

DATA APPENDIX

Return Innovation: The Knowledge Spillovers of the British Migration to the United States, 1870-1940

Davide M. COLUCCIA & Gaia DOSSI
September, 2026

[Latest Version](#) 

C Data Appendix

This section describes the data sources and the construction of the final sample (Appendix C.I), the linking of British immigrants in the US to the UK census (Appendix C.II), and the digitization of the UK patent data and their linking to the census (Appendix C.III).

C.I Summary of Data Sources

C.I.1 Patent Data

US patent data are from Berkes (2018), who digitizes the universe of patents granted between 1836, when the US patent and trademark office was established, and 2010. In this paper, we are interested in the CPC technology class, the issue year, and the coordinates of residence of each inventor. We then assign each patent to US counties at 1900 borders. Depending on the number of inventors, a single patent may be assigned to multiple counties. In the case of patents with more than one inventor, we weigh each by the inverse of the number of inventors to avoid multiple counting. English and Welsh patents after 1900 are available from the European Patent Office. To construct our dataset, we leverage bulk access to the PATSTAT dataset. Information contained in PATSTAT includes the CPC class and the issue year. To retrieve the location of each inventor, we merge the PATSTAT data with the PatCity repository, which contains geo-coded information on the universe of English and Welsh patents during this period (Bergeaud and Verluise, 2024). Data before 1900 are not available. In Appendix C.III, we describe how we digitize the universe of patent documents issued over the period 1853–1899 to fill this substantial gap.

Importantly, we map 3-digit CPC classes to a coarser taxonomy of classes. To do that, we reduce them to functional units using the CPC classification scheme. The scheme is publicly available at the following [link](#). To accommodate the historical context, we divide the transportation category into two classes: “Transporting”, which includes carriages, railways, and cars, and “Ships and Aeronautics”. Moreover, we conflate the “Weapons and Blasting” and the “Mining” classes into the “Metallurgy” category because few patents were observed in those industries. We further augment patent data by defining a measure of “innovativeness” or “importance,” following Ganguli et al. (2024) and Kelly et al. (2021). We describe our approach in Appendix C.III.6.

C.I.2 Migration Data

Disaggregated data on the origin of English and Welsh immigrants—and, more generally, of all other nationalities—do not exist as neither US authorities nor those in the sending countries collected them. As a result, we lack precise information on where British immigrants in the US came from *within* the UK. We fill this gap and link the individual-level UK and US censuses, as described in Appendix C.II. In the US Census, we observe virtually the universe of British emigrants to the US between 1870 and 1930. For these individuals, we observe all information contained in the US census and, after merging it with their records in the UK census prior to emigration, those recorded there as well, including their place of origin. As we discuss in more detail later, we also link return migrants. Since the last publicly available UK census dates to 1911, we can only construct return migration flows for the period 1870–1910.

C.I.3 Data Constructed from the Population Censuses

The main data sources we use are individual-level, non-anonymized UK and US population censuses. The US census features prominently in the economic history literature as a major source of detailed microdata, and we therefore do not discuss it further here (Ruggles et al., 2021). The UK census is relatively less well-known (Schürer and Higgs, 2020). Although not as rich as its US counterpart, the UK population census covers individuals residing in the UK since 1850. The first census was conducted in 1841, but only 1851, 1861, 1881, 1891, 1901, and 1911 are fully digitized.⁴ The census records include name and surname, birth year, division, county, district, parish, precise address of residence, detailed occupation (coded using HISCO), and additional variables not used in this paper, such as dwelling characteristics and fertility information. We augment these data by geo-coding the universe of addresses that appear in the census to precise geographical coordinates.

C.I.4 GIS Shapefiles & Boundary Harmonization

Patents are mapped to 1900 registration district borders using historical GIS files and their geographic coordinates.⁵ All data from the population censuses appear at historical borders. Registration districts have not undergone major boundary changes over the period that we studied. Nevertheless, to ensure consistent geographic units over time, we adapt the method proposed by Eckert et al. (2020) to UK registration districts.

Constructing geographical crosswalks using this method requires assuming that variables are evenly distributed over the area of each geographical unit. The crosswalk is obtained by overlapping geographical units across time. Suppose that unit x in decade d is split, with 80% of its territory remaining in x and 20% assigned to another district y . To construct a crosswalk to period $d + t_2$ for a generic variable between decades $d - t_1$ and $d + t_2$, for $t_1, t_2 > 0$, one needs to multiply the variable measured in district x in $d - t_1$ by $4/5$ and add $1/5$ of the variable in x to that measured in y in the same decade. We map registration districts to their boundaries in 1901. Less than 5% of the overall area of England and Wales is reassigned in this way. We adopt the same methodology to map counties to their 1900 borders.

C.I.5 Geo-referencing the Historical British Census

A notable feature of the UK census is that it contains precise information on the residential address of the universe of the British population. This information is extremely valuable because, in principle, it assigns the finest possible location to each individual. In practice, however, it is highly non-standardized and challenging to use. In this paper, we expand earlier work by Lan and Longley (2019), who analyze the 1901 census and geo-reference the 1851-1911 censuses. The geo-coded census sample is used in the neighborhood individual-level analysis, while all other results do not rely on these data.

There are two ways to geo-reference historical addresses. The first is to manually digitize historical locations, either streets or enumeration units, from historical maps. However, this method does not

⁴The 1921 census is currently being digitized and is partially available. We do not use it because its coverage is still not complete and because it is not available in bulk. All censuses after 1921 are subject to privacy restrictions.

⁵GIS data for the US are provided by NHGIS, while district boundaries have been digitized by the Great Britain Historical GIS Project.

scale up and becomes rapidly unfeasible as the data grows. The second, automated approach uses text-based address matching between historical sources and modern or historical address databases that have already been georeferenced. We follow this latter method, as we need to geo-reference 5,464,578 unique addresses.

To perform the actual matching, we first carry out a preliminary manual cleaning of addresses. First, we remove house numbers because they undergo many changes and re-sequencing over time. Second, we remove uninformative locations, such as “village”, “farm”, and “rectory”. Then, we input the resulting addresses as queries into the geo-referencing engine (MapTiler API). We discard the match if the resulting coordinates are not within the parish’s boundaries where the address is recorded. This consistency check is necessary because homonyms are frequent. Since observing two addresses with the same name within a given parish is extremely rare, this procedure minimizes spurious matches.

We use the geo-coded 1891 census in the neighborhood analysis. We geo-code 85% of the addresses listed in the census, covering 86% of the population. The geo-coding ratio is relatively homogeneous across districts, with the exception of Wales, where many addresses are reported in Welsh. In one robustness exercise, we exclude Wales from the sample and confirm that the results hold.

C.II Linked Migrants Sample

C.II.1 Data Sources

We rely on two externally compiled data sources.⁶ For the US, we use the IPUMS full-count, non-anonymized census (Ruggles et al., 2021). A census was taken in the US every ten years starting in 1790.⁷ Until 1840, this was run at the household level. From 1850 onward, we have detailed *individual*-level information for the universe of the US population.⁸ For confidentiality, these data are available until 1940. Our dataset contains snapshots of the entire US population at any given decade between 1850 and 1940, although we focus on the years 1870-1930 for the purpose of this paper. Because we use the non-anonymized version of the IPUMS data, we observe each individual’s recorded name and surname in addition to the publicly available information.

In the UK, the I-CeM data mirrors the IPUMS content (Schürer and Higgs, 2020). More precisely, it contains information on the universe of individuals living in England, Scotland, and Wales. Similar to the US, the census was conducted at decadal frequency from 1851 to 1911. The records of the 1871 census have not survived. As with the IPUMS data, we use the full-count, non-anonymized version of the dataset. In addition to publicly available information, this includes full names and addresses for the universe of individuals living in the UK at each census year.

C.II.2 Linking Algorithm

Our methodology follows Abramitzky et al. (2021). As neither the US nor the UK—nor any other European countries—recorded where British immigrants came from *within* the UK, we match British

⁶We are deeply thankful to IPUMS and I-CeM for allowing us access to their confidential data. Without their help, this paper would not have been possible.

⁷The individual records of the 1890 census have been lost in a fire.

⁸By US population, we refer to the universe of individuals who *lived* in the US at a given point in time.

immigrants residing in the US with their entry in the UK census, which records where they come from at a granular geographical level.⁹ More precisely, we take the stock of British immigrants residing in the US in a given census year—say, 1900—and match them with their entry in the preceding UK census—in this case, 1891.¹⁰ This implies that we measure the *flow* of British immigrants over time rather than their stock.

We use three variables to link individuals: first name, surname, and birth year. The baseline linked sample consists of individuals who report England or Wales—or analogous denominations, such as Great Britain—as their country of origin in the US census. In the 1900 census, we take all those who immigrated between 1870 and 1899. In the subsequent censuses, until 1930, we retrieve the stock of those who immigrated in the preceding decade. Then, to match each unit in the sample—call the generic one A —to an entry in the UK census, we perform these steps:

1. Take the census that precedes the immigration year of A . Hence, for instance, we match all those who immigrated in 1896 to the 1891 census.
2. Select all records in that census born within a 2-year window of the birth year of A and with the same initial name and surname letter as A . Call the resulting sample $\mathcal{M}^A = \{m_1^A, \dots, m_N^A\}$.¹¹
3. Compute a string-similarity measure between the name and surname of A and that of all elements of $\mathcal{M}^A$. In other words, for every $m_i^A \in \mathcal{M}^A$, compute $d_N(A, m_i^A)$ as the Jaro-Winkler similarity between i 's and m_i^A 's name, and $d_S(A, m_i^A)$ as the surname Jaro-Winkler similarity.¹²
4. The set of candidate matches is defined iteratively. First, we select the matches within $\mathcal{M}^A$ with the highest surname similarity with A :

$$\overline{\mathcal{M}}^A = \left\{ m_i^A \in \mathcal{M}^A \mid d_S^A(A, m_i^A) = \max_{m_{i'}^A \in \mathcal{M}^A} d_S^A(A, m_{i'}^A) \right\}. \quad (\text{C.1})$$

Then, we refine the set of candidate matches by selecting the elements with the highest name similarity with A :

$$\overline{\mathcal{M}}^A = \left\{ m_i^A \in \overline{\mathcal{M}}^A \mid d_N^A(A, m_i^A) = \max_{m_{i'}^A \in \overline{\mathcal{M}}^A} d_N^A(A, m_{i'}^A) \right\}, \quad (\text{C.2})$$

implying that we restrict the set of possible matches to include only those with the highest name similarity among those with the highest surname similarity with record A .

5. Finally, for a given threshold couple $\tau = (\tau_1, \tau_2) \in [0, 1]^2$, we select only the candidate matches

⁹Since women usually changed their name upon marriage, we cannot match them. This is a common problem in linking algorithms (Abramitzky et al., 2021).

¹⁰Since the individual records of the 1871 census have not survived, we link the 1880 US census to the 1861 UK one. This is not overly problematic because we can still match all those aged ten or older in 1871.

¹¹Restricting $\mathcal{M}^A$ to only include records with the same initial name and surname letter as A is recommended by Abramitzky et al. (2021). It is computationally advantageous and unlikely to significantly affect the performance of the matching algorithm because errors in names and surnames are more common at the end than at the beginning.

