Who Becomes a Member of Congress? Evidence From De-Anonymized Census Data*

Daniel M. Thompson,† *Stanford University*

James J. Feigenbaum,‡ *Boston University and NBER*

Andrew B. Hall,§ *Stanford University*

Jesse Yoder,¶ *Stanford University*

August 12, 2019

Abstract

We link future members of Congress to the de-anonymized 1940 census to offer a uniquely detailed analysis of how economically unrepresentative American politicians were in the 20th century, and why. Future members under the age of 18 in 1940 grew up in households with parents who earned more than twice as much as the population average and who were more than 6 times as likely as the general population to hold college degrees. However, compared to siblings who did not become politicians, future members of Congress between the ages of 18 and 40 in 1940 were higher-earners and more educated, indicating that socioeconomic background alone does not explain the differences between politicians and non-politicians. Examining a smaller sample of candidates that includes non-winners, we find that the candidate pool is much higher-earning and more educated than the general population. At the same time, among the candidate pool, elections advantage candidates with higher earnings ability and education. We conclude that barriers to entry likely deter a more economically representative candidate pool, but that electoral advantages for more-educated individuals with more private-sector success also play an important role.

*The protocol for this study was approved by Stanford's IRB (protocol #42719) as well as by NBER. For research assistance, the authors thank Brittany Dutton, Qianmin Hu, Elise Kostial, Anna Nakai, Conor Orton, Josh Rose, and Anish Sundar. For helpful discussion, the authors thank Ran Abramitzky, Avi Acharya, Alberto Alesina, Matilde Bombardini, David Broockman, Nick Carnes, Dan Carpenter, Cesi Cruz, Robert Erikson, Anthony Fowler, Matt Gentzkow, Liz Gerber, Justin Grimmer, Alisa Hall, Tarek Hassan, Seth Hill, Brian Knight, Shiro Kuriwaki, Eddie Lazear, Greg Martin, Jaakko Meriläinen, Daniele Paserman, Vincent Pons, Hunter Rendleman, Dominic Rohner, Jesus Rojas Venzor, Wendy Schiller, Erik Snowberg, Chris Tausanovitch, Danielle Thomsen, Francesco Trebbi, Jessica Trounstine, Jennifer Nicoll Victor, the students of Econ 220 in the Fall 2018-2019 quarter, seminar participants at UBC, the Stanford Economics Brown Bag Lunch, the NBER Spring Political Economy Conference, the members of Stanford's Democracy Policy Lab, the Center for Effective Lawmaking Conference, and the 2019 Congress and History conference at MIT and Harvard.

†Ph.D. Candidate, Department of Political Science. dthomp@stanford.edu.

‡Assistant Professor, Department of Economics. jamesf@bu.edu.

§Corresponding author. Associate Professor, Department of Political Science. andrewbhall@stanford.edu.

¶Ph.D. Candidate, Department of Political Science. yoderj@stanford.edu.

1 Introduction

American politicians are much wealthier and more educated than the general population (e.g., Carnes 2013, 2018; Eggers and Klašnja 2018). To what extent does this reflect a system that blocks people of less privileged backgrounds from seeking political office, and to what extent does it reflect a system that is open to a broad set of candidates but which favors higher-earning and more-educated individuals electorally? Existing data reflects the wealth and educational backgrounds of winning candidates, but because it lacks analogous individual-level data for non-politicians, it does not allow researchers to compare the earnings ability and education of politicians to non-politicians with similar socioeconomic backgrounds. Existing data also rarely details the socioeconomic status of politicians' parents, and it does not allow for the comparison of the backgrounds of losing candidates to winning candidates. These constraints have prevented us from understanding in greater detail the inequality of socioeconomic backgrounds in the political system and the different mechanisms by which the political process might generate a highly unrepresentative legislature.

To overcome these issues, we link future members of Congress to the full-count, deanonymized 1940 U.S. Census, allowing us to observe their education, occupation, earnings, geographical location, and demographic characteristics along with those of roughly 45 million similarly aged Americans.[1] This dataset allows us to go beyond existing work in three main ways.

First, our dataset gives us fine-grained information on the parents of children who go on to become members of Congress, allowing us to offer a uniquely detailed picture of how unrepresentative the backgrounds of members of Congress were in the mid-to-late 20th century. Future members of Congress who were below the age of 18 in 1940 grew up in

[1]We are by no means the first scholars to use historical census records to study American politics. For example, Querubín and Snyder (2013) uses historical census data to evaluate the degree to which 19th-century politicians were able to enrich themselves in office. Hall, Huff, and Kuriwaki (2019) uses the 1860 census to investigate the link between slaveownership and fighting for the Confederacy in the Civil War. But ours is the first paper to our knowledge to use historical census data to analyze the process of political selection in America.

households with average earnings more than twice as large as the population average, and had parents who were more than six times as likely as the general population to hold college degrees.

Second, we take advantage of the scale and detail of our data to evaluate the earnings ability and educational attainment of future politicians compared to socioeconomically similar non-politicians. We track family members across censuses so that we can compare politician and non-politician siblings to one another, thereby holding socioeconomic advantages fixed as in Dal Bó et al. (2017). Using this approach, we find that brothers who entered politics had higher levels of education than their non-politician siblings, on average, had higher earnings, and were substantially more likely to have non-wage income. We also find that future members of Congress earned significantly more in 1940 than non-politicians of the same occupation, education, socioeconomic background, place of birth, and geographical location. These comparisons suggest that members of Congress tend to be highly successful in the private sector before becoming politicians, even compared to non-politicians with the same socioeconomic advantages.

Third, we study the role of elections in this selection process. The pattern of high inequality plus strong selection for individuals with strong private-market skills and high levels of education that we document could indicate that elections favor these types of candidates, or it could reflect a process in which there are high financial barriers to entering politics, so that only high earners can afford to become viable candidates in the first place (Bonica 2017; Carnes 2018). To shed light on this, we use historical newspapers to link candidates who run for office between 1940 and 1950 to the 1940 census (this sample is important because it includes losing candidates in addition to winning candidates.)

Using this sample, we find that the candidate pool as a whole is highly unrepresentative of the population of similar citizens who do not seek political office. We also find that winning candidates are higher earners and have higher levels of education than losing candidates, on average, and this gap is somewhat larger than the gap between the losing candidates

and the population of non-politicians. This suggests to us that there are significant barriers to entry, but also that the electoral process considerably advantages higher-earning and more-educated candidates.

Our work is most closely related to Dal Bó et al. (2017), which studies the process of political selection in Sweden using individual-level registry data. The paper shows that, in Sweden, there are similarly strong patterns of political selection for higher-earning and more-educated individuals (previous work also shows that, in Sweden, earnings ability is correlated with IQ scores and leadership scores from psychological tests (Besley et al. 2017)). However, in Sweden, this process of selection does not lead to a political class that is economically unrepresentative of the population in terms of parental earnings. This key difference between our results and the Swedish results may be the fact that economic mobility is higher in Sweden than in America.[2] In America, political selection for individuals with higher-earning ability and education is likely to lead to a political class with substantially unrepresentative family backgrounds, because there is a strong correlation between parental earnings and child earnings ability. This correlation is weaker in Sweden, allowing for the possibility of political selection for these skills without a cost in terms of representativeness.

Our findings could have important normative implications, but they depend critically on one's views about how earnings ability and education map to political ability, and on one's views regarding the potential tradeoff between favoring higher-ability representatives vs. having an economically representative set of legislators. If earnings ability and education have no bearing on political ability, as Carnes and Lupu (2015) argue, then our findings

[2]Specifically, Bolotnyy and Bratu (2018) finds that the slope of the line relating parental income rank to child income rank is 0.182, while Chetty et al. (2014) finds that this same slope is 0.341 for the United States. Vosters and Nybom (2017) compare the US and Sweden more directly and find significantly more mobility—less persistence—in Sweden across intergenerational measures of income, education, occupation, as well as latent variables built from combinations of these economic status measures, for parent-child links with children born in the 1950s. In the early twentieth century, differences in mobility may not be as stark between the US and Sweden. In Sweden, Lindahl et al. (2012) estimate intergenerational earnings elasticities of 0.356 between the generation born around 1900 and their children born between 1925 to 1930 and an elasticity of 0.303 between those children and their children. Meanwhile, linking parents and children from the Iowa 1915 Census and the Federal 1940 Census, Feigenbaum (2018) found IGE estimates between 0.199 and 0.391, depending on the specification, for that geographically restricted sample of fathers and sons.

suggest that the electoral system's favoring of higher-earning and more-educated candidates is an impediment to having an economically representative legislatures that provides no value. If on the other hand earnings ability and education do correlate with political ability—as an important strand of research in political economy argues (Besley, Montalvo, and Reynal-Querol 2011; Besley and Reynal-Querol 2011; Besley et al. 2017; Meriläinen 2018)—then the relevant question is how much we want our political system to prioritize these abilities at the cost of preventing those with less access to education, and therefore less ability to earn, from becoming politicians. The value of having legislators with strong educational backgrounds and private-market success could be high, if these traits are useful for crafting better policy, but the costs of an unrepresentative legislature are also high, in part because rich legislators may be out of touch with the desires of their constituents (e.g., Gilens 2012), in part because having a broadly representative legislature might add legitimacy to the democratic process (e.g, Mansbridge 1999), and in part because such a legislature might be a normative goal in and of itself. It is not our purpose to offer a judgment on these difficult philosophical questions, but the evidence we offer should help provide empirical grounding for the terms of the debate.

2 Linking Legislators to the 1940 Census

To study the earning ability and socioeconomic backgrounds of legislators, we need to link members of Congress to their records in the 1940 Federal Census. In addition, we find members' siblings and parents by tracing members back to earlier Censuses. Because there are no unique identification numbers that link individuals across Censuses—or to our Congressional biographical data—we rely instead on a fuzzy matching process, making use of both manual matching and algorithmic matching.

2.1 Record Linking Procedures

In this section, we describe the linking procedures we use.

Biographical Information on Members of Congress

To link legislators to the 1940 U.S. Census, we begin by collecting data on the name, birth year, and birth place of every member of the U.S. House and Senate from the Biographical Directory of the U.S. Congress.[3] We match legislators to the 1940 Census because it is the most recent U.S. Census with personally identifiable information.[4] For the main analysis, we link future members of Congress who were between the ages of 18 and 40 in 1940 to their Census records, so we can compare their education and private-sector wages prior to entering Congress to comparable citizens who do not go on to become elected politicians.

Obtaining Potential Matches in 1940 Census

Using the biographical information on members of Congress, we next collect a set of possible matches in the 1940 Census for each legislator. We pull every record from the 1940 U.S. Census where a legislator matches to the Census record exactly on birth state, where the birth year in the Census is within two years of the legislator's birth year in the Congressional biographies, and where both the last name and first name are sufficiently similar using approximate string matching.[5] This outputs a list of Census records that are potential matches for each future legislator.

