II. Details on data construction

In this appendix, we provide more detail on the construction of the linked datasets. The process of building the dataset is as follows: first, build a "master" intragenerational dataset that tracks individuals over their lifetimes after chaining the links available from the Census Tree together. Second, build an intergenerational dataset that attaches the parents' outcomes to the children in this master intragenerational dataset. Third, build a multigenerational dataset that links grandparents to their grandchildren.

Before creating any of these datasets, we download all census-to-census crosswalks from censustree.org that link censuses that are at most 40 years apart (e.g., 1850-1860 links, 1850-1870 links, 1850-1880 links, 1860-1870 links, etc.). This process described below can also be used to create datasets when using alternative linking methods, such as those from the Census Linking Project discussed in (Abramitzky et al., 2021) or the Multigenerational Longitudinal Panel from IPUMS.

II.A. Construction of the intragenerational sample

Our goal is to eventually have an analytical sample where we track individual wealth over a 10-year gap. To get to that sample, we first build a "master" intragenerational panel where we link everyone over their entire life, using data from the Census Tree. Note that the Census Tree provides only census-census links for a given pair of censuses, but does not provide a single dataset that tracks people over their entire lives by chaining links. We then weight this master intragenerational panel to be representative of the population. Finally, we take the subsample from this "master" intragenerational panel that fit the age- and 10-year gap restrictions for analysis.

Building a master intragenerational panel

Our first goal is to create a single "master" unbalanced panel dataset that contains the HISTID for every possible person linked across Census years between 1850 and 1940. This will be done by taking all of the pairwise census links from the Census Tree, and merging them all together to form a single panel. HISTID is an individual-level identifier that is unique to each decennial census but varies within individuals across censuses. Each row of this dataset represents an individual, and each column contains the HISTID for different censuses. The panel is unbalanced because not every person will be found in each census due to linkage failure, death, or out migration out of the US. See Table B.1 as an example. This panel of HISTIDs will then be merged with the full-count census data

from IPUMS to attach observable characteristics, such as whether one has live-in servants in the household.[30]

Table B.1: Example Master Intragenerational Dataset

Person	histid1850	histid1860	histid1870	histid1880	histid1900	histid1910	histid1920	histid1930
1	X	X	.	X	.	.	.	.
2	.	X	X	X	X	.	.	.
3	.	.	.	.	.	X		X
4	.	.	.	.	X	X	X	.
5	.	.	.	X	X	.	.	.

Notes: This is an example of a "master" intragenerational dataset that merges all of the pairwise census links from the Census Tree together. "X" denotes having an observation, and "." denotes missing. This example goes through 1930, but the actual dataset goes through 1940.

To create the master intragenerational panel, we start with the Census Tree's dataset of 1930-1940 links and merge them with the 1920-1930 dataset based on the 1930 HISTID variable. Those successfully merged provide us with three HISTIDs spanning 1920-1930-1940. We also keep unmerged individuals from the 1920-1930 and 1930-1940 datasets. Next, we merge this combined dataset with the 1910-1920 linked dataset, preserving individuals successfully linked across 1910-1920-1930-1940 (four observations). We continue this iterative merging process using adjacent census decades until we reach the 1850 Census. Given the absence of the 1890 microdata due to a fire, we allow for a 20-year merge window between 1880 and 1900. Ultimately, we obtain a master intragenerational dataset comprising individuals successfully linked to at least one adjacent census.

Following this initial merge using adjacent available censuses, we then merge in the datasets that contain links 20 years apart. For example, someone named "Zachary Ward" may not be linked between 1930 and 1940, but is linked between 1920 and 1940. One way this could occur is due to an incorrect age report in 1930, but not in 1920 or 1940 (e.g., the age in 1940 is listed as 43, 30 in 1930, and 23 in 1920). With this merge, we add new links to the master intragenerational panel. We are also filling in links that were not successfully made. While unusual, it is possible that someone is successfully linked between 1930-1940 and 1920-1940, but not between 1920-1930. Thus, a link will be "triangulated" from two other linked datasets. A final possibility is that someone is linked between 1920-1930 and 1930-1940, but a *different* link is made between 1920 and 1940. This could occur for common names, where there are multiple potential links, but a

[30]While we often refer to HISTID, in practice we use a variable we term "HISTID_SHORT," which is a 6-character alphanumeric string variable. With 26 letters and 10 numbers, HISTID_SHORT can uniquely identify up to (36^6) 2,176,782,336 observations. This is in contrast to the 36-character HISTID variable from IPUMS, which adds substantially to computational costs. We create a crosswalk between HISTID and HISTID_SHORT to merge in other observables from IPUMS.

different one is chosen in a 20-year gap than in 10-year gaps. If there is a conflict in the HISTIDs, we drop them from the dataset.

We repeat this iterative process for 30-year links and 40-year links to add to and fill in HISTIDs for the master intragenerational panel. While we only use information on links 40 years apart from the Census Tree, a chain of links could cause an individual to be observed more than 40 years apart. For example, if someone is linked in an 1880-1920 dataset, but also a 1920-1940 dataset, we would chain these links and thus track them between 1880 and 1940.

After completing all the merges, we have a dataset of 134 million individuals.[31] The distribution of the number of links per row is shown in Table **??**, where for most of the dataset, we observe an individual in 2 or 3 Censuses. For over 1,000 individuals, we observe them in every available census between 1850 and 1940.

