Immigration and Social Distance: Evidence from Newspapers during the Age of Mass Migration*

Elliott Ash, Anton Boltachka, Gloria Gennaro,

Dominik Hangartner, Alessandra Stampi-Bombelli†

December 12, 2025

[PLEASE DO NOT CIRCULATE]

Abstract

A constant of human history is the migration of people in search of a better future. How does the arrival of new immigrants affect the perception of outgroup distance among the native majority group? We provide evidence in the context of the Age of Mass Migration (1860-1920). Applying computational linguistics techniques to a newly processed corpus of over 1.8 million newspaper issues (9 million pages) published by 3,675 local outlets in that period, we present and validate a novel text-based measure of perceived social distance between U.S.-born natives and 21 immigrant groups. Using a shift-share instrument for source-specific migrant inflows, we show that an increase in immigration is associated with higher social distance toward a group, as expressed in local newspapers.

Keywords: immigration, social distance, shift-share instruments, text data

*We would like to thank Susanne Keller and David Jiyang Cai for excellent research assistance. We would like to thank Sekou Keita, Ruben Enikolopov, Didier Ruedin, Sulin Sardoschau, Marco Tabellini, and participants of the NCCR Research Day, Copenhagen University, Bocconi University, Universita' Statale di Milano, LISER, IBEO workshop for helpful feedback. Dominik Hangartner and Gloria Gennaro acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant 804307).

†Ash: ETH Zurich, ashe@ethz.ch; Boltachka: Bocconi, anton.boltachka@phd.unibocconi.it; Gennaro: University College London, g.gennaro@ucl.ac.uk; Hangartner: ETH Zurich, dominik.hangartner@gess.ethz.ch; Stampi-Bombelli: University of Zurich, alessandra.stampi-bombelli@uzh.ch.

1 Introduction

One of the constants of human history is that people leave their place of birth in search for a better future for themselves and their families. Upon arrival in destination countries, new immigrants come in contact with both the host population and already established immigrant groups. A vast literature in economics, political science, social psychology, and sociology explores how native citizens react to immigration. Yet, comparatively little is known about how out-group specific attitudes and behavior evolve over time, and how the arrival of a new immigrants reshapes the perceived distance among native citizens to extant immigrant groups.

This paper addresses this gap by studying an important period in U.S. history, the Age of Mass Migration (1860-1920). In this period, a highly diverse group of 30 million Europeans migrants arrived to the United States. During the second half of the 19th century immigration mainly consisted of Northern and Western European immigrants, primarily of Protestant religious background. The turn of the century, starting in around 1890, saw instead a stark increase in Southern and Eastern European arrivals, perceived to be culturally and religiously more distant than those in the former migration wave. The Emergency Quota Act of 1921 and the National Origins Act of 1924 were introduced in response to the influx of Southern and Eastern Europeans, and abruptly put a halt to their entrance.[1]. With its sharp increase and decrease in arrival numbers for different immigrant groups, the Age of Mass Migration serves as an ideal period to study how native citizens react to new immigrants, and how new arrivals moderate natives' relationships to extant immigrant groups.

To shed light on these questions, we have to tackle two empirical challenges. First, we have to produce a measure of immigrant-group specific social distance, as perceived by U.S.-born citizens, that varies across time and locality. Second, we need exogenous

[1]Anecdotal evidence points to the fact that these policies were intended to alleviate the pressure of "undesirable immigrants" on the lives of American-born natives.

local variation in immigration.

With respect to social distance, the core challenge is the absence of panel data on group-specific attitudes at the county level over time. We address this limitation by proxying for local attitudes using views expressed in local newspapers. Our analysis draws on a corpus of scanned daily newspapers from 1860 to 1920. The Chronicling America archive, developed and maintained by the U.S. Library of Congress, has not previously been used for a comprehensive social-science study employing large-scale text analysis. The corpus comprises 9 million pages from 1.8 million newspaper issues published by 3,675 outlets, linked to detailed metadata on county-level circulation.

We next construct a measure of perceived social distance from the newspaper texts. We train a word-embedding model on the full corpus to capture semantic relationships between words and concepts (Pennington, Socher, and Manning 2014; Rodriguez and Spirling 2022). To ensure robustness to OCR-related misspellings, we use an embedding algorithm that incorporates subword information (Bojanowski et al. 2017). We define a semantic dimension spanning from "American" to "Immigrant," which captures variation in expressed social distance (Kozlowski, Taddy, and Evans 2019; Gennaro and Ash 2022). To obtain time-varying, group-specific measures, we adapt the à la carte embeddings approach of Khodak et al. (2018), allowing us to construct context-sensitive vectors for each mention of an immigrant group. We then compute a mention-level distance measure as the cosine distance between the mention vector and the American–Immigrant dimension, capturing the extent to which a given reference frames a group as more foreign versus more native. This procedure yields social distance scores for 82 million mentions of 21 immigrant groups, which we aggregate to the county–year level using newspaper circulation weights.

We validate the measure in several ways. First, qualitative inspection confirms that the words associated with the native and foreign poles of the dimension align with intuitive notions of social distance. Second, quantitative validation using human coders and large language models shows that the measure ranks immigrant-group mentions intuitively. Finally, consistent with expectations, we show that increases in local social

distance toward a group are associated with lower intermarriage rates and higher residential segregation between natives and that group.

A second, more standard empirical challenge concerns the endogeneity of local immigration flows. An OLS regression of perceived social distance on local immigration—whether cross-sectional or panel—would not identify a causal effect, as immigrants' location choices may respond to local attitudes. In particular, immigrants from a given origin may avoid destinations where sentiment toward their group is more negative, biasing OLS estimates downward. To address this concern, we exploit exogenous variation in immigration inflows using a shift-share instrument (Card 2001), adapted to the Age of Mass Migration (Burchardi, Chaney, and Hassan 2019; Tabellini 2020; Fouka, Mazumder, and Tabellini 2021). The instrument interacts pre-period immigrant settlement shares with contemporaneous national inflows by origin. We document a strong first stage and provide evidence supporting the instrument's validity, including balance on pre-period social distance, newspaper circulation, and other covariates, as well as placebo tests showing no effect on lagged social distance.

Our main result shows that increases in arrivals from a given country raise perceived social distance toward immigrants from that origin. A one–standard-deviation increase in immigration inflows increases perceived distance by about 0.09 standard deviations. This effect is robust across a wide range of specifications, including models with additional fixed effects, covariates interacted with time, alternative constructions of the social distance measure, and different implementations of the shift-share instrument (e.g. Burchardi, Chaney, and Hassan 2019). Examining the temporal dynamics, an event-study analysis reveals an inverted-U pattern: social distance rises gradually following immigrant arrivals, peaks roughly 20 years after arrival, and then attenuates over subsequent decades, with no evidence of pre-trends. We further shed light on the mechanisms underlying these changes by examining how immigrant groups are framed in newspaper coverage. Using a topic model applied to the textual contexts in which immigrant groups are mentioned, we find that higher social distance is associated with less frequent references to American values and economic contributions, and more frequent association with themes of foreignness and

social separation. Finally, we show that immigration reshapes relative perceptions across groups. The arrival of new immigrant groups that are relatively more similar to native-born Americans—along dimensions such as language, religion, or genetic ancestry—leads existing, less similar groups to be perceived as more distant. This pattern indicates that group hierarchies in perceived social distance, while structured by ethnic markers, are not fixed but are endogenously re-ranked in response to changing patterns of immigration.

These results add to a large literature on the interplay between immigration, threat, and prejudice (Blalock 1967; Blumer 1958). Our contribution is twofold and relates to the measurement of attitudes toward immigrants. First, we extend measures of perceived social distance to historical periods for which survey data are unavailable. Second, most research on contemporary settings relies on cross-sectional—and, in a smaller number of cases, panel—survey data that capture citizens' attitudes toward immigration in general (for a review, see e.g. Hainmueller and Hopkins 2014). While some recent work, including conjoint experiments, estimates attitudes toward specific immigrant groups (Hainmueller and Hopkins 2015), such data rarely have a panel dimension (for an exception, see Bansak, Hainmueller, and Hangartner 2023). We address this gap by providing fine-grained, immigrant-group–specific measures of perceived social distance that vary over time and space.

These measurement advances allow us to reconsider a central finding in the literature: that natives' attitudes toward different immigrant groups are structured by stable group hierarchies—often referred to as "ethnic hierarchies"—in which prejudice is stronger toward groups perceived as more distant along ethnic, racial, or cultural–religious dimensions (Hagendoorn 1995). Owing to the absence of repeated, group-specific measures of perceived distance, most existing studies implicitly treat these hierarchies as fixed over time, despite theoretical arguments and initial empirical evidence suggesting that they may evolve (Bobo and Hutchings 1996; Hainmueller and Hangartner 2013). Our group- and time-varying measure of perceived distance shows that hierarchies are not fixed but instead evolve in response to arrivals from both new and existing origin countries.

Despite a large literature on attitudes toward immigration, relatively little research

has examined how immigrant arrivals dynamically shape perceived outgroup distance and generate spillover effects across groups. This new measurement approach allows to extend existing work by exploring how immigrant arrivals dynamically shape perceived outgroup distance, and trace spillover effects on attitudes towards other groups. For the latter, we build on recent works of Fouka, Mazumder, and Tabellini (2021) and Fouka and Tabellini (2021), who document how the arrival of a more distant racial or ethnic outgroup can lead to a re-categorization and reduction in perceived distance of an existing but more similar outgroup. Hangartner et al. (2019) provide evidence for the reverse effect and show that the massive, yet transient, arrival of predominantly Muslim refugees from Syria and Afghanistan increased Greek prejudice against their fellow Greek-Muslim citizens. While these extant works are restricted to identifying the impact of one outgroup on another, our fine-grained distance measure allows us to trace ripple effects across the web of relationships between native citizens and 21 distance immigrant groups.

Methodologically, we contribute to the growing literature that uses text data to measure social attitudes and biases (Ash, Chen, and Ornaghi 2021; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Card et al. 2022; Kozlowski, Taddy, and Evans 2019). In research on media discourse, bag-of-words approaches are often employed to identify relatively direct associations—for example, the co-occurrence of terms such as "immigration" and "crime" within the same sentence or document (e.g., Jacobs et al. 2018). While effective at detecting overt forms of bias, these methods are less well suited to capturing more subtle forms of "othering." Such implicit biases—manifested, for instance, when references to minority groups appear in linguistic contexts that resemble those associated with criminal behavior—can be both more pervasive and more enduring (Thijs, Trilling, and Kroon 2024). Our study advances this literature by using word embeddings to move beyond surface-level co-occurrence patterns and to identify both explicit and latent forms of out-group bias in language, thereby providing a flexible approach for measuring social distance and othering across a wide range of social contexts.

The rest of the paper is organized as follows. Section 2 outlines the method for measuring historical social distance. Section 3 describes our identification strategy using

the shift-share instrument design. Section 4 reports the results. Section 5 concludes.

2 Measuring Historical Social Distance

This section outlines the approach to measuring social distance between U.S.-born and different immigrant groups in unstructured text. We start with a conceptual overview and then describe the data. We outline the method and present multiple validation exercises.

2.1 Conceptual Overview

Self-categorization theory (Turner and Reynolds 2011) provides the motivating framework for our measurement approach. According to the theory, an individual self-categorizes herself based on shared attributes with members of certain social categories, thereby distinguishing herself from others. Individuals will tend to self-categorize into a group when, ceteris paribus, the differences perceived among them are smaller than those between them and individuals in other groups (out-groups). This creates the dichotomy of "us" versus "them" or "in-group" versus "out-group", introducing the notion of distance between two opposite poles. Turner's theory posits that individuals classify themselves and others into in-groups and out-groups based on *shared attributes* and perceived *degrees of closeness*.

Our goal is to operationalize these concepts in the realm of natural language. We identify *shared attributes* through *shared contexts* between the in-group and out-groups, as expressed in text. Specifically, we proxy these attributes by analyzing the language used in newspapers surrounding mentions of different immigrant groups (e.g., the words in the immediate vicinity of the term "Italian"). We then compare these to the attributes associated with mentions of "immigrants" or "Americans", enabling an indirect comparison across different contexts and attributes. For instance, if immigrants are frequently mentioned in the context of crime, a single mention of a specific nationality in a similar context will show high similarity to "immigrant", even if it is not explicitly "othered" in that particular sentence. In fact, in the delicate investigation of attitudes, it is funda-

mental to also capture automatic associations and implicit evaluations of different ethnic groups, in addition to reasoned decisions and explicit opinions. By applying word embedding techniques, which have been compared to well-known psychological methods such as Implicit Association Tests (Caliskan, Bryson, and Narayanan 2017), to the words that co-occur with mentions of out-group minorities, we capture the semantic relationships and contextual nuances that reflect underlying social categorizations and perceived distances to each out-group.

2.2 Natural language as a proxy for attitudes

In an attempt to investigate the relationship between immigration and social distance, we turn to the mentions of immigrants measured in the daily newspaper text. A growing body of scholarship supports the use of media language as an indicator of local attitudes. Slant in media language is a response to consumer demand: news outlets tend to match readers' preferences as part of a market-driven choice (Mullainathan and Shleifer 2005; Gentzkow and Shapiro 2006, 2010). Following Mullainathan and Shleifer (2005), our approach is in line with a model of media consumption which holds on two assumptions. First, we assume that readers are biased and hold biased beliefs based on their education, socio-economic status, former prejudices and stereotypes. Second, we assume that newspapers have the ability to slant reporting of news toward these beliefs in order to cater to the biases of their readers. In a two-state world, readers will prefer to consume news that confirms their beliefs rather than confutes them, as they gain utility from confirmation bias. The main driver of slant is thus the distribution of priors in the population. By analyzing newspaper articles and content, we gain understanding of the distribution of readers' preferences and attitudes.[2]

[2]The approach is also consistent with a rational agent model in which readers have limited time to be informed and choose media with a similar ideological position as to maximize the quantity of decision-relevant information they digest (Chan and Suen 2008). For example, a democratic reader extracts more decision-relevant information from a democratic-leaning newspaper than from a republican-leaning newspaper, holding the amount of time fixed. Furthermore, Gentzkow and Shapiro (2006) conclude that a Bayesian consumer who is faced with uncertainty about the quality of a piece of information will believe

Given the historical period under study, it makes sense to measure local attitudes through written newspapers. With radio emerging only in the 1920s and television in the 1940s, local newspapers served as the primary, if not sole source of news during this time. That makes them a unique and comprehensive medium for understanding public sentiment. U.S. literacy rates in the adult population increased from 80% in the 1870s to over 95% in the 1930s.

2.3 Corpus

Our corpus is composed of all of the daily English-speaking newspaper editions in U.S. localities published between 1860 and 1920 and digitized by the *Chronicling America* project. *Chronicling America* is a public repository produced and maintained by the National Digital Newspaper Program, a collaboration between the National Endowment for the Humanities and the Library of Congress. It is the largest corpus of U.S. historical newspapers, with text from 3,675 newspaper outlets in 47 U.S. states between the years 1777 and 1963.

We obtained the digitized text of the corpus from the *Chronicling America* Bulk Data API (chroniclingamerica.loc.gov/about/api/). The unprocessed corpus comprises 2,079,431 published issues and 19,329,230 pages. Each page comes as a plain-text string constructed from scanned images using optical character recognition. The metadata include the newspaper, date, edition (e.g. morning or afternoon), and page number within that edition. Figure 1 shows example scans with associated metadata.

Although the API and bulk data have been available since 2009, it has only been used in a handful of published studies. Yang, Torget, and Mihalcea (2011) train a topic model on all newspapers from the state of Texas (232,000 pages) for use in a descriptive analysis of the resulting topics. In a similarly descriptive analysis, Viola and Verheul (2020) analyze a topic model trained on 4,810 pages of seven Italian-language newspapers.

that the source of information is of higher quality when it confirms the consumer's prior expectations. The more priors favor a given belief, the less likely a newspaper will print a story contradicting that belief.

Figure 1: *Chronicling America* Historical Newspaper Corpus



Screenshot from *Chronicling America* home page, showing some example images and associated metadata.

Smith, Cordell, and Mullen (2015) describe an effort to detect which fictional works are reprinted in pre-Civil-War newspapers (before 1860).[3] Unlike these papers, we consider the full corpus. Our project is the first that we know of to perform a comprehensive social-science analysis using the universe of newspaper articles from *Chronicling America*.

2.4 Method

This section outlines the methodological approach to measuring social distance between U.S.-born and different immigrant groups in unstructured text. Our method for building a county-decade measure of social distance for each immigrant group consists of roughly four steps. First, we prepare the text data for analysis by applying standard pre-processing steps. Second, we train a word embedding model to obtain a geometric representation of historical newspaper language. Third, we obtain vector representations for each mention of immigrant groups in the corpus. Fourth, we scale those mentions

[3]Lee et al. (2020) use computer-vision methods to extract and explore visual features of *Chronicling America*.

against a semantic dimension going from immigrant to native.

Document Pre-Processing. The corpus pre-processing can be summarized as follows. First, given the relatively high rate of OCR error in the corpus (and the presence of some non-English content), we filter out pages that have more than 50% of words that do not appear in the English-language WordNet dictionary.[4] We also filter out any short words (up to three letters) that are not part of WordNet. We then remove punctuation, capitalization, digits, and stopwords. Tokens are then stemmed using the Porter stemmer. Finally, we restrict our vocabulary to include the 100,000 most frequent stemmed-token types. The pre-processing is described in greater detail in Appendix Section B.3. The final corpus consists of 9,476,034 pages.

OCR-Error-Robust Word Embeddings. We use the pre-processed corpus as input to train a word embedding model. Word embeddings are a tool from natural language processing for learning numerical representations of words based on co-occurrence statistics in a given corpus (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). Thus we learn a vector $\boldsymbol{w}$ corresponding to each word w in the vocabulary based on how words co-occur in the newspapers.

The most commonly used word-embedding algorithms are Word2Vec (Mikolov et al. 2013) and GloVe (Pennington, Socher, and Manning 2014). Notably, these standard algorithms do not use the information contained in the letters of a word, so for example the algorithm would not learn any relationship between "immigrant" and "migrant" based on the letters in the words. For large, clean corpora this is not a problem and the relationship between "immigrant" and "migrant" is learned properly from co-occurrences. In our context, though, it is a downside due to the high frequency of OCR errors.[5] We frequently observe digital corruptions like "inmigrant" and "imm1grant",

[4]The OCR error is similarly distributed across decades, even though documents with extremely high error become less frequent over time. Desity plots by decade are reported in Appendix Figure A.1.

[5]Smith, Cordell, and Mullen (2015) undertake a pilot study checking the OCR error rates in 35 poems and stories in *Chronicling America*. They find a character error rate of between 5% and 15%, leading to a word error rate over 25%.

which Word2Vec or GloVe would ignore as being out-of-vocabulary. Given we need to extract embedding information for small subsets of the corpus (county and year), we would like to produce embeddings that are robust to a high OCR error rate.

