C A G E

The problem of false positives in automated census linking: Evidence from nineteenth-century New York's Irish immigrants

CAGE working paper no. 568

June 2021

Tyler Anbinder
Dylan Connor
Cormac Ó Gráda
Simone Wegge

WARWICK
THE UNIVERSITY OF WARWICK

UKRI Economic and Social Research Council

The Problem of False Positives in Automated Census Linking: Evidence from Nineteenth-Century New York's Irish Immigrants[*]

Tyler Anbinder, George Washington University

Dylan Connor, Arizona State University

Cormac Ó Gráda, University College, Dublin

Simone Wegge, College of Staten Island and The Graduate Center—CUNY

ABSTRACT

Automated census linkage algorithms have become popular for generating longitudinal data on social mobility, especially for immigrants and their children. But what if these algorithms are particularly bad at tracking immigrants? Using nineteenth-century Irish immigrants as a test case, we examine the most popular of these algorithms—that created by Abramitzky, Boustan, Eriksson (ABE), and their collaborators. Our findings raise serious questions about the quality of automated census links. False positives range from about one-third to one-half of all links depending on the ABE variant used. These bad links lead to sizeable estimation errors when measuring Irish immigrant social mobility.

[*] We owe a special debt to our genealogist Janet Wilkinson Schwartz, whose expertise and care made this work possible. We are also grateful to Ran Abramitszky, Brian A'hearn, Leah Boustan, Marion Casey, Chris Colvin, Nick Crafts, Joe Ferrie, Eoin McLaughlin, Peter Solar, and Marianne Wanamaker for comments, data, and other support. The usual disclaimer applies. An earlier version was presented at Northwestern University and at the University of Regensburg. Finally, our project would not have been possible without the financial support of the George Washington University, the CUNY Research Foundation (PSC-CUNY Award # 63257-00 51), and the National Endowment for the Humanities (grant RZ-51352-11).

John B. Purcell was one of the best-known Irish immigrants in mid-nineteenth-century America. A native of County Cork who emigrated to the United States in 1820, Purcell must have been a natural-born leader. Four years after Purcell graduated from Mount St. Mary's Seminary in Maryland in 1823, the faculty welcomed him back as a professor. Three years after that, the trustees made him college president. And just three years later, in 1833, Pope Gregory XVI named Purcell the bishop of Cincinnati. In 1850, Pius IX appointed him archbishop and gave him jurisdiction over the entire American Midwest. In that role Purcell became notorious for his strident defence of American Catholics against the attacks of nativist zealots such as Thomas Nast, who caricatured Purcell (See Figure 1) as a power-hungry dictator who sought to impose Catholic dogma on all Americans.

None of that notoriety matters, however, when it comes to automated census linking. For well over half a century, scholars have been trying through one means or another to trace individuals' occupations and locations over the course of their lifetimes in order to measure and compare rates of socio-economic mobility. Harvard's Stephan Thernstrom pioneered this work in the 1960s and inspired many imitators. Yet methodological issues and technological constraints—in particular the inability to demonstrate that traced individuals were representative of entire populations—stymied research in this subfield until computers made it easier to track people as they moved around the United States. In the 1990s, economists and other social scientists began to examine Thernstrom's questions anew, now that they could do so with microprocessors rather than microfilm. Over the past decade or so, mobility studies have gained renewed popularity due to interest in the origins and history of inequality, easy access to online census databases, and the development of algorithms producing data that trace millions of individuals over many decades of census returns. While several groups of scholars have created such algorithms, those written by Ran Abramitzky, Leah Boustan, Katherine Eriksson (hereafter "ABE") and their collaborators have become the most widely used, in part because they have made both the datasets and the coding that generated them easily available on their project website. The ABE data have been used recently for a number of ambitious studies of intra- and inter-generational mobility, focusing in particular on

immigrant groups in the United States (e.g. ABE 2012, 2014; Alexander and Ward 2018; Connor 2019; Pérez 2019; Beck Knudsen 2019).

It was our own interest in American immigrants that led us to the world of automated census linking. Three of us have been working for a decade compiling and analyzing a database tracking the lives of the 15,000 New Yorkers (mostly Irish-born refugees of the Great Famine) who opened accounts at the Emigrant Industrial Savings Bank (hereafter "ESB")[1] from 1850 to 1858. Our findings have included that 1) the New York Irish saved much more money than we had imagined given the prevailing view that Irish immigrants in this era were mired in poverty; and 2) New York's Famine immigrants enjoyed far more upward occupational mobility than we had expected (Anbinder 2012; Anbinder, Ó Gráda, and Wegge 2019). But there was the possibility that our findings stemmed from the positive selection of the savers from the wider immigrant population. It was our search for a way of determining the socio-economic mobility of a true cross-section of New York's Irish immigrant population that led us to the ABE databases. In the combined 1860 to 1870 "crosswalk" from ABE (the one most relevant to our work) and the IPUMS complete-count census data for these years (Ruggles et al., 2020), there are 15,000 theoretically representative Irish immigrants who lived in New York City as of 1860.

At first, we were reassured by the answers found in the ABE-generated database of New York's Irish immigrants. The ESB's customers were no more upwardly mobile than the Irish New Yorkers tracked by the ABE method. But as we looked more closely at the computer-generated census links, we became less sanguine. Not only were the ESB depositors no more upwardly mobile than the computer-generated sample, they were also considerably more *downwardly* mobile. And the ESB customers were eight or nine times less likely to change the state they lived in than the ABE-generated database of Irish New Yorkers. In short, the occupational and geographic mobility of the Irish immigrants in the ABE-generated database defied credibility.

[1] The bank, which still operates in New York, dropped "Industrial" from its name more than a century ago, and we will refer to it by its current and more recognizable name.

That was what led us to Archbishop Purcell. As we examined the names in the ABE-generated database of Irish immigrants and compared their supposed occupations and locations in 1860 and 1870, we noticed Purcell. There he was, living in Cincinnati in 1860, his occupation listed as "RC [Roman Catholic] Archbishop," albeit with his name spelled incorrectly as "Pursell." Yet while we know that the cleric remained in his post in southern Ohio until his death in 1883, the algorithm would have us believe that he had retired by 1870 and had moved to Philadelphia. In fact, Purcell is easy to find in the 1870 census, still in Cincinnati, still listed as "archbishop," although with his surname now spelled correctly, which explains why the algorithm mistakes him for a "John Pursell" of the same age in Pennsylvania.

Our examination of all 95,000 or so of the adult male Irish-born Americans whom the ABE algorithm traces from 1860 to 1870 suggests that about half of those links generated by the ABE "exact" method are false positives like Purcell. Even if one uses the ABE "conservative" variant, designed to limit false positives (also known as "Type I errors"), about a third of the links of Irish immigrants are false. These errors result primarily from rampant age misreporting and surprisingly wide variations in the spelling of names by the original census takers (who wrote each name out by hand) and contemporary census transcribers, who must decipher these sometimes barley legible census returns so that they can be easily searched by scholars, genealogical enthusiasts, and algorithms. This rate of Type I errors raises serious questions about the reliability of studies of immigrants based on algorithmic census linkage.

LITERATURE SURVEY

Systematic attempts at creating longitudinal datasets from census returns began with Thernstrom's *Poverty and Progress* (1964), an analysis of working-class males in the town of Newburyport in northeast Massachusetts. Most of the "hundreds of obscure men" studied by Thernstrom were Irish; he reckoned that neither they nor their compatriots in Boston, nor their children in either place, progressed very far in terms of social mobility (1964, p. 223; 1973, p. 89, 142-3, 247; 1986, p. 42). Thernstrom's method—making a list of every adult male found in the Newburyport census returns from 1850 and then looking for each of those people in the city's subsequent census

schedules—meant he could only hope to track "persisters," i.e. those who chose to remain in Newburyport. Thanks to the creation in the 1980s of searchable state-level census indexes, finding those who moved became feasible, albeit extraordinarily laborious, allowing subsequent work in this area to include "nonpersisters" as well. It emerged that those who moved were more upwardly mobile than "persisters." This corroborated the common presumption that the "best and the brightest" of the poor are the most likely to relocate in search of better employment opportunities. At that stage, however, most of this research was still local, focused on a particular town or county (e.g. Kessner 1977; Griffen and Griffen 1978; Galenson and Pope 1989; Knights 1991).

By developing an iterative linking strategy that matched males of all ages across the entire US with the aid of national census indexes, Ferrie (1995, 1997, 1999; compare Herscovici 1998) set the research agenda for a new generation. From the 1850 U.S. census to 1860 he traced 580 individuals, a number which seems small today but was considered quite impressive a quarter century ago. In the early 2010s Abramitzky, Boustan, and Eriksson (2012, 2014) developed a more powerful variant of Ferrie's algorithm, scaled up to involve the automated linkage of digitized complete-count census data. Subsequent revisions of their method, as well as the release of higher-quality digitized census data, have allowed Abramitzky and Boustan in their most recent work to produce 36 "crosswalks" with millions of males linked across ten censuses (1850-1940). Over the past decade or so, variants of the methods used by ABE have generated research on topics as varied as the economic impact of public policies, the socio-economic progress of African Americans, and the return to schooling. The most popular use of automated census links, however, has been for research on the occupational and geographical mobility of Americans, both native and immigrant (e.g. ABE 2012, 2014; Alexander and Ward 2018; Connor 2019; Pérez 2019; Beck Knudsen 2019; Connor and Storper, 2020).

While automated linkage algorithms mark a huge step forward in efforts to measure socio-economic mobility, they have not been immune to criticism. The primary concern has been the allegedly high rate of false positives (Massey 2017; Ruggles *et al*. 2018; Bailey *et al*. 2020). Fearing that the false-positive rate may be

intolerably high, Ruggles *et al.* (2018), noting that new studies using the ABE links are now "appearing virtually every week," announced that they plan to introduce an alternative linking algorithm that produces fewer false positives. Bailey *et al.* also worry about the likely biases that false positives introduce into the findings of studies based on automated matching. Establishing the prevalence of false positives is not straightforward, however, given the lack of what Bailey *et al.* (2020, p. 998-9) term "ground truth data." In this paper we offer a case-study which pits high quality hand-linked data approximating "ground truth" against automatically linked data for insight into the prevalence of false positives.

THE EMIGRANT SAVINGS BANK AND THE SEARCH FOR "GROUND TRUTH"

One of the greatest challenges for those who wish to evaluate the reliability of automated census linking is that it is not easy to "prove" that a link generated by an algorithm is a false one. It is exceedingly unlikely that a lawyer named John Scanlon, age fifty-four, who owned his own home in Brooklyn in 1860 is (as the ABE algorithm claims) the John Scanlon in the 1870 census who gives his age as age sixty-three and is propertyless day laborer in Scranton, Pennsylvania. But how can we be sure? We take two different approaches to answering this question: First, we use the unique records of the Emigrant Savings Bank in conjunction with census and other genealogical resources to create hand links whose accuracy is much more reliable than those that can be made using the census alone. We then compare the mobility data generated by those links with those of Irish immigrants linked by the ABE algorithm. Second, we demonstrate that it is not actually very hard to prove that many algorithm-generated links are false—not hard, at least, for a professional genealogist such as the one who is part of our project. She found, for example, that John Scanlon the 1870 day laborer was also in Scranton and also a day laborer in 1860 (albeit with his age rounded to fifty), thereby *proving* that the ABE link of the Scanlon the lawyer to Scanlon the day laborer is a false one. It was her analysis of a sample of ABE links that led us to our estimation of the rate of false positives for Irish immigrants.

The ESB records are so useful for those who wish to track ordinary Americans over time because the bank went to extraordinary lengths, in an age before

government-issued photo-identification, to protect its customers' money. To that end, bank officials created "test books," ledgers in which they compiled a wealth of personal information about all depositors, including their address; occupation; townland, parish, and Irish county of birth for Irish-born depositors; the name of the ship that carried them to America and the date of its arrival; their parents' names (including mother's maiden name) and whereabouts; their siblings names and whereabouts; their spouse's names (including wife's maiden name); and their children's names. Then, when people visited the bank and asked to withdraw money, a bank employee would "test" their identity by asking them their mother's maiden name or which of their sisters still lived in Ireland. The bank would periodically update this information, noting new addresses and occupations, spousal deaths, remarriages, and the like. It is not clear how many would-be embezzlers were thwarted by the bank's unusual methods, but the resulting test books are a bonanza for Irish Americans wishing to learn forgotten parts of their family histories (which explains why the test books were digitized and made available on Ancestry.com). They are also a goldmine for those who wish to accurately trace Irish immigrants' lives in the United States.

