Dear PAA session organizer

We include a preliminary version of our analysis based on two of our six samples. We have completed data linkage and preliminary analysis for the remaining samples, and will repeat the analysis to complete the paper before the May conference.

Regards

Evan Roberts (corresponding author)
eroberts@umn.edu

Mining Microdata: Economic Opportunity and Spatial Mobility in Britain and the United States, 1850-1881

Peter Baskerville
Department of History
University of Alberta
Edmonton, Canada
pab@uvic.ca

Steven Ruggles
Department of History & Minnesota Population Center
University of Minnesota
Minneapolis, United States of America
ruggles@umn.edu

Lisa Dillon
Department of Demography
Université de Montréal
Montréal, Canada
ly.dillon@umontreal.ca

Kevin Schürer
Department of History
University of Leicester
Leicester, United Kingdom
ks291@leicester.ac.uk

Kris Inwood
Departments of Economics and History
University of Guelph
Guelph, Canada
kinwood@uoguelph.ca

John Robert Warren
Department of History & Minnesota Population Center
University of Minnesota
Minneapolis, United States of America
warre046@umn.edu

Evan Roberts
Department of History & Minnesota Population Center
University of Minnesota
Minneapolis, United States of America
eroberts@umn.edu

***Abstract*— For almost two centuries social theorists have argued that the fundamental difference in social structure between Europe and North America arises from greater economic and geographic mobility in North America. We study social mobility in three countries across two generations using machine learning techniques to create panels of individuals linked between censuses thirty years apart (1850-1880, 1880-1910). This paper reports on a preliminary analysis of social mobility between 1850 and 1880, finding that mobility was markedly higher in the United States and Canada, compared to Great Britain.**

Keywords—machine learning; social mobility; census

I. Introduction

For almost two centuries, social theorists have argued that differences in economic opportunity and geographic mobility on the two sides of the Atlantic led to fundamental differences in social structure. In the opening line of *Democracy in America,* de Tocqueville stated that "no novelty in the United States struck me more vividly during my stay there than the equality of conditions"[1]. When he visited Canada, de Tocqueville found "the spirit of equality and democracy alive there as in the United States"[2]. Explaining why Americans were "restless in the midst of their prosperity," de Tocqueville expressed amazement at their rootless mobility, claiming that "a man will carefully construct a home in which to spend his old age and sell it before the roof is on . . . He will settle in one place only to go off elsewhere shortly afterwards with a new set of desires" [1]. Nineteenth-century commentators from de Tocqueville to the historian Frederick Jackson Turner maintained that the exceptional level of North American economic mobility was closely tied to geographic mobility: the availability of cheap land in North America allowed economic advancement and promoted high migration [3]. Westward expansion created a "safety valve," which many observers saw as the chief explanation for the failure of the socialist movement in North America [4-13].

In the twentieth century, Canadian and U.S. historians challenged this interpretation. Using linked censuses of more than a dozen communities, historians in both countries argued that despite high geographic mobility, nineteenth-century North America had a rigid class structure with comparatively little upward mobility [14-26]. Some suggested that migrants constituted a "floating proletariat" of declining fortune [17]. In recent years, however, a few studies using national data have argued that the nineteenth-century United States was extremely

fluid compared with nineteenth-century England [27]. The new results suggest that there has been a dramatic decline in the United States in both economic and geographic mobility over the past 150 years. If confirmed, these results would have profound implications for our understanding of social structure and social change on both sides of the Atlantic.

In a project funded by the 2011 application round of the Digging into Data initiative, we apply new data-mining technology to massive new census microdata collections in Britain, Canada, and the United States to address four key questions:

1. What were the relative levels of economic and geographic mobility in Britain, Canada, and the United States in the late nineteenth century?
2. What were the mobility trends in each country?
3. How were economic opportunity and geographic mobility interrelated in each country?
4. What individual and community characteristics were associated with economic and geographic mobility?

II. Data

This project is based on one of the largest microdata collections in existence, the North Atlantic Population Project (NAPP) [28-30]. The NAPP database includes complete enumerations of the populations of Britain, Canada, the United States, and several other countries between 1850 and 1911. The data consist of numerically coded transcriptions of historical censuses for Britain, Canada, and the United States. The files have a hierarchical format, with individuals nested into families and households; within each family and household, the interrelationships of the members are known. The numeric coding system is consistent across countries. Most of the data we intend to use was already incorporated into the NAPP data access system (http://www.nappdata.org) at the inception of the project. The data from which we draw our samples are freely available on the Internet [29]. In addition to the existing NAPP data, during the course of the project, we incorporated new complete-count datasets for Britain in 1911 and a large new sample for Canada in 1852.

Censuses in the United States were conducted every 10 years after 1790. In Canada and Great Britain censuses have been scheduled every 10 years on the '1' years, though Canada's scheduled 1851 census was taken in 1852. Thus our comparison of social mobility over similar generations will be of slightly different years in each country: 1850-1880 and 1880-1910 in the United States, 1852-1881 and 1881-1911 in Canada, and 1851-1881 and 1881-1911 in Great Britain. In the remainder of the text we abbreviate these thirty year intervals as 1850/1-1880/1 and 1880/1-1910/1.

III. Project Goals

The project aims to create representative longitudinal panels of census data in a comparable manner in three countries, and contribute to a long-standing debate on social structure and opportunity in Britain and North America. Given the recent availability of large-scale census databases the challenge *now* in constructing panel data from censuses is the adapting of machine learning teachniques to replace case by case linking pioneered by genealogists. The principal challenge is *not* to find sufficient cases, but ensuring that the panels are representative, unbiased and accurate. False links lead to artifactual social mobility, so it is important to ensure high levels of accuracy. We do this in a similar way across Canada, Great Britain, and the United States taking account of differences in census enumeration methods and questions.

