Wealth Inequality Dynamics in the United States: 1962–2100

Thomas Blanchet*

This version: February 14, 2021

Abstract

I decompose the dynamics of the wealth distribution using a simple dynamic stochastic model that separates the effects of consumption, labor income, rates of return, growth, demographics and inheritance. Based on two results of stochastic calculus, I show that this model is nonparametrically identified and can be estimated using only repeated cross-sections of the data. I estimate it using distributional national accounts for the United States since 1962. I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth (mostly in the form of capital gains), and 2 pp. to lower growth. Under current parameters, the top 1% wealth share would reach its steady-state value of roughly 45% by the 2040s, a level similar to that of the beginning of the 20th century. These conclusions apply to a wide class of models of the wealth distribution, regardless of the exact primitives they use to account for, say, consumption or the labor market. I then use the model to analyze the effect of progressive wealth taxation at the top of the distribution.

*Paris School of Economics.

1 Introduction

This paper develops a new approach to address a basic question: what drives the evolution of the wealth distribution, and how does it react to economic, demographic or policy changes?

The economic literature has already made many contributions to our understanding of the wealth distribution (e.g. Wold and Whittle, 1957; Laitner, 1979; Vaughan, 1979; Benhabib, Bisin, and Zhu, 2011). It has been able to explain key stylized facts, in particular its Pareto-shaped tails. It has uncovered several plausible mechanisms that could explain the levels and changes in wealth inequality. It has emphasized, among others, the role of labor income inequality, unequal rates of return, taxation, demographics through children and the sharing of inheritance, and the spread between the rate of return on capital and the rate of economic growth (Stiglitz, 1969; Cowell, 1998; Favilukis, 2013; Piketty and Zucman, 2014; Böhl and Fischer, 2017; Hubmer and Smith, 2018).

But it remains difficult to untangle these effects and assess their respective importance. Theoretical models tend to focus on a limited set of mechanisms and make simplifying assumptions for the sake of tractability. While understandable, this limits our ability to connect these models to the data beyond the replication of the main stylized facts, and use them for policy purposes.

On the empirical side, most of the literature has been concerned with pure measurement issues (e.g. Saez and Zucman, 2016; Kopczuk, 2015; Bricker, Henriques, and Hansen, 2018). Still, a few papers have tried to test certain mechanisms directly using reduced-form specifications. Acemoglu and Robinson (2015) and Góes (2016) have tested the impact of the difference between the rate of return on capital and the growth rate ($r - g$) that was popularized by Piketty (2014), and found no supporting evidence. But these approaches face certain difficulties. Wealth inequality statistics are still in their infancy, with limited time and geographical coverage, and varying quality. This paucity of wealth inequality data makes it hard to get meaningful variation — let alone exogenous variation — that could be used to capture the effects at hand. This is all the more limiting that theory suggests such effects may be slow and take decades to materialize clearly (Gabaix et al., 2016). The studies have avoided the issue by relying on more widely available but fairly noisy proxies such as income inequality, which complicates the interpretation of the results. In a world with a widespread, cross-country dataset on wealth inequality spanning several centuries, it might be easy to just "let the data speak." But until then, pure reduced-form approaches face considerable challenges.

Another method involves the construction of "synthetic saving rates" (Saez and Zucman, 2016; Garbinti, Goupille-Lebret, and Piketty, 2018; Berman, Ben-Jacob, and Shapira, 2016). These synthetic saving rates are calculated so as to reconcile the wealth changes of the different parts of the distribution with the income of the corresponding groups. They are meant to capture many different effects including mobility, the inequality of savings, and their correlation with wealth. In essence, this is an accounting exercise that looks separately at the different parts of the distribution. That approach constitutes a practical middle ground between the theory and the data. Strictly speaking, however, synthetic saving rates can only be interpreted as structural parameters if we assume no mobility between groups, and homogeneous behavior within groups, which affects the domain of applicability of the method and the generality of its conclusions.

Overall, the answer to many questions remains unclear because the empirical literature has been working with synthetic indicators that are hard to tie explicitly to the individual behavior of people. And, as a consequence, it has had difficulties connecting itself to the theory. This state of affairs is at least partly the result of data limitations. Ideally, to directly integrate theoretical models with the data, we would use a long-run, high-quality panel dataset of both income and wealth. Unfortunately, no such thing exists in most countries, including the United States.

In this paper, I suggest a way to resolve this divide between data and theory. Using only repeated cross-sections, and based on two results of stochastic calculus, I am able to nonparametrically identify and structurally estimate the key parameters that determine the dynamics of the wealth distribution. These parameters directly relate to the individual saving behavior of people, so they do not merely capture reduced form relationships between synthetic indicators of wealth inequality. But they remain very general and mostly agnostic as to the exact reason why people save, making the approach compatible with a wide class of models.

This framework is flexible enough to incorporate a realistic model of income, the tax system, inheritance and demographics. Yet it remains simple enough to allow for a transparent identification of the parameters. The model can reproduce the data, and be used run conditional forecasts and counterfactuals in which we change various economic parameters, such as the growth rate, labor income inequality, or rates of return. The model captures both the steady state and the transitional effects of the different shocks — an important feature given that some shocks can take a lot of time to noticeably affect the wealth distribution.

Models of the wealth distribution that can accurately reproduce its Pareto-shaped fat tails virtually all share the same core idea: that people accumulate wealth through a succession of random multiplicative shocks. These may be preference shocks, shocks to rates of return, to the number of children, and so on. What matters is that, as long as the mean and the variance of these shocks falls into the right range, the steady-state distribution will have a power-law tail (Kesten, 1973; Gabaix, 2009). That leads us to the key insight of the paper: although this process of multiplicative random shocks is very general, it actually makes some sharp predictions regarding the evolution of the wealth distribution. And we can exploit these predictions as a source of identification. To that end, it is important not to solely focus on the steady state. Indeed, there is an infinite number of ways in which we could calibrate the mean and variance of the aforementioned multiplicative shocks so as to reach any given steady-state level of inequality, making the model underidentified. But under these various calibrations, the distribution of wealth would change at widely different speeds. Therefore, as long as we observe the wealth distribution outside of its steady state — which is clearly the case in the United States since the 1960s — we can unambiguously identify the parameters of the underlying wealth accumulation process.

In practice, the approach of the paper is made tractable by the use of the continuous time formalism advocated for instance by Gabaix et al. (2016). This formalism provides access to two highly useful results. First, there is the Fokker-Planck equation, which explicitly relates the evolution of the wealth distribution to the parameters of the underlying accumulation process. Then, there is Gyöngy's (1986) theorem. The process of wealth accumulation is itself a function of stochastic processes for income and consumption that are potentially hard to model accurately. Gyöngy's (1986) theorem shows that, in order to properly model the marginal distribution of wealth, it is not necessary to fully model these processes: all we need to know is their mean and variance conditional on wealth. In essence, the mean and variance of savings conditional on wealth turn out to be "sufficient statistics" which entirely define the evolution of the wealth distribution. This considerably reduces the dimensionality of the problem, and makes the analysis much simpler. In the end, and despite the richness of the model, the estimation reduces to the estimation of a linear relationship between observable quantities: therefore, it provides a clear-cut and visual interpretation that can be used to discuss the quality of the fit, or the presence of structural changes.

I apply the method using the Distributional National Accounts (DINA) data from Piketty, Saez, and Zucman (2018), which I complement using the Survey of Consumer Finances (SCF) and various additional sources to account for demography and inheritance. Piketty, Saez, and Zucman (2018) provide public use samples of their data, available yearly since 1962. This data distributes all of national income and wealth to individuals, making it possible to track their distribution in a way that is consistent with macroeconomic totals, and over a long period that include some major economic changes. I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth (mostly in the form of capital gains), and 2 pp. to lower growth. Over the entire period, rich households appear to have been consuming, on average, a constant fraction of their wealth. At the same time they have seen their income rising, due to both higher labor income inequality and capital gains. Hence, they have been saving a higher fraction of their income, leading to an important accumulation of wealth at the top. Under current parameters, the top 1% wealth share would reach its steady-state value of roughly 45% by the 2040s, a level similar to that of the beginning of the 20th century.

This model of the wealth distribution has some practical applications, in particular for the theory of wealth taxation. Recent contributions have emphasized that the long-run elasticity of the capital stock is a sufficient statistic for optimal capital taxation (Saez and Stantcheva, 2018), yet little is known about its value. I use this paper's model to investigate the issue. It allows me to approach the problem in a way that combines insights from several recent contributions, in particular the role of mobility (Saez and Zucman, 2019), tax avoidance, and saving responses (Jakobsen et al., 2019). I develop a simple formula to estimate how the tax base would react to a wealth tax at the top at the steady-state. This formula suggests that the elasticity can be sizeable, but also that it is higher for small tax rates than for larger ones. As a result, revenue-maximizing tax rates may still be quite high.

The rest of the paper is organized as follows: in section 2, I review the main stylized facts about wealth inequality in the United States, and discuss the mechanisms that account for it. In section 3, I explain how I model the three components that shape the distribution of wealth: income and consumption, inheritance, and demography. In section 4, I explain what data I use and how I estimate demography and inheritance in the model. In section 5, I explain how to identify and estimate the main model. Section 6 discusses the results of the model with an application to wealth taxation, and section 7 concludes.

2 The Distribution of Wealth

Over the past few years, there has been a widespread regain of interest in the topic of wealth and its distribution. In the United States, we know the aggregate level of wealth from the official balance sheets compiled by the Federal Reserve. The distribution of that wealth, on the other hand, is a more complicated issue. Historically, there is no official source for the distribution of household wealth, nor is there any direct administrative data sources that we could use the calculate it. Economists, therefore, have had to devise several indirect methods to estimate wealth inequality.

There is the SCF, a triennial survey of household assets conducted by the Federal Reserve. It exists since 1949 (Schularick, Kuhn, and Steins, 2018), with publicly available data since 1962. The SCF serves as the basis for the recently published Distributional Financial Accounts (DFA) of the United States (Batty et al., 2019). However, it has only been conducted regularly and with a consistent methodology since 1989. While the SCF strongly oversamples the richest households, like all surveys it may suffer from some nonresponse and misreporting — and in fact it explicitly excludes extremely wealthy households from its sampling frame for confidentiality reasons. An alternative approach is the capitalization method (Saez and Zucman, 2016), which estimates wealth from administrative capital income tax data.[1] In this paper, I will primarily rely on estimates from the capitalization method as applied by Saez and Zucman (2016), but also use the SCF for some purposes. I will address divergences between the two sources when necessary.

2.1 Empirical Facts about Wealth in the United States

Since the 1980s, household wealth in the United States has grown larger and more concentrated. From 1980 to 2015, the ratio private of private wealth to national income grew from 310% to 450% (figure 1a), reaching a level not seen since before the Second World War (Piketty and Zucman, 2014). Over the same period wealth inequality has also increased (figure 1b). Using the

[1] A third approach is the estate multiplier method (Kopczuk and Saez, 2004), which estimates wealth from inheritance tax data. The estate multiplier stands out from other methods in that it does not find any increase in wealth inequality over the past decades. However, these estimates are usually considered unreliable for the recent period due to differential mortality and tax avoidance (Saez and Zucman, 2016; Kopczuk, 2015), and by now the estate tax has become too narrow to keep on applying the method.

capitalization method, Saez and Zucman (2016) find that the top 1% owns 37% of private wealth in 2014, compared to 23% in 1980. Data from the SCF shows similar trends.

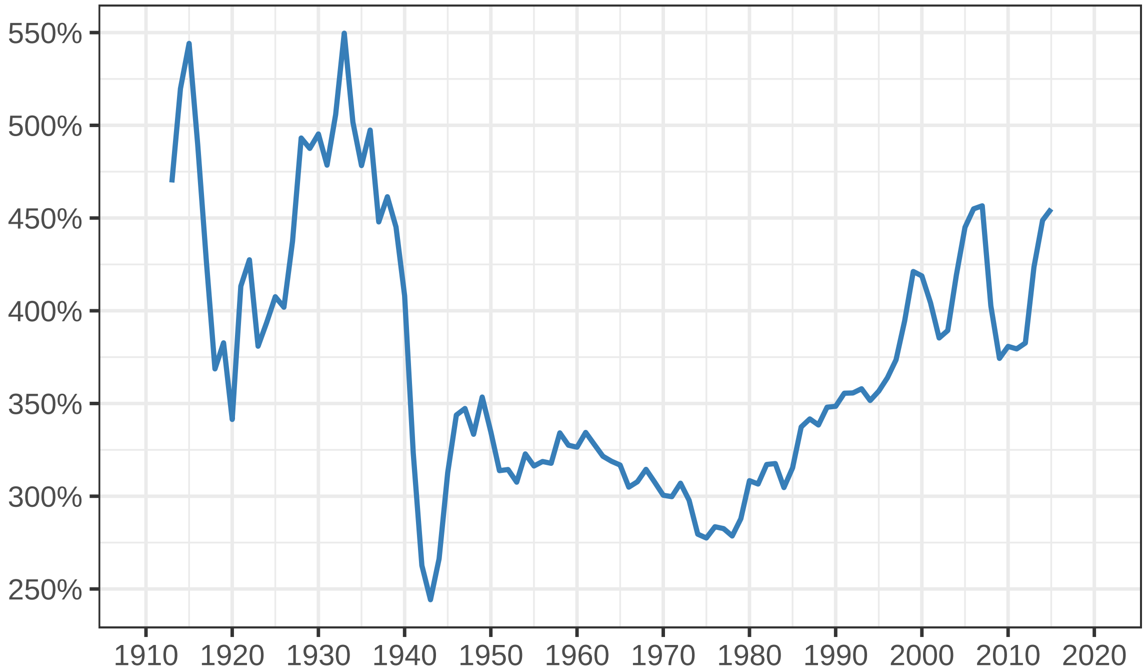
This rise in wealth inequality has been entirely driven by the top tail of the distribution. While there have been some changes for the bottom, notably a rise in the number of indebted households (Wolff, 2010), in practice these dynamics are not central to the topic of increasing inequality. As shown in figure 2a, the top 1% used to hold on average an amount of wealth equal to 70 times average national income in 1980. That amount now exceeds 150 times average national income. During the same period, the bottom 99% owned about 2.5 times the average national income in wealth, a value that remained relatively constant. As a result, if we were to hold constant the wealth/income ratio of the bottom 99% as in figure 1b, the evolution of the top 1% share would be very similar to what we observe in reality. Conversely, if we were to fix the wealth/income ratio of the top 1% at its 1980 level, inequality would not have risen at all.

The rise in wealth inequality has had consequences not only for wealth, but also for income. As shown in figure 3, the share of pre-tax national income owned by the top 1% nearly doubled since 1980, going from 11% to 20%. That increase can only be partially explained by the rise in labor income inequality. Since the early 2000s, the top 1% share of labor income has been mostly flat while the top 1% share of total income has kept on increasing. Similarly, up until 1980, income inequality was slightly decreasing even though labor income inequality was on the rise. These divergences can only be explained by changes in the distribution of capital income, which is directly related to the distribution of wealth.

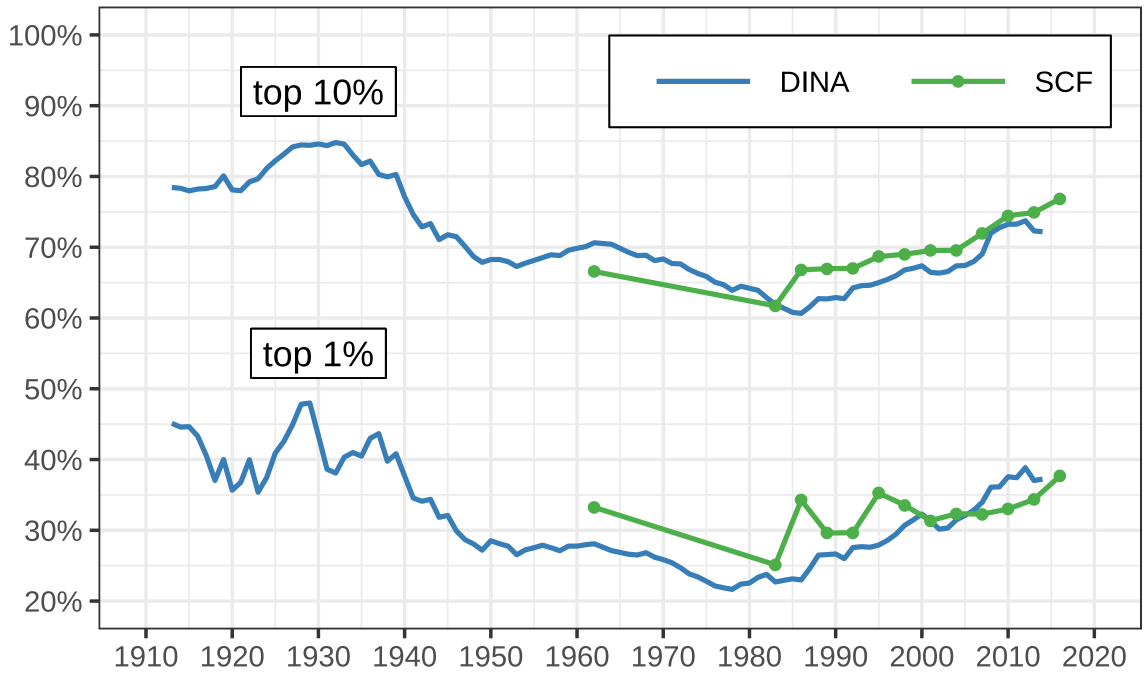
2.2 Mechanisms That Account for Wealth Inequality

The standard models of savings that explain behavior for the bulk of the distribution do not, in general, account well for the shape of the distribution in the tail, which as we have seen in section 2.1 is what explains the increase in inequality. This is true of life-cycle models (Atkinson, 1971) and precautionary saving models (Carroll, 1998). The models that can realistically reproduce the distribution usually incorporate a taste for wealth, either directly (Carroll, 1998; Piketty and Zucman, 2014) or as a bequest motive (Benhabib, Bisin, and Zhu, 2011), and random shocks to preferences (Piketty and Zucman, 2014), number of children (Cowell, 1998), or rates of return (Benhabib, Bisin, and Zhu, 2011). The key feature of all these models is that wealth follows a transition equation of the form $w_{t+1} = a_t w_t + b_t$, where a_t and b_t are random. This type of multiplicative process with random shocks was studied by Kesten (1973), who showed that regardless of the exact distribution of a_t and b_t, w_t converges towards a distribution with a power-law tail.

The Kesten (1973) process justifies why, broadly speaking, power laws arise from multiplicative random shocks with frictions. However, the discrete time formalism of Kesten (1973) quickly gets intractable, so for more elaborate applications it is better to move to continuous time. In continuous time, we can model wealth accumulation as a stochastic differential equation



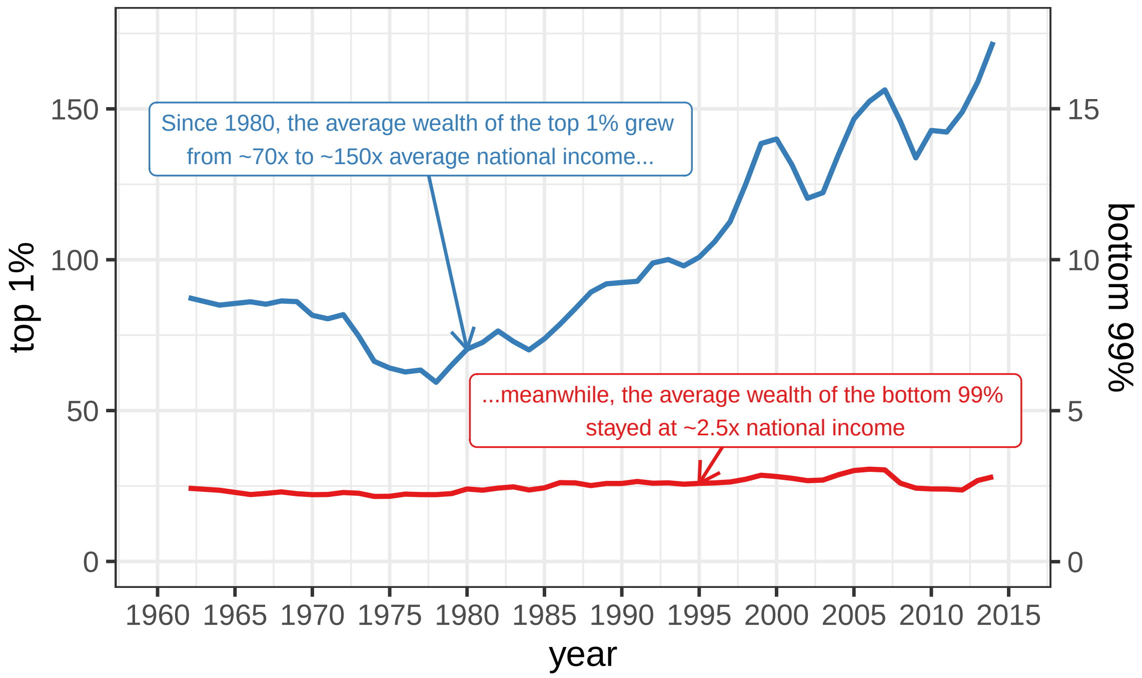
(a) Private Wealth to Income Ratio



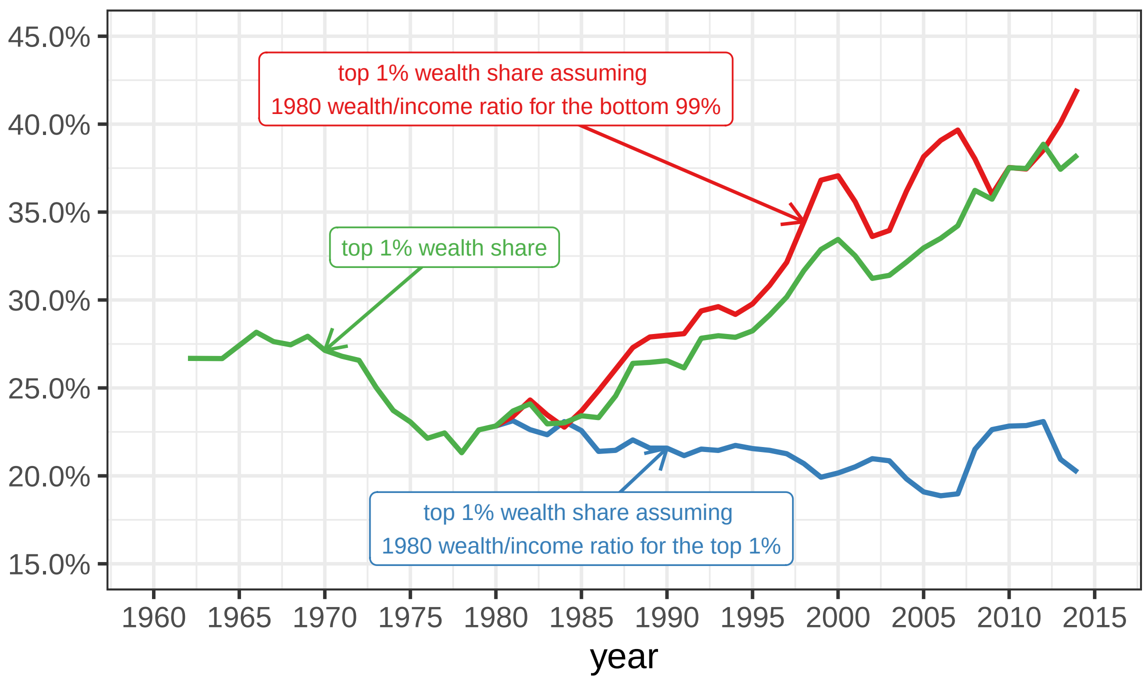
(b) Wealth inequality

Source: Figure 1a: DINA data from Piketty, Saez, and Zucman (2018). Figure 1b: author's computations using DINA data from Piketty, Saez, and Zucman (2018) and SCF data. *Note:* For the wealth-to-income ratios, the income concept for the denominator is net national income. For inequality data, the unit is the adult (20 or older) individual, and wealth is split equally between members of couples.

