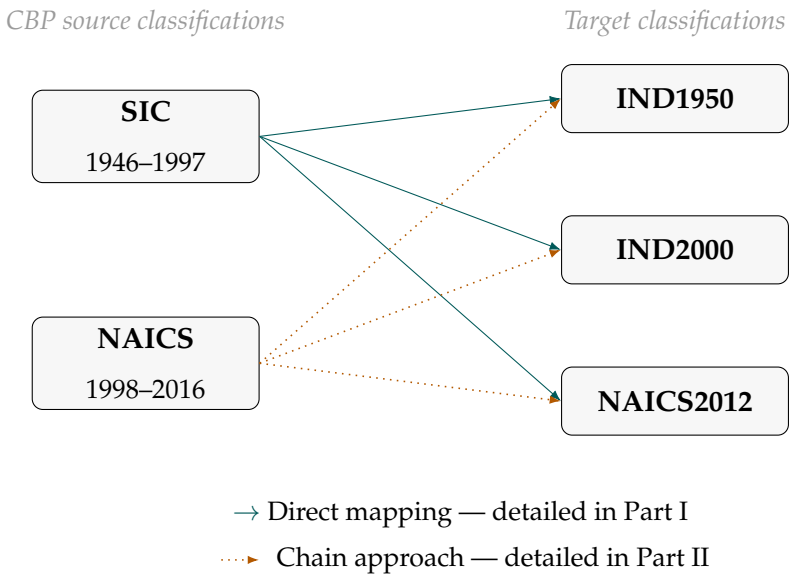


FIGURE A8: Two-pipeline structure for harmonizing CBP industry classifications, 1946–2016. Part I documents the direct mapping of SIC vintages and ICC1942 to all three target classifications. Part II documents the chain approach for modern NAICS vintages (1997–2017), which routes through NAICS2012 as a hub before continuing to IND2000 and IND1950 via the IPUMS and ACS-derived downstream mappings.

E.1.1 Direct mapping: SIC and ICC1942 to IND1950, IND2000, and NAICS2012

For the early CBP files, we crosswalk three sets of source classifications: (i) the 1942 Census of Business Industry Classification Code (ICC1942), which predates the SIC system; (ii) six vintages of the U.S. Standard Industrial Classification (SIC) system, namely SIC1945, SIC1949, SIC1957, SIC1967, SIC1972, and SIC1987; and (iii) a set of **special CBP codes** that appear alongside ICC and SIC codes in CBP annual files from 1946 to 1997 but do not conform to either hierarchy. The IND1950 and IND2000 crosswalks cover all six SIC vintages and ICC1942, while the NAICS2012 crosswalks cover only the four early vintages (SIC1945, SIC1949, SIC1957, SIC1967) and ICC1942, since SIC1972 and SIC1987 are already crosswalked to NAICS by Eckert, Fort, Schott, and Yang (2021a).

We adopt a **direct mapping** approach: each source classification is mapped independently to each target classification, rather than chained through intermediate vintages. This choice is driven by data availability. Full concordances between adjacent SIC vintages, and between ICC1942 and the earliest SIC, do not exist in published form, which precludes a chained construction. We instead build each source–target pair from primary sources, using an automated three-stage pipeline for the SIC and ICC1942 classifications and a manual lookup for the special CBP codes. Stages 1 and 2 are common across all three target classifications. Stage 3 differs by target: manual review

for IND1950 and IND2000 (Section E.1.2 of the online appendix), and an automated hierarchical resolver for NAICS2012 (Section E.1.2). We describe each in turn below.

Source and target classifications. The SIC and ICC1942 source code lists and industry titles are obtained from Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022), except for SIC1987, which we scrape from the Bureau of Labor Statistics. The IND1950 and IND2000 reference lists are scraped from IPUMS; the NAICS2012 reference list is downloaded from the U.S. Census Bureau and contains all 2,209 valid NAICS2012 codes across digit levels two through six. The target classifications are *coarser* than the source classifications for IND: IND1950 contains roughly 160 industries and IND2000 roughly 270, compared with about 670 codes in ICC1942 and roughly 1,000 in each SIC vintage, so most assignments are many-to-one. NAICS2012 is *finer*: each SIC digit level maps to a NAICS2012 code of a specific length (Table OA.5 of the online appendix), which permits one-to-many assignments where a single SIC industry spans multiple NAICS codes.

Stage 1: LLM-assisted mapping. For each source–target pair, we generate a `MAPPING` dictionary that assigns every source code to a target code, together with a match-quality label and explanatory notes. The mapping is constructed using a large language model (Claude), prompted with both the source and target industry titles. The model provides three capabilities that motivate this choice: (i) semantic understanding of industry descriptions, (ii) cross-classification reasoning between the source (SIC/ICC) and target (NAICS) hierarchies, and (iii) structural awareness of the hierarchical logic of the SIC, ICC, and NAICS systems. We classify each assignment into one of four match-quality categories: **Exact** when the source category maps cleanly to a single target label; **Good** when no exact equivalent exists and the closest available category is used; **Split** (or **Partial**) when the source spans multiple target codes; and **No match** when the source has no target equivalent. For NAICS2012, the `MAPPING` dictionary additionally allows one-to-many assignments with a flag field `is_primary`, with a primary code and one or more secondary codes per source entry. The Stage 1 output is a raw CSV per source–target pair containing the source code, source title, assigned target code, match quality, and notes.[55]

Stage 2: Hierarchical consistency flagging. Stage 2 reads the Stage 1 raw CSV and examines whether assignments are internally consistent within the source classifica-

[55]Two implementation details. First, the IND1950 pipeline uses the original SIC1987 code list, in which Division A (Agriculture) three-digit codes are recorded without leading zeros (e.g., "13" for Field Crops). This produces a collision with two-digit major-group codes in other divisions (e.g., "13" for Oil and Gas Extraction under Mining), which we resolve via a division-aware lookup keyed on `(division_letter, code)` tuples. The IND2000 pipeline uses an updated SIC1987 list with leading zeros restored, eliminating the collision. Second, for the NAICS2012 pipeline, the ICC1942 Stage 1 script is generated programmatically from the SIC1949 mappings (for the 458 ICC1942 codes that share a code number with SIC1949) plus an `OVERRIDES` dictionary for the 205 codes unique to ICC1942.

tion's own hierarchy. The SIC and ICC numeric hierarchies share a common three-level structure: two-digit major groups, three-digit industry groups, and four-digit industries. We define a code P as the direct parent of code C when C starts with P and is exactly one digit longer. We then flag each code according to four rules:

1. **Needs-Assignment**: the parent code has no Stage 1 assignment.
2. **Consistent**: the parent's assigned target code matches at least one of its children's target codes.
3. **Parent–Child Contradiction**: the parent's target code matches none of its children's target codes.
4. **Child–Parent Divergence**: the child's target code differs from its parent's target code.[56]

For NAICS2012 specifically, the parent–child consistency has a stricter form: because NAICS itself is hierarchical, a child's NAICS code is consistent with its parent's only when the parent's NAICS code is a prefix of the child's. Therefore, unlike the IND system, we enforce hierarchical consistency for NAICS2012. Stage 2 is purely diagnostic in all three pipelines: no assignments are altered in this stage. The Stage 2 output appends two columns to the Stage 1 file recording the rule applied to each code and supporting detail (the children compared, the modal assignment among them, or the diverging codes observed).

Stage 3 for IND1950 and IND2000: structured human review. Because the IND1950 and IND2000 target classifications are not hierarchical, hierarchical consistency cannot be enforced algorithmically; flags are retained for transparency and inspection. Codes flagged as `Parent-Child Contradiction` or `Child-Parent Divergence` are reviewed manually and corrected as described above; codes flagged as `Needs-Assignment` are filled with the modal child code. A central design principle is that Stage 1 assignments

[56] Parent–Child Contradiction is evaluated from the parent down: it triggers when the parent's target code matches none of its children's target codes. Child–Parent Divergence is evaluated from each child up: it triggers whenever an individual child's target code differs from its parent's target code. Consider, for example, the 2-digit SIC code 50 ("Wholesale Trade"), whose 3-digit children are 501 ("Wholesale Motor Vehicles"), 502 ("Wholesale Furniture"), and 503 ("Wholesale Lumber"). If all three children are assigned to "Wholesale Trade" but the parent is assigned to "Retail Trade," the parent matches no child and is flagged `Parent-Child Contradiction`, while each of the three children simultaneously diverges from the parent and is flagged `Child-Parent Divergence`. The two flags are handled differently in manual review. Parent–Child Contradictions typically signal a Stage 1 error in the parent's assignment, since a higher-level summary should reflect at least one of its constituent children; in such cases, the parent is corrected, often by modal substitution (here, "Retail Trade" would be replaced by "Wholesale Trade," the modal target across the children). Child–Parent Divergences are typically legitimate features of a many-to-one mapping — a parent representing a sector at coarser target resolution will often disagree with sub-industries that fall under different target categories — and are preserved unless review reveals a Stage 1 misassignment.

are otherwise never automatically overridden on the basis of hierarchical disagreement alone, since such disagreements may reflect genuine structural differences between the source and target classifications rather than errors. Corrections are made directly in the Stage 2 output. See Section E.1.2 of the online appendix for the full review protocol.

Stage 3 for NAICS2012: automated hierarchical resolver. Because NAICS2012 is itself hierarchical, parent–child consistency can be enforced algorithmically rather than by manual review. Stage 3 is implemented as an iterative two-pass resolver. *Pass 1* sweeps bottom-up: for each parent SIC code at the three- and two-digit level, we collect the NAICS codes assigned to its direct children and use them to correct the parent — dropping the parent's NAICS entries that disagree with its children, or, if all of the parent's entries disagree, replacing the entire mapping with the modal child NAICS at the appropriate digit length. *Pass 2* sweeps top-down: for each parent, we identify children whose NAICS code falls outside the parent's NAICS subtree and remap each such child to the best-matching candidate within that subtree, where the best match is selected by word-overlap scoring between the source industry title and each candidate NAICS title. The two passes alternate until both produce zero changes. After Stage 3, between 99.3 and 99.8 percent of codes are flagged as `Consistent` across the five source classifications, and every change is logged in a `Correction` column for traceability. The full resolver logic and per-vintage convergence statistics are reported in Section E.1.2 (Table OA.8) of the online appendix.

Special CBP codes. Approximately fifty administrative codes appear in early CBP files but not in any SIC or ICC1942 manual — county totals, top-level industry aggregates, and codes for administrative establishments. Because of their irregular structure and limited number, these codes are crosswalked through a manually constructed lookup table that assigns each to IND1950 and IND2000 directly; codes that span multiple target industries are left unassigned. For the mappings to NAICS2012, we referenced the material provided by Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022). See Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022) for a detailed taxonomy.

Output. The pipeline produces a set of crosswalk files for each source–target pair: fourteen for the IND targets (seven source classifications $\times$ two targets) and five for NAICS2012, plus the special-codes lookup table. Each crosswalk file contains the source code, source title, assigned target code, target title, match-quality label, hierarchy flag, and notes; for NAICS2012, an additional `Correction` column logs every change made by Stage 3.

E.1.2 Chain approach: modern NAICS vintages to IND2000 and IND1950

For modern CBP files, which use NAICS classifications from 1997 onward, we construct crosswalks linking each NAICS vintage — NAICS1997, NAICS2002, NAICS2007, NAICS2012, and NAICS2017 — to IND2000 and IND1950. Pairwise concordances among adjacent NAICS vintages from 1997 to 2012 are taken from Eckert, Fort, Schott, and Yang (2021a); what is missing is the link from NAICS2012 onward to the Census IND classifications. We supply this missing link by constructing two new objects: a NAICS2012-to-IND2000 crosswalk and a NAICS2017-to-NAICS2012 concordance. Combined with an IND2000-to-IND1950 mapping extracted from American Community Survey microdata, these objects form a single chained concordance from each NAICS vintage to IND1950 anchored at NAICS2012 as a hub:

$$\text{NAICS}_{1997} \to \text{NAICS}_{2002} \to \text{NAICS}_{2007} \to \text{NAICS}_{2012} \to \text{IND2000} \to \text{IND1950},$$

with NAICS2017 entering through a backward link to NAICS2012. The full construction is documented in Sections E.1.2–E.1.2 of the online appendix.

NAICS2012-to-IND2000 hub crosswalk. We invert the IPUMS IND2000-to-NAICS2012 crosswalk to obtain initial NAICS2012-to-IND2000 assignments for the codes that appear explicitly in the IPUMS file. For NAICS codes not covered by inversion — predominantly 2- through 4-digit parent codes and a residual set of 6-digit codes — we exploit the hierarchical structure of NAICS: each unmapped code inherits the IND2000 assignment of its closest mapped ancestor, with closer ancestors overriding more distant ones. After inversion and propagation, we manually review the remaining 426 unmapped codes (predominantly high-resolution 6-digit codes whose semantic distance from any covered ancestor is meaningful) and supply IND2000 assignments where a clear correspondence exists. The final crosswalk covers 2,206 of the 2,213 NAICS2012 codes (99.7 percent); see Section E.1.2 of the online appendix for full details.