[3] `http://bioguide.congress.gov/biosearch/biosearch.asp`

[4] Personally identifiable information on the Federal Census is restricted for 72 years after the enumeration of the Census: `https://www.Census.gov/history/www/genealogy/decennial_Census_records/the_72_year_rule_1.html`.

[5] Such blocking on state of birth and range matching on birth year and names is common in the historical census linking literature, for a review see (Abramitzky et al. 2019). For the approximate string matching, we use the Jaro-Winkler distance metric with a penalty parameter $p = 0.1$, implemented by the **stringdist** package in R. Jaro-Winkler string distances are common in record linkage procedures. The distances grow with character edits and transpositions and put more weight on character matching in the beginning of a string. We keep Census records where the last name and first name distances are both less than 0.3.

Manual Disambiguation of Potential Matches

For each future member of Congress we then manually search the list of potential matches and identify the correct Census record.[6] For our comparisons, the main source of bias would be from the risk of false positives in the linking process, where a legislator is mistakenly linked to the Census record of a different individual. The number of future members of Congress is many orders of magnitude smaller than the number of non-members, so falsely linking the Census record of an individual from the general population as a future member of Congress will bias our estimates much more than vice versa. Therefore, our strategy is to minimize the potential for false positives in the linking (in the Appendix, we discuss potential biases associated with false negatives in our linking process). The Census includes a number of other pieces of information about each record that helps us to disambiguate between potential matches. First, the 1940 Census name entries often include middle initial or middle name, which we can cross-reference with the middle name in the legislator's biography. Second, the Census includes an occupation description as of 1940. Third, the Census includes a city code for where the individual resided in 1940, which we can occasionally compare with the member's biographical information.[7] Using this information, along with a manual comparison of name and birth year, we select the Census record that correctly matches to the future member of Congress, if there is one.[8] In Section A.1 of the Appendix, we provide some suggestive evidence that our match improves when we use our manual method.

[6] We discuss various automated record linking methods shortly. While we manually link members of Congress to 1940, we use automated procedures to link the to earlier census waves and to link their parents and siblings across censuses.

[7] Census city codes are listed at `https://usa.ipums.org/usa-action/variables/CITY#codes_section`.

[8] To illustrate why manual matching is necessary, we find Lyndon B. Johnson, who was born in 1908, in the 1940 Census with his first name transcribed as "Lymdan B" and the birth year 1909. His occupation is listed as "Congressman," so we know it is the correct match. Of course, Lyndon B. Johnson does not ultimately enter into our analysis because he was already a member of Congress by 1940.

Finding Parents and Siblings in Earlier Censuses

We also want to collect data on the parents and siblings of the future members of Congress we found as adults in 1940. Because few members are living with these relatives when we observe them in the 1940 Census, this requires us to link backwards to the 1910, 1920, and 1930 Censuses. We start by locating the future members of Congress in these previous Censuses. Because this task multiplies the number of matches we must locate, we are no longer able to do it manually. Therefore, we follow the method described in Feigenbaum (2018), training an algorithm to make links based on a sample of manually linked observations.[9] The procedure is as follows. First, we start with the future member of Congress in 1940. We search for the set of possible matches for that member in the 1910, 1920, and 1930 Censuses.[10] Second, we manually evaluate and identify links for a sample of congress members. Third, we train a record linkage algorithm using the manually constructed data. The algorithm uses various features of a potential match—including string distances in names, year of birth differences, agreement on specific characters, phonetic name differences, and name commonness—to

[9]In addition to the machine learning approach to census linkage, recent work in economics and political science have also used other algorithms, including a fully automated approach based on work by Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2012) and an Expectation-Maximization (EM) algorithm approach (e.g., Abramitzky, Mill, and Pérez 2018; Enamorado, Fifield, and Imai n.d.). Abramitzky et al. (2019) document that empirical analysis based on samples constructed by the three methods—machine learning, Ferrie- or ABE-style, and EM—all tend to produce to similar conclusions: estimates of intergenerational mobility are statistically indistinguishable across the methods in multiple census linking examples. We turn to the machine learning approach for three main reasons. First, our set of matching variables are more limited than in contemporary contexts—like linking voter files as in Enamorado, Fifield, and Imai (n.d.). We observe not date of birth in the Census but reported age (in years) and we do not have information on street address in the legislator's biography. Second, because Census data is spoken, then enumerated in cursive, and then transcribed nearly a century later, we suspect the error rates are much higher in all variables. Third, while both ABE-style methods and the EM algorithm have low rates of false positives in historical data, they do so at the cost of relatively low match rates, failing to recover many links that manual review would code as links, as shown by Abramitzky et al. (2019). While the various state of the art linking methods reviewed in Abramitzky et al. (2019) all sit on the frontier of minimizing false positives and false negatives, the machine learning approach recovers far more of the matches that would be made by manual linking than other methods. In our case, we are not linking a full 100 million record historical census to another census but instead searching for only several thousand specific individuals—MCs and their family members—and higher match rates are essential.

[10]Similar to our Congress to 1940 link, this set of possible matches is anyone born in the same state, within 3 years of the 1940 year of birth, and with Jaro-Winkler string distance in first and last name of less than 0.3. The bounds on year of birth are slightly larger because there is more age and year of birth misreporting in the Census than in the Congressional data.

generate a match score. When a human is record linking, these features are used implicitly to parse matches and non-matches; the algorithm makes these implicit weights explicit and allows us to apply the linkage at scale.[11] Fourth, we identify the matches to each Census—if a sufficiently confident match exists—for each legislator.

Once we have located a future member of congress in a previous Census, we then collect data on his or her parents and siblings. In the historical federal Censuses, households are recorded together. Thus, once we observe a legislator in 1910, we can collect data on his or her parents including name and occupation. Though earnings and education are not collected in the Federal Census until 1940, we make use of occupation score, a variable economic historians and other scholars have relied on to map disparate occupations to a continuous measure of socioeconomic status.[12] Occupation score is based on the median earnings in each occupation in the 1950 Census and constructed by IPUMS. We can also collect information on the legislator's siblings, including first and last names and state and year of birth. We then use this information to link the brothers ahead to the 1940 Census.[13] We use the same machine learning approach to the Census linking described in the previous paragraph, this time searching and training data with reference to the siblings. Linking brothers to the 1940 Census enables us to observe the brothers' earnings, education, and occupation, exactly the data we collected from the 1940 Census about the legislators themselves. We consider potential biases from error in these sibling links in the estimation section below.

[11] Following, Feigenbaum (2018), we train a probit model as the gains from more complex machine learning models are minimal in the historical census linkage context.

[12] Saavedra and Twinam (2017) review recent studies in economic history using occscores as a measure of economic outcomes when income or earnings are unavailable. Prominent examples include Abramitzky, Boustan, and Eriksson (2012) on returns to immigration, Aaronson, Lange, and Mazumder (2014) on Rosenwald schools, Olivetti and Paserman (2015) on intergenerational mobility, and Bleakley and Ferrie (2016) on the Georgia Land Lottery.

[13] We focus on brothers because sisters are likely to change their names if and when they marry, making linking on names impossible and highly biased.

2.2 Dataset of Future Politicians and the General Population

Before proceeding to our analyses, we provide a brief overview of the final linked dataset. The dataset contains information on 557 adults in 1940 who go on to be members of Congress, 359 children in 1940 who go on to be members of Congress, and another roughly 45 million men and women of similar ages from whom we draw our various comparison groups.

Congressional Time Period Covered by the Data

Our dataset is a snapshot of individuals in 1940. Of the future members of Congress in the data, some are middle-aged adults in 1940 and are elected into office in the 1940s; others are young adults who will work in the private sector for decades before seeking office. In thinking through the subsequent analyses, it is useful to understand which congressional terms the results generally reflect. Figure A.2 presents a count of the number of members of Congress in our sample, by the years in which they entered office. The first year anyone in our sample serves in office is 1941, because we restrict to only adults who are not already members of Congress in 1940. The bulk of the adults in 1940 who go on to serve in Congress enter Congress in the 1940s, 1950s, and 1960s, with smaller numbers entering as late as the 1970s and 1980s. Since these are the individuals whose earnings we are able to investigate, the reader can think of the political selection analyses below as speaking primarily to the mid-20th century Congress.

Among the future members of Congress who are children in 1940, the bulk enter Congress in the 1970s and early 1980s. Because we are able to use both youths and adults in 1940 to study the socioeconomic background of members of Congress, the reader can think of the socioeconomic background analyses as covering a wider timeframe, spanning from 1940 through the 1980s, roughly.

Occupations in 1940

To give a flavor of what the data looks like, Table A.1 presents the top five most common occupations in 1940 for future members of Congress vs. the general population.[14] In both cases, the very most common occupation is "Nonclassifiable," a catch-all category. For future members of Congress, the next most common occupation is Lawyer, covering nearly 25% of the future members in our data. The next three for future members of Congress are Unemployed (which may include time spent as a student), Proprietor/Manager, and Clerical. For non-MCs, the remainder of the top 5 after Nonclassifiable is Unemployed, Laborer, Farmer, and Machine Operator.

3 How Unrepresentative are Members of Congress?

We begin by using our data to quantify the unrepresentativeness of members of Congress in the time period of our study. Table 1 studies the families of children in 1940, defined as anyone under the age of 18. Consistent with existing research, we see that future members of Congress come from much higher-earning households. Because individuals earning business income often have no labor earnings to report, we set earnings to missing for individuals who report 0 earnings but who report having non-wage income for this and all subsequent earnings analyses. Average parental labor income is about $2,440 for future members, in the 94th percentile of working men in 1940, but only about $1,204 for the parents of similarly aged children, in the 68th percentile in 1940. As such, parents of future members of Congress earned about twice as much labor income as the average parent of similarly aged children.

Parents of future members also exhibit a large education gap, with 32% of future members living in households where at least one family member had completed college—more than six times the rate at which other similarly aged children lived in such households. Interestingly,

[14] These occupations are coded from the raw census occupation strings by IPUMS and turned into a four-digit occupation category that we use throughout. For more on this process, see: https://usa.ipums.org/usa-action/variables/OCC1950.

Table 1 – Comparing the Socioeconomic Backgrounds of Future Members of Congress to the Population. Compares the earnings, education, and immigration status of the parents of future members of Congress who were children in 1940 to the population of parents of similarly aged children in 1940.

	Individuals Under the Age of 18 in 1940	
Statistic	Future MCs	Population
Family Wage Earnings, Mean	2438.68	1203.55
Family Wage Earnings, 25th Percentile	736.00	360.00
Family Wage Earnings, Median	1773.00	950.00
Family Wage Earnings, 75th Percentile	3000.00	1716.00
Family Member with Earnings > $5000	0.10	0.01
Family Member with Non-Wage Earnings > $50	0.59	0.45
No Family Earnings	0.03	0.04
Family Member with High School Completion	0.70	0.33
Family Member with College Attendance	0.49	0.12
Family Member with College Completion	0.32	0.05
Immigrant Family Member	0.14	0.17
N	359	18,406,686

The first column reports summary statistics for the distribution of annual earnings in 1940 for people who go on to be members of Congress. The second column reports the same statistics for people who never become members of Congress but who were born in the same window.

we find no difference in the probability that future members of Congress live in households with at least one family member who immigrated to the U.S.[15]

Although it is not news that members of Congress are not economically representative of the population, the precise quantification we are able to offer puts this inequality in sharp relief. The remainder of the paper is concerned with understanding the mechanisms by which the American political system generates this extreme unrepresentativeness.