One might imagine that these records would only help track immigrants who remained in New York, but information in the bank records also facilitates finding the immigrants who left New York. In most cases, for example, it would be impossible to trace New Yorker Peter Lynch from 1850 to 1860, given that in the 1860 census there are 123 Irish-born Peter Lynches, dozens of whom are about the right age. But the test books list the names and birth order of Lynch's five brothers and sisters, and the 1885 Minnesota state census lists a Peter Lynch living in the town of Faxon with five siblings whose names and birth order exactly match those of the bank customer. One can use that information to confirm that the Peter Lynch in Faxon in the 1860 census is the New York Peter Lynch from 1850. Michael Egan, a bank customer who also lived in New York in 1850, was traced in a similar but even more circuitous manner. There are 110 Irish-born Michael Egans and Eagans of about the right age in the 1860 census. But when one enters Michael's name into a genealogical search engine along with his wife's maiden name, Ellen Carey (found in the bank records), up pop two death records from the mid-twentieth century of Minnesotans whose parents had those

exact names. This eventually allows us to determine that the Michael Egan found in the 1860 census in, of all places, Faxon, Minnesota is the New York Michael Egan from 1850. And we make this link even though Michael Egan was only a customer of the bank for five months. Our ESB longitudinal database is thus comprised of people who were bank depositors *at some point*, typically not more than a couple of years, rather than of people who remained customers of a bank in New York for a long time.

The reason that the ESB is such a rich resource for research on Irish immigrants and not others[2] is that the institution was created by Irish-American philanthropists in 1850 specifically to provide a safe haven for the savings of the Irish refugees fleeing the Great Famine (Casey 2006, 2013; Anbinder 2012; Ó Gráda 2002). Anyone could open an account at the bank, and there were significant numbers of German, British, and native-born account holders. But while Irish immigrants made up a quarter of the city's population when the bank opened, they comprised 71 percent of the bank's depositors customers in the 1850s. By the end of the decade, more than fifteen thousand people had opened accounts in the bank's offices at 51 Chambers Street directly behind New York's City Hall in lower Manhattan.

Scrutiny of the bank's records reveals that its Irish depositors spanned the spectrum from destitute assisted immigrants to the cream of New York Irish society. Fourteen of those first 15,000 opened their accounts with the minimum deposit of one dollar, equivalent to an unskilled worker's daily wage; and 659 of the 11,147 Irish-born depositors in our database made an initial deposit of ten dollars or less. Still, we need a firmer sense of how typical its customers were. In terms of occupations, the account holders mirrored the New York's Irish population pretty well. Table 1 compares male Irish-born ESB account holders who were living in New York when they opened their accounts and a one-in-ten sample of New York City Irish males taken from the 1855 state census, divided into six broad categories. The professionals were mainly physicians, lawyers, and the like. It should be noted that many of those characterized

[2] See e.g. Ó Gráda 2000; Wegge, Anbinder, and Ó Gráda 2017; Anbinder, Ó Gráda, and Wegge 2019; Anbinder, Ó Gráda, and Wegge 2020. Kelly and Ó Gráda (2000) and Ó Gráda and White (2003) use the records to address the issue of bank panics.

as "business owners" were people of modest means. They were grocers, saloonkeepers, druggists, and the like. Most of the "lower-status white collar" workers were salesman, clerks, overseers, teachers, civil servants, and the like. The "skilled" category is composed mainly of craftsmen such as carpenters, coopers, compositors, masons, and butchers. They are a heterogeneous group; some such as shoemakers and tailors were under severe pressure from automation at this time, while a minority doubled up as manufacturers and store owners. Those immigrants we have classified as "petty entrepreneurs" (pedlars, hucksters, junk dealers, fruit-stand operators, lodging house keepers, and so on) lived in even more precarious circumstances. While petty entrepreneurs and business owners form somewhat higher shares of ESB customers than the labor force as a whole, it is the preponderance of workers, skilled and unskilled, in both sets of data is most striking. The small "others" category consists of those difficult to classify, those with no declared occupation, and those New Yorkers who described themselves as "farmers" in 1855. On the whole, the occupational distributions are similar, but with the distribution of account holders skewed slightly towards business and white-collar workers and away from those in the lowest-paying jobs the city had to offer (compare Alter, Goldin and Rotella 1994). Three-quarters of the immigrant savers had arrived in America in 1846 or later.

Given that savings banks in Ireland catered disproportionately to the lower-middle and middle classes in the bigger towns and cities, few of the ESB customers—who lived overwhelmingly in rural Ireland before coming to America—are likely to have been institutional savers before they emigrated (Ó Gráda 2003). But the savings habit was widespread in the US in the 1850s, and we know that the New York Irish were also enthusiastic savers. Over the course of the 1850s, about 11,000 Irish-born residents of New York City opened accounts at the ESB, a number equal to nearly 8 percent of the city's adult Irish-born population in 1855. Allowing for marriages, perhaps one in nine Irish immigrants were ESB depositors or married to one.

Our project involves tracing the Irish-born customers of the ESB over the course of their lifetimes, and we have employment data on many of the immigrants for thirty, forty, or even fifty years. We use not only federal censuses, but also state population tallies, newspapers, city directories, military enlistment and pension

records, death registries, and probate records. This work requires the skills of an experienced genealogist familiar with economic and social history of mid-nineteenth-century New York, and our project genealogist, Janet Wilkinson Schwartz, fits that description. But in order to use our data to evaluate the accuracy of automated linking algorithms, we had to create a separate database of ESB customers who we found in *consecutive* censuses. Given that all the depositors in our ESB database had arrived in America before the end of 1858, we decided to include in our comparison database only bank customers who we had found in both the 1860 and 1870 censuses. With Schwartz's help, we have managed to identify 947 (for now) of our 6,574 Irish-born male account holders in the censuses of both 1860 and 1870. Another seventeen hundred were found in a census from 1860 or earlier but not 1870, while several hundred more were located in the 1870 census or later but not earlier.

Most of the immigrants found in the 1860 census but missing from the 1870 tally had died during the 1860s, but this was not always the case. Thomas Boran, for example, returned to county Kilkenny and got married there in 1865. Armagh-born Thomas Abbott, a blacksmith living in New York's Ward Five in 1860, could not be located in 1870 but was found, still shoeing horses and still in Ward Five, in the 1880 census. Abbott probably still lived in Ward Five in 1870 but was either skipped by the census taker or had his name so badly recorded that he could not be located. Other immigrants can be tracked through means other than the census. Depositor Timothy Canty from west Cork had already left New York for California by the time a census taker found him in San Francisco in 1860. He could not be located in another census, but state voter registration records and San Francisco's city directory document that he remained there, operating his own tailoring business until his death, which was reported in the city press, in 1882.

One of the clear advantages of the manual method of matching historical records for individuals, so heavily reliant on the skills of the trained genealogist, is that it yields very few false positives. Moreover, as will be clear from the examples given below, it detects many links which would be beyond the reach of the automated linkage algorithms currently in use. Genealogists use a range of information in the census and elsewhere-such as the names of parents, siblings, spouses, and children—

to make matches; linking on recorded ages and place of birth alone often brings poor results. For example, let's say one wants to track Patrick Meagher, who was a twenty-eight-year-old New York gas fitter in 1860. If there was a Patrick Meagher age thirty-eight found in the 1870 census in Chicago, and no one else of that name aged thirty-six, thirty-seven, thirty-eight, thirty-nine, or forty anywhere else in the 1870 census, every currently used census-linking algorithm would consider that a match. But a genealogist would take into account that in 1860, Patrick had a wife named Bridget and three children—Michael age five, Catherine age three, and Mary age one. Seeing that the Patrick in Chicago in 1870 had no wife or kids, but a Patrick Meagher in New York listed as thirty-five years old had a wife named Bridget and five children, the oldest of whom were Michael age fifteen, Kate age thirteen, and Mary age eleven, the genealogist would declare this Patrick to be the match of the 1860 Patrick Meagher. Siblings and in-laws can play a similar role as children in distinguishing true from false matches. These same tools would enable the genealogist to match Patrick Meagher of 1860 to a Patrick Maher or Mahar of 1870. City directory listings and information in marriage and death records also enable genealogists to distinguish true matches from false ones when age and name variations would otherwise make such identifications impossible. Most of these records are posted online nowadays on websites such as Ancestry.com and Familysearch.org.

Here are a few of the many instances among the ESB's depositors where accurate matches could only be made using such methods:

a) William Singleton, account number 3,288: He was recorded as aged 37 in 1860, 25 in 1870, and 31 in 1880. William was a harness maker in 1860 and 1870, but a laborer in 1880. His wife Anne's ages were recorded as 34 in 1860, 46 in 1870, and 50 in 1880.

b) Thomas Kiernan, account number 15,335: Censuses list him as 25 in 1855, 35 in 1860, and 35 in 1870. His wife's recorded ages were 28, 32, and 40; they had no children. In reality, Kiernan was probably 45 in 1870. Opening his ESB account with $70, he held nearly $700 (equal to more than $20,000 in 2021) in the bank at one time.

c) Matthew Quirk, account number 11,102: listed as 30 in 1855, 40 in 1860, and 35 in 1870. Quirk's wife Margaret was recorded in 1855 as aged 25, with children Mary, 3, and Thomas, 10 months. Five years later she was recorded as aged 30, with Mary, 7; Thomas, 5; Ellen, 3; and Margaret, 1. In 1870 she was still listed 30, with Mary, 17; Thomas, 15; Ellen, 13; Margaret, 10 and Jennie, 8.

d) Peter Duggan, account number 3,466: Censuses record his age as 43 in 1855, 25 in 1860, 65 in 1870, and 60 in 1880. In 1855 Peter and his wife Mary lived in the Sixth Ward with their children Michael 14, Charles 12, Mary 9, Dennis 6. Those children allowed the Duggans to be traced despite the erratic recording of Peter's age. Despite his menial occupations—laborer to 1870, junkman by 1880—Peter at one point held over $1,100 in the ESB.

e) James Devanney from Donegal, account 14,364: Documents spell his last name as Devanny, Deviney, and Deveney before James finally settling on Devine. To make matters more confusing, the bank spelled it Devanney and then Deveny. His wife's first name also changed. James told the bank in the 1850s that he was married to Celia, but she was listed as Mary in the 1870 census, and from 1880 went by Sydney. But contemporary records prove that Celia, Mary, and Sydney were all the same person. James was listed as 37 years old in the 1860 census but as 30 in the 1870 census. He told the bank that he was born in 1824, which tallies better with a death notice stating that he was 73 when he died in 1899. The couple could be tracked because of their children's names, especially son Dominick, a rare name among New York's Irish. Dominick was named after his maternal grandfather, Dominick Doherty.