Figure 1: Private Wealth in the United States



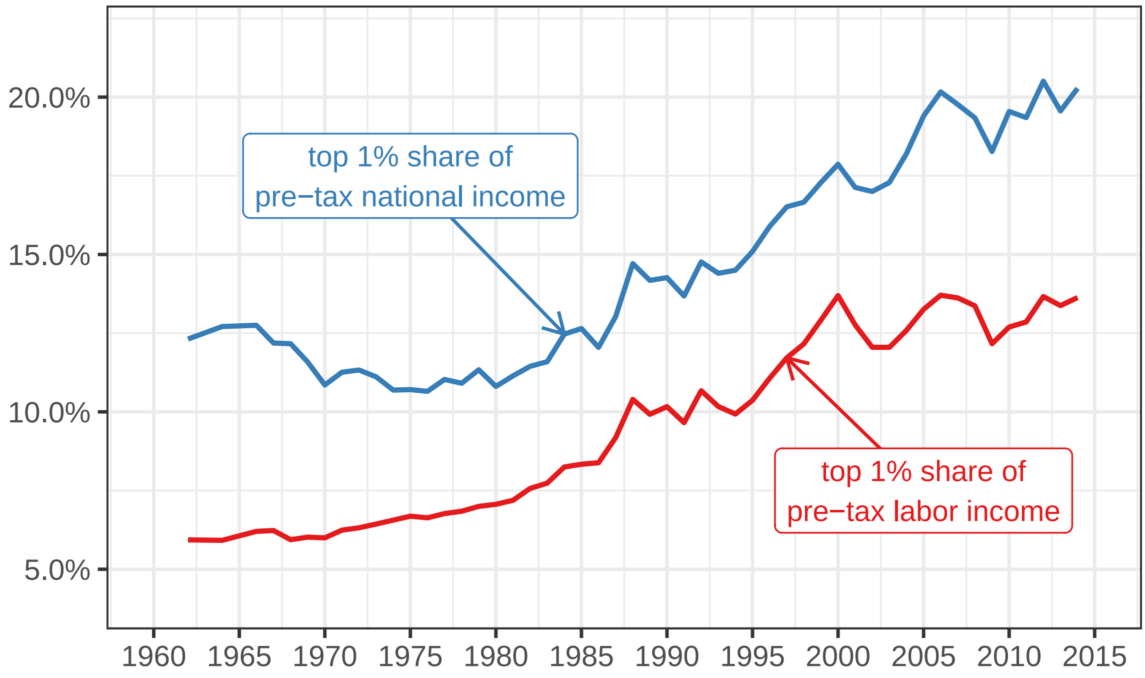
(a) Average Wealth/National Income



(b) Top 1% Share

Source: Author's computation using DINA data from Piketty, Saez, and Zucman (2018). *Note:* For the wealth-to-income ratios, the income concept for the denominator is net national income. For inequality data, the unit is the adult (20 or older) individual, and wealth is split equally between members of couples.

Figure 2: Wealth Inequality: The Role of the Top 1% vs. the Bottom 99%



Source: Author's computation using DINA data from Piketty, Saez, and Zucman (2018). *Note:* The unit is the adult (20 or older) individual, and income is split equally between members of couples.

Figure 3: Income Inequality: Labor and Total Income

(SDE). Like a deterministic differential equation, a SDE relates the current value of a variable to its immediate evolution (i.e. its derivative). But it is stochastic because it assumes that this relationship involves some randomness. Concretely, while a first-order ordinary differential equation for w_t may be written $\frac{\partial}{\partial t} w_t = \mu_t(w_t)$, a SDE formalizes the idea that $\frac{\partial}{\partial t} w_t = \mu_t(w_t) +$ "noise". The proper formalization of "noise" in continuous time is called a Wiener process. Traditionally, we write:

$$\mathrm{d}w_t = \mu_t(w_t)\,\mathrm{d}t + \sigma_t(w_t)\,\mathrm{d}B_t \tag{1}$$

to say the variance of the "noise" over a small amount of time $\mathrm{d}t$ is $\sigma_t^2(w_t)\,\mathrm{d}t$, so that the derivative of w_t is random with mean $\mu_t(w_t)$ and standard deviation $\sigma_t(w_t)$. The value $\mu_t(w_t)$ is called the drift, and $\sigma_t(w_t)$ the diffusion. Using $\mu_t(w_t) = a + bw_t$ and $\sigma_t^2(w_t) = c + dw_t^2$, we get a continuous-time analog to the Kesten (1973) process and, assuming proper parameter values, we converge to a power law. More generally, if we assume $\mu_t(w_t) \propto w_t$ and $\sigma_t(w_t) \propto w_t$ for high w_t, and that some friction prevents w_t from becoming too small, then we converge towards a power law (Gabaix, 2009). The continuous time framework allows us to abstract ourselves from short-term effects that are not relevant in practice but can seriously complicate the analysis.

While the evolution of w_t in equation (1) is random at the individual level, we can characterize the distribution of w_t at the aggregate level using the Fokker-Planck equation:

$$\frac{\partial}{\partial t} f_t(x) = -\frac{\partial}{\partial x}\left[\mu_t(x) f_t(x)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma_t^2(x) f_t(x)\right] \tag{2}$$

This is a deterministic partial differential equation that characterizes the evolution of the density

f_t of w_t at time t, and which will be central the methodology of this paper because it lets us connect the way the wealth distribution evolves with the underlying parameters of the wealth accumulation process.

3 Theoretical Framework

Time is continuous, indexed by t. The distribution of wealth is driven by three factors: income and consumption, birth and death, and inheritance. We treat each of them in turn.

3.1 Income and Consumption

At each time t, the individual i holds W_{it} in wealth, consumes C_{it}, earns Z_{it} in labor income, and gets a rate of return r_{it} on their wealth (including capital gains, if any). At the individual level, wealth follows the differential equation:

$$\frac{\partial}{\partial t} W_{it} = Z_{it} + r_{it} W_{it} - C_{it}$$

Let $\bar{Y}_t$ be the average income (labor and capital), and $g_t \equiv \frac{\partial}{\partial t} \bar{Y}_t / \bar{Y}_t$ is the growth rate of average income. Define $w_{it} \equiv W_{it}/\bar{Y}_t$, $z_{it} \equiv Z_{it}/\bar{Y}_t$ and $c_{it} \equiv C_{it}/\bar{Y}_t$. To stationarize the dynamics of wealth, I will be working with these normalized quantities. The evolution of wealth becomes:

$$\frac{\partial}{\partial t} w_{it} = z_{it} + (r_{it} - g_t) w_{it} - c_{it}$$

Define $y_{it} \equiv z_{it} + (r_{it} - g_t) w_{it}$, so that $\frac{\partial}{\partial t} w_{it} = y_{it} - c_{it}$. I now introduce stochasticity to the income process and the consumption process. Assume, without much loss of generality, that over a small interval of time $[t, t + \mathrm{d}t]$, income (y_{it}) and consumption (c_{it}) are random with mean $\nu_{it}\,\mathrm{d}t$ and $\mu_{it}\,\mathrm{d}t$, and variance $\tau_{it}^2\,\mathrm{d}t$ and $\sigma_{it}^2\,\mathrm{d}t$ respectively (ν_{it}, μ_{it}, τ_{it}^2 and σ_{it}^2 being themselves random processes). Then wealth evolves according to the SDE:

$$\mathrm{d}w_{it} = [\nu_{it} - \mu_{it}]\,\mathrm{d}t + [\tau_{it}^2 + \sigma_{it}^2]^{1/2}\,\mathrm{d}B_{it}$$

where B_{it} is a Wiener process.[2] That SDE has stochastic coefficients, which prevents us from directly applying the Fokker-Planck equation (2). To avoid the need to explicitly model the income and consumption processes separately, I apply a result of stochastic calculus known as Gyöngy's (1986) theorem, which allows us to drastically reduce the dimensionality of the problem to solely focus on wealth.

[2]This formulation implicitly assumes that income and consumption are uncorrelated conditional on wealth. But the analysis still holds if they are. Define $\rho_{it} \equiv \mathrm{Cov}(y_{it}, c_{it})$. Then the equation holds if we redefine σ_{it}^2 to include covariance as $\sigma_{it}^2 + 2\rho_{it}$.

Theorem (Gyöngy, 1986). Let $\boldsymbol{X}_t$ be a n-dimensional stochastic process satisfying:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\alpha}_t \, \mathrm{d}t + \boldsymbol{\beta}_t \, \mathrm{d}\boldsymbol{B}_t$$

where $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are bounded and nonanticipative $n \times 1$ and $n \times m$ stochastic processes, respectively, $\boldsymbol{\beta}_t \boldsymbol{\beta}_t'$ is uniformly positive definite, and $\boldsymbol{B}_t$ is a m-dimensional Wiener process. Then there is a Markov process $\boldsymbol{Y}_t$ satisfying:

$$\mathrm{d}\boldsymbol{Y}_t = \boldsymbol{a}_t(\boldsymbol{Y}_t) \, \mathrm{d}t + \boldsymbol{b}_t(\boldsymbol{Y}_t) \, \mathrm{d}\boldsymbol{B}_t$$

where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ have the same marginal distributions for each t. We can construct $\boldsymbol{Y}_t$ by setting:

$$\boldsymbol{a}_t(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{\alpha}_t | \boldsymbol{X}_t = \boldsymbol{y}] \qquad \boldsymbol{b}_t(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{\beta}_t \boldsymbol{\beta}_t' | \boldsymbol{X}_t = \boldsymbol{y}]^{1/2}$$

Gyöngy's (1986) theorem implies that we can write:

$$\mathrm{d}w_{it} = [\nu_t(w_{it}) - \mu_t(w_{it})] \, \mathrm{d}t + [\tau_t^2(w_{it}) + \sigma_t^2(w_{it})]^{1/2} \, \mathrm{d}B_{it} \tag{3}$$

where $\nu_t(w)$, $\mu_t(w)$ are the means of income and consumption conditional on wealth, and $\tau_t^2(w)$, $\sigma_t^2(w)$ are the variances of income and consumption conditional on wealth.[3]

The Fokker-Planck equation associated to (3) and which describes the density of wealth f_t is:

$$\frac{\partial}{\partial t} f_t(w) = -\frac{\partial}{\partial w} \left[(\nu_t(w) - \mu_t(w)) f_t(w) \right] + \frac{1}{2} \frac{\partial^2}{\partial w^2} \left[(\tau_t^2(w) + \sigma_t^2(w)) f_t(w) \right] \tag{4}$$

3.2 Birth and Death

I extend the model above with a birth and death process. People die randomly according to year, age and sex-specific fertility rates. Let g_t be the density of wealth weighted by these mortality rates. Other people appear with a random initial endowment drawn from a distribution with density h.

Let β_t and δ_t be the overall birth and death rate. The total population N_t grows at a rate $n_t = \dot{N}_t / N_t = \beta_t - \delta_t$. Adding this process turns equation (4) into:

$$\begin{aligned} \frac{\partial}{\partial t} f_t(w) = & \underbrace{-\frac{\partial}{\partial w} \left[(\nu_t(w) - \mu_t(w)) f_t(w) \right] + \frac{1}{2} \frac{\partial^2}{\partial w^2} \left[(\tau_t^2(w) + \sigma_t^2(w)) f_t(w) \right]}_{\text{income and consumption}} \\ & + \underbrace{\beta_t h(w) - \delta_t g_t(w) - n_t f_t(w)}_{\text{birth and death}} \end{aligned}$$

[3] See appendix B.1 for details on how to arrive at that result.

3.3 Inheritance

The wealth of people who die gets redistributed to their spouse or children, after payment of the estate tax, if any. Contrary to income that can be viewed as a continuous flow, inheritance is punctual and introduces a discontinuity in the evolution of wealth. So I model it as a jump process.

The inheritance process is partially connected to the demographic process: it redistributes the wealth of people who die in a given year to their next of kin. The way that inheritance is redistributed depends on the estate tax and additional parameters that capture intergenerational mobility (i.e. do wealthier people inherit more?) I explain how I fully model the process in section 4.2. For now I take the joint distribution of inheritance and wealth as given.

With a probability $\pi_t(w)$, people see their wealth jump from w to $w + \lambda$ where λ is the amount of inheritance received, net of taxes. Let $s_t(\lambda|w)$ be the density of the value of the inheritance, conditional on the value of wealth, and conditional on receiving inheritance. We can model the jump process as a death with rate $\pi_t(w)$ and as an injection with rate $\int \pi_t(w-\lambda)f_t(w-\lambda)s_t(\lambda|w-\lambda)\,\mathrm{d}\lambda$:

$$\begin{aligned}\frac{\partial}{\partial t} f_t(w) = & \underbrace{-\frac{\partial}{\partial w}\left[(\nu_t(w)-\mu_t(w))f_t(w)\right] + \frac{1}{2}\frac{\partial^2}{\partial w^2}\left[(\tau_t^2(w)+\sigma_t^2(w))f_t(w)\right]}_{\text{income and consumption}} \\ & + \underbrace{\beta_t h(w) - \delta_t g_t(w) - n_t f_t(w)}_{\text{birth and death}} \\ & + \underbrace{\int \pi_t(w-\lambda)f_t(w-\lambda)s_t(\lambda|w-\lambda)\,\mathrm{d}\lambda - \pi_t(w)f_t(w)}_{\text{inheritance}}\end{aligned} \tag{5}$$

4 Data, Demography and Inheritance

4.1 Demography

I compute the entire demography of the United States from 1850 to 2100. Although the income and wealth data does not start until 1962, the model requires demographic data that starts much earlier. Indeed, I need to simulate how wealth gets transmitted from one generation to the next. Therefore, if a supercentenarian dies in the 1960s, I have to be able to simulate their entire life history to know how many live children they have, and how old they are. For all years and all ages, I estimate data on the population structure by age and sex, mortality (i.e. life tables), fertility (for both sexes) and intergenerational ties (age and sex of children). Sometimes, data is only available by age groups (e.g. of five years) or a subset of years (e.g. every ten years). Whenever necessary, I interpolate estimates with a monotonic cubic spline (Fritsch and Carlson, 1980) to get data for every single year and age.

Population by Age and Sex Before 1900, I directly estimate the population pyramid using decennial census microdata from the IPUMS USA database (Ruggles et al., 2019). From 1900 to 1932, I use the National Intercensal Tables from the United States Census Bureau. From 1933 to 2016, I use population estimates from the Human Mortality Database.[4] After 2016, I use the projections from the World Population Prospects (United Nations, 2017).

Life Tables Before 1900, I use the historical life tables from Haines (1998). From 1900 to 1932, I use the Human Life Table Database, and from 1933 to 2016, life tables from the Human Mortality Database. After 2016, I rely on projections from the World Population Prospects (United Nations, 2017). All the tables are broken down by sex.

Age-Specific Fertility Rates by Birth Order I estimate age-specific fertility rates by birth order, for both sexes. For women, they are directly available from 1933 to 2016 from the Human Fertility Database. From 1917 to 1932, I use data from the Human Fertility Collection. That same source provides fertility rates until going back to 1895–1899, but without the breakdown by birth order. Therefore, before 1917, I assume that the birth order composition remains constant. Before 1895, there is no age-specific data available, so I use the data on total fertility rate and rescale the age profile from 1895–1899 to that value.[5]

Unlike female fertility rates, male fertility rates are not a standard demographic indicator, so they are not directly available from any source. To estimate them, I combine the age-specific female fertility rates with the joint distribution of the age of opposite-sex couples since 1850 calculated using the decennial census microdata from the IPUMS USA database (Ruggles et al., 2019).

Age and Sex of Children I simulate the distribution of the number, age and sex of living children for in each year after 1962 (when income and wealth data starts), every age and both sexes, which allows me to realistically model how wealth gets transmitted from one generation to the next. To that end, I combine all of the data above. I make every person have children randomly over their past lifetime according to year, age and sex-specific fertility rate. Because I have the breakdown by birth order, I can take into account how the decision to have another child depends on the number of children that one already has. Then, I make each child go through life and die at random according to their year, age and sex-specific mortality rate. As result, I can tie every individual in the database to fictitious descendants that are, on average, representative the true composition of descendants.

[4]See `https://www.mortality.org/hmd/USA/DOCS/ref.pdf` for detailed primary sources.

[5]See Gapminder: `https://www.gapminder.org/news/children-per-women-since-1800-in-gapminder-world/`

4.2 Inheritance and the Estate Tax

Part of the inheritance process is determined by the demographics and the distribution of wealth, while other parts have to be modeled separately. I assume that people die at random, conditional on their age and sex, so that the distribution of inheritances correspond to the distribution of wealth, weighted by mortality rates. I then assume that the wealth of decedents is either redistributed to their spouse (if any) or to their descendants (if they have no living spouse), after payment of the estate tax. The age and sex of decedents are given by the demography (see section 4.1). I assume that inheritance is split equally between children, as is the norm in the United States (Menchik, 1980).

While the demographic aspect of inheritance is endogenously determined by demography, I still need to model separately how wealth gets distributed for a given age and sex. This captures intergenerational wealth mobility in the sense that wealthier people might also have wealthier parents and thus inherit more. There are two aspects to this question: the extensive margin (how likely are you to receive inheritance in a given year?) and the intensive margin (how much inheritance do you receive?) To address this question, I use data from the SCF, which has been recording inheritance consistently since 1989. Note that because the probability of receiving inheritance in a given year is very low overall (about half a percent, see figure 4a), I have to pool all the 1989–2016 waves in order to get sufficient sample sizes.