[15]Because of WWI and immigration restriction acts in 1917, 1921, and 1924, the shares of the foriegn-born in the US fell dramatically during the early 20th century (Abramitzky and Boustan 2017).

Table 2 – Comparing the Earnings and Education of Future Members of Congress to the Population.

Statistic	Individuals Aged 18-40 in 1940	
	Future MCs	Population
Wage Earnings, Mean	2009.79	543.81
Wage Earnings, 25th Percentile	274.50	0.00
Wage Earnings, Median	1600.00	260.00
Wage Earnings, 75th Percentile	3000.00	900.00
Wage Earnings > $5,000	0.10	0.00
Non-Wage Earnings > $50	0.51	0.15
No Earnings and Not in School	0.04	0.28
High School Completion	0.84	0.35
College Attendance	0.72	0.12
College Completion	0.52	0.04
Share White	0.99	0.89
Share Man	0.98	0.49
N	557	49,836,824

The first column reports summary statistics for the distribution of annual earnings in 1940 for people who go on to be members of Congress. The second column reports the same statistics for people who never become members of Congress but who were born in the same window.

4 Characterizing Political Selection in Congress

In this section, we use our data to characterize the differences in earnings, education, and family backgrounds between future members of Congress and the general public.

4.1 Simple Comparisons of Future MCs and the General Public

We first compare the self-reported earnings of future members of Congress to the general public. Table 2 shows key summary statistics for future members of Congress and the whole population of similarly aged individuals, focusing on adults who are 18-40 years old in the 1940 Census. We chose the lower threshold of 18 to focus on future MCs who could conceiv-

ably be earning an informative amount of money in the 1940 Census; we chose the upper threshold of 40 because we wanted to observe future MCs *before* they become politicians, to the extent possible. In addition to this age cutoff, we also remove any individuals who are already members of Congress in 1940.

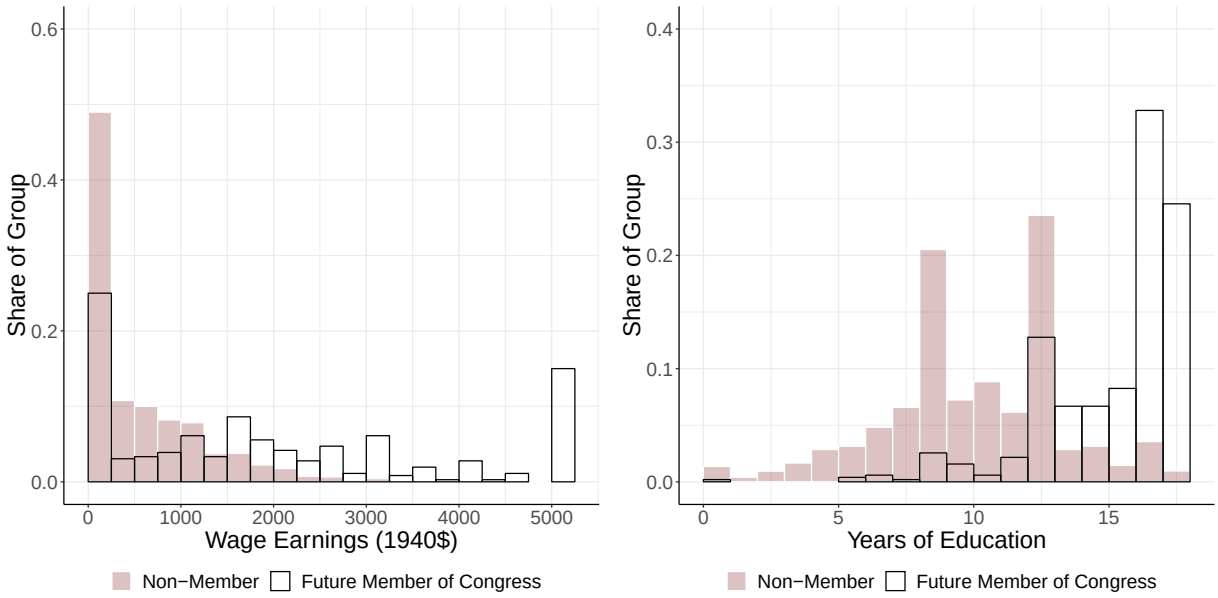
As the table shows, future members of Congress earn far more than the general population, on average. Future members in 1940 earned on average approximately $2,010, in the 90th percentile of labor earnings among working men in 1940, while the average for similarly aged individuals in the population was about $544 (37th percentile in 1940). Future members exhibit a much more right-skewed distribution of earnings, with large differences at the median and the 75th percentile, and with a much higher probability of being top-coded for having earnings of $5,000 or more (only 1% of working men were top coded labor earners in 1940). Future members are also much more likely to have non-wage earnings of $50 or more and less likely to be out of school without any earnings.

Separate from their earnings, future members of Congress were overwhelmingly more likely to complete high school, attend college, and complete college than the population as a whole. Almost three quarters of future members had attended college by 1940, while only 12% of the population of similarly aged individuals had. Finally, and here we are merely stating the obvious for any observer of the history of American politics, future members of Congress in 1940 were almost entirely white men.

4.2 Distribution of Earnings for Future MCs and General Public

Figure 1 presents histograms of the distribution of earnings and education for each of the two groups. As the left panel of the figure shows, future members of Congress tended to out-earn other individuals, with their whole distribution shifted to the right. The large spike at the far right of the plot indicates top-coding, and the histogram shows that members of Congress were far more likely to be top-coded than other individuals. The spike at the far left of the plot indicates individuals who reported no labor income—as a reminder, we set

Figure 1 – Differences in the Earnings and Education of Future Members of Congress and the General Population, 1940.



earnings to missing for individuals who report 0 earnings but who report having non-wage income.

The right panel of the figure likewise shows a pronounced rightward shift in the educational attainment of future members of Congress, as measured by total years of education. A remarkably large proportion of future members of Congress attend or complete college, a level of attainment unusual in the general population in this time period.[16]

These descriptive patterns show that members of Congress substantially out-earn the general public and have much higher average levels of education. What these patterns do not reveal, however, is how much of the overall earnings gap reflects built-in advantages of future members of Congress, vs. how much reflects political selection for those with higher earning ability. Among the relatively privileged elite who have a much greater chance of becoming politicians, do those with the ability to earn more money pursue politics?

[16]Though the return to education varied during the 20th century and may have been relatively lower in 1940 than it had been earlier in the century or in more recent decades (Goldin and Katz 2009), schooling was valued on the labor market in 1940. As Clay, Lingwall, and Stephens Jr (2016) show, exploiting variation in compulsory schooling laws across states and age cohorts, the return to an additional year of education for white men was between 0.064 and 0.079 log points of wage earnings.

4.3 Political Selection Among Male Siblings

In order to hold socioeconomic advantages fixed, we locate the male siblings of future members of Congress who are adults in 1940. Doing so is challenging because, as adults, these future members of Congress are unlikely to be living at home in 1940; their siblings could be anywhere in the census. To solve this issue, we follow future members of Congress backwards in time through the 1910, 1920, and 1930 censuses, until we find them living at home. We then identify their male siblings who also live in these households, and we follow them forwards to the 1940 census so that we can track their earnings alongside those of their politician siblings. Since these siblings grew up in the same households with the same parents, they share precisely the same socioeconomic background. This same strategy is pursued in Dal Bó et al. (2017).

Table 3 presents the results. In the first column, we include no controls; in the latter two columns, we control for age (we cannot match exactly on age since siblings have no overlap in age), either linearly or quadratically. The rows present the estimated differences between future members of Congress and their siblings on a variety of outcome variables.

In the first row, we see that future politicians out-earn their siblings in 1940, whether or not we control for age. In our preferred specification (column 3), future members of Congress are estimated to out-earn their brothers by roughly $280, a substantial gap. Future members of Congress are more likely than their brothers to have top-coded labor income, and to have non-wage income. Finally, they have more years of schooling and are substantially more likely to complete high school (8 percentage points in third column), attend college (11 percentage points in third column), and complete college (12 percentage points in third column), even after controlling for age, suggesting again an important link between education and selection into politics.

It is possible that remaining error in the process that links future members to their siblings overstates these differences. We know that raw differences between future members of Congress and the general population are large; if some of the "siblings" in this analysis

Table 3 – Difference in Earnings and Education for Male Future Members of Congress and Their Brothers.

Variable	Diff Btwn MCs and Brothers		
Wage Earnings	343.90	355.74	279.73
	(110.55)	(108.26)	(93.30)
Earnings > $5,000	0.03	0.03	0.03
	(0.02)	(0.02)	(0.02)
Non-Wage Earnings > $50	0.13	0.13	0.13
	(0.03)	(0.03)	(0.03)
No Earnings and Not in School	0.02	0.02	0.02
	(0.01)	(0.01)	(0.01)
Years of Education	1.44	1.40	1.40
	(0.22)	(0.20)	(0.20)
High School Completion	0.08	0.08	0.08
	(0.03)	(0.03)	(0.03)
College Attendance	0.11	0.11	0.11
	(0.03)	(0.03)	(0.03)
College Completion	0.12	0.12	0.12
	(0.03)	(0.03)	(0.03)
Covs	None	Age	Age, Age^2
Members	371	371	371
N	1871	1871	1871

Each cell reports an estimate of the difference between members of congress and their siblings. Each row presents the differences for a different variable, including earnings from labor (wage earnings), labor earnings at or above the top of the coded earnings distribution (earnings > $5,000), at least $50 in non-wage earnings (non-wage earnings > $50), no reported capital or labor earnings and not in school (no earnings and not in school), years of education, high school completion, college attendance, and college completion. The first column is an unadjusted difference estimated using regressions that include unreported family fixed effect. The second and third columns report differences from the same fixed effects regressions after also adjusting for age differences in the siblings. Robust standard errors reported in the parentheses.

are not true siblings, they may upward bias the resulting comparisons. In the Appendix, we re-estimate the results using a subset for which we are especially confident of the links. Although these results are necessarily less precise, they still suggest important differences. Especially when it comes to non-wage income and years of schooling, future members of Congress still seem importantly different than their non-politician siblings.