COMPARING THE ABE AND ESB DATABASES

Customers of the ESB lived not only in New York, but also in Brooklyn (then a separate city), other parts of Long Island, New Jersey, and the remainder of New York State—its reputation as a safe repository for the savings of Irish immigrants earned it customers far and wide. In order to make a like-to-like comparison with ABE links, however, we chose to consider for this part of our analysis only male ESB customers

men who lived in New York City or Brooklyn in 1860. We limited our focus to men since women are much harder for an algorithm to trace because their surnames change when they marry. [3]

We compare our ESB links to those formed by the two main versions of the ABE algorithm, on which a considerable body of research already rests. In the standard version (ABE "exact"), if the algorithm finds only one person with a certain name and birth year in one census, and then finds only one person with that name and birth year in the second census, this is considered a match. If there is no exact age match, the algorithm looks for someone in the second census who is either nine or eleven years older than the person it is attempting to match. If there is only one such person, this is considered a match. If there is still no match, it tries one more time with people of that name either eight or twelve years older than the person being searched. If there is no unique match, then this person from the 1860 census is eliminated from consideration. The names from census to census do not need to match exactly if the variations are deemed insignificant or are standard abbreviations. In the "conservative" variant of the ABE algorithm, a surname-and-given-name combination must be unique within a five-year age window. In what follows, we focus mainly on the "conservative" variant, but also report some results using the "exact" variant, on which earlier studies rely (Abramitsky, Boustan, and Eriksson 2012, 2014; Ager, Boustan, and Erikkson 2019; Abramitsky 2020; Abramitsky *et al.* 2021).[4] Of 97,573 Irish-born men aged 18 to 64 in 1860 who were recorded in the census of that year as living in Manhattan (then the entirety of New York City) and Brooklyn, the ABE exact method matches 9,691 of them (9.9 percent) to males with the same name living somewhere in the United States in the 1870 census. The ABE conservative method matches 3,803, or 3.9 percent, of the same group.

[3] Unmarried women who later wed are difficult for a genealogist to trace as well, but the information in the bank records has allowed us to track many more of them than would normally be possible for humans or machines.

[4] Note that according to the 1860 U.S. census, there were 64 Bridget Lynchs living in New York City, 87 Bridget Ryans, 134 Bridget Murphy/Murpheys, and 146 Bridget Kelleys/Kellys. And there were three to four times as many Marys with each of these surnames.

CONTRASTING RESULTS

The contrast between the geographic and occupational mobility rates for New York's Irish immigrants as measured by the ABE algorithms and our hand links could not be more stark. To measure class mobility, Table 2 invokes the occupational classification scheme already used in Table 1. We compared the two ABE linkage variants to the "hand links" of ESB customers from the same decade done by our genealogist. The results, found in Table 2, provide a like-for-like comparison between ABE-linked data and the hand-linked ESB customers.

We expected that positive selection might lead our ESB data to differ from ABE's, but not nearly to the extent described. The outcome suggested by hand linking is of significant persistence in all occupational categories between 1860 and 1870. This is reflected in the percentages in the diagonal of Panel 1, ranging from 54 per cent in "Lower Status White Collar" to 79 per cent in "Skilled." The outcomes using either version of ABE are in stark contrast, with only the "Unskilled" category showing persistence. The implications of the algorithm-generated data for upward and downward occupational mobility seem far-fetched as well: ABE Exact (Panel 2) has 30 per cent of those classified as "Professional" in 1860 descending to "Unskilled" by 1870, while ABE Conservative (Panel 3) consigns 44 per cent of those in "Business" in 1860 to the "Unskilled" a decade later.

Table 3 compares what the ABE and hand links approaches predict for the rates at which the New York Irish changed location over the decade. Hand linking suggests that 7 percent of the ESB's customers changed their state of residence in the 1860s, and that 18 per cent changed counties. This may seem somewhat lower than what one would expect for the general population, but the rate of geographic movement of those traced by the ABE links seems much more out of line, with half to three quarters changing county or state over the decade.

As a further "reality check" of sorts, we augmented the ABE-generated data to see if the linked individuals had spouses and, if so, whether or not that spouse's name was the same in the 1860 and 1870 censuses. When we discovered that requiring precise first name matches for the spouse excluded many good links, we modified the

rule so that only the first four letters of the first name had to match and making allowances for abbreviations like Maggie, Lizzie, Kate, etc. For the vast majority of Irish-born people then and much later, divorce was not an option. Irish immigrants may have died at a slightly higher rate than other Americans in this era, yet it is plain from municipal death records that two-thirds to three-quarters of New York's Irish-born men cannot have lost spouses in a ten-year timespan (see Appendix 3). The 2 per cent rate of remarriage implicit in our ESB customer database—that is where the wives' names unambiguously differ--is a tiny fraction of that generated by the both ABE conservative and "exact" variations. Given that our genealogist uses wives' names to help confirm links, there are undoubtedly widowers whom she cannot identify with certainty (though if the first marriage produced several children, then the married man of 1860 can often be found with his children in 1870 even if he has remarried). Nonetheless, this inconceivable rate of remarriage in ABE links is another indication of large numbers of false positives. The ESB data base produces a starkly contrasting picture: less than two per cent of male account holders with an identified spouse in 1860 were married to a different woman in 1870.

The relationship between spouse names and geographic persistence provides further evidence that the ABE Irish links must contain many false positives. Table 4 illustrates this relationship. The immigrants who are married to the same spouse in both 1860 and 1870 (and thus most likely to be a valid link) remain primarily in the same state from census to census, while those listed as married to a different spouse are mostly found in different states. Even if we can imagine that a widower might leave behind his support network and move with his kids to a new state to start over fresh, the rates for such behavior found in Table 4 are not credible. This again implies a massive misidentification of males in the ABE matched sample. An added indication that there are many false positives in the ABE links is that correlation between the occupational classes in 1860 and 1870 are high for those who were found in the same location in 1870 as in 1860 but negligible for those who supposedly migrated. The contrasting outcomes for stayers and movers are given in Table 5, which invokes an occupational classification scheme, HISCO/HISCLASS, which derived its inspiration from the ILO's International Standard Classification of Occupations (ISCO). The

scheme, derived by Dutch scholars (van Leeuwen, Maas, and Miles 2002; van Leeuwen and Maas 2011), has been widely invoked by economic historians (e.g. Breschi *et al*. 2014; Dribe *et al*. 2014; Humphries and Weisdorf 2016; Vickers and Ziebarth 2016; Bengtsson *et al*. 2018; Connor 2019). The scheme ranks occupations from professional, managerial, white-collar (with HISCO values of up to 30000), through farming, skilled, commercial, and artisanal occupations (with values between 30001 and 89999), and unskilled occupations (with values of 90000 and above) who moved during the decade. These loosely represent the upper, mid, and lower steps on the occupational ladder. Table 5 shows that correlations between HISCO_1860 and HISCO_1870 are consistently high for those who stay put whereas they are negligible for those who migrate, an added indication that there is something amiss with the location of 1870 matches.[5]

What of the possibility that those not linked by hand differed systematically from those who were? The tables in Appendix 2 show that while such biases were present, they were not very powerful. Those linked were more likely to be professionals or businessmen, while the less skilled were more likely to be lost. Our hand links also show that those who left New York did have a bit more upward mobility (attributable primarily to the propensity of these men to become farmers) than those who remained in New York, but that overall there was still a strong relationship between one's occupation in 1860 and that followed in their new location. The ABE links would have us believe there was virtually no relationship at all, a finding which is simply not credible. Put another way, the ABE-linked Irish immigrants who supposedly move hardly ever stay in the same occupation when they do so. This is even the case for workers in hard-to-learn, highly sought-after, well-paid trades like baking, butchering, masonry, plumbing, printing, and stonecutting. It

[5] Similarly, the correlation between the declared real property of New York or Kings County residents who remained in those counties over the decade is significant (0.335, N=3,392), as is that of residents of those counties who stayed within New York state (0.281, N=5,431), while the correlation for those who changed state are close to zero (0.029, N=11,637).

is not plausible that 90 to 100 percent of the workers in all these trades who left New York would have abandoned them, as the ABE links would have us believe.

IRISHMEN OF AN UNCERTAIN AGE[6]

Why do the ABE links of Irish immigrants apparently contain so many Type I errors? The erratic recording of ages noted above in the discussion of how our genealogist makes links pointed us to the first reason why automated census linking produces so many false positives for immigrants: the prevalence of age heaping in census entries for the foreign born. Measures of age-heaping have long been used by social scientists and historians as a proxy for numeracy or cognitive ability. The most common measure of age-heaping is the Whipple Index, favored by United Nations demographers since the 1970s. The Whipple value "is obtained by summing the age returns between 23 and 62 years inclusive and finding what percentage is borne by the sum of the returns of years ending with 5 and 0 to one-fifth of the total sum."[7] The value of the index can range from 100 (no age heaping) to 500 (complete age-heaping). The *UN Demographic Yearbook* has proposed: Highly Accurate data [WI<105]; Approximate data [110-124.9]; Rough data [125-174.9]; Very Rough data [175+]. In poor economies, Whipple values are typically high, particularly for females.

In a pioneering comparative study of age heaping in many countries over several centuries, A'Hearn *et al*. (2009, p. 792) found that census data from mid-nineteenth century Ireland produce very high Whipple scores.[8] Yet he found that Irish

[6] With a bow to Blum *et al.* 2017.

[7] J. T. Marten, *Census of India, 1921*, vol.1, part 1 (Calcutta, 1924), pp. 126-127 as cited in United Nations (2017).

[8] In a letter to William Farr, chief architect of the English census of 1841, his Irish counterpart Thomas Larcom described how "it was by no means unusual for the country people more especially, and in the higher ages of life more so still, to call themselves by the nearest ten or five, whether they were above or below it." Larcom published the data on ages for single years

Americans in the same period had even higher Whipple values and concluded that the Irish who came to America must have had "*extremely* low levels of age numeracy" (italics in the original). Census returns for ESB customers appear to corroborate A'Hearn's findings. Irish-born ESB depositors yield extraordinarily high Whipple values by any comparative standard (see Table 6); they match or exceed those inferred from the Irish population census of 1841. The bank customers are no different in this regard than other Irish-born New Yorkers, whose Whipple values place them in the "Very Rough" range.[9] Educational attainment clearly played a role in age heaping. In New York on the eve of the Civil War, ESB customers who were managerial and professional workers yielded lower Whipple values than their blue-collar counterparts. But why should age heaping among the New York Irish in general be higher than among the Irish at home?[10]

Beginning in the late 1850s, the ESB asked depositors to sign their names in the test books when they opened an account, and this data is far more reliable than self-reported reading and writing ability in measuring literacy rates. The ESB test

---

in the census in order to show that although "The ordinates representing these ages are at first sight … formidably irregular … a close inspection will show that the irregularities follow a very constant law, and when reduced to an equated line, exhibit a curve very consistent with the results of established Age Tables" (BPP 1843: xlvi; National Library of Ireland, Ms. 7526, Larcom to Farr, 21 October 1844).In Ireland the authorities intended household heads to fill the forms which were to be collected by an enumerator on the following day; in practice, however, enumerators must have frequently assisted "such persons as may not be able to fill the forms themselves" (BPP 1843, pp. v, xci). On the basis of analyses of workhouse and prison records, Blum *et al*. (2017) suggest that Irishwomen's numeracy was 'hugely overestimated' by the census, perhaps because husbands filled in their wives ages on the family enumeration forms.

[9] The correlations between WI values and Irish ward population shares were very high, as shown in Table A1.1. The tendency for WI values to increase over time is mainly a reflection of the ageing of the population: people tended to heap more around ages 50 and 60 years than ages 30 and 40.

[10] See, e.g. Ahearn, Baten, and Crayen 2009; Crayen and Baten 2010; Baten and Crayen 2010; Baten, Crayen, and Voth 2014; De Moor and van Zanden 2010; Blum *et al*. 2017.

books indicate that 63 percent of adult female Irish-born New Yorkers could not write their names, versus only 20 percent of Irish American men. Yet while innumeracy ought to correlate with illiteracy, Irish-born men in New York are *more likely* to be age-heaped than women. Such outcomes add to the evidence that there is more at play in age heaping than relative numeracy.