¹²We cannot match on the exact same name and surname because coding errors are common in historical census data. Restricting to exact matches would thus reject possible true positive matches.

whose similarity score is above (τ_1, τ_2) . The set of accepted matches thus boils down to:

$$\widetilde{\mathcal{M}}_{\tau}^A = \left\{ m_i^A \in \overline{\mathcal{M}}^A \mid d_N(A, m_i^A) \geq \tau_1 \wedge d_S(A, m_i^A) \geq \tau_2 \right\}. \quad (\text{C.3})$$

Clearly, $\widetilde{\mathcal{M}}_{\tau}^A$ can be empty, meaning that A has no accepted matches. It can have one element, in which case we refer to it as a “single match,” or it can have multiple elements. In our baseline exercise, we set $\tau_1 = \tau_2 = 0.95$ to ensure that the sample only includes high-quality matches.

We evaluate the distance between two strings A and B in terms of their Jaro-Winkler similarity $d(A, B)$:

$$d(A, B) \equiv \widehat{d}(A, B) + \ell p \left[1 - \widehat{d}(A, B) \right], \quad (\text{C.4})$$

where

$$\widehat{d}(A, B) \equiv \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|A|} + \frac{m}{|B|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases}, \quad (\text{C.5})$$

where m is the number of matching characters, $|A|$ and $|B|$ are the lengths of the two strings, t is half the number of matching characters appearing in a different sequence order, ℓ is the length of the common prefix capped at four characters, and $p = 0.1$ is a constant scaling factor. Two characters are matching only if they are identical and are not farther than $\left\lfloor \frac{\max(|A|, |B|)}{2} \right\rfloor - 1$.¹³ The Jaro-Winkler distance has been shown to perform well in linking routines (Abramitzky et al., 2021). In our particular case, this metric outperforms simpler string dissimilarity metrics because coding errors are far more frequent at the end of names and surnames than at the beginning.

C.II.3 Matching Rate

Figure C.1 reports the key matching rate statistics of the linked sample. In particular, Figure C.1a displays the crude matching rate over time, i.e., the share of English immigrants recorded in the US census linked to at least one record in the UK census. Since it is impossible to link British immigrants born after the last census before they migrated (i.e., are born and migrate within two census years), we distinguish between the full sample in blue and the “matchable” sample in red. The matching rate in the matchable sample is approximately 53%. It declines from 65% in the 1870s to 50-55% in the 1890s and remains constant after that. The large jump in 1880 in the crude matching rate is because the 1871 census is unavailable; hence, between 1870 and 1881, we cannot match anyone younger than 20—as opposed to 10 in the rest of the sample—when they migrated. The matching rate is relatively high compared to other algorithms, such as Abramitzky et al. (2021). In the next section, we thus ask whether matched and unmatched immigrants are balanced on observables. Figure C.1b replicates the previous plot but reports the absolute number of emigrants in the sample and those linked instead of the percentage shares.

Figure C.1c displays the distribution of linked immigrants by number of matches. A large fraction ($\approx 35\%$) are paired with a single entry in the UK census. Approximately 50% are matched to one or two entries. Beyond this threshold, multiple matches are more common, and the average number of matches is 16. To ensure that multiple matches do not confound our analysis, we always weigh

¹³The Jaro-Winkler distance has been recently employed in the economic history literature for intergenerational linking purposes by, among others, Abramitzky et al. (2021).

emigrants by the inverse of the number of matches. Additionally, we check that the baseline results remain qualitatively unchanged when we include only single matches in the dataset.

In Figure C.2, we report the distribution of the name and surname similarities between the records of British immigrants in the US census and their links in the UK census. The dashed black lines plot the threshold below which we reject the link (.95). In the sample, we find that if an immigrant is linked to at least one record in the UK census, then the quality of the match is high, as more than 80% of the sample lies right to the .95 line. However, it is important to note that for a link to be accepted in the final sample, we require that *both* the name and the surname similarities be above 0.95.

C.II.4 Internal Consistency: Natural Language Processing Applied to Census Linking

We validate the internal consistency of the linked migrants sample using manual and semi-automated methods. The spirit of these exercises is to verify that the matches produced by the linking algorithm would be delivered by an expert human agent operating with the same information available to the algorithm. We adopt three methods: a manual validation on a random subsample of the data, a validation based on a classification constructed by a state-of-the-art large language model (GPT-5) on the same random subsample, and a validation of the entire dataset using a bidirectional text representation model (BERT) trained on the manually validated random subsample. We now describe these three approaches in detail.

Manual Validation The scale of the dataset precludes a manual verification of all the matches. We therefore take two random samples of 1,000 observations each. The first only comprises matches that we accept according to the Jaro-Winkler classification rule explained above, i.e., all the elements in the set $\widetilde{\mathcal{M}}_{0.95}^A$. The second only comprises matches that we reject according to the same classification rule, i.e., all the elements in the set $\overline{\mathcal{M}}^A \setminus \widetilde{\mathcal{M}}_{0.95}^A$. Our first validation approach is to manually verify that the matches in $\widetilde{\mathcal{M}}_{0.95}^A$ are indeed plausible, and that the matches in $\overline{\mathcal{M}}^A \setminus \widetilde{\mathcal{M}}_{0.95}^A$ are indeed to be rejected. Naturally, we expect some classification error, so we calculate the share of true positives and the share of false negatives as statistics to evaluate the overall performance of the algorithm.

Table C.1a (column 1) reports the results. Within the sample of 1,000 accepted matches (Panel A), we confirm 977. This figure corresponds to a 98% share of true positive matches. Panel B reports the corresponding exercise on the subsample of rejected matches. We find that out of 1,000 matches rejected by our algorithm, a human agent would have rejected 825, and 175 would have been accepted. These figures correspond to a 17% false negative rate. The spirit of setting a 0.95 threshold for the Jaro-Winkler classifier, as in Green (2025), is to minimize the share of false positives, thereby retaining only high-confidence matches. This choice is reflected in the 98% true positive rate. However, we find that this has only a moderate impact on the recall rate. Out of $977+175 = 1,152$ true matches in the two datasets, our classifier correctly identifies 85%. We thus estimate a $175/1,152 = 15\%$ share of missed matches.

Semi-Automated (GPT) Validation A possible concern with this strategy is that the manual verification is conducted by a human researcher and is thus subject to cognitive biases. To address this issue, we adopt a semi-automated approach and employ GPT-5 to establish the matches in place of the human researcher using the prompt below.

GPT-5 Prompt for Record Linkage Verification

You are a native English speaker with expertise in record linkage. Consider the following two individual names:

- name_in_us_census
- name_in_uk_census

Determine if they refer to the same person, taking into account:

- Phonetic similarities despite minor spelling differences,
- Common abbreviations, nicknames, and misspellings,
- Possible OCR or transcription errors,
- If only one record has a middle name or initial middle name, disregard it. If both have it, consider it.

Scoring rules:

- 1 = Certainly or almost certainly the same person: exact match or very minor variation,
- 0 = Otherwise.

Additional guidelines:

- First priority is to minimize false positives,
- Output only valid JSON, with no explanation or extra text.

Respond strictly in this format: `{ "score": < 0 | 1 > }`

We prompt the language model to ensure that it has the same objective function as the researcher, i.e., to minimize the probability of delivering false positive matches.

Column (2) of Table C.1a reports the results. The GPT-powered classifier confirms 978 matches out of the 1,000 accepted by the Jaro-Winkler algorithm, which corresponds to a 98% accuracy rate. The LLM classifier thus exactly replicates the accuracy derived through manual classification. In Panel B, we find that GPT would have accepted 356 out of the 1,000 matches discarded by the Jaro-Winkler classifier, yielding a false-negative rate of 36%. This figure is larger than the 17% we obtained through manual verification, indicating that the LLM is plausibly more lenient than the researcher. However, the Jaro-Winkler classifier still displays a $978 / (978 + 356) = 1,334 = 73\%$ recall rate when considering the GPT-classifier as the ground truth. This is a highly satisfactory performance, given that maximizing the accuracy rate is the main objective.

In principle, it is possible that while the manual and the GPT classifiers give similar accuracy rates, they are actually confirming and rejecting different matches. To check that this is not the case, in Table C.1b, we display the confusion matrix of the manual and the GPT classifiers on the subsample of

accepted (columns 1–3) and rejected (columns 4–6) matches. A confusion matrix tabulates the joint distribution of predicted and actual classifications, displaying the number of instances correctly and incorrectly assigned to each category. It separates correct identifications (true positives and true negatives) from misclassifications (false positives and false negatives). The confusion matrix indicates that the human and the GPT classifiers are essentially concordant on the accepted-matches sample (columns 1–3). On the rejected-matches sample (columns 4–6), they reject and accept the same observations, except that the GPT classifier accepts 189 matches that the manual validation rejects. Overall, however, the two classification approaches yield substantially aligned predictions.

Full Sample Automated Validation These validation approaches are performed on a 2,000 (stratified) random sample of the linked dataset. To validate the entire dataset, we employ a bidirectional text representation model (BERT).¹⁴ The BERT model yields vector representations of the names in the US and the UK censuses. Then, we compute the cosine similarity between two vectors to measure how close they are, and consider them a match if their similarity is above a certain threshold τ^* . This threshold is thus the only hyperparameter of the BERT-based classifier. To obtain τ^* , we take an 80% random sample of the 2,000 matches that we annotated manually for the previous validation exercise. We select τ^* to maximize the F1 score of the BERT classifier on this training set.¹⁵

To evaluate the performance of the BERT classifier, we compare it to the manual and the GPT classifiers on the test set, i.e., the 20% random sample that we annotated manually and using GPT that we did *not* use to compute τ^* . Columns (3–5) of Table C.1a report the results. Panel A of columns (3) and (4) reports the predictions of the manual and the GPT classifiers, whereas column (5) displays the BERT classifier, on the sample of accepted matches according to the Jaro-Winkler classifier. All models deliver a true positive rate in excess of 95%, with BERT yielding a slightly higher number of false positives (6 out of 182). Panel B replicates the exercise on the sample of rejected matches. In this case, BERT yields the lowest rate of false negative matches, which is less than 50% of that of GPT. Panel B of Table C.1b reports the confusion matrix comparing the manual classification outcomes with those obtained from BERT. The two methods deliver virtually the same classification outcomes on both samples of accepted (columns 1–3) and rejected (columns 4–6) matches.

Since the performance of the BERT-based classifier is nearly identical to that of the manual classifier, we can scale it up to the full dataset to evaluate the predictions of the Jaro-Winkler linking algorithm on the entire sample of linked migrants. Table C.2 reports the results. Column (1) shows the number of matches that are accepted by the Jaro-Winkler classifier that the BERT classifier rejects (first row) and accepts (second row). Out of 359,757 accepted matches in the dataset, the BERT classifier accepted 339,135. The implied share of true positive links (taking the outcome of the BERT classification as the ground truth) generated by the Jaro-Winkler classifier is 94%. Column (2) reports the matches that are rejected by the Jaro-Winkler classifier. Out of 202,255 rejected matches, the BERT classifier rejects 182,162 and accepts 20,093, corresponding to a true negative rate of 90% and a false positive rate of

¹⁴Technically, we employ `all-MiniLM-L6-v2`, a distilled sentence-BERT from the `SentenceTransformers` library with six transformer layers and a 384-dimension embeddings. This model is pre-trained on one billion sentence pairs and fine-tuned for semantic similarity tasks.

¹⁵The F1 score of a classifier is the harmonic average of its precision and recall. We BERT classifier attains a 0.91 F1 score on the training set.