NAICS2017-to-NAICS2012 concordance. The link from NAICS2017 back to NAICS2012 is not available as a single file. We construct it from the NAICS Association's 2017 changes documentation, which provides 1,069 direct 6-digit mappings between the two vintages. We then promote these 6-digit mappings to parent codes (2- through 5-digit) via a strict unanimous-child rule: a parent receives a mapping only if all of its 6-digit children agree on the corresponding 2012 parent code. This conservative rule preserves consistency with the underlying 6-digit data and avoids introducing ambiguous many-to-many mappings at parent levels. Of the 2,003 codes in the full NAICS2017 list, 2,001 receive a NAICS2012 mapping; see Section E.1.2 of the online appendix.

IND2000-to-IND1950 mapping. The mapping closing the chain is extracted from the 2000 American Community Survey, in which respondents' industry of employment is coded simultaneously under IND and IND1950. After deduplication on observed (IND2000, IND1950) pairs, the resulting mapping covers all 266 IND2000 codes and is many-to-one (Section E.1.2).

Assembly and coverage. The full concordance is assembled by sequentially joining the six pairwise mappings via outer joins, preserving codes appearing on only one side rather than silently dropping them. From the assembled master table, we extract individual crosswalk files for each NAICS vintage. Every NAICS vintage achieves over 99 percent end-to-end coverage to IND1950, ranging from 99.05 percent for NAICS2017 to 99.31 percent for NAICS1997 (Table OA.14 of the online appendix). The resulting crosswalks are used to construct the harmonized CBP panel that underlies the analysis in Section 5.

Online Appendix: Threading County Business Patterns, 1946–2016

Methodology and Replication Guide

Version 5.0

May 5, 2026

Abstract

This appendix describes the construction of industry concordances used to harmonize the U.S. Census Bureau's County Business Patterns (CBP) panel from 1946 through 2016. Because the CBP files use historical SIC classifications through 1996 and modern NAICS classifications from 1997 onward, harmonization requires two complementary methodologies. **Part I** documents the direct mapping of six SIC vintages (1945, 1949, 1957, 1967, 1972, 1987) and the 1942 Census of Business Industry Classification Code (ICC1942) to three target classifications: the U.S. Census 1950 and 2000 Industry Classifications (IND1950 and IND2000) and the 2012 North American Industry Classification System (NAICS2012). The mappings are constructed via a three-stage automated pipeline that uses a large language model (Claude) to generate initial assignments and enforces hierarchical consistency through structured human review (for IND targets) or an automated two-pass resolver (for NAICS2012). **Part II** documents the chain approach for modern NAICS vintages, in which crosswalks from NAICS1997 through NAICS2017 reach IND1950 and IND2000 through a chain of pairwise concordances anchored at NAICS2012 as a hub, consistent with Eckert, Fort, Schott, and Yang (2021a). Both methodologies are fully replicable from the provided code and source files.

1. Introduction

The County Business Patterns (CBP) is the longest-running establishment-level dataset published by the U.S. Census Bureau, with annual coverage from 1946 to the present. A central obstacle to its use in long-run research is that the underlying industry classification has changed several times: the early files (1946 through 1997) use the U.S. Standard Industrial Classification (SIC) and its predecessor the 1942 Census of Business Industry Classification Code (ICC1942), while files from 1998 onward use the North American Industry Classification System (NAICS), itself revised in 2002, 2007, 2012, and 2017. Linking establishments across this entire span on a common industry basis

requires a unified system of crosswalks that map every source classification to a small set of stable target classifications.

This appendix describes such a system. We harmonize seven source classifications — ICC1942, the six SIC vintages spanning 1945–1987, and a set of **special CBP codes** that appear in CBP annual files from 1946 to 1997 — into three target classifications: the U.S. Census 1950 Industry Classification (IND1950), the U.S. Census 2000 Industry Classification (IND2000), and the 2012 North American Industry Classification System (NAICS2012). For modern CBP files (1998 onward), we additionally provide crosswalks from each NAICS vintage (NAICS1997, NAICS2002, NAICS2007, NAICS2012, NAICS2017) to IND2000 and IND1950, and the 2012 North American Industry Classification System (NAICS2012).

1.1 Two methodologies

The construction is divided into two methodologically distinct pipelines.

Part I — Direct mapping. For the SIC and ICC1942 sources, we map each source classification independently to each target classification using a three-stage automated pipeline. Stage 1 generates a candidate mapping using a large language model (Claude), which provides semantic understanding of industry descriptions, cross-classification reasoning between source and target hierarchies, and structural awareness of the SIC and ICC numeric hierarchies. Stage 2 examines within-source hierarchical consistency and flags codes for review. Stage 3 corrects flagged codes: for the IND targets (which are flat classifications), via structured human review; for NAICS2012 (which is itself hierarchical), via an automated two-pass resolver that enforces parent–child consistency algorithmically. The IND1950 and IND2000 crosswalks cover all six SIC vintages and ICC1942. The NAICS2012 crosswalks cover only the four early SIC vintages (SIC1945, SIC1949, SIC1957, SIC1967) and ICC1942, since SIC1972 and SIC1987 are already crosswalked to NAICS by Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022).

Part II — Chain approach. For the modern NAICS sources, we adopt a chain approach in which each NAICS vintage reaches the IND target classifications through a sequence of pairwise concordances anchored at NAICS2012 as a hub. The pairwise concordances among adjacent NAICS vintages from 1997 to 2012 are taken from Eckert, Fort, Schott, and Yang (2021a); we construct two new objects to complete the chain: a NAICS2012-to-IND2000 crosswalk obtained by inverting and hierarchically propagating the IPUMS IND2000-to-NAICS2012 crosswalk, and a NAICS2017-to-NAICS2012 concordance obtained from the NAICS Association's published changes between the two vintages. An IND2000-to-IND1950 mapping extracted from American Community Survey microdata closes the chain.

The pipeline structure is shown in Figure OA.1. Both methodologies reach all three target classifications: SIC and ICC1942 sources reach the targets directly (Part I, teal), while modern NAICS sources reach them through the chained pairwise concordances anchored at NAICS2012 (Part II, coral).

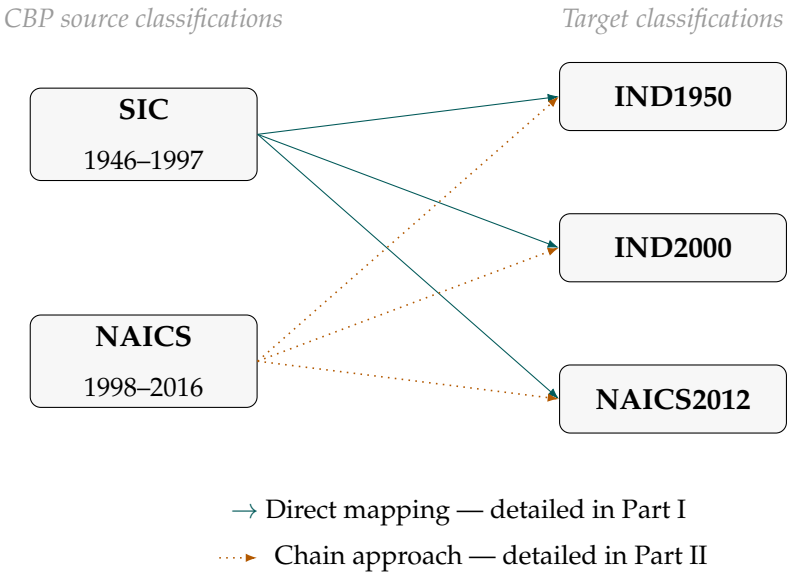


FIGURE OA.1: Two-pipeline structure for harmonizing CBP industry classifications, 1946–2016. Part I documents the direct mapping of SIC vintages and ICC1942 to all three target classifications. Part II documents the chain approach for modern NAICS vintages (1997–2017), which routes through NAICS2012 as a hub before continuing to IND2000 and IND1950 via the IPUMS and ACS-derived downstream mappings.

1.2 Document organization

The remainder of the document is organized as follows. Section 2 describes the source data common to both pipelines. Sections 3–11 contain Part I (direct mapping). Sections 12–19 contain Part II (chain approach). Section 20 lists the references.

2. Source Data

This section provides a unified inventory of the source data used in both pipelines. Files used only in Part I are described in Sections 2.1–2.4. Files used only in Part II are described in Sections 2.5–2.7. Both parts share Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022) as the source for the SIC code lists used in Part I and Eckert, Fort, Schott, and Yang (2021a) as the source for the pairwise NAICS concordances used in Part II.

2.1 ICC1942 Classification File

`ICC1942.csv` (used in Part I). ICC1942 predates SIC and does not include manufacturing codes. Manufacturing was classified under separate census schedules. The file lists ICC1942 codes (with leading zeros) and industry titles, including letter-coded division headers (e.g., A, B). It is downloaded from Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022).

2.2 SIC Classification Files

`SIC{year}.csv` (used in Part I). The IND1950 and IND2000 crosswalks draw on six SIC source files (1945, 1949, 1957, 1967, 1972, 1987); the NAICS2012 crosswalks use only the first four. SIC1945 and SIC1949 are complementary: together they span the full economy for the late 1940s — SIC1945 covers manufacturing and SIC1949 covers non-manufacturing. Files for SIC1945 through SIC1972 are downloaded from Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022). SIC1987 is scraped by the researchers from the Bureau of Labor Statistics.

2.3 Special CBP Industry Codes

`special_industry_codes.xlsx` (used in Part I). 50 special industry codes appear in CBP annual files from 1946 to 1997 alongside standard SIC and ICC codes. These codes do not conform to the standard SIC or ICC1942 hierarchies. As described in Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022), they fall into four types (Table OA.1).

TABLE OA.1: Special CBP codes.

Type	Format	Years Active	Description
County totals	`--`	1946–1997	Four dashes indicating county totals.
Top-level industries	`XX-` (e.g., `20-`)	1946–1997	2-digit code followed by two dashes. Represents top-level industries such as Mining and Manufacturing. Most have no IND or NAICS mapping as they span multiple sub-industries.
Administrative and auxiliary industries	`XXX/` (e.g., `098/`)	1949–1997	3-digit code followed by a slash. Represents auxiliary establishments (administrative offices, warehouses, etc.) associated with an industry.
Other non-standard industries	3- or 4-digit year-specific codes	1946–1974	Defined by CBP to represent industries not classified elsewhere or a group of closely related industries (e.g., `099` for CBP1946, `330` for CBP1947–1950).

2.4 Target Classification Reference Files

`IND1950_list_classified.csv` and `ind2000_codes.csv` are scraped from IPUMS and contain the complete IND1950 and IND2000 code lists with industry titles. `naics_2012.xls` is downloaded from the U.S. Census Bureau and contains all 2,213 valid NAICS2012 codes and their official titles across digit levels two through six.

The IND targets are coarser than the source classifications: IND1950 contains roughly 160 industries and IND2000 roughly 270, compared with about 670 codes in ICC1942 and roughly 1,000 in each SIC vintage. Most assignments to IND targets are therefore many-to-one. NAICS2012 is finer: each SIC digit level maps to a NAICS2012 code's digit is longer (e.g., a 2-digit SIC major group maps to a 3-digit NAICS code, and a 4-digit SIC industry maps to a 6-digit NAICS code; see Table OA.5 for the full digit-level correspondence). This permits one-to-many assignments where a single SIC industry spans multiple NAICS codes.

2.5 Pairwise NAICS Concordances

Four pairwise concordances link adjacent NAICS vintages (used in Part II): `full_naics97_naics02.cs` `full_naics02_naics07.csv`, `full_naics07_naics12.csv` (downloaded from Eckert, Fort, Schott, and Yang (2021a)'s `Final Industry Concordance Files`), and `2017_to_2012_NAICS-Cha` (downloaded from the NAICS Association). The first three are ready-to-use pairwise concordances; the fourth is a raw changes file used as input to the construction in Section 14.

2.6 IPUMS IND2000-to-NAICS2012 Crosswalk and NAICS2017 Code List

`ind_indnaics_crosswalk_2023.xlsx` (used in Part II) is published by IPUMS and maps each IND2000 code to one or more NAICS2012 codes. Section 13 inverts this relationship to produce the NAICS2012-to-IND2000 crosswalk used as the hub of the chain.

`naics2017.txt` (used in Part II) is downloaded from the U.S. Census Bureau and contains the complete NAICS2017 code list at every digit level (2- through 6-digit), totaling 2,003 unique codes.

2.7 IND2000–IND1950 Microdata

`IND2000-IND1950-ACS.csv` (used in Part II) contains observed (IND2000, IND1950) pairs extracted from the 2000 American Community Survey microdata, downloaded from IPUMS USA. It is used in Section 15 to construct the IND2000-to-IND1950 mapping that closes the chain.

Part I — Direct Mapping

SIC and ICC1942 to IND1950, IND2000, and NAICS2012

3. Part I Pipeline Overview

Stages 1 and 2 are identical across all three target classifications. Stage 3 differs because of a structural property of the targets: NAICS2012 is itself hierarchical (a code's parent is the prefix obtained by removing its last digit), so parent–child consistency between source and target can be enforced algorithmically. IND1950 and IND2000 are flat classifications, so consistency must be reviewed manually. Special CBP codes are handled separately via a manual lookup table for all three targets.