4.4 Decomposing the Overall Earnings Advantage of Future MCs

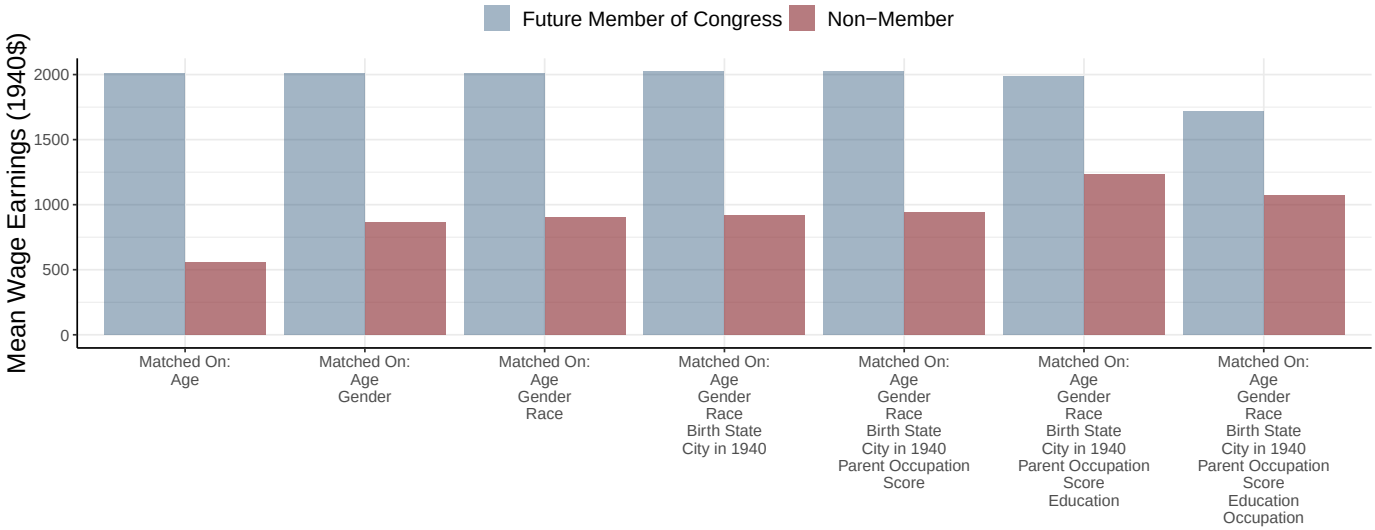
We now compare future members of Congress to increasingly strictly matched sets of individuals who do not become politicians, finding consistent evidence that future members out-earn relevant comparison sets. The purpose here is not to obtain a causal effect, but rather to explore which of the characteristics that might differ between future members of Congress and non-politicians might explain the overall gap in their earnings. This is why we sometimes match on variables that could be the downstream consequences of individual competence and ability, particularly occupation and education.

Figure 2 visualizes these comparisons. Each pair of bars in the plot indicates the mean 1940 earnings of members of Congress and a comparison set of non-members, with labels under each pair of bars indicating the covariates used to construct exact-matched comparisons sets.

In the first pair of bars to the leftmost side of the plot, we match exactly on age. For each member of Congress, we find all individuals in the Census who share the same birth year, compute their average earnings, and pair it with the earnings of the member of Congress before computing the difference in means. As we see, future members of Congress earn substantially more than their matched set of same-aged non-members.

Next, we match exactly on gender in addition to age. Here, we see that the comparison group's mean earnings rises by roughly $300, reducing the overall gap between future members of Congress and non-politicians by roughly 30%. This is largely because members of

Figure 2 – Differences in Earnings Between Future Members of Congress and Various Matched Comparison Groups.



Congress in this time period are overwhelmingly likely to be men, and men are likely to earn more than women in this time period (Goldin et al. 1992).

Next, we further match exactly on race. This leads to only a very modest increase in earnings for the comparison group. Although members of Congress are nearly all white in this time period (see Table 2), the general population is also 90% white in this time period.

Next, we further match exactly on birth state and city of residence in 1940. This barely changes the overall differences, suggesting that differences in place are not an explanation of the overall labor income gap.

Next, we also match on socioeconomic background, which we measure using parental occupation scores. As a reminder, occupation scores are based on the median earnings in each occupation in the 1950 Census and are constructed by IPUMS. Since this measure takes 80 unique values, we find matches whose family occupation scores are in the same 5-unit bin as future members of Congress. Adding socioeconomic background only modestly shrinks the difference between future members of Congress and their peers—perhaps because occupation scores are a coarse measure of socioeconomic background. As a robustness check, we match on the exact value of parent's occupation score and find very similar results with slightly

smaller differences between future members and their peers while losing an additional 20% of members for whom we cannot find exact matches.

The next set of bars suggests that *a major part of the earnings advantage of future members is related to their high levels of education.* When we add years of education to the match set, we see a substantial increase in the comparison group's average earnings. In fact, roughly half of the overall difference between future members of Congress and the general public falls away once we match on education. We explore this pattern in greater detail below. Despite this advantage, though, future members still earn substantially more than other individuals of the same age, race, gender, geographical location, and education level.

Finally, we additionally match exactly on occupation. This actually decreases the comparison group's average earnings—evidently, conditional on education, future members of Congress tend to opt into somewhat lower-paying occupations. However, since the set of future members for whom we can find exact matches also has somewhat lower earnings, the estimated gap in earnings does not change much when we add occupation.

Overall Results: Future Members are Higher Earners

In sum, consistent with the siblings analysis from before, it appears that a significant part of the earnings gap between future members of Congress and non-politicians reflects differences in earnings ability. Comparing the rightmost pair of bars with the full set of matched covariates to the leftmost pair, we see that the included covariates together appear to reduce the overall earnings gap between future members of Congress and non-politicians by a little more than half, leaving a substantial earnings gap even between future members and individuals with the same education level, occupation, and socioeconomic background.

Table 4 – Differences in 1940 Earnings for Future Members of Congress and Comparison Groups.

	Difference in Annual Wage Earnings, 1940, MCs vs. Non-MCs						
Statistic	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	1450.83	1145.34	1109.78	1104.54	1078.80	751.28	642.19
	(76.25)	(72.03)	(71.49)	(44.40)	(50.32)	(71.11)	(113.58)
25th Percentile	274.50	34.50	24.50	45.00	24.50	-148.00	0.00
Median	1320.00	900.00	840.00	850.00	800.00	300.00	682.50
75th Percentile	2094.00	1714.00	1700.00	1664.00	1630.00	1100.00	800.00
Wage Earnings > $5,000	0.10	0.10	0.10	0.10	0.10	0.09	0.09
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Non-Wage Earnings > $50	0.36	0.30	0.30	0.30	0.28	0.22	0.14
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.03)
No Earnings and Not in School	-0.24	-0.04	-0.04	-0.03	-0.03	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Members	557	557	557	539	474	289	142
Matches	49,836,824	36,257,432	33,080,748	631,791	109,364	9,351	2,839
Covariates Matched On							
Age	X	X	X	X	X	X	X
Gender		X	X	X	X	X	X
Race			X	X	X	X	X
Birth State				X	X	X	X
City in 1940				X	X	X	X
Parent Occupation Score					X	X	X
Education						X	X
Occupation							X

Each cell reports the difference between members and their matches on a particular summary statistic of the distribution of income in 1940. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The standard errors come from regressions that include a fixed effect for ever unique set of values of the covariates. The reported statistics for matched groups are calculated by matching future members of Congress with people who never become members of Congress but who have the exact same values on the matching covariates indicated at the bottom of the table. The non-members are reweighted so that the joint distribution of all covariates is identical for the member and non-member groups.

Formal Estimates

Table 4 presents these graphical comparisons in a more formal manner, allowing the reader to see the precise estimated differences and their standard errors, as well as information about sample sizes and more formal details about the matching approach.

4.4.1 Matching on Neighborhood

Parent's occupation is a coarse measure of the socioeconomic environment in which someone grew up. Unfortunately, the 1910, 1920, and 1930 Censuses where we collect family background information did not measure earnings or education, preventing us from matching on more detailed information about parental background. To further consider socioeconomic background, we construct a measure of average occupation score for adult men with occupations in the childhood neighborhood of each future member of Congress and their potential matches.

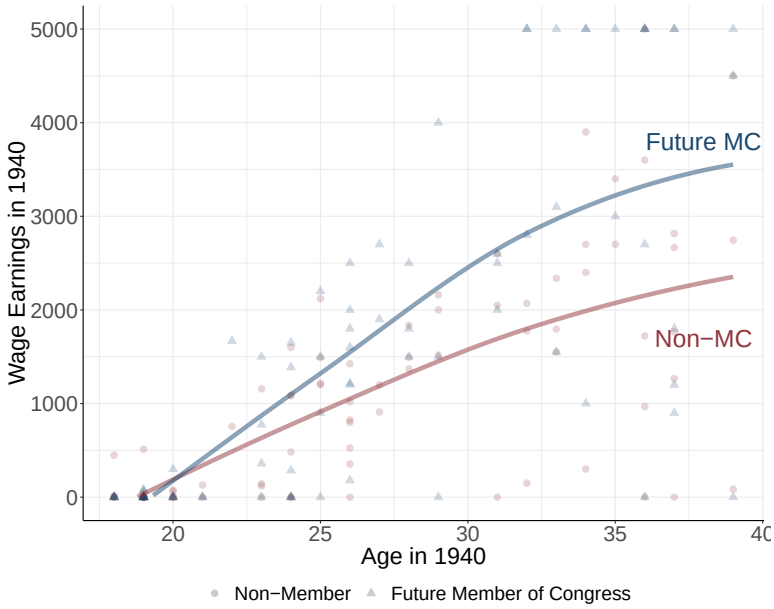
Matching on mean neighborhood occupation score for men does not meaningfully change any of the results. Compared to column 7 in Table 4, the earnings differential *increases* by \$24, and the difference in the percent earning over \$5,000 in wages declined by less than one percentage point. We take this as evidence that matching on a larger and larger set of childhood characteristics would not meaningfully change our main findings. Since one third of the 118 future members of congress with a match in column 7 do not also have a match on mean neighborhood occupation score, we have chosen to leave the formal estimates out of the table.

4.5 The Earnings Gap Across Age in 1940

The results above include all adults from age 18 to 40 in 1940, but we might think that it is harder to observe earnings ability among individuals who are still teenagers. These individuals may still be in school—part of why we find that members of Congress have a larger spike at 0 labor income—or they may not have had an opportunity yet to advance in their careers to a point where their talent manifests itself.

Figure 3 focuses on future members of Congress and their matched comparisons as in column 7 of Table 4, where matches are based on the full set of covariates, and it presents earnings across respondent age in 1940. As the plot shows, we see little difference in earnings for those still in their teens or early 20s. The earnings gap grows substantially for individuals

Figure 3 – Earnings Advantage for Future Members of Congress and Matched Controls Across Age in 1940.



in their 30s, which is also where we see the appearance of top-coded observations. Based on this plot, in the Appendix we pursue a set of analyses where we subset to future members who are 25 years or older in 1940.