To address the OCR error issue, we use the FastText word-embedding algorithm from Bojanowski et al. (2017). That algorithm enriches word vectors with the sub-word information contained in the constituent letters. Formally, words are represented as the sum of embeddings of the set of three-letter subsequences in the word. Compared to other word embedding algorithms, FastText is robust to misspellings from OCR-errors. Appendix Tables A.1 and A.2 demonstrates the better performance of FastText embeddings for misspelled words.

To train the embeddings, we input all pre-processed pages from each published newspaper issue over the time period 1860 through 1920. We use the python gensim implementation with 256-dimensional vectors, an eight-word context window, a vocabulary size of 100,000, and training for 5 epochs. Rodriguez and Spirling (2022) show that different choices for these hyperparameters do not significantly affect the outputs of word embedding models.

Mention-level Embeddings for Immigrant Groups. We are interested in understanding how portrayals of the same immigrant groups change over time and across U.S. counties. Our trained FastText models will not work for this purpose because it encodes associations at the level of the whole corpus. Allowing embeddings to vary across units is an area of active research. For example, Rodman (2020) takes a global model and fine-tunes an additional model for each time period. Ash, Chen, and Ornaghi (2021) train separate models for each judge in a corpus of judicial opinions. Adding additional structure to the embedding network, for example to allow the vectors to depend on group and time categories, is promising in theory but despite many attempts (e.g. Bamman, Dyer, and Smith 2014; Rudolph et al. 2017; Rudolph and Blei 2018; Gillani and Levy 2019), the literature has not produced a robust method that works well in practice across domains. In our context, training separate models for each newspaper and time period would not work because of the small corpus and the computational burden.

To resolve this issue, we apply a recently developed method called *À La Carte* (ALC) embedding (Khodak et al. 2018). This method allows us to calculate context-dependent embeddings for each individual *mention* of an immigrant group in the corpus. In summary, ALC embedding works by taking an average of the embeddings of the words that appear in a word's vicinity in any given occurrence. Then, a learned matrix is used to transform the average of the neighboring word embeddings to a predicted vector for the target word. In this case, the final embedding is mention-specific, and encodes semantic content that can be inferred by the specific context in which it appears.

We focus on measuring social distance between U.S.-born and major immigrant groups of European origins. For this purpose, we create lexicons of nationality markers. For each immigrant nationality, we start by compiling lists of synonyms. We also include the most frequent misspellings of nationalities. The complete list of nationality lexicons are reported in Appendix Table B.1. Using the full set of lexicons, we iterate through the corpus to obtain 82,190,688 mention-specific embeddings, corresponding to all detected instances of immigrant-group mentions.

Scaling Social Distance. Word embeddings encode semantic dimensions in language as geometric dimensions in a vector space. Our goal is to construct a dimension in this space corresponding to the typical portrayal of immigrants at one pole and U.S.-born natives on the other. The resulting social-distance measure toward different immigrant groups can then proxy for variation in local attitudes. Related work taking similar approaches includes Kozlowski, Taddy, and Evans (2019), Ash, Chen, and Ornaghi (2021), and Gennaro and Ash (2022).

To construct the social-distance dimension, we start with word lists for an immigrant pole and a native pole. These words are listed in Appendix Table B.2. For example, immigrant includes "immigrant" and "migrant", while native includes "us-born" and "american".

For each word in each lexicon, we have an embedding vector. Define the native dimension ($\vec{N}$) as the mention-weighted average of the vectors from the native lexicon, with the immigrant dimension ($\vec{I}$) analogously defined as the mention-weighted average

of the vectors of words in the immigrant lexicon. We construct the native-immigrant dimension ($\vec{NI}$) as the vector difference, $\vec{NI} = \vec{I} - \vec{N}$.

We calculate social distance s_{nip} for mention i of immigrant group n in newspaper p as the cosine similarity between each mention vector $\vec{m}_{nip}$ and the immigrant-native dimension $\overrightarrow{NI}$: $s_{nip} = \text{sim}(\overrightarrow{NI}, \overrightarrow{m}_{nip})$. At the end of this process, we have a scalar similarity score s_{nip} for each mention of immigrant-groups in the newspaper corpus. Appendix Table A.3 reports examples of mentions from those that obtained the highest and lowest social distance scores in our corpus.

Social-Distance Outcome. For each immigrant group, we obtain aggregate measures of social distance at the county-decade level, by averaging the mention-level social distance scores for newspapers that circulate in that county-year. We compute:

$$S_{nct} = \frac{1}{N_{nct}} \sum s_{nip} \times \mathbb{1}[p \in C_{ct}] \tag{1}$$

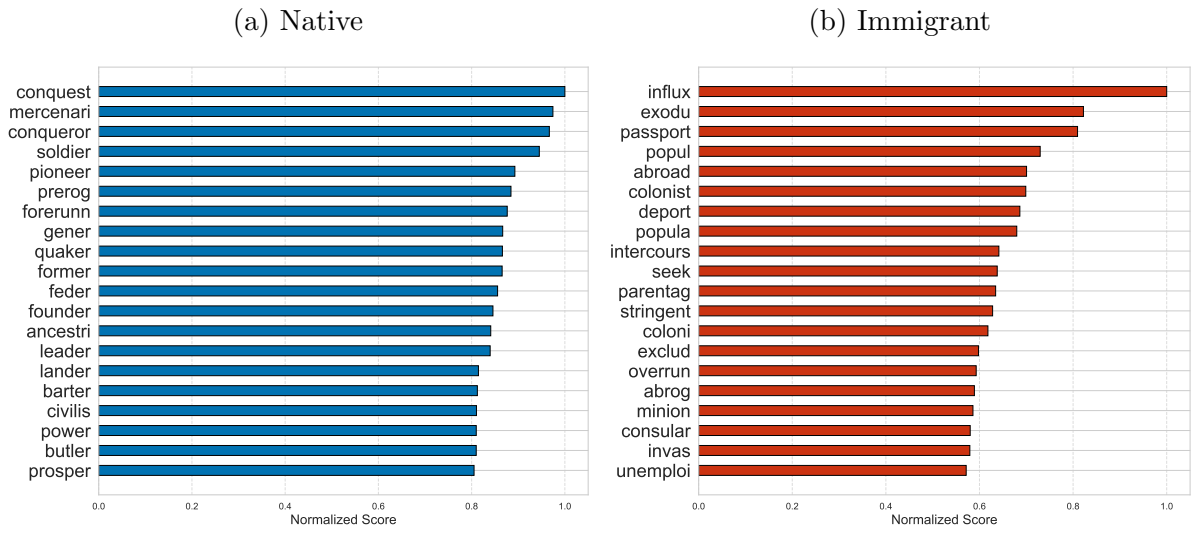
where S_{nct} is social distance for immigrant group n in county c, decade t; s is social distance in mention i in newspaper p; $\mathbb{1}[p \in C_{ct}]$ is an indicator function equal to 1 if newspaper p is among the set of newspapers C that are circulated in county c in decade t; N_{nct} is the total number of mentions for immigrant group n in the given county-decade.

2.5 Validation

In this section, we describe the steps taken to validate our measure of social distance across qualitative and quantitative dimensions. These evaluations include a combination of manual inspection, human validation, and validation against external measures of integration. Through these exercises, we assess the reliability of our measure in capturing meaningful distinctions between native and immigrant portrayals over time and across different news outlets.

Qualitative evaluations. First, we inspect the words that are closest to the native ($\vec{N}$) and immigrant ($\vec{I}$) average vectors in the word embedding space of both corpora,

Figure 2: Semantic Poles for Native and Immigrant in Newspaper Language



Top 20 vocabulary words that are closest to the respective "poles" of the dimension in the embedding space corresponding to native/American (a) and immigrant/foreign (b). The horizonal axis represents the normalized social distance score. Vocabulary words that are included in or contain words from the Immigrant, Native or Nationality Lexicons, as well as some mispellings are dropped. Size indicates proximity (higher cosine similarity) to the respective word-vector centroid.

separately. The closest words to both poles, as observed in this initial validation test, are terms that share a similar contextual meaning to either pole, virtually functioning as near-synonyms for "native" or "immigrant." By extension, in our model, when we compare each mention of an immigrant group to either pole, we indirectly assess their proximity to these terms. The word lists are constructed after excluding tokens that appear in the immigrant and native lexicons, as well as in nationality lexicons and some mispellings.

Figures 2a and 2b plot the top 20 words for the newspaper corpus, after removing the lexicon terms and variants. The native document vector is surrounded by words that refer to being American or native-born, while the immigrant document vector appears close to words related to immigration flows. Intuitively, the native pole exhibits concepts like "pioneer", "founder" and "civilise", while the immigration pole exhibits concepts like "influx" and "deport".

Second, we visually explore random samples of sentences that include a mention of an

immigrant nationality. These samples are extracted from the 5% most extreme contexts on the *Immigrant-Native* dimension. Some of those sentences are reported in Appendix Table A.3. Contexts with higher social distance (Immigrant contexts) refer to members of the immigrant group as homogeneous others – cultural manifestations, group-based labour competition, characterisations of ethnic groups. Contexts with lower social distance (Native contexts) refer to members of the immigrant group mostly as individuals, and cultural manifestations are depicted as desirable from the American point of view.

Third, we rank historical group stereotypes according to the social distance measure and ask whether our method ranks groups intuitively. Katz and Braly (1933) report the top 15 stereotypes attached to selected immigrant groups, including four salient groups in our dataset: Italian, Irish, German, and British. For instance, according to the 100 students surveyed by Katz and Braly (1933), Italian immigrants are associated with "Artistic" and "Impulsive," while Irish immigrants are associated with "Pugnacious" and "Quick-tempered." For each immigrant group, we calculate a stereotype vector as the weighted average of individual word embedding vectors for each adjective. Specifically, the word embedding for each adjective is weighted by the percentage of respondents associating that adjective with a given nationality and then averaged. This allows us to calculate the cosine similarity between each stereotype vector and the immigrant-native dimension to derive a social distance score for each group. We inspect the ranking of those immigrant groups along the social distance measure, obtained by their stereotyped attributes, and include other groups examined in Katz and Braly (1933), such as African-Americans, Jews, and Chinese. Appendix Table A.4 shows that the social distance measure captures meaningful variation across stereotyped images of these groups. While white Americans and British immigrants are closest to the *Native* dimension, we find that Irish, Japanese, and Chinese immigrants obtain the highest levels of social distance, in line with historical accounts (Aarim-Heriot 2003).

Human validation. This subsection reports the results of a human annotation task to assess the validity of our score in capturing social distance. Coders were provided with sets of text snippets extracted from the corpus, which include reference to a foreign nationality.

The specific nationality attribute is masked to the coders' view, and indicated as Group X. For each set, they are asked to pick the snippet that portrays the foreign group more as an out-group, rather than an in-group. In doing that, we provide instructions along those lines "Out-group framing suggests [Group X] is perceived as outsiders, strangers, foreigners, or immigrants. In-group framing suggests [Group X] is perceived as insiders, friends, family members, neighbors, or native-born individuals". Coders know that the snippets come from newspaper articles from the late 1800s and early 1900s, and that Group X is a nationality attribute or social group. No other information is provided about where the snippets come from.

In the baseline validation check, we construct sets of four snippets randomly extracted among the 5% with highest and lowest values in our social distance score. The annotators are fluent in English. If the measure was uninformative, we should observe agreement between coders and our measure around 25% of the time, which correspond to the probability of picking the right snippet at random, absent any signal in the measure. In a sample of 896 snippets, grouped in 224 sets, human coders agree with our measure 60.71% of the time.

We extend this validation with two further exercises, assisted by a large language model (specifically GPT-4o) as an alternative coder. First, we investigate whether accuracy varies by topic. To do that, we apply an unsupervised topic model (Latent Dirichlet Allocation, Blei (2012)) on all snippets containing group references extracted from the corpus, assuming 45 topics.[6] The topics are then manually reviewed and assigned to larger categories, for a total of 16 topic families. Next, we extract a random sample of 504 snippets stratified by topic family, with equal representation, and form random pairs composed of snippets from the same family. GPT-4o-coding agrees with our measure 68.4% of the time on average.[7] However, results reveal large variation in the accuracy by

[6]LDA works by assuming a structural model for language, where documents are distributions over topics and topics are distributions over words. The parameters of these distributions are learned from the corpus and then produce interpretable topics.

[7]Accuracy increases to 69.2% in the same sample when asking GPT to provide motivation. It further increases to 71.9% when allowing GPT to answer "Neither" as a third option.

topic. Intuitively, the least accurate predictions concern topics that do not involve social groups, such as the weather (44%) and shipping schedules (49%). In those cases, our selection algorithm is likely to pick up nationality markers that are not related to social groups (e.g. *Italian* immigrants) but rather to places (e.g. *Italian* harbour) or events (e.g. winds on the *Italian* peninsula). On the other hand, topics concerning business transactions and local community matters obtain the highest accuracy score (83% and 79% respectively). Full results are reported in Appendix Table A.5.

Second, we explore accuracy across the whole distribution of the social distance measure. In particular, we extract 994 snippets completely at random and, based on them, created 497 random snippet pairs. Using the same prompt as with humans, we asked GPT4o to pick the snippet that portrays the group X as outsider. Excluding 40 cases classified as "unclear", GPT4o agrees with our measure 60.6% of the time on average. However, we also find that accuracy is higher when the snippets differ more in their social distance score. While accuracy is only 47% for snippets that are closest to each other (first quintile of the difference in score), it reaches 67% for the most distant snippets (fifth quintile of the difference in score). Full results are reported in Appendix Figure A.2.

Correlation with social integration measures. Next, we show that our social distance measure correlates with known proxies of a related construct – that is, social integration. In particular, we correlate our measure with residential segregation and intermarriage rate. From Eriksson and Ward (2019) we have a measure of residential segregation calculated for each immigrant group with respect to US-born residents. Using the census, we calculate county-decade intermarriage rates for each immigrant group as the group share that is married to a U.S.-born resident over the total number of immigrants in that group. We expect social distance to be positively correlated with residential segregation: that is, the higher the residential segregation of that ethnic group, the more distant they portrayed in language. Similarly, we expect intermarriage rate to be negatively correlated with social distance, such that higher distance is associated to lower intermarriage.

Table 1 reports the results of this validation. We regress social distance for a given immigrant group, in a given county and decade, on residential segregation (columns 1-4)

Table 1: Residential Segregation and Intermarriage

DV: Social Distance	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Residential Segregation	0.830***	0.823***	0.844***	0.201***				
	(0.078)	(0.070)	(0.068)	(0.066)				
Intermarriage					-0.306***	-0.307***	-0.315***	-0.084***
					(0.028)	(0.028)	(0.028)	(0.025)
Observations	14105	14105	14057	14105	21926	21926	21912	21926
R-squared	0.009	0.073	0.233	0.256	0.008	0.060	0.199	0.219
Decade-State FE		✓		✓		✓		✓
Decade-County FE			✓				✓	
Nationality				✓				✓

Notes. Columns 1 to 3 show the regression of group-specific social distance at the county-decade level, on residential segregation for the same group (Eriksson and Ward 2019). Columns 4 to 6 show the regression of group-specific social distance at the county-decade level, on intermarriage rate between group members and American citizens. Columns 2 and 5 include decade-state fixed effects. Columns 3 and 6 include decade-county fixed effects. Intermarriage rate is computed as the share of immigrants from nationality n who married native-born Americans, over the total number of immigrants n, in a county-decade. Residential segregation ranges from 0 to 1 with higher values indicated higher residential segregation. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

and intermarriage rate (columns 5-8) for the same group, county, and decade. The results are consistent with expectations, suggesting that social distance is correlated with more residential segregation and a lower intermarriage rate.

3 Empirical Approach

This section outlines our empirical approach. We use a shift-share instrument to identify exogenous variation in local immigration across different source countries. This section describes the instrument and reports the validity checks.

3.1 Data

Newspaper Circulation Data. The *Chronicling America* collection contains some useful metadata for our purposes. We extract geographic coverage of each newspaper at the county level. This allows us to match each newspaper to a set of county-year observations where it has been distributed. Overall, the dataset covers 924 counties in 47 states. Appendix Table A.23 shows that those counties in 1860 are more populated and less dependent on agriculture than non included counties. However, Appendix Figure A.7 shows that almost all States appear in the dataset since the 1860 already, while the

number of unique newspapers per State increases over time.

Immigration Data. We draw immigration statistics for the 21 largest European immigrant groups from the U.S. Census between 1870 and 1920, corresponding to five Census waves (the Census was recorded every 10 years with the exception of 1890). For each census year, we collect the number of residents from each origin country at the county level. For each immigrant group, we further use immigration flows from the 1860 U.S. Census and national-level immigration inflows for each decade to build our shift-share instrument.

Other Census Statistics. From the U.S. census we obtain a set of socio-economic variables. We use the 1860 census to draw information on the demographic composition and economic performance of each county before our study period.

3.2 Estimation Framework

Immigration In-Flows. The main treatment variable we are interested in is the in-migration of various groups by county of origin, into a local county. We define M_{nct} as the change in the number of residents from immigrant group n between decade $t-1$ and t. Due to the skewed distribution of the inflow variable, we apply the inverse hyperbolic sine transformation.

Empirical Model. Our research design is based on a panel dataset of U.S. counties in the five census years from 1870 to 1920.[8] In particular, we examine how social distance with specific immigrant groups responds to changes in exposure to that same immigrant group. We estimate the following regression model:

$$S_{nct} = \alpha + \beta M_{nct} + \Gamma \vec{X}_{nct} + \rho_{nct} + \nu_{nct} \tag{2}$$

[8]The 1890 census went missing in a fire and is hence unavailable to researchers.

where S_{nct} is social distance for immigrant group n, in county c and decade t; M_{nct} is the migration in-flow for group n, in county c between decade $t-1$ and t; $\vec{X}$ is a vector of time-varying county controls; ρ_{nct} is a vector of fixed effects for county and decade, as well as interactions among those. In our main specification, we control for decade-by-county fixed effects.

Controlling for decade-by-county fixed effects implies that β is estimated from variation in the magnitude of in-flows across different immigrant groups in the same county and decade. Time-varying controls include pre-treatment county characteristics, interacted with time trends. To ease interpretation, we standardize both the dependent and the main independent variables by subtracting the mean and dividing by the standard deviation. Standard errors are clustered at the county level.

Instrumental Variables Approach. While the fixed effects can help adjust for many possible confounding factors, there are still issues of selection due to time-varying local omitted variables. Immigrants do not randomly choose destination locations, but are attracted to areas with, e.g., more favorable labor conditions and higher integration prospects. In particular, immigrants could self-select into American counties which are already relatively more welcoming to those groups. In our empirical setting, that would entail reverse causality from the outcome variable (social attitudes, as proxied from newspaper language) to the treatment variable (migration in-flows). Because a lower measured social distance would increase migration in-flows, the OLS estimate from Eq. (2) is likely to be biased downward.

To overcome these selection problems, we identify the causal effect of increased exposure to an immigrant group using an instrumental variable specification. Specifically, we predict immigrants' location decision using a shift-share instrument commonly adopted in the immigration literature (Card 2001). The shift-share instrument is formed by interacting past local immigrant shares with aggregate national shifts in immigration. This strategy was first introduced by Bartik (1991) to identify growth shocks, whereas Card (2001) popularized it in the immigration literature by exploiting the empirical regularity that near immigrants tend to settle near earlier immigrants of the same country of origin

(due to family ties and social networks). Tabellini (2020) applies this method in the context of the Age of Mass Migration.