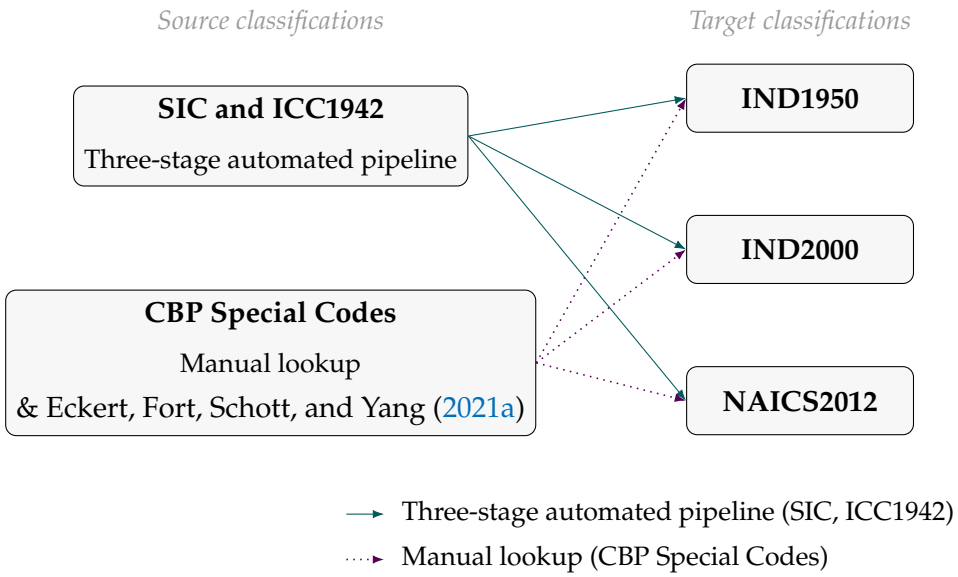


FIGURE OA.2: Direct mapping of source classifications to target classifications (Part I). SIC and ICC1942 sources are mapped independently to each target classification via a three-stage automated pipeline (solid teal arrows). The CBP Special Codes are mapped via a manual lookup table (dotted violet arrows).

3.1 Pipeline Summary

Table OA.2 summarizes the pipeline for both target families.

TABLE OA.2: The three-stage pipeline of ICC/SIC-to-target crosswalk construction.

Stage	Script / Action	Output
1	`build_crosswalk_{SRC}_{TGT}.py` — one per source × target pair (Claude-generated `MAPPING` dictionary).	`crosswalk_{SRC}_{TGT}.xlsx` and `{SRC}_{TGT}_raw.csv`
2	Hierarchical consistency flagging. `build_crosswalks.py` (SIC, IND targets) or `build_crosswalk_{SRC}_{TGT}_stage2.py` (ICC1942 and NAICS pipeline).	`{SRC}_{TGT}.xlsx` with `Hierarchy_Flag` column populated
3 (IND)	Structured human review. Not automated.	Corrected assignments saved in Stage 2 output Excel files
3 (NAICS2012)	Automated two-pass resolver: `resolve_hierarchy.py`. Single script processes all five source vintages.	`{SRC}_NAICS2012_corrected.xlsx` with all hierarchy corrections applied

{SRC} = source classification (SIC1945…SIC1987, or ICC1942). {TGT} = target (IND1950, IND2000, or NAICS2012).

Special CBP codes are crosswalked via a separate manual lookup table: `special_industry_codes.xls` (§7).

3.2 Repository Structure

TABLE OA.3: Repository structure.

Type	Path	Description
Code	`{TGT}/build_crosswalk_SIC{yr}_{TGT}.py`	Stage 1 scripts for SIC vintages, one per source–target pair.
	`{TGT}/build_crosswalk_ICC1942_{TGT}.py`	Stage 1 script for ICC1942.
	`NAICS2012/gen_icc1942.py`	Generator that produces the ICC1942 Stage 1 script for the NAICS2012 pipeline (NAICS2012 only).
	`{TGT}/build_crosswalks.py`	Stage 2 for all SIC vintages in the IND pipelines (runs Stage 1 internally).
	`{TGT}/build_crosswalk_{SRC}_{TGT}_stage2.py`	Standalone Stage 2 scripts (ICC1942 in IND pipelines; all sources in NAICS pipeline).
	`NAICS2012/resolve_hierarchy.py`	Stage 3 automated resolver (NAICS2012 only).
Data	`data/SIC/`	Source SIC code lists (CSV and XLSX).
	`data/ICC/`	Source ICC1942 code list.
	`data/IND/`	Target classification reference lists (IND1950, IND2000).
	`data/NAICS/naics_2012.xls`	NAICS2012 reference code list.
	`data/special_industry_codes.xlsx`	Manual crosswalk for special CBP codes.
Output	`{TGT}/{SRC}_{TGT}.xlsx`	Final IND crosswalk outputs (Stage 2).
	`NAICS2012/{SRC}_NAICS2012_corrected.xlsx`	Final NAICS2012 crosswalk outputs (Stage 3).

4. Stage 1 — Mapping Scripts

Stage 1 scripts assign each source code to the best-fitting target classification category based on a review of the source and target industry descriptions. The mapping scripts are built using Claude, a large language model — a transformer-based neural network trained on a large corpus of text — which provides (1) semantic understanding of industry descriptions, (2) cross-classification reasoning, and (3) structural awareness, such as understanding the hierarchical logic of the SIC and ICC systems. These assignments form the core component of the crosswalk construction; all subsequent stages are quality-control wrappers.

4.1 MAPPING Dictionary

For the IND pipelines, each Stage 1 script contains a `MAPPING` dictionary of the form

```
MAPPING = {source_code:  (target_code, match_quality, notes), ...}
```

For the NAICS2012 pipeline, each entry permits one-to-many assignments:

```
MAPPING = {source_code:  [(naics_code, is_primary, match_quality,
notes), ...]}
```

The first NAICS entry is always the primary mapping; any additional entries are secondary. Match-quality values are consistent across source classifications, with terminology differences across the three targets (Table OA.4).

TABLE OA.4: Types of match quality.

Value	IND2000	IND1950	NAICS2012
Exact	Source category maps cleanly to one target label with no ambiguity.	Same.	Same.
Good	No exact equivalent; closest available category used.	Same.	Same.
Split / Partial	Source spans multiple IND2000 codes; most representative code given; all spanned codes listed in Notes.	Source spans multiple IND1950 codes; single best-fit chosen.	Source does not map cleanly to any single NAICS2012 code; the closest is given as the primary mapping, with secondary rows added for the remaining matches. Also applied by Stage 3 Pass 1 (Section E.1.2) when a parent's assignment is replaced via modal substitution.
No match	No IND2000 equivalent.	No IND1950 equivalent.	No NAICS2012 equivalent.

4.2 SIC1987 Division Handling (IND1950 only)

The two IND target pipelines use different versions of the SIC1987 code list, which accounts for a structural difference in how Division A (Agriculture) codes are handled.

The IND1950 crosswalk was built against the original SIC1987 code list, in which Division A 3-digit codes were recorded without leading zeros (e.g., "13" for Field Crops, "16" for Vegetables and Melons). This creates a collision: the stripped forms of Division A 3-digit codes are identical to 2-digit major group codes in other divisions. For example, "13" also identifies Oil and Gas Extraction under Division B (Mining), and "16" identifies Heavy Construction under Division C. Without further disambiguation, a lookup on "13" is ambiguous. The IND1950 script resolves this via `DIV_MAPPING`, a dictionary keyed on `(division_letter, code)` tuples, so that `("A", "13")` and `("B", "13")` route to different IND codes.

The IND2000 crosswalk was built against an updated version of the SIC1987 code list in which leading zeros had been restored for Division A codes (e.g., "013", "016"). With padded codes, no collision exists and no division-aware disambiguation is needed. The

IND2000 script therefore has no `DIV_MAPPING` step. The NAICS2012 pipeline does not use SIC1987 at all and is unaffected.

Replication note. To reproduce the original crosswalk results, the unpadded SIC1987 code list should be used as input for the IND1950 pipeline and the padded version for the IND2000 pipeline. Using the incorrect version will produce incorrect Division A assignments in IND1950 or spurious no-match errors in IND2000.

4.3 NAICS Digit-Level Correspondence (NAICS2012 only)

Each SIC digit level maps to a NAICS2012 code of a specific length (Table OA.5). This correspondence is used in Stage 1 (to validate `MAPPING` assignments) and again in Stage 3 (to determine the target digit length for modal substitution and candidate sets).

TABLE OA.5: Expected NAICS digit length by SIC digit level.

SIC Digit Level	Expected NAICS Length	Example
Division (letter)	2-digit (or compound)	SIC "E" → NAICS "48-49" (Transportation & Warehousing)
2-digit	3-digit	SIC "42" → NAICS "484" (Truck Transportation)
3-digit	4-digit	SIC "421" → NAICS "4841" (General Freight Trucking)
4-digit	6-digit	SIC "4213" → NAICS "484122" (General Freight Trucking, Long Distance, LTL)

4.4 ICC1942 Generator Script (NAICS2012 only)

Unlike the SIC vintages and the IND pipelines, which each use a single hand-written Stage 1 script, the NAICS2012 pipeline for ICC1942 involves two scripts. The first, `gen_icc1942.py`, is a code generator: it reads the SIC1949 raw CSV (for the 458 ICC1942 codes that share a code number with SIC1949) and an `OVERRIDES` dictionary defined in `prelim_mapping_ICC1942.py` (for the 205 codes unique to ICC1942), validates all NAICS codes against `naics_2012.xls`, and writes the Stage 1 script `build_crosswalk_ICC1942_NAICS2012.py`. The second script is then run to produce the raw CSV and styled reference workbook. Replication therefore requires running only `build_crosswalk_ICC1942_NAICS2012.py`; `gen_icc1942.py` is needed only if the Stage 1 script itself must be regenerated from scratch.

4.5 Division Code Handling (NAICS2012 only)

Division-level codes (single-letter codes, e.g., A, B) appear in SIC1949, SIC1957, SIC1967, and ICC1942. They are present in the Stage 1 `MAPPING` for completeness but are excluded from all Stage 2 and Stage 3 processing in the NAICS2012 pipeline. They are dropped at load time by the resolver and do not appear in the Stage 3 output, since a Division-level NAICS assignment is too broad to constrain its children reliably.

4.6 Output

Each Stage 1 script produces two files in its directory (`{TGT}/`):

- `crosswalk_{SRC}_{TGT}.xlsx` — styled reference workbook with Crosswalk and Summary sheets. Human-readable documentation of the raw mappings.
- `{SRC}_{TGT}_raw.csv` — machine-readable flat file consumed by Stage 2. Columns: `SIC_Code` (or `ICC_Code`), `SIC_Title` (or `ICC_Title`), `Target_Code`, `Match_Quality`, `Notes`. The NAICS2012 pipeline additionally includes an `Is_Primary` column to support one-to-many assignments.

5. Stage 2 — Hierarchical Consistency

Stage 2 reads the raw CSV from Stage 1, examines within-vintage hierarchical consistency, and writes the flagged crosswalk Excel file. The procedure is identical for all source classifications and all three target classifications. In the IND pipelines, Stage 2 is the final automated step; in the NAICS2012 pipeline, Stage 2 is purely diagnostic, with all corrections deferred to Stage 3.

5.1 Stage 2 Scripts

TABLE OA.6: Stage 2 scripts.

Source	Script	Notes
SIC1945–SIC1987 (IND targets)	`{TGT}/build_crosswalks.py`	Single script processes all six SIC vintages. Invokes each Stage 1 script as a subprocess, then applies the consistency rules.
ICC1942 (IND targets)	`{TGT}/build_crosswalk_ICC1942_{TGT}_stage2.py`	Standalone script for ICC1942 only. Reads the existing Stage 1 raw CSV directly; does not re-run Stage 1.
All sources (NAICS2012)	`NAICS2012/build_crosswalk_{SRC}_NAICS2012_stage2.py`	Standalone script per source vintage; reads the Stage 1 raw CSV directly.

5.2 Phase A: Load Raw Mappings

Stage 2 reads the raw CSV and constructs a working DataFrame with source-specific code and title columns, plus a `Digit_Level` column derived from the length of each code string: purely alphabetic codes (SIC1949 division letters A/B; ICC1942 division letters A–J) are classified as "Letter"; numeric codes as 2-digit, 3-digit, or 4-digit.

5.3 Phase B: Hierarchical Consistency Rules

ICC and SIC classifications share the same three-level numeric hierarchy: 2-digit major group, 3-digit industry group, and 4-digit industry. A code P is the direct parent of code C when C starts with P and $\text{len}(C) = \text{len}(P) + 1$. Letter codes are excluded from Phase B.

Phase B applies four rules bottom-up (from 4-digit to 3-digit to 2-digit), then a second pass over all children (Table OA.7).

TABLE OA.7: Hierarchical consistency rules.

Condition	Action	Hierarchy_Flag
Parent has no assigned target code	Assign modal (most frequent) child target code	`Assigned-from-Children` (IND); `Needs-Assignment` (NAICS)
Parent's assigned target code matches at least one of its children's target codes	Keep parent unchanged	`Consistent`
Parent's assigned target code matches none of its children's target codes	Keep parent unchanged (no override at Stage 2)	`Parent-Child Contradiction`
Child's target code differs from its parent's target code	Keep child unchanged at Stage 2	`Child-Parent Divergence`

For NAICS2012, parent–child consistency has a stricter form: a child's NAICS code is consistent with its parent's only when the parent's NAICS code is a prefix of the child's. For example, NAICS 484 (Truck Transportation) is consistent with NAICS 48 (Transportation and Warehousing).