10%. Overall, the Jaro-Winkler classifier thus correctly identifies 94% of $339,135 + 20,093 = 359,228$ true matches.

C.II.5 Balance of Linked Emigrants

Table C.3 reports the correlation between the probability that a British immigrant in the US is linked to an entry in the UK census and a set of individual-level characteristics observed in the US census. Column (1) reports the sample value of each variable for unmatched immigrants, and column (2) refers to matched immigrants. Column (3) reports the difference between the two groups. The linking probability does not correlate with either literacy, for which one may expect a correlation since educated immigrants could present systematically more precise census records, or income, another proxy for educational standing. Matched immigrants are less likely to work as professionals, in clerical occupations, or as sales workers. These differences, however, are small in magnitude and affect a small portion of the population. On the contrary, manufacturing workers (skilled “craftsmen” and unskilled “operatives”) account for almost 60% of the population and do not appear to be selected on their linking probability. In terms of their region of residence, emigrants living in the Northeast are more likely to be linked, although the difference is small (less than 2%). They are also less likely to reside in the South, but again, this difference affects only approximately 1% of the sample. Overall, Table C.3 provides reassuring evidence that linked emigrants seem not to be selected on observable characteristics, as any statistically significant differences are small in magnitude.

C.II.6 External Validation Against Social Security Data

It is challenging to validate our sample against external data because, to the best of our knowledge, primary data on the origins of British emigrants within the UK do not exist.

As a first strategy, we employ Social Security records used, among others, by Althoff et al. (2025). The dataset contains Social Security applications submitted by 266,410 UK-born individuals between 1936 and 2013. We restrict the sample to 27,832 applications submitted in 1936 and 1937. Since naturalization required at least five years of residence in the U.S., this sample comprises individuals who had immigrated no later than 1932 and would therefore plausibly appear in our baseline sample.

Among other information, Social Security data includes the name and surname, year of birth, and city of birth for each applicant. We link this dataset to the 1930 census following the algorithm described in Appendix C.II.2, we georeference the reported city of birth, and use the coordinates to assign it to a registration district.

We use the second linked dataset to assess the plausibility of the links in the baseline sample. Because the Social Security applications data contains only individuals who resided in the United States, linking them to the census records is less prone to false-positive matches than the cross-census linked sample. We thus view this exercise as yielding a more conservative dataset.

The Social Security-census linked dataset covers 14,098 individuals, equal to a 50.65% matching rate, which is comparable to the matching rate we obtain in the baseline linked sample. Each Social Security entry is linked, on average, to 3.08 census entries, and 78.23% of the entries are linked to exactly one census entry. As in the baseline algorithm, we select acceptable matches only if the name and surname Jaro-Winkler similarities exceed 0.95, yielding average name and surname similarities

of 0.9835 and 0.9950, respectively.

At the same time, Social Security data present three disadvantages compared to the cross-census linked sample. First, because the Social Security Administration was established in 1935, applications did not begin until 1936. Hence, the data only covers (i) individuals who were alive in 1836, and (ii) who applied to obtain a Social Security number. Second, the data do not contain information on the immigration year, implying that we can only match the records to the 1930 census and obtain their 1930 location, which is plausibly not relevant to measuring prior exposure to local technology. Third, while in the baseline dataset, we observe the last place of residence, Social Security records contain the city of birth.

Of the 1,278 individuals who appear in both linked datasets, 583 are matched to the same UK registration district, corresponding to a 46% overlap. Using census tabulations, Ravenstein (1885) estimated that, in the late 19th century, approximately 55% of the population lived in a county other than their county of birth. A suggestive calculation indicates that the two linked datasets yield the same origin place for $583 / (0.45 \times 1,437) = 90\%$ of the cases. Given that counties are coarser geographic units than registration districts (there were 52 counties compared to 623 districts), it is plausible that cross-district migration rates were even higher and thus that the share of matches identifying the same place of origin constitutes a lower bound on the true concordance between the two datasets.

C.II.7 External Validation Against Estimated Out-Migration Data

Baines (2002) attempts to estimate decade-level data for British counties between 1880 and 1910. To the best of our knowledge, these estimates are the only alternative source we can use to gauge the plausibility of our data. They nonetheless present important limitations. First, they are based on interpolations from population data tabulated from the population census. Second, the author tries to account for internal migration, which remains an important confounding factor. Third, the estimates refer to aggregate emigration outflows, whereas our data focuses only on emigration to the US. Fourth, the estimates are computed at the county level, which are coarse geographical units in the UK.

Keeping these limitations in mind, in Figure C.3, we aggregate our data at the county-decade level and plot it against the estimates produced by Baines (2002). Both series are shown in logs to reduce the influence of extreme values, and the data is weighted by county population. We find a positive and statistically significant correlation between our data and the Baines series. Our numbers are generally lower than those provided by Baines, which plausibly reflects that our data do not cover emigration to countries other than the US. Overall, we view this graph as providing supporting evidence of the plausibility of our dataset.

C.II.8 Return Migration Data

Following the same linking algorithm described before, we construct a linked sample of return migrants. This identifies English and Welsh immigrants in the US in year t and looks for possible matches in the UK census in year $t + 10$ using the algorithm previously described. Since the last available UK census is in 1911, we face a hard upper bound for the coverage of return migration, as we can only construct return migrants linked samples spanning the period 1870–1910.

C.III Patent Data

C.III.1 Motivation

Despite its historical significance, we lack comprehensive patent data for the Second Industrial Revolution period (1850–1900) in the United Kingdom. In particular, it is impossible to reconstruct the geographical distribution of innovation activity during this period. This data limitation sharply contrasts the effort undertaken to document. While data on UK patents between the inception of the English patent law in 1617 up until the end of the First Industrial Revolution in the 1840s (Nuvolari and Tartari, 2011; Nuvolari et al., 2021) is available in digitized form, we lack comprehensive patent data for the Second Industrial Revolution period (1850–1900). In particular, it is impossible to reconstruct the geographical distribution of innovation activity during this period. To fill this gap, we digitize the universe of English and Welsh patents for the period 1853–1900. These data contain detailed information on the text, geographical location, inventors’ personal information, and grant date.

C.III.2 Data Sources

We access restricted full-page scans of original patent documents granted in England and Wales between 1617 and 1899 from the UK Intellectual Property Office. This paper focuses on the period 1853–1899 for two reasons. First, Nuvolari and Tartari (2011) already digitized patents before 1853 from Bennet Woodcroft’s index, although patent documents contain additional information compared to the index. Second, in 1853, a reform dramatically lowered patent application prices. This makes it challenging to compare patents before and after the reform. Patent documents contain a wealth of unstructured information. We provide two examples in Figure C.5: Figure C.5a shows the patent granted to Henry Bessemer for the eponymous process to produce steel, and Figure C.5b displays the patent granted to John Starley for the first modern safety bicycle. Both patents are in our dataset. The rectangles identify the location of the textual data that we extract. These include (i) a short title, (ii) a long title, (iii) the author(s)’s name(s), (iv) the author(s)’s address(es), (v) the author(s)’s profession(s), (vi) the filing date, (vii) the issue date, (viii) the type of protection, (ix) an indicator of whether the application was filed by an agent on behalf of someone living abroad, and (x) the full text of the patent. Not all (i–x) are available throughout the sample. In particular, (i), (vi), and (viii) are available only until 1873. After that date, a short title is no longer reported, the filing date is reported only sporadically, and the type of protection becomes uninformative as only granted patents are included in the sample.

C.III.3 Digitization

We perform optical character recognition (OCR) on each patent to structure the data in a machine-readable dataset. To ensure state-of-the-art performance, we OCR the first page of each document, where all the (i–ix) variables are located, using Amazon’s commercial textract engine. To retrieve the rest of the text, we use the open-source engine tesseract. An OCR-ed document is a text file. To extract the relevant variables, we harness the flexibility of large language models—specifically, GPT 3.5—to parse all variables (i–ix). The text of patents is relatively standardized, but inconsistencies arise relatively frequently due to idiosyncratic writing, phrasing, and OCR errors. The flexibility of large language models allows us to consistently and precisely extract all the relevant information for

a large (95%) share of all patent applications. Running the same model on the title of each patent, we assign the technological class to patent grants. The final database comprises approximately 303,000 patents granted between 1853 and 1899 and 800,000 between 1853 and 1940.

C.III.4 Geo-Coding

To retrieve each patent's location, we geocode each inventor's listed address using the commercial geocoding engine provided by MapTiler AG. To geocode an address, if a coarse geographical unit is listed on the patent (e.g., the county), we condition the outcome coordinates to lie within that unit. Figure [IIIb](#) reports the resulting distribution of patents per capita and Figure [C.7](#) reports the spatial distribution of patent grants across technologies. These maps are consistent with underlying population and economic development indicators and with the historical evidence (e.g., there is a substantial clustering of textile patents in the Lancashire districts).

C.III.5 Validation of Coverage

We validate the data with the two existing series covering a portion of the period 1853-1899. Hanlon ([2016](#)) digitized an index of patents issued between 1855 and 1883. This data list, for each patent, the inventor(s) and their profession(s), the technology class, and the issue year. The second dataset is the "A Cradle of Invention" (COI) series, published by Finishing Publications ([2018](#)). These data, too, were digitized from indices and therefore only list authors, issue year, and, often, titles. In principle, this series spans the years 1617–1895. However, after 1883, patent applications that were eventually denied protection were also listed. Absent a way to identify granted patents, we do not report statistics for the COI series after 1883.

Table [C.4](#) reports the aggregate number of patents issued and recorded in our series (columns 2 and 6), COI (columns 3 and 7), and Hanlon ([2016](#)) (columns 4 and 8). The three are highly consistent. Our series is closest to Hanlon ([2016](#)), but the COI figures are not too far off either. The table strongly suggests that our series is as complete as the Hanlon ([2016](#)) database. We cannot, however, externally validate it for the later part of the period because there is no data available.

C.III.6 Measuring the Similarity Between US and UK Patents and their Importance

This section describes our approach to measure the similarity of UK patents with previous US patents—i.e., the extent to which UK patents "copied" US innovations—and the novelty of UK patents. We apply state-of-the-art natural language processing (NLP) tools to patent texts to map patents into the "idea space." Once located in the idea space, we can measure the content similarity between patents and define various measures of patent novelty. Our approach builds on recent work by Ganguli et al. ([2024](#)), who, using manually annotated data, document that leading NLP models provide an adequate representation of patent texts.

We employ General Text Embeddings (GTE, Li et al., [2023](#)). GTE embeddings are state-of-the-art open-source vector representations of text trained to map semantically similar inputs close together in a shared embedding space, which we refer to as the "idea space." They substantially outperform more traditional term-frequency inverse document frequency (TF-IDF) bag-of-words representations, employed, among others, by Kelly et al. ([2021](#)), as demonstrated by Ganguli et al. ([2024](#)). Moreover, Ganguli et al. ([2024](#)) document that GTE embeddings outperform embeddings specifically designed

for patent texts, such as PaECTER, and older Sentence BERT models, particularly on historical documents.

Let $\mathbf{v}_i \in \mathbb{R}^{384}$ be the normalized 384-dimensional GTE embedding representation of patent i . For all US patents, $\mathbf{v}_i$ is obtained from the full text. For UK patents, we compute $\mathbf{v}_i$ using the full text for patents dated 1853–1899, and we use the abstract for patents between 1900 and 1940, for which the full text is not available.¹⁶ Since embeddings are normalized, the scalar product $d_{ij} = \mathbf{v}_i^\top \mathbf{v}_j$ is the cosine similarity between patents i and j , which is our baseline measure of semantic similarity between two patent documents.