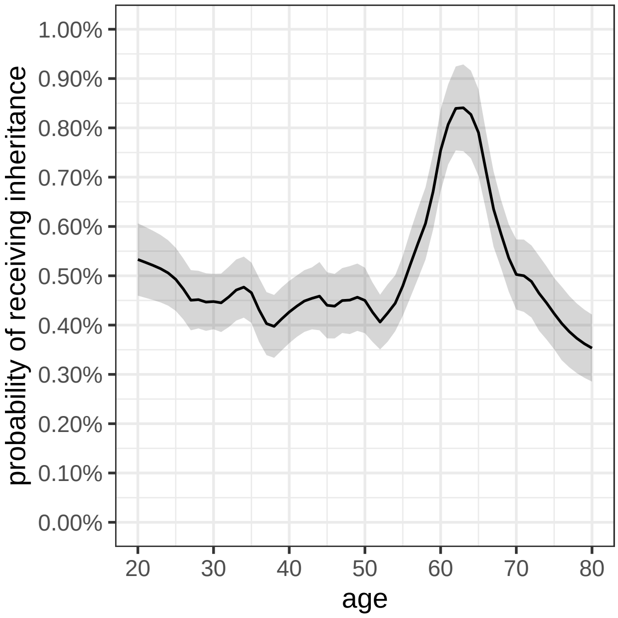
Extensive Margin Let $D_i = 1$ if individual i receives inheritance, and $D_i = 0$ otherwise. Let A_i be their age, and W_i their wealth. Assume that:

$$\mathbb{P}\{D_i = 1 | A_i = a, W_i = w\} = \mathbb{P}\{D_i = 1 | A_i = a\} \phi(F_{A_i=a}(w)) \tag{6}$$

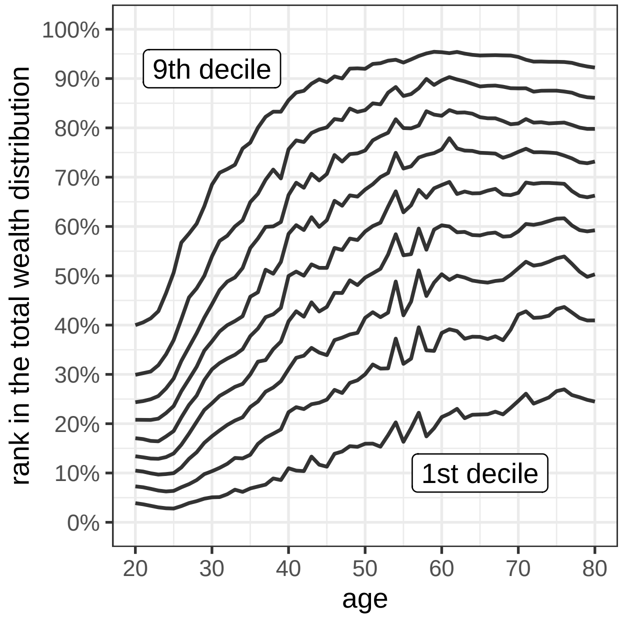
where $F_{A_i=a}$ is the cumulative distribution function (CDF) of wealth conditional on age, and $\int_0^1 \phi(r) \, \mathrm{d}r = 1$. By construction, the expected value of the right-hand side of (6) conditional on age is equal to $\mathbb{P}\{D_i = 1 | A_i = a\}$ so that the specification makes probabilistic sense.[6] Note that $F_{A_i=a}(w)$ is the rank of w in the wealth distribution (conditional on age), which is how we can make the formula (6) consistent regardless of the shape of the wealth distribution.

The value of $\mathbb{P}\{D_i = 1 | A_i = a\}$ is determined by demography, so we only need to estimate ϕ. I start by calculating a rank in the wealth distribution conditional on age by running nonparametric quantile regression of wealth on age for every percentile (see figure 4b). I then regress the dummy D_i for having received inheritance on that rank, multiplied by $\mathbb{P}\{D_i = 1 | A_i = a\}$. I use ordinary least squares (OLS) and a cubic polynomial with coefficients constrained so that its integral over $[0, 1]$ equals one (see figure 4c). As we can see, even after partialling out the effect of age, wealthier people still experience a higher probability of receiving inheritance. I use that polynomial as my estimate of ϕ.

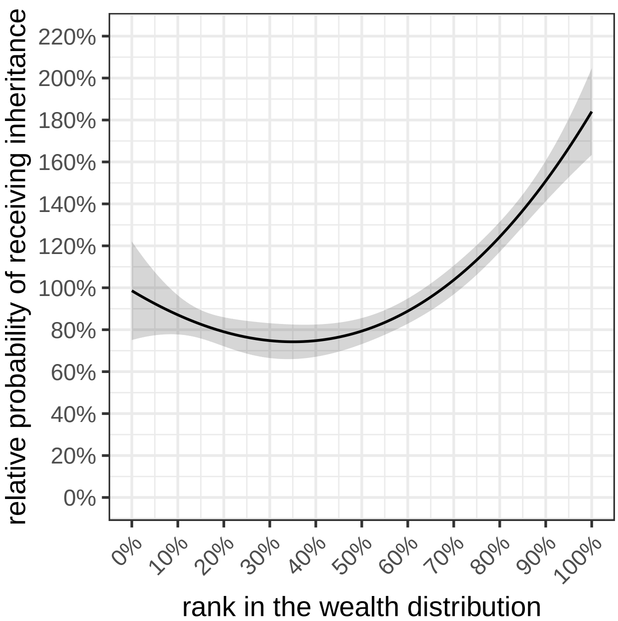
[6] It is the direct result of a change of variable $r = F_{A_i=a}(w)$ and using the fact that $\frac{\partial}{\partial w} F_{A_i=a}(w) = f_{A_i=a}(w)$, so that $\int_{-\infty}^{+\infty} \phi(F_{A_i=a}(w)) f_{A_i=a}(w) \, \mathrm{d}w = \int_0^1 \phi(r) \, \mathrm{d}r = 1$.



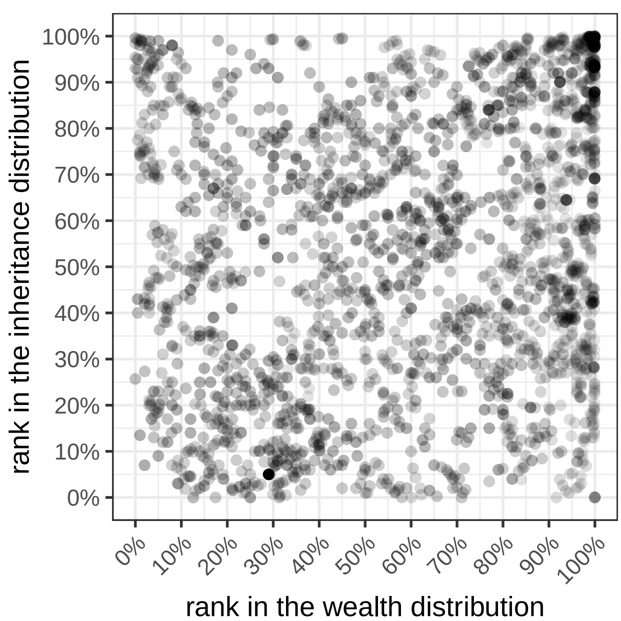
(a) Probability of Receiving Inheritance, Conditional on Age



(b) Rank in the Wealth Distribution, Conditional on Age



(c) Relative Probability of Inheritance, Conditional on Age



(d) Joint Ranks in the Wealth and the Inheritance Distribution, Conditional on Age

Source: Author's computation using the SCF (1989–2016). Gray ribbons correspond to the 95% confidence intervals. In figure 4d, opacity is proportional to the weight of observations.

Figure 4: Modeling of Inheritance

Intensive Margin I account for the intensive margin by modeling the joint distribution of the ranks in the wealth distribution and the inheritance distribution (i.e. the copula), conditional on age and on having received inheritance. I take the subsample of inheritance receivers and calculate their rank in the wealth and the inheritance distribution using nonparametric quantile regression as I did for the extensive margin.

The dependence between the two ranks is weak, but significant (see figure 4d): their Kendall's tau is equal to 7.2%. I represent this dependency parametrically using a bivariate copula. I select the most appropriate model out of a large family of 15 single-parameter copulas by finding the best fit according to the Akaike information criterion (AIC), which is the Joe copula.[7][8] I estimate its parameter so as to match the empirical value for Kendall's tau.

Estate Tax I account for the federal estate tax using the complete estate tax schedule and exemption amount for each year. The top marginal estate tax rate has followed a clear inverted U-shaped pattern over the 20th century (figure 5a), having been reduced by half since its mid-century peak. However, the changes to the overall progressivity of the estate tax are more ambiguous (figure 5b). While the top marginal tax rate was very high in the 1950s, the top bracket did not kick in until extremely high levels of wealth. The 1980s reforms significantly reduced the top tax rate and increased the exemption amount, so that by 1990, the very top and the upper middle of the wealth distribution were facing lower average tax rates. But individual owning about \$10M of wealth were actually facing slightly higher average tax rates. By now, however, the estate has been lowered so much that its profile is unambiguously less progressive than in the 1950s.

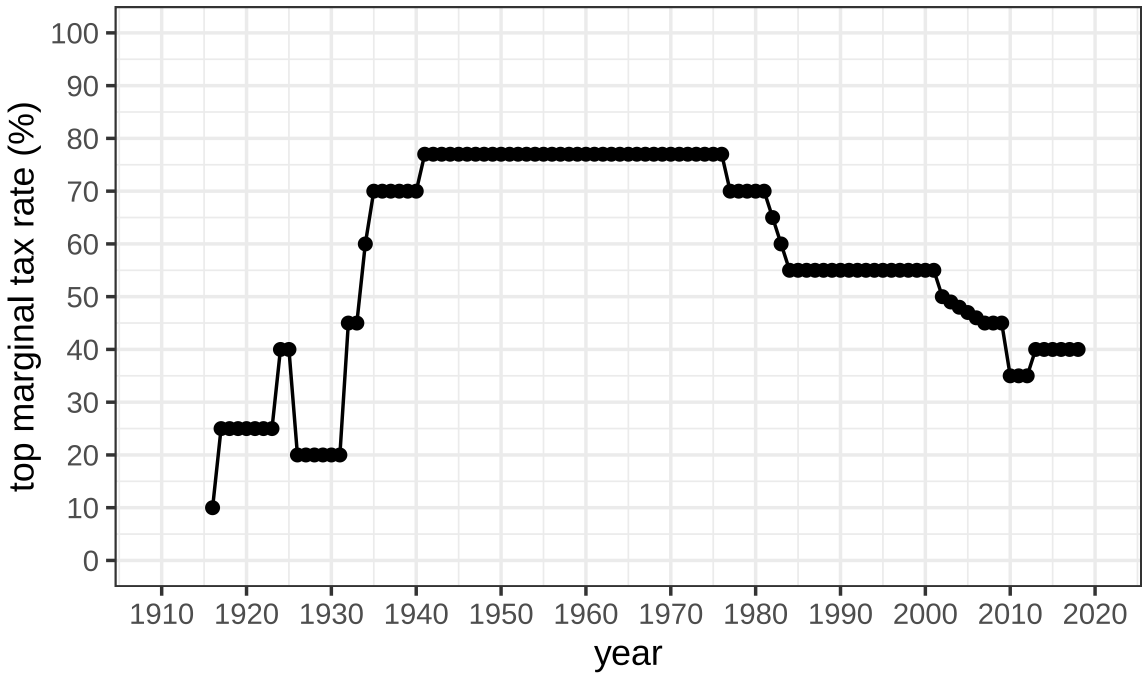
4.3 Income and Wealth

For the income and wealth data, I primarily rely on the DINA public microdata from Piketty, Saez, and Zucman (2018). These files are annual (except for 1963 and 1965) since 1962. Each observation corresponds to an adult individual (20 or older), and each variable correspond to an item of the national accounts, that is distributed to the whole adult population. These files distribute the entirety of the income and wealth of the United States. The public version regroups observations for anonymity, so it has smaller sample sizes than the one they use internally, and does not exactly reproduce results from their more complete internal files (Saez and Zucman, 2018). The discrepancies, however, are small.

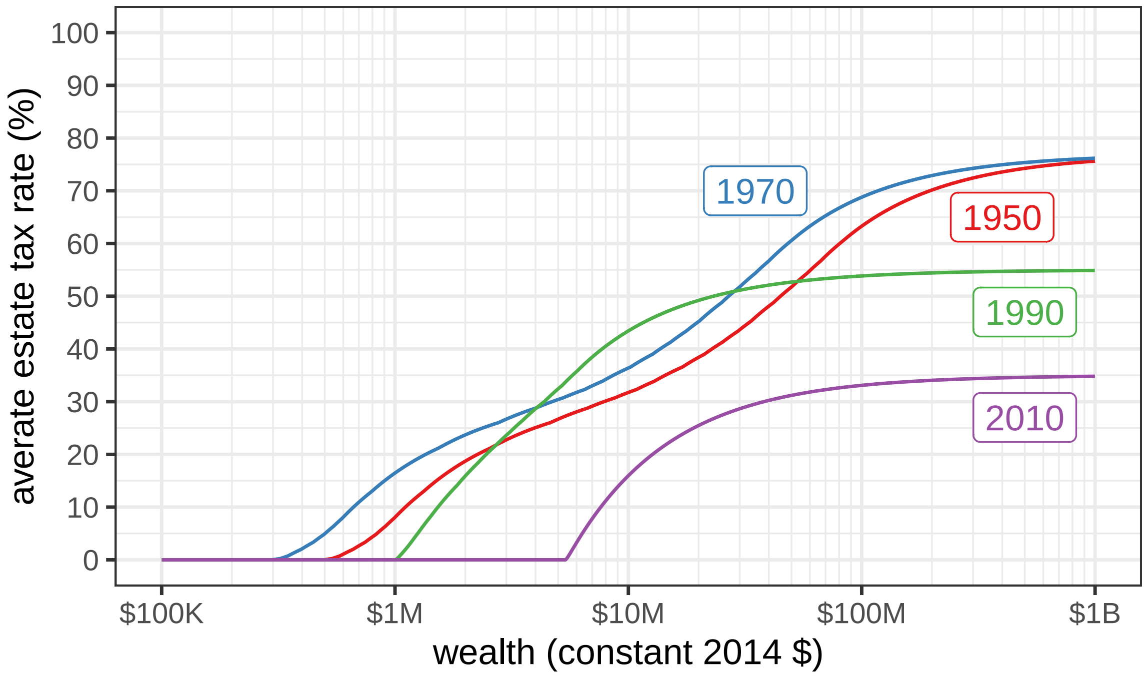
This data has several advantages. It provides distributional estimates that are consistent with macroeconomic aggregates. It has rather large samples (from about 35 000 in the 1960s to about 65 000 today), with oversampling of the richest. And because it is based on tax data, it captures the top tail of the distribution well. It does have some drawbacks, though. First, it has limited

[7]The list of copulas includes the Gaussian copula, Student's t copula, the Clayton copula, the Gumbel copula, the Frank copula, the Joe copula, and rotated versions of these copulas.

[8]The Joe copula has the parametric form $C_\theta(u,v) = 1 - \left[(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right]^{1/\theta}$.



(a) Top Marginal Estate Tax since 1916



(b) Average Estate Tax Rate, by Wealth

Source: Author's computation using the tax schedules of the federal estate tax.

Figure 5: Estate Tax

socio-demographic information: in particular, age information is only available in the form of very broad age groups. Second, it estimates wealth using the capitalization method: that is, it assumes that everyone gets the same rate of return from the same type of asset. Under the right assumptions (Saez and Zucman, 2016), that method provides accurate estimates of the distribution of wealth, and of average income conditional on wealth. But it almost certainly underestimate the variance of capital income conditional on wealth. Third, the data does not include capital gains, because they are not part of national income as defined by the national accounts. For these reasons, I make some adjustments and imputations to these data, using the SCF and national accounts.

I use post-tax national income as my income concept of reference. It corresponds to income after all taxes and transfers. It also distributes government expenditures and the income of the corporate sector to individuals, so as to sum up to net national income.

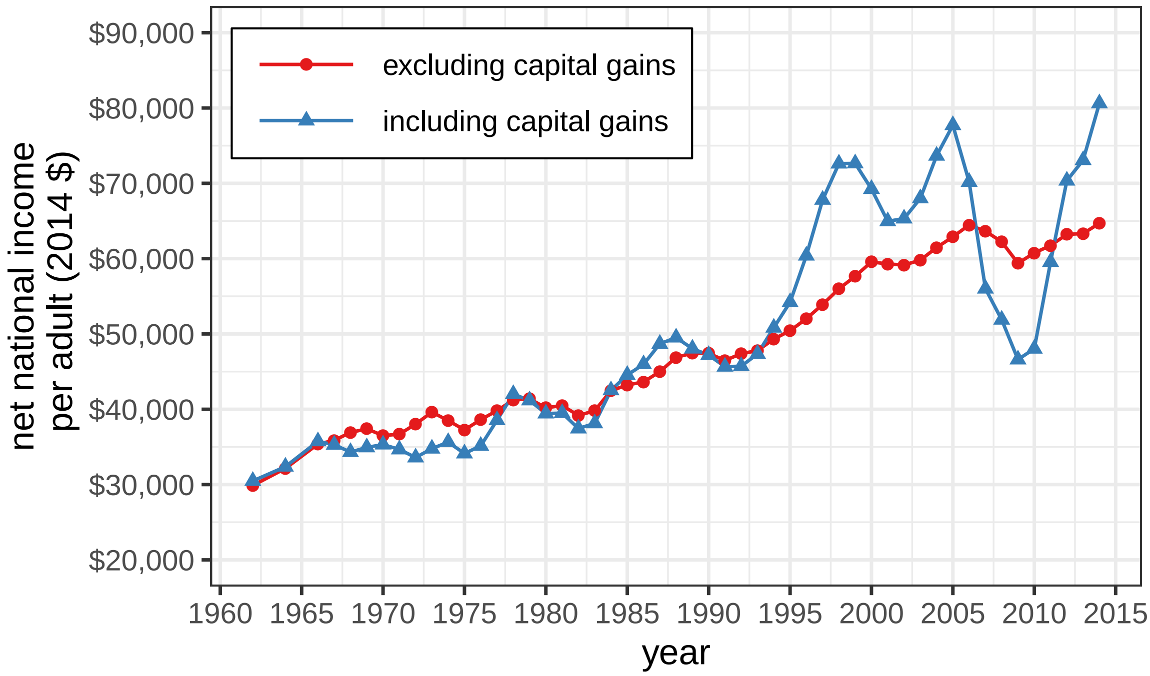
Capital Gains We can measure capital gains when they accrue to individuals, or when they are realized. For our purpose, accrued capital gains are more useful than realized ones, because they are the one that reconcile changes in the value of the balance sheet with national income and savings. Whether a capital gain is realized now or later, on the other hand, is the result of various tax and economic incentives that not relevant here and does not correspond to any meaningful economic aggregate.

The DINA data only records taxable capital gains, which is essentially a measure of realized capital gains. These are a poor proxy for accrued capital gains (Alstadsæter et al., 2017). Instead, I estimate them individually using the capitalization approach of Robbins (2018). I retrieve the rate of capital gains by year and asset type from the national accounts (Piketty, Saez, and Zucman, 2018, table TSD1 in appendix). Then, I assume for a given asset type, everyone gets the same rate of capital gains. By construction, these micro-level estimates of capital gains are consistent with macro totals. Their distribution follows the logic of the Saez and Zucman (2016) capitalization method.[9] Robbins (2018) provides a thorough discussion of why that measure is more appropriate to analyze the role of asset prices changes to inequality and the economy.

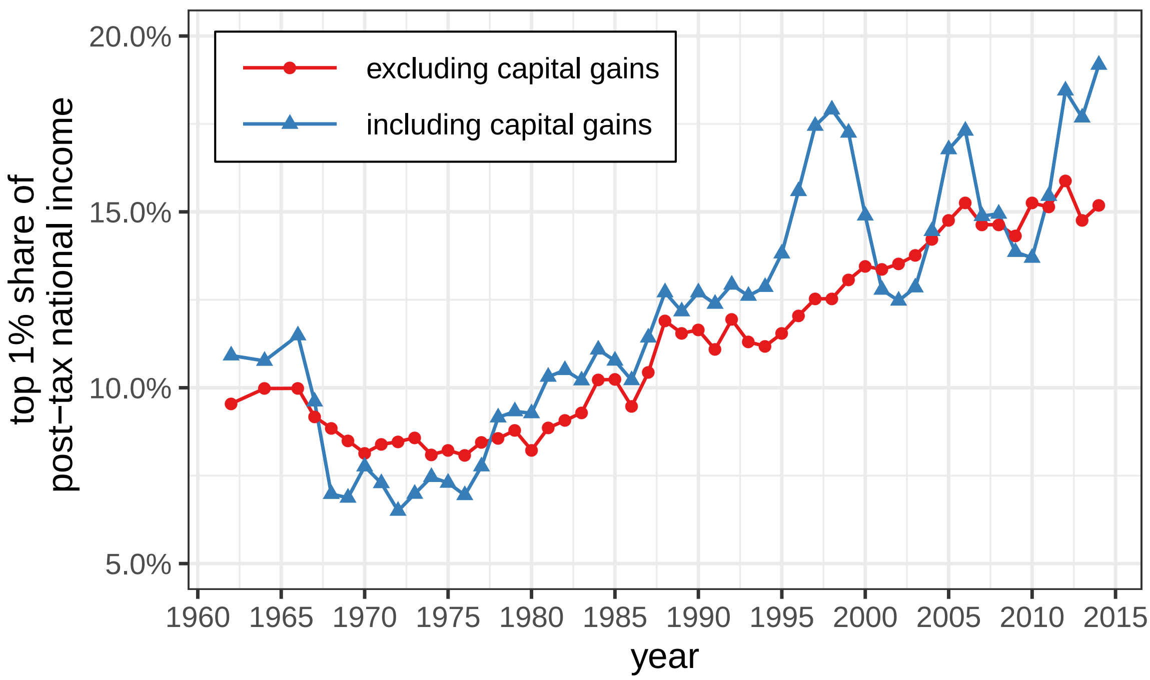
National income including capital gains can be quite volatile (figure 6a), but on average their inclusion matters on several fronts. Robbins (2018) shows that their inclusion overturns certain stylized facts about the United States economy (such as the long run decline of saving rates) and strengthen others (such as the rising capital share and increase of income inequality). As shown in figure 6b, capital gains were dampening the top 1% share of post-tax national income during most of the 1970s, but since then they have consistently increased it.