Key design principle. A parent's Stage 1 assignment is never overridden at Stage 2 solely because it differs from the modal child, as such differences may be valid due to structural differences between source and target. Parents receive an assignment only when none exists, avoiding mechanical corrections without contextual understanding. `Parent-Child Contradiction` flags economically meaningful inconsistencies for downstream review (manual in IND, algorithmic in NAICS) but does not trigger automatic correction at Stage 2. `Child-Parent Divergence` captures legitimate within-group heterogeneity without altering either assignment.

5.4 Output

Stage 2 produces the flagged crosswalk Excel file, `{SRC}_{TGT}.xlsx`, by augmenting the Stage 1 raw CSV with two new columns:

- `Hierarchy_Flag` records the outcome of Phase B processing for each code (except for letter-coded division headers).
- `Hierarchy_Notes` provides supporting detail, including the child codes compared, modal assignment chosen, or diverging target codes observed.

6. Stage 3 — Correction

Stage 3 corrects the assignments flagged by Stage 2. The procedure depends on whether the target classification is itself hierarchical. IND1950 and IND2000 are flat classifications, so consistency cannot be enforced algorithmically and is reviewed manually (Section E.1.2). NAICS2012 is hierarchical, so consistency can be enforced algorithmically by an iterative two-pass resolver (Section E.1.2).

6.1 Stage 3 for IND1950 and IND2000: Structured Human Review

Stage 3 for the IND targets is a structured human review step and is identical for all source classifications. Corrections are made directly in the Stage 2 output Excel files; Stage 1 scripts are not modified. Because IND1950 and IND2000 are not themselves hierarchical, hierarchical consistency is not strictly enforced in the final mapping results; flag variables are retained for transparency and inspection.

What to review.

- **Parent-Child Contradiction**: the parent's assignment is from a different economic sector than all of its children. May indicate an incorrect Stage 1 assignment, or a legitimate edge case (e.g., a major-group header that straddles two sectors).
- **Child-Parent Divergence**: a 4-digit code maps to a different target category than its 3-digit parent. Most cases are legitimate, but some may indicate errors in the original assignment.

How to correct.

1. Open the relevant `{SRC}_{TGT}.xlsx` file produced by Stage 2.
2. Locate the flagged row.
3. Update the target code, target label, and/or `Match_Quality` cells as appropriate.
4. Save and archive the corrected file. Stage 2 should not be re-run after manual corrections, as this would overwrite changes.

Output. Reviewed and corrected versions of the `{SRC}_{TGT}.xlsx` files.

6.2 Stage 3 for NAICS2012: Automated Hierarchical Resolver

Stage 3 for the NAICS2012 pipeline is implemented in `resolve_hierarchy.py`. It reads the Stage 1 raw CSV for each source vintage, enforces hierarchical consistency between

parent and child NAICS assignments, and writes a corrected output Excel file with a full log of every automated change. Because NAICS2012 is hierarchical, parent–child consistency has a precise definition: a child's NAICS assignment is consistent with its parent's if the child's NAICS code has the parent's NAICS as a prefix.

The resolver consists of two passes that alternate until convergence.

Pass 1: Bottom-up correction. Pass 1 fixes *Parent-Child Contradiction*, where a parent's NAICS code is not an ancestor of any of its children's NAICS codes. It sweeps bottom-up: 3-digit SIC codes first, then 2-digit SIC codes. For each parent, the resolver collects the NAICS codes assigned to all of its direct children and classifies each of the parent's NAICS entries as consistent or inconsistent with those children. If some entries are consistent and others are not, the inconsistent entries are dropped; if all entries are inconsistent, the entire mapping is replaced by the modal child NAICS at the appropriate digit length (Table OA.5), and `Match_Quality` is set to `Partial`. The modal code must exist as a valid NAICS2012 code; if not, the next most common is tried. If no valid code is found, the parent is left unchanged. Compound NAICS parents (`31-33`, `44-45`, `48-49`) are intentional cross-sector assignments and are never replaced; division-level SIC codes are excluded entirely.

Pass 2: Top-down correction. Pass 2 fixes *Child-Parent Divergence*, where a child's NAICS code is not a descendant of its parent's NAICS code. It sweeps top-down: 2-digit SIC codes first (correcting their 3-digit children), then 3-digit SIC codes (correcting their 4-digit children). For each parent, the resolver builds the set of valid candidate NAICS codes: all NAICS2012 codes of the correct digit length that are descendants of any of the parent's NAICS assignments. Each child entry that is inconsistent with all parent NAICS codes is remapped to the best-matching candidate within the parent's subtree. If the new code duplicates an existing entry, it is dropped instead of remapped. Children with no available candidates in the parent subtree are preserved with a warning appended to the `Notes` field. Division-level SIC codes are not used as authoritative parents, since their NAICS assignments are too broad to reliably constrain their children.

Semantic matching. When Pass 2 needs to remap a child's NAICS entry to a candidate within the parent's subtree, the best candidate is selected by word-overlap scoring between the source industry title and each candidate NAICS title. Stopwords such as "and," "of," "the," "manufacturing," and "services" are removed before comparison. A small bonus is awarded when the source title is a catch-all ("not elsewhere classified," NEC) and a NEC candidate is available; a penalty is applied when a specific source is matched to a NEC candidate. The candidate with the highest score is selected; ties are broken by the lower NAICS code number. If no candidate achieves a meaningful word overlap, the resolver falls back to the most specific NEC candidate in the set; if no NEC

candidate exists, the original NAICS code is retained.

Convergence. The resolver alternates Pass 1 and Pass 2 in a nested-loop structure. Each pass runs to internal convergence (zero changes) before the other begins; the outer loop terminates when both passes produce zero changes in the same iteration. Convergence is reached within three outer iterations for all five vintages (Table OA.8). After Stage 3, between 99.3% and 99.8% of codes are flagged as `Consistent` across the five source classifications. Every change is logged in a `Correction` column for traceability.

TABLE OA.8: Stage 3 resolver convergence statistics, by source classification.

Source	Total rows	Primary codes	Codes corrected	Outer iterations	Final consistency
SIC1945	657	633	146	2	654/657 (99.5%)
SIC1949	881	846	181	3	879/881 (99.8%)
SIC1957	1,514	1,461	397	3	1,504/1,514 (99.3%)
SIC1967	1,498	1,448	400	2	1,494/1,498 (99.7%)
ICC1942	701	663	189	2	698/701 (99.6%)

ICC1942-specific handling. ICC1942 uses the same resolver logic as the SIC vintages, with one structural difference: the parent function must handle the non-contiguous Division H, whose 2-digit codes are {70,72–76,78–83,86,90} (manufacturing codes 18–39 are absent from ICC1942). Division codes are excluded from all resolver processing, consistent with the treatment of SIC division codes.

Output. The resolver writes one output file per source vintage: `{SRC}_NAICS2012_corrected.xlsx`. Each file contains four sheets: (i) *Crosswalk*, all source codes in depth-first hierarchical order with all columns from the Stage 1 raw CSV plus `Correction`, `Hierarchy_Flag`, and `Hierarchy_Notes`, color-coded by flag and correction status; (ii) *Summary*, count of rows by `Hierarchy_Flag` and count of primary rows by `Match_Quality`; (iii) *Corrections*, log of all codes that received at least one automated correction; and (iv) *Legend*, color-coding guide.

7. Special CBP Industry Codes — Manual Crosswalk

7.1 Background

CBP annual files from the Census Bureau contain, alongside standard SIC and ICC codes, a set of administrative codes that do not appear in any official SIC or ICC1942 manual. These codes represent county totals, top-level aggregates, auxiliary establishments, and other non-standard categories such as early tabulation groupings (Table OA.1). Due to their irregular structure and limited number, they are crosswalked

via a manually prepared lookup table rather than through Stages 1–3. The lookup table assigns each special code to IND1950 and IND2000 directly. Eckert, Fort, Schott, and Yang (2021a) also provides crosswalks from CBP special code to NAICS2012.

Output — The Lookup Table

`special_industry_codes.xlsx` contains one row per special code. The file structure is shown in Table OA.9.

TABLE OA.9: Structure of special CBP codes lookup table.

Column	Description
`Year`	Year range(s) in which the code appears in CBP files (e.g., "1946–1974," "1946 only").
`Code`	The special code as it appears in the CBP file.
`Industry`	Industry description from Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022).
`IND1950`	Manually assigned IND1950 code. Blank for county aggregate and top-level codes spanning multiple industries.
`IND1950_Industry`	IND1950 label corresponding to the assigned code.
`IND2000`	Manually assigned IND2000 code. Blank for county aggregate and top-level codes spanning multiple industries.
`IND2000_Industry`	IND2000 label corresponding to the assigned code.

8. Output File Column Reference

Table OA.10 describes the columns of the IND-target output files (`{SRC}_IND1950.xlsx`, `{SRC}_IND2000.xlsx`). Table OA.11 describes the columns of the NAICS2012 Stage 3 corrected files (`{SRC}_NAICS2012_corrected.xlsx`), which include additional columns to support one-to-many assignments and the correction log.

TABLE OA.10: Output file column reference — IND targets.

Column	Type	Description
`{SRC}_Code`	String	Source classification code. Leading zeros preserved where applicable. ICC1942 letter codes are single uppercase letters.
`{SRC}_Title`	String	Official source industry title.
`Digit_Level`	String	2-digit, 3-digit, 4-digit, or Letter.
`IND1950_Code / IND2000_Code`	Integer	Target classification code. Blank when no match. May be inferred by Phase B for parent codes that were originally unmapped.
`IND1950_Label / IND2000_Label`	String	Label for the target code from the reference CSV.
`Match_Quality`	String	Exact, Good, Split (IND2000) / Partial (IND1950), No match, or Inferred (Phase B).
`Hierarchy_Flag`	String	Blank, Consistent, Assigned-from-Children, Parent-Child Contradiction, or Child-Parent Divergence. Set by Stage 2.
`Hierarchy_Notes`	String	Details of hierarchy action: codes compared, child counts, etc.

TABLE OA.11: Output file column reference — NAICS2012 target (Stage 3 corrected files).

Column	Type	Description
{SRC}_Code	String	Source classification code (SIC_Code for SIC vintages, ICC_Code for ICC1942). Leading zeros preserved.
{SRC}_Title	String	Official source industry title.
Digit_Level	String	2-digit, 3-digit, 4-digit, or Division. Division rows are excluded from processing.
NAICS2012_Code	String	Assigned NAICS2012 code. Blank when no match. May include compound codes (31-33, 44-45, 48-49) at the Division level.
NAICS2012_Title	String	NAICS2012 industry title corresponding to NAICS2012_Code.
Is_Primary	Boolean	"True" for the primary (best-fit) assignment; "False" for secondary (one-to-many) assignments.
Match_Quality	String	Exact, Good, Partial, or No match. Partial is also applied by Stage 3 Pass 1 when a modal substitution is made.
Notes	String	Explanatory notes from the Stage 1 assignment. Stage 3 may append warnings for entries that could not be remapped.
Correction	String	Pipe-separated log of all Stage 3 corrections applied to this code (e.g., P1: [old] → [new] | P2: [old] → [new]). Empty if no correction was made.
Hierarchy_Flag	String	Consistent, Parent-Child Contradiction, Child-Parent Divergence, Needs-Assignment, or blank. Set after Stage 3 corrections are applied.
Hierarchy_Notes	String	Supporting detail for the Hierarchy_Flag: modal child code, parent NAICS codes compared, child count.

Source-Specific Coverage

Table OA.12 reports the number of source codes mapped to each target classification. Coverage is essentially complete across all source–target pairs: every SIC vintage and ICC1942 maps to over 99 percent of its codes in IND1950 and IND2000, and the five source classifications crosswalked to NAICS2012 each achieve at least 98.7 percent coverage. Unmapped codes are predominantly source codes too broad or residual to have a clean target equivalent (e.g., division-level aggregates, "not elsewhere classified" categories, and special CBP codes that span multiple sectors)

TABLE OA.12: Source-specific coverage (Part I).

Source	# of codes	Mapped to IND1950	Mapped to IND2000	Mapped to NAICS2012
ICC1942	672	672 (100%)	672 (100%)	663 (98.7%)
SIC1945	633	632 (99.8%)	632 (99.8%)	633 (100%)
SIC1949	846	845 (99.9%)	845 (99.9%)	846 (100%)
SIC1957	1,470	1,470 (100%)	1,470 (100%)	1,461 (99.4%)
SIC1967	1,457	1,457 (100%)	1,457 (100%)	1,448 (99.4%)
SIC1972	1,509	1,509 (100%)	1,509 (100%)	1,504 (99.7%)
SIC1987	1,447	1,447 (100%)	1,447 (100%)	1,443 (99.7%)
Special CBP codes	59	50 (84.7%)	50 (84.7%)	31 (52.5%)

Notes: For SIC1945, SIC1949, and special CBP codes, unmapped codes are those for industries too broad to have a reasonable equivalent in the target classification. SIC1972 and SIC1987 are not mapped to NAICS2012 directly: standard concordances are provided by Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022). NAICS2012 mapped counts refer to primary mappings only and are reported on the post-Stage 3 *primary codes*.

Discussion (Part I)

Two caveats are worth stating explicitly. First, the Stage 1 mappings are generated by Claude. Although the model performs well in semantic understanding and structural awareness of industry classifications, automated outputs can be incorrect, particularly for industries that changed scope across SIC vintages. The match-quality column makes the Stage 1 confidence explicit, allowing users to filter on Exact assignments if precision is critical. Second, the NAICS2012 resolver enforces parent–child hierarchical consistency by remapping divergent assignments to codes within the parent's NAICS subtree. This guarantees a coherent hierarchy in the output but can reduce semantic accuracy, since a child originally assigned the closest-meaning NAICS code may be remapped to a less semantically similar code. The `Correction` column logs every such change for auditability.