The other source of age-heaping, which the literature tends to ignore or downplay, is shoddy enumeration on the part of census marshals. Indeed, the instructions given to census enumerators allowed for age "approximation." Why would census enumerators have taken advantage of this time-saving shortcut with Irish New Yorkers more than others? Perhaps fearing crime or disease, census marshals might have been afraid of spending too much time in Irish tenements and more likely to guess at ages in order to complete their visits as quickly as possible. Prejudice may also have been involved. Perhaps the Irish—stereotyped as degraded, drunken brutes—were not perceived to *deserve* the careful consideration given to other groups. Eventually, the state would demand a more accurate accounting of its citizens. A'Hearn *et al*. (2019) ascribe the sharp drop in age heaping in southern Italy between 1881 and 1901 to "the state's increased allocation of resources to census operations, its enhanced technical competence, its increasing success in overcoming the suspicions and enlisting the cooperation of its citizens, and its growing ability to monitor and control the actions of local government." Age heaping declined in the American censuses in the same period, probably for the very same reasons.[11]

Researchers are aware of the problem age heaping causes for census linking (Ferrie 1999, p. 22; Cirenza 2011, p. 55, 68; Bailey *et al*. 2020, p. 1001), but perhaps not sufficiently so. Table 7 shows the spread in age differences in more detail for these two censuses as well as 1850 to 1860. The "All" columns report percentages with age differences outside a five-year band centred on 10 (italicized in bold). Note that 46.5

[11] A highlight is the stark contrast between the age heaping of the Italian-born in the US in 1910 as recorded in the census (150 for males, 149 for females) and of Ellis Island arrivals from Italy in 1898-1912 (99 and 105). In the latter case the data were assembled on board and "sloppiness was extraordinarily rare" (A'hearn, Delfino, and Nuvolari 2019, Table 4). For an earlier example of disregard for administrative sloppiness see Mokyr and Ó Gráda (1982).

per cent of all potential links would be under the radar of an algorithm using a 5-year age band in 1860-70, and 52.2 per cent of links if age is heaped in 1860.

This age heaping would not be a problem if it merely led to Type II errors, in which potential links were left unidentified. What happens instead, however, is that if a Patrick Connor in New York is asked his exact age in 1860 by a conscientious census taker, but his lazy successor in 1870 estimates Connor's age, while an enumerator encountering a Patrick Connor in Texas is lazy in 1860 but diligent in 1870, then the algorithm can be led to believe that the New York Connor moved to Texas and the Texas Conner relocated to New York, when in fact neither man moved at all. This may be less of a problem in the twentieth century when census takers had to record each resident's year of birth. But in the nineteenth century, when enumerators only had to record an age and estimation was permitted, the problem was pervasive and it leads to tens of thousands of Type I errors for Irish immigrants in the 1860 to 1870 ABE crosswalk.

NAME VARIATIONS

As the example of Archbishop Purcell made clear, variations in name spelling are the other main cause of false positives generated by the ABE algorithm. This fact became clear when we subjected our ESB links to scrutiny by the ABE algorithm. We identified 102 Emigrant Bank account holders in the 1860 census who were matched to people in 1870 by both manual and ABE methods. Only in 56 did the matches concur. In about half of the instances in which our genealogist determined that the ABE algorithm made a bad match, the age difference in the true match was outside the range of 8-to-12 years permitted by ABE. In most of the others cases, the problem that had tripped up the algorithm was a variation in surname spelling.

Some of these name-spelling variations are so extreme that no automated system will likely ever be able to identify them. For example, when reading the entry of a census enumerator from 1870 with messy handwriting, a modern transcriber recorded Robert Baxter's name as "Robert Bartoo," leading the ABE algorithm to link Baxter, a Michigan brass finisher and former ESB customer, to a book binder named Robert Baxter in New York rather than the correct Robert Baxter, who was still a brass

finisher and still living in the same town in Michigan in 1870. In a more typical case, the ABE algorithm incorrectly concludes that the New York chairmaker Hugh Donohoe from the 1860 census is Hugh Donohoe the Minnesota farmer in the 1870 census because in that latter year, the enumerator in New York spelled Hugh's surname as "Donahue" (Table 8).

In many cases, both a name spelling variation and age heaping prompt the ABE algorithm to make an erroneous match. The algorithm believes, for example, that ESB customer Lawrence Fleming, a New York day laborer living with wife Jane and son Edward in 1860, had moved to Pennsylvania by 1870, become a carter, remarried a Catherine, and was now childless. In fact, Fleming was still in New York, still a day laborer, still married to Jane, and still had a son named Edward, but the ABE method cannot make this link because the sloppy New York census enumerator in 1870 recorded Fleming's age as thirty when in fact he was nine years older. But even if that census taker had correctly documented Fleming's age, the ABE algorithm would have still made the same bad link because in 1870 that careless New York census marshal spelled the immigrant's name as "Flemming," and the ABE method only links Flemings to Flemings and Flemmings to Flemmings. Between the spelling and age variations, most of the 117 Irish-born Flemings and Flemmings the ABE algorithm links from 1860 to 1870 are errors. The same result is found with the many other Irish surnames commonly spelled in more than one way, such as Burns/Byrnes, Conner/Connor, Eagan/Egan, Maher/Meagher, O'Neil/O'Neal/O'Neill, Quin/Quinn, and Riley/Reilly, just to name a few. In all the cases in which the ABE algorithm makes an erroneous link of an ESB customer, both versions of the algorithm should have eliminated these people from consideration because in each case there were two people with the same name of a very similar age. But because of inconsistency in the spelling of their names and the recording of their ages, the ABE method made erroneous matches instead.

A CASE STUDY: CLERGYMEN, DOCTORS, AND LAWYERS

Abramitzky *et al*. claim that their automated matching algorithms typically "generate very low (less than 5%) false positive rates" (2020, abstract), yet our analysis

thus far of the links created by the ABE method for Irish immigrants in the US in the 1860s suggests a far higher error rate. In order to test our suspicion that the ABE method creates more false positives than has been acknowledged, we conducted a case study using the Irish-born Catholic priests found in the 1860 census whom the ABE algorithm match in the 1870 census. Catholic clergymen are a particularly interesting group to consider, since their calling was almost invariably a life-long one. We do not discount the occasional possibility of a priest being defrocked, but in the mid-nineteenth century the mantra of "once a priest, always a priest" rang true. Furthermore, priests create a much larger paper trail—in parish histories and news accounts of their lives (and deaths)—than the average citizen, making it fairly easy to determine with certainty whether or not the ABE links for these priests are accurate.

The ABE-exact algorithm matches 70 Irish-born Catholic priests (once some transcription errors[12] are corrected) found in the 1860 US census to individuals in the 1870 tally. Of those 70, according to the ABE-exact algorithm, only 26 (37 percent) were still priests in 1870. Even if one chooses the ABE conservative algorithm, the percentage of priests who supposedly remained in that line of work ten years later increases only to 48 percent (14 out of 29). In collaboration with our genealogist, we closely investigated the 44 who had supposedly left the priesthood and found that 43 of the 44 matches were demonstrably erroneous.[13] Of these 43, seven could be shown to have died before 1870. Twelve of the supposed former priests had children in 1870 who had been born at times and in places indicating their father could not possibly be the same person as the priest of the same name from 1860. According to the algorithm, for example, Wisconsin priest George Brennan in 1870 had become a

[12] For example, one linked immigrant whose occupation in 1860 was transcribed as "RC paster" and coded as a menial worker was clearly a Roman Catholic pastor and has been included with those whose occupations as Catholic clergymen were accurately transcribed. In other cases, an immigrant is listed only as a "clergyman" and it takes some research to determine if they were Catholic clergymen.

[13] In the forty-fourth case, even though we could not prove that Father John Cassey of California in 1860 was not John Cassey the Philadelphia domestic servant in 1870, the link is certainly erroneous.

leather currier and moved to Massachusetts. But the 1870 census lists the Massachusetts man as having four children born in the Bay State from 1855 to 1861, meaning the leather currier of 1870 was not working as a priest 1,000 miles away in 1860. Furthermore, we can show that twenty of the alleged ex-priests were still clergymen in 1870, in every case in the very state where they had lived in 1860 (Table 9). The remaining links can be proven false because the person linked in 1870 can be found in the 1860 census in an entry different than that for the priest of the same name. Note that nearly all the false matches were to immigrants living in different states than those where the priests were documented to have originally lived.

Still, a sample of 70 is not conclusive, so we expanded our case study to also include Irish-born clergymen who were not Roman Catholics as well as doctors and lawyers. Again, these groups were chosen because they were more likely than the average immigrant to leave a paper trail that would allow us to definitively evaluate the ABE links. The result is a more robust sample of 355 doctors, lawyers, and clergymen. The ABE accuracy rate for this larger group is slightly better than for the priests alone. Still, half of the links for these immigrants formed by the ABE "exact" algorithm (177 of 355) are definitely false positives, while even the conservative ABE method produces a false positive rate of one-third (58 out of 178). The cases of the Irish, clergymen, doctors, and lawyers, to whom we return below, adds direct evidence to the already strong circumstantial case that the ABE method produces many more erroneous links than is generally understood.

DATA QUALITY, HAND LINKING, AND MATCH RATES

Abramitzky *et al*. 2020 have subjected their automated linking algorithm to a variety of tests against hand links. In one exercise comparing the results obtained from automated linking of the entire 1910 and 1920 US censuses to links from Familysearch.org's "Family Tree" data they report that their method "generates very similar results" in terms of false positives—about a 5 percent error rate for their "exact" method and 3 percent for their conservative method (Abramitzky *et al*. 2020, pp. 6, 18-19). Why is it, then, that the false-positive rates we find differ so markedly from what they report in this exercise? The answer probably lies in part in the relative

quality of the underlying data, which is likely to have differed not only over time, but across countries. We have highlighted how age misreporting and the inconsistent spelling of surnames led to both false positives and missed links. We suspect that these issues matter more for data from the second half of the nineteenth century than those from the first half of the twentieth. If that is so, then the older census records are likely to generate more Type I and Type II errors than the latter. Algorithms such as ABE rely on identifiers that should not change over time such as place of birth, birth year, surname, and gender. But as the quality of the underlying data improved over time, the variation in *the gap* in ages between censuses and in spelling errors decreased, leading to better matching. This probably helps to account for the very low match rates found by Ferrie and by ourselves relative to those claimed by ABE and others in different historical contexts.

As noted earlier, the original ABE algorithm matched 9.9 per cent of Irish-born aged 18-64 living in New York and Kings counties in 1860. That is in the same ballpark as the 10.6 percent achieved by Ferrie (1999, p. 22) for immigrants matched in the 1850 and 1860 US censuses, but much lower than the 20 percent matched by Ager *et al*. (2019: 9-10), the 19 percent by Collins and Wanamaker (2014), the 21 percent by Connor and Storper (2020) using later US census data, and the 17 percent matched by Long and Ferrie (2018: F426-7), using British data. For 1880-1910 cohorts of U.S. immigrants from 17 countries, Abramitsky, Boustan, Jácome, and Pérez (2019) find match rates ranging from 15.9 and 27.7 percent, and for 1910-1940 from 20.9 to 34.3 per cent. Such differences and any potential biases they might cause are also of potential interest, particularly if the prevalence of false positives is a function of the match rate (compare Pérez 2019).

Our genealogist, in contrast, has produced a much higher match rate—she has found 54 percent of Irish ESB customers located in the 1860 census in the 1870 count.[14] The ABE exact algorithm purports to find 11 percent of those same immigrants from the 1860 census in 1870. The genealogist's higher linkage rate results primarily from the fact that she can use information other than censuses—such as directory listings, naturalization, birth, and death records—as well as census information on the

[14] Compare Kosack and Ward (2020, 969-970).

depositors' family members to confirm or eliminate potential matches. The poorer quality of the earlier data enhances the value of hand linking.

Perhaps another reason that ABE underestimate their rate of false positives is that there are deficiencies in the design of the tests they use to compare their results to those attainable by humans. In most of their tests, ABE limit their human testers to the same data points that the algorithm has at its disposal—name, age, and birthplace. Those tests merely show that man is no better than machine *at doing precisely what ABE ask a machine to do*. Yet professional genealogists—surely the appropriate yardstick—would not go about the problem of linking people across different censuses in the manner ABE prescribe. They use all the data at their disposal and consequently make much better links and produce many fewer false positives than any algorithm can. An under-appreciated factor is that genealogists know much better than an algorithm when not to make a link at all. They can determine that someone found in one census has died before the next one was conducted or that what seems like a very unusual name is really an enumerator's misspelling of a very common one. The ABE contention that humans in some tests have produced 25 per cent false positives (2020: 6, 37) really only indicates that either the rules of their test are deficient or the wrong humans are being used.