The shift-share instrument is constructed by interacting lagged local shares of immigrant groups from each country of origin with the current national inflow of immigrants from that given country of origin. Formally, the instrument for the in-flow of immigrant group n in county c at census year t as:

$$z_{nct} = \frac{M_{n,c,1860}}{M_{n,1860}}(M_{n,t} - M_{n,t-1}) \tag{3}$$

Where $M_{n,c,1860}$ is the number of immigrants from group n in county c in 1860, $M_{n,1860}$ is the total number of immigrants from the same group in the U.S. in 1860, $M_{n,t} - M_{n,t-1}$ is the change in the total number of immigrants to the U.S. from group n between census year t and $t-1$.

The first stage estimating equation is

$$M_{nct} = \alpha + \gamma z_{nct} + \Gamma \vec{X}_{nct} + \rho_{nct} + \epsilon_{nct} \tag{4}$$

where the other terms are as above. Causal estimates for β are produced by instrumenting migration in-flows in Equation (2) using the first-stage prediction from Equation (4).

3.3 Instrument Validity

2SLS requires relevance in the first stage. Table 2 reports the first-stage relation between the shift-share instrument z_{nct} and migration inflow M_{nct} for the same immigration group. The relationship is positive and statistically significant. A one-standard-deviation increase in the magnitude of the instruments corresponds to 0.3 standard deviation increase in the main independent variable. In Table 3, we report Kleinbergen-Paap cluster-robust first-stage F-statistics and they are consistently above 300, indicating a well-powered first stage.

Beyond relevance, 2SLS imposes three requirements for consistent estimates. First, monotonicity implies that the instrument influences immigration inflows in the same di-

Table 2: First Stage Estimates

DV: Immigrant Inflow	(1)	(2)	(3)	(4)	(5)	(6)
Migration Instrument	0.352***	0.346***	0.346***	0.347***	0.382***	0.381***
	(0.022)	(0.020)	(0.020)	(0.020)	(0.021)	(0.021)
Observations	32424	32424	32424	32424	32424	32424
R-squared	0.261	0.276	0.278	0.280	0.305	0.306
Decade FE	✓	✓	✓	✓		
County FE	✓	✓	✓	✓		
Urban		✓	✓	✓		
Economy			✓	✓		
Demographics				✓		
Decade-County FE					✓	✓
Newspapers						✓

Notes. Each column shows the regression of *Immigrant inflow* M_{nct} in county c, decade d, for nationality n, on the instrument z_{nct}. All columns include decade-county FE. Columns 2 to 6 include also nationality fixed effects. Columns 3 to 6 progressively include urban, demographic, and economic controls measured in 1860 and interacted with time trends, and newspaper-level controls. *Urban* indicates controls for size of place, urban population, and city population; *Demographics* indicates controls for percent of white population, African-American population, Chinese population and the average number of couples in a household (proxy for marriage rate); *Economy* indicates controls for percent in labor force, percent in manufacturing, percent literate, real estate value, asset value, percent in poverty, percent crime. *Newspapers* indicates controls for the number of mentions, the average page number on which the mentions are found, and the average share of OCR error in newspapers. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

rection for all counties. Monotonicity is likely to hold in this context, as the instrument is unlikely to reduce migration flows. Second, exogeneity requires that the instrument and county-level social distance are not endogenously determined. Third, exclusion requires that the instrument affects social distance only through its effect on contemporary immigration inflows. Recent work has highlighted how the credibility of shift-share instruments with respect to exogeneity and exclusion restriction can rest on one of the two components individually, rather than both.

In particular, Goldsmith-Pinkham, Sorkin, and Swift (2020) show that exogeneity of the initial shares, conditional on covariates, is a sufficient condition to determine the exogeneity of the instrument. In this approach, the identifying assumption is that the differential effect of higher exposure (i.e. share) of one nationality (compared to another) only affects the change in the outcome through the change in future migration inflows, and not through any other potential channel. The exclusion restriction is violated if, for example, areas with high versus low nationality-specific shares may have other features

that predict change in the nationality-specific social distance through channels other than the endogenous variable. In our setting, county-specific characteristics that influenced the decision of early migrants may have affected the evolution of social distance in subsequent decades. To support the exogeneity assumption, Appendix Table A.9 shows that the instrument is mostly balanced across pre-period county characteristics. While two out of fifteen pre-period characteristics are positively correlated with our instrumental variable, this could be due to chance. We include all those characteristics in our main regression, interacted with time dummies, with little change to the magnitude of the coefficient. We interpret the stability of the coefficient as a sign in support of a limited role for omitted factors (like in Altonji, Elder, and Taber 2005). Appendix Table A.10 further shows the breakdown of those results by nationality, confirming that minor imbalances are not due to any specific nationality. We further show, in Appendix Table A.16, that the 1860 county-level shares for different immigrant groups do not predict, alone, future levels of social distance for those groups. The small and non-significant coefficients associated with the shares suggest that possible predictors of those shares are not also affecting future social distance, and provides additional support to the assumption that the instrument (and its components) are only related to the outcome through their effect on contemporary migration shares.

On the other hand, Borusyak, Hull, and Jaravel (2022) show that the shift-share instrument is valid also under two sufficient assumptions, that guarantee exogeneity of the shocks (even when the shares cannot be considered exogenous). Shocks need to be as-good-as-randomly assigned, i.e. uncorrelated with the relevant unobservables in expectation, and need to be many and sufficiently independent, i.e. each with sufficiently small average exposure. We provide evidence in support of those assumptions in three main ways. First, in Appendix Table A.12, we report the results obtained with an alternative version of the instrument, where we subtract from the national level immigration inflows those immigrants that locate in the specific county. If aggregate inflows from specific nationalities were generated by localized pull factors in specific localities, then the shocks could be endogenously related to the social distance outcome. This may happen, for ex-

ample, if migrants carry nationality-specific skill sets. Second, we expand this robustness check by subtracting immigrants who settle in the state (Appendix Table A.13) or region (A.14) where the specific county is located.[9] This adjustment accounts for potential pull factors that might originate from areas that are larger than a single county, thereby isolating the national-level shock more effectively. Third, Appendix Figure A.6 reports main results obtained after dropping one nationality at a time. The results are not driven by any particular nationality, providing support to the fact that shocks have small average exposure. Similarly, we show that the results are not due to over-time correlation in shocks, by estimating a placebo specification, where we regress social distance measured in decade $t-1$ on immigrant inflow in decade t. Appendix Table A.11 shows that the estimated coefficient is nearly zero in all specifications and far from conventional levels of statistical significance. This result provides reassurance that the immigration push identified by our instrument is not driven by past levels of perceived social distance.

Despite all the robustness tests, the exclusion restriction remains an untestable assumption with direct means, as there could be other unobservable ways that the forces captured by the shift-share instrument influence newspaper articles. This is an important caveat for our (as any) 2SLS estimates. However, our reduced-form estimates, reported in Appendix Table A.8, have a causal interpretation under exogeneity even if exclusion fails. To capture the direction of the causal effect, the reduced-form estimates are sufficient.

4 Results

This section reports the results of the empirical analysis on how immigration influences expressed social distance in newspapers. First, we report the main results with robustness checks. Second, we unpack the framing of the results by analyzing topics. Third, we look at spillovers between incumbent and arriving ethnic groups.

[9] We use the 9 official U.S. divisions: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific.

4.1 Main Results

This section reports the main results with robustness checks.

2SLS Estimates. Table 3 reports the main results. Column 1 reports the estimates from an ordinary-least-squares regression of social distance on same-group immigration inflow over the previous decade. Columns 2 to 7 report the two-stage-least-squares estimates of the effect of higher immigration inflow on same-group social distance, where immigration inflow is instrumented with the shift-share instrument. All columns include decade and county fixed effects. Columns 3 to 5 include pre-period county variables for urbanization, demographic characteristics, and economic development, interacted with time trends. Column 6 includes county-decade fixed effects, and hence the coefficient is estimated from variation in the magnitude of immigration inflows across different countries of origin, within each county and decade. Column 7 includes as covariates some relevant newspaper text features – the total number of a group's mentions in a county-decade, and the average page number of those mentions in the county decade, and average OCR error rate. Note that these covariates are endogenous controls in the sense that the instrument likely affects them directly.

In all columns, the estimated treatment effects are positive and statistically significant. The OLS correlation between immigration inflows and social distance reported in column 1 is positive and statistically significant: a one standard deviation larger immigration inflow is associated with a 0.054 standard deviation higher social distance. The 2SLS estimates are similar to the OLS estimate and stable in magnitude, and they range from 0.09 in columns 2 to 0.093 in column 7. In particular, estimates do not change significantly when including county-level controls, suggesting that county-level characteristics that may have attracted early immigrants are not likely confounders of the effect of immigration on social distance. The estimates are unchanged when introducing other mention-level characteristics (Column 7), suggesting that the effect of immigration on newspaper reporting about specific immigrant groups is not driven by the magnitude of coverage dedicated to specific groups, nor by location within outlets. The magnitude

Table 3: Effect of Immigration on Social Distance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.054***	0.090***	0.096***	0.096***	0.094***	0.092***	0.093***
	(0.006)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Observations	32424	32424	32424	32424	32424	32424	32424
F-statistic		357.209	435.510	428.905	428.144	455.899	455.645
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Column 1 reports the OLS estimates; columns 2 to 7 report the 2-stage-least-squares estimate. Columns 1 to 5 include decade and county FE. Columns 6 and 7 include county-decade fixed effects. Controls are measured in 1860 and interacted with decade dummies. *Urban* includes the share of urban population, and county population; *Economy* includes share in labor force, in manufacturing, agriculture, and mining, average real estate value, asset value, share on welfare/charity, share with criminal records. *Demographics* indicates controls for percent of white population, African-American population, share of literate and in school, average number of couples per household; *Newspapers* indicates controls for the number of mentions, the average page number on which the mentions are found, and the average share of OCR error in newspapers. Standard errors are clustered by county. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

of the effect is economically meaningful: when immigration increases by one standard deviation, social distance between the same group and U.S. born citizens increases by 5–10% of a standard deviation.

Robustness Checks on Specification and Measurement. Results are robust to a wide range of robustness checks on the specification and measurement of the main variables. First, we test the robustness of our results against different specifications of the immigration inflow variable. Appendix Table A.18 reproduces the main results with the raw independent variable without inverse hyperbolic sine transformation. Next, we specify immigration inflows as changes in the share of a given immigrant group over total county population. Appendix Table A.20 shows that all results are confirmed. Appendix Table A.21 reports instead the regression of social distance for a given immigrant group in a county-decade on the share of that immigrant group on the 1860 county population. Results increase in magnitude and are stable across specifications. Appendix Table A.17

reproduces the main results after excluding nationality-county-decade observations with negative or 0 inflow counts, hence exploiting differences in the magnitude among positive immigration shocks, rather than in the presence or absence of inflows.

Most notably, we provide alternative estimates that gauge a different source of exogenous variation in migrants' settlement decisions. Building on Burchardi, Chaney, and Hassan (2019) and Bursztyn et al. (2024), we isolate variation in the pull and push factors that is plausibly independent of local determinants of social distance towards a specific group. The intuition behind this instrument is that, upon their arrival, migrants are more likely to locate in counties that are more attractive at that specific time. Attractiveness (the pull factor) is measured as the fraction of migrants coming from anywhere in the world who settle in a specific county at a specific time. To ensure that this measure does not capture nationality-specific skill set, we exclude migrants from the same region as the nationality under consideration.[10]. The push factor is measured as the total number of migrants arriving in the U.S. from a given country of origin at a give time. Again, we exclude migrants who settle in the same U.S. region as the county under consideration, to ensure that the national level aggregate captures country-push factors rather than the attractiveness of some U.S region.[11] Results are reported in Appendix Table A.15, alongside more details on the instrument. In this case, the F-statistics of the first stage ranges between 18.1 and 20.6, indicating a weaker instrument compared to the traditional shift-share, yet sufficiently strong to deliver consistent estimates. Despite the difference in the first stage, the magnitude and statistical significance of the treatment effects is virtually unchanged compared to the main results.

[10]European migrants are divided in four regions. West Europe includes Austria, Belgium, France, Germany, the Netherlands, and Switzerland; East Europe includes Bulgaria, Czechoslovakia, Hungary, Russia, Poland; North Europe includes Denmark, England, Ireland, Norway, Scotland, Sweden, and Wales; South Europe includes Greece, Italy, Portugal and Spain.

[11]In their paper, Burchardi, Chaney, and Hassan (2019) and Bursztyn et al. (2024) also include higher-order interactions of these pull and push factors, to model the serial correlation of those factors over time. In their setting, this allows instrumentation of modern day ancestry composition. In our setting, our goal is to instrument the next decade migration inflow in a county. Hence, we drop this component from our instrument.

Next, we test the robustness of the main results with respect to more restrictive measures of social distance. In Appendix Table A.22, we report the regression results after dropping ads,[12] articles on international affairs, or both. We also present results after controlling for positive and negative sentiment and topic fixed effects. The stability of the results suggest that the social distance measure captures more than just those alternative linguistic dimensions.

The results are also robust to alternative clustering of standard errors. Appendix table A.19 shows that all specifications hold when clustering standard errors at the nationality by decade level. That allows for correlations across locations that experience similar shocks.[13]

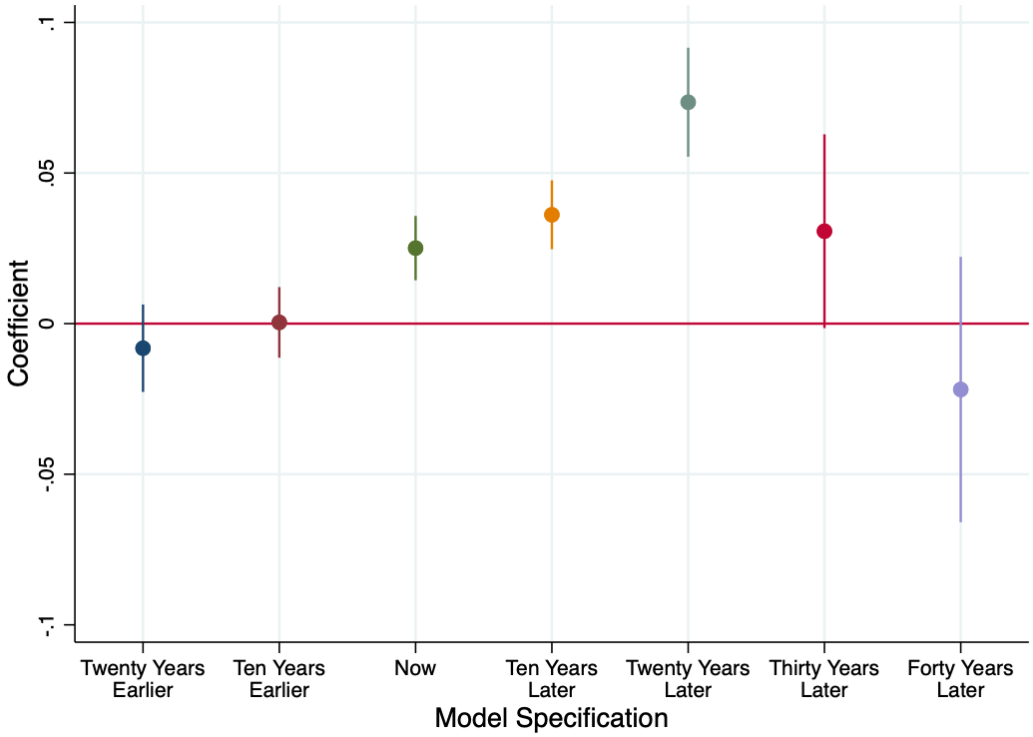
Dynamics over time. Figure 3 illustrates the temporal dynamics of social distance in response to immigration inflows. The figure plots the coefficients from reduced-form regressions of social distance on both the lead and lagged values of the instrument for immigration shocks. The results reveal a distinct inverted-U pattern in the relationship. Social distance initially increases following immigrant arrivals, with the use of increasingly "othering" language to describe immigrants peaking approximately 20 years after their arrival. This tipping point marks the height of perceived social distance.

Over time, the effect diminishes, with the coefficient becoming statistically indistinguishable from zero around 30 years post-arrival. By 40 years, the social distance measure turns negative (yet statistically insignificant), indicating a shift in the portrayal of immi-

[12]The main approach to identifying ads in our dataset is to look for content that is repeated across different issues of the same newspaper, considering potential OCR errors. Content that appears multiple times is classified as ads. Details are reported in the Appendix Section C.12.

[13]We further test the effect of immigration inflows on other outcomes related to the mention of immigrant groups in the newspaper corpus. Appendix Table A.25 reports the effect of immigration inflows on the aggregate number of mentions for a given nationality in newspapers that circulate in any county and decade. While we find that immigration inflows are positively correlated with the number of mentions of a given immigrant group, the causal effect of immigration on the number of mentions is negative. Further, Appendix Table A.25 reports the effect of immigration inflows on the average page number at which those mentions appear. We find that immigration inflows increases the average page number where a given immigrant group is mentioned.

Figure 3: Dynamics of social distance over time



Notes: Coefficients represent coefficients of seperate regression of social distance on leads or lags of the instrument variable. Vertical lines represent 95% confidence intervals. In all regressions, we control for the lagged value of the dependent variable, as well as the log of the lagged county population, and include decade and county fixed effects as well as county interacted with time trends. Standard errors are clustered at the county-level.

grant groups toward greater inclusion. At this stage, immigrants are no longer framed distinctly as "immigrant" or "native," but instead, if anything, as more "native." This temporal progression highlights the dynamic nature of social attitudes, reflecting both initial resistance and eventual integration over a generational timespan.

4.2 Social distance across different topics

What does it mean to become "more immigrant" or "more native"? Do newspapers report on immigrant nationalities in the context of different topics? Or do they change their slant to discuss them in the same contexts, yet with different language? In this section, we address this question by looking at topics. To do that, we use the unsupervised topic model, LDA (Blei 2012), described above. Appendix Section B.8 reports further

Table 4: Effect of Immigration on Prevalence of and Distance by Local Topics

A. Effect of In-Flows on Immigrant Group's Topic Prevalence

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DV: Topic prevalence for:	Crime	Church Services	Diseases	Medicines	Schools	Job Ads	Local News	Personal Stories (I)	Personal Stories (II)
Immigrant Inflow	0.033**	0.059***	0.093***	0.114***	0.062***	0.141***	0.001	-0.078***	-0.018
	(0.014)	(0.021)	(0.015)	(0.018)	(0.012)	(0.023)	(0.011)	(0.015)	(0.017)
Observations	32,424	32,424	32,424	32,424	32,424	32,424	32,424	32,424	32,424
KP F-stat	455.899	455.899	455.899	455.899	455.899	455.899	455.899	455.899	455.899
Decade-County FE	✓	✓	✓	✓	✓	✓	✓	✓	✓

B. Effect of In-Flows on Immigrant Group's Topic-Specific Distance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DV: Social Distance by:	Crime	Church Services	Diseases	Medicines	Schools	Job Ads	Local News	Personal Stories (I)	Personal Stories (II)
Immigrant Inflow	0.028*	-0.058***	0.054***	0.010	-0.007	0.030**	0.021	0.056***	0.044***
	(0.014)	(0.016)	(0.015)	(0.016)	(0.015)	(0.015)	(0.022)	(0.013)	(0.013)
Observations	24,669	18,533	25,118	20,599	22,886	21,340	6,068	28,248	23,984
KP F-stat	440.192	496.624	459.319	444.211	465.451	479.228	346.068	437.403	463.354
Decade-County FE	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes. All variables are standardized. Standard errors are clustered by county.
*,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

information on the method and topic distribution. We assign each paragraph to the most likely topic, and build two new quantities.