Wealth by Age The age information in the DINA data is very limited so I cannot use it. Instead, I import it from the SCF and demographic estimates using constrained statistical

[9] Although the income measure in the DINA data does not include capital gains, it does distribute income from the corporate sector to the owners of capital, with may partly account for changes in asset prices. My measure of capital gains is net of retained earnings, so that there is no double counting.



(a) Net National Income, with and without Capital Gains



(b) Post-tax National Income Share, with and without Capital Gains

Source: Author's computations using the public DINA microdata and table TSD1 (online appendix) from Piketty, Saez, and Zucman (2018). *Note:* The unit of analysis is the adult individual (20 or older). Income is split equally between members of couples. Capital gains are estimated assuming a constant rate of capital gains by asset type. Rates of capital gains by asset types smoothed using a 5-year moving average.

Figure 6: The Impact of Capital Gains on National Income and Its Distribution

matching. I calculate the rank in the wealth distribution in both the DINA and the SCF data, and the rank in the age distribution by sex and household type (single or couple) in the SCF data. Then, I match the DINA observations one by one to SCF observations based on their wealth rank to give them a rank in the age distribution.[10] Finally, I use the population structure from the demographic data to attribute an age to every DINA observation. By construction, the method preserve the wealth distribution in DINA, the population by age and sex from demographic sources, and the copula between wealth and age from the SCF.

Variance of Income by Wealth Because the capitalization approach in the DINA data assumes a fixed rate of return by asset type, it is likely to understate the variance of capital income conditional on wealth. Indeed, it will only account for the "between assets" component of total variance, not the "within assets" component. Given that the variance of income conditional on wealth is one of the drivers of the dynamic of wealth, I make an adjustment the DINA estimates using the SCF.

Since 1989 (previous waves provide insufficient data due to lack of oversampling), the standard deviation of the income/wealth ratio at the very top of the distribution is equal to 10.7%, compared to 5.2% in DINA. I use this difference to compute a "within assets" variance component by wealth that I add to the DINA estimates of the income variance conditional on wealth. Note that the survey estimate of this variance is by no means perfect, and is in fact likely to be inflated for two reasons. First, measurement error for either income and wealth might increase the spread of the income/wealth ratio in the survey for spurious reasons. Second, the income in the SCF refers to the previous year, while the wealth refers to the time of the interview: this disconnect introduces additional noise that will have a tendency to also increase the variance of the income/wealth ratio. However, I stress that by construction this adjustment can only affect the interpretation of some parameters of the model, not the overall dynamics of wealth. Indeed, the evolution of wealth ultimately depends on $\sigma_t^2(w) + \tau_t^2(w)$, the sum of the variance of consumption and income. Therefore, as will be explained in section 5, in effect the model will directly estimate the overall variance $\sigma_t^2(w) + \tau_t^2(w)$, and then estimate $\sigma_t^2(w)$ by subtracting $\tau_t^2(w)$. To the extent that we overestimate the variance of income, we will underestimate the variance of consumption, and *vice versa*. In any case, the results of section 6 will be unaffected.

[10] Note that both datasets are weighted, so that observations end up being duplicated and partially matched to one another. When the samples contain M and N observations respectively, the resulting dataset contains at most $M + N - 1$ observations.

5 Identification and Estimation

For concision, define in equation (5):

$$\phi_t(w) \equiv \beta_t h(w) - \delta_t g_t(w) - n_t f_t(w) \qquad \text{(the birth/death effect)}$$

$$\psi_t(w) \equiv \int \pi_t(w-\lambda) f_t(w-\lambda) s_t(\lambda|w-\lambda)\ \mathrm{d}\lambda - \pi_t(w) f_t(w) \qquad \text{(the inheritance effect)}$$

So that the Fokker-Planck equation (5) becomes:

$$\frac{\partial}{\partial t} f_t(w) = -\frac{\partial}{\partial w}\left[(\nu_t(w) - \mu_t(w)) f_t(w)\right] + \frac{1}{2}\frac{\partial^2}{\partial w^2}\left[(\tau^2(w) + \sigma^2(w)) f_t(w)\right] + \phi_t(w) + \psi_t(w)$$

I will use uppercase letters to denote integrated quantities, in particular:

$$F_t(w) = \int_{-\infty}^{w} f_t(s)\ \mathrm{d}s \qquad \Phi_t(w) = \int_{-\infty}^{w} \phi_t(s)\ \mathrm{d}s \qquad \Psi_t(w) = \int_{-\infty}^{w} \psi_t(s)\ \mathrm{d}s$$

5.1 Identification

General Result I integrate the Fokker-Planck equation with respect to w, borrowing a suggestion from Lund, Hubbard, and Halter (2014) in the context of physical chemistry.[11] After reordering terms, I get:

$$\frac{\frac{\partial}{\partial t} F_t(w)}{f_t(w)} - \frac{\Phi_t(w)}{f_t(w)} - \frac{\Psi_t(w)}{f_t(w)} + \nu_t(w) - \frac{1}{2}\frac{\partial}{\partial w}\tau_t^2(w) - \frac{1}{2}\tau_t^2(w)\frac{\frac{\partial}{\partial w} f_t(w)}{f_t(w)} = \mu_t(w) + \frac{1}{2}\frac{\partial}{\partial w}\sigma_t^2(w) + \frac{1}{2}\sigma_t^2(w)\frac{\frac{\partial}{\partial w} f_t(w)}{f_t(w)} \tag{7}$$

The left-hand side of the equation only contains estimable quantities, while the right-hand side is a linear function of $\frac{\partial}{\partial w} f_t(w)/f_t(w)$ whose slope and intercept relate to the unknown parameters $\mu_t(w)$ and $\sigma_t^2(w)$.

Therefore, if these quantities are stable over time, then for a level of wealth w, we should expect $\frac{\partial}{\partial w} f_t(w)/f_t(w)$ and the left side of the equation to fall alongside a straight line. Assuming that there is some variability of both sides of the equation, we are able to estimate the parameters of interest simply by fitting a line. This leads to the following result.

[11] To integrate the equation, we must be able to invert the time derivative with the integral sign, which is allowed either it we assume that the support of wealth is bounded from below, or if the density of wealth is Lipschitz-continuous (i.e. has a bounded derivative).

Theorem (Identifiability of the Model)**.** Assume that there is at least two dates, t_1 and t_2, for which:

(i) For all w, we observe all the quantities in (7), except $\mu_t(w)$, $\sigma_t^2(w)$ and $\frac{\partial}{\partial w}\sigma_t^2(w)$.

(ii) The parameters $\mu_t(w)$ and $\sigma_t^2(w)$ are the same in t_1 and t_2: for all w, $\mu_{t_1}(w) = \mu_{t_2}(w) = \mu(w)$ and $\sigma_{t_1}^2(w) = \sigma_{t_2}^2(w) = \sigma^2(w)$.

(iii) The distribution is different between t_1 and t_2, such that $\frac{\partial}{\partial w}f_{t_1}(w)/f_{t_1}(w) \neq \frac{\partial}{\partial w}f_{t_2}(w)/f_{t_2}(w)$ almost surely.

Then the functions $\mu(w)$ and $\sigma^2(w)$ that satisfy (7) are unique, i.e. the model is identified.

The assumptions required to estimate the model are relatively innocuous. Assumption (i) states that we can observe, or at least separately estimate, all the relevant quantities except consumption, which we seek to identify. Assumption (ii) states that we need some stability in the consumption process over time to be able to estimate it. And assumption (iii) states that we need some variability in the distribution of wealth, so that we cannot already be at the steady state. This is clearly the for the United States since the 1960s. In theory only two observations are needed to estimate the model. In practice it is better to have many more. First, because we need to estimate the time derivative of the CDF of wealth in (7), which requires several data points. Second, because there will always be some measurement error for the different quantities, which can be averaged out when using many data points.

Interpretation in a Simplified Case To better understand the dynamics implied by the estimating equation, consider the following simplified case, which nonetheless capture all the main intuitions of the more complete setting. Ignore the role of demographics ($\Phi_t(w) = 0$), inheritance ($\Psi_t(w) = 0$) and the conditional variance of income ($\tau_t^2(w) = 0$). Consider a high level of wealth w, and assume that at these levels the mean and the standard deviation of consumption are proportional to wealth ($\mu_t(w) \equiv \mu w$ and $\sigma_t(w) \equiv \sigma w$). Define the conditional income-to-wealth ratio $\gamma(w) \equiv \nu_t(w)/w + g$ (note the apparition of the economy's growth rate that was previously included in $\nu_t(w)$ because wealth was normalized by average income). After dividing both sides by w, the estimating equation (7) simplifies to:

$$\frac{\frac{\partial}{\partial t}F_t(w)}{wf_t(w)} + \gamma(w) - g = \mu - \sigma^2 \left(-\frac{1}{2}\frac{w\frac{\partial}{\partial w}f_t(w)}{f_t(w)} - 1 \right) \tag{8}$$

Under these circumstances, the top tail converges towards a power law (Gabaix, 2009). Thus, assume that wealth is Pareto-distributed with Pareto coefficient $\alpha > 1$, i.e. $f_t(w) \propto x^{-\alpha-1}$. Then $-\frac{1}{2}w\frac{\partial}{\partial w}f_t(w)/f_t(w) - 1 = (\alpha-1)/2 > 0$ can serve as a proxy for inequality: the higher it is, the lower inequality.[12] On the left-hand side, inequality increases when $\frac{\partial}{\partial t}F_t(w)/(wf_t(w))$ is negative, and decreases otherwise. We can write equation (8) as $Y_t(w) = \mu - \sigma^2 X_t(w)$.

Figure 7 describes the situation. The Pareto coefficient is on the x-axis: the right side of the figure

[12]We can assume in general that $\alpha > 1$, otherwise mean is infinite.

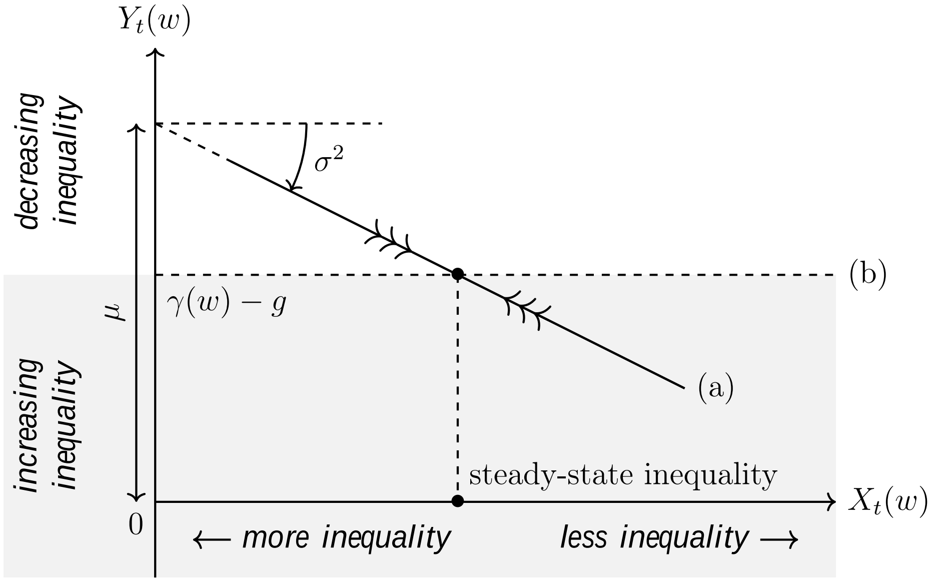


Figure 7: Simplified Dynamics of Wealth Inequality in the Top Tail

corresponds to low inequality, and the left side to high inequality. The y-axis relates to changes in inequality. We can separate the plane into two regions: the gray one where $\frac{\partial}{\partial t}F_t(w)/(wf_t(w)) < 0$ and therefore inequality increases, and the white one where it decreases. The system moves alongside the (a) line, either up or down depending on whether we are in the gray area or the white area. We keep moving up or down until we meet the (b) line that delimits both these areas: thus, the intersection between (a) and (b) indicates the steady-state level of inequality. The slope of (a) is determined by the diffusion coefficient σ^2, which captures mobility, while the intercept μ corresponds to the average consumption/wealth ratio.

This diagram helps perform some comparative statics. An increase in mean consumption at the top implies that the line (a) shifts upwards, leading to a steady state with lower inequality. A higher mobility (that is, an increase in σ^2) increases the slope of (a) while keeping its intercept constant: so it meets the line (b) at a lower value of $X_t(w)$, which implies higher steady-state inequality. If labor income is negligible at the top, then $\gamma(w) \approx r$, so that the line (b) is positioned at $r - g$. Therefore, inequality is an increasing function of $r - g$ (see Piketty, 2014; Piketty and Zucman, 2015).

We can also use the figure to explain what makes the model identifiable. If we solely focus on the steady state, there is an infinity of values of (μ, σ^2) that can reach a given level of inequality, making the model impossible to estimate. Yet these different parameter values would yield very different dynamics of inequality. Having both low consumption and low mobility means that (a) is very flat, therefore $\frac{\partial}{\partial t}F_t(w)/(wf_t(w))$ is very small, and we converge very slowly to the steady state. Reaching the same steady state by having both high consumption and high mobility happens a lot faster. That line of reasoning breaks down if we are already at the steady state, however, which explains why assumption (ii) is required to identify the model.

5.2 Estimation

In essence, the estimation of the model involves fitting the line (a) from figure 7. This section covers how to do so in practice.

Transformation of Wealth Because of its fat tail, it can be difficult to estimate the density of wealth. To overcome the problem, I will be working with wealth transformed using the inverse hyperbolic sine function: $x \mapsto \operatorname{asinh}(x)$. This practice is common is the literature on wealth inequality (e.g. Thompson and Suarez, 2015; Kakar, Daniels, and Petrovska, 2019; Steinbaum, 2019). The transformation is bijective, strictly increasing, behaves linearly from low values and logarithmically for high values. Hence, it acts as a logarithmic transform for the top tail without creating problems for zero or negative wealth. I use Itō's lemma to move from the dynamics of wealth to that of its transform:

$$\mathrm{d}\operatorname{asinh}(w_t) = \left[\frac{\nu_t(w_t) - \mu_t(w_t)}{\sqrt{1+w_t^2}} - \frac{1}{2}\frac{\tau_t^2(w_t) + \sigma_t^2(w_t)}{1+w_t^2}\frac{w_t}{\sqrt{1+w_t^2}}\right]\mathrm{d}t + \frac{(\tau_t^2(w_t) + \sigma_t^2(w_t))^{1/2}}{\sqrt{1+w_t^2}}\,\mathrm{d}B_t$$

There are two changes compared to the dynamics of untransformed wealth. First, all quantities are divided by $\sqrt{1+w_t^2}$, meaning that we use ratio quantities for high values of wealth, and absolute quantities for low values. Second, the drift term is adjusted by a factor that depends on the diffusion. I use tildes to designate to quantities that pertain to transformed wealth: that is, I will write $\tilde{\nu}_t$, $\tilde{\mu}_t$, etc. to denote the variables ν_t, μ_t, etc. divided by $\sqrt{1+w_t^2}$, and use $\tilde{F}_t$ and $\tilde{f}_t$ for the CDF and density of transformed wealth.

Estimating Equation Assume that μ_t and σ_t^2 are the same for all t. To simplify analysis and limit the number of parameters, assume that $\sigma(w) = \tilde{\sigma}\sqrt{1+w^2}$. This assumption meets the usual requirements of the literature for models of the wealth distribution. Because the standard deviation of consumption is scale invariant at the top, it can produce Pareto-shaped tails (Gabaix, 2009). At the same time, by breaking the scale invariance at the bottom, it makes it possible to get a stationary process. This is similar in spirit to what was done by Gabaix (1999) with a strictly positive reflecting barrier, but smoother (see Saichev, Sornette, and Malevergne (2010, p. 16), for more details on that approach).

With that assumption, the estimating equation (7) for transformed wealth becomes:

$$\frac{\frac{\partial}{\partial t}\tilde{F}_t(w)}{\tilde{f}_t(w)} - \frac{\tilde{\Phi}_t(w)}{\tilde{f}_t(w)} - \frac{\tilde{\Psi}_t(w)}{\tilde{f}_t(w)} + \tilde{\nu}_t(w) - \frac{1}{2}\frac{\partial}{\partial w}\tilde{\tau}_t^2(w) + \frac{1}{2}\tilde{\tau}_t^2(w)\left[-\frac{\frac{\partial}{\partial w}\tilde{f}_t(w)}{\tilde{f}_t(w)} - \frac{w}{\sqrt{1+w^2}}\right] = \tilde{\mu}_t(w) - \frac{1}{2}\tilde{\sigma}_t^2\left[-\frac{\frac{\partial}{\partial w}\tilde{f}_t(w)}{\tilde{f}_t(w)} - \frac{w}{\sqrt{1+w^2}}\right] \quad (9)$$

For concision, write this equation as $\tilde{Y}_t(w) = \tilde{\mu}(w) - \tilde{\sigma}^2 \tilde{X}_t(w)$. To see the link with the simplified estimating equation (8), note that the logarithm of a Pareto distributed variable follows an exponential distribution. Consider that for the top of the distribution, $f(w) \propto w^{-\alpha-1}$. Then, transformed wealth approximately follows an exponential distribution with coefficient α, so that $-\frac{\partial}{\partial w}\tilde{f}_t(w)/\tilde{f}_t(w) \approx \alpha$. Furthermore, $w/\sqrt{1+w^2} \approx 1$. Therefore, we have $\tilde{X}_t(w) \approx (\alpha - 1)/2$, as in equation (8). Matters are somewhat more complicated for the left-hand side, though the intuition is similar. The time derivative $\frac{\partial}{\partial t}\tilde{F}_t(w)/\tilde{f}_t(w)$ is equal to zero when the rest of the left-hand side equals the right hand side, which determines the steady state. However, it is not possible anymore to separate the plane neatly into two fixed regions because the effects of demographics ($\tilde{\Phi}_t(w)$) and inheritance ($\tilde{\Psi}_t(w)$) are endogenous to the distribution of wealth, so the steady-state can only be determined through simulations.

Fitting the Model Assume that we observe the system over at a series of dates $(t_1, \ldots, t_k)$. Define a grid of wealth values $(w_1, \ldots, w_n)$. Using equation (9), the estimation of the model reduces to the estimation of a fixed-effect regression with the specification:

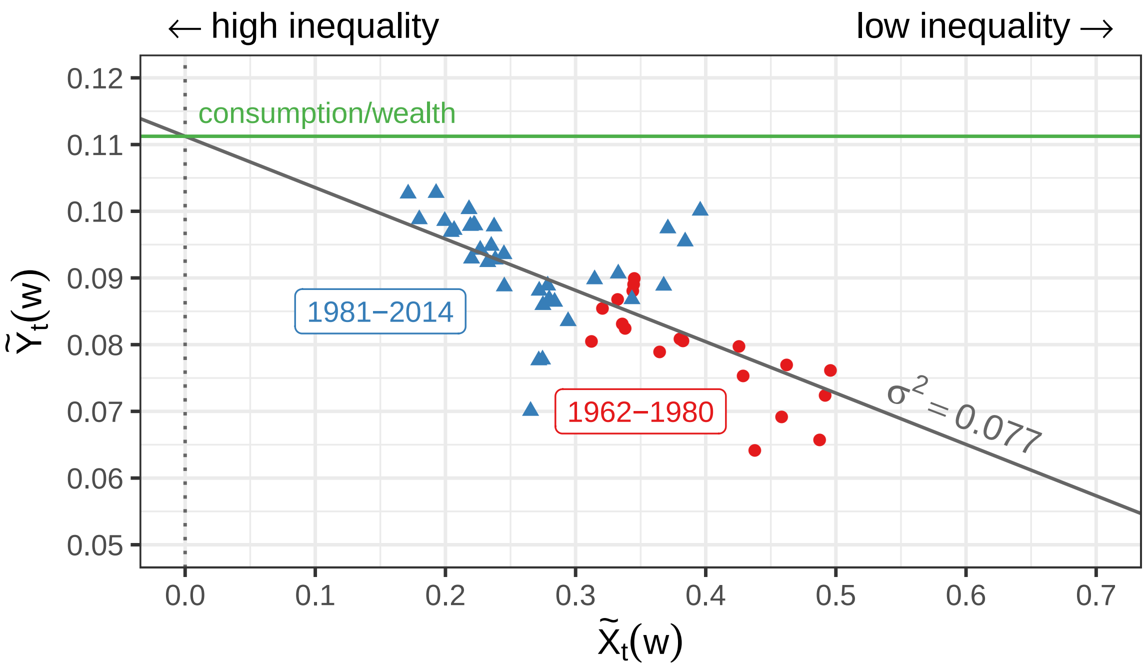
$$\forall t \in \{t_1, \ldots, t_k\} \quad \forall i \in \{1, \ldots, n\} \qquad \tilde{Y}_t(w_i) = \tilde{\mu}(w_i) - \tilde{\sigma}^2 \tilde{X}_t(w_i) + \varepsilon_{it}$$

where $\tilde{\mu}(w_i)$ is a wealth-specific fixed effect, $\tilde{\sigma}^2$ is the opposite of the slope, and ε_{it} captures measurement error.[13] I estimate the CDF and density for the distribution of transformed wealth using kernel density estimators. I simulate the demographic and inheritance effects (see section 4.2 and 4.1), and also estimate resulting distributions using kernel density. For derivatives, both with respect to time and wealth, I run a local polynomial estimator of degree one.[14] For income conditional of wealth, I separate the sample into two periods (1962–1980, and 1981–2014) that constitute the two branches of the U-shaped pattern of wealth. I average income over these two periods so that the model focuses on the long-run mechanisms, rather than short-run dynamics that would introduce noise. I perform the estimation for levels of wealth greater than 50 times average national income, which roughly correspond to the top 1% wealth threshold today. This follows from the fact that the evolution of the top 1% has determined the trajectory of wealth inequality in the United States (see section 2.1) and that the model requires meaningful variation in the distribution to properly identify the effects at hand (cf. assumption (ii)).