Let $\mathcal{B}_i^{-k}$ be the corpus of (UK or US) patents issued no more than k years before patent i . Analogously, let $\mathcal{F}_i^{+k}$ be the corpus of patents issued no later than k years after patent i . We define the backward and forward similarity of patent i as the average similarity with previous and later patents. Formally,

$$\begin{aligned} B_i^{-k} &= \frac{1}{|\mathcal{B}_i^{-k}|} \sum_{j \in \mathcal{B}_i^{-k}} d_{ij}, \\ F_i^k &= \frac{1}{|\mathcal{F}_i^{+k}|} \sum_{j \in \mathcal{F}_i^{+k}} d_{ij}. \end{aligned} \tag{C.6}$$

Here, $|\mathcal{B}_i^{-k}|$ and $|\mathcal{F}_i^{+k}|$ respectively denote the number of patents in the backward and forward sets.

We measure the similarity between a given UK patent and previous US patents as B_i^{-k} , for various levels of k , by including US patents in the set $\mathcal{B}_i^{-k}$. This variable quantifies the semantic proximity between UK patents and the stock of previously issued US patents.

Second, we define a proxy for “importance,” following Kelly et al. (2021), who, however, employ bag-of-words instead of embedding representations. Intuitively, a patent should be “important” if it introduces concepts that were not present in previous patents and that are present in later innovations. Building on this idea, we define our proxy for patent importance as

$$Q_i^k = \frac{\tilde{F}_i^k}{\tilde{B}_i^k} \tag{C.7}$$

where $\tilde{F}_i^k$ and $\tilde{B}_i^k$ are the forward and backward similarity measures obtained by comparing UK patent i with later and previous UK patents issued over a k -year window around the issue year of patent i . Following the practical recommendation of Ganguli et al. (2024) and Kelly et al. (2021), we residualize the baseline forward and backward similarity against year and technology fixed effects. This is motivated by the fact that some technological fields are inherently more cumulative or standardized, leading patents in those areas to exhibit systematically higher similarity to both prior and subsequent innovations. Year fixed effects, in turn, absorb mechanical variation in similarity arising from changes in the size and composition of the patent corpus over time, including the expansion in the number of patents (which increases the likelihood of close matches), shifts in the distribution of technologies being patented, and secular changes in document length and drafting conventions that affect embedding-based similarity independently of underlying inventive content.

¹⁶In practice, and following the literature, the GTE model we employ truncates patent texts to the first 384 tokens (≈ 500 words). Since abstracts are typically longer than 384 tokens, this difference is immaterial.

C.III.7 Validation of the Technology Class Imputation

We apply a flexible large language model to patent texts to impute CPC technology classes, which we then map to the coarser technology classification we employ in the analysis. This approach leverages the flexibility of LLMs to tackle OCR errors and linguistic nuances that may be challenging for less sophisticated classification algorithms to capture. At the same time, LLM-based classification may suffer from “hindsight” bias because LLMs are trained on a predominantly modern language corpus.

To explore the severity of this concern, we compare the coarse technology classification delivered by the LLM to a classifier trained exclusively on historical US patent texts. Patent data for the US by Berkes (2018), in fact, contains CPC technology classes, which we map to our coarser technology classification. Then, we construct a supervised classifier that maps US patent text embeddings into the technology categories.¹⁷ The classifier is multi-label because one patent can be assigned to more than one technology class.

Let $\mathbf{v}_i$ denote the embedding of patent i , and let Y_i be a vector of binary indicators for technology membership. The problem is a multi-label classification problem with 19 technologies and a large number of observations. We estimate a one-vs-rest linear classifier. Specifically, for each technology k , we fit a logistic model of the form

$$\Pr(Y_{ik} = 1 \mid \mathbf{v}_i) = \sigma(\mathbf{v}_i^\top \beta_k), \quad (\text{C.8})$$

where $\sigma(\cdot)$ is the logistic function. The multi-label structure estimates one binary classifier per technology and combines their predicted probabilities.

We split the sample of US patents into training (70%), validation (15%), and test (15%) sets. Let τ be the probability threshold for converting predicted probabilities into discrete labels. For a given threshold τ , we assign technology k to patent i if $\hat{p}_{ik} \geq \tau$. To avoid empty predictions, we impose that each patent is assigned at least one technology by selecting the highest-probability label when no probability exceeds the threshold. We choose τ to maximize the F1 score on the validation set.

After selecting the optimal threshold, we re-estimate the classifier on the combined training and validation samples and evaluate performance on the held-out test set. Table C.5 reports the results. Column (1) reports the results on the train set we use to estimate the model. Column (2) reports the results we obtain on the validation set, which we do not use to estimate the model but use to optimally select the threshold for technology assignment. Column (3) refers to the test set, i.e., the set of patents we exclude from both model estimation and fine-tuning. Panel A reports technology-level statistics: precision, recall, and the harmonic mean between the two. Panel B reports patent-level statistics: precision, recall, and the Jaccard overlap.¹⁸ The model delivers the same performance on training, validation, and test sets, indicating no evidence of overfitting. Despite the simple specification, the classifier correctly predicts approximately 72% of technology classes with a recall rate of roughly 70%.

In column (4), we compare the technology classes of UK patents obtained with GPT with those of the classifier. The classifier’s ability to predict GPT-predicted technology classes is as strong as its

¹⁷We employ the same embeddings used to construct the patent similarity and novelty measures, as described in Appendix C.III.6.

¹⁸Jaccard overlap is defined as $|A \cap B| / |A \cup B|$, where A is the set of technologies assigned by the classifier and B is the reference set (observed US labels in columns (1–3), GPT labels in column 4).

ability to predict actual technology classes. This indicates that the classifier reproduces the broader technology patterns identified by GPT and that its performance remains stable when moving from observed labels in the US sample to GPT-based labels in the UK sample. Overall, the similarity in classifier performance when evaluated against observed US technology classes and against GPT-predicted UK technology classes provides validation for the GPT-based classification procedure. If the GPT assignments were noisy, agreement in the UK sample would be substantially weaker. Instead, the comparable performance across the two settings suggests that GPT assigns technology classes in a systematic and meaningful way.

C.III.8 Census Linking

To perform the individual-level analysis on neighborhood emigration, we link the inventors listed on patent documents issued between 1881 and 1900 to the 1891 population census. The availability of census data determines the sample restriction. First, we aim to follow the inventing activity of individuals over time. Therefore, we link patent records to a single census. When linking inventors to the census, we use the location listed on the document to restrict the pool of potential matches in the census. This assumes that the inventor did not move between the time when the patent was granted and the closest census year, which is a plausible assumption only for patents issued not too far from the closest census year. Since census records for 1871 are not available, we restrict our attention to patents granted no earlier than ten years before the closest census, i.e., 1891.

Matching Algorithm Given a patent p , define the set of its inventors as $\mathcal{A}_p = \{A_1, \dots, A_{n_p}\}$. Most patents are solo-authored in this period, meaning $|\mathcal{A}_p| = 1$. Call $\mathcal{L}_p = \{\ell_1, \dots, \ell_{m_p}\}$ the set of locations patent p is associated to. Each ℓ is a pair of latitude-longitude coordinates. Let $\mathcal{L}_p^{\text{parish}}$ be the set of parishes associated with each coordinate. Analogously, let $\mathcal{L}_p^{\text{neighbor parishes}}$, $\mathcal{L}_p^{\text{district}}$, $\mathcal{L}_p^{\text{neighbor districts}}$, and $\mathcal{L}_p^{\text{county}}$ be the set of, respectively, neighboring parishes, districts, neighboring districts, and counties where each coordinate locates. These are progressively coarser units: parishes are contained within districts, which are nested within counties. Because we do not observe the inventor-location pair directly, we proceed as follows to match a generic A_p :

1. With a slight abuse of notation, let $\mathcal{L}_p^{\text{parish}}$ —and, analogously, $\mathcal{L}_p^{\text{district}}$ and $\mathcal{L}_p^{\text{county}}$ —denote the set of census records in each parish, district, and county within the respective sets.
2. Take all entries i within the set of parishes $\mathcal{L}_p^{\text{parish}}$ that are at least 18 when the patent p is issued. Let year_i and t_p denote, respectively, the birth year of i and the issue date:

$$\mathcal{M}_{A_p}^{\text{parish}} = \left\{ i \in \mathcal{L}_p^{\text{parish}} \mid t_p - \text{year}_i \geq 18 \right\} \quad (\text{C.9})$$

3. For each $i \in \mathcal{M}_{A_p}^{\text{parish}}$, compute the distance between the name and surname of i and that of A_p :

$$\delta_i^{A_p} = \alpha \times d_N(A_p, i) + (1 - \alpha) \times d_S(A_p, i) \quad (\text{C.10})$$

for some $\alpha \in [0, 1]$. In our baseline setting, we pick $\alpha = 0.3$ as in the migrant linked algorithm to assign a larger weight to the surname. The terms $d_N(A_p, i)$ and $d_S(A_p, i)$ denote the name and surname Jaro-Winkler similarities between record A_p and i .

4. Define the set of acceptable matches as those with the highest similarity with the given A_p :

$$\overline{\mathcal{M}}_{A_p}^{\text{parish}} = \left\{ i \in \mathcal{M}_{A_p}^{\text{parish}} \mid \delta_i^{A_p} = \bar{\delta}^{A_p} \equiv \max_{i' \in \mathcal{M}_{A_p}^{\text{parish}}} \delta_{i'}^{A_p} \right\} \quad (\text{C.11})$$

where $\bar{\delta}^{A_p}$ denotes the similarity between all elements in $\overline{\mathcal{M}}_{A_p}^{\text{parish}}$ and A_p . Notice that this is the same across all $i \in \overline{\mathcal{M}}_{A_p}^{\text{parish}}$.

5. Set a threshold τ such that if $\bar{\delta}^{A_p} < \tau$, $\overline{\mathcal{M}}_{A_p}^{\text{parish}} = \emptyset$, otherwise pass. As in the UK-US census-linked sample, we select $\tau = 0.95$.

6. If $\overline{\mathcal{M}}_{A_p}^{\text{parish}}$ is not empty, then inventor A_p is matched to all records in $\overline{\mathcal{M}}_{A_p}^{\text{parish}}$. If empty, repeat steps 2–4 conditioning on records in the coarser matching set.

Patent data have the clear advantage of providing geographical information on the location of inventors. However, inventors are mobile, and there may be a considerable time lag between the patent grant date and the 1891 census. For these reasons, we incrementally exploit geographical information on the inventor's location. First, we look for high-quality matches within the same parish where the patent is filed. We then progressively expand the set of records by coarsening their geographic location to their neighboring parishes, districts, neighboring districts, and counties.

Evaluation of the Matching In Figure C.8, we report the matching rate from this procedure. We successfully match approximately 75% of inventors in the sample, and this share remains stable over time. Almost 50% of the matches are attained at the parish level—the smallest geographical unit we consider—and about half of the remaining matches are obtained within neighboring parishes of the location indicated in the patent document. These figures indicate that mobility remained relatively low within the relatively narrow time frame around the census year we analyze.