Part II — Chain Approach

NAICS1997 through NAICS2017 to IND1950, IND2000, and NAICS2012

9. Part II Overview

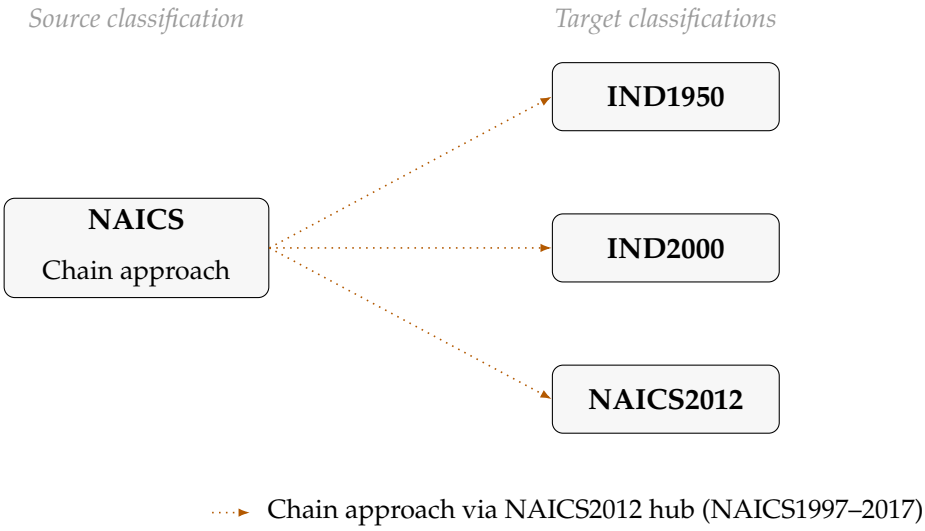


FIGURE OA.3: Chain approach for harmonizing modern NAICS source classifications (Part II). NAICS vintages from 1997 through 2017 are linked to each target classification via a sequence of pairwise concordances anchored at NAICS2012 as a hub (dotted coral arrows). The full construction is documented in Sections E.1.2–E.1.2.

Part I documented the direct mapping of historical SIC and ICC1942 source classifications to IND1950, IND2000, and NAICS2012. Part II addresses the complementary task of harmonizing the modern, NAICS-based portion of the County Business Patterns panel. We construct crosswalks that link every vintage of the North American Industry Classification System — NAICS1997, NAICS2002, NAICS2007, NAICS2012, and NAICS2017 — to IND2000 and IND1950.

The construction follows a *chain approach*, anchored at NAICS2012 as a hub. Pairwise concordances among adjacent NAICS vintages already exist as published files (Section 2.5); what is missing is the link from NAICS2012 onward to the Census IND classifications. We supply this missing link by constructing two new objects: a NAICS2012-to-IND2000 crosswalk obtained by inverting and hierarchically propagating the IPUMS IND2000-to-NAICS2012 crosswalk (Section 13), and a NAICS2017-to-NAICS2012 concordance obtained from the NAICS Association's published changes between the two vintages (Section 14). Combined with the NAICS-to-NAICS pairwise concordances of

Eckert, Fort, Schott, and Yang (2021a) and an IND2000-to-IND1950 mapping extracted from American Community Survey microdata (Section 15), these objects form a single chained concordance from each NAICS vintage to IND1950, with NAICS2012 as the connecting hub:

$$\text{NAICS}_{1997} \rightarrow \text{NAICS}_{2002} \rightarrow \text{NAICS}_{2007} \rightarrow \underbrace{\text{NAICS}_{2012}} \rightarrow \text{IND2000} \rightarrow \text{IND1950},$$

with NAICS2017 entering through a backward link to NAICS2012. The resulting crosswalks achieve over 99 percent end-to-end coverage from every NAICS vintage to IND1950 (Section 17).

The remainder of Part II is organized as follows. Sections 13–15 document the construction of each new mapping. Section 16 describes how the pairwise concordances are chained together. Section 17 reports coverage statistics. Section 18 discusses caveats. Section 19 contains brief replication notes.

10. Assembling the Chained Concordance

The full concordance is constructed by sequentially joining the six pairwise mappings on their shared key columns:

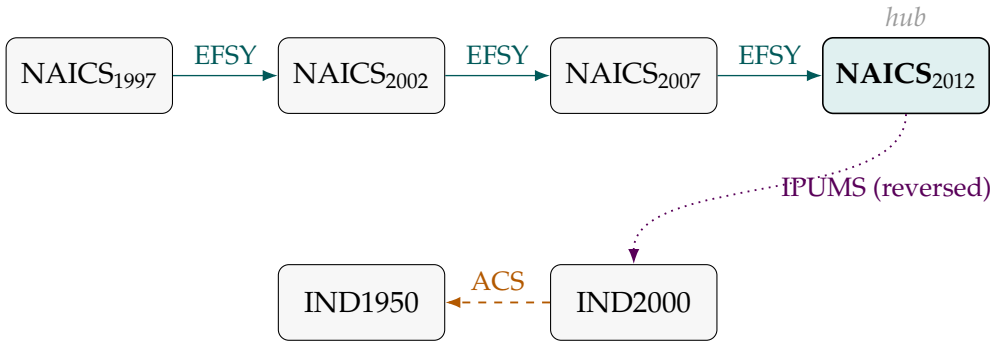


FIGURE OA.4: Chain structure for harmonizing modern NAICS source classifications. NAICS2012 serves as the hub: the three pairwise NAICS-to-NAICS concordances of **eckert2021imputing** (EFSY) link earlier vintages forward to NAICS2012, the inverted IPUMS IND2000-to-NAICS2012 crosswalk (Section E.1.2) links NAICS2012 to IND2000, and the 2000 American Community Survey link extends from IND2000 to IND1950 (Section E.1.2).

with NAICS2017 entering at the NAICS2012 hub via the link constructed in Section E.1.2. We use outer joins at every step so that codes appearing on only one side of a pairwise concordance are preserved rather than silently dropped; codes for which no path through the chain exists are retained with null entries downstream and reported as such.

Prior to joining, all code columns are normalized to a common string representation:

trailing decimal artifacts (e.g., "101.0" from float-typed columns) are removed, non-digit suffixes are stripped, and parent-code notation specific to NAICS2017 (dashes for 2-digit sectors, slashes for 3- through 5-digit groups) is reduced to pure numeric form. This normalization is necessary because the source files originate from different providers and use different conventions for representing identical codes.

The result is a wide table whose rows trace paths through the chain and whose columns hold the corresponding code in each NAICS vintage along with IND2000 and IND1950. From this master table, we extract individual crosswalk files for each NAICS vintage, each containing the (vintage code, IND2000, IND1950) triples deduplicated.[57]

10.1 Constructing the NAICS2012-to-IND2000/IND1950 Crosswalk

We construct two complementary mappings that together link NAICS2012 to both IND classifications. The first, from NAICS2012 to IND2000, requires the bulk of the construction effort and combines IPUMS data, hierarchical propagation, and manual review. The second, from IND2000 to IND1950, is extracted directly from American Community Survey microdata. We describe each in turn.

NAICS2012-to-IND2000 mapping

The IPUMS file maps each IND2000 industry to one or more NAICS2012 codes; we require the inverse direction, and we require coverage at every digit level of NAICS2012. Direct inversion of the IPUMS file falls short on both counts: many NAICS codes (particularly 2- through 4-digit parent codes and certain 6-digit codes) do not appear explicitly in the IPUMS crosswalk and therefore receive no IND2000 mapping. We address these gaps in three steps.

1. *Inversion.* The IPUMS crosswalk is reorganized into a dictionary keyed by NAICS2012 code. For each NAICS code that appears explicitly in the IPUMS file, the set of IND2000 codes mapped to it is recorded as the direct mapping. This step yields exact, IPUMS-sourced assignments for the subset of NAICS codes covered by the original file.

2. *Hierarchical propagation.* For NAICS codes not covered by inversion, we exploit the hierarchical structure of NAICS: a longer code is a child of any shorter code that forms its prefix (e.g., 1111 is a child of 111, which is a child of 11). We sort the full NAICS2012 code list by length and traverse it from shortest to longest. A code that already has a direct IPUMS mapping is left unchanged; a code without

[57] The chained construction and per-vintage extraction are implemented in the notebook `crosswalk_creation_chained.ipynb`. The output files are named `naics{vintage}_to_ind1950_chained.csv`.

a direct mapping inherits the IND2000 assignment of its closest ancestor that does. Inheritance is monotone in specificity: a closer (longer) ancestor's mapping always overrides an earlier one inherited from a more distant ancestor. For each (NAICS, IND2000) pair, we record the NAICS code that was the ultimate *direct* source of the mapping in a provenance column, allowing downstream users to assess the resolution at which each mapping was determined.

3. *Manual review.* After inversion and propagation, 426 of the 2,213 NAICS2012 codes (24 percent) remain unmapped. These are predominantly high-resolution 6-digit codes whose semantic distance from any covered ancestor is meaningful, and a residual set of broad sector aggregates (e.g., 11, 31, 99) for which no IND2000 equivalent exists. We review the unmapped codes manually, supplying IND2000 assignments where a clear correspondence exists. Where a single NAICS code straddles multiple IND2000 categories, all corresponding IND2000 codes are recorded as separate rows. After manual review and consolidation, 2,206 of the 2,213 NAICS2012 codes (99.7 percent) carry an IND2000 mapping.[58]

The IND2000-to-IND1950 Mapping Closing the chain requires a mapping from IND2000 to IND1950. We extract this mapping from the 2000 American Community Survey, in which respondents' industry of employment is coded simultaneously under IND and IND1950. After deduplication on the (IND2000, IND1950) pairs observed in the microdata, the resulting mapping covers all 266 IND2000 codes and assigns each to one of 133 IND1950 codes. Because IND1950 is the coarser classification, the mapping is many-to-one. Coverage is exact: every IND2000 code observed in the 2000 ACS receives an IND1950 assignment.

10.2 Constructing the NAICS2017-to-NAICS2012 Concordance

Unlike the earlier NAICS-to-NAICS concordances, the link from NAICS2017 back to NAICS2012 is not available as a single file. We construct it from the NAICS Association's changes documentation in two steps.

Direct 6-digit mappings. The changes file enumerates 1,069 six-digit NAICS2017 codes whose definitions changed between the 2012 and 2017 revisions, paired with their 2012 counterparts. These pairs constitute the foundation of the concordance and are tagged with provenance `direct_6digit`.

[58] The 20 NAICS codes that remain unmapped after manual review are predominantly 2-digit sector aggregates and residual or government categories (e.g., 95, 99) that have no natural equivalent in IND2000. The implementing notebook `naics12_to_ind2000_crosswalk_creation.ipynb` produces an intermediate file flagging these codes for review and a final consolidated crosswalk `naics12_to_ind2000_consolidated.csv`.

Unanimous-child promotion. The changes file covers only 6-digit codes, but the full NAICS2017 list contains 2- through 5-digit parent codes that also require mappings. We promote 6-digit mappings to parent codes via a strict unanimous-child rule. For each parent length $L \in \{5,4,3,2\}$, processed from longest to shortest, we group every 6-digit NAICS2017 code by its L-digit prefix and collect the set of distinct L-digit prefixes among the 2012 counterparts of its 6-digit children. If the set has exactly one element — that is, all of the parent's 6-digit children map to NAICS2012 codes sharing the same L-digit prefix — we record the parent-level mapping at that prefix. Otherwise no parent mapping is created. This rule preserves consistency with the underlying 6-digit data and avoids introducing ambiguous many-to-many mappings at parent levels. Promoted mappings are tagged with provenance `parent_unanimous_L{L}` for traceability.

Coverage. Of the 2,003 codes in the full NAICS2017 list, 2,001 receive a NAICS2012 mapping after these two steps. The two unmapped codes are the header row and code 99 ("industries not classified"), neither of which represents a substantive industry.[59]

11. Coverage

Table OA.13 reports coverage of each pairwise concordance — the share of unique source codes that successfully receive a mapping in the immediate downstream classification.

TABLE OA.13: Pairwise coverage of the chained concordance.

From	To	Source codes	Mapped	Coverage
NAICS1997	NAICS2002	2,323	2,323	100.0%
NAICS2002	NAICS2007	2,345	2,345	100.0%
NAICS2007	NAICS2012	2,332	2,332	100.0%
NAICS2017	NAICS2012	2,003	2,001	99.9%
NAICS2012	IND2000	2,213	2,206	99.7%
IND2000	IND1950	266	266	100.0%

The three Eckert et al. concordances among adjacent NAICS vintages each achieve full coverage, as does the IND2000-to-IND1950 mapping extracted from ACS microdata. The two new concordances we construct — NAICS2017-to-NAICS2012 and NAICS2012-to-IND2000 — both achieve coverage above 99.5 percent. The small residuals in each case are concentrated in non-substantive codes: the NAICS2017 residual is the header row and the "industries not classified" code 99; the NAICS2012 residual is a small set

[59] The construction is implemented in the notebook `NAICS17-12.ipynb` and produces the file `merged.csv`, which serves as the NAICS2017-to-NAICS2012 link in the assembled chain.

of 2-digit sector aggregates and government or residual categories with no IND2000 counterpart.