IV. Record Linkage Approach

Our linkage strategies build on recent research in data mining and machine-learning [31]. The theoretical framework for probabilistic record linkage derives from Fellegi and Sunter, who demonstrated that it is possible to define an optimal linkage rule that minimizes the number of false links [32]. Major extensions and refinements of record-linkage theory were contributed by Jaro, Winkler, Belin, Rubin, and Larsen [33-36]. Recent research has focused on using machine-learning techniques instead of fixed linkage rules [37]

Our record linking procedures build on these innovations. Our goals, however, differ significantly from those of most data mining applications of record linkage. The primary goal of most data mining has been to maximize the number of valid links. Our objective is different: we do not focus on maximizing the linkage rate. Instead, our procedures are designed to maximize the *representativeness* of the linked cases and the *accuracy* of the links. This means we pay close attention to potential sources of selection bias, and ignore information routinely used by other record-linkage procedures. Although we cannot eliminate selection bias for unobserved characteristics, we can adopt procedures that greatly reduce the potential for bias compared with previous approaches.

Our algorithm relies exclusively on characteristics that should not change over time. At minimum, these variables are first name, last name (for men and for women who do not marry between observations), birth year, sex, and place of birth. Most record linkage software makes use of a broader range of characteristics to confirm links and resolve ambiguities, but that approach introduces bias. For example, if we use spouse's characteristics to confirm linkages, we would bias the sample in favor of persons who remained married to the same person for multiple decades, and such persons are not representative with respect to either occupational or geographic mobility. Wisselgren et al provide a recent discussion and evaluation of these issues in historical census record linkage [38].

A challenge posed by our approach is that the limited set of variables we use cannot uniquely identify all individuals. To take the worst-case scenario—the most common male name with the most common birthplace—the 1880 U.S. census has 17 white men aged 33, named John Smith, and born in New York. Even this example understates the problem, because it assumes an exact match of name and age. Errors in enumeration and transcription cause a significant proportion of matches to be imperfect: linking must be carried out probabilistically, allowing for imperfect correspondence of name and age. Whenever there is more than one possible

match, we must exclude all potential matches. This eliminates many true matches, but is necessary to minimize false matches. False matches would lead to systematic upward bias for transition rates—such as migration and occupational mobility—and therefore must be avoided at all cost.

Because our linking strategy must rely heavily on names, we need an approximate string comparison algorithm. We favor the Jaro string comparator as modified by Winkler [39, 40]. This algorithm computes a similarity measure between 0.0 and 1.0 based on the number of common characters in two strings, the lengths of both strings, and the number of transpositions, accounting for the increased probability of typographical errors towards the end of words. In addition to using a string comparator, we standardize given names to account for diminutives and abbreviations (e.g., "Willie" and "Wm." are transformed into "William.") Such name-cleaning techniques are language-specific and must be customized for each language of enumeration. This work draws on the rich body of research on name cleaning [39-44]. Finally, we use both NYSIIS and Double-Metaphone phonetic name coding, which provide multiple encoded strings corresponding to variant pronunciations [45, 46].

We use two approaches to calculate similarity measures, including Jaro/Winkler indices and age similarity scores. We use both the open-source "Freely Extensible Biomedical Record Linkage" (FEBRL) software [47-50] and a new implementation of distance function routines written by Guelph post-doctoral researcher Luiza Antonie customized for large historical datasets [51-53]. Other linking variables—such as birthplace and sex—do not pose string comparison problems because they are numerically coded to eliminate spelling variation. Thus, for example, we do not worry about the innumerable spelling variations of Aberystwyth, or variant names for the same location.

We assume every pair of records drawn from two files are either matches referring to a single individual or non-matches describing two different persons. Optimal matching requires every individual be compared with every possible match. It is not computationally feasible, however, to assess every potential match. For example, using such a linking algorithm for the full U.S. 1880 census and 1900 U.S. sample would involve over 15 trillion comparisons. To reduce the computational load, we use "blocking factors"—such as birthplace, sex, and race—limiting comparisons to people sharing blocking factors.

To estimate parameters for the record linkage algorithm, we need training data. Training data are cases where true links are known. We obtain training data by having multiple research assistants hand-link the same sets of data, and combine the results to obtain a set of highly-reliable links. We use the training data to implement a Support Vector Machine (SVM) on the full set of unlinked census data to classify each potential match [54-56]. We implement the SVM using the open-source library of tools developed by Chang and Lin [57]. Based on the training data, the SVM calculates a confidence score for every potential match; when one and no more than one potential match exceeds the threshold, we establish a link. We have extensively tested our procedures against known links, and we estimate that the false link rate averages less than 3%. Once we have established the full set of links, we weight the cases to represent the potentially linkable population with respect to age, sex, birthplace, whether related to head, occupational group, and size of place in the terminal year.

V. Measuring social mobility

We adopted the Historical International Standard Classification of Occupations (HISCO) as our basic framework for occupational classification.[58-60] The HISCO system is a modification of the 1968 United Nations occupational classification system with extensions to accommodate historical occupations. HISCO was developed by an international committee with representatives from Belgium, Canada, England, France, Germany, the Netherlands, Norway, Sweden and the United States. We modified and extended the HISCO system to accommodate the additional detail available in the North Atlantic database.[61] To ensure that we coded the millions of occupations comparably across each country, we traded random samples of the occupation dictionary across countries, so that part of each country's occupations were independently coded by researchers in each other participating country. We then reconciled all differences of interpretation, which sometimes involved lengthy discussion and debate.