A plausible concern is that the probability of obtaining a link is not random. This may be the case if, for instance, more successful inventors were more educated and therefore more likely to report their names correctly in the census. On the other hand, if successful inventors were relatively more mobile, we may fail to link them because obtaining a match would require expanding the search to a national level, which would likely result in multiple potential matches and lead us to drop them. These hypotheses are difficult to test directly. Figure C.9 reports the overall distribution of the number of matches in the sample. Approximately 60% of the inventors listed on patent documents are linked to a single census record, and the share of inventors linked to more than 10 records is negligible. In the baseline analysis, we focus on the subsample of single matches to reduce the probability of spurious matches, but all the results are quantitatively similar when using multiple matches and weighting them by the inverse of the number of matches.

Table C.6 reports the correlation between the number of matches and a set of individual-level characteristics. Panel A displays the age and the residence divisions, while Panel B lists the occupation reported in the census. The number of matches is correlated with some of these characteristics, but the magnitude of this association is small, except for the fact that inventors living in Wales have substantially more matches in the sample. To ensure that Wales does not drive our results, we replicate the main results, excluding them from the sample, and confirm that the results hold.

C.III.9 A Method to Measure Technological Leadership from Patent Texts

We construct a measure of British and American technological leadership using the textual content of patents. The baseline intuition extends the logic of Kelly et al. (2021) to the cross-country setting. In our context, a country is closer to the technology frontier in a given sector when patents granted later in that sector resemble its current patents more than they resemble patents granted in the other country.

We start from the embeddings constructed in Appendix C.III.7. Let patent i be represented by the embedding vector $\mathbf{v}_i$. The vector is normalized to have unit length. Hence, for any two patents i and j , the scalar product $\mathbf{v}_i^\top \mathbf{v}_j$ is their cosine similarity. Let $\mathcal{P}_{ckt}$ denote the set of patents granted in country $c \in \{\text{UK}, \text{US}\}$, technology class k , and decade t . For every patent i granted in the same technology class in the following decade, we compute its proximity to the contemporaneous knowledge base of country c as follows:

$$A_{c,kt}(i) \equiv \frac{1}{K} \sum_{j \in \mathcal{N}_K(i; \mathcal{P}_{ckt})} \mathbf{v}_i^\top \mathbf{v}_j, \quad (\text{C.12})$$

where $\mathcal{N}_K(i; \mathcal{P}_{ckt})$ is the set of the K patents in $\mathcal{P}_{ckt}$ that are most similar to i . In the baseline specification, we set $K = 100$ and select the most similar patents through a K-Nearest Neighbors search. This choice allows the measure to compare each future patent with a broad set of related contemporaneous patents, while avoiding cases where proximity is dominated by patents in remote parts of the same technology class.

We then compare the proximity of future patents to the British and American contemporaneous corpora. Let $\mathcal{F}_k^{t+10}$ denote the set of future patents granted in technology class k in decade $t + 10$. We define relative leadership in k and t as

$$L_{kt} \equiv 2 \left[\frac{1}{|\mathcal{F}_k^{t+10}|} \sum_{i \in \mathcal{F}_k^{t+10}} 1\{A_{\text{UK},kt}(i) > A_{\text{US},kt}(i)\} \right] - 1. \quad (\text{C.13})$$

The term in square brackets is the share of future patents that are closer to contemporaneous British patents than to contemporaneous American patents. We re-center it around zero to ease interpretation. Positive values indicate that subsequent innovation is relatively closer to British patents; negative values indicate that it is relatively closer to American patents; values close to zero indicate no clear leadership.

To ensure that the measure does not mechanically reflect differences in patenting volume, we compute L_{kt} on balanced samples. For each technology class and decade, we draw the same number of British and American patents from the contemporaneous period and from the future period. We repeat this procedure across multiple draws ($T = 200$) of equal size ($N = 500$) and average the resulting estimates. Cells without sufficient support in either country are excluded.

Panel C.11a reports a Kalman-smoothed technology-decade series. Darker shades of red indicate higher values of L_{kt} , which in turn display sector-decade pairs where the UK leads the US. Darker shades of blue, by contrast, indicate technologies and decades where the US is the technological leader. Consistent with historical evidence, we find that, between 1870 and 1900, the UK retained a relative advantage in technologies associated with the First Industrial Revolution, including textiles,

engines and pumps, metallurgy, and mechanical engineering. The US frontier position strengthened in sectors that became central to later industrialization, such as electricity, instruments, chemistry, and transport, in the second part of the sample period (Rosenberg, 1982; Nelson and Wright, 1992). By 1930, the US led in virtually all technologies.

Panel C.11b provides a patent-level validation exercise. We focus on well-known inventions whose historical relevance is clear from the subsequent development of their industries, including Bakelite (which we described in Section 5.1), the Tesla alternating-current system, the safety bicycle, Bell's telephone, Edison's electric lamp, and the Westinghouse air brake. For each patent, we compute its similarity to subsequent patents in the same technology class and locate it in the corresponding distribution of future-proximity scores. The bars report the percentile of each patent in that distribution. Higher percentiles indicate patents whose language is close to later innovations in the same field. All these pivotal innovations lie in the top 75% of patents that are closest to the joint US-UK technology frontier. In the case of Tesla's and Marconi's patents, for example, this proximity approaches the top 1%, attesting to their influence in later work in telecommunications and electricity.

Table C.1. Validation of the Linked Migrants on Random Subsample

(a) Classification Report

	Full		Test Set		
	Random Sample		Random Sample		
	(1)	(2)	(3)	(4)	(5)
	Manual	GPT	Manual	GPT	BERT
Panel A. Accepted Matches (Jaro-Winker Classifier)					
Rejected	23	22	4	6	10
Accepted	977	978	178	176	172
True Positive Rate	0.98	0.98	0.98	0.97	0.95
Panel B. Not Accepted Matches (Jaro-Winker Classifier)					
Rejected	825	644	180	141	188
Accepted	175	356	38	77	30
False Negative Rate	0.17	0.36	0.17	0.35	0.14

(b) Confusion Matrix

	Jaro-Winkler Classifier					
	Accepted Matches			Rejected Matches		
	(1)	(2)	(3)	(4)	(5)	(6)
	Rejected	Accepted	Total	Rejected	Accepted	Total
Panel A. GPT Classifier (Full Random Sample)						
Rejected	11	11	22	636	8	644
Accepted	12	966	978	189	167	356
Total	23	977	1000	825	175	1000
Panel B. BERT Classifier (Test Set Random Sample)						
Rejected	1	9	10	170	18	188
Accepted	3	169	172	10	20	30
Total	4	178	182	180	38	218

Notes. Table C.1a reports the number of accepted and rejected matches on two random 1,000 samples (columns 1–2) and on the test set of the BERT classifier (columns 3–5). In columns (1) and (3), we report the results of the manual verification; columns (2) and (4) refer to the classification generated by GPT, and column (5) refers to the BERT classifier. Panel A reports the results on the 1,000 random sample of matches that we accept in the linked dataset following the Jaro-Winkler-based classifier described in Appendix C.III.8. Panel B refers to the other 1,000 sample of matches that we reject according to the same Jaro-Winkler classifier. Table C.1b reports confusion matrices comparing the GPT (Panel A) and BERT (Panel B) classifiers against manually verified classifications within subsamples defined by the baseline Jaro-Winkler classifier. Panel B refers to the test set of the BERT classifier. Columns (1–3) refer to the subsample of matches accepted by the Jaro-Winkler classifier; columns (4–6) refer to matches rejected by the Jaro-Winkler classifier. Referenced on pages: 10, A41, A42, A43.

Table C.2. Validation of the Linked Migrants Full Sample

BERT Classifier	Jaro-Winkler Classifier	
	(1) Accepted Matches	(2) Rejected Matches
Rejected	20622	182162
Accepted	339135	20093
Total	359757	202255
True Positive Rate	0.94	
False Negative Rate		0.10

Notes. This Table reports the results of validating the full sample of linked migrants using the BERT classifier trained on the manually annotated random subsample. In column (1), in the first row, we report the number of matches that the Jaro-Winkler classifier identifies as valid that the BERT classifier rejects. In the second row, we report the number of matches that both the Jaro-Winkler and the BERT classifier classify as valid. In column (2), we replicate on the sample of matches that the Jaro-Winkler algorithm rejects: in the first row, we report those that are also rejected by the BERT classifier, while in the second row, we report those that are accepted by the BERT classifier. The Table also reports the share of true positive matches identified by the Jaro-Winkler classifier, when taking the results of the BERT classifier as ground truth. Referenced on pages: [10](#), [A43](#).

Table C.3. Balance of Linked Emigrants Sample

	Correlation with Matching Status			
	= 1 if Not matched (1)	= 1 if Matched (2)	Diff. (3)	Std. Err. (4)
Panel A. Individual Characteristics				
Literacy	0.898	0.965	0.066	(0.057)
Income	3.146	3.141	-0.005	(0.009)
Panel B. Occupation				
Professional	0.052	0.041	-0.011***	(0.002)
Farmer	0.087	0.091	0.004	(0.008)
Manager	0.065	0.059	-0.006*	(0.003)
Clerical	0.056	0.044	-0.012***	(0.003)
Sales	0.060	0.046	-0.014***	(0.001)
Craftsman	0.263	0.281	0.018**	(0.008)
Operative	0.310	0.333	0.023*	(0.012)
Service	0.050	0.049	-0.001	(0.003)
Laborer	0.057	0.056	-0.001	(0.003)
Panel C. Region of Residence				
North East	0.518	0.543	0.025***	(0.006)
Midwest	0.278	0.269	-0.009***	(0.003)
South	0.056	0.044	-0.012***	(0.001)
West	0.147	0.144	-0.004	(0.007)

Notes. This Table reports the correlation between the matching probability in the migrants-linked sample and a set of individual-level characteristics observed in the US census. The dependent variable equals one if the immigrant is linked and zero otherwise. We report the average value of the row variable for unmatched (column 1) and matched immigrants (column 2), as well as the difference between the two groups (column 3) along with its standard error clustered at the census year level (column 4). All row variables are indicators except income, which is the log of the occupational income score. Standard errors are clustered at the census year level and are displayed in parentheses. Referenced on pages: 37, 9, A44.

*: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table C.4. External Validation of Newly Digitized Patent Data

Years 1853–1876				Years 1877–1899			
Year (1)	New Data (2)	Hanlon (3)	COI (4)	Year (5)	New Data (6)	Hanlon (7)	COI (8)
1853	2926		3042	1877	4937	4940	5133
1854	2737		2715	1878	5338	5333	5258
1855	2868	2955	2883	1879	5312	5325	5442
1856	3204	3102	3003	1880	5531	5509	5211
1857	3175	3197	3108	1881	5762	5745	5760
1858	3071	2999	2988	1882	6244	6233	6187
1859	3031	2998	3008	1883	6118	5981	6075
1860	3231	3190	3161	1884	9809		
1861	3279	3272	3303	1885	8695		
1862	3479	3486	3485	1886	8913		
1863	3312	3308	3411	1887	9070		
1864	3245	3257	3286	1888	9283		
1865	3391	3378	3436	1889	10315		
1866	3422	3452	3481	1890	10376		
1867	3720	3720	3723	1891	10768		
1868	3976	3984	3955	1892	11454		
1869	3837	3781	3781	1893	11986		
1870	3459	3405	3323	1894	11664		
1871	3574	3525	3542	1895	12243		
1872	3951	3967	4013	1896	13619		
1873	4261	4282	4336	1897	14304		
1874	4419	4491	4533	1898	13105		
1875	4576	4557	4537	1899	13267		
1876	5048	5064	5085				

Notes. This Table reports the total number of patents in England and Wales between 1853 and 1899. Columns (2) and (6) report the series constructed from our newly digitized dataset; columns (3) and (7) report data from Hanlon (2016); columns (4) and (8) tabulate data from *A Cradle of Inventions* (Finishing Publications, 2018). The *A Cradle of Inventions* series potentially stretches until 1899. However, after 1883, there is no way to distinguish between patents granted and applications. Hence, we do not report figures for these later years (Nicholas, 2014). Data from Hanlon (2016) only cover the years 1855–1883. Referenced on page: A47.

