Refinancing Frictions, Mortgage Pricing and Redistribution*

David Berger† Konstantin Milbradt‡ Fabrice Tourre§
Joseph Vavra¶

April 2023

Abstract

There are large cross-sectional differences in how often US borrowers refinance mortgages. In this paper, we develop an equilibrium mortgage pricing model that allows us to explore the consequences of this heterogeneity. We show that equilibrium forces imply important cross-subsidies from borrowers who rarely refinance to those who refinance often. Mortgage reforms can potentially reduce these regressive cross-subsidies, but the equilibrium effects of these reforms can also have important distributional consequences. For example, many policies that lead to more frequent refinancing also increase equilibrium mortgage rates and thus reduce residential mortgage credit access for a large number of borrowers.

*We would like to thank Morris Davis, David Zhang (both discussants) and Fernando Alvarez for helpful discussions and the seminar participants at USC, UIUC, Northwestern University, UCLA, Copenhagen Business School, UIC, Princeton, Baruch College, McGill, Cleveland Fed, University of Chicago Booth, Society of Economic Dynamics (2022), Chicago Fed Housing and Racial Bias Workshop (2022), Rio FGV (2022), TAU (2022), NYU Stern, and BU Finance for helpful discussions and feedback. Fabrice Tourre gratefully acknowledges financial support from the Danish Finance Institute and the Center for Financial Frictions (FRIC) (grant no. DNRF-102).

†Duke University and NBER; david.berger@duke.edu
‡Northwestern University and NBER; milbradt@northwestern.edu
§Copenhagen Business School; ft.fi@cbs.dk
¶University of Chicago and NBER; joseph.vavra@chicagobooth.edu

1 Introduction

There are large cross-sectional differences in how often US borrowers refinance their fixed-rate mortgages. Some "fast" borrowers refinance frequently. Other "slow" borrowers do not refinance despite substantial financial incentives.[1] In this paper, we argue that this heterogeneity has important *equilibrium* implications. We develop and characterize an equilibrium model of the mortgage market with persistent borrower heterogeneity, estimate it using US mortgage micro data, and show that heterogeneous refinancing leads to equilibrium forces that amplify inequality.

Institutional features of US mortgage markets limit lenders' ability to price-discriminate, so borrowers with very different refinancing propensities face the same mortgage rates at origination. We show that this pooling equilibrium leads to substantial cross-subsidies from slow to fast borrowers: slow borrowers pay higher rates and fast borrowers lower rates at origination than if lenders were to price-discriminate.

These equilibrium forces are also important for evaluating alternative mortgage market designs and policies. Since heterogeneous refinancing leads to substantial inequality, it is natural to think that policies leading to more frequent refinancing would improve borrower welfare and reduce inequality. However, we show that the same equilibrium forces that play an important role in the current market also matter for evaluating the distributional consequences of various policy counterfactuals. For example, "automatically refinancing" mortgages eliminate refinancing disparities across borrowers but would also lead lenders to charge higher rates on newly originated mortgages and thus reduce mortgage credit access for a large number of borrowers. It is important to account for these equilibrium effects in addition to the more commonly studied direct effects of policy reforms.

While our insight—the fact that the consequences of mortgage market design depend on equilibrium effects—is not new,[2] systematic analysis of effects on inequality in equilibrium has been limited by the complexity of equilibrium environments with heterogeneity. Beyond our specific mortgage application, an important contribution of our paper is thus the development of a tractable framework that can be used to study equilibrium environments featuring ex ante household heterogeneity.

[1] See Keys, Pope, and Pope (2016) and Andersen et al. (2020) for evidence of low refinancing propensities on average and Gerardi, Willen, and Zhang (2020) and Zhang (2022) for evidence of cross-borrower heterogeneity.

[2] See, e.g., Campbell (2006).

We develop this framework in three key steps. In the first step, we characterize in partial equilibrium the optimal refinancing decisions of borrowers facing the two main frictions identified by the past literature (e.g., Andersen et al. (2020)). Specifically, we allow both for "inattention" or other nonmonetary frictions (which generate time-dependent inaction) and for fixed monetary costs of refinancing (which generate state-dependent inaction) and solve for the optimal behavior of borrowers.

In the second step, we embed this household refinancing problem into an equilibrium model of the mortgage market under the assumption that borrowers are ex-ante identical. We assume that risk-neutral competitive investors purchase mortgage-backed securities (MBS), which pool together monthly payments made by borrowers (net of any intermediation fees), and we characterize the relative role of different frictions for new-issue MBS prices and resulting mortgage rates in equilibrium. We show that monetary fixed costs have small effects on equilibrium mortgage pricing since these costs primarily reduce refinancing for borrowers with small "rate gaps" (the difference between the coupon and current mortgage market rate), and closing small gaps via refinancing barely affects lender profits. In contrast, inattention has large effects on mortgage rates since it reduces refinancing even for borrowers with large gaps and this substantially changes lender profits and pricing.

The third step introduces heterogeneous refinancing frictions across borrowers into this equilibrium environment, which allows us to explore the redistributive effects of various mortgage market interventions. Consistent with institutional features of the US agency MBS market, we focus primarily on a "pooling" equilibrium in which lenders do not price-discriminate based on borrowers' refinancing speed. We make two simplifying assumptions to increase tractability and approximate the cross-sectional distribution over households' coupons and attention rates, which would otherwise arise as an additional state variable in the investors' pricing problem. First, we assume that fixed costs are not paid up front but are instead capitalized into a higher interest rate for the new loan. This is a strong assumption, but it broadly aligns with actual US mortgage markets and greatly simplifies borrower decisions.[3] Second, we assume that investors exhibit a simple form of bounded rationality: they value mortgages based on the average distribution of attention among those refinancing at

[3]More than 80% of origination costs in the US are rolled into higher rates rather than paid up front (Zhang, 2022). Since up-front fixed costs have small pricing effects in our model, modeling the remaining 20% would substantially complicate the analysis but have little quantitative impact.

the current mortgage rate rather than using the entire history of rates to infer the current attention distribution.[4] These two assumptions simplify the pooling equilibria substantially, allowing us to then characterize sufficient conditions for existence and many other important properties. For example, in equilibrium, borrower heterogeneity affects mortgage pricing through a simple covariance adjustment term.

We next turn to the model's quantitative results. We estimate the cross-sectional distribution of borrower attention using a monthly borrower-level panel of mortgages from 2005 to 2017 and explore the implications of this heterogeneity in our equilibrium model. We start by testing whether the equilibrium outcomes implied by our model match corresponding time series in the data. We take US treasury yields from 2005 to 2017 together with estimates of intermediation costs from the literature and calculate the equilibrium mortgage rates and refinancing patterns implied by the model. These align well with the data, giving us confidence in the model's implications.

We then explore the quantitative implications of heterogeneous refinancing propensities across borrowers. For a given mortgage environment, our model lets us measure both: 1) ex-post coupon inequality 2) ex-ante cross-subsidies from charging identical rates to heterogeneous borrowers. We can then measure how both inequality and cross-subsidies change in response to various policy changes.

Consistent with Gerardi, Willen, and Zhang (2020), we estimate substantial borrower heterogeneity and resulting ex-post coupon inequality in the current US mortgage market. Although fast and slow borrowers face the same rates at origination, the fastest borrowers refinance more frequently over time and so ultimately pay coupons which are on average 94 bps below the slowest borrowers. While interesting, this simply measures the ex-post inequality realized in the current rate environment and so does not require an equilibrium model. Our other results rely crucially on our model's counterfactual equilibrium analysis.

First, we can compute the mortgage coupons that fast and slow borrowers would pay in a counterfactual "separating" equilibrium.[5] We find that the fastest borrowers pay 127 bps higher coupons on average in the separating equilibrium than they do in

[4] As usual, the quantitative importance of this assumption cannot be fully evaluated without solving the true (intractable) pricing problem. However, we provide some some evidence that this simplifying assumption likely has little quantitative effect on our conclusions.

[5] The relevance of this counterfactual mortgage market equilibrium depends on the extent to which borrower attention is observable and thus potentially priced ex-ante and not just ex-post. Evidence in Gerardi, Willen, and Zhang (2020) as well as our own analysis in the online appendix suggest that ex-ante observable heterogeneity is indeed substantial.

the pooling equilibrium. This 127 bps change in coupons is a measure of the cross-subsidies received by fast borrowers from slow borrowers through pooling.[6] This cross-subsidy is large both relative to the difference in ex-post average coupon paid by fast vs. slow borrowers in the pooling equilibrium, and relative to the average coupon (391 bps) paid by the fastest borrowers in that equilibrium.

We then explore the equilibrium effects of a frequently discussed alternative contract design aimed at reducing mortgage inequality: the "automatically refinancing" mortgage. This contract refinances automatically with no active borrower intervention when rates decline. This effectively makes all borrowers infinitely fast and so eliminates cross-subsidies and inequality across borrower types, but how much does this contract actually benefit formerly slow borrowers? Automatically refinancing mortgages lead to much more refinancing for slow borrowers but also lead to an increase in mortgage rates at origination of about 91 bps that offsets some of this benefit. In equilibrium, automatically refinancing mortgages yield individual time-paths of mortgage coupons that decline more rapidly but that start from higher initial values. Over the loan life, these mortgages reduce average coupons by 78 bps for the slowest borrowers, but this decrease is substantially smaller than the 136 bps reduction that would arise without the equilibrium rate increase at origination. This rate increase is also likely to have important implications for access to housing markets: higher rates may force households that are at debt-to-income (DTI) limits to downsize their purchases or exclude these households from the housing market entirely. Borrowers benefit from the more frequent refinancing induced by automatic refinancing only if they are able to afford a mortgage at the initial higher rate in the first place.[7]

In addition to evaluating alternative mortgage contracts, we show that other refinancing related trends can have important equilibrium effects. For example, our data suggest that the rise of fintech and other nonbank lenders has spurred more frequent refinancing. Loans originated by nonbank lenders have 100 bps greater effective refinancing attention than loans originated by banks.[8] If these same cross-sectional

[6]The slowest borrowers pay 63 bps lower rates in the separating equilibrium than in the pooling. The magnitude of the rate subsidy and penalty is not symmetric because we estimate that the slowest group has many more borrowers than the fastest group.