First, we measure the prevalence of each topic by nationality, county, and decade, as the average probability that a group mentions belongs to any given topic. Out of 45 topics, we manually identified those that had most to do with local matters. We regress the prevalence of "local" topics on immigration, and investigate whether immigration of a given group increased the likelihood that that group is mentioned in the context of those topics, from crime to job ads. Table 4 Panel A reports the results. As more immigrants arrive, they are mentioned more often in the news in relation to local topics. This is true across almost all selected topics, with the exception of personal readers' stories. Intuitively, their probability of being discussed in international affairs is instead negatively affected (see Appendix Table A.6)

Second, we measure topic-specific social distance. In this case, we aim at understanding if once immigrants arrive, their portrayal across contexts changes. Table 4 Panel B

shows that social distance increases across all local topics, with the exception of (protestant and catholic) church services. Topics related to international affairs also display higher social distance after immigration increases (see Appendix Table A.7).

To sum up, immigration changes both the topics under which the new arrivals are discussed and how they are portrayed within a given topic.

4.3 Inter-group spillovers

This section addresses the question of how the arrival of new immigrant groups affects attitudes toward extant groups. We explore the hypothesis that migration can trigger a re-categorization spillover of extant groups, causing established immigrants to be perceived as either closer or more distant from natives, and that the direction of the re-categorization depends on whether the distance between the new immigrants and native citizens is smaller or larger compared to the distance between the extant group and native citizens (Hangartner et al. 2019; Fouka and Tabellini 2021; Fouka, Mazumder, and Tabellini 2021). In particular, building on Fouka, Mazumder, and Tabellini (2021), we hypothesize that when the new group is more proximate to native citizens, relative to the extant groups, the extant groups suffer from negative re-categorization. In this case, they are considered more distinct from native citizens and their social distance decreases. Vice versa, when the new group is more distinct from native citizens, relative to the extant group, then the social distance for extant groups decreases.

While there are different conceptualizations and measurements of group-wide distance across groups, we focus on three main attributes: language, genetics, and religion. Importantly, all those conceptualizations and measurements of group-wide distance correlate positively among each other, as well as with other metrics of baseline attribute differences, such as geographic distance (Spolaore and Wacziarg 2009).

Proximity Measures. We use three proximity measures. First, the basic idea behind **linguistic proximity** is to capture time-invariant intelligibility across language families and, by extension, across groups. Quantitative measures of linguistic proximity have been

shown to predict migrants' trajectories of language acquisition and integration (Isphording and Otten 2014). We leverage measures of linguistic proximity between the language of each immigrant group and American English, as proposed by Spolaore and Wacziarg (2009). In particular, our measure builds on linguistic trees, that classify languages into families starting from the Ethnologue (Fearon 2003). The number of common nodes in the linguistic tree is a measure of linguistic similarity. Matching countries to their linguistic composition, it is then possible to construct indices of linguistic proximity between the U.S. and immigrants' origin countries. Following Spolaore and Wacziarg (2009), we use two strategies to aggregate to the group level: First, we calculate the linguistic proximity, relative to American English, for the main language spoken in the origin country. Second, we calculate a weighted average of the linguistic proximity of all languages spoken in the origin country.

Next, **genetic proximity** captures the passing of time since when two groups shared common ancestors. This is relevant to the extent that not only it relates to possibly visible genetic differences across groups, but also it approximates the existence of clusters of inter-generational transmission of cultural values and norms (Cavalli-Sforza 1994; Spolaore and Wacziarg 2009). Technically, genetic proximity measures the similarity in gene distributions between two populations, as the probability that two alleles at a given random locus are different across two populations Cavalli-Sforza (1994). Similarly to linguistic proximity, genetic proximity can be mapped into modern nationality definitions, and aggregated to represent either the genetic proximity, relative to Americans, for the main ethnic group in the origin country, or as a weighted average of the genetic proximity of all ethnic groups in the origin country.

Religious differences have been shown to hamper immigrant integration across contexts and time periods (Lucassen 2005; Adida, Laitin, and Valfort 2016). We capture **religious proximity** as a simple binary variable that captures nationalities whose origin country is majority protestant. The data come from Spolaore and Wacziarg (2009) who refer to Mecham, Fearon, and Laitin (2006) for their religious nomenclature. European migration to the U.S. covered in our sample only comprises origin countries whose major-

Table 5: Spillovers on Extant Immigrant Groups

DV: Social distance	(1)	(2)	(3)	(4)	(5)
Less Distant	0.147***	0.168***	0.090***	0.101***	0.113***
	(0.005)	(0.005)	(0.014)	(0.006)	(0.005)
Observations	30752	30752	30752	29516	30752
Group proximity measured as:	Linguistic proximity	Linguistic proximity (weighted)	Genetic proximity	Genetic proximity (weighted)	Religious proximity

Notes. Each column shows the regression of (standardized) Social Distance in county c, decade d, for nationality n, on *Relative Proximity*, i.e. the proximity to the U.S. of the largest nationality in the migration inflow minus the proximity to the U.S. of the extant nationality n. Proximity refers to linguistic proximity in columns 1-2, genetic proximity in columns 3-4, and religious proximity in column 5. We exclude observations where the extant nationality and the largest incoming nationality coincide. All regressions include county-decade fixed effects. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

ity was either protestant or catholic. We code migrant groups originating from protestant countries as more distant than the ones originating from catholic countries.

Relative Group Differences. We calculate the relative measure of proximity for each dyad between an extant and a new immigrant group (defined by nationality). For each county and decade, we focus on the immigrant group with the largest number of arrivals over the last ten years (hereinafter *new group*) (Cikara, Fouka, and Tabellini 2022). For each existing immigrant group (hereinafter *extant group*), we code a variable capturing the extent to which the new group has *higher* linguistic, genetic or religious proximity than the extant group (we refer to this predictor as *Relative Proximity* in Table 5).

Results. We produce OLS estimates of social distance for each immigrant group, county and decade, on each of the five indicators of relative group-wide proximity in attributes. The coefficients estimate the effect of a more proximate group compared to extant ones. We always include county-by-decade fixed effects, as in our most conservative specification in column 6 of table 3.

Table 5 reports the results. We use linguistic proximity in columns 1-2 (using the

main language as reference or using a weighted average of linguistic proximity for all local languages), genetic proximity in columns 3-4 (again, based on the main ethnic group of using a weighted average for all local ethnic groups), and religious proximity in column 5 (using the main country religion as reference).

Across specifications, we see that the arrival of groups that share more attributes with natives is associated with an increase in social distance for extant groups. In all columns, the estimated effects are positive and significant. The result holds for linguistic, genetic, and religious proximity.

5 Conclusion

This paper investigates the relationship between immigration in-flows and social distance during the Age of Mass Migration, using a combination of computational linguistic methods, causal inference methods, and historical newspaper corpora. By validating our measure on newspapers, we construct a novel and robust indicator of social distance that captures nuanced shifts in how immigrant groups were perceived by the native in-group. We find a positive and statistically significant relationship between immigration in-flows and social distance, a result that is robust to various model specifications and alternative measures of distance. Increased local exposure to immigration prompts natives, as reflected in newspaper language, to employ more "othering" terms when discussing immigrant minorities, identifying them more with the out-group. Our conceptualization of social distance is robust to extensive validation tests.

Through topic modeling, the analysis uncovers two mechanisms by which social distance is shaped: first, through changes in the topics under which immigrant groups are discussed, and second, through shifts in the tone and framing of these discussions. This dual dynamic—greater topical integration alongside increased social distance—may arise for several reasons. The increased association of immigrant minorities with local topics could reflect responses to the practical implications of immigration, such as competition in labor markets or public health concerns. However, the persistence of social distance

within these contexts suggests that while immigrants were discussed in local debates, they were not fully integrated into the "in-group" in public discourse. Indeed, our results indicate that local immigration exposure broadens the social distance gap in their representation concerning these topics.

In addition, this paper explores how the arrival of new immigrant groups triggered a re-categorization of attitudes toward existing groups. Drawing on theories of re-categorization, we hypothesize and find evidence that the relative proximity of new immigrant groups to native citizens influenced whether established groups were perceived as closer to or more distinct from natives. When new groups were more culturally or socially proximate to natives, established groups faced negative re-categorization, becoming perceived as more distant. Conversely, when new groups were more distinct, established groups experienced positive re-categorization on average. These dynamics underscore the fluidity of social boundaries (Davenport 2020; Fouka and Tabellini 2021; Fouka, Mazumder, and Tabellini 2021) and the contingent nature of group relations in periods of demographic and cultural change.

The analysis opens several avenues for future research. First, while the primary focus is not to *quantify* biases embedded in language, researchers could extend our methodology to quantify stereotypes by comparing linguistic depictions against demographic benchmarks (Garg et al. 2018). Another extension would investigate implicit dehumanizing metaphors (Card et al. 2022) and their relationship to group-specific immigration exposure. Second, newspapers' multimodal nature offers a promising frontier: integrating computer vision to analyze photographs and illustrations (Lee et al. 2020) could reveal how visual depictions complemented or diverged from linguistic portrayals.

Further analyses using the newspaper corpus could also enrich our findings. For instance, incorporating information on newspaper political affiliations would allow the study of partisanship. The impact of immigration restriction policies could be evaluated, shedding light on how such policies influenced social distance over time. Advertising inserts could also be examined to identify explicit exclusionary rhetoric, such as "No Irish Need Apply" notices.

Our study has several limitations. First, the newspaper corpus is subject to high digitization error, which introduces noise. Ideally, the analysis could be replicated using an extended and cleaned newspaper corpus.Second, the attitudes captured here primarily reflect the literate population, potentially excluding perspectives of less-educated and marginalized groups. Third, limited documentation on the sampling methodology of Chronicling America raises concerns about preservation bias. Finally, while à la carte embeddings (Khodak et al. 2018) allow us to construct measures at the nationality-mention level, census data restrict aggregation to the decade level, limiting temporal granularity.

References

Aarim-Heriot, Najia. 2003. *Chinese immigrants, African Americans, and racial anxiety in the United States, 1848-82.* Vol. 147. University of Illinois Press.

Abramitzky, Ran, and Leah Boustan. 2017. "Immigration in American economic history." *Journal of economic literature* 55 (4): 1311–45.

Adida, Claire L, David D Laitin, and Marie-Anne Valfort. 2016. *Why Muslim integration fails in Christian-heritage societies.* Harvard University Press.

Altonji, Joseph G, Todd E Elder, and Christopher R Taber. 2005. "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools." *Journal of political economy* 113 (1): 151–184.

Ash, Elliott, Daniel L Chen, and Arianna Ornaghi. 2021. "Gender attitudes in the judiciary: Evidence from us circuit courts." *Center for Law & Economics Working Paper Series* 2019 (02).

Bamman, David, Chris Dyer, and Noah A Smith. 2014. "Distributed representations of geographically situated language." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 828–834.

Bansak, Kirk, Jens Hainmueller, and Dominik Hangartner. 2023. "Europeans' support for refugees of varying background is stable over time." *Nature* 620 (7975): 849–854.

Bartik, Timothy J. 1991. *Who benefits from state and local economic development policies? Kalamazoo, MI: WE Upjohn Institute for Employment Research.*

Blalock, Hubert M. 1967. *Toward a theory of minority-group relations.* Vol. 325. New York: Wiley.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55 (4): 77–84.

Blumer, Herbert. 1958. "Race prejudice as a sense of group position." *Pacific sociological review* 1 (1): 3–7.

Bobo, Lawrence, and Vincent L Hutchings. 1996. "Perceptions of racial group competition: Extending Blumer's theory of group position to a multiracial social context." *American sociological review,* 951–972.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5:135–146.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-experimental shift-share research designs." *The Review of economic studies* 89 (1): 181–213.

Burchardi, Konrad B, Thomas Chaney, and Tarek A Hassan. 2019. "Migrants, ancestors, and foreign investments." *The Review of Economic Studies* 86 (4): 1448–1486.

Bursztyn, Leonardo, Thomas Chaney, Tarek A Hassan, and Aakaash Rao. 2024. "The immigrant next door." *American Economic Review* 114 (2): 348–384.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356 (6334): 183–186.

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. "Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration." *Proceedings of the National Academy of Sciences* 119 (31): e2120510119.

Card, David. 2001. "Immigrant inflows, native outflows, and the local labor market impacts of higher immigration." *Journal of Labor Economics* 19 (1): 22–64.

Cavalli-Sforza, Luigi Luca. 1994. "The history and geography of human genes." *Princeton UP.*

Chan, Jimmy, and Wing Suen. 2008. "A spatial theory of news consumption and electoral competition." *The Review of Economic Studies* 75 (3): 699–728.

Cikara, Mina, Vasiliki Fouka, and Marco Tabellini. 2022. "Hate crime increases with minoritized group rank."

Cohn, Raymond L. 2005. "The transition from sail to steam in immigration to the United States." *The Journal of Economic History* 65 (2): 469–495.

Davenport, Lauren. 2020. "The fluidity of racial classifications." *Annual Review of Political Science* 23 (1): 221–240.

Eriksson, Katherine, and Zachary Ward. 2019. "The residential segregation of immigrants in the United States from 1850 to 1940." *The Journal of Economic History* 79 (4): 989–1026.

Fearon, James D. 2003. "Ethnic and cultural diversity by country." *Journal of economic growth* 8 (2): 195–222.

Fouka, Vasiliki, Soumyajit Mazumder, and Marco Tabellini. 2021. "From immigrants to Americans: Race and assimilation during the Great Migration." *Review of Economic Studies.*

Fouka, Vasiliki, and Marco Tabellini. 2021. "Changing In-Group Boundaries: The Effect of Immigration on Race Relations in the United States." *American Political Science Review,* 1–17.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115 (16): E3635–E3644.

Gennaro, Gloria, and Elliott Ash. 2022. "Emotion and Reason in Political Language." *The Economic Journal* forthcoming.

Gentzkow, Matthew, and Jesse M Shapiro. 2006. "Media bias and reputation." *Journal of political Economy* 114 (2): 280–316.

———. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78 (1): 35–71.

Gillani, Nabeel, and Roger Levy. 2019. "Simple dynamic word embeddings for mapping perceptions in the public sphere." *arXiv preprint arXiv:1904.03352.*

Goldin, Claudia, et al. 1994. "The political economy of immigration restriction in the United States, 1890 to 1921." *The regulated economy: A historical approach to political economy,* 223–257.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik instruments: What, when, why, and how." *American Economic Review* 110 (8): 2586–2624.

Guglielmo, Thomas A, et al. 2003. *White on arrival: Italians, race, color, and power in Chicago, 1890-1945.* Oxford University Press, USA.

Hagendoorn, Louk. 1995. "Intergroup biases in multiple group systems: The perception of ethnic hierarchies." *European review of social psychology* 6 (1): 199–228.

Hainmueller, Jens, and Dominik Hangartner. 2013. "Who gets a Swiss passport? A natural experiment in immigrant discrimination." *American political science review* 107 (1): 159–187.

Hainmueller, Jens, and Daniel J Hopkins. 2014. "Public attitudes toward immigration." *Annual review of political science* 17:225–249.

———. 2015. "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants." *American Journal of Political Science* 59 (3): 529–548.

Hangartner, Dominik, Elias Dinas, Moritz Marbach, Konstantinos Matakos, and Dimitrios Xefteris. 2019. "Does exposure to the refugee crisis make natives more hostile?" *American Political Science Review* 113 (2): 442–455.

Isphording, Ingo E, and Sebastian Otten. 2014. "Linguistic barriers in the destination language acquisition of immigrants." *Journal of economic Behavior & organization* 105:30–50.

Jacobs, Laura, Alyt Damstra, Mark Boukes, and Knut De Swert. 2018. "Back to reality: The complex relationship between patterns in immigration news coverage and real-world developments in Dutch and Flemish newspapers (1999–2015)." *Mass Communication and Society* 21 (4): 473–497.

Katz, Daniel, and Kenneth Braly. 1933. "Racial stereotypes of one hundred college students." *The Journal of Abnormal and Social Psychology* 28 (3): 280.

Khodak, Mikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. "A la carte embedding: Cheap but effective induction of semantic feature vectors." *arXiv preprint arXiv:1805.05388.*

Kozlowski, Austin C, Matt Taddy, and James A Evans. 2019. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 84 (5): 905–949.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S Weld. 2020. "The newspaper navigator dataset: extracting headlines and visual content from 16 million historic newspaper pages in chronicling America." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management,* 3055–3062.

Lucassen, Leo. 2005. *The immigrant threat: The integration of old and new migrants in Western Europe since 1850.* University of Illinois Press.

Mecham, R Quinn, James Fearon, and David Laitin. 2006. "Religious classification and data on shares of major world religions." *unpublished, Stanford University* 1 (5): 18.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems,* 3111–3119.

Mullainathan, Sendhil, and Andrei Shleifer. 2005. "The market for news." *American economic review* 95 (4): 1031–1053.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* 1532–1543.

Rodman, Emma. 2020. "A timely intervention: Tracking the changing meanings of political concepts with word vectors." *Political Analysis* 28 (1): 87–111.

Rodriguez, Pedro L, and Arthur Spirling. 2022. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research." *The Journal of Politics* 84 (1): 000–000.

Rudolph, Maja, and David Blei. 2018. "Dynamic embeddings for language evolution." In *Proceedings of the 2018 World Wide Web Conference,* 1003–1011.

Rudolph, Maja, Francisco Ruiz, Susan Athey, and David Blei. 2017. "Structured embedding models for grouped data." *Advances in neural information processing systems* 30.

Smith, David A, Ryan Cordell, and Abby Mullen. 2015. "Computational methods for uncovering reprinted texts in antebellum newspapers." *American Literary History* 27 (3): E1–E15.

Spolaore, Enrico, and Romain Wacziarg. 2009. "The diffusion of development." *The Quarterly journal of economics* 124 (2): 469–529.

Tabellini, Marco. 2020. "Gifts of the immigrants, woes of the natives: Lessons from the age of mass migration." *The Review of Economic Studies* 87 (1): 454–486.

Thijs, Guusje, Damian Trilling, and Anne C Kroon. 2024. "Contextualized Word Embeddings Expose Ethnic Biases in News." In *Proceedings of the 16th ACM Web Science Conference,* 290–295.

Turner, John C, and Katherine J Reynolds. 2011. "Self-categorization theory." *Handbook of theories in social psychology* 2 (1): 399–417.