Our measure of social background outcomes is occupation in early adulthood, measured for the subjects' fathers when the subjects are 0-19 years old, and for the subjects at age 30-49. Occupations are the only measure of social and economic status collected in a consistent manner across time and space in pre-World War II statistical sources. While earnings varied within occupations, there is a relatively stable ordering of earnings across occupations over time [62]. We classify our occupations initially into a modified version of the Historical International Standard Classification of Occupations coding scheme and then aggregate occupations into four categories to measure social class [58, 59, 61]. In this paper we combine occupations into four broader groupings for analysis: (1) white collar workers: a broad group encompassing professionals, clerical workers, and sales people, (2) farmers (3) skilled workers or supervisory workers, such as foremen or overseers, and (4) unskilled workers, encompassing various industrial sectors from service work to farming to manufacturing. Our classification mirrors that in Ferrie and Long's recent analysis of social mobility in the same countries [63].

VI. Results

In this paper we report on an initial analysis of social mobility between 1850/1 and 1880/1 in Great Britain and the United States. Our sample for analysis is boys aged 0-19 in 1850/1, who were living with a co-resident father. In both countries we obtain a sample of slightly under 4000 young men, who we are able to follow into their own adult lives thirty years later. The demographic characteristics of the panels are fairly similar (Great Britain, Table 1; United States, Table 2).

Several demographic aspects of the two samples are interesting. Family size at a comparable stage of the life-course

dropped significantly between generations in both countries. In the second generation family size in 1880/1 averaged 5 (prototypically, a husband, wife and three children). Yet in Great Britain these men had come from families with, on average, 1.4 more children in 1851. In the United States, these men had hailed from families with an average family size of 7.2. Thus, the family context of these men became more similar in the second generation. In many other respects the demographic characteristics of the two samples are remarkably similar: compare for example the average ages of fathers and sons, and the fertility of the second generation by 1880/1.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF BRITISH SAMPLE

Variable	Mean	StdDev	CV	Min	Max	N
Age, 1850/1	8.4	5.7	0.68	0	20	3919
Age, 1880/1	38	5.7	0.15	30	51	3919
Family size, 1850/1	6.4	2	0.32	2	16	3919
Family size, 1880/1	5	2.7	0.54	1	61	3919
Siblings, 1850/1	3.3	2	0.61	0	13	3919
Working kids	0.58	0.93	1.6	0	6	3919
Has kids, 1880/1	0.71	0.45	0.63	0	1	3919
Num. Children, 1880/1	2.6	2.3	0.89	0	9	3919
if has kids	3.7	1.9	0.53	1	9	2800
Youngest child	4	4.6	1.2	0	27	2800
Eldest child	12	5.9	0.51	0	37	2800
Number < 5	0.83	0.99	1.2	0	5	3919
Father's age, 1850/1	42	18	0.43	20	999	3919
Married, 1880/1	0.82	0.39	0.47	0	1	3919

TABLE II. DEMOGRAPHIC CHARACTERISTICS OF AMERICAN SAMPLE

Variable	Mean	StdDev	CV	Min	Max	N
Age, 1850/1	8.8	5.8	0.65	0	20	3715
Age, 1880/1	39	5.8	0.15	30	51	3715
Family size, 1850/1	7.2	2.4	0.33	2	17	3715
Family size, 1880/1	5	2.3	0.46	1	16	3715
Siblings, 1850/1	3.9	2.3	0.59	0	9	3715
Working kids	0.48	0.84	1.7	0	5	3715
Has kids, 1880/1	0.77	0.42	0.55	0	1	3715
Num. Children, 1880/1	2.5	2.2	0.86	0	9	3715
if has kids	3.3	1.9	0.58	1	9	2854
Youngest child	4.6	4.8	1	0	25	2854
Eldest child	12	5.8	0.5	0	31	2854
Number < 5	0.72	0.88	1.2	0	6	3715
Father's age, 1850/1	42	9.6	0.23	17	81	3715
Married, 1880/1	0.86	0.35	0.41	0	1	3715

A. Geographic mobility over thirty years

Particularly in the nineteenth century geographic and social mobility were strongly related. Young men often significant distances to seek new work. Indeed, the restlessness that de Tocqueville and other observers noted about North America was a geographic one. Just over half (52%) of the American sample moved counties between 1850 and 1880. In Britain 36% of men moved counties between 1851 and 1881. Yet this overstates movement in Britain relative to the United States, since the geographic size of British counties was substantially smaller. The pattern of moves was dispersed. No origin-destination pair of states accounted for more than 2.2% of all those who moved. Yet, there was a consistent pattern to geographic mobility in the United States: nearly everyone who moved headed west. Thus, by 1880 the population of this sample had spread widely across the contiguous United States (Fig. 1, Fig. 2). The most common moves in Britain were to adjacent counties, whereas in the United States many movers had skipped entire adjacent states.