[7]Our model does not analyze initial home purchases and instead focuses on cross-subsidies across borrowers from refinancing, but a simple back-of-the-envelope calculation suggests that the increase in interest rates arising from a move to automatically refinancing mortgages might force approximately 16% of borrowers to select smaller homes requiring a smaller initial mortgage balance.

[8]The patterns that we identify are cross-sectional correlations and do not necessarily isolate

patterns were to hold with a move from a mortgage market dominated by banks to one dominated by nonbanks, it would have important equilibrium implications: a 100 bps increase in attention leads equilibrium mortgage rates to rise by 35 bps.

While our paper focuses on mortgage markets, our modeling framework lends itself to the analysis of many other settings and to future quantitative research. These environments all share the following features: on one side of the market, ex ante heterogeneous economic agents make dynamic discrete choices about entering into or renewing a long-term, non–state-contingent contract subject to some frictions, and the other side of the market is competitive but cannot, for informational or legal reasons, price-discriminate. For example, consider the classic labor market environment of Harris and Holmstrom (1982), in which risk-neutral firms set wages to insure risk-averse workers who cannot commit to turning down outside offers. We can use our framework to analyze the wage implications of heterogeneity in outside offer arrival rates.[9] In a pooling wage equilibrium, workers with infrequent outside offers receive lower wages than they would in a separating equilibrium and thus effectively subsidize the wages of less loyal workers who receive frequent outside offers.

The remainder of the paper is structured as follows: Section 2 discusses the related literature. Section 3 characterizes households' refinancing behavior given exogenous mortgage rates. Section 4 introduces the equilibrium, first with homogeneous and then with heterogeneous households. Section 5 lays out our key policy counterfactuals, Section 6 describes the data and estimation of household heterogeneity necessary to discipline these counterfactuals, and Section 7 quantifies the pricing and distributional consequences of this heterogeneity. Finally, Section 8 discusses other applications of our framework. The online supplementary materials contain all proofs.

2 Related literature

A growing literature provides evidence that households fail to refinance their mortgages optimally. Keys, Pope, and Pope (2016) argue that approximately 20% of US

causal effects. However, some of this effect is likely causal since these lenders profit primarily from mortgage origination and so actively encourage refinancing.

[9]There are many reasons the arrival rate of outside offers differs even for workers with identical productivity. Some workers are less "loyal" and solicit outside offers more aggressively, while others have constraints related to children's schooling or spousal employment that make potential outside employers view them as less "movable." Pooling is likely to arise both because "loyalty" cannot be directly observed and because it is illegal to set wages based on many of these characteristics.

households fail to refinance despite substantial benefits, and they provide some survey evidence supporting inattention and behavioral explanations. Agarwal, Rosen, and Yao (2016) provide empirical evidence that US households fail to refinance their mortgages optimally and correlate these patterns with various observable proxies for financial sophistication (see also Amromin et al. (2018)). Andersen et al. (2020) use even more detailed micro data from Denmark to show that both fixed costs and inattention are important for understanding individual refinancing patterns.

In complementary work, Fisher et al. (2021) and Zhang (2022) analyze the distributional impacts of heterogeneous refinancing rates. Fisher et al. (2021) analyze the UK mortgage market setting in which mortgages come with a short teaser rate that later resets to the market rate. Using a partial equilibrium consumption model, they estimate the distributional consequences of moving from this teaser system to a fixed-rate product that would generate the same revenue for lenders. Zhang (2022) uses US data to study cross-subsides arising from interactions between heterogeneous refinancing propensities and purchase points. He analyzes how closing fees change the equilibrium between mortgage originators and heterogeneous borrowers but takes MBS prices as fixed at their empirical values for the pooling equilibrium and computes the equilibrium only for the counterfactual separating environment. Our analysis is motivated by this same household heterogeneity; however, we develop an equilibrium mortgage pricing framework that endogenizes MBS prices and mortgage rates and show that these equilibrium forces have important redistributive consequences, irrespective of the presence of up-front closing costs.

Two related papers study models with equilibrium mortgage pricing but without permanent borrower heterogeneity. Guren, Krishnamurthy, and McQuade (2021) study mortgage market reforms in an equilibrium model with borrower refinancing and risk-neutral competitive mortgage investors. Ex post heterogeneity arises in their model from income and moving shocks, but households are ex ante identical. This means their model cannot speak to the distributional issues that are the focus of our paper. The model in Berger et al. (2021) is most similar to ours, but they focus on entirely different questions. Our relative contribution is twofold: first, we more fully analyze equilibrium and the importance of various frictions for pricing. Second, and more importantly, we study environments with permanent borrower heterogeneity and show that this heterogeneity generates important effects on inequality in equilibrium.

A large literature studies the impact of heterogeneous capital returns for the asset

side of households' balance sheets (see, e.g., Benhabib, Bisin, and Zhu (2011), Bach, Calvet, and Sodini (2020) and Fagereng et al. (2016)). We complement this work by showing that refinancing frictions contribute to wealth inequality via realized return heterogeneity on the *liability* side of households' balance sheets. This heterogeneity is more modest than return heterogeneity on the asset side but is very persistent and so can have a non-negligible effect on wealth inequality.

3 Households' refinancing behavior

In this section, we present a model of household mortgage refinancing. Given our focus on the US mortgage market, we study fixed-rate mortgage contracts that can be refinanced at any time. We consider households that face two types of potential refinancing frictions, which lead to *state-dependent* and *time-dependent* inaction. We initially take mortgage rates as given before endogenizing them in Section 4.

3.1 Setup

Time t is continuous. We consider a continuum of risk-neutral, long-lived households of measure 1, discounting flow utility at rate ρ. Each household has a long-term fixed-rate prepayable mortgage with coupon rate c_t and constant unit balance. We denote as m_t the prevailing mortgage interest rate, i.e., the rate that a household refinancing at time t can lock in. Refinancing is hindered by two different frictions. First, households are inattentive and make decisions only at discrete times, modeled as i.i.d. Poisson events occurring with intensity χ — the *attention rate*. Second, they bear upfront closing costs ψ when refinancing. Last, households move from one house to another at intensity ν; when doing so, they must reset their mortgage coupon to the prevailing mortgage rate.[10]

Given our focus on a Markovian environment, we assume that the aggregate uncertainty is summarized by a latent state vector x_t, a possibly multidimensional, time-homogeneous Itô process with drift $\mu(x)$, diffusion $\sigma(x)$ and infinitesimal gen-

[10] ν can be viewed as the sum of a moving intensity and an amortization intensity—under the assumption that contractual mortgage balances amortize exponentially (an approximation of the actual amortization profile of a standard 30-year mortgage contract). Moving-related fixed costs could be added to the model without changing any of our conclusions.

erator $\mathcal{L}$.[11] The mortgage interest rate is then a function $m_t = m(x_t)$ of this latent state vector. For now, we assume that $m(\cdot)$ is continuous in x, and in Section 4, we prove that the equilibrium of our economy must satisfy this property.

Later, we consider the consequences of heterogeneity in attention rates χ for mortgage rates. In partial equilibrium, this heterogeneity is irrelevant, and so for now our notation abstracts from any potential household-specific χ.

3.2 Interpreting the refinancing frictions

The inability to make decisions continuously is sometimes referred to as *time-dependent* inaction and has featured in a vast range of applications.[12] The attention parameter χ should be viewed as a stand-in for various nonmonetary frictions. Some households, for example, cannot refinance even if it is beneficial, because they have insufficient home equity or income (see Beraja et al. (2019)). Other households have low financial literacy and might only partially understand the mechanics of refinancing a mortgage. Thus, while we refer to households' *inattention*, this friction should be understood as encompassing a wide set of environmental and behavioral factors.

The up-front closing costs when refinancing lead to *state-dependent* household inaction decisions, which change with the economic environment. These up-front closing costs include application fees and the "points" payable out of pocket by borrowers on the transaction closing date; they also represent a component of the revenues collected by lenders upon mortgage origination.

3.3 Household optimal behavior

Let $V(x, c)$ be the valuation of all future mortgage liabilities for a household paying a coupon c on its mortgage, when the latent state is x. Such a household solves

$$V(x,c) := \inf_{a \in \mathcal{A}} \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(a)}dt + a_t \psi dN_t^{(\chi)}\right)\right], \tag{1}$$

$$\text{s.t.} \qquad dc_t^{(a)} = \left(m(x_t) - c_{t-}^{(a)}\right)\left(a_t dN_t^{(\chi)} + dN_t^{(\nu)}\right),$$

[11] $\mathcal{L}$ is defined over functions f of class $\mathcal{C}^2$ via $\mathcal{L}f(x) = \mu(x) \cdot \partial_x f(x) + \frac{1}{2}\text{trace}\left(\sigma'(x)\partial_{xx'}f(x)\sigma(x)\right)$.