Table C.5. Validation of GPT Technology Imputation

	US Patents			UK Patents
	Train (1)	Validation (2)	Test (3)	(4)
Panel A. Technology-level				
Precision (True Positives / All Classifier Assignments)	0.717	0.716	0.717	0.757
Recall (True Positives / All Reference Assignments)	0.696	0.696	0.696	0.679
F1 (Harmonic Mean)	0.707	0.706	0.706	0.715
Panel B. Patent-level				
Within-Patent Precision	0.720	0.719	0.720	0.759
Within-Patent Recall	0.711	0.711	0.711	0.716
Jaccard (Patent-Level Overlap in Assignments)	0.705	0.705	0.705	0.712

Notes. This Table compares the true technology classes of US patents (columns 1–3) and the GPT-imputed technology classes of UK patents (column 4) against the technology classes predicted by the classifier described in Appendix C.III.7. Columns (1–3) report agreement between classifier-predicted and observed technology assignments for US patents in the training, validation, and test splits, respectively. Column (4) reports agreement between classifier-predicted and GPT-assigned technologies for UK patents. Panel A computes technology-level precision, recall, and F1 from patent-technology pair matches. Panel B reports patent-level averages of within-patent precision, within-patent recall, and Jaccard overlap. Predictions are generated using the trained one-vs-rest classifier with the selected probability threshold and at least one technology enforced per patent. Referenced on page: [A49](#).

Table C.6. Balance of Linked Inventor Sample

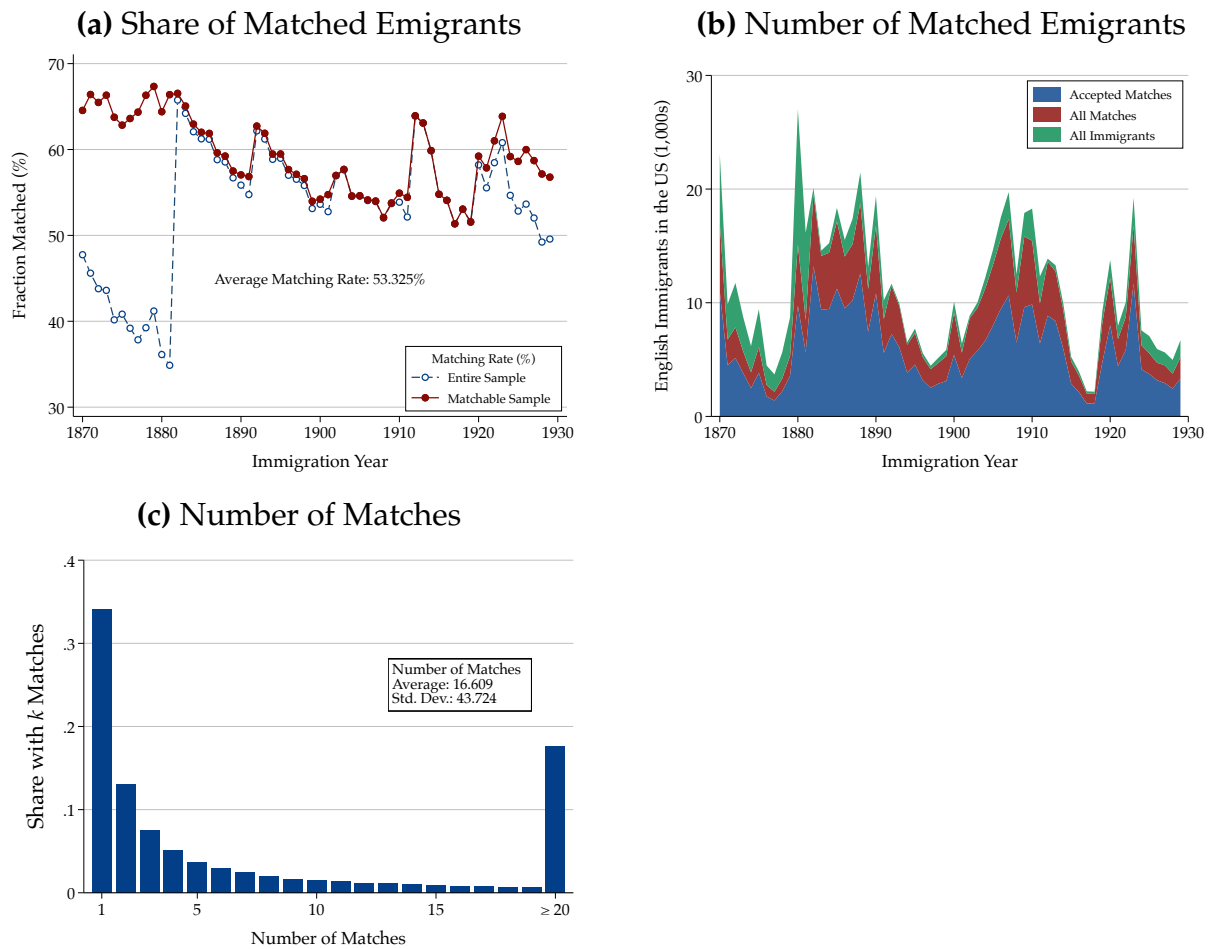
	Unconditional		Year FE	
	Mean (1)	Std. Err. (2)	Mean (3)	Std. Err. (4)
Panel A. Age and Place of Residence				
Age	-0.002	(0.002)	-0.002	(0.001)
London	-0.197	(0.144)	-0.177	(0.128)
South East	-0.362***	(0.132)	-0.342***	(0.114)
East	-0.246*	(0.149)	-0.242*	(0.136)
Yorkshire	-0.324**	(0.137)	-0.308**	(0.122)
South West	-0.412***	(0.124)	-0.404***	(0.114)
West Midlands	0.022	(0.199)	0.022	(0.186)
East Midlands	-0.246*	(0.135)	-0.230*	(0.119)
North East	-0.260**	(0.133)	-0.234**	(0.110)
North West	-0.165	(0.159)	-0.169	(0.154)
Wales	1.809***	(0.227)	1.752***	(0.186)
Panel B. Occupation				
Agriculture	0.528	(0.365)	0.494	(0.320)
Chemicals	-0.209**	(0.105)	-0.189**	(0.083)
Construction	-0.061	(0.076)	-0.055	(0.068)
Engineering	-0.070	(0.062)	-0.063	(0.052)
Liberal Professions	-0.186***	(0.071)	-0.180***	(0.064)
Metallurgy	0.114	(0.081)	0.109	(0.075)
Public Administration	-0.150	(0.094)	-0.148*	(0.089)
Textiles	-0.119	(0.093)	-0.109	(0.083)
Trade	-0.164*	(0.084)	-0.155**	(0.072)
Transports	0.016	(0.027)	0.018	(0.024)
Utilities	-0.178**	(0.080)	-0.170**	(0.071)

Notes. This Table reports the association between the number of matches in the linked inventor sample and a set of individual-level covariates observed in the census. In columns (1–2), we report the estimated coefficients from unconditional regressions and the associated standard errors. In columns (3–4), we repeat the exercise controlling for the patent issue year. Standard errors are clustered at the county level and are displayed in parentheses. Referenced on pages: [A51](#), [A7](#).

*: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

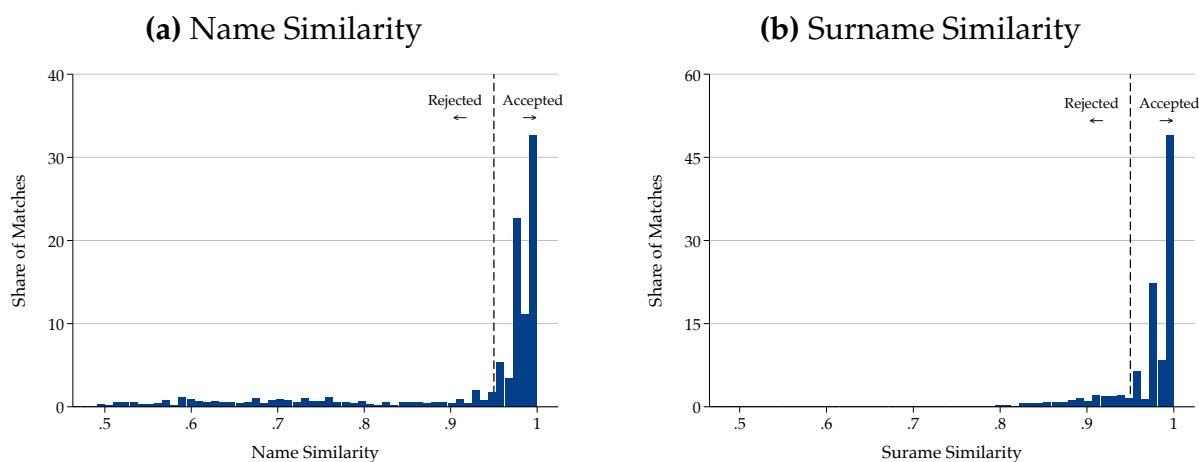
Figures

Figure C.1. Matching Rate and Number of Matches



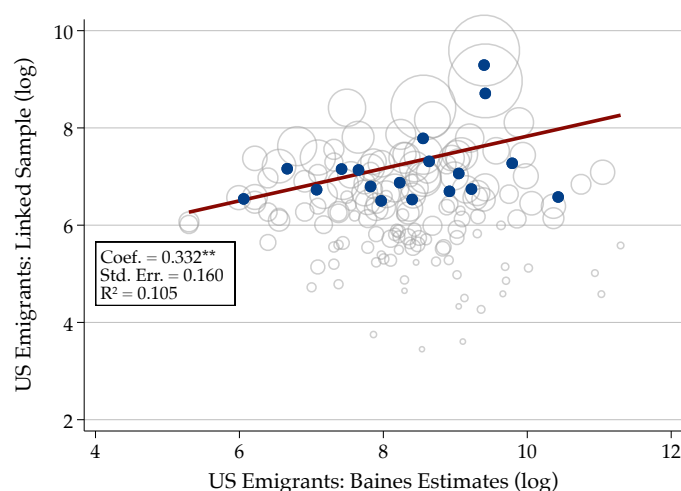
Notes. This Figure reports information on the matching performance of the emigrants' linked sample. In Panel C.1a, we display the matching rate, i.e., the share of emigrants recorded in the US census that are successfully linked to at least one entry in the UK census. The blue dots refer to the overall sample, and the red dots refer to the emigrants who could be recorded in the UK census (see main text for more precise information). In Panel C.1b, we plot the overall number of immigrants in the US census (in green), the number of those that are matched (in red), and the number of matches that we accept (in blue). Panel C.1c reports the distribution of the number of matches, where the last bin collects all those with more than 20 matches. Referenced on pages: 9, A40.