Table B.2: Number of observations in master intragenerational panel

Number of Censuses attached	Frequency	Percent	Cumulative
2	58,132,671	43.33	43.33
3	31,237,640	23.29	66.62
4	21,742,111	16.21	82.83
5	15,965,665	11.90	94.73
6	5,000,832	3.73	98.45
7	1,787,078	1.33	99.79
8	275,237	0.21	99.99
9	10,574	0.01	100.00
Total	134,151,808		100.00

Weighting the master intragenerational panel for analysis

After linking individuals, we weight the data using inverse propensity weights, following the method described by (Bailey et al., 2020) and (Bailey et al., 2020). To do so, we take all individuals who are successfully linked to a given census year and ten years prior (e.g., 1860 links who are also observed in 1850). We then merge in observable characteristics via HISTID from the given full-count census, such as age, race, marital status and literacy/education level. We then pool this linked sample with the full-count

[31] Technically, it is possible that the same person could appear as two different individuals. For example, if "Zachary Ward" was linked between 1910-1920 and 1930-1940, but no successful link was made in overlapping censuses.

census data from the same year, and create an indicator for being in the linked dataset or the full-count data. We only keep those between 10 and 75 years of age in this second census, as those younger than 10 should not be linked 10 years prior and those older than 75 are outside the age range for our analysis.

We then estimate a probit model that uses the observable characteristics to predict successful linkage. We do this process separately for each decennial census (e.g., 1860, 1870, 1880, 1910, 1920, 1930, 1940). This probit model includes the following predictors:

1. Indicator for being in the top 1% of wealth, when available in 1870, 1930, or 1940 (personal and real property in 1870, home values in 1930, 1940)

2. Indicator for having multiple servants in the household, when available between 1880-1920. If wealth data is available, we use that variable to mark elite households rather than the servant measure.

And then we *fully interact* the above indicators with the below predictors, when available:

1. Indicators for age, rounded to the nearest decade (e.g., 15-24 rounded to 20)

2. Literate, or if in 1940, education level[32]

3. White

4. Ever married

5. Relationship to the household head[33]

6. Lives in an urban area

7. Male

8. Male interacted with ever married

9. Census division of residence

[32]Educational attainment is based on six categories: 0-3 years of schooling, 4-6 years, 7-8 years, 9-11 years, 12 years, more than 12 years

[33]Indicators for head; spouse of head; other relative of head; non-relative

Because each of the above variables are interacted with both being in the top 1% or having multiple servants, the data will be weighted to be representative of elite families, and also be weighted to represent the bottom 99%.

Based on these observables and the probit coefficients, each observation in the linked sample has a predicted probability $\hat{p}$ of being in the linked sample. From this, we create the weights for our linked sample as $w = \frac{1-\hat{p}}{\hat{p}} \times \frac{l}{1-l}$ where l is the fraction of the linked and full-count sample that is in the linked sample (Bailey et al., 2020). As a final step, we winsorize the weights at the 1st and 99th percentiles to ensure that extreme weights do not influence analysis.

See Tables B3-B9 for the representativeness of the master intragenerational linked sample. Generally, these samples indicate that we are more likely (depending on the year) to link individuals who are white, married, literate/educated, and from the Midwest. While these biases are not large, the weighting process addresses them. Although our weighted sample still shows statistically different observables from the full-count sample due to the large sample sizes, these differences are economically insignificant.

Final sample restrictions for analysis

This master intragenerational panel, now with weights, can be used for any project that requires linked data. For the purpose of our paper, we take the final steps of narrowing the "master" intragenerational sample to the analytical sample that is used for this paper's question of the persistence of the top 1% of wealth. We keep links that meet these criteria: (1) observed 10 years apart, (2) age 25-55 at the first observation, (3) age 35-65 at the second observation, (4) are in non-group quarters, and (5) we can create a weight for the individual (that is, they have observables such as age, urban, birthplace, etc.).

Table B.3: Representativeness of Intragenerational Linked Sample in 1860

	Bottom 99% wealth				Top 1% wealth			
	Linked Sample				Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	29.589	32.196	29.734	0.000	30.726	33.301	30.891	0.006
ECDF of wealth	0.495	0.495	0.491	0.000	0.995	0.995	0.995	0.000
Male	0.512	0.506	0.514	0.000	0.831	0.840	0.831	0.852
White	0.978	0.988	0.980	0.000	0.995	0.997	0.995	0.935
Black	0.019	0.012	0.020	0.000	0.004	0.003	0.005	0.489
Other race	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.052
Urban	0.249	0.179	0.240	0.000	0.220	0.207	0.218	0.381
Ever married	0.450	0.501	0.450	0.928	0.530	0.585	0.531	0.547
Literate	0.626	0.643	0.622	0.000	0.784	0.799	0.784	0.769
Head of household	0.264	0.302	0.265	0.001	0.663	0.698	0.663	0.991
Northeast	0.400	0.404	0.401	0.002	0.255	0.261	0.256	0.629
Midwest	0.319	0.314	0.320	0.000	0.215	0.203	0.214	0.817
South	0.256	0.273	0.259	0.000	0.510	0.528	0.511	0.542
West	0.025	0.009	0.021	0.000	0.020	0.008	0.018	0.025
Sample size	19311095	7936663	7936663		195065	80172	80172	

Table B.4: Representativeness of Intragenerational Linked Sample in 1870

	Bottom 99% wealth				Top 1% wealth			
		Linked Sample				Linked Sample		
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	30.014	32.680	30.341	0.000	31.277	34.217	31.560	0.000
ECDF of wealth	0.495	0.499	0.492	0.000	0.995	0.995	0.995	0.005
Male	0.502	0.491	0.516	0.000	0.824	0.835	0.828	0.008
White	0.876	0.982	0.905	0.000	0.975	0.995	0.977	0.085
Black	0.121	0.018	0.093	0.000	0.024	0.005	0.023	0.115
Other race	0.003	0.000	0.003	0.444	0.000	0.000	0.000	0.174
Urban	0.273	0.228	0.280	0.000	0.363	0.368	0.362	0.690
Ever married	0.447	0.496	0.448	0.000	0.570	0.622	0.574	0.038
Literate	0.782	0.867	0.811	0.000	0.935	0.952	0.940	0.000
Head of household	0.267	0.294	0.273	0.000	0.650	0.690	0.654	0.012
Northeast	0.331	0.369	0.348	0.000	0.399	0.431	0.403	0.040
Midwest	0.330	0.364	0.343	0.000	0.371	0.367	0.372	0.476
South	0.312	0.249	0.283	0.000	0.197	0.176	0.192	0.004
West	0.027	0.018	0.026	0.000	0.033	0.025	0.032	0.434
Sample size	27618162	10179147	10179147		278975	105440	105440	