4.6 Political Selection Among Lawyers

One problem with the analyses thus far is that occupation is somewhat coarsely measured. Roughly one quarter of future members of Congress in our sample have an occupation that can't be classified in the 1940 Census, so matching on occupation is not as restrictive a match as we might like. To see how this might affect the results, Table 5 replicates the analysis restricting only to lawyers—constituting roughly 25% of all future MCs in our sample—in the 1940 Census. Now all comparisons are between lawyers who go on to become members of Congress, and other lawyers who do not. As the table shows, we continue to find large

Table 5 – Differences in 1940 Wage Earnings for Future Members of Congress and Comparison Groups, Including Lawyers Only.

	Diff in Wage Earnings, 1940, MCs vs. Non-MCs				
Statistic	(1)	(2)	(3)	(4)	(5)
Mean	573.08	556.10	566.60	816.54	729.18
	(182.58)	(182.52)	(182.32)	(199.14)	(277.72)
25th Percentile	100.00	100.00	144.00	59.50	722.00
Median	590.00	590.00	590.00	1000.00	1000.00
75th Percentile	1850.00	1800.00	1800.00	2058.00	1400.00
Wage Earnings > $5,000	0.09	0.09	0.09	0.13	0.14
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Members	131	131	131	79	42
Matches	71,329	69,290	68,900	2,615	542
Covariates Matched On					
Age	X	X	X	X	X
Gender		X	X	X	X
Race			X	X	X
Birth State				X	X
City in 1940				X	X
Parent Occupation Score					X

Each cell reports the difference between members and their matches on a particular summary statistic of the distribution of wage earnings in 1940. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The standard errors come from regressions that include a fixed effect for ever unique set of values of the covariates. The reported statistics for matched groups are calculated by matching future members of Congress with people who never become members of Congress but who have the exact same values on the matching covariates indicated at the bottom of the table. The non-members are reweighted so that the joint distribution of all covariates is identical for the member and non-member groups.

differences in earnings.[17] Again, this seems to suggest that individuals with higher earning capacity selected into political careers.

4.7 Political Selection in Terms of Education

We now turn to studying differences in education between politicians and non-politicians in greater detail. Table 6 shows the estimated differences in education levels between future members of Congress and various comparison groups in the general population, similar to previous tables.

[17] We omit the specification that additionally exact matches on education because we are left with only 8 future MCs in this specification. In this specification, the difference between future MCs and non-MCs is larger ($1,227.66; se=425.8), but the small sample size is tiny.

Table 6 – Differences in 1940 Education for Future Members of Congress and Comparison Groups.

	Average Differences, MCs vs. Non-MCs					
Statistic	(1)	(2)	(3)	(4)	(5)	(6)
Years of Education	5.24	5.37	5.08	4.73	4.33	2.86
	(0.11)	(0.10)	(0.10)	(0.07)	(0.09)	(0.14)
High School Completion	0.57	0.60	0.58	0.54	0.49	0.36
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
College Attendance	0.67	0.67	0.66	0.63	0.60	0.40
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
College Completion	0.53	0.52	0.52	0.50	0.48	0.28
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
Members	509	509	509	492	430	251
Matches	48,609,040	35,330,240	32,247,202	557,737	93,469	12,521
Covariates Matched On						
Age	X	X	X	X	X	X
Gender		X	X	X	X	X
Race			X	X	X	X
Birth State				X	X	X
City in 1940				X	X	X
Parent Occupation Score					X	X
Occupation						X

Each cell reports the difference between members and their matches on a particular summary statistic of the distribution of educational attainment in 1940. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The standard errors come from regressions that include a fixed effect for ever unique set of values of the covariates. The reported statistics for matched groups are calculated by matching future members of Congress with people who never become members of Congress but who have the exact same values on the matching covariates indicated at the bottom of the table. The non-members are reweighted so that the joint distribution of all covariates is identical for the member and non-member groups.

In column 1, we present raw comparisons to the whole population of similarly aged individuals. The first row shows the difference in total years of education between future members and the population, finding a large difference. The second, third, and fourth rows re-estimate these differences in terms of the probability of completing high school, attending college, and completing college, respectively, again finding extremely large differences.

As we go across the columns, we continue to find large differences. In column 5, we match on age, gender, race, state of birth, city of residence in 1940, and parent occupation score, and we still find massive differences in education levels.

Finally, in column 6, we match on occupation (a "post-treatment" variable from the perspective of education, if we were focused on a question of causal inference.) Here, the

differences do shrink, but they are still extremely large. Compared to individuals of the same age, gender, race, birth state, city of residence in 1940, parent occupation score, and occupation, future members of Congress have almost 3 more years of schooling, on average, are 36 percentage points more likely to complete high school, and are 28 percentage points more likely to finish college.

In sum, future members of Congress possess much higher levels of education, compared both to the general population and to matched individuals with the same demographic characteristics and family backgrounds.

5 The Candidate Pool and Electoral Selection

The selection of high-earning and highly educated individuals for Congress could reflect an electoral system that favors these types of candidates, or it could reflect a system that discourages other types of candidates from running, preventing voters from supporting them. To understand these mechanisms, we need to observe both who runs for office and, from among the candidate pool as a whole, who wins office. This is challenging because finding losing candidates in the 1940 census is more difficult than finding winning candidates. Most names are not unique in the census; previously, when we linked winning candidates to the census, we took advantage of biographical information provided by Congress to disambiguate between potential name matches. Finding this biographical information for losing candidates is much harder, because no centralized repository exists.

We solve this problem by developing a manual workflow in which we find potential name matches of people who lived in the same state in 1940 in which the candidate ran. We focus on races close in time to 1940—specifically, 1942 to 1950—in order to minimize the likelihood of individuals having moved states between the 1940 census and when they ran for office.

Our workflow involves searching the names of potential matches in newspapers.com, focusing on the year and state of each election in our dataset. We use newspaper articles

from the time to obtain useful information about candidates, including in various cases their city of residence, their age, their occupation, and so forth. We use whatever information we are able to obtain to then locate their record in the 1940 census on ancestry.com, which in turn allows us to link their record to our dataset.

Using this approach, we are able to locate a total of 549 candidates, of which 398 lose the election in which they ran (we are only able to use all 549 candidates for the education analysis, as quite a few have earnings set to missing because they report no earnings but have business income; we have 306 candidates in the earnings analyses, for which 82 are winners). We then proceed with the same exact matching approach employed throughout the paper, running regressions in which we compare earnings and years of education for losing candidates, winning candidates, and matched sets of non-politicians. Table 7 presents the results.

As the first row of the table shows, the set of people who run for office are substantially higher earners than those who do not run. This is true when we match only on age, gender, and race (first column), when we also match on birth state and city in 1940 (second column), and when we also control for education and occupation (third column). Looking at the third column, we estimate that the candidate pool earns roughly $340 more than matched non-politicians, roughly 29% higher earnings than the mean for matched non-politicians.

The second row shows that there is positive electoral selection from among the pool. Winning candidates are estimated to earn roughly $570 more than losing candidates (third column), a substantial increase.

The final two columns show that there are large gaps in terms of education, too. The candidate pool possesses 4.38 more years of education than matched non-politicians (fifth column), a 48% increase from the overall mean for non-politicians, and among those who run, winners have on average an additional 1.31 years of education, roughly (second row, fifth column).

Table 7 – Differences in 1940 Earnings and Education for Future Congressional Candidates and Comparison Groups.

	Wage Earnings			Years of Educ	
	(1)	(2)	(3)	(4)	(5)
Ran	960.00	864.32	339.70	4.94	4.38
	(87.58)	(57.69)	(55.51)	(0.15)	(0.09)
Ran × Won	730.34	785.03	570.92	0.95	1.31
	(184.06)	(111.61)	(109.36)	(0.25)	(0.18)
Mean for Matches	1106.12	1177.27	1177.27	8.73	9.20
Winners	82	82	82	146	146
Losers	224	224	224	403	403
Matches	37,133,080	214,844	214,844	37,133,080	214,844
Covariates Matched On					
Age	X	X	X	X	X
Gender	X	X	X	X	X
Race	X	X	X	X	X
Birth State		X	X		X
City in 1940		X	X		X
Fixed Effects					
Covariate Cell	X	X	X	X	X
Education			X		
Occupation			X		

Each cell reports coefficients from regressions of income and education on indicators for whether the individual ran (Ran) for congress and whether they won if so (Ran × Won). Members of the public who did not run only enter the regression if the share the exact same profile as someone who ran based on a set of covariates. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The members of the public are reweighted so that the joint distribution of all covariates is identical for the candidate and non-candidate groups.

Together, these results suggest an electoral advantage for higher-earning and more-educated candidates. At the same time, both losing and winning candidates tend to be substantially higher-earning and more educated than other individuals of the same age, gender, and race who were born in the same state and live in the same city in 1940. Because individuals may run for office in anticipation of their electoral prospects, it is hard for the data to distinguish whether the candidate pool is unrepresentative because of barriers to entry, anticipated electoral defeat for lower-earning and less educated individuals, or both. Whatever the political equilibrium, the electoral advantage for higher-earning and more-educated candidates appears to play an important role in making Congress unrepresentative of the general population in this time period.

6 Conclusion

It is a well-known fact that American legislators are far wealthier and more educated than the general population. Exactly why this is the case remains a matter of debate. American legislators could be richer and more educated because the costs of running for office prevent normal people from seeking political office, or because elections favor individuals with particularly high levels of education and earnings ability, or because of some combination of the two. This paper has contributed to this topic by studying the earnings ability, education, and socioeconomic backgrounds of future members of Congress as of 1940, taking advantage of a unique historical dataset allowing us to make fine-grained comparisons between future politicians and over 40 million comparably aged citizens.

Our data has allowed us to observe the parents of future members of Congress in much greater detail than existing work. We find that parents of future members of Congress earn much more than the general public, and are much, much more likely to have attended college. This is a key difference from previous findings in Sweden (Dal Bó et al. 2017). Because intergenerational mobility is much lower in the United States than in Sweden, even a pure political meritocracy on the basis of earnings ability and education in America would translate into high inequality in political opportunities.

Turning to earnings ability, we first showed that future members of Congress earn more and are more-educated than their siblings who do not go into politics, suggesting that there is political selection for these traits over and above socioeconomic background. We also found that future members of Congress earned substantially more money in 1940 than did individuals in the same occupation, with the same level of education, living in the same area and sharing the same race and gender, and coming from comparable socioeconomic backgrounds. Moreover, future members of Congress are substantially more educated than comparable individuals who do not go into politics. Despite reasons to think that democratic systems might struggle to attract highly skilled and highly educated individuals who have

lucrative outside options in the private market, the U.S. system is evidently able to make such individuals a compelling offer.

Finally, after characterizing these patterns, we have offered some preliminary evidence to explain their roots in the electoral system. The set of candidates who run for office are much higher-earning and more educated than comparable non-politicians, suggesting barriers to running for office could be an important driver of the differences between members of Congress and the general public. However, we also found evidence that elections favor candidates with higher earnings ability and higher levels of education. The electoral advantage for these types is likely to be an important part of the inequality we observe.