DO ABE FALSE POSITIVES AFFECT RESULTING OUTCOMES?

As Abramitzky *et al.* (2021) rightly point out, false positives are inevitable in any automated linking project and do not really matter unless they bias the outcome of analysis for which the data are used. ABE argue that the damage done is small, but our case study of Irish immigrants linked by the ABE algorithm provides evidence that the foreign-born may produce many more false positives than previously understood. This matters a great deal given that studying the social mobility of immigrants is one of the most popular uses of linked-census databases (see Ferrie 1999; Abramitzky, Boustan, and Eriksson 2012; Abramitzky, Boustan, and Eriksson 2014; Pérez 2017, 2019; Collins and Zimran 2019a; Collins and Zimran, 2019b; Connor 2019; Kosack and ward 2020; Aaronson, Davis, and Schulze 2020; Abramitzky, Boustan, Jácome and Pérez, 2021).

To test the extent to which outcomes can be affected by ABE false positives, we return to the occupational and geographic mobility of Irish immigrants in the United States from 1860 to 1870. First, in Table 10 we compare outcomes using the "exact" and conservative ABE algorithms for Irish males aged 18 or more living in New York or Kings counties in 1860 with hand-linked Emigrant Bank account holders, also aged 18 or more and living in the same counties in 1860. The contrasts in outcomes are staggering. Not only does the algorithm return proportions moving out of New York state that are 8 to 9 times as high as the EISB sample: it produces even more improbable contrasts for changing wives over the decade. And, whereas the bank data suggest about a quarter changed occupational category (as defined above), the algorithm suggests that 50-55 per cent did so. Now, undoubtedly, some of these differences are explained by selection in the Emigrant Bank data emanating from both bank depositors being positively selected relative to the New York Irish in general (see Table 1) and from selection in those we have successfully linked (see Appendix 2). Because linked account holders were on average a few years older than those in the ABE databases[15], they were less likely to move states; and because linked bank account holders were somewhat better off than those not linked, they were likely to be more upwardly mobile. But, as explained earlier, the biases emanating from these differences are second order compared to the huge differences between hand-linking and the algorithm, which are primarily due to false positives generated by both versions of the algorithm.

We can do better, however. We can compare the social mobility of the 355 Irish-born clergymen, lawyers, and physicians in 1860 who were tracked to 1870 by the ABE algorithm with genealogist hand linking of the same 355 individuals. This is a genuine "like-to-like" comparison. Table 11 shows that the ABE method generates results that drastically misrepresent the propensity of these immigrants to relocate or change vocations. The ABE Exact links overstate the geographic and occupational mobility of this group by 475 per cent, while even the ABE Conservative variant overstates their propensity to move or change occupational categories by 350 per cent. Furthermore,

[15] For those aged 18 and over in 1860 the averages are: ESB 37.8 years, ABE-exact 34.3 years, ABE-conservative 35.7 years.

there is no question about whose links are correct. In each of the 298 cases out of 355 (84 per cent) in which the genealogist has made a link, there is proof positive that her link is correct; this represents the gold standard that Bailey *et al.* (2020: 998) define as "data obtained by direct observation of the true link."[16]

The ABE algorithm's results are similarly distorted when tracking the Irish immigrants who opened accounts at the Emigrant Savings Bank. When comparing the ABE links of the bank's Irish-born customers to genealogist hand links of only those depositors who ABE link, we find that the automated technique overstates the percentage who changed their states of residence from 1860 to 1870 by 350% using the Conservative version of the algorithm and by 450% with the Exact version. The ABE methods overstate the bank customers' propensity to move up or down among our six occupational categories by about 100% (Table 12). These results suggest that false positives generated by the ABE method of linking significantly distort social mobility analysis.

CONCLUSIONS

Our examination of the reliability of the ABE method for linking Americans between censuses, using mid-nineteenth-century Irish immigrants as a test case, yields several important findings. It shows that even very slight variations in name transcriptions, which are quite common, can play havoc with the efforts of automated census linking algorithms to make accurate matches. Age heaping and age misreporting generally pose another major challenge to automated linking efforts. The propensity to age heap was widespread—at least in the nineteenth century—and significantly impacts the quality of matching results. We find, consequently, that there are more false positives produced using the ABE method than has been recognized. Our case study of Irish doctors, lawyers, and clergymen found that 50 percent of the links created by the ABE-exact method were false positives, and that even the ABE "conservative" variant produced 33 percent false positives. We have no reason to believe that analyses of other subsets of Irish immigrants would produce different results.

[16] Furthermore, in all but a handful of the remaining 57 cases, our genealogist could prove that the ABE link is incorrect even though she could not come up with a verifiable alternative link.

We also found that the false positives generated by automated linking significantly affect the socio-economic outcomes implied by those links. The large number of false positives generated by the ABE method for Irish immigrants linked from 1860 to 1870 produced far too much occupational mobility—both upward and downward. Geographic mobility was even more drastically distorted by false positives.

To what extent such distortions apply to other linkage exercises remains to be seen. We recognize that linking mid-nineteenth-century immigrants may produce more false positives than matching later immigrants or the native born in the ABE system. Abramitzky *et al*. (2021) acknowledge this fact, but we found a much higher rate of false positives than they suggest would occur for foreign-born Americans. Yet we also expected that Irish immigrants would be less difficult for the ABE algorithm to handle than German or Eastern European immigrants—whose names would be difficult for American-born census takers to spell. That was not the case. Instead, the very fact that so many Irish immigrants had the same common given names and surnames made the Irish *more* likely than other Americans to be inaccurately matched. We are thus currently at a juncture where we lack full understanding of the factors that generate false positives, an issue that will continue to bias our findings from economic and demographic research.

We do not contend, however, that automated linking cannot be made more accurate, nor are we proposing professional genealogists as a substitute for the vast potential of computational record linkage. In this spirit, we have been experimenting with a version of the ABE method that requires men linked from one census to the next to have a spouse with the same name in both censues. As Table 13 shows, this variation cuts down the likely false-positive rate dramatically; the proportions predicted to have changed states fall from 64 and 55 per cent to 30 and 21 per cent, respectively, and the proportions changing occupational category also fall significantly. But even this variation still implies more movement than our genealogist's hand links.[17] The striking improvement generated by this tweak to the

[17] We recognize that this spouse-match variation has its limitations—it will not count single men, who constituted 20 percent of Irish-born American men age thirty or older in 1860 and a much higher proportion of those under thirty, and its first name recognition rules need

algorithm prompts the following ecumenical summary: the gap between the strikingly high false-positive rate that we found and the 5 per cent reported by Abramitzky *et al*. stems from three factors: 1) our matched database refers to an earlier, more error-prone crosswalk than theirs; 2) we are testing a more error-prone group, the Irish, who are more prone to age heaping and more name variations; and 3) imperfections in the algorithm itself.

Debates about some of the issues raised in this paper have already prompted a search for other improved automated census-linking systems (Feigenbaum 2016; Massey 2017; Thomas 2018; Bailey *et al*. 2020; Abramitzky, Mill, and Perez 2020; Abramitzky *et al*. 2021). Price *et al*. (2021) have used the family-tree links posted on Familysearch.org in an attempt to "train" computers to make more accurate census matches (Price *et al.* 2021). As part of a collaborative effort at IPUMS, Helgertz *et al.* (2020) are developing a census-linking algorithm that tries to emulate the methods of professional genealogists by considering the names of spouses and children, as well as other characteristics, when determining matches. We hope that by pointing out the surprising degree to which age heaping and name spelling variation can trip up the linking algorithms, we can contribute to improvements in automated census linking that may one day make the data they generate almost as accurate as that of genealogical experts. Doing so would allow us to answer questions central to the work of many social scientists, allowing economists, historians, and others to better understand how mobility and inequality have changed from the past to the present and what those changes may portend for the future.

sharpening. But we feel that through sample weighting or other means, these deficiencies will be remedied to some extent, and that until some better solutions come along, the benefits of using a system that generates only a fraction of the false positives currently being created far outweigh the costs.

REFERENCES

Aaronson, Daniel, Jonathan Davis, and Karl Schulze, 2020. "Internal Immigrant Mobility in the Early 20th Century: Evidence from Galveston, Texas." *Explorations in Economic History* 76 (2020) 101317.

Abramitsky, Ran. 2019. "Historical Record linking" [https://ranabr.people.stanford.edu/matching-codes], 2019.

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102, no. 5 (2012), 1832-56.

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122, no. 3 (2014), 467-506.

Abramitsky, Ran, Leah Boustan, Katherine Eriksson, James J. Feigenbaum, and Santiago Pérez. "Automated Linking of Historical Data." *Journal of Economic Literature*, forthcoming 2021.

Abramitzky, Ran, Leah Boustan, Elisa Jacome, and Santiago Pérez. "Intergenerational Mobility of Immigrants in the United States over Two Centuries." NBER Working Paper 26,408, 2019.

Ager, Philipp, Leah Boustan, and K. Eriksson. "The Intergenerational Effects of a Large Wealth Shock: White Southerners after the Civil War." NBER WP 25,700, 2019.

Abramitzky, Ran, Roy Mill, and Santiago Pérez. "Linking Individuals Across Historical Sources: A Fully Automated Approach." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53, no. 2 (2020), 94-111.

A'Hearn, Brian, Jörg Baten and Dorothee Crayen. "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital. *Journal of Economic History*, 69, no. 3 (2009), 783-808.

A'Hearn, Brian, Alexia Delfino, and Alessandro Nuvolari. "Cognition, Culture and State Capacity: Age-heaping in XIX Italy." CEPR Discussion Paper 14,261, 2019.

Alcorn, Richard S. and Peter R. Knights, "Most Uncommon Bostonians: A Critique of Stephan Thernstrom's The Other Bostonians," *Historical Methods Newsletter*, 8 (1975): 98-114.

Alter, George, Claudia Goldin, and Elyce Rotella. "The Savings of Ordinary Americans: The Philadelphia Saving Fund Society in the Mid-nineteenth Century." *Journal of Economic History* 54, no. 4 (1994), 735-767.

Anbinder, Tyler, "From Famine to Five Points: Lord Lansdowne's Irish Tenants Encounter North America's Most Notorious Slum." *American Historical Review*, 107, no 4 (2002), 350-387.

Anbinder, Tyler. "Moving beyond "Rags to Riches": New York's Irish Famine Immigrants and Their Surprising Savings Accounts. *Journal of American History*, 99, no. 3 (2012), 741-770.

Anbinder, Tyler, C. Ó Gráda, and Simone Wegge. "Networks and Opportunities: A Digital History of Ireland's Great Famine Refugees in New York." *American Historical Review*, 124, no. 5 (2019), 1591-1629.

Bailey, Martha J., Connor Cole, M. Henderson, and C. Massey. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature*, 58, no. 4 (2020), 997-1044.

Baten, Jörg and Dorothee Crayen. "New Evidence and New Methods to Measure Human Capital Inequality Before and During the Industrial Revolution: France and the US in the Seventeenth to Nineteenth Centuries." *Economic History Review* 63, no. 2 (2010), 452-478.

Baten, Jörg, Dorothee Crayen, and Hans-Joachim Voth. "Numeracy and the Impact of High Food Prices in Industrializing Britain, 1780-1850." *Review of Economics and Statistics* 96, no. 3 (2014), 418-430.

Bengtsson, Tommy, Martin Dribe, and Bjoern Eriksson. "Social Class and Excess Mortality in Sweden During the 1918 Influenza Epidemic." *American Journal of Epidemiology* 187, no. 12 (2018), 2568-2576.

Blum, Matthias, Christopher L. Colvin, Laura McAtackney, and Eoin McLaughlin. "Women of an Uncertain Age: Quantifying Human Capital Accumulation in Rural Ireland in the Nineteenth Century," *Economic History Review* 70, no. 1 (2017), 187-223.

BPP [British Parliamentary Papers]. *Report of the Commissioners Appointed to Take the Census of Ireland for the Year 1841*, vol. XXIV (1843).

BPP. *The Census of Ireland for the Year 1851, Part VI: General Report*, vol. XXI (1856).