Viola, Lorella, and Jaap Verheul. 2020. "Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920." *Digital Scholarship in the Humanities* 35 (4): 921–943.

Yang, Tze-I, Andrew Torget, and Rada Mihalcea. 2011. "Topic modeling on historical newspapers." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 96–104.

Immigration and Social Distance: Evidence from Newspapers during the Age of Mass Migration

Supporting Information

Table of Contents

A Historical Background **47**

B Details on the Social Distance Measure **49**

B.1 Nationality Lexicons 49

B.2 Immigrant and Native Lexicons 49

B.3 Pre-processing Details 49

B.4 Robustness to OCR errors 51

B.5 Extreme Immigrant and Native Contexts 53

B.6 Validation of Social Distance Measure with Historical Stereotypes . . . 54

B.7 GPT Validation 57

B.8 Topics 58

C Robustness of the IV **63**

C.1 Reduced Form 63

C.2 Balance tables 63

C.3 Placebo Test 66

C.4 Leave-one-out instrument 66

C.5 Push-pull instrument 69

C.6 Shares endogeneity 70

C.7 Results dropping one nationality at a time 71

C.8 Results on the intensive margin 72

C.9 Results without i.h.s. transformation 72

C.10 Clustering at the Decade by Nationality Level 74

C.11 Relative Size of the Immigrant Group 74

C.12 Removing Ads 75

C.13 Newspaper Coverage 77

D Additional Results **81**

D.1 Heterogeneity by Region of Origin 81

D.2 Newspaper outcomes 81

D.3 Recategorization 81

D.4 Including distinctive names 84

E Summary Statistics **84**

A Historical Background

The Age of Mass Migration was characterized by virtually no restrictions on immigration from Europe to the US. Over 30 million residents from Europe alone entered the United States between 1850 and 1920.

Within this period, we can distinguish two main waves of immigration from Europe. A first wave, between 1850 and 1890, was dominated by the arrival of Northern and Western Europeans. A second wave, starting approximately in 1890, brought along poorer and less skilled Southern and Eastern Europeans. Discrimination and prejudice were more often directed towards the latter new immigrant groups (Fouka, Mazumder, and Tabellini 2021). These groups, although phenotypically white, were often treated as belonging to an inferior race (Guglielmo et al. 2003).

The shift from sail to steam maritime transportation provided another reason for mass migration. This technology was introduced in the 1850s and the share of immigrants arriving by steamship started increasing monotonically, reaching 97% by 1873 (Cohn 2005). This greatly reduced the time of the trans-Atlantic crossing and the morality risk of the journey; furthermore, the increased number of immigrants travelling by ship reduced the overall costs of migrating from Europe to the US (Abramitzky and Boustan 2017). The decrease in transport costs not only increased the total number of European immigrants entering the US, but also changed its composition. In fact, less wealthy Southern-Eastern immigrants reacted more strongly to the drop in costs than Northern-Western Europeans (Abramitzky and Boustan 2017). While in 1850 Northern and Western immigrants, primarily Protestant, occupied 90% of the migrant stock, this number halved by 1920, whereas the share of Southern and Eastern European immigrants, primarily Catholic, increased to 41%.

It has been claimed that the tension between the need for labor on the one hand, and the fear for the outsider on the other, structured immigration policy throughout this entire period. This time interval also coincided with the onset of the eugenics movement, nativism, and the reemergence of the Ku Klux Klan, which openly opposed, among others, - the ideology of Catholics and Jews. The publication of the influential book *The*

Passing of the Great Race by American eugenicist Madison Grant in 1916 was used to advocate for restricted immigration as justification for so-called "scientific racism". Alcohol prohibition also played a role by allegedly fueling embedded anti-immigrant sentiments towards specific ethnic groups. There existed indeed a stark distinction in alcohol consumption habits according to religious beliefs: Protestants, mainly American and also those of German, Scandinavian and British descent, would not practice drinking and rather enforced the law within their communities; in contrast, in the Roman Catholic immigrant culture it was not only tolerated but also common to consume alcohol. Historical evidence points to the fact that while there was a big effort to increase the ability of the government to enforce prohibition laws, the country simply did not have enough resources to fully control the issue. American citizens partially took it upon themselves to enforce the laws informally, through social norms. This led, in part, to further igniting anti-Catholic sentiments.

The raise in migration levels increased tension and translated into an escalation of restrictive immigration policies. The 1917 Immigration Act was passed with an overwhelming majority and required adult immigrants to undertake a literacy test at entry while virtually barring most immigration from Asia. A perhaps lesser known fact is that this law expanded the list of "undesirable" migrants to be excluded from entry such as, for instance, those described as "idiots", "criminals", "alcoholics" and "political radicals".

Shortly after, the Emergency Quotas Act of 1921, amended in 1924 to become the National Origins Act, for the first time introduced a set of country-specific immigration quotas. These were set on 2% of the 30-year prior (1890) level of immigrants per country of origin. Furthermore, a total immigration quota was capped at 165,000 individuals: the law was engineered to specifically address and reduce the inlet of Eastern and Southern Europeans, whose prevalence across the US territory was very low in 1890, while favoring the older wave of migrants such as Germans, Scandinavians and the British. This act deeply reflected the hostility felt towards newer, more recent immigrants as support for it was mainly based upon economic concerns, xenophobia and antipathy towards the "new" ethnic enclaves formed of Southern and Eastern Europeans (Goldin et al. 1994).

B Details on the Social Distance Measure

B.1 Nationality Lexicons

Table B.1 reports the nationality lexicons used in the main analysis. All lexicons include a lexicon for the country, the adjective indicating its inhabitants, and the most common mispellings. Mispellings were chosen based on non-polysemic largest nearest neighbors. Selected polysemic terms have been manually removed (e.g. polish).

1	*Austria*	austria,austrian,austrians,austrohungarian,austroilnngarian,austrial,austriahungary
2	*Belgium*	belgian,belgians,belgium,flemish
3	*Czechoslovakia*	czech,czechs,czechoslovakian,czechoslovaks,czechoslovakia
4	*Denmark*	denmark,danish,dane,denmark,denck,demark
5	*England*	english,british,brits,englishman,englishmen,england,britain,uk,sbritish,leabritish,britisb
6	*France*	french,france,frenchl,frenchg,frenchy,frenchs
7	*Germany*	german,germany,germans,bavarian,germanlkqud,austrogerman,germanv,hagerman,germanstr
8	*Greece*	greeks,greece,greek,greeubaeh,greeuyi,greeh
9	*Hungary*	hungarian,hungarians,hungary
10	*Ireland*	irishman,irishmen,ireland,irishl,irishs,qirishk,irishwo
11	*Italy*	italian,italians,italy,ltalian,ciitalian,itallian,cvnalian,iiitalian,iitalian,talian
12	*Netherlands*	dutch,netherlands,nederlandse,dutchman,nederlands,holland,dujttch,duwtch,dujmtch,dupttch
13	*Norway*	norwegian,norwegians,norway
14	*Poland*	poland,polishidg
15	*Portugal*	portugues,portuguese,portugal,portngal
16	*Russia*	russian,russians,russia,prussian,kussian,eussian,rlussian,ussian,pussian
17	*Scotland*	scotland,scottish,scots,scottland,scotchirish
18	*Spain*	spanish,spaniards,spain,snanish,spauish
19	*Sweden*	swedish,swedes,sweden,wedish
20	*Switzerland*	swiss,switzerland,swissconfederation
21	*Wales*	wales,welshs

B.2 Immigrant and Native Lexicons

Table B.2 reports the immigrant and native lexicons used in the main analysis. Both lexicons include synonyms and and words that are semantically related to

Immigrant	immigrant,immigrants,migrant,migrants,foreign, foreigner,foreigners,alien,aliens expatriate,expatriates,emigrant,emigrants, nonnative,nonnatives,stranger,strangers
Native	america,united_states,american,americans,yankees, us-born,native,natives,yankee yankee,yankees,domestic,americans,usa,indigenous,citizen, citizens,national,nationals, local,locals,resident,residents

B.3 Pre-processing Details

We report the pre-processing steps to obtain the word vectors: (1) Remove punctuation, (2) Remove capitalization, (3) Tokenize, (4) Remove words with less than three

letters, (5) Remove stopwords, (6) Stem. The model is trained on all newspaper pages between the years 1860 and 1920. Additionally, if the OCR occurrence is higher than 50% of words on any given page, we do not consider that page. The final vocabulary is restricted to include the 100,000 most frequent words in the corpus.

The same pre-processing steps are followed for the training of the (i) the FastText model (Bojanowski et al. 2017), (ii) the *A La Carte* model (Khodak et al. 2018) as well as the LDA topic modeling (Blei 2012).

Stopwords are taken from NLTK and include: ('i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't").

FastText training parameters In training the FastText embeddings, we set the window size (i.e. the amount of words placed to the left and the right of a given target word) to 8, the embedding size to 256, and epochs to 5. Vocabulary is also limited to the 100,000 most common words.

A La Carte training parameters In training the mention-level A La Carte embeddings, we set the window size to 8, the embedding size to 256, and epochs to 5.

B.4 Robustness to OCR errors

Optical character recognition (OCR) errors arise are defined as the difference between the original printed characters on paper and the digitized characters. It is common to encounter high rates of OCR errors in historical corpora. We account for this by applying the FastText embedding approach and by (Bojanowski et al. 2017) and dropping newspaper pages with high OCR errors.

In Table A.1 below we show the terms with highest cosine similarity for the terms "American" in the left panel and "German" in the right panel. We can see that OCR errors lead to mispellings of the same word and that our approach is able to correctly capture this.

Table A.1: Fasttext Capturing OCR Errors

FastText	American		German	
	Term	Cosine Similarity	Term	Cosine Similarity
	ameri	0.931	germanv	0.883
	erican	0.894	germau	0.855
	americau	0.891	germana	0.840
	americ	0.884	germai	0.825
	americani	0.882	francogerman	0.818
	merican	0.879	austrogerman	0.816
	america	0.878	germani	0.813
	americn	0.839	qerman	0.806

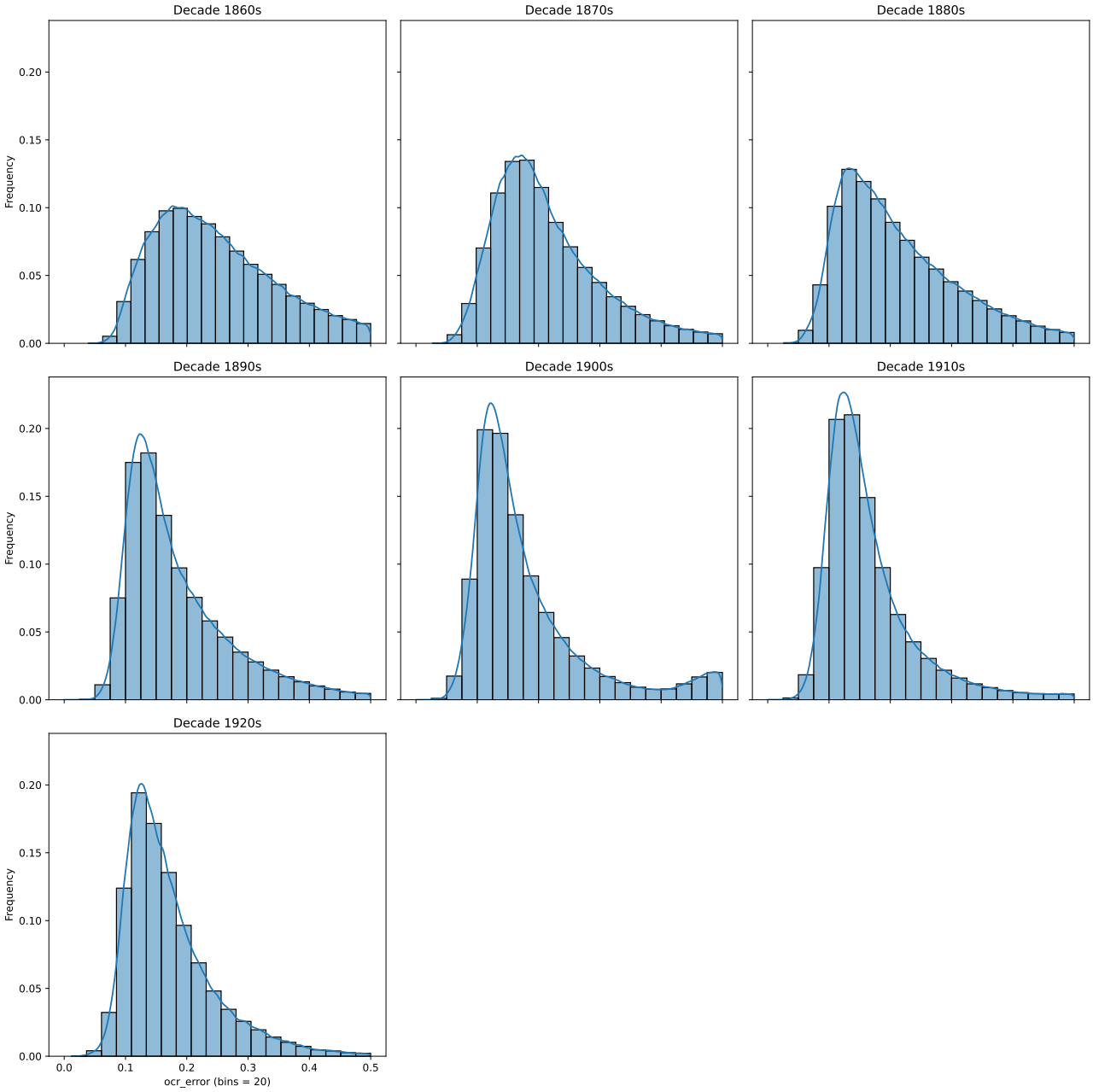
In Table A.2 we report the same results as in Table A.1 but using Word2vec on a pre-trained model on Google News. We can see that other models, such as Word2vec, are less effective in recognizing misspellings or OCR errors as having similar meaning.

Table A.2: Word2vec and OCR Errors

Word2Vec	American		German	
	Term	Cosine Similarity	Term	Cosine Similarity
	amercian	0.691	austrian	0.742
	america	0.612	germany	0.718
	amercan	0.561	germans	0.685
	u.s.	0.531	bavarian	0.680
	amer_ican	0.528	hungarian	0.667
	americans	0.518	dutch	0.649
	african	0.515	belgian	0.644
	ameican	0.510	czech	0.636

Finally, we report they yearly distribution of the OCR error across documents in Figure A.1. The plot shows that, while the quality of text slightly improves over time, there is no massive change in the error distribution.

Figure A.1: Distribution of OCR error across decades



B.5 Extreme Immigrant and Native Contexts

In Table A.3, we manually check for the most extreme immigrant and native contexts, respectively, according to our social distance measure. In the panel on the left below, we can see that, in these extreme examples, immigrant groups are associated with topics like labour market competition, changing names to assimilate to local culture, and distinctive

cultural manifestations. Contexts on the right are classified as highly native and tend instead to include depictions of notable individuals and desirable cultural trends.

Table A.3: Context Examples

Immigrant Contexts	Native Contexts
1. "[...] many of which are imbued with the peculiar rhythm and plaintive sentiment of the African race. The effect of this pronunciation of the **French** language produced a patois which we can scarce pretend to translate." 2. "[...] Congressional legislation which will prevent, when wages become high in this country, the importation of Hungarians, **Italians** or others to take the place of our laboring men in the interests of the capitalists." 3. "He suggested they should select for surnames European or other cities; hence why **French** Jews are called Hustibomie. Creinieuve. Lisbonnc, etc."	1. "I know him personally a little, by reputation still better, as an honorable, straightforward functionary, who would hardly lend himself to anything underhand. **French** notaries, as a rule, bear a deservedly high character." 2. "[...] the books in which he is particularly interested being "The Nomadic Life," by M. Jusseroud, the new **French** ambassador who is to succeed M. Cambon." 3. "LADIES' G. D. JUSTRITE CORSETS Made of finest **French** white, nicely boned; corsets that are perfection of fit."

B.6 Validation of Social Distance Measure with Historical Stereotypes

We use historical surveys on immigrant group stereotypes to compare the ranking of immigrant groups along the social distance dimension, with the ranking of historical stereotypes. In particular, we rely on the study by Katz and Braly (1933), who ask college students to associate adjectives to immigrant groups, and report the top 15 stereotypes most associated to each group. Examples include Italian immigrants being more associated with "Artistic" and "Impulsive", Irish immigrants being associated with "Pugnacious" and "Quick-tempered", and German immigrants being associated with "Scientifically-minded" and "Industrious".[14] For each immigrant group, we calculate a

[14]Some of the stereotypes are expressed as composite words or preceded by qualifiers. To simplify the analysis, in the following cases we drop the qualifier: "very religious", "extremely nationalistic",

stereotype vector as the weighted average of individual embedding vectors associated to each adjective. Specifically:

$$\vec{v_n} = \frac{1}{15} \sum p_{wn} \vec{w}$$

Where $\vec{w}$ is the word embedding for each adjective in the list of nationality n stereotypes; p_{wn} is a the percentage of respondents who associated adjective w to nationality n. The sum is dived by 15 that is the total number of adjectives associated to each nationality. We then calculate the cosine similarity between $\vec{v_n}$ and our immigrant-native dimension ($\vec{IN}$) to obtain a social distance score for each stereotype.

This exercise shows that our method ranks groups in an intuitive way. Results are presented in Table A.4

"physically dirty", "loyal to family ties". We then modify the following composite words, to simpler forms that appear in the model vocabulary: "scientifically-minded" to "scientific", "quick-tempered" to "irritable", "happy-go-lucky" to "carefree', "tradition-loving" to "traditional", "pleasure-loving" to "hedonistic".