[12] Some of the many applications include consumption-savings decisions (Reis, 2006), stock market investment (Abel, Eberly, and Panageas, 2007), and sticky prices (Calvo, 1983).

where $\mathcal{A}$ is a set of progressively measurable binary actions $a = \{a_t\}_{t\geq 0}$ such that $a_t \in \{0, 1\}$ at all times, $N_t^{(\chi)}$ (resp., $N_t^{(\nu)}$) is a counting process with jump intensity χ (resp., ν), $c_t^{(a)}$ is the coupon rate on the mortgage for a household following strategy a, and the subscript on the expectation indicates that it is conditional on the information available at time t. At the random points in time when the household pays attention, the household choice $a_t = 1$ represents a decision to refinance, while $a_t = 0$ means that the household chooses to keep its existing mortgage. V captures all mortgage liabilities—including the *current* mortgage (at rate c) but also all *future* mortgages arising from future refinancing decisions. Going forward, let $z_t := c_t - m_t$ be the refinancing incentive, or *rate gap*, of a given household at time t. In Online Appendix 1.1, we establish the following result:

Proposition 1. *V is twice continuously differentiable in x and continuous and strictly increasing in c. It satisfies the Hamilton–Jacobi–Bellman (HJB) equation*

$$(\rho + \nu + \chi)\, V(x,c) = c + \mathcal{L}V(x,c) + \nu V(x, m(x)) + \chi \min\left[V(x,c), V(x,m(x)) + \psi\right]. \quad (2)$$

The optimal refinancing choice satisfies

$$a^*(x,c) = \mathbb{1}_{\{c - m(x) \geq \theta(x)\}}, \quad (3)$$

where the rate gap threshold $\theta(x)$ satisfies the indifference condition

$$V(x, m(x)) + \psi = V(x, m(x) + \theta(x)). \quad (4)$$

Our proof relies on standard results for continuous time stochastic control problems. Proposition 1 holds for any arbitrary (continuous) mortgage function $m(\cdot)$, not just the equilibrium one. It states that a household refinances optimally by following a state-dependent rate-gap cutoff $\theta(x)$ when it pays attention to mortgage rates. HJB (2) admits an analytical solution only in some special cases that we discuss now.

First, consider the environment where households do not bear any up-front closing costs. In this case, households optimally refinance as soon as they pay attention and their contractual coupon is above the mortgage market rate. This environment will soon become the main focus of our paper.

Corollary 1. *Absent upfront closing costs ($\psi = 0$), the rate gap threshold is $\theta(x) = 0$, and the optimal refinancing choice is $a^*(x, c) = \mathbb{1}_{\{c \geq m(x)\}}$.*

Next, consider the case where the mortgage rate is a Brownian motion. This simplified environment allows us to derive analytic expressions for the value function and rate gap threshold and leads to several important insights.

Proposition 2. *Assume that m_t is a Brownian motion with volatility σ. The (state-independent) rate gap threshold $\theta > 0$ satisfies the implicit equation*

$$e^{-\eta_0 \theta} + (\eta_0 + \epsilon_\chi)\,\theta = 1 + (\eta_0 + \epsilon_\chi)\,(\rho + \nu)\psi, \tag{5}$$

with constants η_0, η_χ and ϵ_χ that are equal to:

$$\eta_0 := \frac{\sqrt{2\,(\rho + \nu)}}{\sigma} \qquad \eta_\chi := \frac{\sqrt{2\,(\rho + \nu + \chi)}}{\sigma} \qquad \epsilon_\chi := \frac{(\rho + \nu)\,(\eta_0 + \eta_\chi)}{\chi}.$$

The threshold θ increases with the attention rate χ, and asymptotically:

$$\lim_{\chi \to 0} \theta = (\rho + \nu)\psi \tag{6}$$

$$\lim_{\chi \to +\infty} \theta = \frac{1}{\eta_0}\left[1 + \eta_0 \psi(\rho + \nu) + W\left(-\exp\left(-1 - \eta_0 \psi(\rho + \nu)\right)\right)\right], \tag{7}$$

where W is the Lambert W function. A Taylor expansion of (5) around $\theta = 0$ yields an approximation $\hat{\theta}$ of the threshold θ with an explicit formula:

$$\hat{\theta} = \sqrt{\frac{2}{\eta_0}\left(1 + \frac{\epsilon_\chi}{\eta_0}\right)(\rho + \nu)\,\psi + \left(\frac{\epsilon_\chi}{\eta_0^2}\right)^2} - \frac{\epsilon_\chi}{\eta_0^2}. \tag{8}$$

Our proof (Online Appendix 1.2) relies on the observation that the value function can be decomposed into (a) the present value of all future interest payments c/ρ (based on the current mortgage coupon) minus (b) the value of a prepayment option which, given the unit root behavior of mortgage rates, only depends on the rate gap z. Proposition 2 generalizes the results of Agarwal, Driscoll, and Laibson (2013) (hereafter ADL) to the case where households are inattentive.

While the rate gap threshold reduces to the ADL formula (7) if households are infinitely attentive, a decrease in χ reduces the rate gap threshold: households optimally refinance at smaller rate gaps in environments where they only pay attention

to rates sporadically. Figure 1 illustrates how the cutoff θ varies with attention. In our later empirical work, we estimate that average χ in the data is approximately 31% per year, i.e., $\log(\chi) \approx -1.2$. Figure 1 shows that for this attention level, the gap threshold for refinancing is only 40% as large as that in ADL. Interestingly, this might help explain, within an optimizing framework, the frequent empirical finding that borrowers refinance at rate gaps which are smaller than the ADL threshold.[13] More importantly for our analysis, the fact that inattention frictions dampen the importance of up-front closing costs will be one of three quantitative arguments we rely on to justify abstracting these costs in our full model with heterogeneity.

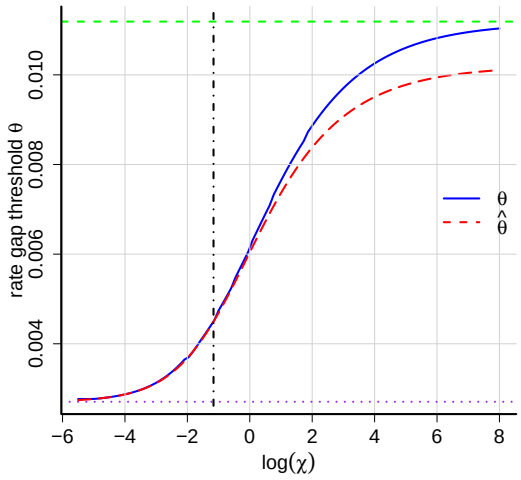

Figure 1: **Rate gap threshold θ vs. attention rate χ.** The solid blue line shows the threshold θ when m_t is a Brownian motion. The horizontal dashed green (dotted purple) line shows the limit of θ when $\chi \to +\infty$ ($\chi \to 0$). The vertical dot-dashed black line shows $\chi = 31\%$, our estimate of the average attention rate (see Section 6). The figure is computed for $\rho = 5\%$, $\nu = 8.5\%$, $\psi = 2\%$ and $\sigma = 1\%$.

In Online Appendix 1.3, we explore the sensitivity of our conclusions to the random walk assumption for mortgage rates. This is important since in our equilibrium analysis, m_t is endogenous and will ultimately be mean-reverting. We show that mean reversion induces state-dependence in refinancing decisions and leads to an increase in refinancing thresholds but that these effects are quantitatively modest.

[13] See, e.g., Agarwal, Rosen, and Yao (2016) and Fuster et al. (2019).

4 Mortgage market equilibrium

We now introduce mortgage investors and discuss the equilibrium environment. While borrowers pay coupon c_t on their mortgage, investors receive only $c_t - f$, with a wedge f capturing the fees charged by intermediaries for providing various services. At the time of origination, mortgage pools are sold by the initial lender to (secondary market) investors at a price of $1 + \pi$, where the "gain on sale" π represents revenues generated by the original lender (in addition to those arising from up-front closing costs ψ paid by borrowers).[14] Thus, originators collect total revenues $\psi + \pi$ per mortgage originated; with perfect competition, this sum must equal marginal origination costs, i.e., the *price of intermediation*, as defined in Fuster, Lo, and Willen (2017).

We initially focus on borrowers that are ex ante homogeneous in their attention rate χ. As we discuss in more detail in Section 6, our micro-data rejects the hypothesis of homogeneous attention. Nevertheless, this homogeneous environment serves as an important building block for the empirically relevant case in which borrowers exhibit ex ante attention heterogeneity.

We start with an environment that includes both state- and time-dependent inaction arising from up-front closing costs and inattention. However, three quantitative arguments will lead us to ultimately exclude up-front closing costs from our full model with heterogeneity. First, we already showed that up-front closing costs have muted effects on refinancing in the presence of time-dependent inattention frictions. We next show that realistic up-front closing costs have only small effects on equilibrium mortgage rates in the homogeneous environment. Finally, most borrowers in the US do not pay closing costs up front and instead roll them into higher rates. Together, these three observations motivate us to abstract from state-dependent frictions in our model with heterogeneity, which dramatically simplifies our subsequent analysis.