Table B.5: Representativeness of Intragenerational Linked Sample in 1880

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	30.750	32.235	30.831	0.000	33.078	34.150	33.155	0.495
Number of Servants	0.009	0.010	0.010	0.000	2.577	2.541	2.544	0.000
Male	0.510	0.481	0.512	0.000	0.478	0.462	0.480	0.645
White	0.876	0.919	0.878	0.000	0.984	0.991	0.985	0.926
Black	0.121	0.081	0.122	0.000	0.014	0.009	0.015	0.366
Other race	0.003	0.000	0.001	0.000	0.001	0.000	0.000	0.000
Urban	0.281	0.232	0.277	0.000	0.598	0.565	0.596	0.561
Ever married	0.460	0.479	0.460	0.128	0.460	0.473	0.460	0.906
Literate	0.828	0.864	0.829	0.000	0.972	0.978	0.972	0.985
Head of household	0.274	0.280	0.276	0.000	0.262	0.262	0.262	0.884
Northeast	0.305	0.313	0.304	0.000	0.414	0.390	0.413	0.782
Midwest	0.348	0.372	0.348	0.001	0.189	0.197	0.188	0.865
South	0.310	0.288	0.312	0.000	0.363	0.381	0.366	0.532
West	0.037	0.028	0.035	0.000	0.034	0.032	0.034	0.597
Sample size	36267169	14918455	14918455		61367	29452	29452	

Table B.6: Representativeness of Intragenerational Linked Sample in 1910

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	32.300	31.419	32.357	0.000	36.694	35.046	36.693	0.980
Number of Servants	0.021	0.025	0.024	0.000	2.665	2.608	2.611	0.000
Male	0.518	0.474	0.517	0.000	0.479	0.448	0.480	0.418
White	0.891	0.927	0.893	0.000	0.990	0.994	0.990	0.908
Black	0.105	0.072	0.105	0.000	0.008	0.006	0.009	0.049
Other race	0.004	0.001	0.002	0.000	0.002	0.000	0.001	0.000
Urban	0.476	0.421	0.471	0.000	0.799	0.800	0.799	0.892
Ever married	0.471	0.472	0.474	0.000	0.504	0.486	0.504	0.778
Literate	0.915	0.946	0.919	0.000	0.995	0.997	0.995	0.863
Head of household	0.278	0.253	0.281	0.000	0.299	0.247	0.298	0.971
Northeast	0.289	0.258	0.285	0.000	0.528	0.510	0.528	0.983
Midwest	0.330	0.373	0.333	0.000	0.207	0.220	0.207	0.936
South	0.303	0.299	0.306	0.000	0.199	0.204	0.199	0.906
West	0.078	0.070	0.076	0.000	0.067	0.066	0.066	0.784
Sample size	70312494	35690840	35690840		307841	178547	178547	

Table B.7: Representativeness of Intragenerational Linked Sample in 1920

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	33.239	31.908	33.244	0.131	37.692	35.649	37.671	0.748
Number of Servants	0.011	0.013	0.013	0.000	2.728	2.671	2.671	0.000
Male	0.512	0.485	0.512	0.000	0.476	0.457	0.477	0.512
White	0.898	0.935	0.901	0.000	0.993	0.997	0.994	0.710
Black	0.098	0.063	0.096	0.000	0.005	0.003	0.005	0.275
Other race	0.004	0.002	0.003	0.000	0.002	0.001	0.001	0.003
Urban	0.530	0.482	0.527	0.000	0.822	0.825	0.821	0.869
Ever married	0.498	0.487	0.499	0.000	0.527	0.514	0.528	0.846
Literate	0.939	0.963	0.940	0.000	0.996	0.998	0.997	0.685
Head of household	0.293	0.263	0.295	0.000	0.325	0.271	0.325	0.972
Northeast	0.285	0.262	0.283	0.000	0.541	0.531	0.541	0.968
Midwest	0.326	0.359	0.327	0.000	0.195	0.205	0.195	0.820
South	0.303	0.295	0.303	0.000	0.194	0.194	0.194	0.800
West	0.087	0.084	0.087	0.000	0.070	0.070	0.070	0.901
Sample size	81338129	45737483	45737483		183600	113206	113206	

Table B.8: Representativeness of Intragenerational Linked Sample in 1930

	Bottom 99% wealth				Top 1% wealth			
	Linked Sample				Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	33.936	32.769	33.939	0.188	35.064	34.327	35.220	0.000
Number of Servants	0.015	0.017	0.016	0.000	0.287	0.290	0.301	0.000
ECDF of wealth	0.495	0.498	0.496	0.000	0.995	0.995	0.995	0.008
Male	0.507	0.503	0.509	0.000	0.402	0.417	0.403	0.190
White	0.900	0.938	0.905	0.000	0.978	0.983	0.978	0.681
Black	0.095	0.060	0.092	0.000	0.021	0.017	0.021	0.817
Other race	0.004	0.003	0.004	0.000	0.001	0.001	0.001	0.423
Urban	0.577	0.541	0.575	0.000	0.795	0.759	0.795	0.852
Ever married	0.505	0.491	0.505	0.000	0.808	0.797	0.807	0.057
Literate	0.955	0.971	0.957	0.000	0.987	0.991	0.987	0.802
Head of household	0.299	0.274	0.300	0.000	0.379	0.374	0.380	0.530
Northeast	0.285	0.269	0.284	0.000	0.420	0.375	0.420	0.947
Midwest	0.318	0.349	0.320	0.000	0.306	0.331	0.306	0.775
South	0.297	0.285	0.297	0.000	0.175	0.188	0.175	0.924
West	0.099	0.097	0.100	0.004	0.099	0.106	0.099	0.835
Sample size	96097479	57522674	57522674		970683	581798	581798	