Breschi, Marco, Stanislao Mazzoni, Massimo Esposito, and Lucia Pozzi. "Fertility Transition and Social Stratification in the Town of Alghero, Sardinia (1866-1935)." *Demographic Research* 30 (2014), 823-852.

Casey, Marion R. "Refractive History: Memory and the Founders of the Emigrant Savings Bank." In *Making the Irish American: History and Heritage of the Irish in the United States*, J. J. Lee & M. R. Casey, eds. New York: NYU Press, 2006, pp. 302-331.

Casey, Marion R. "Emigrant as Historian: Records, Banking and Irish American Scholarship." *American Journal of Irish Studies*, 10 (2013), 145-163.

Cirenza, Peter. "Geography and Assimilation: A case study of Irish immigrants in late nineteenth century America." In R. Hsu & C. Reinprecht, eds., *Migration and Integration New Models for Mobility and Coexistence.* Vienna University Press, 2016, pp. 173-200.

Cirenza, Peter. "Melting Pot or Salad Bowl? Assessing Irish immigrant assimilation in late nineteenth century America," London School of Economics PhD dissertation, 2011, available at: http://etheses.lse.ac.uk/90/1/Melting_pot_or_salad_bowl_assessing_Irish_immigrant_assimilation_in_late_nineteenth_century_America_%28Author%29_with_tables.pdf

Collins, William J. and Marianne H. Wanamaker. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics*. 6, no. 1 (2014): 220-252.

Collins, William J. and Ariell Zimran. "The Economic Assimilation of Irish Famine Migrants to the United States." *Explorations in Economic History*, 74, no. 1 (2019a): 1-22.

Collins, William J. and Ariell Zimran. "Working Their Way Up? US Immigrants' Changing Labor Market Assimilation in the Age of Mass Migration." NBER Working Paper 26414, 2019b.

Costa, D. L., H DeSomer, E. Hanss, C. Roudiez, S. E. Wilson, and N. Yetter. "Union Army Veterans, All Grown Up." *Historical Methods*. 50, no. 2 (2017): 79-95.

Conley, Timothy G. and David Galenson. 1998. "Nativity and Wealth in Mid-nineteenth Century Cities." *Journal of Economic History* 58, no. 2 (1998): 468-93.

Connor, Dylan S. "The Cream of the Crop? Geography, Networks, and Irish Migrant Selection in the Age of Mass Migration." *Journal of Economic History*. 79, no. 1 (2019): 139-175.

Connor, Dylan S. "In the Name of the Father: Fertility, Religion and Child Naming during the Demographic Transition." *Demography*, forthcoming.

Connor, D. S. and M. Storper. "The changing geography of social mobility in the United States." *Proceedings of the National Academy of Sciences, 117, no.* 48 (2020): 30309-30317.

Crayen, Dorethee and Joerg Baten. "Global Trends in Numeracy 1820–1949 and its Implications for Long-term Growth." *Explorations in Economic History* 47, no. 1 (2010): 82–99.

De Moor, Tine and Jan Luiten van Zanden. 2010. "'Every Woman Counts': A Gender-Analysis of Numeracy in the Low Countries during the Early Modern Period." *Journal of Interdisciplinary History*. 41, no. 2 (2010): 179-208.

Dribe, Martin, J. David Hacker, and Francesco Scalone. "The Impact of Socio-economic Status on Net Fertility During the Historical Fertility Decline: A Comparative Analysis of Canada, Iceland, Sweden, Norway, and the USA." *Population Studies* 68, no. 2 (2014): 135-149.

Ernst, Robert. *Immigrant Life in New York City, 1825-1863*. New York: I. J. Friedman, 1965.

Feigenbaum, James J. "A Machine Learning Approach to Census Record Linking* James J. Feigenbaum." Working Paper, March 28, 2016.

Fernihough, Alan and C. Ó Gráda. "Across the Sea to Ireland: Return Atlantic Migration before the First World War." UCD School of Economics Working Paper 19/29.

Ferrie, Joseph P. "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules." *Historical Methods*, 4 (1996), 141-156.

Ferrie, Joseph P. *Yankeys Now: Immigrants in the Antebellum U.S., 1840-1860*. New York, Oxford University Press, 1999.

Fitzpatrick, David. *The Americanisation of Ireland: Migration and Settlement, 1841-1925*. Cambridge: Cambridge University Press, 2020.

Gould, J. D. "European Inter-continental Emigration - the Road Home: Return Migration from the U.S.A." *Journal of European Economic History,* 9, no. 1 (1980): 41-112.

Helgertz, Jonas and Joseph R. Price, Jacob Wellington, Kelly Thompson, Steven Ruggles and Catherine Fitch. "A New Strategy for Linking Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel." IPUMS Working Paper 2020-03, doi: https://doi.org/10.18128/IPUMS2020-03 .

Herscovici, Steven. "Migration and Economic Mobility: Wealth Accumulation and Occupational Change among Antebellum Migrants and Persisters." *Journal of Economic History* 58, no. 4 (1998): 927-956.

Herscovici, Steven. "Progress Amid Poverty: Economic Opportunity in Antebellum Newburyport," *Journal of Economic History*, 57, no. 2 (1997): 484-488

Humphries, Jane and Jacob Weisdorf. "The Wages of Women in England, 1260-1850." *Journal of Economic History* 75, no. 2 (2015): 405-447.

Hough, Franklin B. *Census of the State of New-York for 1855*. Albany: Charles van Benthuysen, 1857.

Kelly, Morgan and C. Ó Gráda. 2000. "Market Contagion: Evidence from the Panics of 1854 and 1857." *American Economic Review*, vol. 90, no. 5 (2000): 1110-1120.

Kessner, Thomas. *The Golden Door: Italian and Jewish Immigrant Mobility in New York City, 1880-1915*. New York: Oxford University Press, 1977.

Kosack, Edward and Zachary Ward. "El Sueño Americano? The Generational Progress of Mexican Americans Prior to World War II." *Journal of Economic History* 80, no. 4 (2020): 961-995.

Long, Jason and Joseph Ferrie. "Grandfathers Matter(ed): Occupational Mobility Across Three Generations in the US and Britain, 1850-1911." *Economic Journal* 128 (2018): F422-445.

Maguire, John Francis. *The Irish in America*. London: Longmans, Green, 1868.

Massey, Catherine G. "Playing with Matches: An Assessment of Accuracy in Linked Historical Data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50, no. 3 (2017), 129-143.

Mokyr, Joel and C. Ó Gráda. 1982. "Emigration and Poverty in Pre-famine Ireland. *Explorations in Economic History*." 19, no. 4 (1982): 360-384.

Ó Gráda, C. "The Famine, the New York Irish, and Their Bank." In Antoin E. Murphy and Renée Prendergast (eds.), *Contributions to the History of Economic Thought – Essays in Honour of RDC Black*. London: Routledge, 2000, pp. 227-248.

Ó Gráda, C. "Savings Banks as an Institutional Import: The Case of Nineteenth-century Ireland. *Financial History Review*, 10, no. 1 (2003): 31-55.

Ó Gráda, C. "The New York Irish in the 1850s: Locked in by Poverty?" *New York Irish History*, 19 (2005): 5-13.

Ó Gráda, C. 2019. "The Next World and the New World: Relief, Migration, and the Great Irish Famine." *Journal of Economic History* 79[2]: 319-355.

Ó Gráda, C., Tyler Anbinder and Simone Wegge. "Assisted Emigration as Famine Relief: Lessons from the Lansdowne Estate." In Maurice Bric, ed. *Kerry: History and Society*, Dublin: Geography Publications, 2020, pp. 367-390.

Ó Gráda, C. and Eugene N. White. "The Panics of 1854 and 1857: A View from the Emigration Industrial Savings Bank." *Journal of Economic History* 63, no. 1 (2003): 213-40.

Olmstead, Alan L. *New York Mutual Savings Banks, 1819-1861*. Chapel Hill: UNC Press, 1974.

Pérez, Santiago. "The (South) American Dream: Mobility and Economic Outcomes of Firstand Second-Generation Immigrants in Nineteenth-Century Argentina." *Journal of Economic History* 77, no. 4 (2017): 971–1006.

Pérez, Santiago. "Intergenerational Occupational Mobility Across Three Continents," *Journal of Economic History*, 79, no. 2 (2019): 383-416.

Price, Joseph Kasey Buckles, Jacob Van Leeuwen, Isaac Riley. "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set." *Explorations in Economic History* (2021), forthcoming, doi: https://doi.org/10.1016/j.eeh.2021.101391.

Steven Ruggles, Catherine A. Fitch, Ronald Goeken, Josiah Grover, J. David Hacker, Matt Nelson, Jose Pacas, Evan Roberts, and Matthew Sobek. IPUMS Restricted Complete Count Data: Version 2.0 [dataset]. Minneapolis: University of Minnesota, 2020.

United Nations. *Demographic Yearbook Special Census Topics, Volume 1 – Basic Population Characteristics*, "Table 1c – Special Topic Volume 1." New York, 2017.

't Hart, Marjolein. "Irish Return migration in the Nineteenth Century." *Tijdschrift voor Economische en Sociale Geografie*. 76, no. 3 (1985): 223-31.

Thernstrom, Stephen. *The Other Bostonians: Poverty and Progress in the American Metropolis 1880- 1970*. Cambridge, MA: Harvard University Press, 1973.

Thernstrom, Stephan. "Poverty and Progress" Revisited: A Response to Riess, Frisch, and Pessen *Social Science History*, 10[4] (1986): 33-44.

Thernstrom, Stephan. *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Cambridge, MA: Harvard University Press, 1964.

Thomas, Ryah. "The Victorian Intergenerational Migrant Cohort: Socioeconomic outcomes 1851-1901." M.Phil. dissertation, University of Oxford, 2018 [https://ora.ox.ac.uk/objects/uuid:25918beb-701a-4c8a-865c-4e9225f3f6e0/download_file?file_format=pdf&safe_filename=MPhil%2BEconomic%2BHistory%2B-%2BDissertation%2B-%2BRyah%2BThomas%2B-%2BORA.pdf&type_of_work=Thesis].

van Leeuwen, Marco H. D., Ineke Maas and Andrew Miles. *HISCO. Historical International Standard Classification of Occupations.* Louvain: Leuven University Press, 2002.

van Leeuwen, Marco H. D., Ineke Maas. *HISCLASS: A Historical International Social Class Scheme*. Leuven: Leuven UP, 2011.

Vickers, Chris and Nicolas L. Ziebarth. "Economic Development and the Demographics of Criminals in Victorian England." *Journal of Law and Economics* 59, no. 1 (2016), 191-223.

Ward, Zachary. “Intergenerational Mobility in American history: Accounting for Race and Measurement Error.” Working Paper, Baylor University, 2020.

Wegge, Simone, Tyler Anbinder, and C. Ó Gráda. “Immigrants and Savers: A Rich New Database on the Irish in 1850s New York.” *Historical Methods*, 50, no. 3 (2017), 144-155.

Tables

TABLE 1
IRISH-BORN MALES BY OCCUPATIONAL CATEGORY: ESB'S NEW YORK CITY CUSTOMERS (1850-1858) AND ALL NEW YORKERS IN 1855

	ESB		**All New York Irish**	
Occupational Category	*N*	*%*	*N*	*%*
Professionals	20	0.4	16	0.3
Business Owners	397	8.0	310	5.8
Lower-Status White-Collar	370	7.4	260	4.9
Skilled Workers	1,738	34.8	2,022	37.9
Petty Entrepreneurs	255	5.1	90	1.7
Unskilled Workers	2,117	42.4	2,567	48.1
Others	91	1.8	73	1.4
Total (N)	4,988	100	5,338	100

Sources: Emigrant Savings Bank database; 1855 New York State Census [10 percent sample of employed Irish-born adults]. Both are available at TK. Note that the 1855 NY Census columns do not include one of the city's twenty-two wards because the returns for Ward Seventeen are not extant. The occupational categories are explained in the text on pp. 7-8.