Table A.4: Social Distance Measure Captures Known Immigrant Stereotypes

Groups	Stereotype Social Distance	Stereotype adjectives from Katz and Braly (1933)
USA	-0.173	industrious, intelligent, materialistic, ambitious, progressive, pleasure-loving, alert, efficient, aggressive, straightforward, practical, sportsmanlike
UK	-0.125	sportsmanlike, intelligent, conventional, tradition-loving, conservative, reserved, sophisticated, courteous, honest, extremely nationalistic
Negroes	-0.124	superstitious, lazy, happy-go-lucky, ignorant, musical, ostentatious, very religious, stupid, physically dirty, naïve, slovenly, unreliable
Turkey	-0.096	cruel, very religious, treacherous, sensual, ignorant, physically dirty, deceitful, sly, quarrelsome, revengeful, conservative, superstitious
Italy	-0.082	artistic, impulsive, passionate, quick-tempered,musical, imaginative, very religious, talkative, revengeful, physically dirty, lazy, unreliable
Germany	-0.066	scientifically-minded, industrious, stolid, intelligent, methodical, extremely nationalistic, progressive, efficient, jovial, musical, persistent, practical
Jew	-0.023	shrewd, mercenary, industrious, grasping, intelligent, ambitious, sly, loyal to family ties, persistent, talkative, aggressive, very religious
China	-0.003	superstitious, sly, conservative, tradition-loving, loyal to family ties, industrious, meditative, reserved, very religious, ignorant, deceitful, quiet
Japan	0.178	intelligent, industrious, progressive, shrewd, sly, quiet, imitative, alert, suave, neat, treacherous, aggressive
Ireland	0.205	pugnacious, quick-tempered, witty, honest, very religious, industrious, extremely nationalistic, superstitious, quarrelsome, imaginative, aggressive, stubborn

B.7 GPT Validation

Figure A.2: Distribution of GPT-4o Accuracy Scores Across Quintiles

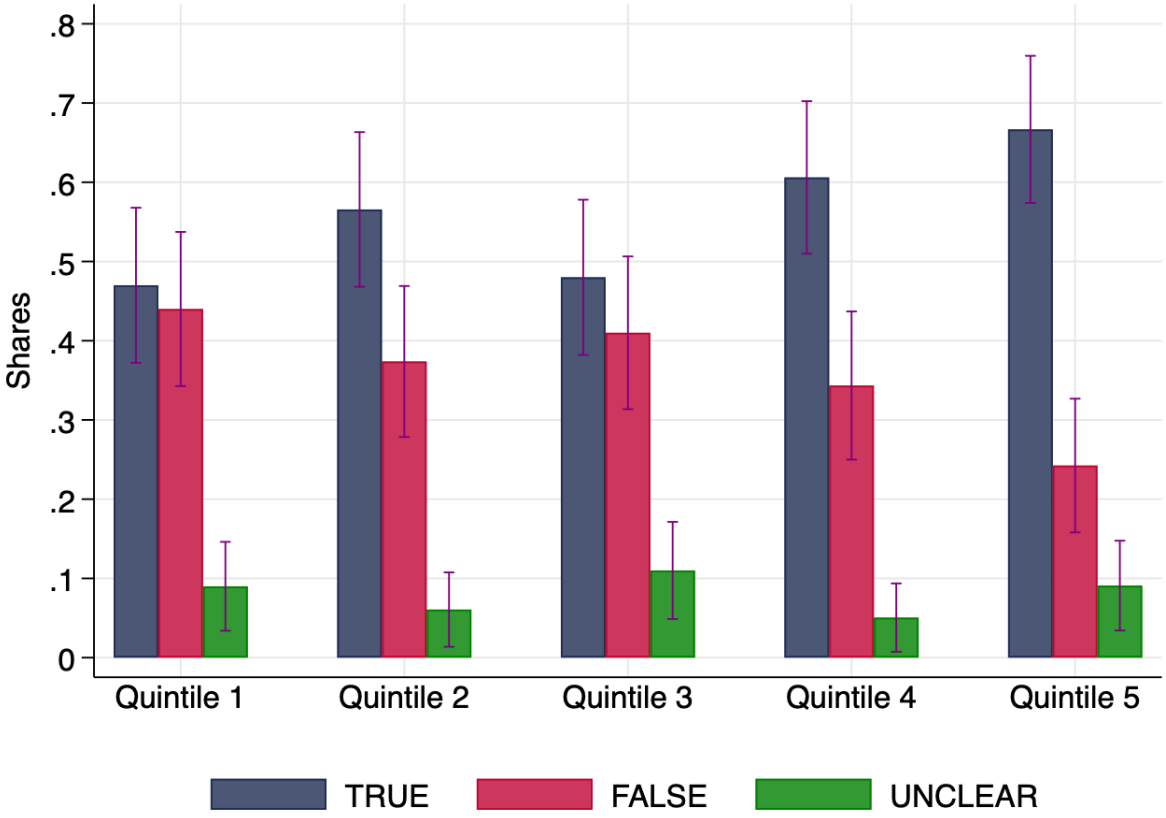


Table A.5: GPT Agreement by Topic Family

Topic Family	**Accuracy (GPT-4o)**
Business	0.825
Government	0.778
Local community	0.794
Sport	0.746
Health	0.746
Law and Order	0.746
Trade	0.762
Random	0.714
Politics	0.683
Economy	0.667
Local Trade	0.651
International affairs	0.635
Traveling and Transportation	0.635
Warfare	0.619
Shipping	0.492
Weather	0.444
Overall Accuracy	**0.684**

B.8 Topics

We estimate a Latent Dirichlet Allocation (LDA) model (Blei 2012) on the entire set of extracted nationality mentions, including the surrounding paragraphs for context.

To train the model, we first preprocess the text data by removing stop words, punctuation, and other irrelevant characters. We then perform tokenization, lemmatization, and lowercasing of all text. The final dictionary is limited to 10,000 unique tokens, excluding nationality names and geographic names. The filtering process for nationality names employs the same variations in spelling that we use to identify nationality-related contexts. To remove geographic names, we utilize the 'geonamescache' Python library, which provides functions to retrieve names of continents, countries, cities, and US states.

The optimal number of topics (45) was determined through a combination of statistical methods and manual evaluation. Specifically, we used coherence scores and perplexity to assess the statistical quality of the models. In addition, we ran multiple models with varying numbers of topics to ensure sufficient granularity and minimize topic duplication. The final model was selected based on a balance between these metrics and manual inspection of the topic coherence.

The LDA model is implemented using the gensim library in Python, which provides efficient algorithms for topic modeling. The preprocessing steps are conducted using the nltk and spaCy libraries.

The topic-specific probability distribution is showed in Appendix Figure A.3.

Descriptive plots Appendix Figure A.4 reports the average social distance score by topic. Multiple accounts of the same topic have been collapsed. The plot shows that topics concerning international affairs carry the highest social distance score. Yet, within topics that concern the local community, we still find higher social distance in topics like Crime and Jobs, compared to Local church and culture. Next, Appendix Figure A.5 shows the prevalence of different topics over time, after aggregating them into 16 main categories [15]. Among other patterns, WWI can be clearly distinguished in the increase in prevalence of warfare and international affairs between 1914 and 1918, at the detriment of local topics.

[15]We manually aggregate the topics into 16 topic families based on a careful inspection of the most frequent words associated with each topic. Additionally, we extract original texts from newspapers for each topic to ensure that the topics within family topics have similar backgrounds.

Figure A.3: Distribution of Topic Probabilities

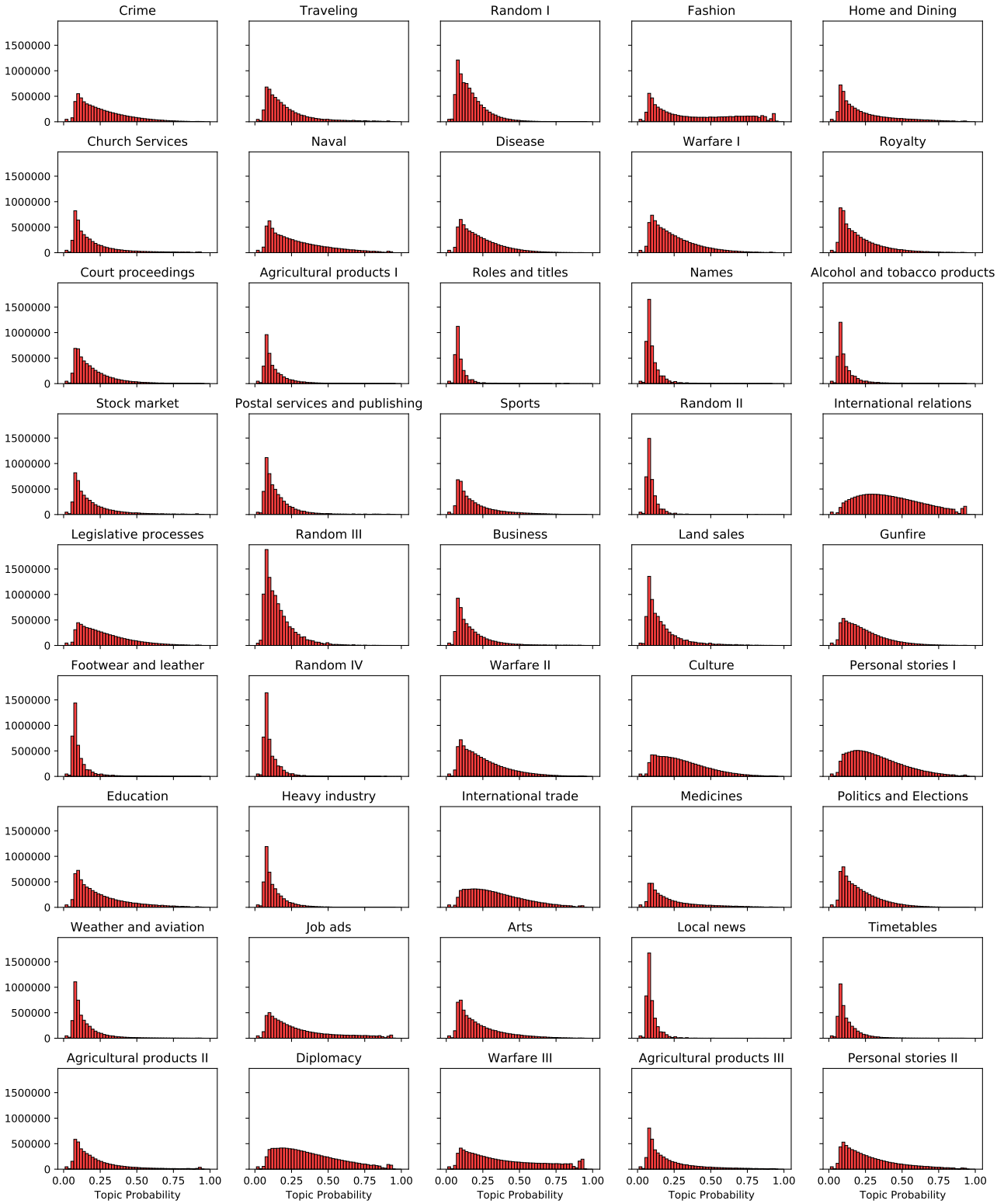


Figure A.4: Social Distance by Topic

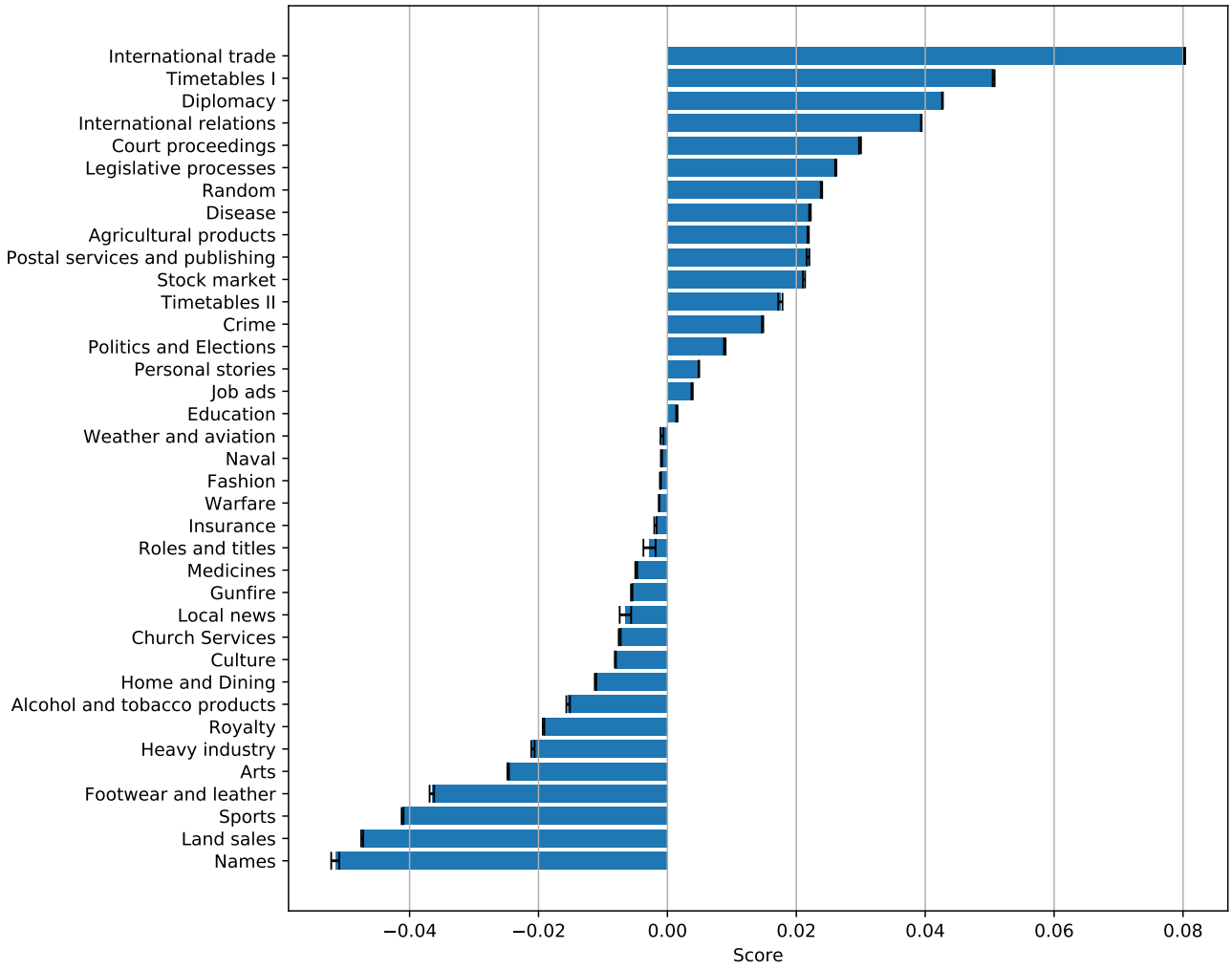


Figure A.5: Prevalence by Topic Family

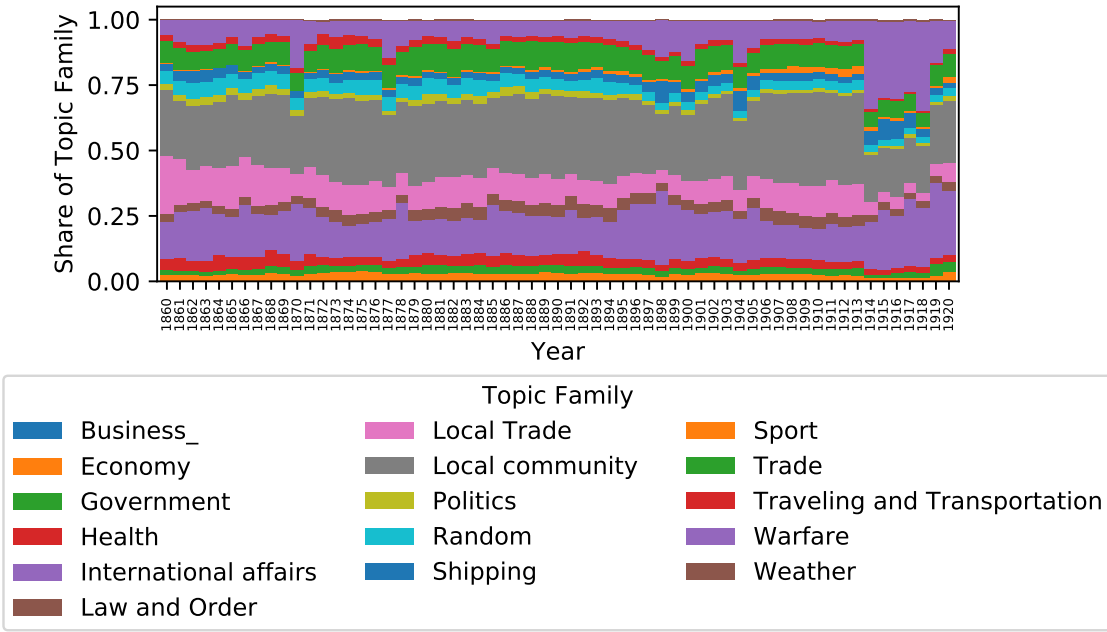


Table A.6: Effect of Immigration on Prevalence of International Topics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Topic Prevalence:	Diplomacy	Royalty	IR	Gunfire	Warfare (I)	Warfare (II)	Warfare (III)
Immigrant Inflow	-0.050***	-0.023*	-0.115***	-0.045***	-0.028*	0.027**	-0.008
	(0.016)	(0.013)	(0.015)	(0.013)	(0.016)	(0.012)	(0.011)
Observations	32,424	32,424	32,424	32,424	32,424	32,424	32,424
KP F-stat	455.899	455.899	455.899	455.899	455.899	455.899	455.899
Decade-County FE	✓	✓	✓	✓	✓	✓	✓

Notes. All variables are standardized. Standard errors are clustered by county. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.7: Effect of Immigration on Social Distance in International Topics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Social Distance on:	Diplomacy	Royalty	IR	Gunfire	Warfare (I)	Warfare (II)	Warfare (III)
Immigrant Inflow	0.130***	0.073***	0.079***	0.088***	0.117***	0.076***	0.069***
	(0.015)	(0.016)	(0.014)	(0.017)	(0.016)	(0.015)	(0.017)
Observations	25,009	24,344	27,164	24,963	23,406	22,569	20,496
KP F-stat	490.307	429.472	445.751	444.575	460.082	486.592	476.236
Decade-County FE	✓	✓	✓	✓	✓	✓	✓

Notes. All variables are standardized. Standard errors are clustered by county. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

The effect of immigration on international topics This section investigates the counterpart of the analyses presented in Table 4 and 4. In particular, it looks at topics that can be considered as international affairs. Appendix Table A.6 shows that, as more immigrants arrive, their nationalities are mentioned less often in the news in relation to international affairs. Note that this corresponds to a rise of attention to local topics. Appendix Table A.7 shows that, as more immigrants arrive, their nationalities are treated with higher social distance in the context of international affairs.

C Robustness of the IV

C.1 Reduced Form

Table A.8 reports the reduced form estimates of the effect of the instrument on social distance. The specification of controls is the same as in our main Table 3.

Table A.8: Reduced Form: Effect of Instrument on Social Distance

DV: Social Distance	(1)	(2)	(3)	(4)	(5)	(6)
IV	0.031***	0.033***	0.033***	0.032***	0.032***	0.032***
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	32424	32424	32424	32424	32424	32424
R-squared	0.077	0.077	0.079	0.081	0.145	0.147
Decade FE	✓	✓	✓	✓		
County FE	✓	✓	✓	✓		
Urban		✓	✓	✓		
Economy			✓	✓		
Demographics				✓		
Decade-County FE					✓	✓
Newspapers						✓

Notes. Notes. Each column shows the regression of Social Distance in county c, decade $d-1$, for nationality n, on *Migration Instrument*, i.e. the shift-share instrument calculated as in equation 6. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.2 Balance tables

Table A.9 reports the regressions of our instrument on several pre-period variables extracted from the 1860 U.S. Census. In particular, we look at whether our instrument correlates with indicators of urbanization (column 1), economic development (column 2) and demographic composition (column 3). Column 4 includes all variables in the same regression. To account for the fact that our main specification includes county fixed effects, the outcome variable is the instrument in first difference. While the instrument is uncorrelated with most pre-period county characteristics, a handful of covariates are significant at the 1 percent level, as one would expect due to chance. Our main specification captures those variables by including county-level fixed effects.