Table B.9: Representativeness of Intragenerational Linked Sample in 1940

	Bottom 99% wealth				Top 1% wealth			
	Linked Sample				Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	35.191	34.148	35.154	0.000	36.277	35.756	36.444	0.000
Number of Servants	0.013	0.014	0.013	0.464	0.213	0.213	0.222	0.000
ECDF of wealth	0.495	0.499	0.496	0.000	0.995	0.995	0.995	0.000
Male	0.502	0.514	0.504	0.000	0.412	0.445	0.412	0.703
White	0.903	0.935	0.907	0.000	0.980	0.984	0.980	0.186
Black	0.093	0.062	0.090	0.000	0.019	0.015	0.019	0.574
Other race	0.004	0.003	0.004	0.000	0.001	0.001	0.001	0.002
Urban	0.572	0.550	0.571	0.000	0.647	0.614	0.648	0.109
Ever married	0.514	0.497	0.515	0.000	0.818	0.808	0.819	0.313
Educ. (years)	8.579	8.722	8.597	0.000	12.075	11.911	12.062	0.428
Head of household	0.314	0.297	0.316	0.000	0.405	0.420	0.405	0.404
Northeast	0.280	0.274	0.279	0.000	0.341	0.313	0.340	0.244
Midwest	0.307	0.332	0.309	0.000	0.298	0.316	0.298	0.707
South	0.306	0.287	0.304	0.000	0.230	0.233	0.230	0.705
West	0.107	0.107	0.108	0.000	0.132	0.138	0.133	0.103
Sample size	107300508	69765338	69765338		1083849	714448	714448	

II.B. Construction of the intergenerational dataset

In this section, we discuss how we build the intergenerational dataset used for the analysis of parents with servants and children with servants. We follow a similar process as when we build the intragenerational data. First, we build a "master" intergenerational panel that links children to their parents. We create custom weights for this sample based on the observables of linked children. Finally, we limit the sample to the one used in this paper for analytical purposes.

Building a Master Intergenerational Panel

Building the master intergenerational panel from the master intragenerational panel is straightforward. We locate children HISTIDs in the intragenerational panel and merge in their parents' HISTIDs as separate columns (See Table B.10). To do this, we need to create two crosswalks for each decennial census: (1) between the HISTID of the child and the HISTID of the father, and (2) between the HISTID of the child and the HISTID of the mother. These crosswalks are based on the POPLOC and MOMLOC variables from IPUMS, which identify the person number of the father and mother within the household. Note that parents are social relationships rather than biological ones, and so we include adoptive and step-parents. We only merge in parents information for children 14 years old or younger for this crosswalk, as children older than 14 who still live in the household may be a selected group.

Table B.10: Example Master Intergenerational Dataset

Person	histid_G2_1850	...	histid_G1_F_1850	...	histid_G1_M_1850
1	X	.	X	.	X
2	.	.	X	.	.
3	X	.	X	.	X
4	X	.	.	.	X
5	X	.	X	.	X

Notes: This is an example of the master intergenerational dataset, where an "X" denotes having an observation, and "." denotes missing. HISTID is a unique identified for each person and census. HISTID_G2 is the second generation (or child's) HISTID. HISTID_F_1850 is the father's HISTID and HISTID_M_1850 is the mother's. This example is just of 1850, but the actually dataset includes columns for each decennenial census until 1940.

This first step of merging in the father's and mother's HISTIDs can be done using cross-sectional census data. However, it is also useful to have multiple observations of the parents, such as averaging the father's occupational status in order to account for measurement error (Ward, 2023). To fill in the rest of the father and mother's HISTIDS

between 1850 and 1940, we merge the father's and mother's HISTIDs back into the master intragenerational panel, which linked people over their entire lives.

Weighting the Master Intergenerational Panel

Getting into the Intergenerational Panel is nonrandom because people are not randomly linked. We create custom weights for the panel broadly following the method discussed in Bailey et al. (2020) where we pool the linked sample with full-count census data, predict successful linkage based on observable characteristics, and then use the predicted probability as inverse probability weights, same as the intragenerational sample.

Our process is as follows. For each census year between 1870 and 1940, we retain the HISTIDs of successfully linked second-generation (G2) adults. Since we start with G2 adult outcomes, we weight based on their adult characteristics, not their childhood ones. Thus, the dataset is conceptually similar to modern surveys like the General Social Survey, which asks adults about their father's occupation when the respondent was 16 and then weights the sample of respondents to reflect population characteristics. However, of course, our data has a direct observation of the father's occupation instead of a recall.

Next, for a given census (e.g., 1940 Census), we keep observations where both the child and a parent is observed either 20, 30, or 40 years earlier. Thus, our intergenerational sample will also be based on parent-child links observed at least 20 and at most 40 years apart.

One issue is that most Black children cannot be linked to the 1850 and 1860 Censuses, pre emancipation. However, it is important for the racial composition of intergenerational sample to reflect the population to accurately capture an inequality statistic like intergenerational mobility (Ward, 2023). For the 1870 and 1880 censuses, we append a random sample of adult Black males and females to account for the formerly enslaved population observed as adults but not as children. Finally, we append the full-count census data from IPUMS for each year, keeping those in the same age, birthplace, and occupation range as the G2 adults.