[14] Total revenues—the up-front closing cost ψ and the gain on sale π—compensate the lender for all costs incurred in connection with mortgage origination. These origination costs include (a) legal and underwriting, (b) broker commissions, (c) hedges of mortgage locks, (d) future servicing, and (e) the portion of guarantee fees related to "loan-level price adjustment" (for Fannie Mae) or "credit fees for mortgages with special attributes" (for Freddie Mac) and payable up front by the original lender. See Fuster et al. (2013) for a detailed description of mortgage lenders' costs of origination.

4.1 Homogeneous borrowers

In this section, all borrowers share the same attention parameter χ. When pricing mortgage debt, investors take borrowers' refinancing decisions as given. Let $P(x, c; \chi)$ denote the market price of a unit mortgage with coupon c whose borrower has attention intensity χ, when the latent state is x:

$$P(x, c; \chi) := \mathbb{E}_x \left[\int_0^\tau e^{-\int_0^t r(x_s)ds} (c - f)\, dt + e^{-\int_0^\tau r(x_s)ds} \right], \tag{9}$$

where τ is the (random) prepayment time. Competitive mortgage lenders must break even when extending a new loan and immediately selling it to secondary market investors and consequently need to generate a gain on sale π at the time of loan origination to recoup their costs. This yields the equilibrium condition

$$P(x, m(x); \chi) = 1 + \pi. \tag{10}$$

We are now equipped to define an equilibrium in this environment.

Definition 1. *A Markov perfect equilibrium (MPE) is defined as (i) a borrower value function V that satisfies (2), (ii) the associated optimal refinancing policy satisfying (3), (iii) a pricing function P defined via (9) and (iv) a mortgage rate function $m(x)$ that satisfies (10).*

In some of our subsequent analysis, we will narrow down our focus to one-dimensional processes for x_t. In that case, we can define a monotone equilibrium as follows.

Definition 2. *When x is uni-dimensional and $r(\cdot)$ is increasing, an MPE is "monotone" if the mortgage rate $m(\cdot)$ is increasing in x.*

Since the definition of P in (9) implicitly depends on a mortgage rate function $m(x)$ (via the prepayment time τ), and since the equilibrium condition (10) defines $m(x)$ implicitly via the function P, the MPE is a fixed-point problem. Our equilibrium concept is then analogous to the Markov perfect equilibria studied in the sovereign or dynamic corporate debt literature.[15] In these environments, the existence and uniqueness of the equilibrium frequently depend on various assumptions. In the

[15] See Chatterjee and Eyigungor (2012) for an example of MPE in the context of a sovereign default model or DeMarzo and He (2021) in the context of a corporate dynamic capital structure model.

context of mortgage prepayments, the special case without up-front closing costs allows us to derive the following sharp result (see Online Appendix 2.1).

Proposition 3. *Assume a finite attention rate (i.e., $\chi < \infty$) and assume that short-term rates r_t are positive and bounded. Absent up-front closing costs (i.e., $\psi = 0$),*

i. If the gain on sale $\pi = 0$, there exists a unique MPE;

ii. If the gain on sale $\pi > 0$, and if x is uni-dimensional, there exists a unique monotone MPE.

Borrowers optimize over their refinancing decisions, taking mortgage rates as given. Investors price mortgages competitively, taking borrowers' refinancing behavior as given. Proposition 3 tells us that this fixed-point problem, absent closing costs, always admits a unique solution. In this special case, borrowers' decisions can be decoupled from the investors' pricing problem: irrespective of how rates evolve, borrowers want to refinance whenever their coupon is above the current mortgage rate.[16] In this environment, we can also derive the following comparative static result.

Proposition 4. *Under the assumptions of Proposition 3 for which a unique MPE exists, the mortgage market interest rate $m(\cdot)$ is increasing in the attention rate χ.*

Proposition 4 (see Online Appendix 2.2) implies that higher borrower attention is worse for mortgage market investors. With higher attention, borrowers exercise their prepayment option more optimally, and since mortgage investors are short this option, these investors react by raising mortgage market interest rates. The left-hand side of Figure 2 illustrates the sensitivity of the equilibrium mortgage rate function to a range of attention parameters. When χ increases from 50% to 150% of the average value estimated in the data (Section 6.2), $m(r)$ becomes steeper and the resulting ergodic average equilibrium mortgage rate increases by 49 bps.

When inattentive borrowers bear up-front closing costs, we can study numerically the extent to which these costs influence equilibrium mortgage rates. The right-hand side of Figure 2 illustrates this sensitivity for fixed costs ranging from 0% to 200% of empirically relevant average closing costs[17]; equilibrium mortgage rates decrease on

[16] This holds irrespective of the gain on sale $\pi \geq 0$. When origination costs are rolled into a higher rate rather than paid up-front, the problem still simplifies to comparing the current mortgage coupon to the market rate. Mortgage market rates then adjust so that π covers origination costs.

[17] Zhang (2022) estimates average origination costs of around 4%, 80% of which are financed via higher rates, with the balance financed by households via upfront closing costs

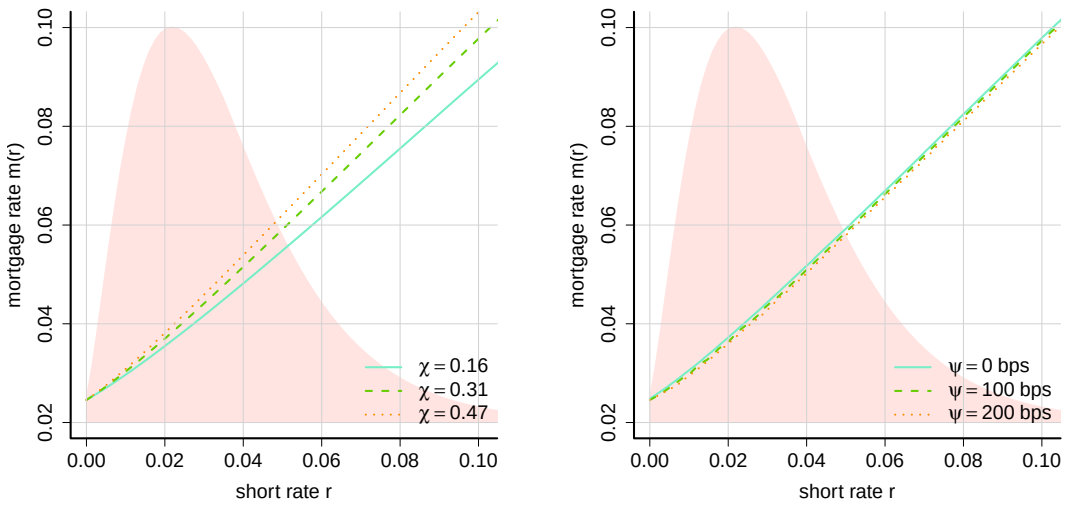

Figure 2: **Equilibrium mortgage rates vs. χ and ψ.** The left panel shows the sensitivity of $m(r)$ to the attention rate χ when $\psi = 0$. The right panel shows the sensitivity of $m(r)$ to the fixed-cost parameter ψ when the attention rate χ equals its estimated average value 31% (see Section 6.2). All other parameters are given in Table 1. The ergodic distribution of r is shown in the shaded pink area.

average by 13 bps when going from one extreme to the other. Figure 2 suggests that equilibrium mortgage rates have low sensitivity to up-front closing costs. It turns out theory backs this conclusion, as we discuss next (see Online Appendix 2.3).

Proposition 5. *Consider the case with no gains on sale ($\pi = 0$), and assume the sequence of MPEs indexed by ψ exists and is sufficiently smooth (in ψ). Denote $m_0(x)$ the mortgage rate in the MPE with $\psi = 0$; in the asymptotic expansion of the MPE when $\psi > 0$ and small, the mortgage rate $m(x)$ satisfies*

$$m(x) \underset{\psi \to 0}{=} m_0(x) + \psi m_1(x) + o(\psi),$$

with the first order correction term $m_1(x) = 0$.

Intuitively, small positive fixed costs induce inaction for households that have a small and positive rate gap. For investors, whether these mortgages get refinanced or not is irrelevant, as in either case, the value of their investment is close to par; this minimal effect of fixed costs onto investors' profits then translates into a low of sensitivity (at the first order) of the equilibrium mortgage rate to ψ.

Thus, up-front closing costs have little quantitative effect on equilibrium mortgage rates in the environment with homogeneous attention. Moreover, up-front closing costs represent only a small fraction of total origination costs in practice: as documented in Zhang (2022), 80.5% of these costs are rolled into a higher mortgage coupon. Lenders' origination costs are then recouped via the gain on sale π. Beyond simple motivations related to liquidity, this choice to roll fixed costs into the rate can also be cast as an optimal decision in an environment with bounded rationality: with sufficiently limited information processing capacity, this choice leads to a simple refinancing problem, which dominates the alternative—paying up-front closing costs and then solving a more complex option exercise problem.

These observations prompt us to make the following assumption, which simplifies our numerical computations in the case of ex ante heterogeneous borrowers and will apply for the remainder of the paper:

Assumption 1. *Borrowers do not face any up-front closing costs (i.e., $\psi = 0$).*

We end this section with a discussion of the interpretation of the MPE in Definition 1, connecting the homogeneous environment that we have studied until now to the heterogeneous environment we explore next. If borrowers are heterogeneous in their attention rate but investors can screen on χ, mortgage prices and mortgage market interest rates are type specific, i.e., $m(x, \chi)$, with each type's mortgage price determined by equation (9), and mortgage market interest rates determined by the break-even condition (10) just as in the homogeneous case. Thus, we will sometimes refer to the MPE in the homogeneous case as the *separating MPE*. When investors do not observe χ (i.e., in a "pooling" environment), significant complexities emerge.