Our study has two main limitations. First, the ability to earn more than comparable people measures only one component of personal ability, and could be uncorrelated or even negatively correlated with the attributes that make individuals "good" representatives. While the positive results of our study are important for understanding how the American political system, and democracies in general, operate, on their own they have no clear normative implications for American government or democracy.

Second, because the U.S. Census is only made available to researchers after a 72-year gap, we can only study people who were alive—and, in many of our analyses, already adults—in 1940. This means that our analyses primarily speak to political selection in the United States from roughly 1950 to 1970. While an important period in its own right, many factors have changed about American politics from then to today. In particular, it appears that running for office today, in an era of severe campaign-finance demands and 24-hour tabloid media coverage is more difficult than it used to be. Moreover, salaries for members of Congress have not kept up with inflation, meaning that holding a seat in Congress may have higher opportunity cost than it used to (Hall 2019). This may deter high-earners more than it did in the past.

These limitations aside, our paper adds to the relatively new literature on political selection (for a recent review, see Dal Bó and Finan n.d.) by characterizing how strongly

the American system favors the selection of highly educated individuals with strong private-market skills over the selection of a broadly representative set of legislators who share the economic backgrounds of their constituents. We hope that documenting these patterns will encourage additional research on the mechanisms of political selection in the United States.

References

Aaronson, Daniel, Fabian Lange, and Bhashkar Mazumder. 2014. "Fertility Transitions Along the Extensive and Intensive Margins." *American Economic Review* 104(11): 3701–24.

Abramitzky, Ran, and Leah Boustan. 2017. "Immigration in American Economic History." *Journal of Economic Literature* 55(4): 1311–45.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102(5): 1832–56.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James J Feigenbaum, and Santiago Pérez. 2019. Automated Linking of Historical Data. Technical report National Bureau of Economic Research.

Abramitzky, Ran, Roy Mill, and Santiago Pérez. 2018. "Linking Individuals Across Historical Sources: a Fully Automated Approach." NBER Working Paper 24324. `http://www.nber.org/papers/w24324`.

Besley, Timothy. 2004. "Paying Politicians: Theory and Evidence." *Journal of the European Economic Association* 2: 193–215.

Besley, Timothy, and Marta Reynal-Querol. 2011. "Do Democracies Select More Educated Leaders?" *American Political Science Review* 105(3): 552–566.

Besley, Timothy, and Stephen Coate. 1997. "An Economic Model of Representative Democracy." *The Quarterly Journal of Economics* 112(1): 85–114.

Besley, Timothy, Jose G Montalvo, and Marta Reynal-Querol. 2011. "Do Educated Leaders Matter?" *The Economic Journal* 121(554): F205–227.

Besley, Timothy, Olle Folke, Torsten Persson, and Johanna Rickne. 2017. "Gender Quotas and the Crisis of the Mediocre Man: Theory and Evidence from Sweden." *American Economic Review* 107(8): 2204–42.

Bleakley, Hoyt, and Joseph Ferrie. 2016. "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations." *The Quarterly Journal of Economics* 131(3): 1455–1495.

Bolotnyy, Valentin, and Cristina Bratu. 2018. "The Intergenerational Mobility of Immigrants and the Native-Born: Evidence from Sweden." Working Paper. `https://scholar.harvard.edu/bolotnyy/publications/intergenerational-mobility-immigrants-and-native-born-evidence-sweden`.

Bonica, Adam. 2017. "Professional Networks, Early Fundraising, and Electoral Success." *Election Law Journal* 16(1): 153–171.

Carnes, Nicholas. 2013. *White-collar Government: The Hidden Role of Class in Economic Policy Making*. University of Chicago Press.

Carnes, Nicholas. 2018. *The Cash Ceiling: Why Only the Rich Run for Office–and What We Can Do about It*. Princeton University Press.

Carnes, Nicholas, and Noam Lupu. 2015. "What Good is a College Degree? Education and Leader Quality Reconsidered." *Journal of Politics* 78(1): 35–49.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *The Quarterly Journal of Economics* 129(4): 1553–1623.

Clay, Karen, Jeff Lingwall, and Melvin Stephens Jr. 2016. Laws, educational outcomes, and returns to schooling: Evidence from the Full Count 1940 Census. Technical report National Bureau of Economic Research.

Dal Bó, Ernesto, and Frederico Finan. n.d. "Progress and Perspectives in the Study of Political Selection." *Annual Review of Economics*. Forthcoming.

Dal Bó, Ernesto, Frederico Finan, Olle Folke, Torsten Persson, and Johanna Rickne. 2017. "Who Becomes A Politician?" *The Quarterly Journal of Economics* 132(4): 1877–1914.

Eggers, Andrew C., and Marko Klašnja. 2018. "Wealth, Fundraising, and Voting in the U.S. Congress." Working Paper.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. n.d. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review*. Forthcoming.

Feigenbaum, James J. 2018. "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940." *The Economic Journal* 128(612): F446–F481.

Ferrie, Joseph P. 1996. "A new sample of males linked from the public use microdata sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules." *Historical Methods* 29(4): 141–156.

Gagliarducci, Stefano, and Tommaso Nannicini. 2013. "Do Better Paid Politicians Perform Better? Disentangling Incentives from Selection." *Journal of the European Economic Association* 11(2): 369–398.

Gilens, Martin. 2012. *Affluence and Influence: Economic Inequality and Political Power in America*. Princeton University Press.

Goldin, Claudia, and Lawrence F Katz. 2009. *The Race between Education and Technology*. Harvard University Press.

Goldin, Claudia et al. 1992. "Understanding the gender gap: An economic history of American women." *OUP Catalogue* .

Hall, Andrew B. 2019. *Who Wants to Run? How the Devaluing of Political Office Drives Polarization*. University of Chicago Press.

Hall, Andrew B., Connor Huff, and Shiro Kuriwaki. 2019. "Wealth, Slave Ownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War." *American Political Science Review* .

Key, V.O. 1949. *Southern Politics in State and Nation*. Alfred A. Knopf.

Lindahl, Mikael, Mårten Palme, Sofia Sandgren Massih, and Anna Sjögren. 2012. "The Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations." Working Paper. `http://ftp.iza.org/dp6463.pdf`.

Mansbridge, Jane. 1999. "Should Blacks Represent Blacks and Women Represent Women? A Contingent "yes"." *The Journal of Politics* 61(3): 628–657.

Meriläinen, Jaakko. 2018. "Politician Quality, Ideology, and Fiscal Policy." Working Paper.

Olivetti, Claudia, and M Daniele Paserman. 2015. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." *American Economic Review* 105(8): 2695–2724.

Osborne, Martin J., and Al Slivinski. 1996. "A Model of Political Competition with Citizen-Candidates." *The Quarterly Journal of Economics* 111(1): 65–96.

Palmer, Maxwell, and Benjamin Schneer. 2016. "Capitol Gains: The Returns to Elected Office from Corporate Board Directorships." *Journal of Politics* 78(1): 181–196.

Palmer, Maxwell, and Benjamin Schneer. 2018. "Post-Political Careers: How Politicians Capitalize on Public Office." Working Paper.

Querubín, Pablo, and James M. Snyder, Jr. 2013. "The Control of Politicians in Normal Times and Times of Crisis: Wealth Accumulation by U.S. Congressmen, 1850-1880." *Quarterly Journal of Political Science* 8: 409–450.

Saavedra, Martin, and Tate Twinam. 2017. "A Machine Learning Approach to Improving Occupational Income Scores." *arXiv preprint arXiv:1704.08299* .

Vosters, Kelly, and Martin Nybom. 2017. "Intergenerational Persistence in Latent Socioeconomic Status: Economicsvidence from Sweden and the United States." *Journal of Labor Economics* 35(3): 869–901.

Online Appendix

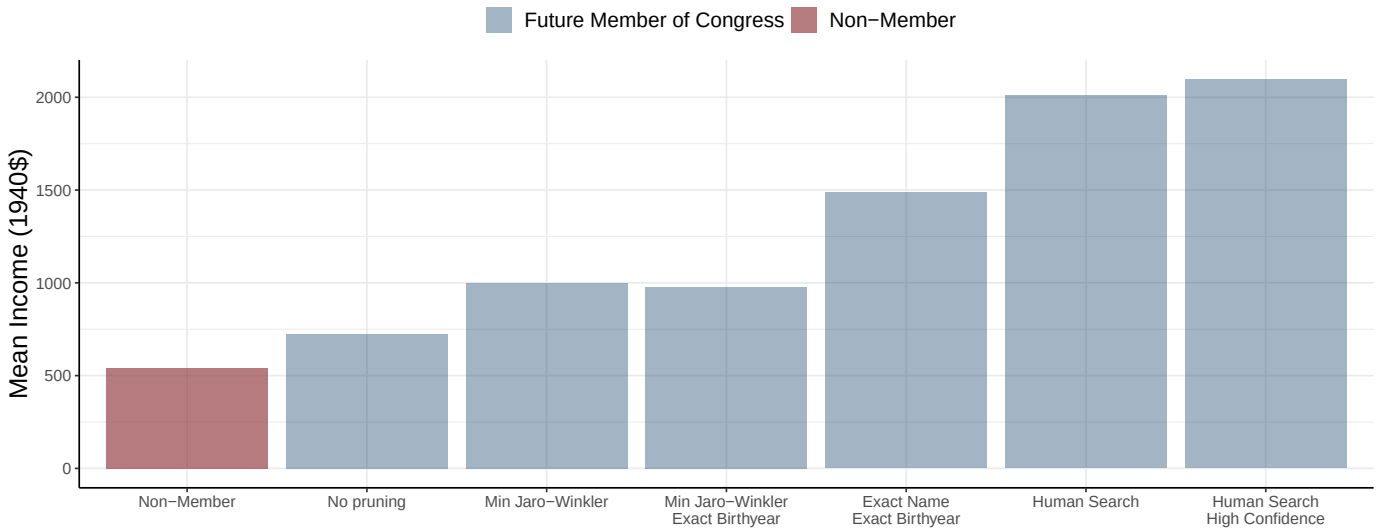
Intended for online publication only.

Contents

A.1 Comparing Techniques for Record Linkage

Here, we compare a variety of linking strategies to validate our hand matching procedure. The red bar in Figure A.1 shows the mean earnings in 1940 dollars among the general population, while each of the blue bars shows the mean earnings in 1940 dollars among members of Congress using increasingly constrained linking criteria. We find that the difference in earnings between the members of Congress and the general population grows as we become more confident in the match accuracy. For example, the "No pruning" bar computes the mean earnings of all Census records we identify in our first pass as potential matches for members of Congress. Specifically, this includes all Census records with the same birth state, birth year within two years of the legislator's, and both last and firstname Jaro-Winkler string distance less than 0.3.[18] Using that linking procedure, most legislators have multiple potential matches in the 1940 Census, so we know there are false positives when computing the mean earnings for the "No pruning" bar. As expected, this attenuates the estimated difference between mean earnings for MCs compared to the general population.

Figure A.1 – Difference from Population Grows with Match Accuracy.