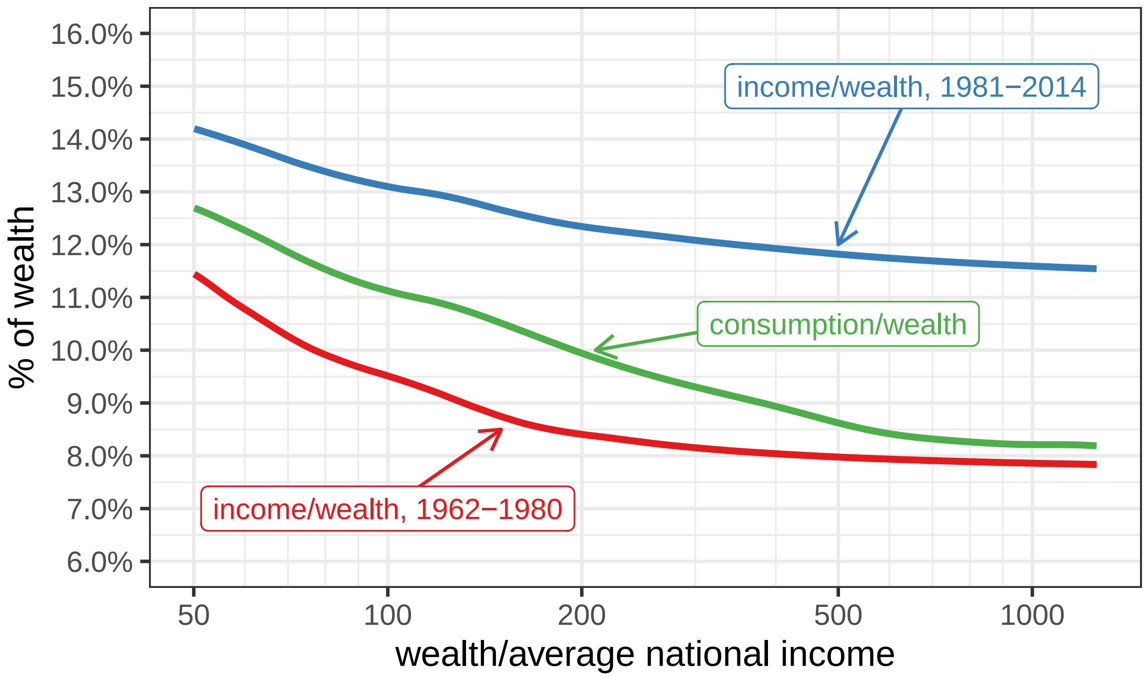
Figure 8 graphically shows the results from the estimation. Panel 8a shows a diagram somewhat similar to figure 7, but with actual data, for a level of wealth corresponding to 100 times average national income. As we see, the data points for the periods 1962–1980 and 1981–2014 are more or less spread alongside the same line, despite the sharp change in the wealth/income ratio between

[13] In fact, measurement error should affect both the dependent and the independent variable, so the standard within estimator for fixed effect regression may give biased results. That being said, I have tested alternative estimators such as orthogonal least squares that account for error on both terms of the equation, and results were virtually identical.

[14] For the term $\frac{\partial}{\partial w}\tilde{f}_t(w)/\tilde{f}_t(w)$, note that it is equal to $\frac{\partial}{\partial w}\log \tilde{f}_t(w)$, so that I directly estimate the derivative of the logarithm of density. Because the logarithm of the density of an exponential distribution is linear, this yields more robust results.



(a) Dynamics of the Wealth Distribution at 100 × average national income



(b) Estimated Consumption/Wealth and Income/Wealth Profiles

Source: Author's computations. *Note:* The unit is the adult (20 or older) individual, and wealth is split equally between members of couples.

Figure 8: Estimation of the Main Model

both periods that impacts the left-hand side of the equation. This suggests that there hasn't been any strong structural changes since 1962 in terms of consumption/wealth profiles. There is a handful of points that stand out: these all correspond to the 1981–1989 period. That can be attributed to the reversal of many dynamics, so that several derivatives change sign at the same, making it harder to estimate them properly. In practice, removing or including these points does not change the results.

The slope is the same for all levels of wealth, and correspond to the variance of consumption/wealth. It is equal to $\sigma^2 = 0.077$. The fixed effects capture the average consumption/wealth ratio: they correspond to the intercept in panel 8a. In panel 8b, I plot that consumption/wealth ratio for all levels of wealth. That profile is a decreasing function of wealth. During the 1962–1980 period, that ratio was consistently higher than income, so that the decrease in wealth inequality was driven by dissavings at the top. Since 1981, however, income has increased, allowing people at the top to maintain the same relative levels of consumption while still accumulating wealth.

Bottom of the Distribution To fit the model, I restricted myself to wealth below 50 times average national income. To account for the distribution of wealth below that level in simulations (see section 6), I assume that the diffusion parameter $\tilde{\sigma}$ is the same for the whole distribution. Under that assumption, we can directly estimate $\tilde{\mu}(w)$ for all levels of wealth by taking the average of $Y_t(w_i) + \tilde{\sigma}^2 \tilde{X}_t(w_i)$ over all time periods.

That approach is an approximation, because it assumes that we can infer wealth mobility throughout the entire distribution based on mobility in the top tail. But in practice it is the simplest and most robust way to match the actual dynamics of wealth at all wealth levels. In particular, it is preferable to the inclusion of low and medium wealth when fitting the model, or to the fitting of a separate complete model for these levels of wealth. Indeed, some quantities for the bottom and middle are harder to estimate (especially the derivative of the density), and the amount of meaningful variation is lower (because the largest changes have happened to the top tail, see section 2.1). In addition, there may be some other, time-varying phenomenons that we doe not properly capture. All of this makes the signal-to-noise ratio less favorable. As a result, if we tried to include that part of the data with the top when fitting the model, we would lower the quality of the fit for the top — which has been driving most of the increase in inequality — and therefore diminish our ability to reproduce the main facts about wealth inequality. If we tried to estimate both the diffusion and the drift by fitting a complete model separately, we would get unstable and problematic results (including negative variances in some cases).

However, this approximation has a very limited impact on the overall results. First, because what matters to the shape of the wealth distribution is not the value of diffusion itself, but the joint effect of both drift and diffusion. By taking the average of $Y_t(w_i) + \tilde{\sigma}_t^2 \tilde{X}_t(w_i)$ to estimate average consumption, I ensure that, taken together, the estimate of drift and diffusion reproduce observed patterns. Second, because what matters the most is to faithfully reproduce the top of the distribution, which has driving the dynamics of inequality (see section 2.1). In practice, the assumptions for the bottom and the middle are here to ensure a roughly stable wealth distribution the bottom.

6 Results

Having estimated the model of wealth accumulation for the United Sates economy, I can now reproduce the observed trajectory of wealth inequality. More importantly, I can also change certain parameters and observe how these changes would affect wealth inequality. I can also make projections of how high wealth inequality is likely to go under current circumstances, and see what could affect that level. I start by studying the past evolution of the wealth distribution in section 6.1, and then the future in section 6.2.

6.1 Past Evolution (1962–2014)

For the past evolution of wealth, we may first check that if we feed the model the actual parameters of the economy, we can reproduce the observed evolution of the wealth distribution. That is, I assume that simulated observations get the same labor income as people with the same rank in the true wealth distribution, and that they the rate of return on their capital income. I also use observed demographic parameters, and the actual estate tax schedule.

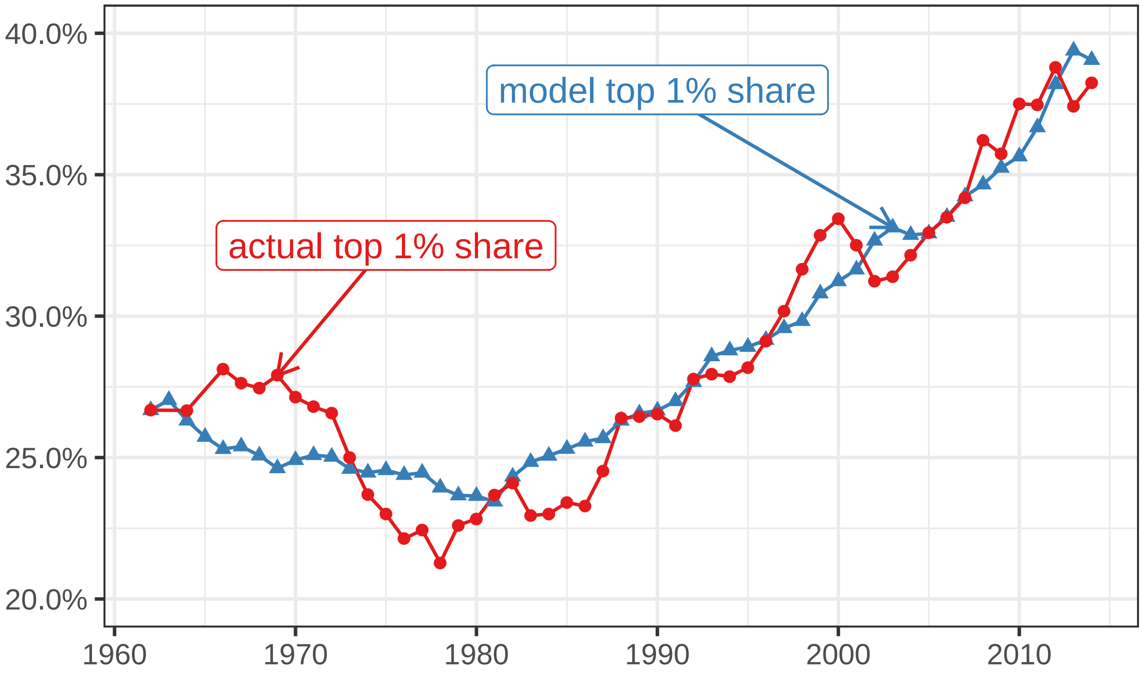
As shown in panel 9a, I match the U-shaped pattern of wealth inequality since 1962. Because the model focuses on long-run dynamics, it does not reproduce the small variations that are primarily driven by short-run changes in asset prices. In the long run, however, the model matches the data well.

Panel 9b further compares the simulated data with the whole distribution for all levels of wealth by looking at the density. The tail is getting increasingly fatter, as expected given the rise in inequality: the model matches that rise, but also reproduces the overall shape of the distribution for lower levels of wealth.

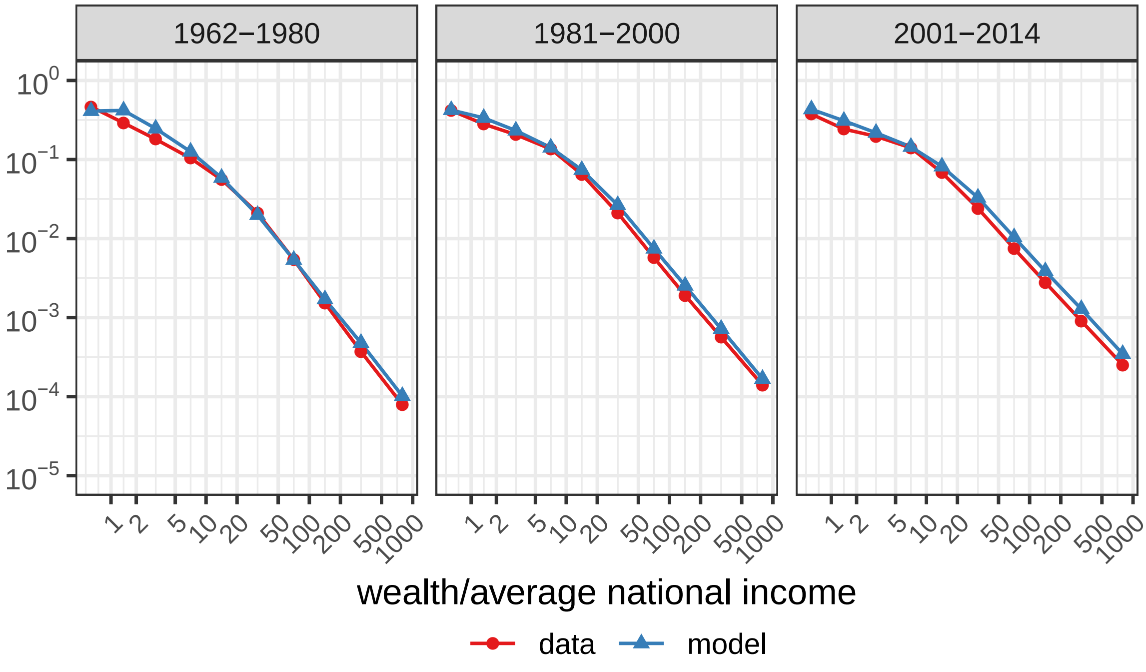
6.1.1 Labor and Capital Income

In figure 10, I estimate what the distribution of wealth would look like today if the distribution of labor income or returns on capital had stayed the same after 1980 as it was over the 1962–1980 period. That is, in panel 10a, I give people with a given rank in the wealth distribution after 1980 the average mean and variance of labor income from people with the same rank over 1962–1980. By construction, this implies that the distribution of labor income is held fixed after 1980. In panel 10b, I do the same for the rates of return on capital (including capital gains).

Both labor income and capital returns have been significant drivers of wealth inequality: taken together, they account for most of the 15 pp. increase in the top 1% wealth share observed since



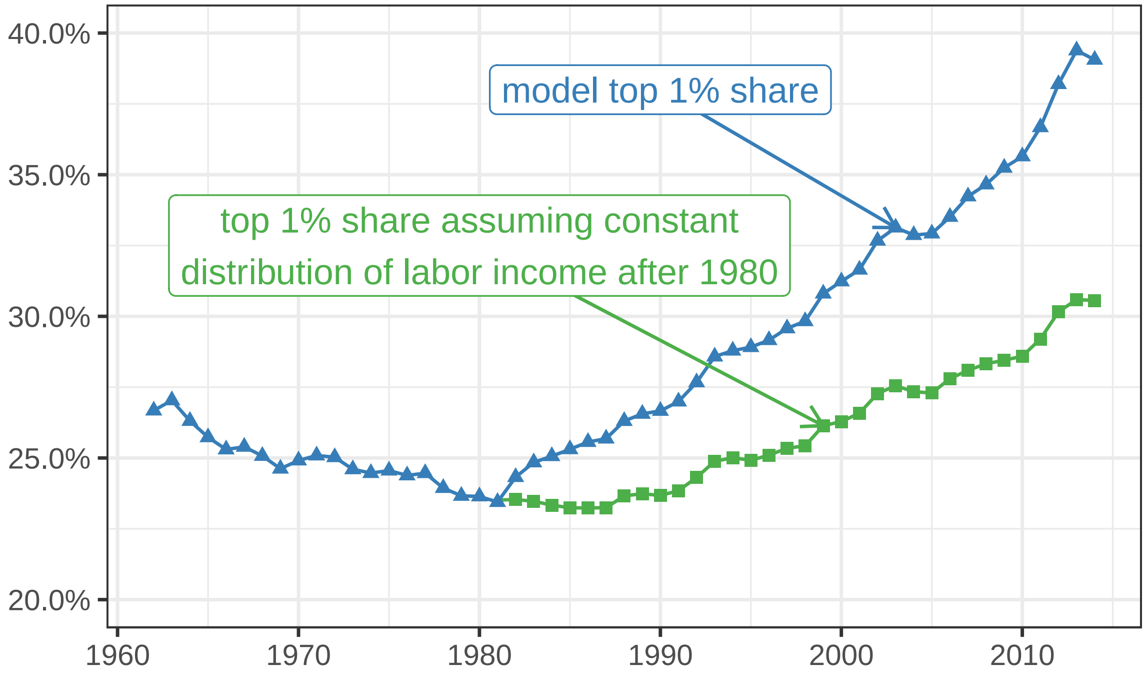
(a) Top 1% Wealth Share: Model vs. Data



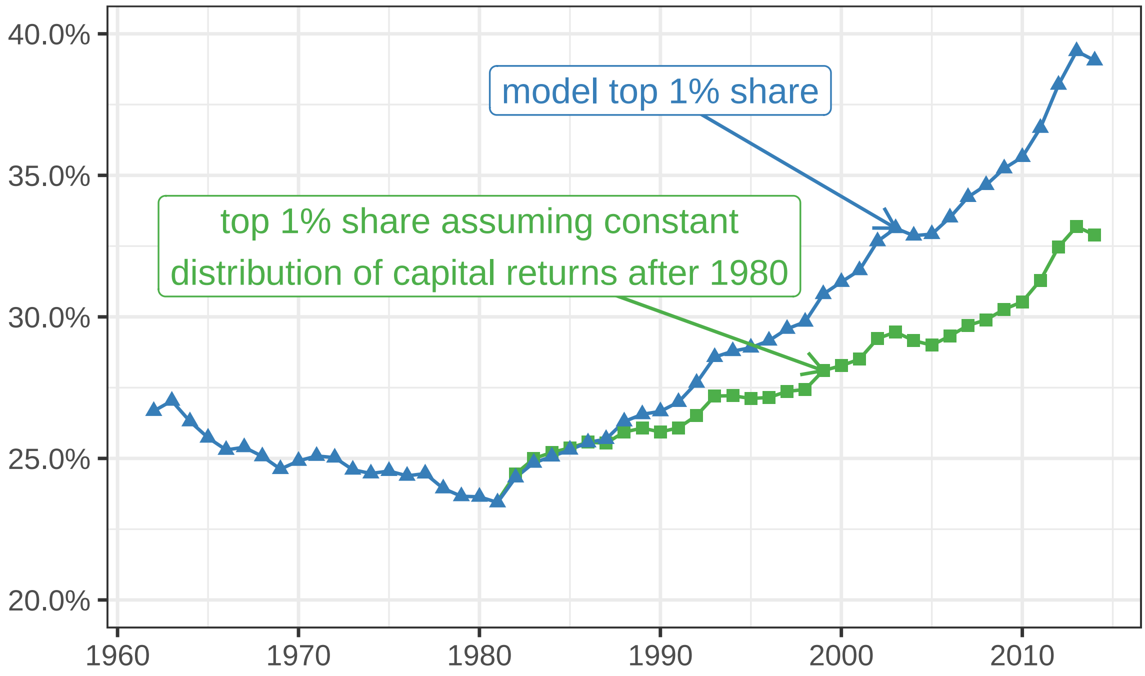
(b) Density of Wealth: Model vs. Data

Source: Author's simulations and DINA data from Piketty, Saez, and Zucman (2018). *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 9: Past Evolution of Wealth: Data vs. Model



(a) Top 1% Wealth Share: Impact of Labor Income



(b) Top 1% Wealth Share: Impact of Capital Rates of Return

Source: Author's simulations. *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 10: Past Evolution of Wealth: The Role of Labor and Capital Income

1980. Most of it can be attributed to increases in mean income conditional on wealth, as the conditional variance of income has not changed much. The role of labor income inequality is somewhat larger, but both factors are major contributors.

6.1.2 Deciphering the Role of $r - g$: Capital Gains and the Growth Rate

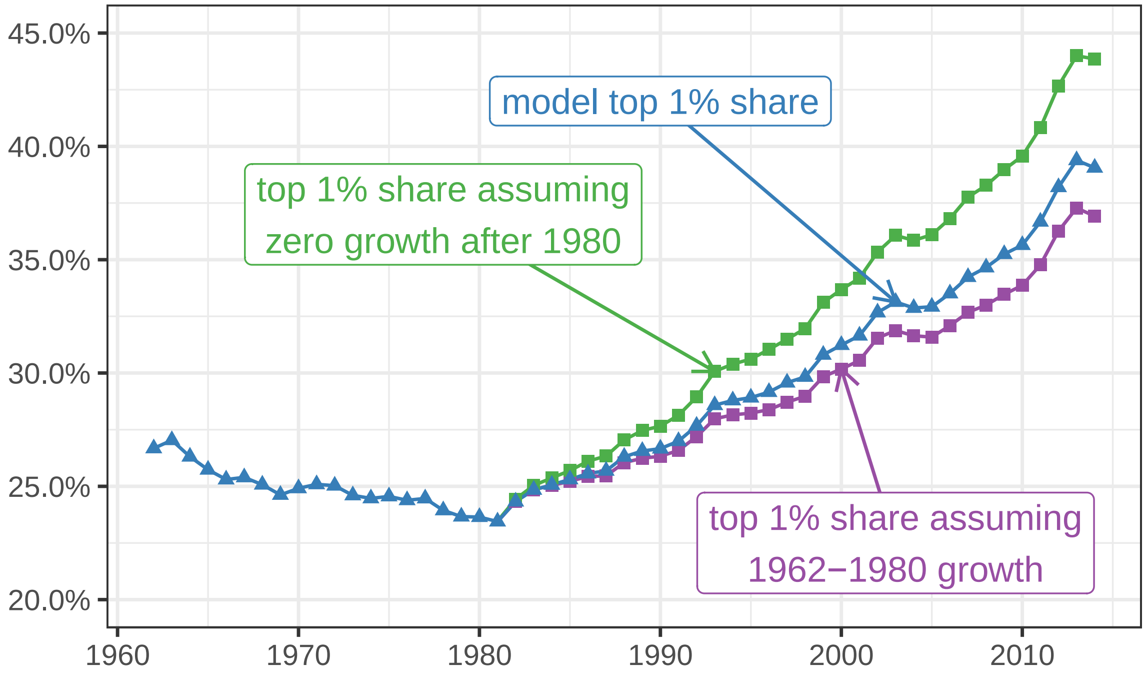
The role played by capital rates if return in figure 10b is directly connected to the impact of the spread between the rate of return on capital and the growth rate ($r - g$) that was popularized by Piketty (2014) (see also Piketty and Zucman, 2015).