End-to-end coverage from each NAICS vintage to IND1950 is reported in Table OA.14. Every NAICS vintage achieves over 99 percent end-to-end coverage. The minor differences across vintages reflect the cumulative effect of the small gaps in the NAICS2012-to-IND2000 link and, for NAICS2017, the additional NAICS2017-to-NAICS2012 step.

TABLE OA.14: End-to-end coverage from each NAICS vintage to IND1950.

NAICS vintage	Coverage to IND1950
NAICS1997	99.31%
NAICS2002	99.23%
NAICS2007	99.10%
NAICS2012	99.10%
NAICS2017	99.05%

Part II Discussion

Three caveats are worth stating explicitly. First, hierarchical propagation in the NAICS2012-to-IND2000 construction assigns to each unmapped child code the IND2000 mapping of its closest mapped ancestor. The semantic distance between a 6-digit NAICS code and a 3-digit ancestor can be substantial, and propagated mappings are correspondingly less precise than direct mappings. The provenance column makes this explicit, allowing users to filter on direct mappings if precision is critical for a given application. Second, the unanimous-child rule used to construct the NAICS2017-to-NAICS2012 concordance is conservative: when a NAICS2017 parent's children map to multiple NAICS2012 parents, no parent-level mapping is created. The alternative — recording a many-to-many mapping at the parent level — would preserve coverage at the cost of introducing ambiguity that is difficult for downstream researchers to handle correctly. We prefer the loss of coverage. Third, the pairwise concordances preserve all many-to-many mappings rather than collapsing them to a single representative code. Researchers requiring a unique assignment per NAICS code should disambiguate using application-specific criteria such as employment-weighted modal mappings.

Part III — Assigning Weights

SIC Vintages to NAICS Vintages to NAICS2012

Part III Overview

Harmonizing the County Business Patterns (CBP) files from 1946 to 2016 requires resolving a long sequence of changes in the U.S. industrial classification system. The CBP has used six distinct industry classifications across this period: the Standard Industrial Classification (SIC) of 1945, 1949, 1957[60], 1967, 1972, 1977, and 1987, followed by the North American Industry Classification System (NAICS) revisions of 1997, 2002, 2007, and 2012. The very early CBP files (1946-1948) additionally report establishments under an industrial classification used by the 1942 Census of Manufactures, which we refer to as ICC1942. Each revision introduced reorganizations of varying severity, from minor relabelings to the wholesale restructuring that accompanied the 1997 transition to NAICS. Without a unified concordance, empirical work on long-run industry dynamics is forced either to restrict attention to a short subsample of the CBP, to aggregate to coarse classifications that mask within-sector variation, or to rely on piecemeal mappings that vary in methodology and quality across vintage pairs.

With the industry classification schemes harmonized to a common target, the seventy years of CBP data spanning different classification systems can now be studied within a unified analytical framework. This harmonization, however, is incomplete without a consistent method of assigning weights when industries split or merge across vintages. When a 1945 classification splits a single source industry into multiple successor industries by 1957, or when several 1949 categories are consolidated into a single 1957 code, an analyst must decide how to allocate the economic content of the source across its successors. This allocation choice is not innocuous: it determines what fraction of a source industry's measured employment, establishments, or payroll counts as contributing to each successor and ultimately to the modern NAICS2012 categories used in empirical work. Tracking the rise of new industries and the decline of obsolete ones over an seven-decade horizon requires that these allocations be transparent, internally consistent, and applied uniformly across every transition in the chain.

[60]The Census Bureau also issued a 1963 supplement to SIC1957 that modified approximately 97 four-digit codes and introduced 15 new ones; CBP files for the years 1964–1966 use this supplemented version. Because the supplement is a small, localized amendment of SIC1957 rather than a full revision, we do not treat it as a separate vintage. Instead, we handle the supplement-affected codes as a code-level substitution against SIC1957 at the CBP-harmonization step, as described in Section OA.0.4.

The problem is compounded by a second feature of the historical record: existing concordances cover the CBP only from 1957 forward. The two earliest SIC classifications — SIC1945 for manufacturing and SIC1949 for nonmanufacturing — and the still-earlier ICC1942 classification used in the 1946 CBP file have no published machine-readable concordance to SIC1957. Researchers who wish to extend their analyses into the early postwar period therefore face two problems simultaneously — the absence of a concordance for the 1946–1956 CBP files, and the absence of a uniform weighting framework that makes the surviving concordances composable across the full seven decades. Without addressing both, empirical work on long-run industry dynamics is forced either to restrict attention to a short subsample of the CBP, to aggregate to coarse classifications that mask within-sector variation, or to rely on piecemeal mappings whose methodology varies across vintage pairs.

We address both problems in this paper by constructing a chain of pairwise concordances and applying a single weighting rule throughout. The chain is the central data construction of this paper. The three new links we contribute extend the harmonized CBP to cover the 1946–1956 period that has lacked direct concordance pathways in prior work. This ten-year window is the golden era of American manufacturing: the postwar boom in which U.S. industrial production, employment, and household consumption of durable goods reached historic peaks before the long structural transformation of the late twentieth century. The absence of a four-digit concordance for these years has meant that fine-grained empirical study of American industry during its mid-century peak has been infeasible.

OA.0.2 The chain

Figure OA.5 summarizes the chain. The three early-vintage links that we construct (ICC1942→SIC1957, SIC1945→SIC1957 for manufacturing, and SIC1949→SIC1957 for nonmanufacturing) converge on SIC1957, which then becomes the starting point of the longer downstream chain through 1957, 1967, 1972, 1977, 1987, and the four NAICS vintages.

Table OA.15 reports the structure of each link. The number of source codes varies substantially across links. The SIC1957→SIC1967 and SIC1967→SIC1972 links are sourced from the EFSY concordance workbooks (`conc_sic57_sic67.xlsx` and `conc_sic67_sic72.xlsx`), which cover essentially the full four-digit universe of each vintage. The 1963 supplement to SIC1957, which modified roughly 97 codes and introduced 15 genuinely new ones, is not included in the EFSY workbooks: they map pre-supplement SIC1957 codes directly to SIC1967. ELMMSS provides a separate `SIC1963_Supplement.csv` file describing the SIC1957→SIC1963 changes, which we apply as a code-level substitution at the CBP-harmonization step rather than as an additional chain link; see Section OA.0.4.

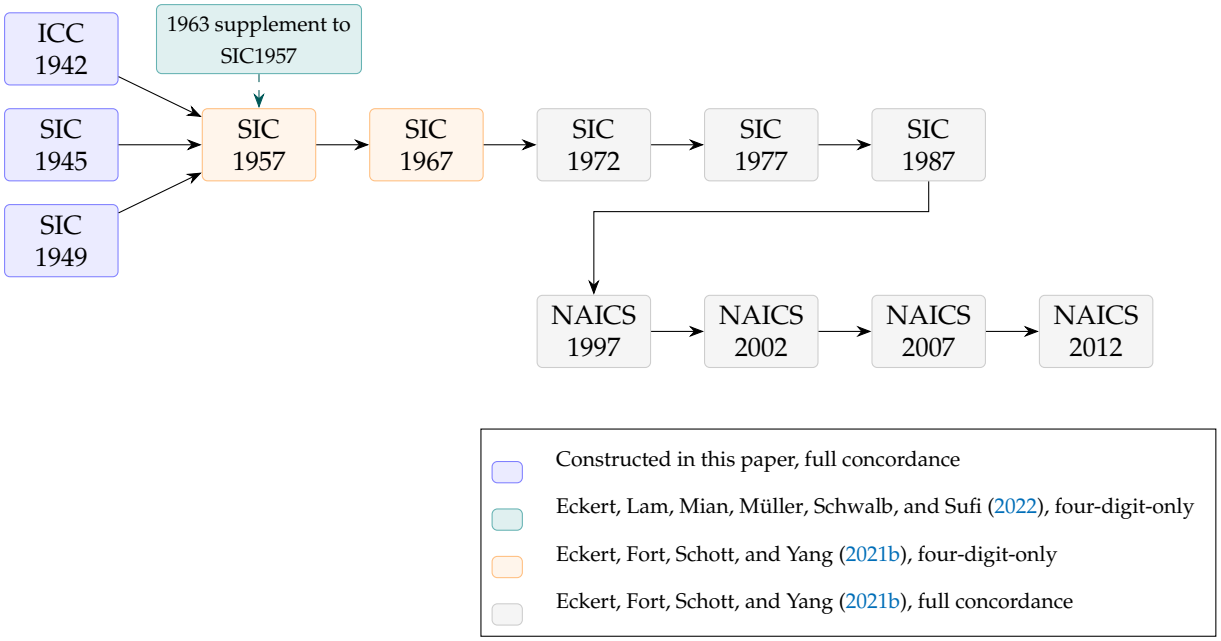


FIGURE OA.5: The industry concordance chain with three early-vintage entry points and the SIC1957 supplement. The three early-vintage links that we construct (ICC1942→SIC1957, SIC1945→SIC1957 for manufacturing, and SIC1949→SIC1957 for nonmanufacturing) extend the chain back to the 1946 CBP. The downstream links to NAICS2012 are all sourced from Eckert, Fort, Schott, and Yang (2021b). The 1963 supplement to SIC1957 (sourced from Eckert, Lam, Mian, Müller, Schwalb, and Sufi (2022)) is applied as a code-level substitution against the SIC1957 mapping at the CBP-harmonization step rather than as a separate chain link (dashed arrow); see Section OA.0.4. All chain links are composed under a single count-based weighting rule (see Section OA.0.5).

The proportion of source codes that correspond to a single target (1:1 mappings, denoted $n = 1$) is high across most links and falls only at three specific transitions: the 1967 revision, the 1987 revision, and the 1997 transition from SIC to NAICS, the last of which is the single most consequential break in the chain.

OA.0.3 Construction of the early-vintage crosswalks

The most substantive contribution of this paper is the construction of three new concordances at the earliest end of the chain: ICC1942→SIC1957 (for the 1942-vintage codes still present in the 1946-1948 CBP files), SIC1945→SIC1957 (for manufacturing industries from 1946 to 1956), and SIC1949→SIC1957 (for nonmanufacturing industries from 1946 to 1956). These three links extend the harmonized CBP back by roughly ten years and make the early postwar period available at the four-digit NAICS industry level for the first time.

TABLE OA.15: Structure of the chain links

Link	Sources	$n = 1$	Splits	Notes
ICC1942 → SIC1957	346	304	42	1942 Census vintage; constructed here
SIC1945 → SIC1957	462	433	29	Manufacturing-only; constructed here
SIC1949 → SIC1957	540	522	18	Nonmanufacturing-only; constructed here
SIC1957 → SIC1967	922	880	42	EFSY workbook
SIC1967 → SIC1972	917	783	134	EFSY workbook
SIC1972 → SIC1977	1,003	1,001	2	EFSY; minor revision
SIC1977 → SIC1987	1,004	885	119	EFSY; substantial revision
SIC1987 → NAICS1997	1,005	639	366	EFSY; SIC→NAICS transition
NAICS1997 → NAICS2002	1,169	1,071	98	EFSY; modest revision
NAICS2002 → NAICS2007	1,179	1,166	13	EFSY; minor revision
NAICS2007 → NAICS2012	1,175	1,169	6	EFSY; minor revision

Notes: "Sources" is the number of distinct source codes at the most-detailed level of the source vintage (four-digit SIC/ICC, six-digit NAICS). "$n = 1$" is the number of source codes that map to a single target. "Splits" is the number of source codes that map to multiple targets. The remaining sources have specialized characteristics (e.g., aggregate-level codes) and are not used in the most-detailed-level chain.

OA.0.3.1 The gap in the literature

Existing concordance work covers the U.S. industrial classification system from SIC1957 forward. Eckert, Lam, Mian, Müller, Schwalb, and Sufi (2022) provide concordances for the SIC1957→SIC1967 and SIC1967→SIC1972 transitions as part of their effort to digitize the early postwar CBP files.[61] Eckert, Fort, Schott, and Yang (2021b) released more complete concordances for the same two transitions that include explicit identity rows, and we use the EFSY versions in this chain. Eckert, Fort, Schott, and Yang (2021b) also provide concordances from SIC1972 onward through NAICS2012, drawing on Fort and Klimek (2018) for the SIC1972→NAICS1997 transition. The pre-1957 SIC vintages (the 1945 revision for manufacturing and the 1949 revision for nonmanufacturing) and the ICC1942 classification used in the 1946–1948 CBP files have no published machine-readable concordance to SIC1957. As a result, CBP files for the years 1946 through 1956 cannot be harmonized at the four-digit level without constructing such a concordance first.

The CBP files for this early period use ICC1942 for nonmanufacturing industries in the 1946–1948 files, SIC1945 for manufacturing industries (Divisions D through F in the 1957 numbering) across the 1946–1956 files, and SIC1949 for nonmanufacturing industries (Divisions A through C and G through I) in the 1951–1956 files. SIC1945 covers manufacturing throughout 1946–1956 and is complementary to whichever nonmanufacturing classification is in use that year. For the nonmanufacturing side, ICC1942 is used in 1946–1948 and is then succeeded by SIC1949 from 1951 onward,

[61] The ELMMSS files list only codes that changed between vintages, leaving unchanged codes implicit as identity mappings, and a few codes that did change are also omitted.

so these two vintages are sequential covers of the same set of industries rather than complementary. We treat all three as distinct links in the chain, all terminating at SIC1957.