TABLE 2
PERCENTAGES STAYING IN SAME OCCUPATIONAL CATEGORY, 1860-1870: HAND LINKED VERSUS ALGORITHM LINKED

A. ESB hand linked

Occ. category in 1860	*Occ. category in 1870 (%)*					
	Prof.	*Business*	*LSWC*	*Skilled*	*PE*	*Unskilled*
Professional	***78***	2	0	0	0	0
Business	0	***62***	23	10	9	11
LSWC	6	0	***54***	4	3	5
Skilled	11	27	4	***79***	0	5
PE	0	4	5	1	***64***	4
Unskilled	6	6	14	6	24	***75***
Observations	*18*	*52*	*79*	*287*	*33*	*283*

B. ABE Exact

Occ. category in 1860	*Occ. category in 1870 (%)*					
	Prof.	*Business*	*LSWC*	*Skilled*	*PE*	*Unskilled*
Professional	***36***	1	1	1	0	0
Business	13	***34***	18	16	19	16
LSWC	5	4	***14***	3	4	3
Skilled	17	14	20	***32***	18	18
PE	0	1	2	1	***11***	2
Unskilled	30	45	45	48	48	***62***
Observations	*64*	*803*	*584*	*3277*	*234*	*4791*

C. ABE Conservative

Occ. category in 1860	*Occ. Category in 1870 (%)*					
	Prof.	*Business*	*LSWC*	*Skilled*	*PE*	*Unskilled*
Prof	***55***	1	0	0	0	0
Business	3	***50***	18	15	18	16
LSWC	6	7	***26***	4	4	3
Skilled	9	14	18	***43***	19	17
PE	0	2	3	1	***23***	2
Unskilled	24	44	45	44	54	***68***
Observations	*31*	*366*	*244*	*1358*	*107*	*1925*

Notes: Occupational categories as in Table 1. LSWC = lower-status white collar; PE = petty entrepreneur.
Sources: ESB hand links from Emigrant Savings Bank database; ABE data from database of Irish-born men linked from the 1860 to 1870 U.S. censuses generated using the ABE Exact and Conservative algorithms, available at www.censuslinkingproject.org. The underlying census data are drawn from the complete count censuses published by IPUMS (Ruggles et al., 2020).

TABLE 3
GEOGRAPHICAL MOBILITY FROM 1860 TO 1870 OF MALE IRISH IMMIGRANTS LIVING IN NEW YORK OR KINGS COUNTIES IN 1860, BY LINKAGE METHOD

	ABE EXACT	**ABE CONS.**	**ESB Hand Links**
Changed state (%)	64	55	7
Changed county (%)	78	63	18
Total (N)	9,470	3,747	1,047
Description	Full ABE Exact match sample	Full ABE Conservative match sample	Sample hand-linked by genealogist

Note: Restricted to males aged 18 and above.
Source: See Table 2.

TABLE 4
RELATIONSHIP BETWEEN REMAINING MARRIED TO THE SAME WOMAN AND GEOGRAPHIC PERSISTENCE FOR IRISH IMMIGRANTS, 1860-1870, IN LINKED ABE CONSERVATIVE DATABASE

	Kings County		**NY County**	
	Stayers	Movers	Stayers	Movers
Same Wife, 1860-70 (%)	73.8	17.7	63.6	17.6
Different Wife, 1860-70 (%)	12.9	58.7	18.8	54.4
No Wife, 1870 (%)	13.3	24.7	17.6	8.0
Total (N)	271	462	880	1474

Source: See Table 2.

TABLE 5
CORRELATIONS BETWEEN HISCO 1860 AND 1870 VALUES FOR MALES IN NY AND KINGS COUNTIES, BY AGE

Age in 1860	*Stayers*		*Movers*	
	State	County	State	County
A. ABE Exact				
< 30	.153	.198	.024	.038
30-39	.423	.534	-.006	.038
40-49	.299	.440	-.002	-.014
50-59	.306	.336	-.015	.018
60-69	.389	.492	.056	.051
B. ABE Conservative				
< 30	.235	.304	.058	.066
30-39	.560	.636	.044	.104
40-49	.367	.491	.031	.018
50-59	.367	.390	.034	.067
60-69	.513	.517	.137	.180

Source: 1860-1870 matched Irish-born males generated by combination of ABE links with complete-count census data from IPUMS..

TABLE 6
WHIPPLE VALUES FOR IRISH-BORN ESB CUSTOMERS, 1850-1880

	Census Year				
	1850	1855	1860	1870	1880
Male					
Whipple	215	229	243	261	246
N	492	1447	1717	1213	651
Female					
Whipple	206	233	237	278	242
N	*248*	*756*	*822*	*530*	*281*

Source: Emigrant Savings Bank database.

TABLE 7
AGE GAPS AND AGE HEAPING AMONG ACCOUNT-HOLDERS (% OF TOTAL)

	Years					
	1850-1860		1860-1870		1870-1880	
Age difference [years]	All	0's+5's	All	0's+5's	All	0's+5's
<8	29.0	27.3	23.1	24.2	30.0	25.3
8-12	***51.4***	***47.1***	***53.5***	***47.8***	***51.3***	***48.3***
>12	19.7	25.6	23.4	28.1	18.7	26.5
Not linked	*48.6*	52.9	*46.5*	52.2	*48.7*	*51.7*
Total	407	172	1,337	645	947	487

Notes: the numbers in bold refer to observations within a five-year band of 10. The rows above and below refer to number outside the band.
Source: Emigrant Savings Bank database.

TABLE 8

ESB CUSTOMERS INCORRECTLY LINKED BY THE ABE CONSERVATIVE METHOD

	Name	Age	State	Occupation	Spouse	Children	Cause of ABE mistake
1860 census	Robt Baxter	31	MI	Brass finisher	Eliz.	Isabel, Robt., Mary, Amelia	
1870 ABE match	Robert Baxter	42	NY	Book binder	Letitia	none	
1870 JWS match	Robert Bartoo	43	MI	Brass finisher	Elize	Isabella, Robt., Maria, Mina, Alex	Mis-transcription of Baxter
1860 census	Patrick Berrigan	37	NY	Gardener	Mary	Wm., Sarah, Jas., Thos., Robt.	
1870 ABE	Patrick Berrigan	47	NY	Patient	none	none	
1870 JWS	Patrick Berrigan	51	NY	Laborer	Mary	Wm., Sarah, Thos., Josephine	Age discrepancy
1860 census	Denis Corcoran	60	NY	Laborer	Johanah	Michael (age 10)	
1870 ABE	Dennis Corcoran	70	PA	None	Mary	John (40)	
1870 JWS	Dennis Corcoran	60	NY	None	Johanna	John (30) and Michael (20)	Age discrepancy
1860 census	Hugh Donohoe	38	NY	Chairmaker	Jane	None	
1870 ABE	Hugh Donohoe	46	MN	Farmer	None	None	
1870 JWS	Hugh Donahue	50	NY	Tea peddler	Jane	None	Surname spelling variation
1860 census	Timothy Eagan	35	NY	Market man	Bridget	Wm., Stephen, John	
1870 ABE	Timothy Eagan	45	MI	Laborer	Mary	John, Michael, Mary, Rosa, Timothy	
1870 JWS	Timothy Egan	50	NY	Produce dealer	Bridget	Wm., Stephen, John, Mary A., Cath., Timothy, Jas., Jos., Edw.	Age & name spelling discrepancies
1860 census	Lawrence Fleming	29	NY	Laborer	Jane	Edw. (9 months old)	
1870 ABE	Lawrence Fleming	40	PA	Carter	Cath.	Jane (10), Cath., Annie, Ellen, Hugh, Julia	
1870 JWS	Lawrence Flemming	30	NY	Laborer	Jane	Edw. (10), Ann, Ellen, Cath.	Age & name spelling discrepancies

TABLE 9
PRIESTS IN 1860 CENSUS INCORRECTLY LINKED TO 1870 CENSUS

1860 Census Name	1860 State	Reported 1870 State	Reported 1870 Occupation	Evidence link is bad
George Brennan	WI	MA	Currier	Age/birthplace of children in 1870
Thos. Burk	IL	PA	Domestic Servant	Still a priest in IL in 1870
John Burnes	PA	AL	Retired Laborer	Died in 1866
Nicholas Byrne	NY	IA	Farmer	Found farmer, same name in Iowa, 1860
Richard Carroll	CA	NY	Laborer	Died 1861.
John Cody	PA	OH	Miner	Still a priest in PA in 1870
Michael Colton	IL	IL	Laborer	Age/birthplace of 1870 ch.
John Cullen	ME	IA	Retired Mason	Still a priest in ME in 1870
Thom. Cunningham	NY	NY	Laborer	Still a priest in NY in 1870
Patrick Donohoe	WI	NY	Mason	Still a priest in WI in 1870
Peter Eagan	MA	NY	Mason's Laborer	Died 1864
James M. Earley	NY	PA	Laborer	Still a priest in NY in 1870
Jas Elliott	KY	PA	"Old Gent"	Still a priest in KY in 1870
Daniel P. Falvey	NY	MA	Laborer	Died in 1866
Timothy O. Farrell	NY	MA	Tailor	Still a priest in NY in 1870
Wm Feely	IL	NY	Shoe Factory Hand	Age/birthplace of 1870 ch.
Cornelius Fitzpatrick	NY	PA	Works in Quarry	Still a priest in NY in 1870
Edward P Flaherty	IN	MA	Laborer	Died in 1868
Patrick J. Foran	MD	KY	Farm Laborer	Age/birthplace of 1870 ch.
Patk Gainor	CT	MA	Rolling Mill Worker	Died in 1869
John P. Macken	NJ	VT	Farmer	Age/birthplace of 1870 ch.
James Mackey	NY	OH	Laborer	Still a priest in NY in 1870
John Maginnis	CA	IL	Grocer	Grocer found in IL in 1860
Edwd McClusky	NJ	PA	Shoemaker	Age/birthplace of 1870 ch.
Jas McGuiness	NY	RI	Belt Maker	Age/birthplace of 1870 ch.
H. McLaughlin	PA	WI	Farmer	Age/birthplace of 1870 ch.
Daniel Moore	NY	CA	Laborer	Still a priest in NY in 1870
Patrick Moran	NJ	OH	None	Still a priest in NJ in 1870
Wm Noland	PA	ME	Marble Worker	Age/birthplace of 1870 ch.
Patrick Noman	PA	CT	Laborer	Priest surname incorrect
James O'Donnell	MA	NY	Shoemaker	Died in 1861
Joseph O'Keefe	PA	MN	Farmer	MN farmer found in 1860
Edward O'Neil	CT	NY	Laborer	Age/birthplace of 1870 ch.
Edward E.J. O'Riley	NY	NY	Works at Lime Kiln	Still a priest in NY in 1870
Michael O'Riley	CT	WI	Farmer	Age/birthplace of 1870 ch.
John B. Pursell	OH	PA	"At Home"	Still a priest in OH in 1870
Thomas O. Rielly	GA	NY	Liquor Dealer	Still a priest in GA in 1870
Patrick Riley	DE	PA	Laborer on RR	Still a priest in DE in 1870
Dennis Shean	NY	MA	Laborer	Still a priest in NY in 1870
Michael Sheridan	PA	NY	Coppersmith	Still a priest in PA in 1870
John W. Tiernan	WI	MO	Wagon Maker	Age/birthplace of 1870 ch.
Nicholas Walsh	PA	IL	Stone Cutter	Still a priest in PA in 1870
Francis Welch	IA	PA	Laborer	Still a priest in IA in 1870

Sources: See Table 2.