Table A.10 reports the same analysis, separately for each nationality in the sample. Imbalances do not appear to be attributable to any specific nationality.

Table A.9: Instrument Balance - Analysis Sample

DV: Instrument	(1) Urbanization Controls	(2) Economics Controls	(3) Demographics Controls	(4) All Controls
Labour force participation		0.007		0.009
		(0.010)		(0.010)
Manufacturing		-0.028*		-0.023
		(0.017)		(0.016)
Agriculture		0.044***		0.043***
		(0.017)		(0.016)
Mining		-0.011		-0.015
		(0.024)		(0.023)
Real estate property		-0.003		-0.004
		(0.006)		(0.006)
Personal property		-0.004		-0.010
		(0.008)		(0.009)
Share in poverty		0.003		0.002
		(0.005)		(0.006)
Share in crime		0.022**		0.022**
		(0.010)		(0.010)
Share urban	0.162			0.111
	(0.100)			(0.104)
County population	-0.206*			-0.136
	(0.111)			(0.116)
White			-0.074	-0.083
			(0.069)	(0.066)
Black			-0.061	-0.062
			(0.066)	(0.062)
Number of couples in household			0.000	-0.012
			(0.018)	(0.017)
Literacy			-0.007	0.008
			(0.011)	(0.011)
Share in schooling			0.024	0.013
			(0.015)	(0.015)
Observations	474	474	474	474
R-squared	0.230	0.264	0.212	0.290
Joint F-test	3.249**	3.329***	1.057	2.200***
P-val of the F-test	0.040	0.001	0.384	0.006

Notes. Each column shows the regression of the instrument averaged at the county level (in first difference), on 1860 variables. All variables are standardized. All specifications include state fixed effects. Robust standard errors are in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.10: Instrument Balance by Nationality - Analysis Sample

	(1) AT	(2) BE	(3) CZ	(4) DK	(5) EN	(6) FR	(7) GE	(8) GR	(9) HU	(10) EI	(11) IT	(12) NL	(13) NO	(14) RUS	(15) PO	(16) PT	(17) SCO	(18) SP	(19) SE	(20) CH	(21) WE
Urban	0.337	-0.025	-0.473	0.084	0.265	-0.567*	0.388	0.003	0.087	0.694*	-0.155*	0.034	0.251	0.232	-0.167	-0.063	0.339	0.015	0.234	0.224	0.544
	(0.419)	(0.028)	(0.445)	(0.077)	(0.283)	(0.322)	(0.550)	(0.050)	(0.338)	(0.379)	(0.087)	(0.032)	(0.168)	(0.382)	(0.656)	(0.045)	(0.218)	(0.074)	(0.247)	(0.156)	(0.419)
County pop	-0.420	0.034	0.571	-0.089	-0.320	0.688*	-0.438	0.020	-0.175	-0.871**	0.174*	-0.038	-0.326*	-0.301	0.191	0.092*	-0.410*	0.067	-0.304	-0.279	-0.630
	(0.462)	(0.028)	(0.485)	(0.084)	(0.312)	(0.354)	(0.602)	(0.044)	(0.373)	(0.416)	(0.092)	(0.034)	(0.183)	(0.419)	(0.711)	(0.047)	(0.243)	(0.058)	(0.270)	(0.174)	(0.457)
White	-0.413*	-0.022	0.210	-0.008	-0.309*	-0.099	-0.032	-0.057	-0.179	0.031	0.011	-0.012	-0.203	-0.283	0.366	0.003	-0.223*	0.034	-0.260	-0.071	0.031
	(0.222)	(0.050)	(0.223)	(0.055)	(0.172)	(0.281)	(0.313)	(0.065)	(0.218)	(0.179)	(0.052)	(0.038)	(0.123)	(0.214)	(0.365)	(0.027)	(0.134)	(0.046)	(0.178)	(0.089)	(0.126)
Black	-0.394*	-0.016	0.232	0.018	-0.226	-0.098	0.034	-0.050	-0.125	0.017	0.023	-0.011	-0.186	-0.220	0.447	0.007	-0.179	0.037	-0.265	-0.055	0.053
	(0.219)	(0.046)	(0.202)	(0.053)	(0.161)	(0.265)	(0.287)	(0.061)	(0.209)	(0.168)	(0.051)	(0.035)	(0.121)	(0.207)	(0.344)	(0.025)	(0.123)	(0.044)	(0.175)	(0.083)	(0.117)
Couples	-0.044	0.018**	-0.017	0.017	-0.028	-0.146***	0.050	0.017*	-0.039	-0.023	0.025	0.010	0.017	0.041	0.018	-0.010	-0.004	-0.004	-0.008	0.012	-0.008
	(0.082)	(0.009)	(0.064)	(0.016)	(0.050)	(0.054)	(0.074)	(0.010)	(0.067)	(0.039)	(0.022)	(0.008)	(0.039)	(0.091)	(0.118)	(0.007)	(0.040)	(0.016)	(0.055)	(0.023)	(0.023)
Literacy	0.009	0.006	-0.029	0.028**	-0.006	0.033	0.027	-0.001	0.059	-0.112***	0.018	-0.002	-0.014	0.033	0.196**	0.006	0.011	0.002	-0.002	0.000	-0.018
	(0.054)	(0.006)	(0.059)	(0.014)	(0.039)	(0.053)	(0.068)	(0.011)	(0.048)	(0.034)	(0.015)	(0.005)	(0.019)	(0.048)	(0.097)	(0.005)	(0.023)	(0.016)	(0.033)	(0.016)	(0.020)
Schools	0.126	0.002	-0.048	0.011	0.035	-0.043	0.101	0.001	0.047	-0.074	-0.002	0.008	-0.000	0.045	-0.102	-0.003	0.028	0.003	0.020	0.032	-0.002
	(0.080)	(0.008)	(0.071)	(0.013)	(0.055)	(0.046)	(0.073)	(0.008)	(0.043)	(0.049)	(0.019)	(0.006)	(0.030)	(0.052)	(0.107)	(0.008)	(0.038)	(0.012)	(0.041)	(0.023)	(0.039)
Labour force	0.039	0.006	-0.037	-0.009	0.071**	0.011	0.005	-0.001	0.058*	-0.005	0.001	-0.001	0.022	0.039	0.005	0.000	0.040*	-0.001	-0.006	0.002	0.002
	(0.054)	(0.005)	(0.057)	(0.011)	(0.036)	(0.029)	(0.046)	(0.005)	(0.031)	(0.028)	(0.011)	(0.005)	(0.019)	(0.040)	(0.081)	(0.003)	(0.024)	(0.008)	(0.033)	(0.013)	(0.016)
Manufacturing	-0.117	0.011	-0.119	0.011	0.022	0.054	-0.130*	0.008	-0.071	0.004	0.023	-0.005	-0.030	-0.090	0.016	0.007	-0.004	0.021	-0.097**	-0.044**	-0.058
	(0.082)	(0.009)	(0.089)	(0.014)	(0.055)	(0.044)	(0.076)	(0.011)	(0.045)	(0.045)	(0.021)	(0.007)	(0.031)	(0.068)	(0.096)	(0.007)	(0.038)	(0.014)	(0.048)	(0.022)	(0.042)
Agriculture	0.068	-0.002	-0.118	0.036**	0.153***	-0.044	0.097	-0.014	0.075	0.133**	-0.010	-0.004	0.073*	0.095	-0.002	-0.014*	0.100***	-0.069***	0.075	0.046**	0.060**
	(0.084)	(0.008)	(0.088)	(0.016)	(0.054)	(0.051)	(0.072)	(0.015)	(0.055)	(0.054)	(0.021)	(0.007)	(0.041)	(0.075)	(0.105)	(0.008)	(0.037)	(0.021)	(0.049)	(0.022)	(0.025)
Mining	-0.063	-0.010	-0.153*	0.014	-0.051	0.071	0.004	-0.021	-0.005	-0.088	0.004	-0.010	-0.022	-0.031	0.098	0.007	0.016	0.017	-0.028	-0.001	-0.065
	(0.077)	(0.008)	(0.092)	(0.021)	(0.069)	(0.071)	(0.080)	(0.015)	(0.062)	(0.066)	(0.017)	(0.008)	(0.048)	(0.094)	(0.105)	(0.007)	(0.039)	(0.018)	(0.067)	(0.026)	(0.040)
Real estate	-0.009	0.007	-0.051	-0.013	-0.016	-0.020	-0.004	0.018	-0.035	-0.043	-0.010	-0.000	0.001	-0.009	0.136	0.006	0.005	-0.014	-0.057	0.002	0.008
	(0.036)	(0.009)	(0.036)	(0.016)	(0.031)	(0.040)	(0.049)	(0.015)	(0.036)	(0.032)	(0.018)	(0.003)	(0.010)	(0.026)	(0.125)	(0.009)	(0.018)	(0.012)	(0.040)	(0.011)	(0.006)
Property	-0.016	-0.010	0.075	-0.010	-0.017	0.049	-0.110*	-0.019	-0.037	0.048	0.017	-0.001	0.021	-0.006	-0.158	-0.015	-0.024	-0.029*	0.055	-0.007	-0.007
	(0.053)	(0.011)	(0.047)	(0.020)	(0.030)	(0.053)	(0.057)	(0.015)	(0.046)	(0.040)	(0.022)	(0.005)	(0.017)	(0.043)	(0.118)	(0.009)	(0.022)	(0.016)	(0.039)	(0.016)	(0.013)
Poverty	0.017	0.002	-0.019	0.006	0.012	-0.014	0.026	-0.004	-0.061	-0.039*	0.014	0.002	-0.008	0.032	0.018	0.001	0.013	0.000	0.000	-0.003	-0.008
	(0.025)	(0.003)	(0.026)	(0.005)	(0.018)	(0.020)	(0.030)	(0.004)	(0.041)	(0.023)	(0.010)	(0.002)	(0.009)	(0.025)	(0.030)	(0.003)	(0.013)	(0.004)	(0.017)	(0.008)	(0.011)
Crime	0.075**	0.021*	0.180	0.023***	0.038	-0.072	0.042	0.039*	0.073**	-0.041**	0.053***	0.012	0.015	0.043	0.038	0.006	0.019	0.006	0.028	0.013	0.009
	(0.033)	(0.011)	(0.169)	(0.007)	(0.027)	(0.044)	(0.047)	(0.022)	(0.028)	(0.019)	(0.013)	(0.008)	(0.015)	(0.031)	(0.032)	(0.006)	(0.019)	(0.009)	(0.023)	(0.016)	(0.006)
Obs	450	442	294	457	470	471	468	462	450	465	464	465	452	464	434	445	448	465	450	452	458
R2	0.183	0.255	0.184	0.248	0.310	0.215	0.180	0.352	0.216	0.415	0.199	0.126	0.391	0.209	0.113	0.409	0.357	0.515	0.301	0.302	0.323
F-test	2.043**	1.732**	.866	1.850**	2.158***	3.079***	1.983**	1.301	2.180***	4.012***	2.835***	1.011	1.959**	1.374	.971	2.105***	2.224***	5.017***	1.630*	1.869**	4.201***

Notes. Each column shows the regression of the instrument averaged at the county level (in first difference), on 1860 variables. All variables are standardized. All specifications include state fixed effects. Robust standard errors are in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.3 Placebo Test

This section reports a placebo specification. In particular, we want to rule out that social distance in the previous decade is predicted by immigration inflow in the next decade. Should we find a relationship between immigration and past social distance, this could be an indication that immigrants self-select in locations depending on their levels of perceived social distance, or that both social distance and immigrant inflow are co-determined by time-varying county level factors that are not captured by our controls. Table A.11 reports the effect of immigrant inflow between decades d and $d-1$ on the social distance measured in decade $d-1$. All model specifications yield small and highly insignificant estimates.

Table A.11: Placebo Test: Effect of Immigration on Lagged Social Distance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance $d-1$	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.025***	-0.015	-0.018	-0.018	-0.021	-0.014	-0.011
	(0.007)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Observations	15119	15119	15119	15119	15119	15118	15118
F-statistic		284.420	338.759	335.325	337.062	364.923	363.504
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade $d-1$, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.4 Leave-one-out instrument

This section reports a series of three robustness checks on the shift-share instrument, that consist in excluding an increasingly larger portion of the migrants from the national-level shock. The idea is that exogeneity of the shock is best guaranteed if we exclude those migrants that may have been attracted to the US precisely in order to locate to

a specific county. If that was the case, the factors that make a county (or larger area) specifically attractive, may also affect social distance in the medium run.

The shift-share instrument is modified to obtain more robust shocks in the following way. The instrument for the in-flow of immigrant group n in county c at census year t as:

$$z_{nct} = \frac{M_{n,c,1860}}{M_{n,1860}}[(M_{n,t} - M_{n,t-1}) - (M_{n,t,d} - M_{n,t-1,d})] \tag{5}$$

Where $M_{n,c,1860}$ is the number of immigrants from group n in county c in 1860; $M_{n,1860}$ is the total number of immigrants from the same group in the U.S. in 1860; $M_{n,t} - M_{n,t-1}$ is the change in the total number of immigrants to the U.S. from group n between census year t and $t-1$; $M_{n,t,d} - M_{n,t-1,d}$ is the change in the total number of immigrants to the U.S. from group n between census year t and $t-1$ that settle in d. Depending on the specification, d coincides with the county c, its state, or its U.S. sub-region.

Appendix Table A.12 reports the results obtained with an alternative version of the instrument, where we subtract from the national level shift those immigrants that locate in the specific county c. Appendix Table A.13 reports the results obtained after subtract from the national level shift those immigrants that locate in the same state as c. Appendix Table A.14 reports the results obtained after subtract from the national level shift those immigrants that locate in the same region as c. Although excluding a larger share of immigrants from the national shocks slightly reduces the F-statistics of the first stage, the results remain virtually unchanged.

Table A.12: Leave-one-County-out instrument

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.054***	0.090***	0.095***	0.095***	0.093***	0.092***	0.093***
	(0.006)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Observations	32428	32428	32428	32428	32428	32424	32424
F-statistic		367.725	443.492	436.726	436.149	465.189	465.018
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.13: Leave-one-State-out instrument

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.054***	0.093***	0.099***	0.099***	0.097***	0.095***	0.096***
	(0.006)	(0.017)	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)
Observations	32424	32424	32424	32424	32424	32424	32424
F-statistic		340.844	414.912	408.662	408.414	436.334	436.240
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.14: Leave-one-Region-out instrument

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.054***	0.099***	0.105***	0.105***	0.103***	0.100***	0.101***
	(0.006)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Observations	32424	32424	32424	32424	32424	32424	32424
F-statistic		311.467	378.803	372.992	373.331	402.665	402.324
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.5 Push-pull instrument

The push-pull instrument is obtained in the following way. The instrument for the in-flow of immigrant group n in county c at census year t as:

$$z_{nct} = \frac{M_{-n,c,t-1} - M_{-n,c,t-2}}{M_{-n,t-1} - M_{-n,t-2}}[(M_{n,t} - M_{n,t-1}) - (M_{n,t,-c} - M_{n,t-1,-c})] \tag{6}$$

The first component captures the share of migrants arriving between $t-2$ and $t-1$ from any region of origin except the region of country n and who settle in county c. This is a measure of county attractiveness. The second component is national level shock, i.e. the number of migrants arriving between t and $t-1$ from country of origin n, excluding those who settle in the U.S region of c.

Results obtained with this instrument are reported in Appendix Table A.15, and very similar to the main ones in Table 3.

Table A.15: Push-pull instrument

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.053***	0.089***	0.107***	0.107***	0.109***	0.101***	0.098***
	(0.007)	(0.031)	(0.028)	(0.027)	(0.027)	(0.022)	(0.023)
Observations	27411	27411	27411	27411	27411	27411	27411
F-statistic		20.551	18.327	18.212	18.506	18.096	18.188
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.6 Shares endogeneity

Table A.16 reports the results from regressing the social distance measure on the 1860 immigration shares that are used to build the instrument. The small and non-significant coefficient provides additional support to the assumption that the instrument (and its components) are only related to the outcome through their effect on contemporary migration shares.

Table A.16: Testing endogeneity of the shares

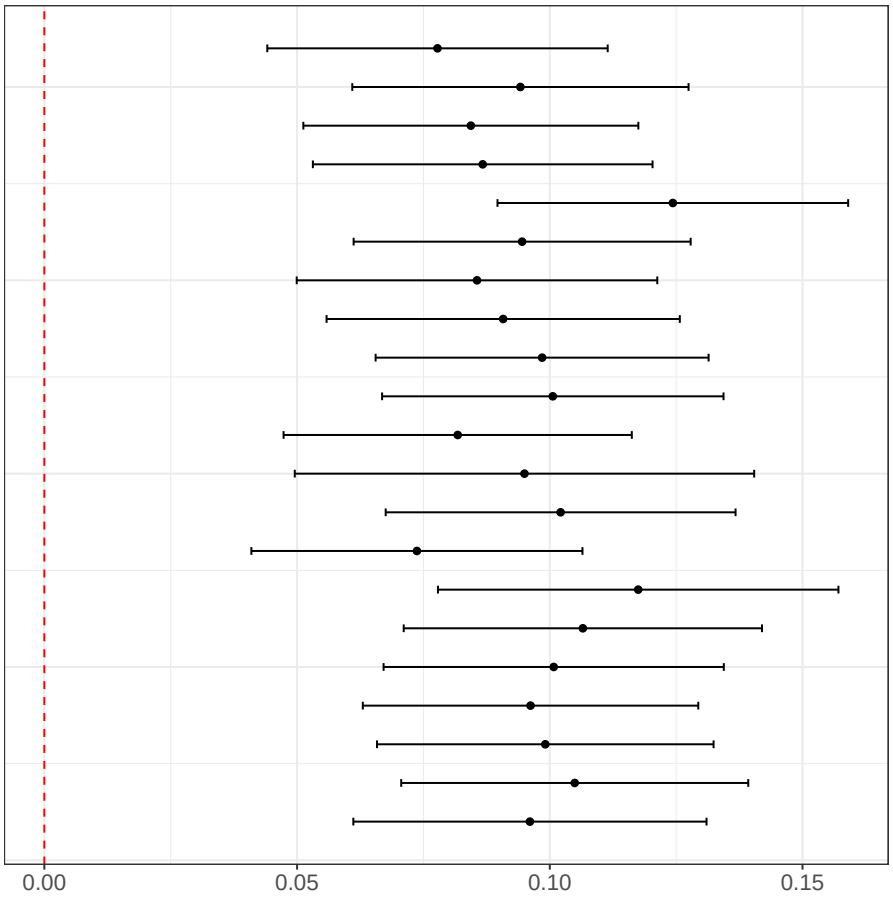
DV: Social distance	(1)	(2)	(3)	(4)	(5)	(6)
Shares (of the instrument)	-0.006	-0.006	-0.006	-0.006	-0.008	-0.007
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Observations	32,428	32,428	32,428	32,428	32,424	32,424
R-squared	0.076	0.077	0.078	0.080	0.144	0.146
Decade FE	✓	✓	✓	✓		
County FE	✓	✓	✓	✓		
Urban		✓	✓	✓		
Economy			✓	✓		
Demographics				✓		
Decade-County FE					✓	✓
Newspapers						✓

Notes. Standard errors are clustered by county. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.7 Results dropping one nationality at a time

Figure A.6 reports the baseline specification as in column 3 of Table 3, dropping one nationality at a time. The table shows that results hold and are not driven by one particular nationality.