Figure C.2. Quality of Matches



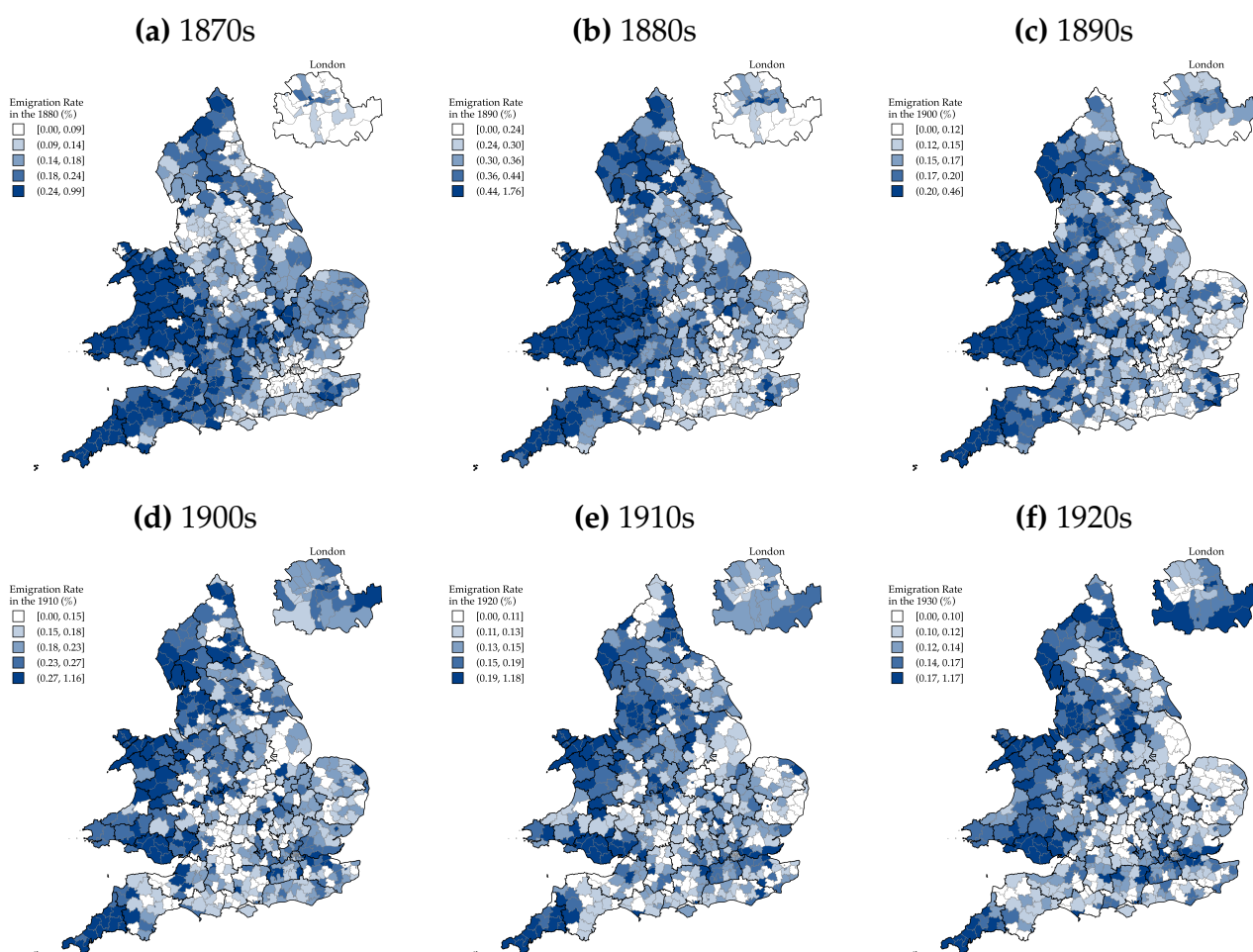
Notes. This Figure reports the distribution of the match quality in terms of name and surname similarity for the set of records with no more than two matches in the baseline sample. The similarity measure we use to construct the links is the Jaro-Winkler. This string metric measures the edit distance between the name and surname of the British immigrant recorded in the US census and their match(es) in the UK census. Panel C.2a reports the distribution of the name similarity; Panel C.2b refers to surnames. The vertical lines mark the quality thresholds we impose for a match to be part of the final linked sample. Referenced on pages: 9, A41.

Figure C.3. External Validation with Out-Migration Estimates



Notes. This Figure reports the correlation between out-migration in our linked inter-census sample and the estimates produced by Baines (2002). These estimates are at the county level and span 1880–1910. We thus aggregate our data by county. Furthermore, we exclude the London area because we cannot map our data to Baines’ geographical divisions. Both variables are expressed in log terms. The gray dots display county-decade observations. Counties are weighted by their population. The blue dots report binned means. The red line overlays a linear fit between the variables. The Figure also reports the regression coefficient, its robust standard error, and the R^2 of the regression. Referenced on pages: 10, A45.

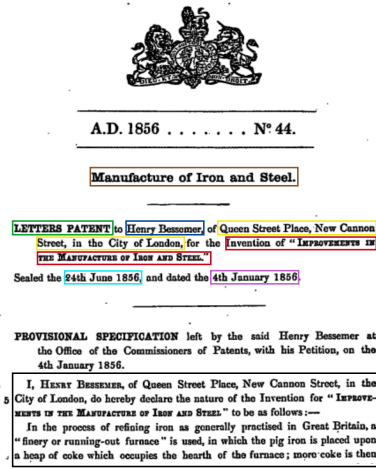
Figure C.4. Geographic Distribution of Emigration Rates Over Time



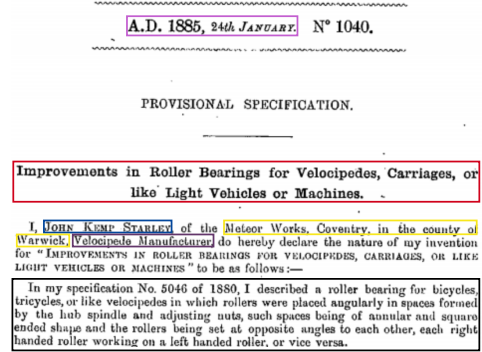
Notes. This Figure reports the distribution of US emigrants across districts in England and Wales over the period 1870–1940 by decade. Data are from the matched emigrants sample. The number of emigrants in each decade is normalized by population in 1891 and is expressed in percentage terms. Districts are displayed at their 1891 historical borders. Black edges also display historical county borders. Out-migration is cross-walked to consistent historical borders. Lighter to darker shades of blue indicate increasing quantiles of the emigration rate. The London area is displayed separately. Referenced on page: 11.

Figure C.5. Sample Annotated Patent Documents

(a) Henry Bessemer's 1856 Patent



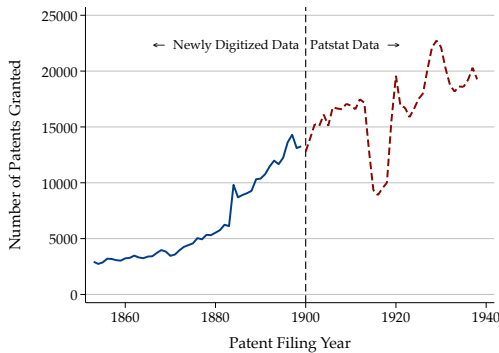
(b) John K. Starley's 1885 Bicycle Patent



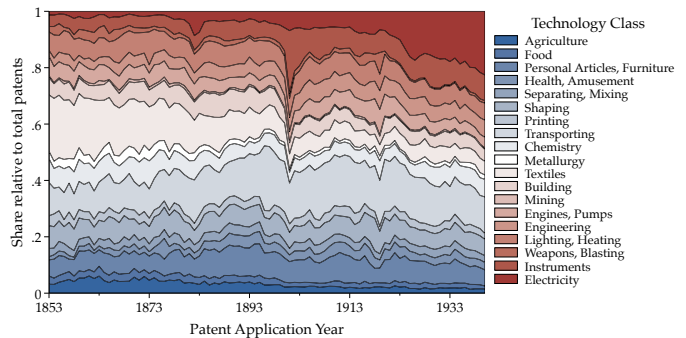
Notes. This figure displays two sample patent documents in our dataset. Panel C.5a was granted to Henry Bessemer in 1856 to invent the famous eponymous process for the mass production of steel from molten pig iron. Panel C.5b was granted to John Starley in 1885 to invent the first modern bicycle, which would soon revolutionize mobility in Europe and the US. Colors mark different variables that we structure in the dataset: (i) in brown, the short title; (ii) in red, the complete title (iii) in green, the type of protection granted; (iv) in blue, the author(s) name(s); (v) in yellow, the author(s)'s address(es); (vi) in light blue, the application date; (vii) in purple, the issue date; (viii) in black, the patent text that continues in the rest of the patent document; (ix) in dark purple, the author(s) profession(s). Not all (i–ix) data are available on every patent and in each year. Referenced on page: A46.