Fig. 1. Residence of U.S. sample in 1850

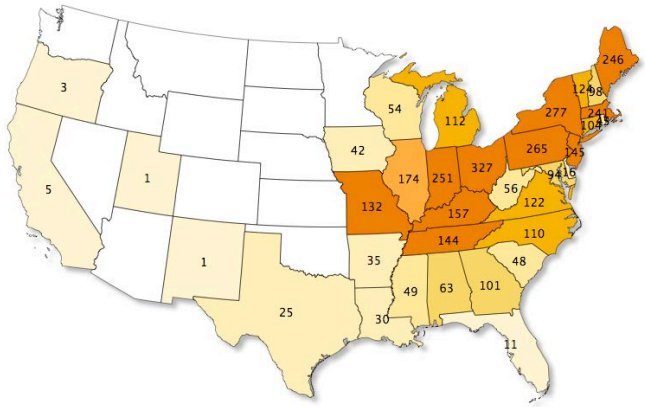
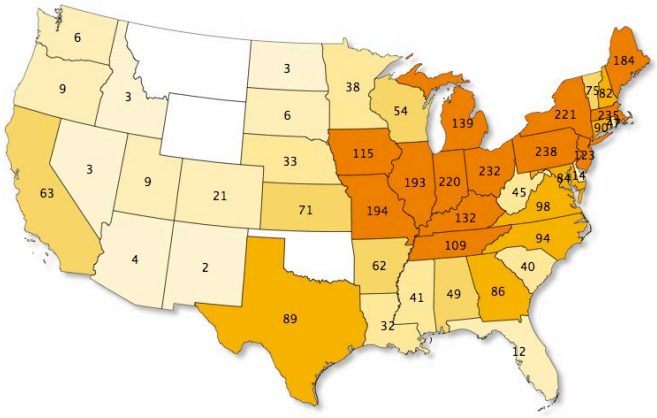


Fig. 2. Residence of U.S. sample in 1880



VII. Occupational Status and Transitions

In both countries, the information on occupations of fathers in 1850/1 and sons in 1880/1 allows rich description and analysis of the pattern of occupational change across generation. As well as the occupation, both countries's censuses also included other information indicating social and economic status. In the United States census of 1880 information on literacy and unemployment was also collected. Again, the sample is broadly representative of American white men of this era, who had achieved nearly universal literacy. Unemployment was also low, with just 5.5% of this sample having experienced unemployment in the year preceding the census.

A major difference in the occupational structure of the two countries is the radically different proportion of the workforce, and of these representative samples, in farming (Table III). In the United States, 62% of the fathers were farmers in 1850, declining only to 47% among their sons in 1880. Britain's occupational structure was quite different, with the industrial revolution much further advanced. In Britain just 9.4% of fathers were farmers and 4.6% of their sons in 1881. In both countries this mirrored broader trends in the changing occupational structure. The proportion of farmers among American men was not below 10% until well into the 1920s, showing the dramatic differences in occupational structure between the two countries. Despite a large drop in the proportion of American men farming, nearly half the sons in 1880 were still farmers. Though not all were the sons of farmers, many were. Thus, in the United States a far greater proportion (42%) of sons had the exact same occupation as their father than in Britain (23%). Yet this highlights a limitation of the occupational information in the census. Although both countries supported a diversity of farming, the census recorded nearly all as "Farmer," omitting to record the crop or animal farmed.

TABLE III. Occupational Status of Sons and Fathers

Great Britain						
Variable	**Mean**	**StdDev**	**CV**	**Min**	**Max**	**N**
Exact occ as dad	0.23	0.42	1.8	0	1	3919
Same major group	0.4	0.49	1.2	0	1	3919
Father farmer, 1850/1	0.094	0.29	3.1	0	1	3919
Father, acres farmed	132	158	1.2	1	1141	325
Son farmer, 1880/1	0.046	0.21	4.5	0	1	3919
Son, acres farmed	184	209	1.1	1	1400	156
United States						
Can read and write	0.95	0.22	0.23	0	1	3715
Unemployed in 1879/80	0.055	0.23	4.1	0	1	3715
Sick on 1880 census day	0.014	0.12	8.4	0	1	3715
Exact occ as dad	0.42	0.49	1.2	0	1	3715
Same major group	0.5	0.5	1	0	1	3715
Father farmer, 1850/1	0.62	0.49	0.78	0	1	3715
Son farmer, 1880/1	0.47	0.5	1.1	0	1	3715

Particularly for farmers, sharing the exact same occupational description is likely to overstate the extent to which sons were actually doing the exact same work as their father. A broader measure of occupational inheritance between generations is the proportion of men who had an occupation in the same "major group" as their father. The HISCO occupational codes identify nine major groups of occupations: Professionals, Managers, Clerical workers, Sales workers, Service workers, Agricultural workers, Manufacturing workers, Transport workers, and Laborers. Sons who had jobs in the same major group as their father were likely to be doing something similar, either in terms of what they were producing, or the level of education and skill brought to the job. To make the concept more concrete, a father who was a carpenter and a son who was a painter would both be in the same major group. Both might have worked in the construction industry. Similarly, a father who was a lawyer and a son who was a doctor are both professionals, both occupations requiring a high level of education and thus similar in that respect.

Occupations provide a great deal of detail on what fathers and sons were doing, but this very detail can inhibit understanding of how father's occupations influenced son's occupations. In order to make sense of how closely a father's occupation influenced his son's occupation, we need to aggregate occupations into a smaller number of categories. To assess occupational mobility between generations we combine occupations into four broader groupings for analysis: (1) white collar workers: a broad group encompassing professionals, clerical workers, and sales people, (2) farmers (3) skilled workers or supervisory workers, such as foremen or overseers, and (4) unskilled workers, encompassing various industrial sectors from service work to farming to manufacturing. Our classification mirrors that in Ferrie and Long's recent analysis of social mobility in the same countries [63].

Our results are summarized in Table IV, describing occupational mobility from 1850/1 to 1880/1 in both countries. The layout of the panel for the two countries is identical. Occupations of the father are described in the columns, and of sons in each row. The classification of occupations is the same for fathers and sons. For each cell we list the number of sons of fathers in that occupational group who end up in a given occupational group. Percentages are calculated within columns for each country. For example, in Britain, 484 fathers had white collar occupations, and 274 of their sons (56.6%) also had a white collar occupation.