Similar as the intragenerational panel, we then estimate a probit model that uses the observable characteristics to predict successful linkage. We do this process separately for each decennial census for adult children (e.g., 1870, 1880, 1900, 1910, 1920, 1930, 1940). This probit model includes the following predictors:

1. Indicator for being in the top 1% of home values, when available in 1930 or 1940

2. Indicator for having multiple servants in the household, between 1900-1920

And then we *fully interact* the above indicators with the below predictors, when available:

1. Indicators for age, rounded to the nearest decade (e.g., 15-24 rounded to 20)

2. Literate, or if in 1940, education level[34]

3. White

4. Ever married

5. Relationship to the household head[35]

6. Lives in an urban area

7. Male

8. Male interacted with ever married

9. Census division of residence

Because each of the above variables are interacted with both being in the top 1% or having multiple servants, the data will be weighted to be representative of elite families, and also be weighted to represent the bottom 99%.

Based on these observables and the probit coefficients, each observation in the linked sample has a predicted probability $\hat{p}$ of being in the linked sample. From this, we create the weights for our linked sample as $w = \frac{1-\hat{p}}{\hat{p}} \times \frac{l}{1-l}$ where l is the fraction of the linked and full-count sample that is in the linked sample (Bailey et al., 2020). As a final step, we winsorize the weights at the 1st and 99th percentiles to ensure that extreme weights do not influence analysis. The main effect of this is to reduce the weights of some Black

[34]Educational attainemnt is based on six categories: 0-3 years of schooling, 4-6 years, 7-8 years, 9-11 years, 12 years, more than 12 years

[35]Indicators for head; spouse of head; other relative of head; non-relative

individuals, who are very difficult to link, such that our weighted sample still has a lower fraction Black than the population.

Tables B11-B15 report the representativeness of the linked sample by year of the child's observation (1900-1940). The table is split by whether the child is observed without multiple servants in adulthood (the left-side columns) or with multiple servants (the right-side columns). In general, we find little meaningful economic difference in average observable characteristics. However, we will note that often the Black share of the sample does not reach its target of the full population due to the winsorizing of weights at the 99%, as Black individuals are less likely to be linked and thus are more highly weighted.

Choosing the primary parent and child observations for analysis

At this point, the master intergenerational panel includes multiple observations for the father, mother, and child, along with the weights. However, this dataset is not yet suitable for analysis as it is unclear which of the multiple father or child observations to use in the analysis when estimating the relationship between parental economic status and child economic status.

We choose the primary child observation to be the one that is observed closest to age 40. We pick age 40 as it is in the middle of the lifecycle, which is important for reducing lifecycle bias in intergenerational mobility estimates. We choose the primary parental observations as the ones when they are in the same house as the child. If the child is observed in the parental household multiple times (e.g., age 2 in 1900 and age 12 in 1910), then we pick the later observation as the parent observation.

Final Sample Restrictions for Analysis

This construction of the intergenerational panel can be used for any type of intergenerational analysis, but is not fully suited for this paper's focus on the persistence of elite families. We need to make final adjustments for the analysis, partially based on when the servant measure is available between 1880 and 1940. We keep observations where the child is (1) between age 25-55, (2) observed in non-group quarters, (3) the primary parent observation is between 1880 and 1920, (4) the primary child observation is between 1900 and 1940, and (5) children have a weight.

Table B.11: Representativeness of Intergenerational Linked Sample in 1900

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	32.592	32.345	32.452	0.000	35.008	34.049	34.859	0.002
Number of Servants	0.032	0.039	0.036	0.000	2.606	2.565	2.569	0.000
Male	0.505	0.517	0.520	0.000	0.463	0.462	0.466	0.134
White	0.864	0.952	0.874	0.000	0.994	0.998	0.995	0.582
Black	0.133	0.048	0.126	0.000	0.005	0.002	0.005	0.792
Other race	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.030
Urban	0.374	0.329	0.374	0.208	0.734	0.724	0.734	0.752
Ever married	0.572	0.608	0.564	0.000	0.557	0.549	0.556	0.424
Literate	0.892	0.941	0.901	0.000	0.990	0.993	0.991	0.662
Head of household	0.315	0.332	0.317	0.000	0.280	0.257	0.276	0.034
Northeast	0.255	0.257	0.257	0.000	0.514	0.502	0.512	0.547
Midwest	0.340	0.393	0.343	0.000	0.245	0.257	0.245	0.857
South	0.352	0.299	0.348	0.000	0.198	0.197	0.199	0.601
West	0.052	0.051	0.052	0.000	0.043	0.043	0.043	0.947
Sample size	31,141,261	10,919,187	10,919,187		200,700	83,669	83,669	

Table B.12: Representativeness of Intergenerational Linked Sample in 1910

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	39.619	39.305	39.569	0.000	40.883	40.135	40.669	0.000
Number of Servants	0.028	0.034	0.032	0.000	2.664	2.631	2.636	0.000
Male	0.514	0.538	0.530	0.000	0.461	0.478	0.464	0.219
White	0.874	0.955	0.880	0.000	0.991	0.995	0.991	0.541
Black	0.123	0.045	0.118	0.000	0.009	0.005	0.008	0.827
Other race	0.003	0.001	0.001	0.000	0.001	0.000	0.000	0.092
Urban	0.446	0.396	0.446	0.043	0.797	0.798	0.796	0.796
Ever married	0.730	0.776	0.731	0.155	0.736	0.753	0.740	0.176
Literate	0.924	0.960	0.928	0.000	0.997	0.998	0.997	0.764
Head of household	0.434	0.462	0.442	0.000	0.388	0.382	0.385	0.215
Northeast	0.250	0.246	0.251	0.000	0.530	0.522	0.530	0.799
Midwest	0.328	0.381	0.331	0.000	0.202	0.213	0.203	0.582
South	0.344	0.297	0.341	0.000	0.202	0.199	0.202	0.793
West	0.078	0.076	0.078	0.000	0.065	0.066	0.065	0.957
Sample size	23929557	8371673	8371673		143614	60812	60812	