Figure A.6: Effect of Immigration on Social Distance Leaving One Nationality Out



C.8 Results on the intensive margin

Table A.17 reports the main results, on the subset of county-decade-nationality observations for which immigration flow is larger than zero. Results hence leverage on the intensive, rather than extensive, margin of migration flows. Results are largely consistent with our main results in Table 3

C.9 Results without i.h.s. transformation

Table A.18 reports the main results, as in Table 3, without transforming the immigration inflow variable with the inverse hyperbolic sine transformation.

Table A.17: Effect of Immigration on Social Distance - Only positive shocks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DV: Social Distance	OLS	IV	IV	IV	IV	IV	IV
Immigrant inflow	0.150***	0.219***	0.229***	0.231***	0.227***	0.224***	0.225***
	(0.018)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)
Observations	12777	12777	12777	12777	12777	12777	12777
F-statistic		528.339	582.440	581.090	580.733	605.710	629.825
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. The sample only includes county-decade-nationality observations with strictly positive migration inflows. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.18: No i.h.s. transformation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Immigrant inflow	0.013**	0.015**	0.019***	0.019***	0.019**	0.018***	0.018***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	32428	32428	32428	32428	32428	32424	32424
F-statistic		64.246	59.351	59.430	59.394	60.481	60.388
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. No variable is i.h.s. All variables are standardized. Controls are included as in the main Table. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.19: Alternative clustering

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Immigrant inflow	0.054***	0.090***	0.096***	0.096***	0.094***	0.092***	0.093***
	(0.005)	(0.015)	(0.015)	(0.016)	(0.016)	(0.015)	(0.015)
Observations	32424	32424	32424	32424	32424	32424	32424
F-statistic		1170.258	1206.632	1194.003	1201.033	1314.597	1321.389
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. Controls are included as in the main Table. Standard errors are clustered at the nationality-decade level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.10 Clustering at the Decade by Nationality Level

Table A.19 reports the main results, as in Table 3, but with standard errors clustered at the nationality-decade level, that is the level at which the shock is measured.

C.11 Relative Size of the Immigrant Group

In Table A.20, we replicate the main analysis regressing social distance for a given immigrant group in a county-decade, on the share of that immigrant group over total population in a given county-decade. Results are unchanged. Table A.20 uses the IV as a share of the total county population in a given county-decade. Table A.21 uses the IV as a share of the total county population in a 1860.

Table A.20: Effect of Shares of Immigrants over County Population

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Immigrant share	0.026***	0.194***	0.196***	0.196***	0.193***	0.176***	0.173***
	(0.006)	(0.056)	(0.057)	(0.057)	(0.056)	(0.049)	(0.048)
Observations	32382	32382	32382	32382	32382	32382	32382
F-statistic		7.272	7.308	7.318	7.355	6.910	6.912
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant share*, i.e. the change in the share of immigrants of nationality n in county c between decade d and $d-1$ over total county population. All variables are standardized. Controls are included as in the main table. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.21: Effect of Shares of Immigrants over 1860 County Population

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Immigrant share	0.007	0.329***	0.330***	0.336***	0.344***	0.346***	0.341***
	(0.007)	(0.090)	(0.090)	(0.092)	(0.096)	(0.092)	(0.092)
Observations	32424	32424	32424	32424	32424	32424	32424
F-statistic		13.624	13.589	13.477	12.823	13.886	13.862
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant share*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$, normalized by county population in 1860. All variables are standardized. Controls are included as in the main table. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.12 Removing Ads

To identify advertisements (or any other recurring content) in our dataset, we employ a systematic approach that uses text similarity measures. In particular, we use pairwise comparisons of every context extracted from the same newspaper on different dates or the same date but across different editions (e.g., morning and evening editions).

We begin by parsing contexts specific to each newspaper, identified by unique newspaper IDs. Potential advertisements are flagged by comparing textual content across various dates and editions. If two contexts from the same newspaper have a high degree of similarity—quantified by a *rapidfuzz* partial ratio match exceeding a threshold of 91%—while differing in publication date or edition, they are classified as likely advertisements. The threshold of 91% accounts for any OCR errors that may be present in the contexts. This threshold is chosen based on experimentation with different percentages to optimize the accuracy of advertisement detection.

To validate our method, we manually review a random sample of 100 detected advertisements by examining the raw newspaper pages. In all 100 cases, the identified contexts are either advertisements or train/boat timetables, confirming the accuracy of our detection method. This validation ensures that our approach reliably identifies recurring content indicative of advertising across large text corpora.

In Table A.22, we test the robustness of the main results with respect to more restrictive measures of social distance. In particular, we replicate the main specification as in Table 3, column 6, on different versions of the outcome variable that drop elements of the newspapers that should be less relevant to the measurement of social distance. We drop ads in column 1, international affairs in column 2, and both in column 3. Results are unchanged. Columns 4 and 5 further control for sentiment, and topic fixed effects.

Table A.22: Robustness on Topics and other Newspaper Elements

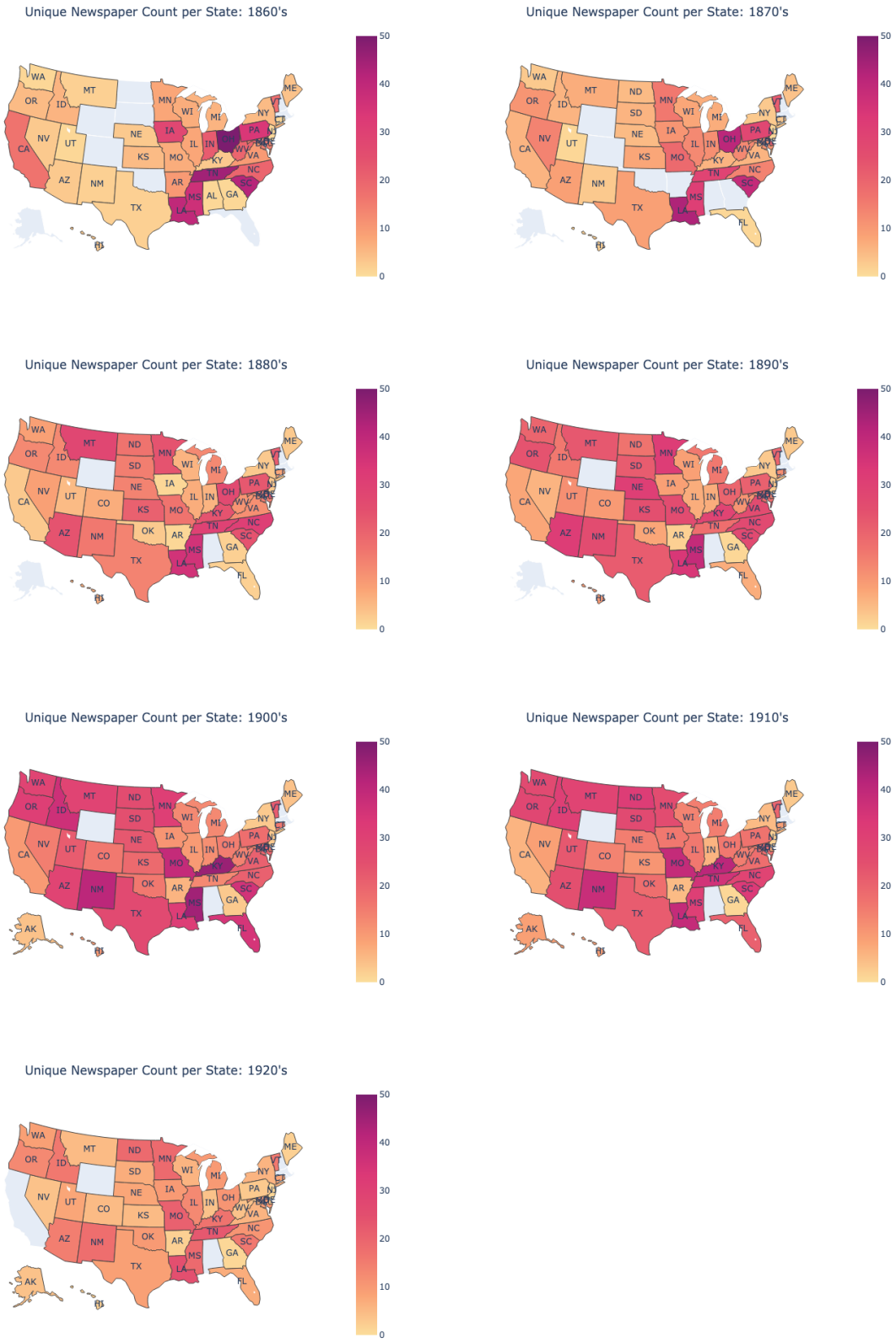
	(1) No Ads	(2) No International Affairs	(3) No Ads or International Affairs	(4)	(5)
Immigrant share	0.092***	0.099***	0.099***	0.105***	0.125***
	(0.016)	(0.017)	(0.017)	(0.017)	(0.017)
Observations	32424	32283	32283	32424	32424
F-statistic	455.899	454.727	454.727	457.800	462.992
Decade-County FE	✓	✓	✓	✓	✓
Sentiment FE				✓	✓
Topic FE					✓

Notes. Each column shows the regression of the social distance score in county c, decade d, for nationality n, on *Immigrant share*, i.e. the change in the share of immigrants of nationality n in county c between decade d and $d-1$ over total county population. Social distance is calculated after dropping ads (column 1), international relation topics (column 2) or both (column 3). All variables are standardized. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.13 Newspaper Coverage

Figure A.7 shows the geographic distribution of the newspaper data, across the different decades.

Figure A.7: Unique newspaper counts by state and decade



Recall that the corpus Chronicling America covers 924 counties out of the approximately 3000 counties of the time. The counties were 2049 in 1860. Table A.23 shows how 1860 counties that are covered by Chronicling America differ from non covered counties. In particular, we regress a dummy variable equal to 1 if the 1860 county is covered by the corpus, and 0 otherwise, on county characteristics extracted from the 1860 census. Covered counties have larger population, marginally higher literacy, and are less dependent on agriculture.

Table A.23: 1860 counties covered by Chronicling America

DV: Instrument	(1) Urbanization Controls	(2) Economics Controls	(3) Demographics Controls	(4) All Controls
Labor force participation			-0.130*	-0.135*
			(0.0769)	(0.0809)
Manufacturing			-0.0739	-0.190
			(0.220)	(0.236)
Agriculture			-0.292***	-0.308***
			(0.0672)	(0.0744)
Mining			0.507**	0.715***
			(0.200)	(0.214)
Real estate property			0.000	0.000
			(0.000)	(0.000)
Personal property			0.000	0.000*
			(0.000)	(0.000)
Share in poverty			1.873	0.355
			(9.041)	(9.414)
Share in crime			0.304	0.366
			(4.284)	(4.152)
White		0.235*		0.472***
		(0.137)		(0.134)
Black		0.560**		0.540**
		(0.233)		(0.244)
Number of couples in household		-0.136		0.0949
		(0.105)		(0.126)
Literacy		0.207***		0.0204
		(0.0786)		(0.0848)
Share in schooling		0.0102		0.0338
		(0.144)		(0.149)
Urban population	-0.000			0.000
	(0.000)			(0.000)
County population	0.000*			0.000
	(0.000)			(0.000)
Observations	2,049	2,068	2,068	2,049
R-squared	0.004	0.006	0.028	0.032
Joint F-test	3.743	2.866	7.200	5.021
P-val of the F-test	0.024	0.014	0.000	0.000

Notes. Each column shows the regression of the social distance score in county c, decade d, for nationality n, on *Immigrant share*, i.e. the change in the share of immigrants of nationality n in county c between decade d and $d-1$ over total county population. Social distance is calculated after dropping ads (column 1), international relation topics (column 2) or both (column 3). All variables are standardized. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

D Additional Results

D.1 Heterogeneity by Region of Origin

Table A.24: Heterogeneity by region of origin

DV: Social Distance	(1) East Europe	(2) North Europe	(3) South Europe	(4) West Europe
Immigrant inflow	0.023	-0.039	1.071***	-0.028
	(0.039)	(0.028)	(0.152)	(0.022)
Observations	5310	11047	6325	9617
F-statistic	140.346	438.053	90.419	467.545
Decade-County FE	✓	✓	✓	✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. All variables are standardized. All columns report the main specification as in Column 3 in Table 3. The sample include different regions. West Europe includes Austria, Belgium, France, Germany, the Netherlands, and Switzerland. East Europe includes Bulgaria, Czechoslovakia, Hungary, Russia, Poland. North Europe includes Denmark, England, Ireland, Norway, Scotland, Sweden, and Wales. South Europe includes Greece, Italy, Portugal and Spain. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

D.2 Newspaper outcomes

Table A.25 reports the effect of the immigration inflow on a number of newspaper outcomes. First, the number of mentions for a given immigrant group, county and decade. Second, it reports the effect of the immigration inflow on the average page number where mentions for a given immigrant group appear, by county and decade. Third, it reports the effect on sentiment. Fourth, it reports the effect of the immigration inflow on the average share of OCR errors on pages where mentions for a given immigrant group appear, by county and decade.

D.3 Recategorization

Table A.26 reports the results of regressing social distance on a dummy equal to 1 when the largest immigrant group is more proximate to American-born citizens than the extant group. The results are opposite to the ones presented in the main paper table 5.

Table A.25: Other newspaper outcomes

	(1)	(2)	(3)	(4)
DV:	Number of mentions	Avg page number	Sentiment	Avg OCR error
Immigrant inflow	-0.100***	0.016**	-0.047***	-0.012***
	(0.027)	(0.008)	(0.013)	(0.004)
Observations	32424	32424	32424	32424
F-statistic	455.899	455.899	455.899	455.899
Decade-County FE	✓	✓	✓	✓

Notes. Each column shows the regression of the average page number in county c, decade d, for nationality n, on *Immigrant share*, i.e. the change in the share of immigrants of nationality n in county c between decade d and $d-1$ over total county population. All variables are standardized. Controls are included as in Table 3. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.26: Recategorization - Less distant immigrant group

DV: Social distance	(1)	(2)	(3)	(4)	(5)
Less Distant	0.221***	0.290***	0.126***	0.083***	0.223***
	(0.015)	(0.014)	(0.014)	(0.013)	(0.013)
Observations	30752	30752	30752	29516	30752
Group proximity measured as:	Linguistic proximity	Linguistic proximity (weighted)	Genetic proximity	Genetic proximity (weighted)	Religious proximity

Notes. Each column shows the regression of (standardized) Social Distance in county c, decade d, for nationality n, on *Less Distant*, i.e. a dummy equal to 1 if the largest nationality in the inflow has lower group distance than nationality n. *Less Distant* refers to linguistic distance in columns 1-2, genetic distance in columns 3-4, and religious distance in column 5. All regressions include county-decade fixed effects. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.27: Including distinctive names

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Immigrant inflow	0.054***	0.167***	0.178***	0.179***	0.176***	0.166***	0.164***
	(0.006)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Observations	33883	33883	33883	33883	33883	33883	33883
F-statistic		355.760	430.552	423.836	423.427	452.430	452.189
Decade FE	✓	✓	✓	✓	✓		
County FE	✓	✓	✓	✓	✓		
Urban			✓	✓	✓		
Economy				✓	✓		
Demographics					✓		
Decade-County FE						✓	✓
Newspapers							✓

Notes. Each column shows the regression of Social Distance in county c, decade d, for nationality n, on *Immigrant inflow*, i.e. the inflow of immigrants of nationality n in county c between decade d and $d-1$. Social distance is measured on nationality mentions, and distinctive names. All variables are standardized. Controls are included as in the main Table. Standard errors are clustered at the county level. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

D.4 Including distinctive names

E Summary Statistics

Table A.28: Summary Statistics

	Count	Mean	SD	Min	Max
Main Variables					
Social distance	32424	0.020	0.033	-0.264	0.405
Immigrant inflow	32424	97.650	2880.389	-114550.000	179834.000
Immigrant inflow (i.h.s)	32424	0.422	3.093	-12.342	12.793
IV	32424	76.756	2204.754	-81982.219	191544.109
Instrument (i.h.s.)	32424	0.753	2.567	-12.007	12.856
Mention-level variables					
Number of mentions	32424	1683.014	9783.420	1.000	536422.000
Average page number	32424	3.652	1.966	1.000	26.951
Pct. OCR error	32424	0.202	0.075	0.051	0.500
District Variables in 1860					
Total county population	32424	18838.818	52930.187	35.000	813154.000
Share in urban areas	32424	0.085	0.205	0.000	1.000
Share in labor force	32424	0.457	0.117	0.125	0.991
Share in manufacturing	32424	0.072	0.047	0.000	0.269
Share in agriculture	32424	0.520	0.199	0.008	0.951
Share in mining	32424	0.012	0.060	0.000	0.724
Average real estate value	32424	580.656	925.470	0.000	9646.845
Average property value	32424	578.261	842.019	3.368	9454.968
Share pauper	32424	0.001	0.002	0.000	0.022
Share crime	32424	0.000	0.002	0.000	0.033
Share White	32424	0.974	0.061	0.199	1.000
Share Black	32424	0.023	0.057	0.000	0.801
Avg couples in a household	32424	0.880	0.107	0.219	1.111
Share literate	32424	0.885	0.132	0.136	1.000
Share in school	32424	0.177	0.072	0.000	0.347

Notes. Main summary statistics for all variables included in the IV analysis.

Table A.29: Summary Statistics - Nationality Mentions

	Count	Mean	SD	Min	Max
Austria	1575	696.957	2626.597	1	74057
Belgium	1538	487.783	2266.068	1	69194
Czechoslovakia	721	34.158	149.653	1	3346
Denmark	1559	361.255	1147.963	1	18599
England	1660	8340.011	26894.834	1	523458
France	1656	6067.024	20874.386	1	506732
Germany	1649	5058.360	19496.717	1	536422
Greece	1602	584.753	1851.434	1	41835
Hungary	1520	292.764	992.589	1	23842
Ireland	1620	1995.381	6883.593	1	89844
Italy	1619	1663.970	5683.989	1	133770
Netherlands	1622	953.483	3086.769	1	76731
Norway	1527	390.975	1476.207	1	24305
Russia	1621	2337.331	8134.456	1	196766
Poland	1517	235.181	771.801	1	18758
Portugal	1508	144.759	501.984	1	8508
Scotland	1556	207.898	689.119	1	11433
Spain	1625	2059.974	7730.390	1	174885
Sweden	1557	444.957	1599.950	1	23062
Switzerland	1587	540.819	2131.145	1	44616
Wales	1585	545.059	1621.295	1	25567
Total	32424	1683.014	9783.420	1	536422

Notes. Main summary statistics for the number of mentions for each nationality in any county-decade. The sample includes county-nationality-decade observations that are included in the main analysis.