An assessment of occupational mobility requires us to measure how closely associated son's occupations were the occupation of their father's. In a symmetrical table the natural measure of association is a cross-product. However, as discussed earlier the occupational structure of the two countries differed significantly. We follow Long and Ferrie in calculating the Altham statistic for the tables of fathers' and sons' occupations [63, 64]. By multiplying one of the tables by a series of arbitrary constants the marginal frequencies are made identical, allowing us to compare only the degree to which the rows and columns are associated, i.e. the extent to which fathers occupations influence their son's occupations.

TABLE IV. INTERGENERATIONAL OCCUPATIONAL MOBILITY IN GREAT BRITAIN AND THE UNITED STATES, 1850/1-1880/1

	Father's occupations (1850/1)				
Great Britain	**White collar**	**Farmer**	**Semi/skilled**	**Unskilled**	**Total**
Son's occupation (1881)					
White collar	274	57	368	83	782
	56.61	15.24	18.1	8.07	19.95
Farmer	9	134	29	18	190
	1.86	35.83	1.43	1.75	4.85
Semi/skilled	158	109	1,438	472	2,177
	32.64	29.14	70.73	45.91	55.55
Un-skilled	43	74	198	455	770
	8.88	19.79	9.74	44.26	19.65
Total	484	374	2,033	1,028	3,919
	100	100	100	100	100
United States					
Son's occupation (1880)					
White collar	150	298	183	33	664
	48.86	12.78	23.4	11.22	17.87
Farmer	71	1,439	186	92	1,788
	23.13	61.71	23.79	31.29	48.13
Semi/skilled	66	358	323	90	837
	21.5	15.35	41.3	30.61	22.53
Un-skilled	20	237	90	79	426
	6.51	10.16	11.51	26.87	11.47
Total	307	2,332	782	294	3,715
	100	100	100	100	100

Note: Each cell reports frequency (e.g. 274) and column percent (e.g. 56.61)

Some aspects of the different occupational structure and transitions can be seen just from Table IV. In Great Britain, 44% of sons of unskilled workers remained in the same unskilled class, whereas in the United States just 27% of sons of the unskilled remained in that class. Thus, upward mobility for the sons of the lowest skilled was approximately half as likely again in the United States.

In both countries occupational inheritance was strong, with high percentages along many of the diagonals of the table. The exceptions to this are relatively low inheritance of farming in Britain, and the higher upward mobility of the unskilled in the United States. While occupational inheritance of farming occupations was high in the United States—61% of farmers' sons were farmers—more than 20% of the sons of other occupational classes also ended up in farming. The most similar aspect of the two countries occupational structure was entry into white collar work. In both countries occupational inheritance was relatively high, with around half of the sons of white collar workers being white collar workers themselves thirty years later. The proportion of sons of other occupational classes who ended up as white collar workers was relatively similar in both countries (compare the top row of each panel of Table IV).

TABLE V. ASSOCIATION BETWEEN FATHERS' AND SONS' OCCUPATION IN GREAT BRITAIN & THE UNITED STATES, 1850/1-1880/1

(1) Comparison	(2) M	(3) M'	(4) d(P,J)	(5) d(Q,J)	(6) d(P,Q)	(7) d^i(P,Q)
Ferrie/Long GB 1881 (P)	42.6	35.5	22.7 ***		13.2 ***	4.5
Ferrie/Long US 1880 (Q)	45.4	47.9		11.9 ***		
This paper GB 1881 (P)	41.2	33.8	25.2 ***		12.2 ***	2.5
This paper US 1880 (Q)	46.4	50.4		14.9 ***		

Note: *** indicates statistical significance at p=0.01.

Table V summarizes occupational mobility in Britain and the United States over a similar period of thirty year. We compare our results to Long and Ferrie, who created samples over a similar time period using alternative linkage methods. Column M reports the proportion of off-diagonal entries in each country, sons who ended up in a different occupational group than their father. Overall levels of mobility are similar, with the higher occupational inheritance of farming in the United States being balanced out by higher occupational inheritance in other categories in Britain.

However, the occupational structure differed in the two countries over time. Thus Column M' reports adjusted mobility statistics where the American marginal totals have been adjusted to match the British, and vice-versa. This comparison shows mobility in the United States to have been substantially greater than in Britain: son's occupations were not as tightly related to their father's occupations in the United States.

The underlying association between fathers' and sons' occupations is measured by the Altham statistic, which calculates the distance from independence of the occupational structure. In a simple 2 x 2 matrix the Altham statistic is the familiar cross-product ratio (ac/bd). If the rows and columns are independent, then the cross product ratio is 1. A matrix where all elements are ones satisfies these conditions, or indeed any matrix of constants. Matrices with more than 2 rows and columns have multiple cross product ratios, and the Altham statistic incorporates all the cross-product ratios into a single statistic.

$$d(\mathbf{P},\mathbf{Q}) = \left[\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{l=1}^{r}\sum_{m=1}^{s}\left|log\left(\frac{p_{ij}p_{lm}q_{im}q_{lj}}{p_{im}p_{lj}q_{ij}q_{lm}}\right)^2\right|\right]^{1/2} \quad (1)$$

The statistic has a chi-squared distribution, and the statistical significance of the metric can be calculated. The Altham statistics for Britain and the United States are presented in Columns 4 and 5 of Table V. In both countries the occupations of fathers and sons were strongly related, as the Altham statistic are significantly different from 0 in both cases. That is, comparing the frequencies for each country to a matrix of

identical constants (independent occupations) shows that both countries father-to-son occupational transitions differed significantly from the baseline of independence. However, the Altham statistics for Britain were 2/3 as large again as in the United States. Just as we can calculate the difference between each country's matrix and the null hypothesis of independence, we can also calculate the difference between the Altham statistics for each country, and whether it is statistically significant. This statistic is displayed in Column 6: d(**P,Q**), and we compare our results with Ferrie and Long's prior work on the same time period.