4.2 Heterogeneous borrowers

Suppose now that there is a cross-section of attention types in the population, with cumulative distribution $H(\chi)$ and associated density h. Crucially, we assume that investors cannot screen on χ. We discuss this assumption and relate it to institutional features of the US agency MBS market in Section 4.4.

4.2.1 Infinite-dimensional problem

Similar to (1), define $V(S, c; \chi)$ as the valuation of all future mortgage liabilities for a type-χ borrower with current mortgage coupon c when the state of the economy is S.

Under Assumption 1, borrowers refinance whenever they pay attention and $m_t \leq c_t$ —just like in the homogeneous case.

Let $F_t(c, \chi)$ be the joint cumulative distribution over outstanding coupon rates c and types χ in the population at time t, with associated joint density $f_t(c, \chi)$. The relevant state of the economy, from the point of view of mortgage investors who cannot observe the type of individual borrowers, is $S_t := (x_t, F_t)$. This consists of the exogenous latent state x that determines current short rates together with the infinite-dimensional endogenous cross-sectional distribution F over current coupons and types. The mortgage market interest rate is then $m_t = m(S_t)$.

In a Markov perfect equilibrium, we need to specify the dynamics of the state vector S_t. x_t is exogenous and follows a time-homogeneous Markov process. The density f_t, instead, evolves endogenously over time with borrowers' refinancing decisions, according to equations described in Online Appendix 2.4.

We denote as $P(S, c; \chi)$ the *shadow* price of a mortgage with coupon c, conditional on the knowledge that the related borrower has attention rate χ, as defined in (9). We refer to $P(S, c; \chi)$ as a shadow price since investors do not observe χ and thus cannot trade conditional on χ.

The rate for newly originated mortgages depends on the characteristics of borrowers refinancing at time t. These borrowers have a type distribution with density

$$g_t(\chi) = \frac{\int_c \left(\nu + \chi \mathbb{1}_{\{c > m_t\}}\right) f_t(c, \chi)\, dc}{\int_\chi \int_c \left(\nu + \chi \mathbb{1}_{\{c > m_t\}}\right) f_t(c, \chi)\, dc d\chi}, \tag{11}$$

with corresponding cumulative distribution function $G_t(\chi)$. In low-rate states, this attention distribution G_t of *refinancers* is tilted towards higher-attention types relative to the distribution H of attention in the population. For example, consider the case where everyone's refinancing option is in the money at time t. The origination distribution g_t is then given by

$$g_t(\chi) = \frac{(\nu + \chi) h(\chi)}{\int_y (\nu + y) h(y) dy} = \left(\frac{\nu + \chi}{\nu + \bar{\chi}_H}\right) h(\chi), \tag{12}$$

where $\bar{\chi}_H := \mathbb{E}^H[\chi]$ is the average degree of attention in the population. Thus, g_t over-represents high-χ types relative to the population distribution h, in particular in low rate states. Conversely, when no one's refinancing option is in the money at time

t, the origination distribution g_t then coincides with the population distribution,

$$g_t(\chi) = h(\chi). \tag{13}$$

Our perfect competition assumption imposes the following restriction on the mortgage rate function $m(S_t)$:

$$\mathbb{E}^{G_t}\left[P\left(S_t, m(S_t); \chi\right)\right] := \int_{\chi} P\left(S_t, m(S_t); \chi\right) dG_t\left(\chi\right) = 1 + \pi, \tag{14}$$

subject to g_t given by (11) and where the superscript on the expectation indicates the distribution of borrower types χ over which the cross-sectional average is computed. We can then define a pooling Markov perfect equilibrium of this economy as follows.

Definition 3. *A pooling MPE is defined as (i) a refinancing policy satisfying (3), (ii) a shadow pricing function P defined via (9), (iii) a joint density f_t with evolution consistent with borrowers' refinancing decisions, (iv) a mortgage rate function $m(S_t)$ that satisfies (14), with (v) an origination distribution G that satisfies (11).*

This pooling MPE, which features an infinite-dimensional state space, is reminiscent of problems in heterogeneous agent models in macroeconomics (see Krusell and Smith (1998)), but with the additional complexity of a zero-profit pricing condition. Rather than addressing the computation of the pooling MPE in general, we will instead make simplifying assumptions that yield tractability while still capturing the main economic forces underlying the mortgage market equilibrium.

4.2.2 Simplifying assumption

The equilibrium computation in the pooling environment is *significantly* more complex than in the separating MPE, as it involves the determination of a fixed point in the space of functions of infinite-dimensional objects. To make progress, and for the remainder of the paper, rather than attempting to find such a fixed point, we make the following simplifying assumption:

Assumption 2. *Regardless of the path of r_t, investors price mortgages assuming a cross-sectional origination distribution that is either (i) a constant $G\left(\chi\right)$ or (ii) a state-dependent function $G\left(\chi|x\right)$.*

Assumption 2 restricts the origination distribution G used for pricing purposes to be dependent at most on the latent state x rather than on the full time-varying density f_t, and it thus reduces substantially the dimensionality of the relevant state space. While we make this assumption largely for computational tractability, it can also be justified when investors are engaged in k-level thinking, so that they understand the impact of refinancing incentives on prepayments but do not fully consider how this prepayment behavior then affects the attention distribution of refinancers over time. When we turn to the equilibrium definition, we will impose a consistency condition, in that the distribution G must be either the (i) unconditional or (ii) conditional ergodic average origination distribution G_t; this ensures that investors, while potentially making gains or losses upon their mortgage purchases at each point in time, break even on average.[18] The strength of Assumption 2 depends on how much the actual origination distribution G_t dynamically changes and differs from the distribution G assumed for pricing purposes; in Online Appendix 5.2, we compute the pricing errors made by investors and show that they remain quantitatively modest.

4.2.3 Mortgage pricing in the simplified environment

Under Assumption 2, the only relevant aggregate state variable for the investors' pricing problem is the latent state x_t, and we thus write the mortgage market interest rate $m_t = m(x_t)$. We continue to use $P(x, c; \chi)$ for the shadow price of a type-χ mortgage. Let $\bar{P}_G(x, c)$ be the expectation of $P(x, c; \chi)$ under the origination distribution G, and let the market price of a newly issued mortgage pool be

$$\bar{P}_G(x, c) := \mathbb{E}^G[P(x, c; \chi)]. \tag{15}$$

Under Assumption 2, the market equilibrium condition is given by

$$\bar{P}_G(x, m(x)) = 1 + \pi. \tag{16}$$

[18] Our approach resembles the approximation method developed in Krusell and Smith (1998); when G is the unconditional (conditional) origination distribution over attention, borrowers' decisions feed back into an aggregate behavior that leads to a constant (state-dependent) distribution f over coupons and types (c, χ) and thus a constant (state-dependent) attention distribution for newly originated mortgages G. Our algorithm thus leads us to solve a fixed-point problem, as in Krusell and Smith (1998); in the version of our approximation where the origination distribution is $G(\chi|x)$, the dynamics of this distribution are nonlinear, whereas the dynamics of the first moments of the relevant cross-sectional distribution in Krusell and Smith (1998) are assumed to be log-linear.

Finally, borrowers' optimal refinancing behavior combined with the mortgage rate function $m(\cdot)$ implies an ergodic cross-sectional distribution $f_\infty(x, c, \chi)$ and thus an ergodic marginal type distribution for refinancers. The unconditional origination distribution is given by

$$g(\chi) = \frac{h(\chi) \int_x \left[\left(\nu + \chi \int_{c \geq m(x)} f_\infty(c|x, \chi) dc \right) \right] f_\infty(x) dx}{\int_\chi h(\chi) \int_x \left(\nu + \chi \int_{c \geq m(x)} f_\infty(c|x, \chi) dc \right) f_\infty(x) d\chi dx}, \tag{17}$$

while the conditional origination distribution is given by

$$g(\chi|x) = \frac{h(\chi) \left(\nu + \chi \int_{c \geq m(x)} f_\infty(c|x, \chi) dc \right)}{\int_\chi h(\chi) \left(\nu + \chi \int_{c \geq m(x)} f_\infty(c|x, \chi) dc \right) d\chi}. \tag{18}$$

These distributions, as well as Assumption 2, help us build our equilibrium definition:

Definition 4. *An approximate pooling MPE is defined as (i) a borrower refinancing policy satisfying (3), (ii) a shadow pricing function P defined via (9), (iii) an ergodic joint density $f_\infty(x, c, \chi)$ and its corresponding ergodic marginal density over refinancers g satisfying either consistency condition (17) (in the unconditional case) or (18) (in the conditional case), (iv) a newly originated pool pricing function $\bar{P}_G$ defined via (15), and (v) the break-even condition (16).*

The separating MPE and the approximate pooling MPE are similar in that they both have a single aggregate state variable, x_t. However, they differ in two aspects. First, the break-even condition of originators in the heterogeneous case is a cross-sectional expectation version of that in the homogeneous case. Second, and most importantly, our approximate pooling MPE requires a consistency condition: the cross-sectional origination distribution G used by investors when pricing new issue mortgages needs to be consistent with the marginal density of refinancers, as implied by borrowers' behavior and the corresponding joint density f_∞ over the latent state x, coupon c and inattention χ. The approximation imposed by Assumption 2 allows us to establish some useful theoretical results and simplifies our numerical calculations.