OA.0.3.2 The two-stage pipeline

We construct each concordance through a pipeline of two stages. Stage 1 produces an LLM-generated mapping in which every source code is assigned a list of one or more (target code, primary-flag, quality label, note) tuples; multi-target "Split" sources are enumerated in full at this stage, with the primary target flagged so the canonical (s,t) pair is unambiguous. Stage 2 validates every (s,t) pair in the Stage 1 mapping using sentence-transformer embeddings, flagging entries where the LLM's quality label appears inconsistent with the semantic similarity of the source and target titles. The output of the pipeline is a structured dictionary mapping each four-digit source code to a list of target tuples, from which the count-based weights of equation (OA.1) are computed directly.

Stage 1: LLM-generated mapping with splits enumerated. We use Claude to produce the Stage 1 mapping. The prompt provides the model with the full ICC1942 (or SIC1945, or SIC1949) source code list and the full SIC1957 target code list, both with titles, and asks the model to produce a one-to-one or one-to-many mapping for every source code at every digit level (letter division, two-digit major group, three-digit industry group, and four-digit industry). The output is a Python dictionary where each entry takes the form

$$\begin{aligned} \texttt{"source\_code"}: \big[& (\mathit{target}_1, \mathit{is\_primary}_1, \mathit{quality}_1, \mathit{note}_1), \\ & (\mathit{target}_2, \mathit{is\_primary}_2, \mathit{quality}_2, \mathit{note}_2), \\ & \ldots \big] \end{aligned}$$

with *quality* taking one of four values: EXACT (the source and target codes have identical titles and scope), GOOD (the target is the most natural successor but with some change in title or scope), SPLIT (the source maps to multiple targets; the primary target carries *is_primary* = `True`), and NO MATCH (the source has no SIC1957 successor at the same digit level). The *note* field is free text in which the model justifies the assignment.

The Stage 1 prompt is constructed to constrain the model's output in ways that improve reliability. The prompt enforces a strict same-digit-level rule (each source code is mapped to a target at the same digit level rather than rolled up to a coarser parent or split down to a finer child), requires that every source code in the input list appear in the output dictionary (so coverage is complete by construction), asks the model to provide a note for every entry (so the rationale is auditable), and requires the model to emit *every* target code in the case of multi-target SPLIT sources (so the downstream

weighting computation does not depend on parsing free text out of the *note* field). These constraints turn what would otherwise be a free-form classification task into a structured labeling task with a fixed output schema.

Stage 2: Sentence-transformer validation. The Stage 1 output is reliable but requires validation: LLM outputs can contain plausible-sounding errors that are difficult to catch by visual inspection alone, particularly at scale, and the cost of incorrect mappings propagating through the chain is substantial. We use a sentence-transformer model (`all-mpnet-base-v2`) to embed each source and target title into a 768-dimensional vector space, then compute the cosine similarity between the embedded titles of *every* (*source*, *target*) pair in the Stage 1 mapping — primaries and non-primary split targets alike. The cosine similarity provides an independent semantic measure that is uncorrelated with the LLM's own quality label.

We compare the similarity score to the LLM's quality label and flag inconsistencies. Entries labeled EXACT should have similarity scores at or above 0.85; entries labeled GOOD should generally fall between 0.55 and 0.95; entries labeled SPLIT should remain above 0.50; entries labeled NO MATCH have no expected similarity range. Entries that fall outside these ranges are flagged for manual review. The thresholds are conservative: a flagged entry is not automatically incorrect, but the flag indicates that the LLM's confidence does not align with the embedded similarity and that the assignment warrants closer inspection.

The validator did not catch substantive errors in our particular case, but this should not be interpreted as evidence that LLM-generated concordances do not need validation. The result reflects three properties of our specific setting: the source code lists are small enough (under 900 entries per source vintage) that the LLM can attend to all of it; the source and target vintages are temporally close enough that most assignments are either Exact or Good; and the LLM had access to both classification systems' published documentation. Validation remains essential as standard practice: an unvalidated LLM mapping is not a defensible methodological choice, even when ex-post the LLM turns out to have been accurate.

OA.0.3.3 A worked example: the dry-goods and apparel case

To illustrate count-based composition on a many-to-many case, consider the SIC1949→SIC1957 transition for the two four-digit codes 5113 and 5132, both titled "Dry goods and apparel" in SIC1949. SIC1949 carried these two codes under separate wholesale channels (manufacturers' sales branches versus agents and brokers). SIC1957 abandoned the channel distinction at the four-digit level and split the dry-goods/apparel category into three product subgroups: 5032, 5035, and 5039. Each SIC1949 source thus maps

to the same three SIC1957 targets, and under the count-based weighting rule each (source, target) pair receives weight 1/3:

$$\begin{aligned} w(5113 \to 5032) &= 1/3, \\ w(5113 \to 5035) &= 1/3, \\ w(5113 \to 5039) &= 1/3, \\ w(5132 \to 5032) &= 1/3, \\ w(5132 \to 5035) &= 1/3, \\ w(5132 \to 5039) &= 1/3. \end{aligned}$$

Figure OA.6 shows the structure of this transition.

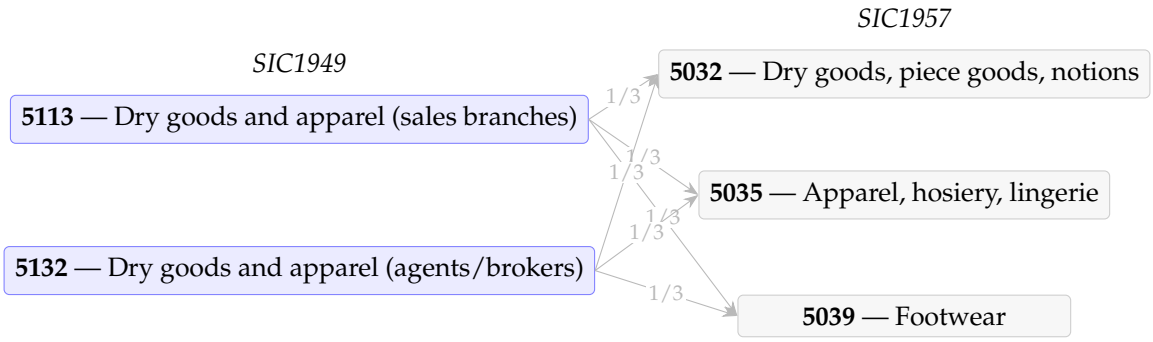


FIGURE OA.6: The dry-goods and apparel case. SIC1949 codes 5113 and 5132 both represented "Dry goods and apparel," distinguished only by wholesale channel (manufacturers' sales branches versus agents and brokers). The 1957 revision dropped the channel distinction at the four-digit level and split the category into three product subgroups (5032, 5035, 5039). Under the count-based weighting rule, each of the six (source, target) pairs receives weight 1/3.

OA.0.3.4 Coverage and quality

Table OA.16 reports summary statistics for the three new crosswalks. The ICC1942 crosswalk covers 351 four-digit source codes, of which 304 (87%) map to a single SIC1957 successor (1:1) and 42 (12%) are SPLIT cases with multiple targets. The SIC1945 crosswalk covers 462 four-digit manufacturing codes, of which 433 (94%) are 1:1 and 29 (6%) are SPLIT. The SIC1949 crosswalk covers 542 four-digit nonmanufacturing codes, of which 522 (96%) are 1:1 and 18 (3%) are SPLIT.

The three crosswalks together cover the full universe of CBP industry codes from the 1946–1956 period (SIC1945 for manufacturing, ICC1942 and SIC1949 for nonmanufacturing), allowing every four-digit industry code in this early period to be harmonized to a NAICS2012 target once composed with the downstream chain. The validator's null result on substantive errors gives us confidence that the three crosswalks are reliable for empirical work.

TABLE OA.16: Summary statistics for the three constructed crosswalks

	ICC1942 → SIC1957	SIC1945 → SIC1957	SIC1949 → SIC1957
Four-digit source codes	351	462	542
of which: 1:1 mappings	304 (87%)	433 (94%)	522 (96%)
multi-target splits	42 (12%)	29 (6%)	18 (3%)
no match	5 (1%)	0	2 (0.4%)

Notes: "Four-digit source codes" counts the number of distinct four-digit source codes in each crosswalk.

The data artifacts are made available as enriched CSV files (`weights_ICC1942_SIC1957_enriched.csv`, `weights_SIC1945_SIC1957_enriched.csv` and `weights_SIC1949_SIC1957_enriched.csv`) that include source and target codes, full industry titles for both, the count-based weight, the LLM's quality label, the number of targets per source, and the LLM's free-text note explaining each assignment. The full pipeline of construction scripts, including the Stage 1 prompts, the Stage 2 sentence-transformer validator, the per-link weight computer, the chain composer, and the lossy-path diagnostics, is provided as a reproducible kit at the project's replication repository.

OA.0.4 The 1963 supplement to SIC1957 handling

The Census Bureau released a 1957 supplement in 1963 to the SIC1957 manual that modified approximately 97 codes and introduced 15 genuinely new codes; the supplement is structurally a minor revision rather than a new vintage. Adding it as a separate chain link would introduce an additional count-based weight layer that is substantively redundant for codes the supplement merely renamed, and would over-count for codes the supplement consolidated.

We instead apply the 1963 supplement to SIC1957 at the CBP-harmonization step as a *code-level substitution* against the SIC1957 mapping. For CBP records from 1964–1967 (the years when the supplement was in force), codes that exist in both SIC1957 and SIC1957 supplement — the vast majority — are looked up directly in the SIC1957 weights table. For the small number of SIC1957 supplement codes that do not exist in SIC1957 (the 15 new codes), the supplement file is parsed to identify each new code's primary SIC1957 ancestor (preferring the ancestor whose numeric code matches the new code's number). The CBP record's code is replaced with this primary ancestor before the SIC1957 lookup. This substitution introduces no additional weight layer and keeps the chain at the same length as for the 1959–1962 CBP files.

OA.0.5 Count-based weighting

We treat each chain link as a stochastic transition matrix and compose the chain by multiplication. This subsection formalizes the weighting rule used to populate each

link, illustrates it with a concrete historical example, and describes how composition produces an end-to-end mapping from any source code to NAICS2012.

The weighting rule. Consider a single link mapping vintage V to vintage V'. Let s denote a source code at the most-detailed level of V, and let $\mathcal{T}(s) \subseteq V'$ be the set of distinct target codes to which s is concorded. Define $m(s) = |\mathcal{T}(s)|$. Under the count-based weighting rule, the weight from s to each target $t \in \mathcal{T}(s)$ is

$$w(s \to t) = \frac{1}{m(s)}. \tag{OA.1}$$

This rule allocates the economic content of each source equally across its targets. Weights are nonnegative and satisfy $\sum_{t \in \mathcal{T}(s)} w(s \to t) = 1$ for every s, so each link defines a stochastic transition from V to V'.

The count-based rule is one of several possible interpretations of how to assign weights when no employment, payroll, or other empirical allocation is available at the link level. Eckert, Fort, Schott, and Yang (2021b) discuss two alternatives in detail: empirical weighting (using contemporaneous employment shares from the Economic Census, when available) and what they term *Interpretation 1* (treating each (s,t) pair as if it represents the entire economic content of s, which over-counts when s splits). Empirical weighting requires employment data at the granularity of the transition and is only feasible for the SIC1987→NAICS1997 link, where the 1997 Economic Census provides direct allocations. For the early-vintage links we construct, no such data exists. The count-based rule of equation (OA.1) is the natural choice when the only information available is the structure of the concordance itself.

An illustration: the postwar household appliance industry. Figure OA.7 illustrates the rule on SIC1945 code 3621 ("Electrical appliances"). In the 1945 classification, the entire household appliance sector was a single four-digit industry. By the 1957 revision, the postwar diversification of household goods had fragmented this category into seven distinct industries reflecting the specific appliances Americans were beginning to purchase in volume: cooking equipment (3631), refrigerators and freezers (3632), laundry equipment (3633), electric housewares and fans (3634), vacuum cleaners (3635), sewing machines (3636), and a residual category for other household appliances (3639).

Under the count-based rule, the SIC1945 code 3621 maps to each of these seven SIC1957 successors with weight $1/7 \approx 0.143$. The total weight leaving 3621 is $7 \times (1/7) = 1$, conserving the economic content of the source category across its now-disaggregated successors.