Ferrie and Long's matching method relied to a greater extent on exact similarity in the spelling of names, and a more rigid treatment of age discrepancies between censuses. Our linking methodology allows slightly greater tolerance for discrepancies in names and ages, particularly when there are no other potential matches that could be made. Ferrie and Long's method is slightly more likely to lead to false positive matches, and a higher degree of mobility. The differences in the Altham statistics between our results and theirs lie consistently in this direction (Columns 4 and 5). We find that both Great Britain and the United States were further from independence than Ferrie and Long do: in our results fathers' occupations exerted a slightly greater constraint on their sons' occupations than Ferrie and Long found. However, as can be observed the differences are relatively small, and do not attain statistical significance. Indeed, the difference that we find between Great Britain and the United States is very similar to what Ferrie and Long found (Column 6).

Finally, looking at the off-diagonal elements only (Column 7), we find only small differences in the overall degree of association between the countries. Thus, the differences in mobility between the two countries are mostly due to differences in occupational inheritance within the same occupational groups. In only one case (white collar to white collar) are the diagonal elements similar across the two countries, and the differences along the diagonals are fundamental to the differences between the two countries.

VIII. Agricultural Inheritance in Great Britain

Although relatively few men farmed in late nineteenth century Britain, compared to the United States, the transition of sons out of farming was socially significant. Many people in late nineteenth century British society were concerned about concentrated wealth holding, and the continuing control of farms by a landed elite. Data on overall patterns of land inheritance within British farming are scarce, yet the census returns contain information that allows much greater exploration of these questions than in the existing literature.

Instructions to British census enumerators asked them to record the acreage of farms, and the number of employees that a farmer had. Thus, farmers in the British census typically have occupational responses of the following form:

Farmer of x_1 acres employing y

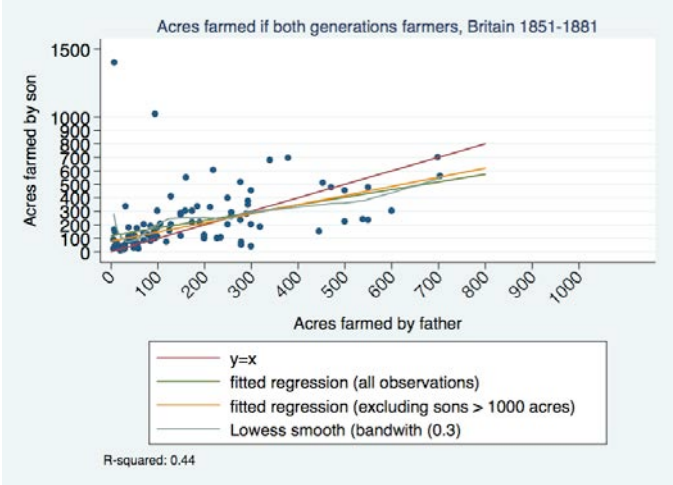
Farmer of x_2 acres employing y_1 men and y_2 girls

The expressions are regular, with the number of acres almost always preceding the word acres, or a limited number of spelling variations. There is slightly less regularity of the expressions describing employees, but the number of variants of ways to describe employees is finite and straightforward to identify. Our linked sample is small, matching a 2% sample of the 1851 census with a 100% database of the 1881 census. Thus, we have 367 fathers who are farmers, and 182 sons. Complete databases of all British censuses from 1851-1911 will soon be available with occupational information transcribed, and it will be feasible to parse out information on acres farmed and employees on farms from occupational descriptions.

To do this, we first identify variants of the word "acres" that are found in the data, such as "ac", "acr", "acers", "acres)", "acs", "acre", "acrs", and "a". The program then reads each occupational description and extracts the word before acres to place in a new variable measuring acres farmed. We do this for both fathers' and sons' occupational descriptions. This new data allows us to examine how acres farmed by the father affected son's occupational chances using a simple probit model. For sons who remained in farming, we can compare the acres farmed between generations.

Fig. 3. Relationship of son staying in farming to fathers' acres farmed

Fig. 4. Relationship of sons' and fathers' acres farmed



Several conclusions are apparent from this analysis, which we emphasize is more suggestive of the potential for application to pending complete-count databases of the British census, than a definitive analysis of agricultural inheritance in nineteenth century Britain. First, the chances that a son stayed in farming was strongly related to how many acres a father farmed. It was not until a father farmed more than 400 acres that his son had a greater than even chance of remaining in farming (Figure 3). Among sons who remained in farming, however, most ended up farming more acres than their father (Figure 4). This pattern is suggestive of selection into farming, where sons with the best chances of acquiring land stayed in the occupation; and also confirms that land ownership became more concentrated in late nineteenth century Britain. Finally, it is notable that the relationship between fathers' and sons' acres farmed is relatively consistent across the range of observations, with the fitted OLS regression line and a locally weighted regression line remaining close to each other over the range of the data.

IX. Conclusion

We began this paper by noting the fundamental historical question dating back to de Tocqueville, if not earlier, that motivates our research: was (is) America a more mobile society than Europe. Social mobility is an issue of special significance to the humanities, reflecting the extent to which societies organize themselves to allow either many or few of their citizens to exercise the full extent of their talents. We apply linking methods new to the historical literature. Historical linking has either been done by hand without specified rules, or by machine with exact matching or with rigid criteria for deviations from exact matching.

In this paper we use samples of the American and British 1850/1 censuses linked to complete databases of the 1880/1 censuses, but the methodology is scalable to forthcoming complete databases of these populations that will increase the number of potential and achieved matches significantly.

Despite the differences in our linking methodology to the research of Ferrie and Long [63] we find relatively small differences in the substantive conclusion that late nineteenth century America was a significantly more mobile society than Britain at the same time. That this finding is robust to alternative methods of constructing linked census samples only strengthens the conclusion about social differences across the Atlantic.