Next, we try constraining the linking criteria in other ways. In the "Min Jaro-Winkler" bar, for each legislator we select the potential Census match that minimizes our name distance metric. We take the Jaro-Winkler distance between the legislator's last name and Census record's last name along with the Jaro-Winkler distance between the legislator's first name and the Census record's first name, and we sum the absolute value of those two quantities to construct our distance metric. The difference in mean earnings between the members of Congress and general population grows when we use this more constrained linking procedure. We try a few other linking strategies, matching on exact name and exact birth year,

[18] We use the **stringdist** package in R to compute the Jaro-Winkler distance, setting the penalty parameter $p = 0.1$.

for example, but for all results in the body of the paper we hand match members to the list of potential matches in the Census. In the paper, we describe the criteria we use to select hand matches. As we can see in Figure A.1, the earnings difference grows even more using the hand matching procedure, which suggests that we are minimizing the presence of false positives in the record linkage. In the final bar, we show the mean earnings for the hand matches where we have high confidence in the match, but to keep as many observations as possible we use all of the hand matched legislators in the body of the paper.

A.2 Years Served in Congress in Our Sample

Our analyses rely on observing future members of Congress in the 1940 Census. In our main analyses, we focus on the earnings of future members of Congress who are adults in 1940, restricting the sample to future members who are between the ages of 18 and 40 as of 1940. Later, we also study those who are under the age of 18 in 1940, in order to understand socioeconomic background. Given the range of ages in 1940, what years do these individuals go on to serve in Congress? Figure A.2 shows the distribution of years that our sample serves in Congress, plotted separately for youths and adults in 1940.

Figure A.2 – Years Served in Congress in Our Sample.

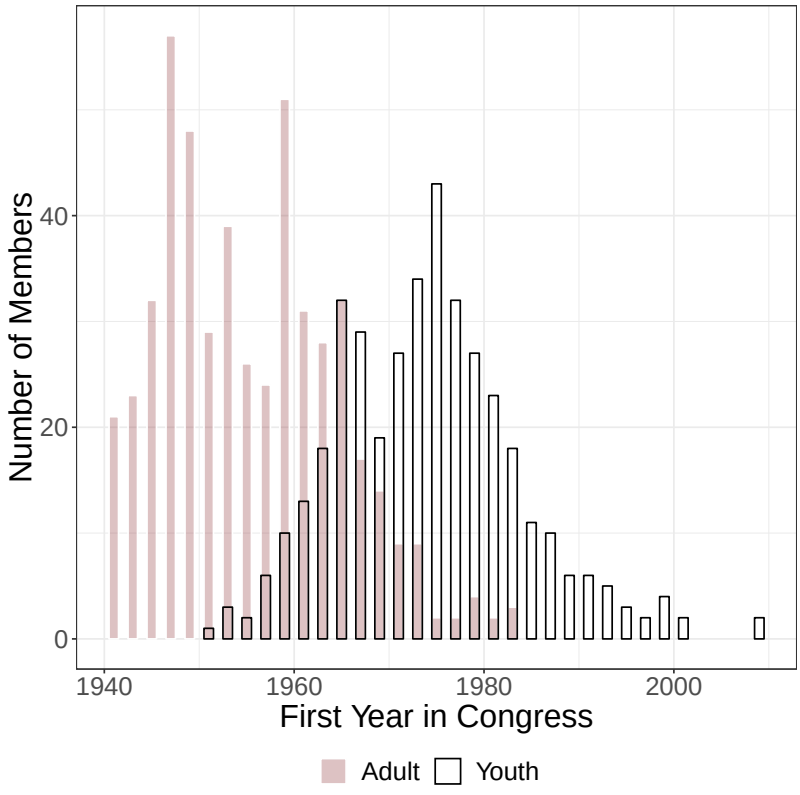


Table A.1 – Top Occupations for Future Members of Congress and Non-Members.

Rank	Future MCs		Non-MCs	
	Occupation	Share of Pop	Occupation	Share of Pop
1	Nonclassifiable	0.26	Nonclassifiable	0.20
2	Lawyer	0.24	Unemployed	0.10
3	Unemployed	0.17	Laborer	0.10
4	Proprietor, Manager	0.06	Farmer	0.06
5	Clerical	0.03	Machine Operator	0.06

A.3 Most Common Occupations in 1940

Table A.1 reports the top five occupations for future members of Congress and the population. 26% of future members of Congress and 20% of the population are working in a role that did not fit into the classifications offered by the Census or were looking for their first job. The second most common group of occupations for future members of Congress was in the law, with 24% of future members working as lawyers or judges. 17% of members of Congress were unemployed. The remaining top occupations were both professional roles. Among the population, 10% of people were unemployed. The remaining top occupations for non-members were all blue collar: laborer, farmer, and machine operator.

A.4 Comparing Members of the House vs. the Senate

As we saw in Table A.6, the earnings differential is much larger among future senators than representatives. In table A.2, we track the size of the gap as we match on more and more covariates. The earnings differential between future senators and members of the public their same age in 1940 is roughly $1,550, while that same difference is $925 for future representatives. This leaves a gap between senators and representatives of about $625. By column five, when we are matching on age, gender, race, birth place, city, and education level, the difference between senators and representatives shrinks to $180, but is still present. The remaining columns are difficult to interpret given the relatively small set of future senators who have a match on all covariates in the Census.

A.5 Replicating the Results For 25-40 Year Olds

Table A.3 reports the earnings differentials for future members of Congress who are between 25 and 40 in 1940. Looking at Figure 3, it is clear that this should increase the differences, since the future members earn a similar amount as their peers at 18. The differences between future members and their peers start to appear in their mid 20s. We confirm that the differences are larger when future members are older in Table A.3. Across all seven columns, the differences are larger for future members over 25 than for future members over 18. The

earnings differential for future members is approximately two-thirds larger when we exclude people under 25, based on the estimates using the full set of covariates in column seven.

A.6 Socioeconomic Advantages Across Generations

In Table 1 we saw that future members of Congress who were between the ages of two and seventeen in 1940 came from families with higher earnings and more education. Using the 1910, 1920, and 1930 Censuses, we can compare the families of future members who were between ages two and seventeen in 1940 to those who were between eighteen and forty in 1940. Table A.4 presents these results. We find that youth in 1940 come from families working in higher-earning occupations than the future members who are adults in 1940. This result must be taken with a grain of salt: getting the occupation scores for the parents of adults requires automated record linkage that may introduce some error, whereas the occupation scores the parents of a young person living at home are easily available.

Table A.2 – Differences in 1940 Wage Earnings for Future Members of Congress and Comparison Groups.

	Senators						
Statistic	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	2360.24	2004.32	1953.42	2040.75	1968.46	1443.29	1622.44
	(221.81)	(216.22)	(215.50)	(159.69)	(209.63)	(382.18)	(928.08)
Wage Earnings > $5,000	0.21	0.21	0.21	0.21	0.23	0.18	0.20
	(0.04)	(0.04)	(0.04)	(0.02)	(0.03)	(0.06)	(0.10)
Non-Wage Earnings > $50	0.42	0.35	0.35	0.34	0.30	0.23	0.32
	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.07)	(0.11)
No Earnings and Not in School	-0.26	-0.04	-0.04	-0.03	-0.01	0.02	0.03
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.08)
Members	80	80	80	75	60	33	15
Matches	45,306,392	22,337,936	20,093,224	31,930	2,700	139	21
	Representatives						
Statistic	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	1410.33	1107.50	1073.02	1066.25	1063.81	756.10	668.28
	(79.61)	(75.23)	(74.67)	(45.69)	(51.81)	(73.06)	(119.98)
Wage Earnings > $5,000	0.09	0.09	0.09	0.09	0.09	0.09	0.08
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Non-Wage Earnings > $50	0.36	0.30	0.29	0.29	0.28	0.22	0.12
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.03)
No Earnings and Not in School	-0.24	-0.04	-0.04	-0.03	-0.03	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Members	509	509	509	495	438	272	133
Matches	49,836,824	36,257,432	33,080,748	609,915	107,343	9,256	2,824
Covariates Matched On							
Age	X	X	X	X	X	X	X
Gender		X	X	X	X	X	X
Race			X	X	X	X	X
Birth State				X	X	X	X
City in 1940				X	X	X	X
Parent Occupation Score					X	X	X
Education						X	X
Occupation							X

Each cell reports the difference between members and their matches on a particular summary statistic of the distribution of earnings in 1940. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The standard errors come from regressions that include a fixed effect for ever unique set of values of the covariates. The reported statistics for matched groups are calculated by matching future members of Congress with people who never become members of Congress but who have the exact same values on the matching covariates indicated at the bottom of the table. The non-members are reweighted so that the joint distribution of all covariates is identical for the member and non-member groups.

Table A.3 – Differences in 1940 Wage Earnings for Future Members of Congress and Comparison Groups, Aged 25 to 40 in 1940.

	Difference in Annual Wage Earnings, 1940, MCs vs. Non-MCs						
Statistic	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	2116.17	1718.22	1674.29	1665.08	1646.45	1281.75	1249.94
	(108.51)	(103.25)	(102.50)	(57.64)	(67.29)	(104.26)	(190.03)
25th Percentile	1490.00	1075.00	1022.00	1068.00	1010.00	500.00	980.00
Median	2125.00	1565.00	1525.00	1500.00	1526.00	1056.00	950.00
75th Percentile	3460.00	3060.00	3000.00	2900.00	2910.00	2780.00	2400.00
Wage Earnings > $5,000	0.14	0.14	0.14	0.14	0.14	0.14	0.15
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.03)
Non-Wage Earnings > $50	0.46	0.39	0.39	0.39	0.38	0.29	0.18
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.04)
No Earnings and Not in School	-0.26	-0.03	-0.03	-0.02	-0.03	-0.01	-0.03
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Members	387	387	387	377	329	193	83
Matches	30,817,102	24,520,858	22,332,688	433,320	85,036	7,293	2,145
Covariates Matched On							
Age	X	X	X	X	X	X	X
Gender		X	X	X	X	X	X
Race			X	X	X	X	X
Birth State				X	X	X	X
City in 1940				X	X	X	X
Parent Occupation Score					X	X	X
Education						X	X
Occupation							X

Each cell reports the difference between members and their matches on a particular summary statistic of the distribution of earnings in 1940. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The standard errors come from regressions that include a fixed effect for ever unique set of values of the covariates. The reported statistics for matched groups are calculated by matching future members of Congress with people who never become members of Congress but who have the exact same values on the matching covariates indicated at the bottom of the table. The non-members are reweighted so that the joint distribution of all covariates is identical for the member and non-member groups.

Table A.4 – Difference in Parent Occupation Score Between Future Members of Congress and Their Peers, Adults v Youth.