In figure 10b, I fixed r but not g. In figure 11a, I do the opposite exercise and fix g but not r. We can see that lower economic growth since 1980 has played some role in increasing wealth inequality, but that this role remains more limited than that of capital returns. Still it implies that the overall impact of increasing $r - g$ has been an important contributor to wealth inequality, on par with the rise of labor income inequality.

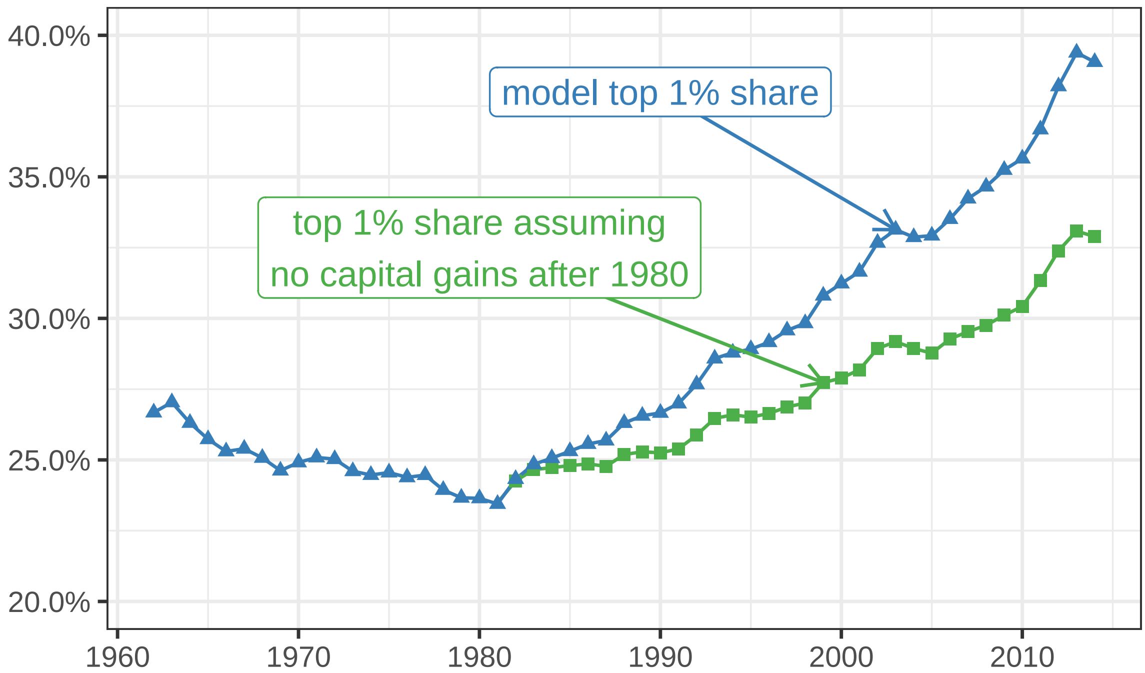
Though there is a twist. The usual story behind $r - g$ emphasizes normal capital returns (i.e. excluding capital gains), but these cannot explain rising wealth inequality. In fact, according to the DINA data, the average rate of return at the top has been somewhat ***lower*** since 1980 than before 1962. It is capital gains that explain most of the increase: as figure 11b shows, the rise of wealth inequality assuming no capital gains after 1980 is essentially the same as that assuming the overall rate of return as the 1962–1980 period.

The crucial role of capital gains is somewhat at odds with many models of wealth accumulation in the long run that tend to focus on normal capital returns: capital gains tend to be treated as short-run phenomenons that can be ignored when it comes to long-term trends. One of the reasons behind this view is that capital gains represent a change in relative prices. And, almost by definition, relative prices should not be changing when the economy is at its steady state, so there cannot be capital gains in the long run. That view is challenged by the fact that capital gains have been a persistent and economically meaningful phenomenon for the United States economy, especially since the 1980. It remains possible that the role of capital gains will eventually disappear, and the period from 1980 to today will become a historical anomaly. But there is an alternative view put forward by Robbins (2018), who shows that in a neoclassical model with imperfect competition, it is possible for capital gains to represent a meaningful fraction of national income at the steady state. Which of these view holds true would have a significant impact on the future evolution of the wealth distribution.

Panel 11a further shows how lower growth can increase wealth inequality. Assuming, at the extreme, that the growth rate fell to zero after 1980, top 1% wealth share would be about 5 pp. higher than it is today.



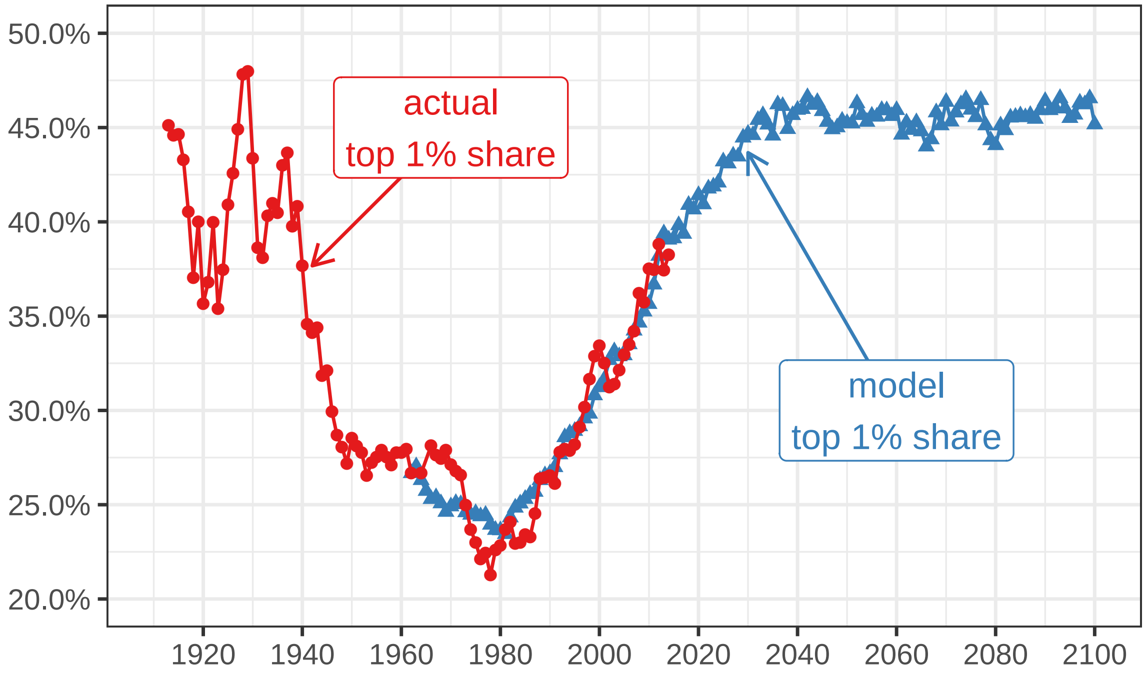
(a) Top 1% Wealth Share: Impact of Growth



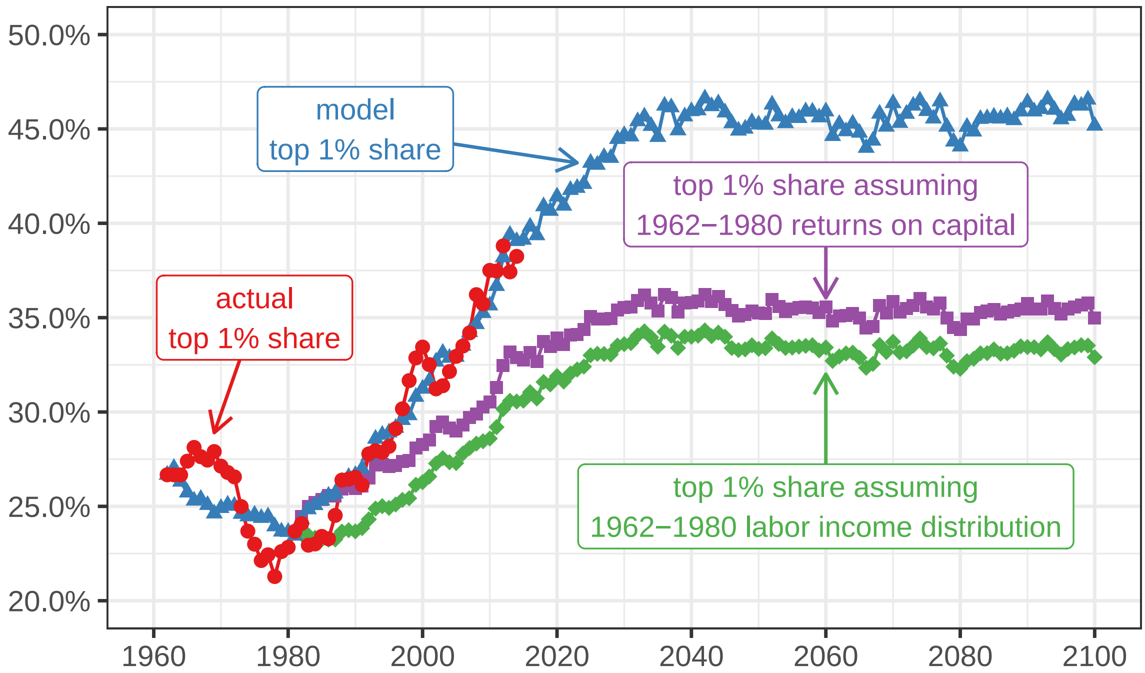
(b) Top 1% Wealth Share: Impact of Capital Gains

Source: Author's simulations. *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 11: Past Evolution of Wealth: The Role of Growth and Capital Gains



(a) Top 1% Wealth Share, 1913–2100, Assuming Current Parameters



(b) Top 1% Wealth Share: Long-Term Impact of Labor and Capital Income

Source: Author's simulations and DINA data from Piketty, Saez, and Zucman (2018). *Note:* The unit is the adult individual (20 or older), and wealth is split equally between members of couples.

Figure 12: Future Evolution of Wealth

6.2 Future Evolution (2015–2100)

We can run the model to get projections of future inequality levels under various scenarios, and determine the steady state of wealth inequality, if any. I stress that these forecasts are always conditional on various parameters (regarding consumption, income, etc.) I do not attempt to endogenize these parameters: the point is to get some idea of how the wealth distribution reacts to them in the long run, and how high inequality can go under their current value.

The first result is presented in figure 12a. Under current parameters, the wealth distribution in the United States would reach its steady state by the 2040s, with a top 1% share around 45%. This would put it at a level similar to that of the early 20th century — or even slightly higher.

The steady state would correspond to a much lower level of inequality, had the distribution of labor or the distribution of capital rates of return stayed at its 1962–1980 level. The level of inequality in the long run would correspond to a top 1% wealth share of 33% and 35%, respectively.

6.3 The Taxation of Wealth

In this section, I use the model of this paper to assess the long run effect of wealth taxes at the top of the distribution. The literature on the topic has grown significantly over the past few years. Recent theoretical contributions have stressed that the long-run elasticity of wealth with respect to the net-of-tax rate is a sufficient statistic for optimal capital taxation (Saez and Stantcheva, 2018; Piketty and Saez, 2013).[15] Unfortunately little is known about the value of that elasticity.

Several empirical papers have used quasi-experimental settings to estimate the ***short-run*** elasticity of wealth with respect to the net-of-tax rate: Seim (2017) in Sweden, Londoño-Vélez and Àcila-Mahecha (2018) in Colombia, Brülhart et al. (2016) in Switzerland, and Jakobsen et al. (2019) in Denmark. With the exception of Switzerland, these elasticities tend to be small. This is consistent with the view that a government trying to raise revenue with a one-off, unexpected wealth tax can indeed choose a very high marginal rate.

But the ability to raise revenue sustainably from a wealth tax depends on the ***long-run*** elasticity. That elasticity is likely to be larger than in the short run. The short-run elasticity only captures tax avoidance or short-run saving responses. But over time, wealth taxes also keep people from accumulating wealth, either through mechanical (lower post-tax rates of return) or behavioral effects (lower savings). This leads to a slow erosion of the tax base. Because it takes a long time to materialize, it is hard to get a clean identification of this effect in the data. As a result, we lack a clear understanding of how the stock of capital would react to wealth tax in the long run.

[15] The famous result of Chamley (1986) and Judd (1985) — tax the optimal tax rate on capital is zero in the long-run — can be interpreted as the result of an implicit assumption that wealth is infinitely elastic. Various contributions have overturned the result by introducing, for example, uncertainty (Aiyagari, 1994), incomplete markets (Farhi, 2010) or heterogenous altruism (Farhi and Werning, 2013).

Recently, two papers have tackled that question. Jakobsen et al. (2019) use their short-run elasticity estimates to calibrate a structural model of savings at the top. They indeed find a higher elasticity in the long-run. Saez and Zucman (2019) consider the problem of taxing the very top of the wealth distribution (billionaires) using data from the Forbes rankings. These two papers provide models that shed different lights on the problem. Jakobsen et al. (2019) model wealth accumulation using an deterministic model of intertemporal choice. This model features standard preferences over a consumption path and a taste for end-of-life wealth (i.e. bequests). They use it to derive analytical expression linking the long-run elasticity of wealth to the short-run elasticity and preference parameters. This model emphasizes the role of behavioral responses on consumption, but it is deterministic so it does not account for the role that mobility plays in shaping the distribution of wealth. This stands in contrast to the model of Saez and Zucman (2019). They focus on billionaire wealth, and therefore assume that the role of consumption is negligible. They consider a simple model in which billionaires are subjected to a given tax rate on their total wealth (not just above a threshold), while everything else remains the same. In that model, the sole determinant of the elasticity of wealth in the long-run has to do with mobility. If wealth mobility is low, then a wealth tax ends up taxing the same people again and again: as their wealth mechanically goes down, so does the tax base. Therefore, the elasticity of taxable wealth is high, and the ability to tax wealth is the long-run is limited. However, if mobility is high, the tax base often gets renewed. Individual people are subjected to the tax during shorter periods of time, with new, previously untaxed wealth entering the tax base on a regular basis: as a result, the elasticity is lower.

I contribute to that literature by providing a simple, practical and transparent method to determine how the tax base reacts to a wealth tax in the steady-state. It connects short-run elasticities with the dynamics of wealth using the dynamic model of this paper to estimate a long-run elasticity. This allows me to incorporate insights from Jakobsen et al. (2019) and Saez and Zucman (2019) into single formula. In the short run, I account for behavioral response on savings and tax avoidance using reduced-form elasticities. Then, I use the model of this paper to compute how these effects accumulate over time to produce long-run responses of the tax base to the wealth tax.

I show that, under very general conditions, the steady-state density of wealth ***with*** a wealth tax is equal to the steady-state density of wealth ***without*** a wealth tax, multiplied by an additional term that only depends on the tax schedule, wealth mobility, and some behavioral elasticities. This makes it easy to simulate how the tax base would eventually react to any given wealth tax.

I start by considering the pure mechanical effect of a wealth tax to present the key result of this section. Then I show how we can account for various behavioral response by using the same result with a modified "effective" tax schedule that is slightly different from the statutory one.

Dynamic Mechanical Effect Absent a wealth tax, assume that the dynamic of wealth follows the SDE:

$$\mathrm{d}w_{it} = a(w_{it})\,\mathrm{d}t + b(w_{it})\,\mathrm{d}B_{it} \tag{10}$$

where $a(w_{it}) \equiv (\nu_t(w_{it}) - \mu_t(w_{it}))$ correspond the average saving by wealth, and $b(w_{it}) \equiv (\tau_t^2(w_{it}) + \sigma_t^2(w_{it}))^{1/2}$ is the standard deviation of savings by wealth. For the rest of this section, I neglect the impact of demographics and inheritance for the sake of tractability. Note that these processes have a limited impact on the long-run dynamics of wealth, so this should not significantly affect the conclusions. We can assess their impact using simulations.

I then introduce a wealth tax with rate α for wealth above the threshold w_0. The dynamic of wealth becomes:

$$\mathrm{d}w_{it} = (a(w_{it}) - \alpha(w_{it} - w_0)_+)\,\mathrm{d}t + b(w_{it})\,\mathrm{d}B_{it} \tag{11}$$

where $(x)_+ = \max\{x, 0\}$. I will further assume that, for $w \geq w_0$, the standard deviation of shocks is proportional to wealth, i.e. $b(w) = bw$. That last assumption is not very restrictive, since it is required at the top for Pareto-shaped tails to arise, in line with the literature and the findings of this paper.

At this stage, I do not assume any behavioral response: yet, in the long-run, the distribution of wealth changes in response to that wealth tax, because it lowers post-tax returns on capital. The following result gives the steady-state distribution of wealth with the tax as a function of the steady-state distribution of wealth without the tax (see appendix **??** for proof).

Theorem (Steady-State Distribution With a Wealth Tax)**.** Assume that, without a wealth tax, the dynamic of wealth follows the equation (10). Introduce a wealth tax with rate α on wealth above w_0, so that wealth now evolves according to (11). Let f_α be the steady-state density of wealth with the tax, and f_0 the steady-state density without the tax. Define:

$$\zeta(w) \equiv \begin{cases} \exp\left\{-\frac{2\alpha}{b^2}\left(\frac{w_0}{w} - 1\right)\right\}\left(\frac{w}{w_0}\right)^{-2\alpha/b^2} & \text{if } w \geq w_0 \\ 1 & \text{if } w < w_0 \end{cases}$$

and $K^{-1} \equiv \int_{-\infty}^{+\infty} \zeta(w) f_0(w)\,\mathrm{d}w$. We have $f_\alpha(w) = K f_0(w)\zeta(w)$.

That result makes it possible to estimate how the tax base would react to a wealth tax in the long-run, effectively by reweighting the steady-state distribution of untaxed wealth using the function ζ. I have considered the effect of a linear tax above an exemption threshold, but the result could be extended to an arbitrary number of brackets with different rates without difficulties. The setting mentions the introduction of a new wealth tax where there previously was none, but we could apply the same result to an increase or a decrease of an existing wealth tax by redefining α as a change in the rate of the wealth tax.

The result emphasizes the role of mobility, as explained by Saez and Zucman (2019). As we can see, the impact on the tax base depends on α/b^2 and not just α. Therefore, doubling the parameter b quadruples the parameter b^2, which implies the same change in the tax base despite a tax rate four times as high. The intuition is the same as in Saez and Zucman (2019): high mobility means that people only gets taxed for a short period of time and that new, previously untaxed wealth keeps entering the tax base. As a result, the tax base does not react too much to wealth taxation. When mobility goes to zero, however, the same wealth from the same people is taxed repeatedly, so that the tax base eventually goes to zero.

Let $B = \int_{w_0}^{+\infty}(w - w_0)f_\alpha(w)\,\mathrm{d}w$ be the steady-state tax base. We can calculate it as follows:

$$B = \frac{\mathbb{E}[(w - w_0)_+\zeta(w)]}{\mathbb{E}[\zeta(w)]}$$

where the expectations should be taken according to the steady-state distribution without tax, i.e. f_0. If we know f_0, then this quantity is directly estimable. Finding that true steady-state density does require some assumptions and additional modelling, as was done in previously paper. The steady-state tax revenue is equal to αB.

Behavioral Response Through Tax Reporting People can react to a wealth tax by hiding some of their wealth, either through tax evasion or tax avoidance. Assume that, in response to a tax α, people only report a fraction $(1-\alpha)^\varepsilon$ of their wealth. The parameter ε is the elasticity of declared wealth to the net-of-tax rate $1-\alpha$. For a small rate $\alpha \ll 1$, people react by approximately hiding a fraction $\alpha\varepsilon$ of their wealth. When $\varepsilon = 0$, people truthfully report all of their wealth. As ε goes to infinity, people start hiding all of their wealth to avoid paying the tax. With tax avoidance, people that own w in wealth pay:

$$\alpha[(1-\alpha)^\varepsilon w - w_0]_+ = \alpha(1-\alpha)^\varepsilon[w - w_0(1-\alpha)^{-\varepsilon}]_+$$

instead of $\alpha(w - w_0)_+$. In effect, this is equivalent to having a wealth tax with a lower rate $\alpha(1-\alpha)^\varepsilon$ and a higher exemption threshold $w_0(1-\alpha)^{-\varepsilon}$. Therefore, the results for the purely mechanical model hold with minimal modifications. It suffices to replace the true tax parameters α and w_0 by their effective counterparts $\alpha(1-\alpha)^\varepsilon$ and $w_0(1-\alpha)^{-\varepsilon}$.