Our research also highlights the importance of large samples for investigating questions of social mobility, and indeed other historical questions. While we summarize the overall differences between occupational mobility in Britain and the United States in a single statistic, the statistic can be decomposed into a smaller number of component statistics that show more precisely where the two countries diverged. In the late nineteenth century, those differences lay largely in greater American persistence in farming across generations, and a significantly greater chance for sons of unskilled men to end up in farming, white collar work or skilled occupations. Moreover, in the United States sons of farmers who left farming were much more likely to avoid ending up in unskilled work than their peers in Britain. Taken together, these results suggest that young men in the late nineteenth century United States had significantly better life chances than their British peers. Were these differences the result of institutions—such as government and educational opportunities—or environments—with more abundant land in the United States? The next phase of our research will incorporate Canadian data for the same time period, and for all three countries for a subsequent generation (1880/1 – 1910/1) to address these questions.

References

[1] A. de Tocqueville, *Democracy in America*. Paris, 1831.

[2] A. de Tocqueville, *Regards sur le Bas-Canada*. Montréal: Typo, 2003 (1836).

[3] F. J. Turner, "The Significance of the Frontier in American History," *Proceedings of the State Historical Society of Wisconsin,* vol. 41, pp. 79-112, 1893.

[4] R. Archer, *Why is there no labor party in the United States?* Princeton: Princeton University Press, 2007.

[5] J. Heffer and J. Rovet, Eds., *Why Is There No Socialism in the United States?* Paris, 1988, p.^pp. Pages.

[6] A. Bosch, "Why Is There No Labor Party In The United States? A Comparative New World Case Study: Australia And The U.S., 1783-1914," *Radical History Review,* vol. 67, pp. 35-78, 1997.

[7] W. Sombart, *Why Is There No Socialism in America?* New York, 1976 (1906).

[8] S. M. Lipset, *Continental divide : the values and institutions of the United States and Canada*. New York: Routledge, 1990.

[9] S. M. Lipset and G. W. Marks, *It didn't happen here : why socialism failed in the United States*, 1st ed. New York: W.W. Norton & Co., 2000.

[10] R. Archer, "Labour Politics in the New World: Werner Sombart and the United States," *Journal of Industrial Relations,* vol. 49, pp. 459-482, 2007.

[11] L. Cox, "Review Essay: Revisiting the Labour Question in the United States," *Thesis Eleven,* vol. 100, pp. 168-178, 2010.

[12] H. D. Forbes, "Hartz-Horowitz at Twenty: Nationalism, Toryism and Socialism in Canada and the United States," *Canadian Journal of Political Science,* vol. 20, pp. 287-315, 1987.

[13] S. M. Lipset, *Agrarian socialism*. Berkeley,: University of California Press, 1950.

[14] S. Blumin, "Mobility and Change in Ante-Bellum Philadelphia," in *Nineteenth-Century Cities*, S. Thernstrom and R. Sennett, Eds., ed New Haven: Yale University Press, 1969, pp. 165-208.

[15] P. R. Knights, *The plain people of Boston, 1830-1860: A study in city growth*. New York: Oxford University Press, 1971.

[16] J. Modell, "The Peopling of a Working-Class Ward: Reading, Pennsylvania, 1850," *Journal of Social History,* vol. 5, pp. 71-95, 1971.

[17] S. Thernstrom, *The other Bostonians; poverty and progress in the American metropolis, 1880-1970*. Cambridge: Harvard University Press, 1973.

[18] H. M. Gitelman, *Workingmen of Waltham: Mobility in American urban industrial development, 1850-1890*. Baltimore: Johns Hopkins University Press, 1974.

[19] M. B. Katz, *The people of Hamilton, Canada West: Family and class in a mid-nineteenth-century city*. Cambridge: Harvard University Press, 1975.

[20] D. Gagan, "Geographical and social mobility in nineteenth-century Ontario: a microstudy*," *Canadian Review of Sociology/Revue canadienne de sociologie,* vol. 13, pp. 152-164, 1976.

[21] M. B. Katz, M. J. Doucet, and M. J. Stern, "Migration and the Social Order in Erie County, New York: 1855," *The Journal of Interdisciplinary History,* vol. 8, pp. 669-701, 1978.

[22] L. Glasco, "Migration and Adjustment in the Nineteenth-Century City: Occupation, Property, and Household Structure of Native-Born Whites, Buffalo, New York, 1855," in *Family and Population in Nineteenth Century America*, T. Hareven and M. A.