Proposition 6. *Let x be uni-dimensional and $r(\cdot)$ be monotone increasing. Define*

the candidate mortgage rate

$$m(x;G) := f + \frac{1+\pi-\mathbb{E}^G\left[PO(x;\chi)\right]}{\mathbb{E}^G\left[IO(x;\chi)\right]}, \tag{19}$$

where G is a distribution defined in (17) (unconditional) or (18) (conditional), where

$$IO(x;\chi) := \mathbb{E}_x\left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds} dt\right], \qquad PO(x;\chi) := \mathbb{E}_x\left[e^{-\int_0^{\tau_{x,\chi}} r_s ds}\right],$$

and where, for any arbitrary x, $\tau_{x,\chi}$ is a stopping time with arrival intensity $\nu + \chi \mathbb{1}_{\{x_t \leq x\}}$. If $m(x;G)$ is monotone in x, there exists a unique monotone approximate pooling MPE of this economy, with $m(x;G)$ being the equilibrium mortgage rate.

See Online Appendix 2.5. If a monotone equilibrium exists, we compute its related unconditional and conditional origination distributions $G(\chi)$ and $G(\chi|x)$, and show that the mortgage rate must satisfy (19). Conversely, if the object m, defined in (19), is monotone increasing in x, then an approximate pooling MPE exists and is unique. What are the properties of equilibrium mortgage rates in this environment with permanent attention heterogeneity? Our next proposition (proven in Online Appendix 2.6) allows us to be more specific about the impact of cross-sectional heterogeneity on mortgage rates in the case of the unconditional approximate pooling MPE.

Proposition 7. *In a monotone unconditional approximate pooling MPE, the pool price $\bar{P}_G$ satisfies*

$$\bar{P}_G(x,c) = P(x,c;\bar{\chi}_G) - \mathbb{E}_x\left[\int_0^{\tau} e^{-\int_0^t r(x_s)ds} \mathbb{1}_{\{m(x_t)\leq c\}} \mathbb{C}\mathrm{ov}^G\left(\chi, P(x_t,c;\chi)\right) dt\right], \tag{20}$$

where τ is the prepayment time for a borrower with attention rate $\bar{\chi}_G := \mathbb{E}^G[\chi]$.

Thus, the pool price $\bar{P}_G$ in the market behaves *as if* it were made up of homogeneous borrowers with attention $\bar{\chi}_G$, with an adjustment equal to the average (conditional on the rate gap being positive) discounted cross-sectional covariance between (a) shadow mortgage prices and (b) attention rates. If the shadow price P is decreasing in χ in expectation whenever the prepayment option is in the money, this correction term is positive. This yields the following corollary:

Corollary 2. *In a monotone unconditional approximate pooling MPE, if the average (conditional on the rate gap being positive) discounted cross-sectional covariance*

between (a) shadow mortgage prices and (b) attention rates is negative, then the equilibrium mortgage rate $m(\cdot)$ when borrowers have a nondegenerate origination distribution G is lower than when borrowers are homogeneous with attention $\bar{\chi}_G$.

In all our numerical computations of the approximate pooling MPE, we find that the correction term in equation (20) is indeed positive. Intuitively, holding the average attention rate $\bar{\chi}_G$ constant, faster borrowers have a shorter effective maturity than slower borrowers. Investors make money off slower borrowers while making losses off faster borrowers. Since the average maturity of slower borrowers is higher than that of faster borrowers, a mean-preserving spread benefits investors by increasing $\int_\chi P(x, c; \chi) dG(\chi)$. The zero-profit condition then forces investors to pass this benefit on to borrowers in the form of lower mortgage rates.

We end this section by discussing how the interaction between the current interest rate and heterogeneity affects mortgage pricing and the state dependence of mortgage interest rates.

Proposition 8. *In a monotone approximate pooling MPE, let $\underline{x}$ be the lowest attainable latent state. Then the lowest mortgage rate $m(\underline{x})$ is invariant to the distribution over permanent heterogeneity H.*

Proposition 8 (proven in Online Appendix 2.7) delivers some intuition about how the mortgage rate function $m(\cdot)$ changes as the variance of the distribution H increases: m is relatively insensitive to attention heterogeneity when rates are low but substantially more sensitive in high-interest-rate states. The proof relies on the observation that if a borrower locks in the lowest attainable mortgage coupon $m(\underline{x})$, it will never refinance for strategic reasons, only due to an exogenous move. This necessarily means that such a lowest-coupon mortgage has a shadow price that is independent of the attention rate χ. The break-even condition at $x = \underline{x}$ allows us to conclude that $m(\underline{x})$ is invariant to H.

4.3 Redistribution via the mortgage market

We then consider the distributional effects of the pooling equilibrium by household type. Let the equilibrium mortgage rate be $m(x, G)$ in the approximate pooling MPE and let it be $m(x, \chi)$ in the separating MPE for type-χ households.

In a pooling environment, fast borrowers face lower mortgage rates and slow borrowers face higher mortgage rates than they would in a separating equilibrium. Since

investors break even on average, this means the pooling environment necessarily leads to cross-subsidies between borrowers. One simple measure of the extent of redistribution is thus $m(x,\chi) - m(x,G)$, i.e., the difference in mortgage rates that a type-χ household faces in the separating vs. approximate pooling MPEs.[19] When this difference is positive, type-χ households benefit from a subsidy, and it coincides with investors losing money on the specific mortgage, i.e. $P(x, m(x,G);\chi) < 1+\pi$.

However, this static difference in mortgage rates at origination provides only a partial picture of cross-subsidies over time since attention is a permanent household attribute. Indeed, households with different attention rates have different effective maturities of their current mortgages, and they benefit from (if they are of the fast type) or are hurt by (if they are of the slow type) pooling not just in their current mortgage but also every time they refinance in the future. As an alternative measure of redistribution, we thus consider $\mathbb{E}[c_t|\chi, \text{pooling}]$,[20] the ergodic average coupon paid by type-χ households in the approximate pooling MPE relative to its cross-sectional average $\mathbb{E}[c_t|\text{pooling}]$. This calculation takes into account not only the subsidies/taxes obtained by a household for a given mortgage but also those obtained for all future mortgages *on average*. The difference in these ergodic averages across household types stems from ex post differences in refinancing rates rather than from equilibrium forces. If one wants to factor equilibrium effects into our measure of redistribution, one needs to instead consider $\mathbb{E}[c_t|\chi, \text{separating}]$—i.e., the ergodic average coupons paid by household types in the separating MPE.

These various measures of redistribution will be considered when we study policy proposals and perform counterfactual calculations.

4.4 Pooling in the US Mortgage Market

We have described a number of general properties of our mortgage pooling environment and the resulting implications for redistribution. In this section, we discuss why this pooling equilibrium is relevant for the US mortgage market that we study in our empirical applications.

[19] Alternatively, assume that borrower type χ is not directly observable in the data, but suppose instead that one could separate borrowers along observables into $i \in I$ distinct groups with group distribution G_i, so that $G(\chi) = \sum_{i\in I} w_i G_i(\chi)$, with w_i being the weight of each respective group. Then, $\Delta_i m(x) := m(x,G_i) - m(x,G)$ captures the cross-subsidy to group i.

[20] The expectation operator $\mathbb{E}$ integrates over the ergodic density of the dynamic system, conditional on the equilibrium being the approximate pooling MPE.

First, the majority of mortgage lending in the US is funded through the agency MBS market. Fuster, Lo, and Willen (2017) document that between 2009 and 2014, only 20% of loans originated were kept on banks' balance sheets. As of 2020, 70% of conforming mortgages were originated by speciality mortgage lenders rather than deposit-taking institutions; these finance companies' sole objective is to originate conforming mortgages and immediately distribute them to investors via the agency MBS market (see Jiang (2019) or Buchak et al. (2018)).

To hedge their pipeline, these finance companies sell their origination book forward via the to-be-announced (TBA) market. TBA buyers do not know the exact mortgage pool that they will receive at settlement. Rather, they know only 5 characteristics of the pool: the agency (Fannie Mae or Freddie Mac), the average coupon, the maturity, the face value, and the settlement month. Thus, interest rates on mortgages originated by those finance companies are indirectly linked to the TBA market, in which prices take into account the fact that investors do not know the specific pool characteristics beyond those described above.[21]

Second, using our micro data (see Section 6), conditioning on the time a mortgage is originated, on the FICO score, and on the LTV ratio explains 95% of the cross-sectional variation in mortgage coupons, suggesting that mortgage originators do not price-discriminate on any dimension other than these two key variables (see also Hurst et al. (2016) for additional evidence on the lack of spatial heterogeneity of mortgage interest rates).

These considerations provide the rationale for our argument that a pooling equilibrium is relevant in the context of the US conforming mortgage market.

5 Policy evaluations and counterfactuals

We now turn to various policy proposals put forth in the academic literature and in policy circles, primarily aimed at improving mortgage borrowers' welfare by reducing costs induced by widespread financial mistakes and the lack of financial literacy. Our structural model allows us to evaluate the consequences of these policies in equilibrium. In this section we describe the theoretical mechanisms at play, and in Section 6 and Section 7, we take this analysis to the data.

[21] See Fuster, Lucca, and Vickery (2022) for a detailed discussion on the institutional features of the US MBS market and, in particular, the role of the TBA market.