Table B.13: Representativeness of Intergenerational Linked Sample in 1920

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	33.437	32.517	33.464	0.000	36.994	35.967	36.864	0.033
Number of Servants	0.012	0.014	0.014	0.000	2.705	2.645	2.656	0.000
Male	0.501	0.522	0.516	0.000	0.461	0.477	0.464	0.237
White	0.878	0.954	0.887	0.000	0.994	0.998	0.994	0.649
Black	0.120	0.045	0.111	0.000	0.006	0.002	0.006	0.853
Other race	0.002	0.000	0.001	0.000	0.001	0.000	0.000	0.274
Urban	0.512	0.469	0.510	0.000	0.821	0.823	0.820	0.800
Ever married	0.608	0.617	0.605	0.000	0.634	0.621	0.634	0.960
Literate	0.954	0.977	0.958	0.000	0.998	0.998	0.998	0.560
Head of household	0.325	0.329	0.328	0.000	0.327	0.304	0.322	0.050
Northeast	0.240	0.234	0.241	0.040	0.530	0.521	0.529	0.816
Midwest	0.331	0.381	0.333	0.000	0.198	0.211	0.198	0.879
South	0.344	0.299	0.341	0.000	0.204	0.200	0.205	0.678
West	0.085	0.087	0.085	0.000	0.068	0.068	0.068	0.980
Sample size	44608027	16819134	16819134		111352	46226	46226	

Table B.14: Representativeness of Intergenerational Linked Sample in 1930

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	33.806	30.248	30.787	0.000	37.014	31.924	33.085	0.000
Number of Servants	0.013	0.015	0.014	0.000	2.625	2.586	2.596	0.002
ECDF of wealth	0.500	0.500	0.500	0.179	0.670	0.644	0.656	0.000
Male	0.498	0.552	0.515	0.000	0.465	0.496	0.440	0.000
White	0.882	0.952	0.889	0.000	0.995	0.998	0.995	0.604
Black	0.115	0.047	0.109	0.000	0.004	0.002	0.004	0.899
Other race	0.003	0.001	0.003	0.000	0.001	0.001	0.001	0.187
Urban	0.567	0.536	0.568	0.000	0.789	0.791	0.811	0.000
Ever married	0.623	0.620	0.599	0.000	0.677	0.629	0.647	0.000
Literate	0.967	0.984	0.971	0.000	0.998	0.998	0.998	0.959
Head of household	0.332	0.334	0.308	0.000	0.346	0.292	0.274	0.000
Northeast	0.244	0.232	0.243	0.000	0.525	0.507	0.541	0.000
Midwest	0.323	0.366	0.325	0.000	0.211	0.231	0.207	0.081
South	0.335	0.303	0.336	0.000	0.178	0.176	0.171	0.007
West	0.097	0.099	0.096	0.000	0.085	0.086	0.081	0.008
Sample size	56069867	21086662	21086662		117260	45314	45314	

Table B.15: Representativeness of Intergenerational Linked Sample in 1940

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	34.076	34.274	33.940	0.000	37.015	36.613	36.854	0.027
Number of Servants	0.012	0.014	0.012	0.000	2.479	2.425	2.438	0.000
ECDF of wealth	0.500	0.501	0.499	0.000	0.645	0.647	0.656	0.000
Male	0.496	0.589	0.510	0.000	0.468	0.543	0.441	0.000
White	0.894	0.951	0.904	0.000	0.993	0.996	0.992	0.159
Black	0.103	0.047	0.094	0.000	0.006	0.004	0.007	0.029
Other race	0.003	0.001	0.003	0.000	0.001	0.000	0.000	0.012
Urban	0.570	0.534	0.568	0.000	0.689	0.682	0.710	0.000
Ever married	0.622	0.641	0.620	0.000	0.650	0.662	0.666	0.000
Educ. (years)	9.531	9.970	9.618	0.000	15.025	15.757	15.185	0.039
Head of household	0.337	0.399	0.340	0.000	0.349	0.378	0.313	0.000
Northeast	0.254	0.238	0.252	0.000	0.488	0.477	0.515	0.000
Midwest	0.310	0.349	0.314	0.000	0.258	0.276	0.249	0.003
South	0.330	0.301	0.326	0.000	0.152	0.146	0.140	0.000
West	0.106	0.112	0.108	0.000	0.102	0.101	0.096	0.002
Sample size	66769326	30750057	30750057		73043	36726	36726	

II.C. Construction of the Multigenerational Sample

In this section, we describe how we create a "Master" multigenerational sample, which is similar to the process as building the master intragenerational dataset and master intergenerational dataset. First, we create a master multigenerational panel that links third-generation children (G3) to their second-generation fathers and mothers (G2), as well as their first-generation paternal and maternal grandparents (G1). In essence, this process will add more columns to the intergenerational sample (recall Table B.10, but with more columns for maternal and paternal grandparent HISTIDs). After building this panel, we create custom weights based on the observable characteristics of grandchildren. We then identify the "primary" grandchild observation used in our analysis, as well as a primary second-generation and first-generation observations. Finally, we apply sample restrictions, which results in the dataset used for analytical purposes.

Building the Master Multigenerational Panel

To build the multigenerational sample the process is simple: take the parents in the intergenerational dataset, and then check whether they also children in the same dataset.