Why count-based, uniformly. A single weighting rule applied across all chain links has two practical advantages. First, it permits direct matrix multiplication of link

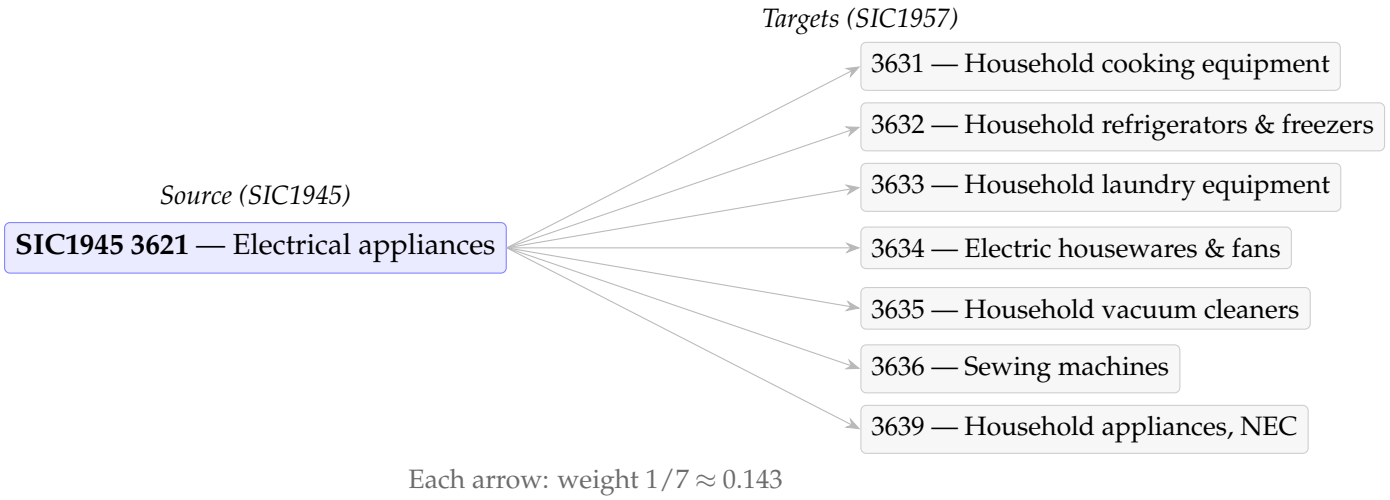


FIGURE OA.7: Count-based weighting applied to SIC1945 code 3621. The 1945 "Electrical appliances" category fragmented into seven distinct household appliance industries in the 1957 SIC revision. Under equation (OA.1), the SIC1945 source maps to each SIC1957 successor with weight 1/7.

transitions to produce end-to-end weights with the standard interpretation of a Markov-chain composition. Second, it ensures that the weight from any particular source code in any vintage to any NAICS2012 target has a single, transparent derivation rather than a hybrid of methods whose interactions would be difficult to justify. The cost is that empirical weights from the 1997 Economic Census are not used; for the SIC1987→NAICS1997 link, EFSY's data appendix provides several pre-computed weight columns, and we use `weight_mapping_all`, which implements the count-based rule of equation (OA.1) on the union of EFSY's two underlying concordance sources (the official Census concordance and the Economic Census coverage mapping). This column is fully populated for all 3,121 source-target pairs at the six-digit NAICS level and reproduces the weighting rule we apply elsewhere in the chain.

Composition. Let W_ℓ denote the transition matrix for link $\ell \in \{1,\dots,L\}$, where $W_\ell[s,t] = w(s \to t)$ under equation (OA.1) and $L = 10$ for the full chain from SIC1945 (or SIC1949 or ICC1942) to NAICS2012. The end-to-end transition is

$$W_{\text{total}} \;=\; W_1 W_2 \cdots W_L. \tag{OA.2}$$

For any source code s_0 in the initial vintage and any target code s_L in NAICS2012, the end-to-end weight is the sum over all chain paths of the product of link-level weights along each path:

$$W_{\text{total}}[s_0, s_L] \;=\; \sum_{\text{paths } s_0 \to s_1 \to \cdots \to s_L} \;\prod_{\ell=1}^{L} w(s_{\ell-1} \to s_\ell). \tag{OA.3}$$

When multiple paths from s_0 converge on the same s_L, their contributions add. Returning to the electrical-appliances example: each of the seven SIC1957 successors of 3621 flows downstream through the SIC1967, SIC1972, SIC1977, and SIC1987 vintages and into the modern NAICS2012 "Household Appliance Manufacturing" family of codes (335220, 335221, 335222, 335224, 335228). Several of these intermediate chains converge on the same NAICS2012 targets, and the end-to-end weight to each target is the sum of the path products.

Many-to-many cases also arise. For example, SIC1949 codes 5113 and 5132, both wholesalers of building materials, each split across the SIC1957 codes 5032, 5035, and 5039, generating six (source, target) pairs at the first link of the chain that the composition then carries forward. We describe this case in detail in Section OA.0.3, where it arises naturally in the construction of the SIC1949→SIC1957 crosswalk.

The composition is exact under one assumption: that every intermediate code in a path has a corresponding entry as a source in the next link. For the two EFSY-workbook links (SIC1957→SIC1967 and SIC1967→SIC1972), a small number of codes still do not appear as sources — the workbooks are more complete than the partial ELMMSS files but not exhaustive. We interpret absent intermediate codes in these two links as implicit identity mappings ($s \to s$ with weight 1.0). Without this convention, codes that remained unchanged across the two transitions would be falsely treated as dead paths and the chain coverage would be substantially understated. Section OA.0.6 reports the residual coverage after this correction.

OA.0.6 Coverage and attrition

The composition described in Section OA.0.5 produces an end-to-end weight from every source code in every CBP vintage to one or more NAICS2012 targets. Most source codes reach NAICS2012 with weights that sum to 1.0 by construction. A small number of source codes, however, exhibit genuine weight loss along the chain. These cases reveal something substantive about the discontinuities in U.S. industry classification over the 1942–2012 period, and we document them here.

We distinguish genuine weight loss from floating-point noise. When a source code splits 1:m into m targets at some intermediate stage and m does not divide cleanly (e.g., $m = 3$), the recombined weights at later stages can fall short of 1.0 by a numerically small amount due to floating-point arithmetic. We treat any source whose total weight to NAICS2012 differs from 1.0 by less than 10^{-3} as numerically clean rather than genuinely lossy.

Table OA.17 reports coverage statistics by source vintage. The genuinely lossy paths fall into four clusters that reflect substantive industry attrition over the seven decades of the chain.

TABLE OA.17: Coverage of source codes by vintage

Source vintage	Source codes	Lossy	Lossy %
ICC1942	346	2	0.6%
SIC1945	462	0	0.0%
SIC1949	540	2	0.4%
SIC1957	922	1	0.1%
SIC1967	917	1	0.1%
SIC1972	1,003	1	0.1%
SIC1977	1,004	0	0.0%
SIC1987	1,005	1	0.1%
NAICS1997	1,169	0	0.0%
NAICS2002	1,179	0	0.0%
NAICS2007	1,175	0	0.0%
NAICS2012	1,065	0	0.0%
Total	10,787	8	0.07%

Notes: "Source codes" counts distinct four-digit (SIC/ICC) or six-digit (NAICS) industry codes at the most-detailed level. "Lossy" counts source codes whose end-to-end weights to NAICS2012 sum to less than $1.0 - 10^{-3}$. The threshold excludes floating-point rounding artifacts that arise mechanically from $1 : m$ splits where m does not divide 1.0 cleanly.

Railway express service. Code 4041 ("Railway express service") persists as a four-digit code from ICC1942 through SIC1972 and was renumbered to 4010 in SIC1977; the SIC1977→SIC1987 transition then has no successor for 4010. This single economic story accounts for one lossy source in each of ICC1942, SIC1949, SIC1957, SIC1967, and SIC1972 — five lossy paths in total — and reflects the collapse of railway-based parcel delivery as trucking and air freight expanded.

Retail-bakery reclassification. SIC1949 code 5462 ("Retail bakeries – manufacturing") maps to SIC1957 code 5461 and propagates forward as 5461 through SIC1972; the SIC1972→SIC1977 transition then has no four-digit successor. The retail side of the bakeries classification was absorbed into the manufacturing side at that step, leaving no distinct four-digit code in SIC1977.

Early-vintage government establishments. ICC1942 code 9411 ("Regular government establishments") propagates forward as SIC1972 code 9190 and has no SIC1977 successor at the four-digit level. The path dies at the SIC1972→SIC1977 step.

Residual codes. SIC1987 code 9999 ("Nonclassifiable Establishments") has no NAICS1997 successor by design. NAICS handles unclassified establishments through a different mechanism, and the SIC1987 residual does not appear in EFSY's SIC1987→NAICS1997 concordance.

For empirical applications using the harmonized CBP, the implications are bounded.

TABLE OA.18: CBP industry classification systems, 1946–2016.

CBP Year	Industry Classification System	Most-Detailed Resolution
1946	ICC1942 / SIC1945	2-digit
1947–1948	ICC1942 / SIC1945	3-digit
1949–1950	SIC1945	3-digit
1951–1953	SIC1945 / SIC1949	3-digit
1956	SIC1945 / SIC1949	4-digit
1959–1962	SIC1957	4-digit
1964–1967	SIC1957 & SIC1963 Supplement	4-digit
1968–1973	SIC1967	4-digit
1974–1987	SIC1972[a]	4-digit
1988–1997	SIC1987	4-digit
1998–2002	NAICS1997	6-digit
2003–2007	NAICS2002	6-digit
2008–2011	NAICS2007	6-digit
2012–2016	NAICS2012	6-digit

Source: Vintage assignments follow Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022) for 1946–1974 and Eckert, Fort, Schott, and Yang (2021a) for 1975–2016.
Notes: "Most-detailed resolution" is the digit level at which we apply the count-based weighting — the finest source-code digit level published in the CBP for that year.
[a] CBP data from 1977 to 1987 follow SIC 1977, which is not a major SIC revision and only contains a few code updates from SIC 1972. Therefore, we do not include SIC 1977 in the chain.

The total mass of lost weight across the 10,787 source-vintage codes amounts to less than 0.1% of total chain coverage and is concentrated in a handful of codes with coherent economic interpretations. Researchers interested specifically in any of the affected industries should be aware that those lines of business cannot be cleanly traced into modern NAICS codes. For aggregate analyses of employment by sector, geographic concentration, or broad industry composition, the loss is substantively negligible.

OA.0.7 Harmonization and Collapse of the CBP Panel

The chain of pairwise concordances developed in Sections OA.0.2–OA.0.6 delivers, for every most-detailed-level source code in every vintage, a count-based weight $w(s,t) \in (0,1]$ that maps the source code s to one or more NAICS2012 targets t. Applying these weights to the annual CBP county files produces a harmonized national panel covering the nearly 70 published CBP years from 1947[62] to 2016.

Harmonization at the most-detailed level. For each county–industry observation in the raw CBP [63] we restrict attention to the most-detailed source code available in that

[62] The 1946 CBP file publishes source codes only at most at the 2-digit SIC level, which roughly corresponds to 3-digit NAICS — coarser than the 4-digit NAICS harmonization target as described later — so we omit 1946 from the harmonized panel.

[63] Two systematic corrections are applied to the raw CBP files before harmonization. *(i) Scale adjustment, 1975–1985.* The CBP data for 1975–1985 require removing the last three digits of numeric fields to bring

year's classification — as summarized in Table OA.18 — and match it against the chain to retrieve its NAICS2012 target(s) and associated weight $w(s,t)$. Every numeric measure — employment, establishment counts, first-quarter and annual payroll, and the establishment-size bins — is then multiplied by $w(s,t)$ before being aggregated to the NAICS2012 target. Three considerations motivate operating at the most-detailed level. First, the count-based weighting rule is a brute-force mechanical device for splitting the economic content of a single source industry across multiple NAICS2012 targets; applying it at the finest available level of source disaggregation minimizes the distortion this mechanical split introduces, since the unit being divided is itself as narrow as the underlying classification allows. Second, the early-vintage NAICS-side concordances of Eckert, Fort, Schott, and Yang (2021a) — in particular SIC1957→SIC1967 and SIC1967→SIC1972 — are constructed only at the 4-digit level, so harmonizing at any coarser or finer source-digit level would require substantial ad-hoc reconstruction outside the existing chain. Third, the CBP reports each county–industry cell at every level of the SIC/NAICS hierarchy; restricting attention to a single digit level is therefore necessary to avoid double-counting at the collapse step. Special CBP codes that the chain does not cover, principally the administrative codes for auxiliary establishments and a small set of year-specific industry breakouts in the early period, are mapped through a separately curated 1-to-1 lookup with $w(s,t) = 1$. Employment is suppressed for some county–industry cells in the raw data to protect confidentiality. We fill these suppressed cells with the imputations of Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022) for 1947–1974 and Eckert, Fort, Schott, and Yang (2021a) for 1975–2016; for the latter we take the midpoint of the imputed lower and upper employment bounds as the cell value.

Collapse to 4-digit NAICS2012. The weighted observations are then summed across all U.S. counties and across all most-detailed source codes that map into a given (year, 4-digit NAICS2012) cell. For observations whose chain-mapped target is at the 6-digit level (the 1954–2016 portion of the panel), the cell is identified by the first four digits of the 6-digit code.

References

Eckert, Fort, Schott, and Yang (2021a): Fabian Eckert, Teresa C. Fort, Peter K. Schott, and Natalie J. Yang, "Imputing Missing Values in the US Census Bureau's County Business

values in line with neighboring vintages. *(ii) Annual-payroll inflation, 1979–1980.* The county record layout provided by National Archives and Records Administration (NARA) reports that for a small set of (year, county) records the annual-payroll field `TANPAY` was published without the standard 1,000 adjustment, leaving those values inflated by a further factor of 1,000. The affected records are Los Angeles County, CA in 1979, and Los Angeles County, CA, New York County, NY, and Cook County, IL in 1980 (see User Note 1, Margaret O. Adams, December 12, 2003); for these we apply an additional division by 1,000 to bring the values into line with the surrounding panel.

Patterns," *NBER Working Paper* (2021).

Eckert, Lam, Mian, Muller, Schwalb, and Sufi (2022): Fabian Eckert, Ka-leung Lam, Atif R. Mian, Karsten Müller, Rafael Schwalb, and Amir Sufi, "The Early County Business Pattern Files: 1946–1974," *NBER Working Paper* 30578 (2022).