5.1 Automatically refinancing mortgages

Consider the introduction of automatically refinancing mortgages (auto-RMs), as suggested for instance by Keys, Pope, and Pope (2016) or Campbell et al. (2011). With an auto-RM, a borrower pays the minimum realized mortgage rate since the mortgage's inception at time τ:

$$\underline{m}_t \equiv \min_{\tau \leq s \leq t} \{m_s\}. \tag{21}$$

The contractual rate of this product is thus tied to the minimum process of the mortgage market interest rate. The auto-RM appears particularly beneficial for inattentive borrowers since it means they can take advantage of rate reductions they would otherwise miss.[22] However, discussions around this proposal are usually cast in terms of the *partial equilibrium* and thus fail to take into account the *equilibrium* response of mortgage rates. Our model can speak to this response.

To ensure the existence of an MPE in this environment, we make the following *smart-contract* assumption:

Assumption 3. *No origination costs are incurred at the time of automatic rate resets. The equilibrium rate m_t is such that the price of a newly issued auto-RM is equal to $1 + \pi$.*

Under Assumption 3, a change in rates can then be viewed as a rate reset, just like under adjustable rate mortgages. However, unlike under an adjustable rate mortgage, this adjustment process is asymmetric: rates adjust down when the market rate declines but do not adjust up when the market rate rises. Origination costs are incurred only at the time households move, and thus take on a new mortgage at the then-current auto-RM rate, denoted (with an abuse of notation) $m(x, \infty)$.[23] We relegate all technical details to Online Appendix 3.1.

We make three observations about this environment. First, even though borrowers still have heterogeneous attention rates, this heterogeneity is irrelevant for pricing purposes due to the mortgage contract design. This case is thus equivalent to an

[22] See Agarwal, Rosen, and Yao (2016) for a quantification of the cost of these mistakes.

[23] Under Assumption 3, $m(x, \infty)$ is the limit of the separating MPE's mortgage market rate $m(x, \chi)$ as $\chi \to +\infty$ when the gain on sale $\pi = 0$. We also use this notation for $\pi > 0$, cognizant of the fact that we assume no origination costs are incurred upon rate reset under Assumption 3. Without this assumption, we would not have an equilibrium in the limit, as discussed in the online appendix.

economic environment in which borrowers no longer face any refinancing frictions, i.e., an environment without cross-subsidies.

Second, traditional fixed-rate prepayable mortgages trigger origination costs upon refinancing that are recovered by lenders via a combination of (i) up-front closing costs ψ paid by borrowers and (ii) the gain on sale π extracted from secondary market mortgage investors. Under Assumption 3, the auto-RM must be a more efficient contract since it removes these dead-weight origination costs at rate resets.

Third, when starting from the same rate at origination, borrowers almost always pay less under an auto-RM relative to a more traditional adjustable rate mortgage which adjusts to both rate decreases and increases. This means that lenders must charge higher rates at origination for auto-RM in order to break even.[24] We formalize this in our next proposition, which we prove in Online Appendix 3.2:

Proposition 9. *The auto-RM rate satisfies* $m(x, \infty) \geq r(x) + f$ *for all* x.

Next, suppose that borrowers with heterogeneous χ all initially pool in a traditional fixed-rate prepayable mortgage and consider what the introduction of the auto-RM does in this environment. The slowest borrowers overpay for traditional mortgages in the pooling equilibrium, while the fastest borrowers underpay. Thus, the former find it beneficial to migrate to the auto-RM when the opportunity arises since they can obtain an actuarial "fair" rate with no cross-subsidies. As these slow borrowers migrate to the auto-RM, the effective attention rate of borrowers left in traditional mortgage contract increases, pushing their mortgage rates higher in equilibrium. The slowest remaining borrowers in the traditional mortgage now subsidize the fastest ones and so now find it beneficial to migrate to the auto-RM, further pushing up traditional mortgage rates and raising effective attention in the traditional mortgage pool even more. This unraveling continues until only the highest type is left in the traditional mortgage market. This leads to the following proposition:

Proposition 10. *With heterogeneous attention rates, no financial constraints, and the ability of borrowers to choose between (i) traditional fixed-rate prepayable mortgages or (ii) auto-RMs, all borrowers migrate to the auto-RM.*

[24] Our auto-RM framework has similarities with models of wage determination with stochastic productivity, risk-averse workers and one-sided firm commitment—see, for instance, Harris and Holmstrom (1982) and Section 8 for a greater discussion on the mapping between the two models.

What could undo this unraveling? First, some borrowers might not fully understand the value of the refinancing option embedded in the auto-RM. When faced with rates $m(r, G) < m(r, \infty)$, they might gravitate towards the cheaper rate, even though the expected net present value of future interest costs is lower under the auto-RM than under the traditional mortgage.[25]

Second, the presence of financial constraints might lead some borrowers to pick traditional mortgages which come with lower initial rates even if this means paying cross-subsidies to other borrowers, if this allows them to *just* purchase their target home. In other words, the disutility of a suboptimal home allocation might outweigh the cross-subsidies and deadweight costs associated with refinancing inherent in the traditional mortgage.

Third, the auto-RM might not be the most desirable option if a borrower is risk averse, given the high associated cash-flow volatility in comparison to that under a more traditional mortgage. The welfare impact of this type of contract, in the presence of risk-averse borrowers, depends on the comovement of income with rates.

5.2 Improving financial decisions

The attention distribution H is a stand-in for a range of frictions, one of which is financial literacy. The cross-subsidies embedded in the pooling MPE—from inattentive to attentive borrowers—suggest that policies raising financial literacy for the least sophisticated borrowers may appear attractive to policymakers interested in reducing mortgage inequality. Our framework, however, emphasizes that this type of intervention—even when properly targeted—has equilibrium effects that tend to hurt untreated borrowers, via an increase in equilibrium mortgage rates.

A shift in the attention distribution H can also occur absent policy intervention, for instance with the evolution of the mortgage market. Buchak et al. (2018) document a significant increase in the share of all mortgages originated by specialty finance companies in recent years, reaching 50% in 2015. Financial technology (fintech) lenders' market share has also been rising during this time period, accounting for 8% of total US mortgage issuance as of 2016 (see Fuster et al. (2019)). In Section 7.4.2, we argue that these recent trends in mortgage origination and servicing

[25] With no gain on sale ($\pi = 0$), $m(r, G) < m(r, \infty)$ is a natural outcome. When gains on sale are required for originators to recoup costs (i.e., $\pi > 0$) then under assumption Assumption 3, this might not hold if $\bar{\chi}_G$ is large enough.

have improved the effective attention rate of the borrower population, with an impact on equilibrium mortgage rates that can be quantified through the lens of our model.

6 Household attention in mortgage prepayment data

We now estimate attention rates for the population of US conforming mortgage borrowers — an essential input in determining the equilibrium impact of the mortgage market interventions we are interested in. We use two separate datasets for this purpose: (1) a monthly *borrower*-level panel from Equifax Credit Risk Insight Servicing McDash (CRISM), which allows us to track various mortgage statistics for a sample of borrowers over time, and (2) a monthly *loan*-level panel from Fannie Mae's Single-Family Loan Performance (SFLP), which offers a longer sample period and more detailed covariates than CRISM but tracks loans rather than borrowers.[26] Online Appendix 4.1 provides full details on both data sets as well as on our sample construction.

6.1 Are households homogeneous in their attention rate?

A non-parametric regression of prepayments on rate gaps reveals that *all* prepayment types — not only pure rate refinancings, but also moves and cash-outs — are affected by rate gaps. Thus, the empirical counterpart to households' attention rate in the model is their *strategic prepayment intensities*, defined as prepayment intensities that are affected by the rate gap.

Using CRISM data, we then estimate a model in which our N_h households share a common prepayment intensity $\nu + \chi \mathbb{1}_{\{gap>\theta\}}$.[27] Our MLE delivers point estimates of $\hat{\nu} = 0.074$ (with s.e. of [xxx]) and $\hat{\chi} = 0.132$ (with s.e. of [xxx]).

In addition, our CRISM data delivers, for each household $i \leq N_h$, an empirical average prepayment rate $\hat{p}_i := s_i/t_i$, where s_i (resp. t_i) denotes household i number of prepayment events (resp. observed time periods). We can then compare the empirical cross-sectional distribution of prepayment rates $\{\hat{p}_i\}_{i \leq N_h}$ to the theoretical distribu-

[26] We use a 0.5% random sample of the CRISM data, covering around $N_h = 250,000$ borrowers.

[27] We choose $\theta > 0$ since this allows for some refinancing inertia to arise from up-front fixed costs and not just from inattention and shows that our conclusions are not sensitive to this choice. While we abstract from up-front fixed costs when solving for equilibrium (since they have small effects on pricing), they do affect refinancing decisions at small gaps and thus estimated *levels* of inattention.

tion $\{p_i\}_{i\leq N_h}$ of prepayment rates that would arise if households were homogeneous w.r.t. their prepayment intensities and household i was observed for t_i periods. A Kolmogorov-Smirnoff test rejects the hypothesis that the empirical distribution of average prepayment rates arises from a homogeneous group of households.