	Diff in Parent Occ Score [0-80]	
Statistic	Adults	Youth
Mean	3.93	5.78
	(0.69)	(0.41)
25th Percentile	0.00	2.00
Median	1.00	3.00
75th Percentile	7.00	10.00
Members	446	349
Matches	46,407	534,925
Covariates Matched On		
Age	X	X
Gender	X	X
Race	X	X
Birth State	X	X
City in 1940	X	X

Each cell reports the difference between the childhood households of future members and their matches on a particular summary statistic of the distribution of occupation score. Occupation score is an estimate of the typical annual wage for an occupation in 1950 divided by 100 and is top coded at 80. Each column corresponds to a particular set of covariates. Robust standard errors in parentheses. The standard errors come from regressions that include a fixed effect for ever unique set of values of the covariates. The reported statistics for matched groups are calculated by matching future members of Congress with people who never become members of Congress but who have the exact same values on the matching covariates indicated at the bottom of the table. The non-members are reweighted so that the joint distribution of all covariates is identical for the member and non-member groups.

A.7 Link Error in the Siblings Analysis

As we discussed above, we find siblings through an automated record linkage approach while we find future members of Congress using a more manual strategy. Previous work finds that the approximately 15% of the links made by this automated record linkage strategy would be rejected by humans (Feigenbaum 2018). Given the large difference we report between future members of Congress and people with very similar background, labeling an average member of the public as a future member of Congress's brother would increase the gap between future members of Congress and the people we think are their brothers. To address this we change the target of the algorithm–rather than maximizing the average of the positive predictive value (PPV) and the true positive rate (TPR), we minimize the $.02 * TPR + .98 * PPV$. This focus on PPV while essentially ignoring TPR brings our false positive rate below 9%.

Despite using a much higher threshold of certainty for counting a member of the public as the future member of Congress's brother, we find largely similar results. As expected, most of the estimates shrink toward zero to some degree, but all of the estimates continue to show a small but noticeable positive selection.

A.8 Bias From Future Members We Can't Link

One threat to our interpretation of our results throughout the paper is error in our manual process for finding future members of Congress. If it is easier to find higher earners, this would bias our estimate of the raw earnings differential between future members of Congress and the general population. To provide a bound on this, imagine that all of the future members of Congress who we could not find in the 1940 Census were average earners. Since we were able to find 63% of all members, this would reduce our estimate of the raw earnings difference by about 40% to $595. This is a large potential bias in the overall results, but it relies on a set of assumptions about the future members who we failed to match that is incredibly strong and may even suggest the wrong bias. Perhaps instead it is easier to find future members of Congress who earn less on average; this error would produce a bias in the opposite direction. More importantly, these potential biases are not a major concern in the main analysis of the paper.

The main goal of this paper is to measure the degree of positive selection into politics and decompose its sources. The bulk of this exercise involves comparing future members of Congress to people who look like them on many observables and measure the remaining differences in private market earnings and education levels. In these analyses in which we condition on a large set of observables, we only care about biases in our linkage that still exist conditional on that large set of observables. We cannot directly test for this form of bias, but we have no reason to believe this bias exists or is driving our results.

A.9 Size of Selection Across Contexts

In this section, we investigate variation in the size of the selection effects across contexts. The key result is that selection does not appear to be lower in the south, despite the lack of general-election competition in that region in this time period, again consistent with the idea

that electoral competition is not the mechanism selecting individuals with higher earnings ability and education for political office.

Before looking at the south and non-south comparison, we start with a sanity check on the benefits of office. A common theme in many models of electoral accountability and running for office is the idea that larger benefits of office encourage the entry and selection of highly skilled individuals (Besley and Coate 1997; Besley 2004; Osborne and Slivinski 1996).[19] Empirical work also suggests that larger politician salaries leads to stronger electoral selection for quality (Gagliarducci and Nannicini 2013) and ideological moderation (Hall 2019). Although we do not have any variation in salaries to exploit—members of the House and Senate generally earn the same salary, except if they hold leadership positions—we can take advantage of the fact that the Senate is a more prestigious and more sought-after office than the House. Senators serve longer terms and hold more power by virtue of their lesser numbers, making the office more appealing. Competition is greater for senate seats, senatorial candidates raise more money, on average, and they serve on corporate boards after retiring at higher rates than members of the House (Palmer and Schneer 2016, 2018).

Table A.6 compares the earnings gap between future members of Congress and non-members across contexts. In the first two columns the outcome variable is labor income in 1940, while in the second two columns the outcome is an indicator for attending college. For comparison, column 1 simply replicates the overall earnings gap for future members from column 4 of Table 4, for comparison purposes. This is the specification matching on age, gender, race, and geographical location. We do not match on education in this table because it is an outcome variable in the second two columns.

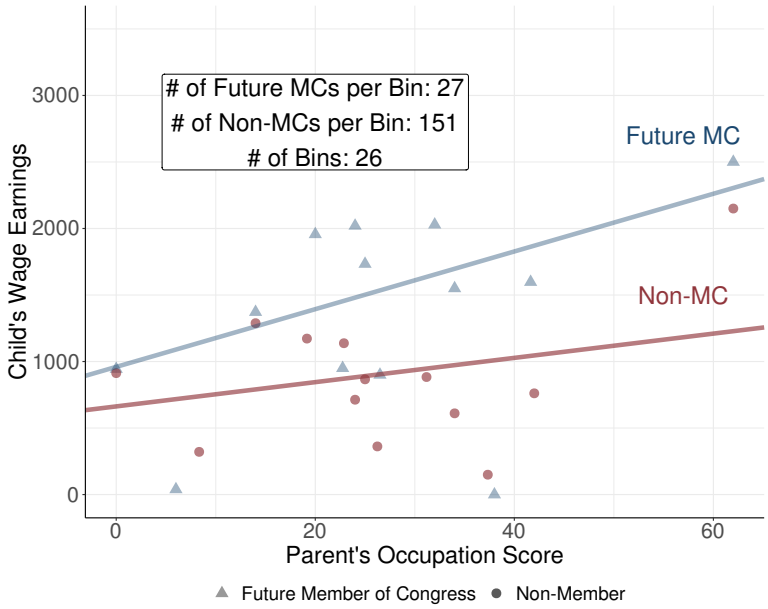
In the second row of the table, looking across the columns, we see that future senators have a much larger earnings gap than future House members—in fact, the earnings gap is more than twice as large for future senators. This indicates that, consistent with the findings in Dal Bó et al. (2017), selection is larger where the benefits of office are higher. However, this does not separate the two mechanisms from one another because where the benefits of office are higher, the barriers to entry may also be higher. It is much more expensive to be a viable senatorial candidate than to be a viable candidate for the House.

The third row is where we examine the south/non-south difference. In column 2 we see that the earnings doesn't appear to be meaningfully different in the south vs. in the rest of the country. In this time period, the south is dominated by the Democratic party, with virtually no general-election competition (e.g., Key 1949). That there is equally large selection even in this setting suggests that electoral competition is not a necessary condition for positive selection.

Columns 3 and 4 replicate these analyses for college attendance. As the second row shows, we again find a premium for future senators. The gap in rates of college attendance for future Senators is roughly 7 percentage points higher than the gap for future members of the House. The other comparisons are more mixed. We see that future members from the south are more likely to have attended college, in contrast to the earnings analysis in column 2. We find no evidence that future Democratic members have more education than future Republican members.

[19] In citizen-candidate models, higher office benefits leads to equilibria with more moderate candidates, a particular kind of "ability."

Figure A.3 – Parental Earnings, Child's Earnings, and Selection into Political Office.



A.10 Comparing America and Sweden

Figure A.3 examines the relationship between parental occupation scores and earnings for adults ages 18 to 40 in the 1940 Census. Lines are estimated separately for future members of Congress and for the comparable set of non-members exactly matched on age, race, gender, birth state, city of residence in 1940, parental occupation score, years of education, and occupation as in column 7 of Table 4. As the figure shows, intergenerational labor income is highly correlated in the United States. Evidence for political selection can be seen in the plot, as the line for future members of Congress is shifted up from the line for non-members; across levels of parents' earnings, future members of Congress earn more than comparable non-politicians.

This figure is important because it helps to square our evidence from the United States with the Swedish evidence from Dal Bó et al. (2017). In Sweden, intergenerational mobility is high, so that parental income does not correlate strongly with children's income. Positive electoral selection in terms of earning ability then means that Swedish politicians earn more than comparable non-politicians, but can still be drawn from across the parental income spectrum.

In the United States, in contrast, positive electoral selection in terms of earning ability—even if it were purely meritocratic—makes an economically representative legislature unlikely, because parental earnings and children's earnings are correlated relatively strongly.

Table A.5 – Difference in Earnings and Education for Male Future Members of Congress and Their Brothers, Certain Matches.

Variable	Diff Btwn MCs and Brothers		
Wage Earnings	218.11	233.26	187.36
	(110.82)	(108.08)	(96.37)
Earnings > $5,000	0.02	0.02	0.02
	(0.02)	(0.02)	(0.01)
Non-Wage Earnings > $50	0.11	0.11	0.12
	(0.03)	(0.03)	(0.03)
No Earnings and Not in School	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)
Years of Education	1.08	1.07	1.07
	(0.23)	(0.21)	(0.20)
High School Completion	0.04	0.04	0.04
	(0.03)	(0.03)	(0.03)
College Attendance	0.07	0.07	0.07
	(0.03)	(0.03)	(0.03)
College Completion	0.08	0.08	0.08
	(0.03)	(0.03)	(0.03)
Covs	None	Age	Age, Age^2
Members	341	341	341
N	1442	1442	1442

Each cell reports an estimate of the difference between members of congress and their siblings. Each row presents the differences for a different variable, including earnings from labor (wage earnings), labor earnings at or above the top of the coded earnings distribution (earnings > $5,000), at least $50 in non-wage earnings (non-wage earnings > $50), no reported capital or labor earnings and not in school (no earning and not in school), years of education, high school completion, college attendance, and college completion. The first column is an unadjusted difference estimated using regressions that include unreported family fixed effect. The second and third columns report differences from the same fixed effects regressions after also adjusting for age differences in the siblings. Robust standard errors reported in the parentheses.

Table A.6 – Comparing Political Selection Effects Across Relevant Dimensions of Heterogeneity.

	Wage Earnings, 1940		College Attendance	
	(1)	(2)	(3)	(4)
Future Member of Congress	1104.54	883.38	0.56	0.52
	(44.40)	(66.24)	(0.01)	(0.02)
Future Member of Congress × Senator		1077.01		0.07
		(164.15)		(0.03)
Future Member of Congress × South		-8.57		0.15
		(102.98)		(0.03)
Future Member of Congress × Democrat		132.17		-0.02
		(89.27)		(0.02)
Members	344	344	539	539
Observations	406,223	406,223	632,330	632,330
Covariates Matched On				
Age	X	X	X	X
Gender	X	X	X	X
Race	X	X	X	X
Birth State	X	X	X	X
City in 1940	X	X	X	X

Regressions reported in all four columns include unreported fixed effects for each unique set of covariate values. In columns two and four, these fixed effects absorb the main effects of Senator, South, and Democrat. Robust standard errors are reported in parentheses below each estimate.