Vinovkis, Eds., ed Princeton: Princeton University Press 1978, pp. 154-178.
[23] C. Griffen and S. Griffen, *Natives and newcomers : the ordering of opportunity in mid-nineteenth-century Poughkeepsie*. Cambridge: Harvard University Press, 1978.
[24] G. Darroch, "Migrants in the Nineteenth Century: Fugitives or Families in Motion?," *Journal of Family History,* vol. 6, pp. 257-277, 1981.
[25] D. Gagan, *Hopeful Travelers: Families, Land and Social Change in Mid-Victorian Peel County, Canada West*. Toronto: University of Toronto Press, 1981.
[26] T. Dublin, "Rural-Urban Migrants in Industrial New England: The Case of Lynn, Massachusetts, in the Mid-Nineteenth Century," *The Journal of American History,* vol. 73, pp. 623-644, 1986.
[27] J. Long and J. Ferrie, "The path to convergence: intergenerational occupational mobility in Britain and the US in three eras," *Economic Journal,* vol. 117, pp. C61-C71, 2007.
[28] S. Ruggles, E. Roberts, S. Sarkar, and M. Sobek, "The North Atlantic Population Project: Progress and Prospects " *Historical Methods,* vol. 44, pp. 1-6, 2011.
[29] Minnesota Population Center, *North Atlantic Population Project: Complete Count Microdata. Version 2.0*. [machine readable database]. Minneapolis, MN: Minnesota Population Center [distributor], 2008.
[30] E. Roberts, S. Ruggles, Lisa Y. Dillon, Ó. Gar∂arsdóttir, J. Oldervoll, G. Thorvaldsen*, et al.*, "The North Atlantic Population Project: An Overview," *Historical Methods,* vol. 36, pp. 80-88, 2003.
[31] L. Gu, R. Baxter, D. Vickers, and C. Rainsford, "Record linkage: Current practice and future directions," Canberra2003.
[32] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association,* vol. 64, pp. 1183-1210, 1969.
[33] M. A. Jaro, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association,* vol. 84, pp. 414-420, 1989.
[34] T. R. Belin and D. B. Rubin, "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association,* vol. 90, pp. 694-707, 1995.
[35] M. D. Larsen and D. B. Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association,* vol. 96, pp. 32-41, 2001.
[36] W. E. Winkler, "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," presented at the American Statistical Association Proceedings of the Section of Survey Research Methods, 1993.
[37] P. Christen, "Automatic Record Linkage Using Seeded Nearest Neighbour And Support Vector Machine Classification," presented at the Proceedings of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, Las Vegas (NV), 2008.
[38] M. J. Wisselgren, S. Edvinsson, M. Berggren, and M. Larsson, "Testing Methods of Record Linkage on Swedish Censuses," *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* vol. 47, pp. 138-151, 2014.
[39] E. H. Porter and W. E. Winkler, "Approximate String Comparison and its Effect on an Advanced Record Linkage System," U.S. Bureau of the Census, Washington D.C.1997.
[40] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," presented at the American Statistical Association Proceedings of the Section of Survey Research Methods, 1990.
[41] R. Vick and L. Huynh, "The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway," *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* vol. 44, pp. 15 - 24, 2011.
[42] P. Christen, T. Churches, and J. X. Zhu, "Probabilistic Name and Address Cleaning and Standardisation," presented at the Proceedings of the Australasian Data Mining Workshop, Canberra, 2002.
[43] J. I. Maletic and A. Marcus, "Data cleansing: Beyond integrity analysis," presented at the Proceedings of the Conference on Information Quality, Boston, 2000.
[44] L. Nygaard, "Name Standardization in Record Linkage: an Improved Algorithmic Strategy," *History and Computing,* vol. 4, pp. 63-74., 1992.
[45] L. Philips, "The Double-Metaphone Search Algorithm," *C/C++ User's Journal,* vol. 18, 2000.
[46] A. J. Lait and B. Randell, "An Assessment of Name Matching Algorithms," Department of Computing Science, University of Newcastle upon Tyne, Newcastle1993.
[47] P. Christen, "Development and user experiences of an open source data cleaning, deduplication and record linkage system," presented at the SIGKDD, 2009.
[48] P. Christen, *Data matching*. Berlin: Springer, 2012.
[49] Z. Fu, P. Christen, and J. Zhou, "A Graph Matching Method for Historical Census Household Linkage," in *Advances in Knowledge Discovery and Data Mining*, ed: Springer, 2014, pp. 485-496.
[50] Z. Fu, M. Boot, P. Christen, and J. Zhou, "Automatic Record Linkage of Individuals and Households in Historical Census Data," *International Journal of Humanities and Arts Computing,* 2014.
[51] L. Antonie, P. Baskerville, K. Inwood, and A. Ross, "Creating Longitudinal Data from Canadian Historical Censuses," Department of Economics, University of Guelph, Guelph2011.
[52] L. Antonie, P. Baskerville, K. Inwood, and A. Ross, "An Automated Record Linkage System – Linking 1871 Canadian census to 1881 Canadian Census," Department of Economics, University of Guelph, Guelph2010.
[53] L. Antonie, K. Inwood, D. J. Lizotte, and J. A. Ross, "Tracking people over time in 19th century Canada for longitudinal analysis," *Machine Learning,* vol. 95, pp. 129-146, 2014.
[54] S. Abe, *Support vector machines for pattern classification*, 1st ed. New York: Springer, 2010.
[55] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2000.
[56] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.
[57] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," Department of Computer Science, National Taiwan University, Taipei2007.
[58] M. H. D. van Leeuwen, I. Maas, and A. Miles, "Creating a Historical International Standard Classification of Occupations," *Historical Methods,* vol. 37, pp. 186-197, 2004.
[59] M. H. D. van Leeuwen, I. Maas, and A. Miles, *Historical International Standard Classification of Occupations*. Leuven: Leuven University Press, 2002.
[60] S. Edvinsson and J. Karlsson, "Recoding occupations in the Demographic Data Base into HISCO," HISMA Berlin1998.
[61] E. Roberts, M. Woollard, C. Ronnander, L. Y. Dillon, and G. Thorvaldsen, "Occupational Classification in the the North Atlantic Population Project," *Historical Methods,* vol. 36, pp. 89-96, 2003.
[62] M. Sobek, "Work, Status, and Income: Men in the American Occupational Structure since the Late Nineteenth Century," *Social Science History,* vol. 20, pp. 169-207, 1996.
[63] J. Long and J. Ferrie, "Intergenerational Occupational Mobility in Britain and the U.S. Since 1850," *American Economic Review,* vol. 103, pp. 1109-1137, 2013.
[64] P. M. E. Altham and J. E. Ferrie, "Comparing Contingency Tables," *Historical Methods,* vol. 40, pp. 3-16, 2007.