Tax evasion has two effects on the dynamic of wealth. Most importantly, it directly lowers the tax base since people under-report their assets. But as a secondary effect, it increases the post-tax rate of return, allowing people to accumulate more, which grows the tax base in the long-run.

Behavioral Response Through Savings People may also react to a wealth by actually accumulating less wealth. Changes to savings have different implications than tax evasion. Indeed, tax evasion affects both the dynamic of wealth and the tax base. Savings, on the other hand, affect the dynamic of wealth but do not directly reduce the tax base.

Theory provide little constraints regarding how a wealth tax ought to affect saving rates, given the vast number of settings and mechanisms that we could consider. The following reduced-form specification can nonetheless account for the overall effect in a direct and intuitive way. Assume that, in response to a tax rate α on wealth above w_0, people reduce their savings by an amount $(1-(1-\alpha)^\eta)(w-w_0)_+$. The parameter η captures the elasticity of savings with respect to the net-of-tax rate $1-\alpha$. If $\eta = 0$, savings do not respond to the wealth tax. If $\eta > 0$, people start to consume some of their wealth in excess of w_0 rather than pay taxes. At the limit, when α approaches one or η approaches infinity, people immediately consume all wealth above w_0 to avoid paying the wealth tax.[16]

Under those circumstances (and ignoring tax evasion for now), the drift term in the dynamic of wealth become:

$$a(w_{it}) - (\alpha + 1 - (1-\alpha)^\eta)(w_{it} - w_0)_+$$

so the results from the pure mechanical model still hold, except that we need to replace the tax rate α by $\alpha + 1 - (1-\alpha)^\eta$. The behavioral response on savings amplifies the impact of the wealth tax.

Complete Model When combining the behavioral response through savings and tax evasion, it makes sense to assume that savings respond to the effective tax schedule (which accounts for tax evasion) rather than the statutory one. That is, people increase their consumption by an amount $(1-(1-\alpha(1-\alpha)^\varepsilon)^\eta)[w - w_0(1-\alpha)^{-\varepsilon}]$. Therefore, the drift term for the dynamic of wealth is:

$$a(w_{it}) - [\alpha(1-\alpha)^\varepsilon + 1 - (1-\alpha(1-\alpha)^\varepsilon)^\eta][w_{it} - (1-\alpha)^{-\varepsilon} w_0]_+$$

So the results from the mechanical model still apply if we replace the statutory exemption threshold w_0 by $w_0(1-\alpha)^{-\varepsilon}$, and if we replace the statutory tax rate α by $\alpha(1-\alpha)^\varepsilon + 1 - (1-\alpha(1-\alpha)^\varepsilon)^\eta$.

Estimates for Behavioral Elasticities To calibrate ε and η, I rely on the recent empirical literature that exploit various quasi-experimental settings to assess behavioral reactions to a wealth tax.

Several of these papers present bunching evidence (Seim, 2017; Londoño-Vélez and Àcila-Mahecha, 2018; Jakobsen et al., 2019). Bunching provides the cleanest estimates of pure tax avoidance elasticity. Indeed, the true value of wealth in the short run tend to follow unpredictable asset movements, so that it would be very hard for a household to precisely bunch at kink points. Seim (2017) finds an elasticity of 0.5 in Sweden, and Jakobsen et al. (2019) find elasticities that are even lower in Denmark. Londoño-Vélez and Àcila-Mahecha (2018) find a higher estimate (2–3) in Colombia.

As their main identification strategy, Jakobsen et al. (2019) pursue a difference-in-difference approach that exploit various tax reforms. This allows them to compute elasticities that

[16] I will ignore the cases where $\eta < 0$, even though they are a theoretical possibility, because it is problematic to assume in a taxation context that the tax base respond positively to the tax. Moreover, the elasticity has to change sign at some point, otherwise a 100% wealth tax would correspond to infinite savings. However, if true, it would imply that wealth tax rates could be higher.

incorporate dynamic and saving responses over larger time spans. Over an 8-year time frame, they find a sizeable elasticity at the top of about 18 with respect to the net-of-tax rate. The authors argue that most (90%) of it can be attributed to a behavioral effect (as opposed to a mechanical effect). Assuming that the elasticities cumulates multiplicatively over time, this would correspond to a yearly behavioral elasticity of 1.4 for both the saving and tax avoidance response. Seim (2017) also analyze saving responses to the wealth tax, but does not find any.

Brülhart et al. (2016) find a much higher elasticity (23–34) in Switzerland using both between canton variations of the tax rate and within variation in the Bern canton. They also look at bunching evidence, but find much lower effects there.

Note that the tax avoidance elasticity is not a pure structural parameter, but also results from how strongly a wealth tax is enforced. For the baseline calibration, I will consider a limited tax avoidance response ($\varepsilon = 1$), which is around the values found by Seim (2017), Londoño-Vélez and Àcila-Mahecha (2018) and Jakobsen et al. (2019). I will also consider a medium savings response ($\eta = 1$), in line with Jakobsen et al. (2019), but higher than zero as opposed to Seim (2017). Then I consider alternative scenarios with a higher saving response ($\eta = 2$) and a higher tax avoidance response ($\varepsilon = 10$). I could consider even higher tax avoidance responses ($\varepsilon = 20$ or $\varepsilon = 30$), as found by Brülhart et al. (2016), but the interest would be limited. Indeed, with such a severe tax avoidance, the dynamic effects under study become negligible compared to static tax avoidance.

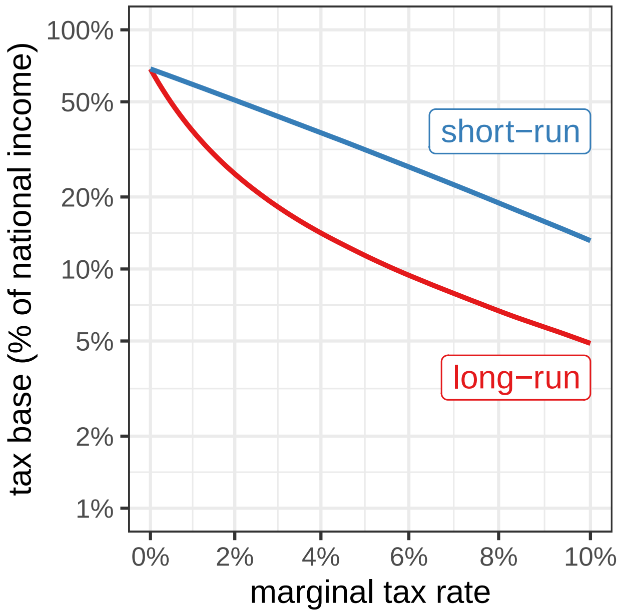
Implications for a Wealth Tax For the different values of ε and η, I simulate how the tax base would react to various marginal tax rates. I consider a linear tax with rate α on wealth above \$50m (in 2014 dollars). The tax applies to equal-split wealth (meaning that the threshold for couples is actually \$100m). I assume that, in the long run, the threshold would rise in line with average income (so that there is a stationary solution), and look at the value of the tax base as a fraction of national income for different values of α. I use the steady-state wealth distribution under current parameters as estimated in section 6.2. Note, however, that results are very similar if we use the last year of data available (the levels would change, but the elasticities would be the same). Therefore, the results of this section do not depend too much on the outcome of the long-run simulations, and one can use the method with the current distribution of wealth as a short-cut. Figure 13 shows the results.

The long-run response of the tax base is naturally stronger than the short-run, which only accounts for tax avoidance. In the baseline specification, a 1% wealth tax decreases the tax base by 50% in the long, which would imply a long-run elasticity around 50. That number may seem high, yet is not out of line with the findings of Jakobsen et al. (2019). At the top of the distribution, they find an elasticity of 18 when looking 8 years after the reform. Using their structural model, this translates into an elasticity of 33 after 30 years in a low return environment, and even higher in a high return environment (which would be closer to the present setting).

However, that elasticity is not constant: it is indeed high for low wealth taxes, but quickly

(a) Baseline ($\eta = 1$, $\varepsilon = 1$)

(b) Pure Mechanical Effect ($\eta = 0$, $\varepsilon = 0$)



(c) High Tax Avoidance ($\eta = 1$, $\varepsilon = 10$)

(d) High Saving Response ($\eta = 2$, $\varepsilon = 1$)

Source: Author's computation. *Note:* Results correspond to the steady-state tax base for a linear wealth tax above $50m in 2014 US dollars. I assume that the tax threshold rises in line with average national income. I assume that the tax applies to equal-split wealth (each members of a couple pay tax on half the wealth of the couple). The parameter ε is the elasticity of tax reporting, and η is the elasticity of the saving response. The short-run response only accounts for tax reporting, while the long-run response also incoporates dynamic mechanical effects et saving effects.

Figure 13: Impact on the Tax Base of a Linear Wealth Tax over $50m

tempers off. As a result, if we were to compute it based on a high 10% wealth tax, the elasticity would be lower (around 21). This nonlinearity results from the dynamic mechanical effects. It implies that one should be careful when extrapolating empirical estimates of the long-run response of the capital stock that are based on relatively small tax changes.

The nonconstant elasticity does impact the tax rate that would maximize revenue in the long run. Indeed, for small marginal tax rates, because we start from a baseline of zero, the wealth tax does raise revenue even though the tax base diminishes quickly. By the time adverse revenue effects arise, the tax base has become less elastic. As consequence, the revenue-maximizing tax rate may in theory be quite high, in several cases north of 10%. In these cases, we nonetheless tend to quickly reach a relatively flat revenue plateau after 10%, so that revenue gains from a wealth tax above 10% are limited.

The pure dynamic mechanical model (figure 13b) is closest to the model Saez and Zucman (2019). Using data from the Forbes 400 rankings and a simplified model, they find an elasticity of 16 (their elasticity is constant by construction). For a 1% wealth tax, I find a somewhat higher elasticity of 27, but for higher wealth tax of 10%, I get a similar value of 15. Note that their estimates are concerned with the extreme top of the distribution (billionaires), whose dynamic of wealth may arguably differ from the rest of the very rich.

I stress that is still significant uncertainty regarding these values. In particular, behavioral elasticities are extrapolated from quasi-experiments based on rather small tax rates, and it is difficult to know what the true reactions would be with more radical policies such as a 10% wealth tax. Note also that I focused on the steady-state. In practice that steady-state may take a very long time to materialize, so it is not necessarily the relevant time horizon. The model nonetheless carries useful insights on the underlying dynamics of the wealth distribution and wealth taxes. Finally, note that I only consider partial equilibrium effects. I do not consider how rates of return or the labor market may react to changes in the capital stock. The estimated elasticities are still useful to calibrate more complex general equilibrium models.

7 Conclusion

In this paper, I have presented a simple model of the wealth distribution that can decompose the impact of labor income, rates of return, growth, savings, demography and inheritance on the wealth distribution. In spite its simplicity, that model can incorporate a realistic modeling of these various factors. I show that this model can be estimated solely using repeated cross-sectional data, and I estimate it using DINA data on income and wealth for the United States since 1962.

I find that, out of the 15 pp. increase in the top 1% wealth share observed since 1980, about 7 pp. can be attributed to rising labor income inequality, 6 pp. to rising returns on wealth, and 2 pp. to lower growth. Importantly, the role played by rising rates of return on wealth can entirely be attributed to capital gains. In the future, and holding constant the present parameters of the economy, the United States economy would reach a steady state with a top 1% wealth share about 45%. I use the model to investigate the impact of progressive wealth taxes on the capital stock and the wealth distribution. I develop a simple a simple formula to characterize how the tax base would react to a wealth tax in the long-run in terms of observable quantities and key behavioral elasticities. I find that the elasticity of wealth with respect to the net-of-tax rate is sizeable, but also nonconstant, so that revenue-maximizing tax rates may be quite high.

These findings are very general in the sense that they apply to any model of the wealth distribution that generates Pareto-shaped fat tails using an accumulation of multiplicative random shocks (to income, consumption, etc.), as is usually the case. Note that all of the counterfactuals coming out of the model assume that everything else remains equal in the economy, which in reality would not always be the case. This is where more tightly specified, fully microfounded models of the wealth distribution can be useful. Such model can endogenously account for the way in which, say, savings might react to a change in the labor income distribution, and include that is its predictions. Yet, it remains true that the findings of this paper have to apply to the more microfounded model. In that sense, both approaches are complementary, and the methodology of this paper is useful to discipline more complex models.

The key insight of this paper — that the Fokker-Planck equation can be used as an empirical tool to identify certain parameters — may be applied to a wide set of problems. For wealth inequality, it could be used to analyze the dynamics of various phenomenons, such as the racial wealth gap. But in theory it could also be applied to any economic situation that involves stochastic growth, such as the income distribution, or the distribution of firms and city sizes.

A Comparison to Synthetic Saving Rates

Saez and Zucman (2016) introduced a different method to decompose the dynamics of inequality. Their approach was extended and used by several authors (Garbinti, Goupille-Lebret, and Piketty, 2017; Berman, Ben-Jacob, and Shapira, 2016) to perform decomposition and simulations similar to section 6 of this paper.

The setting is more or less the same, in that they observe cross-sections of the joint distribution of income and wealth at different points in time, and use this information to analyze the dynamics of wealth. The difference of their approach is that they do not seek to estimate structural parameters that relate to individual behavior. Instead, they construct "synthetic saving rates" that relate the amount of income that accrue to the various percentiles of the wealth distribution to their evolution over time.

Assume zero growth ($g_t = 0$) for simplicity. Synthetic saving rates are defined as follows. Take n brackets of the wealth distribution. Describe the evolution of wealth for each bracket as:

$$\begin{cases} W_{t+1}^{(1)} & = \quad W_t^{(1)} + r_t^{(1)} W_t^{(1)} + Z_t^{(1)} - C_t^{(1)} \\ \quad \vdots & \\ W_{t+1}^{(n)} & = \quad W_t^{(n)} + r_t^{(n)} W_t^{(n)} + Z_t^{(n)} - C_t^{(n)} \end{cases}$$

where $Z_t^{(i)}$, $r_t^{(i)}$ and $C_t^{(i)}$ refers to labor income, rate of return and consumption for bracket (i). Define income $Y_t^{(i)} \equiv Z_t^{(i)} + r_t^{(i)} W_t^{(i)}$. All variables but $C_t^{(i)}$ can be observed in the data, so $C_t^{(i)}$ is estimated as a residual. The ratio $s_t^{(i)} = 1 - C_t^{(i)}/Y_t^{(i)}$ is the synthetic saving rate of bracket (i).

Using various assumptions on how $s_t^{(i)}$ relates to the average income or wealth of the bracket (i), we can perform forecasts and counterfactuals on the evolution of the wealth distribution. Note, however, that the synthetic saving rate is not the average saving rate of the corresponding bracket: it can only be interpreted as such under stringent assumptions (no mobility between brackets and homogeneous behavior within brackets). The synthetic saving rate is a reduced-form parameter that captures mobility, the inequality of savings, their correlation with wealth, demography, intergenerational wealth mobility, etc.

The framework of this paper can shed some light on the underlying mechanics of the synthetic saving rates method. We can use it to explicitly show what synthetic savings rates capture, and how they are capturing it. In doing so I will clarify how it compres the work in this paper.

Consider the following continuous time formulation of the synthetic saving rates. For all t, and for all $0 \leq p < 1$, define the p-th wealth fractile $W_t(p)$. Also write $Z_t(p) = \mathbb{E}[Z_t | W_t = W_t(p)]$ and $r_t(p) = \mathbb{E}[r_t | W_t = W_t(p)]$. I will write:

$$\frac{\partial}{\partial t} W_t(p) = r_t(p) W_t(p) + Z_t(p) - C_t(p) = Y_t(p) - C_t(p) \tag{12}$$

which defines "synthetic consumption" $C_t(p)$. That specification differs from that of this paper in two respects (if we set aside the role of inheritance and demographics). The first issue is the lack of explicit randomness — and, therefore, mobility. If we differentiate equation (12) with respect to p and use the change of variable $p = F_t(w)$, we end up with a special case of the Fokker-Planck equation in which the diffusion term ($\sigma^2(w)$) is equal to zero. Therefore, the synthetic saving rates approach is analogous to a stochastic differential equation with only the drift term: we can view it as a specific, somewhat degenerate case of the more general model used in this paper.

The second issue is the formulation of the saving rate: do we consider savings out of income or out of wealth? Are they a function of the level of wealth, or a function of the rank in the wealth distribution?

Formulation of the Saving Rate There are several justifiable ways of expressing saving rates. Traditionally, the literature writes $C_t(p) = (1 - s(p))Y_t(p)$, so that saving out of income is a function of the rank in the wealth distribution. We could also write $C_t(p) = (1 - s[W_t(p)])Y_t(p)$ to make savings a function of the wealth level rather than the wealth rank. We could also consider savings out of wealth instead of income: $C_t(p) = (1 - s(p))W_t(p)$ or $C_t(p) = (1 - s[W_t(p)])W_t(p)$. That last specification is closest to the one adopted in this paper.

Which formulation to choose depends in part on the intention behind the model. For descriptive purposes, it does not matter. All of them describe the same reality in a different way. Things are different when it comes to running forecasts or counterfactuals. Such an exercise requires setting certain parameters constant, so that the way we parameterize the problem has an impact on the outcome. Utimately, the dynamic of wealth depends entirely on the difference between income and consumption at various points of the wealth distribution: nothing would change if we were to increase everybody's income and consumption by the same amount.

To fix ideas, assume that people at the top of the wealth distribution earn a return of 10% on their wealth, and save 50% of their income. This is identical to saying that they consume 5% of their wealth. Now, increase their income by \$100. Assuming a constant saving rate out of income means that saving increase by \$50. However, assuming a constant saving rate out of wealth means that saving increases by the total amount, i.e. \$100. Therefore, an increase in income has a higher impact with the second specification.

Which specification is better? The first concern is that C_t should accurately describe the behavior of agents. All functional forms for C_t considered here are simplified rule of thumbs that only approximate the true saving behavior, yet it remains important to know which is closest to reality. The macroeconomics and household finance literature would suggest a saving rate out wealth that depends on wealth, i.e. $C_t(p) = s[W_t(p)]W_t(p)$. While there is still considerable disagreement regarding the proper model of household saving (e.g. Browning and Lusardi, 1996), models in which consumption depends on "cash-on-hand" (i.e. wealth plus current income) are commonplace. Such a rule can be microfounded using models of precautionary savings (e.g. Weil, 1993) — especially for the top of the distribution — or models with a preference for wealth (e.g.

Piketty and Zucman, 2014). Models in which agents always consume a given fraction of their current income is harder are harder to justify theoretically, and rarely seen in the literature (with the atypical exception of "hand-to-mouth" households, that have in effect a saving rate of zero). It is also much more common to assume that behavior depends on the absolute level of wealth, rather than the rank in the wealth distribution.

The second concern is that the choosen parameters remain constant over time, which makes it more likely that they will remain so in the future. As we've seen in this paper, we can reproduce the evolution of the wealth distribution since 1962 by assuming constant parameters for consumption out of wealth. However, the average income/wealth ratio at the top has changed between 1962–1980 and 1981–2014. Therefore, the saving rate out of income has changed, even though the saving rate out of wealth has remained stable. This also renders the latter specification preferable.

Mobility The synthetic saving rate parameter is meant to capture both average savings and mobility. We can use the stochastic model of this paper to clarify what that entails. Assume that the dynamic of wealth at the top follows:

$$\mathrm{d}w_{it} = [z(w_{it}) + r(w_{it})w_{it} - \mu(w)w_{it}]\,\mathrm{d}t + \sigma w_{it}\,\mathrm{d}B_{it} \tag{13}$$

using the notations of the paper (in particular, $\mu(w)$ is the average consumption out of wealth, σ^2 is its variance, and we ignore income-induced diffusion $\tau^2(w)$ for simplicity.) Assume that, at time t and for high w, wealth follows a Pareto distribution with coefficient α (i.e. $f_t(w) \propto w^{-\alpha-1}$). The Fokker-Planck equation combine the effect of the drift and the diffusion:

$$\frac{\partial}{\partial t} f_t(w) = \underbrace{-\frac{\partial}{\partial w}\left[(z(w) + w(r(w) - \mu(w)))w^{-\alpha-1}\right]}_{\text{drift}} + \underbrace{\frac{1}{2}\frac{\partial^2}{\partial w^2}\left[\sigma^2 w^2 w^{-\alpha-1}\right]}_{\text{diffusion}}$$

Imagine that, following the synthetic saving rates approach, we estimate a "synthetic" value of consumption $\mu(w)$, noted $\mu^*(w)$, by only taking the drift into account. When observing the evolution of wealth, we still see the role that mobility plays, but we will attribute it to the drift. Thus, rewrite the diffusion term as:

$$\frac{1}{2}\frac{\partial^2}{\partial w^2}\left[\sigma^2 w^2 w^{-\alpha-1}\right] = -\frac{1}{2}\sigma^2(\alpha-1)\frac{\partial}{\partial w}\left[w w^{-\alpha-1}\right]$$

That way, we can include the diffusion term into the drift term such that:

$$\frac{\partial}{\partial t} f_t(w) = -\frac{\partial}{\partial w}\left[\left(y(w) + w\left(r(w) - \mu(w) + \frac{1}{2}\sigma^2(\alpha-1)\right)\right) f_t(w)\right]$$

Hence, the synthetic consumption $\mu(w)$ that we estimate is $\mu^*(w) = \mu(w) - \sigma^2(\alpha-1)/2$. It differs from the true parameter $\mu(w)$ in two respects. First, mobility makes synthetic consumption lower (and therefore makes savings higher) than true average consumption. Second, and perhaps more problematically, this difference depends on the shape of the wealth distribution itself. For example, assume, as found in this paper, that $\sigma^2 \approx 0.08$. Assume that we move from the level of wealth inequality of the 1970s ($(\alpha-1)/2 \approx 0.5$) to today's level ($(\alpha-1)/2 \approx 0.2$). In the 1970s, the synthetic consumption out of wealth $\mu^*(w)$ will be $0.08 \times 0.5 = 4\%$ lower than true consumption, while today it would be only $0.08 \times 0.2 = 1.6\%$ lower. That correspond to a 2.4% increase in synthetic consumption out of wealth, despite no change in the underlying wealth accumulation process. Assuming an income/wealth ratio of 10% at the top, that represents a spurious change of 24 pp. to the synthetic saving rate out of income.