To illustrate this process with an example, take "Zachary Ward," who is a father of "Xavier Ward" in 1900. Zachary's 1900 HISTID is denoted as a G1 (first generation) HISTID, and Xavier's 1900 HISTID is denoted as a G2 (second generation) HISTID. Further suppose that Zachary Ward was also child in 1870, who has a father named Jeffrey Ward. This would show up as a separate row in the intergenerational dataset, but now Zachary's 1870 HISTID is labelled as a *second-generation* (G2) HISTID rather than a G1 HISTID. Since the intergenerational dataset tracks children across all available observations, for the row where 1870 Zachary is observed with his father, he might also be observed with a 1900 G2 HISTID as an adult. Thus, we have a situation where Zachary has born a G2 HISTID where he a child, and a G1 HISTID when he is a father. To create the multigenerational sample, we simply keep all HISTIDs where someone is both a G1 and G2 HISTID in the same year.

Because the Census Tree has links for both men and women, we are able to create three-generational chains along the mother and father lines, and thus have both paternal and maternal grandfathers.

For the purposes of the multigenerational dataset, it is useful to relabel generations to cover three generations. Thus the format of the "Master" multigenerational panel has for

each year: HISTIDs for the third-generation grandchildren (G3), HISTIDs for the second-generation fathers (G2_F) and mothers (G2_M), and HISTIDS for the first-generationl paternal fathers (G1_F_F) and mothers (G1_F_M) and maternal fathers (G1_M_F) and mothers (G1_M_M).

Weighting the master multigenerational sample

We weight the data in the exact same way as in the master intergenerational sample described above. However, rather than weighting to be representative on second-generation characteristics, we weight based on the adult characteristics of third-generation grandchildren. The dataset is thus conceptually similar to a sample of individuals who recall their parents' and grandparents' occupations, which researchers then weight to be representative on observables. Of course, because we can directly link across censuses, we can directly observe parental and grandparental observables. Therefore, our dataset likely contains less error than data based on recall.

Choosing the primary grandparent, parent, and child observations for analysis

The resulting dataset has multiple third-generation, second-generation and first-generation observations, making it unclear which one to use. It also has observations for two parents and four grandparents. Of course, not everyone is always observed due to linkage failure or underenumeration, but it remains a question as to which observations we should use in our analysis.

We follow a similar process as the intergenerational sample described above where we use the following observations as our "primary" observations: (1) the third-generation observation when they are closest in age to 40, (2) the second-generation observation when the third-generation is observed in the household, and (3) the first-generation observation when the second-generation child is observed in the household.

We are interested in the persistence of elite households across three generations, and so a further question is whether we should use the parental or maternal households' wealth as a proxy for descending from an elite household. If both maternal and grandparents are observed, we use the maximum number of servants or the maximum amount of real estate and personal property. If only one of the sets of grandparents is observed, then we use that set's information.

Final Sample Restrictions for Analysis

This construction of the master multigenerational panel can be used for any type of multigenerational analysis but is not fully suited for this paper's focus on the persistence of elite families across three generations. We need to make final adjustments for the analysis when our wealth proxies are available. We create two different samples:

1. Wealth sample: Links grandparental wealth between 1850-1870 and grandchild home values between 1930-1940.

2. Servant sample: Links grandparent servants in 1880-1900 with grandchild servants in 1920-1940.

For both samples, we include observations where:

- The grandchild is between 25 and 65 years old.

- The primary maternal/paternal grandfather is between 25 and 65 years old.

Table B.16: Representativeness of multigenerational Linked Sample in 1920

	No multiple servants Linked Sample				Multiple Servants Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	33.437	31.277	33.088	0.000	36.994	33.918	36.647	0.000
Number of Servants	0.012	0.014	0.015	0.000	2.705	2.637	2.656	0.000
Male	0.501	0.507	0.515	0.000	0.461	0.466	0.466	0.138
White	0.878	0.978	0.907	0.000	0.994	0.999	0.995	0.324
Black	0.120	0.022	0.092	0.000	0.006	0.001	0.005	0.353
Other race	0.002	0.000	0.001	0.000	0.001	0.000	0.001	0.741
Urban	0.512	0.417	0.512	0.813	0.821	0.815	0.819	0.521
Ever married	0.608	0.634	0.602	0.000	0.634	0.590	0.635	0.671
Literate	0.954	0.982	0.965	0.000	0.998	0.999	0.998	0.610
Head of household	0.325	0.313	0.322	0.000	0.327	0.263	0.322	0.139
Northeast	0.240	0.201	0.244	0.000	0.530	0.505	0.530	0.946
Midwest	0.331	0.377	0.339	0.000	0.198	0.214	0.196	0.464
South	0.344	0.333	0.331	0.000	0.204	0.217	0.206	0.453
West	0.085	0.089	0.087	0.000	0.068	0.064	0.068	0.831
Sample size	44608027	9465923	9465923		111352	27722	27722	

Table B.17: Representativeness of multigenerational Linked Sample in 1930

	Bottom 99% wealth				Top 1% wealth			
		Linked Sample				Linked Sample		
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	33.802	30.718	31.412	0.000	34.858	31.800	32.546	0.000
Number of Servants	0.015	0.018	0.018	0.000	0.360	0.395	0.395	0.000
ECDF of wealth	0.495	0.495	0.494	0.000	0.995	0.995	0.995	0.000
Male	0.499	0.535	0.524	0.000	0.365	0.415	0.361	0.042
White	0.881	0.972	0.886	0.000	0.990	0.998	0.991	0.733
Black	0.116	0.027	0.112	0.000	0.009	0.002	0.009	0.564
Other race	0.003	0.001	0.002	0.000	0.001	0.000	0.001	0.389
Urban	0.565	0.481	0.560	0.000	0.834	0.808	0.826	0.000
Ever married	0.620	0.662	0.603	0.000	0.944	0.968	0.954	0.000
Literate	0.967	0.986	0.970	0.000	0.998	0.999	0.999	0.281
Head of household	0.332	0.343	0.318	0.000	0.395	0.428	0.382	0.000
Northeast	0.243	0.186	0.238	0.000	0.425	0.354	0.413	0.000
Midwest	0.323	0.370	0.324	0.051	0.315	0.353	0.323	0.000
South	0.337	0.340	0.342	0.000	0.163	0.192	0.166	0.018
West	0.097	0.104	0.096	0.000	0.098	0.102	0.097	0.508
Sample size	55625252	12383059	12383059		561875	125085	125085	