TABLE 10
MOBILITY FROM 1860 TO 1870 OF MALE IRISH IMMIGRANTS LIVING IN NEW YORK OR KINGS COUNTIES IN 1860, BY LINKAGE METHOD

	ABE Exact (%)	*ABE Conservative* (%)	*ESB Hand Links (%)*
Changed state	64	55	7
Changed occupational class	55	51	27
Moved up occupational class	28	26	18
Moved down occupational class	28	25	9
Changed wife	71	65	2
Changed state if wife same	27	18	8
Changed state if wife not same or no wife	71	67	6
N	9,753	4,031	752
Description	Full ABE exact match sample	Full ABE conservative match sample	Sample hand-linked by genealogist

Notes: Restricted to males aged 18 and above; *N* represents number of observations when data on occupational class are given. The reported ABE shares for this category are based on considering a match of the first four letters of a wife's name to be enough to designate the spouse in 1870 to be the same spouse. We allow for variations such as Margaret/Peggy, Ann/Nancy, Elizabeth/Betsy/Lizzie, or Catherine/Kate/Katharine, which were quite frequent. Nonetheless, the automated estimates of the rate of change in wives and occupations incorporate errors due to the coding of occupations and the real variants in the reporting of the names of wives over time.
Sources: See Table 2.

TABLE 11
THE GEOGRAPHIC AND OCCUPATIONAL MOBILITY OF IRISH-BORN DOCTORS, LAWYERS, AND CLERGYMEN LINKED BY THE ABE ALGORITHM FROM 1860 TO 1870

	Changed State (%)	Changed Occupation Category (%)	N
ABE Exact	57	57	355
ABE Conservative	43	40	178
Genealogist	12	12	298

Sources: See Table 2.

TABLE 12
THE GEOGRAPHIC AND OCCUPATIONAL MOBILITY OF EMIGRANT SAVINGS BANK CUSTOMERS LINKED BY THE ABE ALGORITHM FROM 1860 TO 1870

	Changed State	Changed Occupation Category
ABE Exact	61%	62%
N	306	265
ABE Conservative	53%	55%
N	130	112
Genealogist	8%	29%
N	194	185

Notes: The genealogist and algorithm agreed on at least one census entry in 1860 or 1870 for 306 bank customers, but the genealogist decided that good links could be made for only 194 of those 306. In about two-thirds of those 194 cases, the genealogist and the ABE algorithm made different links. The N is smaller for occupations than for state of residence because in some cases a person has been traced but the census listed them as unemployed or left the space for employment blank.
Sources: See Table 2.

TABLE 13
MOBILITY FROM 1860 TO 1870 OF MALE IRISH IMMIGRANTS LIVING IN NEW YORK OR KINGS COUNTIES IN 1860, BY LINKAGE METHOD

	SW Exact (%)	*SW Conservative (%)*	*ESB Customers, Hand Links (%)*
Changed state	30	21	7
Changed occupational category	39	35	27
Moved up occupational category	18	17	18
Moved down occupational category	20	18	9
Changed wife	n/a	n/a	2
Changed state if wife same	n/a	n/a	8
Changed state if wife not same or no wife	n/a	n/a	6
N	1,385	787	752
Description	Same wife in 1860 and 1870		Sample hand-linked by genealogist

Notes: Restricted to males aged 18 and above; *N* represents number of observations when data on occupational class are given.
Sources: ESB hand links and our links generated with our "same wife" variation of the ABE algorithms are available at TK.

Figures

Figure 1. Archbishop John B. Purcell of Cincinnati, as portrayed by Thomas Nast on the cover of *Harper's Weekly*, August 28, 1875.

Appendix 1. Age Heaping by Ward

Table A1.1 reports WI values for residents of New York by ward in 1855.

Table A1.1. WI values for New York City wards in 1855						
Ward	Male WI	Female WI	% Irish	% German	% US	Total
1	230	217	46.0	14.7	31.6	13,486
2	199	221	35.8	10.6	39.2	3,249
3	179	199	28.9	9.0	52.0	7,909
4	209	201	45.6	11.7	30.0	22,895
5	186	191	22.5	12.2	52.4	21,617
6	197	203	42.4	14.0	30.3	25,562
7	193	191	34.2	8.7	49.2	34,422
8	177	169	21.2	11.0	56.3	34,052
9	165	171	19.8	5.5	65.8	39,982
10	155	153	13.0	28.6	49.1	26,378
11	166	160	17.5	33.5	44.3	52,979
12	189	187	33.0	12.1	47.3	17,656
13	164	161	18.7	22.3	52.8	26,597
14	197	194	36.2	13.1	42.6	24,754
15	184	195	26.1	4.4	58.6	24,046
16	179	181	29.6	5.9	55.1	39,083
17	.	.	24.9	27.2	41.3	59,548
18	203	197	37.1	8.9	46.9	39,509
19	196	196	35.4	10.0	46.1	17,866
20	181	170	27.3	15.8	47.9	47,055
21	206	199	29.7	5.3	58.7	27,914
22	190	175	25.4	20.9	46.0	22,605
All	185	182	27.9	15.2	48.2	629,904
Source: Hough 1857, pp. 110, 117.						

Appendix 2. Hand linking and Type II errors

Selection is the original sin of much economic history. Our worry that bank customers might be, say, more driven or more prudential than Irish immigrants in general was somewhat alleviated after comparing the occupational profiles of EISB account holders with those of all Irish-born New Yorkers as reflected in the 1855 New York census. An added concern is that those successfully linked might be atypical of the bank customers in general. And, sure enough, comparing matches and non-matches for 1860 and 1870 for all the bank's Irish-born customers reveals some differences between them (Table A2.1). Among males with New York addresses when they opened an account for whom we have an occupational category at the outset, the "business owners" and "professionals" categories were significantly overrepresented among those matched. Knowing the order of the biases guards against undue generalisation.

Table A2.2, which compares the some of the saving patterns of those linked and those not linked, offers some further evidence of selection. The opening and peak deposits of linked account holders were likely to be higher; they were more likely to be in a joint account and to be held by women; they were held for longer and produced more transactions. These features are consistent with some positive selection, but they are not big.

A final linking anomaly was that the unlinked were much more likely to live in Wards 1-4 at the southern tip of Manhattan, though, interestingly, not in Ward 6, the city's most impoverished district. A significant proportion of the bank's customers in Ward 6 came from a single estate in County Kerry and tended to stay in Ward Six for many years, which may explain why they were easier to link than residents of other wards who had weaker ties to their neighborhoods (on this enclave, see Anbinder 2002).

Table A2.1. Male ESB Account Holders with NY addresses in 1860				
Occupational Category	Linked 1860-70	linked as % of total	Not linked	not linked as % of total
Professionals	5	0.7	15	0.4
Business Owner	93	12.2	304	7.2
Lower-Status White-Collar	54	7.1	316	7.5
Skilled Workers	208	36.8	1,458	34.5
Petty Entrepreneurs	38	5.0	217	5.1
Unskilled Workers	283	37.2	1,834	43.4
Others	8	1.1	83	2.0
Total	761	100	5,659	100

Table A2.2. Some characteristics of linked and non-linked depositors: median values		
	Linked	*Not linked*
Age in 1860	35	38
Peak savings [$]	411	171
Peak savings [$], males only	412	200
Peak savings [$], arrived pre-1846	524	320
Peak savings [$], arrived post-1845	331	150
Peak savings if has other account	498	302
Opening deposit [$]	100	60
Opening deposit [$], males only	100	70
Opening deposit [$], pre-1846	120	100
Opening deposit [$], post-1845	80	50
Years to highest deposit	2.83	1.23
Year of arrival	1849	1850
Joint account? (%)	31.4	22.4
Other account? (%)	58.0	31.7
Median duration of account (years)	5.33	2.56
Number of transactions	15	9
Female (%)	30.2	42.5
Number of accounts	c. 1,260	c. 9,830
Note: “Linked” includes all those matched in either 1860-70 or 1870-80		

Appendix 3. Note on Migration and Deaths in the 1860s[18]

Two groups of people unavoidably not captured by our use of the ABE algorithm are those who migrated elsewhere and those who died between 1860 and 1870 (Ferrie 1999: 22-26). Systematic data on migrants returning to Ireland from the US in the 1860s are lacking, but it is widely acknowledged that return migration was very much a minority phenomenon. The "American wake" wasn't for nothing.

Just after the period that interests us, the halving of fares increased the reverse flow but most of that increase probably consisted of people who returned home temporarily. Gould's estimate of the ratio of permanent return migration to immigration for Ireland in 1907-14, 6.7 percent, is only about half that implied by snapshot data for 1912-13 (Gould 1980: 57; Fernihough and Ó Gráda 2019; Fitzpatrick 2020: 13). Let us assume that the return migration rate during the 1860s was 5 percent.[19]

The second group missing in 1870 would have been those who died in the US since 1860. Here we apply male death rates per thousand population in New York City (defined as King's and New York counties) as recorded in the New York state census of 1855 (Hough 1857) to those in our linked database of Irish males living in NYC in 1860 (N = 15,109). Using 1855 data for the 1860s probably exaggerates mortality in the 1860s; against this, the life expectancy of the Irish was probably lower than that of New York's population as a whole. This suggests that about 15 percent of those present in 1860 would have died by 1870 (Appendix 3 below; compare Bailey *et al*. 2020: 1009). Table A3.1 summarises the calculations.[20] Since it is likely that the life expectancy of

[18] Our thanks to Michael Haines for advice on mortality in the 1860s. Any errors are ours.
[19] Marjolein 't Hart's 1985 study of Irish return migrants c. 1858-1865 focuses on their socio-economic characteristics, rather than the size of the reverse flow. 't Hart's returnees were part of the reverse flow described by The *Irish Times* as follows: "from the United States many are returning. One hundred were landed yesterday from the Glasgow, at Queenstown. Every vessel has its full complement, and we understand that thousands are anxious to return to this altered land, if they could find the means" (21 Sept 1861); however, a year later "Comparatively few Irish are now returning to Ireland. The rush homewards was made some weeks since, immediately on the first promulgation of the order for conscription" (1 Sept 1862).
We have not tried to search systematically for ESB account holders who returned to Ireland permanently, but have identified several.
[20] The final column was calculated as follows:
18.3 = 26*(95.1+45.98)*10/(1000*2);

the Irish was lower than that of native-born Americans due to living in less healthy neighbourhoods and also, perhaps, bearing the burden of malnutrition in the 1840s, this may underestimate attrition in the 1860s.

Adding losses due to migration and deaths, therefore, we may assume that at least 15 percent of the 1860 cohort would have "gone missing" by 1870.

Table A3.1. A Rough Estimate of Mortality in the 1860 Cohort

Age_1855	*POP1855_NYC*	*DTHS_55_NYC*	*DR_NYC*	*ABE_1860*	*CALC_DEATHS*
0	14,332	1,363	95.10	26	18.3
1-4	44,956	2,067	45.98	161	46.2
5-9	41,971	478	11.39	442	35.9
10-14	38,613	187	4.84	1,203	66.1
15-19	36,156	222	6.14	1,652	133.1
20-24	43,015	429	9.97	1,956	186.9
25-29	46,943	429	9.14	2,254	220.4
30-34	42,325	441	10.42	2,019	242.6
35-39	29,603	403	13.61	1,613	240.9
40-44	24,231	394	16.26	1,345	292.2
45-49	14,970	407	27.19	899	238.5
50-59	17,902	463	25.86	1,006	344.4
60-69	7,158	305	42.61	445	246.6
70-79	2,096	143	68.23	78	83.6
80 +	527	77	146.11	10	7.3
				15,109	2,385

46.2 = 161*(45.98+11.39)*10/(1000*2);
and so on.

APPENDIX 4. Further Comparison of ABE Links to Hand Linking

THE GEOGRAPHIC AND OCCUPATIONAL MOBILITY OF EMIGRANT SAVINGS BANK CUSTOMERS LINKED BY THE ABE ALGORITHM FROM 1860 TO 1870, OVERLAPPING OBSERVATIONS ONLY

	Changed State	Changed Occupation Category
ABE Exact	51%	51%
N	116	109
ABE Conservative	39%	44%
N	59	55
Genealogist	10%	34%
N	116	112

Sources: See Table 2.

Note: This table mirrors Table 12 but restricts the sample to "Overlapping observations" only. These are cases where the genealogist and the ABE algorithm started with the same observation in 1860 and both made links, even though they may have made different links to the 1870 Census. The genealogist links in the overlapping observations refer to links that overlap with ABE Exact. The N is smaller for occupations than for state of residence because in some cases a person has been traced but the census listed them as unemployed or left the space for employment blank.