Figure C.6. Number and Composition of Patents Over Time

(a) Time Series of Patents

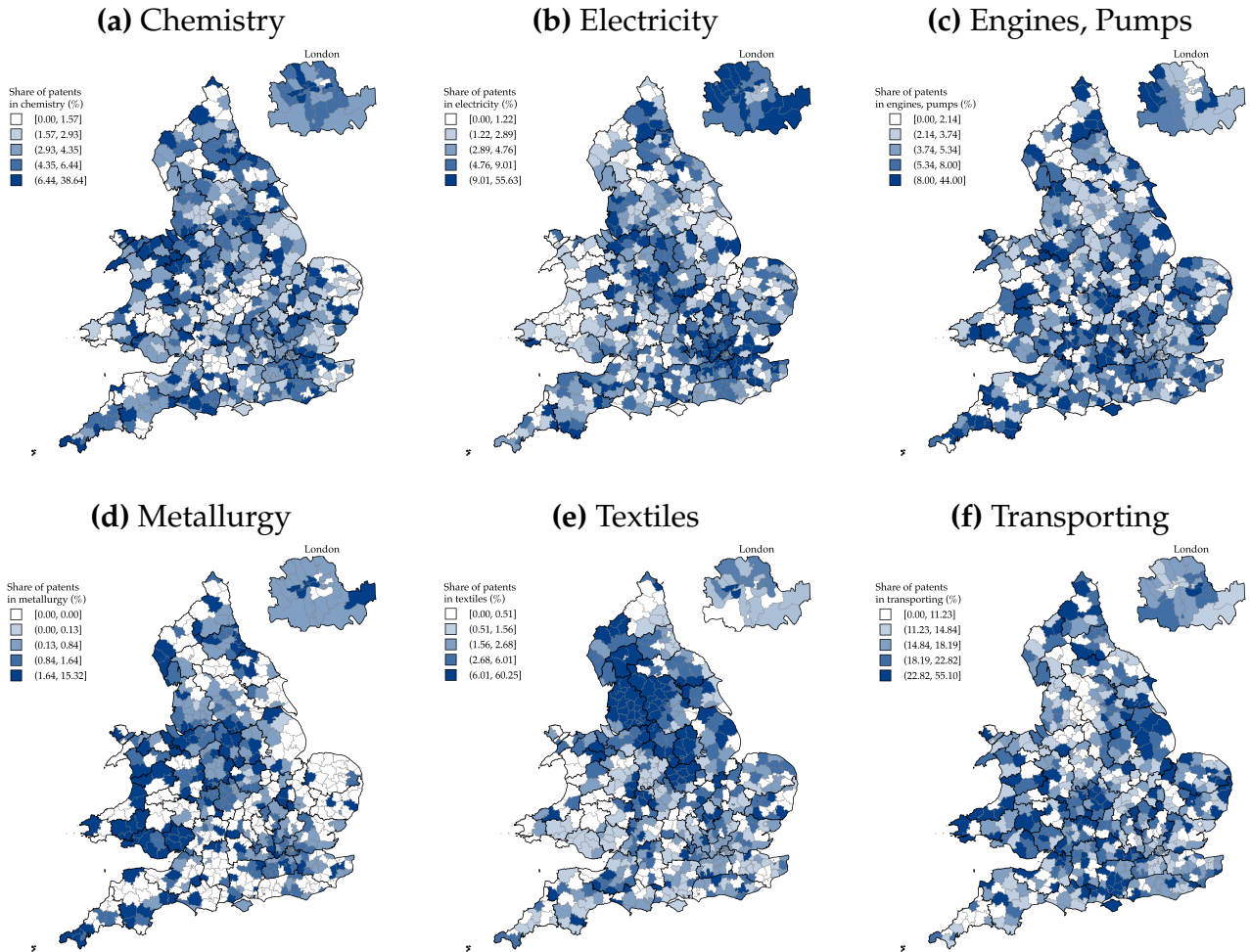


(b) Composition of Patents Across Technologies



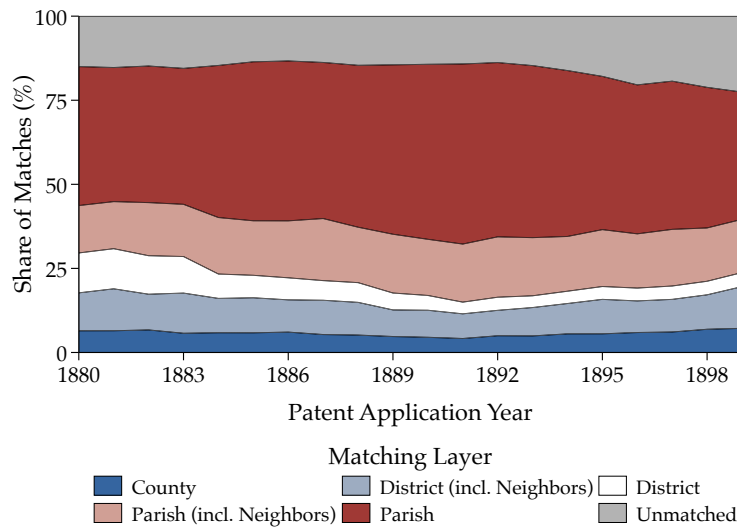
Notes. This Figure reports time-series information on the innovation activity in Britain between 1853 and 1939. In Panel C.6a, we report the total number of patents granted in the UK over the period. The blue dots report the newly digitized data that we assembled for this paper; the red dots report tabulations from the Patstat repository. In Panel C.6b, we plot the share of patents granted over time across technology classes. Referenced on page: 12.

Figure C.7. Geographic Distribution of Patents Across Technologies



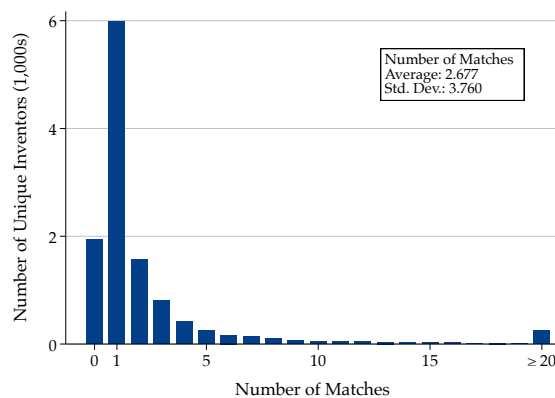
Notes. This Figure reports the intensity of patenting activity across districts over 1880–1939 for selected technology classes. Districts are displayed at 1891 borders. To assign patents to districts, we geo-reference the address of each author listed in the patent document and assign districts based on historical district borders. Black edges display county borders. The London area is displayed separately. Darker shades of blue indicate increasing quantiles of the patenting rate, defined as the percentage ratio between the number of patents in a given technology class and the overall number of patents produced. Referenced on page: [A47](#).

Figure C.8. Matching Rate of Linked Inventors Sample



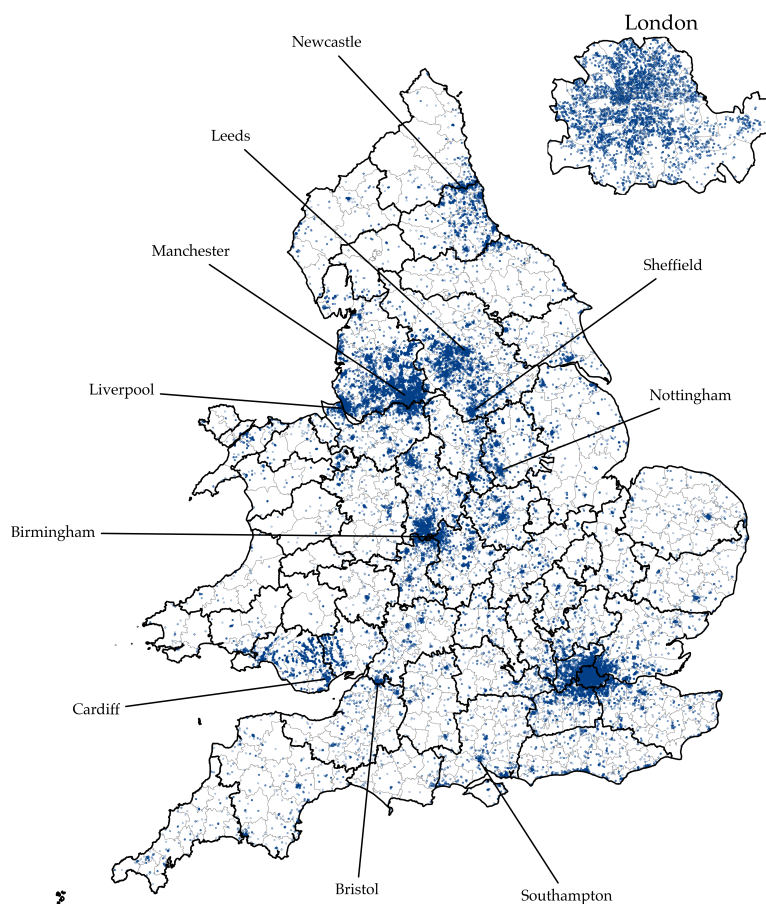
Notes. This Figure reports the matching rate of the linked sample of inventors. The records of the inventors who obtained a patent between 1880 and 1899 are linked to the 1891 population census, as detailed in the main text. The matching rate, i.e., the share of inventors successfully matched to the census, is reported on the y-axis in percentage points. The matching rate is broken down by the geographic layer of aggregation where the match is attained. Hence, we match over 75% inventors to the census throughout the period, and among those, the census record of slightly more than 50% of them is found in the same parish where the inventor is recorded living on the patent document. Referenced on page: [A51](#).

Figure C.9. Number of Matches in Linked Inventor Sample



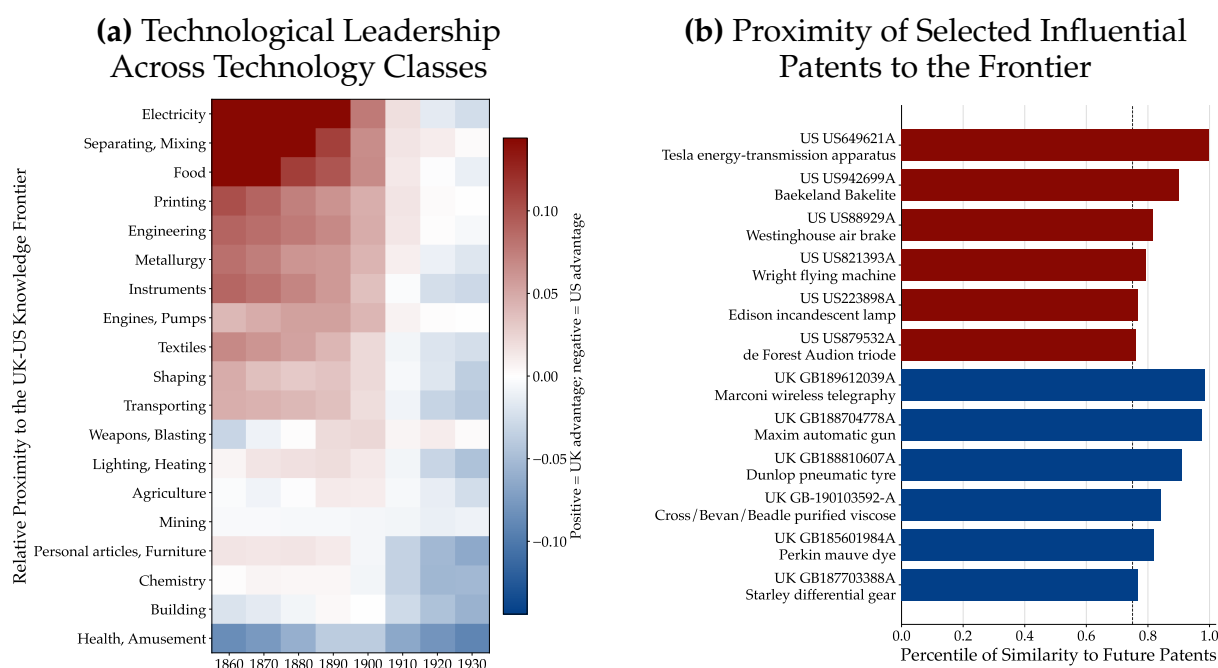
Notes. This Figure reports the number of census entries each inventor in the linked sample is matched to, including the share of inventors that we cannot match. We separately report the average number of matches and its standard deviation. Referenced on page: [A51](#).

Figure C.10. Geographic Distribution of Linked Inventors



Notes. This Figure displays the spatial distribution of inventors across districts between 1880 and 1900. Each marker reports one inventor, defined as an individual who obtains at least one patent over the sample period. To retrieve the coordinates of the inventors, we first link population censuses, whose entries are, in turn, geo-referenced. The background map displays black counties and gray districts at historical borders in 1891. We highlight the ten largest urban centers at the time. Referenced on page: [14](#).

Figure C.11. The UK-US Knowledge Frontier: 1880–1930



Notes. This Figure displays the results of the algorithm employed to characterize British and American technological leadership across technologies, as described in Appendix C.III.9. Panel C.11a shows relative leadership across the technology taxonomy. The measure is constructed as the share of future patents within a technology class that are closer to contemporaneous British patents than to contemporaneous American patents. Positive values indicate that subsequent innovation is more similar to British patents; negative values indicate greater proximity to American patents. Panel C.11b displays a patent-level validation exercise. Each bar corresponds to a patent's percentile in the distribution of similarity to future patents in the same technology class. Patents are selected from among the most influential innovations of the period. Higher percentiles correspond to patents located closer to the subsequent technology frontier. Referenced on page: 30.