Table B.18: Representativeness of multigenerational Linked Sample in 1940

	Bottom 99% wealth				Top 1% wealth			
	Linked Sample				Linked Sample			
	Full	Unweighted	Weighted	Full=Weight	Full	Unweighted	Weighted	Full=Weight
Age	34.069	34.611	33.990	0.000	35.135	35.571	35.109	0.397
Number of Servants	0.012	0.015	0.013	0.000	0.243	0.286	0.263	0.000
ECDF of wealth	0.495	0.495	0.486	0.000	0.995	0.995	0.995	0.079
Male	0.497	0.569	0.516	0.000	0.377	0.477	0.382	0.000
White	0.893	0.970	0.904	0.000	0.993	0.998	0.994	0.003
Black	0.104	0.030	0.094	0.000	0.006	0.002	0.006	0.077
Other race	0.003	0.001	0.002	0.000	0.001	0.000	0.000	0.000
Urban	0.569	0.477	0.564	0.000	0.667	0.628	0.662	0.000
Ever married	0.618	0.670	0.617	0.000	0.943	0.957	0.945	0.000
Educ. (years)	9.497	10.067	9.616	0.000	13.491	14.446	13.564	0.008
Head of household	0.337	0.398	0.342	0.000	0.409	0.497	0.412	0.039
Northeast	0.254	0.183	0.249	0.000	0.331	0.265	0.324	0.000
Midwest	0.310	0.359	0.315	0.000	0.309	0.334	0.311	0.158
South	0.331	0.341	0.329	0.000	0.222	0.252	0.225	0.007
West	0.105	0.118	0.108	0.000	0.138	0.149	0.140	0.050
Sample size	66173940	17996206	17996206		668429	182716	182716	

III. Linking the IPUMS samples to the full-count data

In this Appendix, we discuss how we link the IPUMS public-use samples to the full-count versions. First, we download the 1% 1850 IPUMS sample, 1.2% 1860 and 1870 IPUMS samples, as well as the 5% 1930 sample. We choose these samples because first name and last name strings are available in the public-use samples from IPUMS-USA. Unfortunately, names are unavailable for the 1940 1% sample.

We then link the samples to the full-count versions.[36] To link the datasets, we use the ABE fully automated approach using code available from the Census Linking Project as well as Ran Abramitzky's website.[37] The typical approach when linking is to only block on variables that are stable across a 10 or more year period, which excludes useful information like state and county of residence, because people may move between censuses. But because we are linking two transcriptions of the same underlying dataset in a given year, we can block on more variables that should be stable across sources. We block on the NYSIIS version of first name and last name, birth place, state, county of residence and sex.

After linking the versions of the data, calculating the wealth percentiles is straightforward. These percentiles are based on real estate in 1850, real estate plus personal estate in 1860 and 1870, and home value in 1930. Figure A.7 shows this disagreement. We find that disagreement rates increase as the percentile rises, which is expected since it is more difficult to transcribe multiple numbers correctly than a few. Top values in the 1930 Census stand out with potentially significant error, with about 59% of those in the top 0.1% of home values in the full-count data being in a smaller percentile in the sample.

When we perform out checks on whether 10-year transitions are impacted by these digitization issue, we find within the linked 10-year data the subset that is also linked to the samples. We then limit this 10-year linked sample to the subsample where the digitizations of wealth agree. We do not find substantially different downward mobility rates, as shown in Figure 3.

[36] We do this linking on the National Bureau of Economic Research's server because we need to use the restricted-access versions of the full-count data to observe names.
[37] See https://ranabr.people.stanford.edu/historical-record-linking

IV. Cleaning the Panel Study of Income Dynamics

We use the Panel Study of Income Dynamics (PSID) to measure downward mobility from the top 10% and top 1% for both home equity and total net worth. The PSID data have been used previously to estimate the intergenerational correlation of wealth (Charles and Hurst, 2003; Pfeffer and Killewald, 2018). Wealth information is available from 1984 to 2021, with 5-year intervals between 1984 and 1999, and 2-year intervals between 1999 and 2021.

The PSID provides information on total family wealth, both including and excluding home equity. This is done by summing the responses to various questions about assets (e.g., cash, stocks/bonds, home value) and then subtracting total debts. Because the historical data only have information on home values, we use the difference between the PSID measures of wealth with home equity and without home equity to create a measure of wealth based solely on home equity.

We aim to identify those in the top percentiles of wealth in the PSID and then measure their transitions over 10-year intervals. To do this, we first flag household heads between the ages of 25 and 65 in a given survey year. We then calculate each observation's place in that year's net wealth cumulative distribution, as well as their place in that year's home equity distribution. These values are created within 10-year age bands (e.g., 25-34, 35-44, etc.) and are weighted based on PSID-provided family weights.

With these measures, it is straightforward to determine if those who are in the top 10% of the net worth distribution remain there 10 years later. Estimates of downward mobility are plotted in Figure 4. Estimates of odds ratios and upward mobility rates are in Figures A.5 and A.6.