Steady-State The lack of an explicit diffusion mechanism (i.e. mobility) in the synthetic saving rates approach also has an impact on whether, why and how a steady-state distribution can emerge.

In the stochastic model used in this paper, a steady-state power-law distribution arises naturally from scale invariance at the top. That mechanism does not apply in the absence of diffusion. To fix ideas, assume:

$$\frac{\partial}{\partial t} W_t(p) = r(p)W_t(p) + Z(p) - (1 - s(p))W_t(p)$$

The steady state, if any, is given by setting $\frac{\partial}{\partial t}W_\infty(p) = 0$, so that $W_\infty(p) = Z(p)/(1-s(p)-r(p))$. Assuming scale invariance (i.e. $s(p) \equiv s$ and $r(p) \equiv r$) and constant labor income ($Z(p) \equiv Z$), the distribution of wealth can either diverge or collapse onto the single value $Z/(1-s-r)$. To retrieve a smooth steady-state distribution, it is crucial that at least of one of s, r or Z be a smooth function of the rank in the wealth distribution, and not just the level of wealth. Even then, whether a power-law emerges from this type of model will be a direct consequence of the shape of $s(p)$, $r(p)$ or $Z(p)$, not something that the approach explains on its own.

The existence of a non-degenerate steady state requires $1 - s - r > 0$ and $Z > 0$. So it is not possible in this model to have a stationary state in which people at the top of the wealth distribution have no labor income (i.e. a strict separation between workers and capitalists). Wealth at the top is directly proportional to the labor income earned by the same group. This stands in contrast to the steady-state requirements of the stochastic model, which are more general (especially once we introduce demographics as an additional stabilizing force, see Gabaix (2009)). In particular, with a stochastic model it is possible to sustain a stationary steady-state even if labor income plays no role at the top.

B Omitted proofs

B.1 Application of Gyöngy's (1986) Theorem

Recall that wealth at the individual level follows the SDE:

$$\mathrm{d}w_{it} = [y_{it} - c_{it}]\,\mathrm{d}t + [\tau_{it}^2 + \sigma_{it}^2]^{1/2}\,\mathrm{d}B_{it}$$

Consider a small time interval $[t, t + \mathrm{d}t]$. Over that interval, the income process y_{it} has mean $\nu_{it}\,\mathrm{d}t$ and variance $\tau_{it}^2\,\mathrm{d}t$, while the consumption process c_{it} has mean $\mu_{it}\,\mathrm{d}t$ and variance $\sigma_{it}^2\,\mathrm{d}t$. Following Gyöngy's (1986) theorem, we can write:

$$\mathrm{d}w_{it} = [\nu_t(w_{it}) - \mu_t(w_{it})]\,\mathrm{d}t + [\tau_t^2(w_{it}) + \sigma_t^2(w_{it})]^{1/2}\,\mathrm{d}B_{it}$$

where:

$$\nu_t(w) = \mathbb{E}[\nu_{it}|w_{it} = w] \qquad \tau_t^2(w) = \mathbb{E}[\tau_{it}^2|w_{it} = w]$$
$$\mu_t(w) = \mathbb{E}[\mu_{it}|w_{it} = w] \qquad \sigma_t^2(w) = \mathbb{E}[\sigma_{it}^2|w_{it} = w]$$

To simplify notations, consider all expectations conditional on $w_{it} = w$. We can write, somewhat informally, $c_{it} = \mu_{it}\,\mathrm{d}t + \sigma_{it}\,\mathrm{d}B_{it}$. For the drift term, we have directly $\mathbb{E}[c_{it}] = \mathbb{E}[\mu_{it}]\,\mathrm{d}t = \mu_t(w)\,\mathrm{d}t$. For the diffusion term:

$$\begin{aligned}
\mathrm{Var}(c_{it}) &= \mathbb{E}[(c_{it} - \mu_t(w)\,\mathrm{d}t)^2] \\
&= \mathbb{E}[(\mu_{it}\,\mathrm{d}t - \mu_t(w)\,\mathrm{d}t + \sigma_{it}\,\mathrm{d}B_{it})^2] \\
&= \underbrace{\mathbb{E}[(\mu_{it}\,\mathrm{d}t - \mu_t(w)\,\mathrm{d}t)^2]}_{=0 \text{ because } (\mathrm{d}t)^2=0} + \underbrace{\mathbb{E}[\sigma_{it}^2\,\mathrm{d}B_{it}^2]}_{\substack{=\mathbb{E}[\sigma_{it}^2]\,\mathrm{d}t \\ \text{because } \mathrm{d}B_{it}^2 = \mathrm{d}t}} + 2\underbrace{\mathbb{E}[\sigma_{it}(\mu_{it} - \mu_t(w))\,\mathrm{d}B_{it}\,\mathrm{d}t]}_{=0 \text{ because } \mathrm{d}B_{it}\,\mathrm{d}t = 0}
\end{aligned}$$

Therefore, $\mu_t(w)\,\mathrm{d}t = \mathbb{E}[c_{it}]$ and $\sigma_t^2(w)\,\mathrm{d}t = \mathrm{Var}(c_{it})$. Similarly, $\nu_t(w)\,\mathrm{d}t = \mathbb{E}[y_{it}]$ and $\tau_t^2(w)\,\mathrm{d}t = \mathrm{Var}(y_{it})$.

References

Acemoglu, Daron and James Robinson (2015). "The Rise and Decline of General Laws of Capitalism". In: *Journal of Economic Perspectives* 29.1, pp. 3–28. URL: https://www.jstor.org/stable/43194693 (cit. on p. 2).

Aiyagari, S. R. (1994). "Uninsured Idiosyncratic Risk and Aggregate Saving". In: *The Quarterly Journal of Economics* 109.3, pp. 659–684. ISSN: 0033-5533. DOI: 10.2307/2118417. URL: https://academic.oup.com/qje/article-lookup/doi/10.2307/2118417 (cit. on p. 34).

Alstadsæter, Annette et al. (2017). "Accounting for Business Income in Measuring Top Income Shares: Integrated Accrual Approach Using Individual and Firm Data from Norway". URL: http://www.nber.org/papers/w22888 (cit. on p. 18).

Atkinson, A. B. (1971). "The Distribution of Wealth and the Individual Life-Cycle". In: *Oxford Economic Papers* 23.2, pp. 239–254. ISSN: 1537-6591. DOI: 10.1093/oep/gpq012. URL: https://www.jstor.org/stable/2662236?seq=1#page_scan_tab_contents (cit. on p. 6).

Batty, Michael et al. (2019). "Introducing the Distributional Financial Accounts of the United States". URL: https://doi.org/10.17016/FEDS.2019.017 (cit. on p. 5).

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu (2011). "The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents". In: *Econometrica* 79.1, pp. 123–157. DOI: 10.3982/ECTA8416. URL: http://dx.doi.org/10.3982/ECTA8416 (cit. on pp. 2, 6).

Berman, Yonatan, Eshel Ben-Jacob, and Yoash Shapira (2016). "The dynamics of wealth inequality and the effect of income distribution". In: *PLoS ONE* 11.4, pp. 8–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0154196 (cit. on pp. 2, 43).

Böhl, Gregor and Thomas Fischer (2017). "Can taxation predict US top-wealth share dynamics?" (Cit. on p. 2).

Bricker, Jesse, Alice Henriques, and Peter Hansen (2018). "How much has wealth concentration grown in the United States? A re-examination of data from 2001-2013". URL: https://doi.org/10.17016/FEDS.2018.024. (cit. on p. 2).

Browning, Martin and Annamaria Lusardi (1996). "Household Saving: Micro Theories and Micro Facts". In: *Journal of Economic Literature* 34.4, pp. 1797–1855. ISSN: 00220515 (cit. on p. 44).

Brülhart, Marius et al. (2016). "Taxing Wealth: Evidence from Switzerland". In: *NBER Working Paper Series*, p. 37 (cit. on pp. 34, 39).

Carroll, Christopher (1998). "Why Do the Rich Save So Much?" URL: http://www.nber.org/papers/w6549.pdf (cit. on p. 6).

Chamley, Christophe (1986). "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives". In: *Econometrica* 54.3, p. 607. ISSN: 00129682. DOI: 10.2307/1911310 (cit. on p. 34).

Cowell, Frank A (1998). "Inheritance and the Distribution of Wealth" (cit. on pp. 2, 6).

Farhi, Emmanuel (2010). "Capital taxation and ownership when markets are incomplete". In: *Journal of Political Economy* 118.5, pp. 908–948. ISSN: 00223808. DOI: 10.1086/657996 (cit. on p. 34).

Farhi, Emmanuel and Iván Werning (2013). "Estate taxation with altruism heterogeneity". In: *American Economic Review* 103.3, pp. 489–495. ISSN: 00028282. DOI: 10.1257/aer.103.3.489 (cit. on p. 34).

Favilukis, Jack (2013). "Inequality, stock market participation, and the equity premium". In: *Journal of Financial Economics* 107.3, pp. 740–759. ISSN: 0304405X. DOI: 10.1016/j.jfineco.2012.10.008. URL: http://dx.doi.org/10.1016/j.jfineco.2012.10.008 (cit. on p. 2).

Fritsch, F. N. and R. E. Carlson (1980). "Monotone Piecewise Cubic Interpolation". In: *SIAM Journal on Numerical Analysis* 17.2, pp. 238–246. URL: https://doi.org/10.1137/0717021 (cit. on p. 12).

Gabaix, Xavier (1999). "Zipf's Law for Cities: an Explanation". In: *Quarterly Journal of Economics* 114.August, pp. 739–767 (cit. on p. 24).

– (2009). "Power Laws in Economics and Finance". In: *Annual Review of Economics* 1.1, pp. 255–293. ISSN: 1941-1383. DOI: 10.1146/annurev.economics.050708.142940 (cit. on pp. 3, 9, 22, 24, 46).

Gabaix, Xavier et al. (2016). "The Dynamics of Inequality". In: *Econometrica* 84.6, pp. 2071–2111. ISSN: 1098-6596. DOI: 10.3386/w21363 (cit. on pp. 2, 4).

Gapminder (2019). *Children per women since 1800*. URL: https://www.gapminder.org/news/children-per-women-since-1800-in-gapminder-world/.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty (2017). "Accounting for Wealth Inequality Dynamics: Methods, Estimates and Simulations for France (1800-2014)" (cit. on p. 43).

– (2018). "Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)". In: *Journal of Public Economics* 162.June, pp. 63–77. URL: http://piketty.pse.ens.fr/filles/GGP2016DINA.pdf (cit. on p. 2).

Góes, Carlos (2016). "Testing Piketty's Hypothesis on the Drivers of Income Inequality: Evidence from Panel VARs with Heterogeneous Dynamics". In: *IMF Working Papers* 16.160, p. 1. ISSN: 1018-5941. DOI: 10.5089/9781475523249.001 (cit. on p. 2).

Gyöngy, I (1986). "Mimicking the one-dimensional marginal distributions of processes having an Ito differential". In: *Probability Theory and Related Fields* 71.4, pp. 501–516. URL: https://link.springer.com/article/10.1007/BF00699039 (cit. on pp. 4, 10, 11, 47).

Haines, Michael R. (1998). "Estimated life tables for the united states, 1850–1910". In: *Historical Methods* 31.4, pp. 149–169. ISSN: 19401906. DOI: 10.1080/01615449809601197 (cit. on p. 13).

Hubmer, Joachim and Anthony A Smith (2018). "A Comprehensive Quantitative Theory of the U.S. Wealth Distribution" (cit. on p. 2).

Human Fertility Collection (2019). Max Planck Institute for Demographic Research in Rostock, Germany and Vienna Institute of Demography in Vienna, Austria. URL: https://www.fertilitydata.org/.

Human Fertility Database (2019). Max Planck Institute for Demographic Research in Rostock, Germany and Vienna Institute of Demography in Vienna, Austria. URL: https://www.humanfertility.org/.

Human Life Table Database (2019). Max Planck Institute for Demographic Research in Rostock, Germany, Department of Demography at University of California at Berkeley, USA and Institut d'études démographiques (INED) in Paris, France. URL: https://www.lifetable.de/.

Human Mortality Database (2019). Max Planck Institute for Demographic Research in Rostock, Germany, Department of Demography at University of California at Berkeley, USA. URL: https://www.mortality.org.

Jakobsen, Katrine et al. (2019). "Wealth Taxation and Wealth Accumulation: Theory and Evidence from Denmark". In: *Quarterly Journal of Economics*. DOI: 10.2139/ssrn.3138451. URL: http://gabriel-zucman.eu/files/JJKZ2019.pdf (cit. on pp. 4, 34, 35, 38, 39).

Judd, Kenneth L. (1985). "Redistributive taxation in a simple perfect foresight model". In: *Journal of Public Economics* 28.1, pp. 59–83. ISSN: 00472727. DOI: 10.1016/0047-2727(85)90020-9 (cit. on p. 34).

Kakar, Venoo, Gerald Eric Jr. Daniels, and Olga Petrovska (2019). "Does Student Loan Debt Contribute to Racial Wealth Gaps? A Decomposition Analysis". In: *The Journal of Consumer A airs*, pp. 1–28. DOI: 10.1111/joca.12271. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joca.12271 (cit. on p. 24).

Kesten, Harry (1973). "Random Difference Equations and Renewal Theory for Products of Random Matrices". In: *Acta Mathematica* 131.1, pp. 207–248. DOI: 10.1007/BF02392040 (cit. on pp. 3, 6, 9).

Koenker, Roger (2019). *quantreg: Quantile Regression*. R package version 5.40. URL: https://CRAN.R-project.org/package=quantreg.

Kopczuk, Wojciech (2015). "What Do We Know about the Evolution of Top Wealth Shares in the United States?" In: *Journal of Economic Perspectives* 29.1, pp. 47–66. ISSN: 0895-3309. DOI: 10.1257/jep.29.1.47 (cit. on pp. 2, 5).

Kopczuk, Wojciech and Emmanuel Saez (2004). "Top Wealth Shares in the United States, 1916-2000: Evidence from Estate Tax Returns". In: *National Tax Journal* 57.2, Part 2, pp. 445–487. ISSN: 00280283. DOI: 10.17310/ntj.2004.2s.05 (cit. on p. 5).

Laitner, John (1979). "Household Bequest Behaviour and the National Distribution of Wealth". In: *The Review of Economic Studies* 46.3, pp. 467–483. ISSN: 00346527. DOI: 10.2307/2297014 (cit. on p. 2).

Londoño-Vélez, Juliana and Javier Àcila-Mahecha (2018). "Can Wealth Taxation Work in Developing Countries? Quasi-Experimental Evidence from Colombia" (cit. on pp. 34, 38, 39).

Lund, Steven P, Joseph B Hubbard, and Michael Halter (2014). "Nonparametric Estimates of Drift and Diffusion Profiles via Fokker–Planck Algebra". In: *The Journal of Physical Chemistry B* 118, pp. 12743–12749. DOI: 10.1145/1520340.1520741 (cit. on p. 21).

Menchik, Paul L (1980). "Primogeniture, Equal Sharing, and the U.S. Distribution of Wealth". In: *The Quarterly Journal of Economics* 94.2, pp. 299–316. URL: https://www.jstor.org/stable/1884542 (cit. on p. 14).

Piketty, Thomas (2014). *Capital in the Twenty First Century*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press. ISBN: 978-0674430006 (cit. on pp. 2, 23, 31).

Piketty, Thomas and Emmanuel Saez (2013). "A Theory of Optimal Inheritance Taxation". In: *Econometrica* 81.5, pp. 1851–1886. ISSN: 0012-9682. DOI: 10.3982/ECTA10712. URL: http://doi.wiley.com/10.3982/ECTA10712 (cit. on p. 34).

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). "Distributional National Accounts: Methods and Estimates for the United States". In: *Quarterly Journal of Economics* 133.May, pp. 553–609. DOI: 10.1093/qje/qjx043.Advance (cit. on pp. 4, 7–9, 16, 18, 19, 29, 33).

Piketty, Thomas and Gabriel Zucman (2014). "Capital is Back: Wealth-Income Rations in Rich Countries 1700–2010". In: *Quarterly Journal of Economics* 129.3, pp. 1255–1310. ISSN: 0033-5533. DOI: 10.1093/qje/qju018.Advance (cit. on pp. 2, 5, 6, 44).

– (2015). "Wealth and Inheritance in the Long Run". In: *Handbook of Income Distribution* 2, pp. 1303–1368. ISSN: 15740056. DOI: 10.1016/B978-0-444-59429-7.00016-9 (cit. on pp. 23, 31).

Robbins, Jacob A (2018). "Capital Gains and the Distribution of Income on the United States" (cit. on pp. 18, 31).

Ruggles, Steven et al. (2019). *IPUMS USA: Version 9.0 [dataset]*. Minneapolis (cit. on p. 13).

Saez, Emmanuel and Stefanie Stantcheva (2018). "A simpler theory of optimal capital taxation". In: *Journal of Public Economics* 162.September 2017, pp. 120–142. ISSN: 00472727. DOI: 10.1016/j.jpubeco.2017.10.004. URL: https://doi.org/10.1016/j.jpubeco.2017.10.004 (cit. on pp. 4, 34).

Saez, Emmanuel and Gabriel Zucman (2016). "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data". In: *Quarterly Journal of Economics* 131.May, pp. 519–578. DOI: 10.1093/qje/qjw004.Advance (cit. on pp. 2, 5, 6, 18, 43).

– (2018). "Creating Homogeneous Synthetic Individual Tax Files for Distributional Analysis". URL: https://www.irs.gov/pub/irs-soi/18rpsynthindtaxfiles.pdf (cit. on p. 16).

– (2019). "Progressive Wealth Taxation". URL: https://www.brookings.edu/wp-content/uploads/2019/09/Saez-Zucman_conference-draft.pdf (cit. on pp. 4, 35–37, 41).

Saichev, Alexander, Didier Sornette, and Yannick Malevergne (2010). *Theory of Zipf's Law and Beyond*. Vol. 632. Lecture Notes in Economics and Mathematical Systems, p. 170. ISBN: 9783642029455. DOI: 10.1007/978-3-642-02946-2{_}1 (cit. on p. 24).

Schepsmeier, Ulf et al. (2018). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.1.8. URL: https://CRAN.R-project.org/package=VineCopula.

Schularick, Moritz, Moritz Kuhn, and Ulrike I. Steins (2018). "Income and Wealth Inequality in America, 1949-2016". Minneapolis. URL: https://doi.org/10.21034/iwp.9 (cit. on p. 5).

Seim, David (2017). "Behavioral responses to wealth taxes: Evidence from Sweden". In: *American Economic Journal: Economic Policy* 9.4, pp. 395–421. ISSN: 1945774X. DOI: 10.1257/pol.20150290 (cit. on pp. 34, 38, 39).

Steinbaum, Marshall (2019). "Student Debt and Racial Wealth Inequality" (cit. on p. 24).

Stiglitz, Joseph E (1969). "Distribution of Income and Wealth Among Individuals". In: *Econometrica* 37.3, pp. 382–397. URL: https://www.jstor.org/stable/1912788 (cit. on p. 2).

Thompson, Jeffrey P. and Gustavo A. Suarez (2015). "Updating the Racial Wealth Gap". Washington. URL: https://doi.org/10.17016/FEDS.2015.076r1 (cit. on p. 24).

United Nations (2017). *World Population Prospects*. Tech. rep. URL: https://population.un.org/wpp/Publications/ (cit. on p. 13).

United States Census Bureau (n.d.). *National Intercensal Tables: 1900-1990*. URL: https://www.census.gov/content/census/en/data/tables/time-series/demo/popest/pre-1980-national.html (cit. on p. 13).

Vaughan, R. N. (1979). "Class Behaviour and the Distribution of Wealth". In: *The Review of Economic Studies* 46.3, pp. 447–465. ISSN: 00346527. DOI: 10.2307/2297013 (cit. on p. 2).

Weil, Philippe (1993). "Precautionary Savings and the Permanent Income Hypothesis". In: *Review of Economic Studies* 60.2, pp. 367–383. URL: https://academic.oup.com/restud/article-abstract/60/2/367/1574342 (cit. on p. 44).

Wold, H. O. A. and P. Whittle (1957). "A Model Explaining the Pareto Distribution of Wealth". In: *Econometrica* 25.4, pp. 591–595. DOI: 10.2307/1905385 (cit. on p. 2).

Wolff, Edward N (2010). "Recent Trends in Household Wealth in the United States: Rising Debt and the Middle-Class Squeeze". URL: https://ssrn.com/abstract=1585409%20or%20http://dx.doi.org/10.2139/ssrn.1585409 (cit. on p. 6).