6.2 Estimating the attention distribution $H(\chi)$

In order to estimate the cross-sectional attention heterogeneity in our CRISM data, we use a clustering algorithm and assume that each household belongs to one of $N \ll N_h$ homogeneous groups. For given N, we use a maximum likelihood procedure to estimate a non-strategic prepayment intensity ν and group-specific attention rates $\{\chi_k\}_{k\leq N}$, and to allocate each individual i into a group k. If $\alpha : \{1,\dots,N_h\} \to \{1,\dots,N\}$ denotes a group assignment function, the log-likelihood of a prepayment observation y_{it} for household i in period t is then

$$\mathcal{L}_{it} = y_{it} \log\left\{1 - \exp\left(-\left(\nu + \chi_{\alpha(i)}\mathbb{1}_{\{gap_{it}>\theta\}}\right) dt\right)\right\} - (1-y_{it})\left(\nu + \chi_{\alpha(i)}\mathbb{1}_{\{gap_{it}>\theta\}}\right) dt,$$

where $dt = 1/12$ is the length of a time period. The log-likelihood of the entire data is then maximized over (a) the parameters $(\nu, \chi_1, \dots, \chi_N)$ and (b) the assignment function α. We choose $N = 5$ but show robustness to other choices in Online Appendix 4.3.[28] Figure 3 displays the results of our estimation in the form of the population distribution $H(\chi)$. 40.0% of households in our sample are almost never paying attention, while 6.0% are instead "hyper-attentive", with an estimated intensity of 221% p.a. The remainder of households have attention rates between these two extremes and fall into the remaining three groups. The resulting average attention rate is $\bar{\chi}_H = 31.3\%$ p.a., yielding an average 2.5% monthly attention probability.

We then recover the ergodic unconditional (conditional) distribution of refinancers $G(\chi)$ ($G(\chi|x)$) in our approximate pooling MPE using the estimated distribution H, under the assumption that interest rates and other model parameters are those used in Section 7 and described in Table 1.[29]

The left panel of Figure 3 (red bars) shows the ergodic unconditional origination

[28] Our baseline results use $\theta = 0.25\%$, and we also show robustness to this choice.

[29] In particular we assume a unidimensional interest rate process $r(x) = x$ and we verify that the resulting approximate pooling MPE is monotone for both (i) the unconditional and (ii) the conditional case and thus unique in both (i) and (ii). Our numerical derivation of G closely follows our theoretical derivation from the proof of Proposition 6.

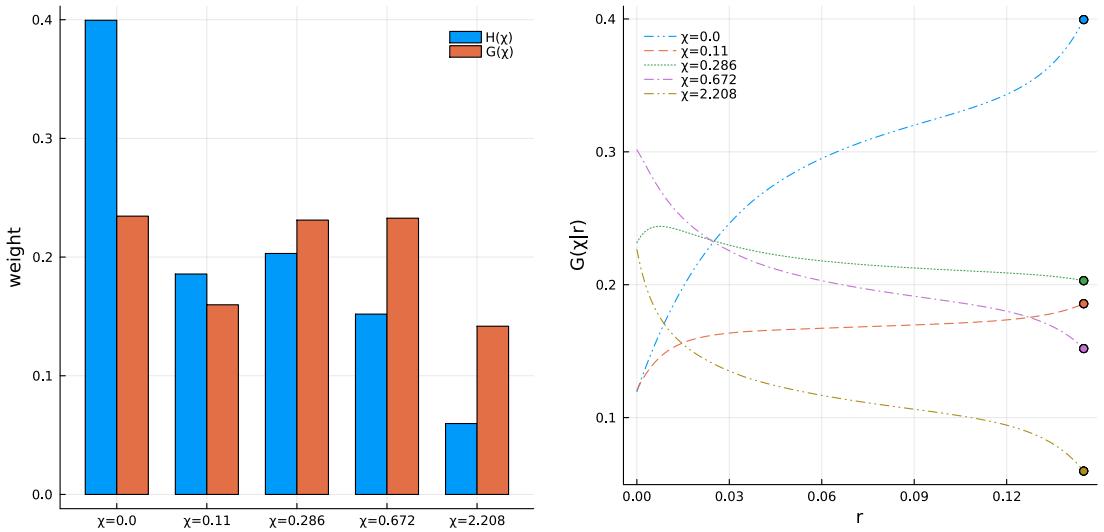

Figure 3: **Ergodic origination distribution $G(\cdot)$ implied by $H(\cdot)$.** The left panel shows the unconditional population (origination) distribution H (G) in the left blue bars (right orange bars). The estimation focuses on households and months with $gap > 0.25\%$, weighted by the average loan amount. The right panel shows the conditional origination distribution $G(\chi|r)$ (lines) and the unconditional population distribution $H(\chi)$ (thick dots).

distribution $G(\chi)$ that we will use for our quantitative analysis in Section 7, while the right panel shows the corresponding *conditional* distribution $G(\chi|r)$, which we will use for robustness purposes. Both origination distributions over-represent high-χ types and under-represent low-χ types relative to the population distribution H, as discussed in Section 4.2.1. This observation holds for the conditional distribution $G(\chi|r)$ for all but the highest interest rate state, at which point no one strategically prepays.[30] As the right panel of Figure 3 shows, the distortion in representation in the conditional case is especially strong in low-interest-rate environments. The unconditional origination distribution G is also skewed towards more attentive households, with $\bar{\chi}_G = 55\%$ that is greater than the population average $\bar{\chi}_H = 31\%$.

[30]The state $r = \bar{r}$ corresponds to the situation described by (13); the ergodic average conditional distribution $G(\chi|\bar{r})$ is thus equal to the distribution $H(\chi)$ of permanent heterogeneity.

7 Quantitative implications

In this section, we use our equilibrium model of mortgage rate determination to study quantitatively the various policies and counterfactuals discussed in Section 5.

7.1 Estimation/calibration of remaining model parameters

We use the cross-sectional attention distribution estimated in Section 6.2 via our clustering procedure. The short-term interest rate r_t follows a one-factor, square-root diffusion process as in Cox, Ingersoll Jr, and Ross (1985). We take as the relevant short rate the 3-month treasury rate and estimate the parameters of our term structure model via MLE on a sample covering 1971 to 2021.[31]

The parameter ν can be interpreted as the sum of unconditional prepayment and maturity intensities. We set the unconditional prepayment intensity to 5.2% p.a., as estimated Section 6.2. Since our empirical work focuses on 30-year mortgages, we assume a maturity intensity of 3.3% p.a. Thus, we set $\nu = 8.5\%$.

We set the wedge between mortgage payments made by borrowers and cash receipts by mortgage investors to $f = 0.45\%$, consistent with the estimated ongoing portion of G-fees paid to the GSEs as of 2019 (see the 2019 Federal Housing Finance Agency (FHFA) report on guarantee fees).[32] Finally, since we assume no closing costs borne by the household ($\psi = 0$), we set the gain on sale to $\pi = 80\% \times 4.6\% = 3.68\%$ since 80% of origination costs are financed via higher rates, and since the average cost of mortgage intermediation is 4.6% (see Zhang (2022)). Table 1 summarizes our parameter choice. We solve our model using a standard finite-difference method.

7.2 Validation of our equilibrium mortgage pricing

We first compare our model-implied mortgage rate function $m(x, G)$ with its data counterpart. Since we estimated the borrower attention distribution using a sample of borrowers observed between May 2005 and April 2017, we use this time period for comparison. Given our one-factor term structure model of interest rates, the

[31] In particular, we set $r(x) = x$, $\mu(x) = \kappa(\mu - x)$ and $\sigma(x) = \sigma\sqrt{x}$. The parameters to estimate are thus the long-run mean μ, the speed of mean reversion κ, and the volatility parameter σ.

[32] In general, f reflects (a) G-fees paid to the GSEs and (b) 25 bps servicing fee paid to mortgage servicers. Since the servicing fee is usually sold off separately by the originator and is thus already captured by the gain on sale π, we drop these servicing fees from f in our calibrations to avoid double counting. See Fuster, Lo, and Willen (2017) for related discussion.

Parameter	Value	Interpretation
μ	0.035	Long-run short rate mean
κ	0.13	Mean reversion coefficient
σ	0.06	Volatility
ν	0.085	Total unconditional prepay rate
f	0.0045	Ongoing portion of G-fees
π	0.0368	Gain on sale

Table 1: Baseline parameter values

yield at a single maturity characterizes the entire term structure and reveals the latent state x. We use the 10-year constant maturity zero-coupon Libor swap rate to retrieve x_t, which we then use to compute the relevant model-implied mortgage rate $m_t = m(x_t, G)$.[33]

In the left panel of Figure 4, we plot the time series of the model-implied mortgage rate as well as the 30-year mortgage rate from Freddie Mac's PMMS, as reported by the St. Louis Fed. Our model-implied mortgage rate is highly correlated with its data counterpart and has a time-series average that is only slightly above what we measure in the data. The only notable difference between the model and data occurs during the financial crisis period of 2009, when stress in financial markets and questions surrounding the implicit government backing of the GSEs resulted in very wide mortgage spreads. We could better capture this episode by modeling a credit spread factor as in Chernov, Dunn, and Longstaff (2018), but modeling this risk is outside the focus of our analysis. Overall, the model shows a good fit to the actual mortgage data given our simple one-factor model of interest rates.

In the right panel of Figure 4, we show the weighted average outstanding coupon in the data and through the lens of our model.[34] The model fit is once again good,

[33]We use the Libor swap rate as our benchmark since agency MBSs trade at an option-adjusted spread relative to the Libor swap curve (which our model assumes is zero). To construct the 10-year constant maturity zero-coupon Libor swap rate, we add the 23 bps swap spread from the sample period 2008–2017 to the 10-year constant maturity treasury rate reported by the St. Louis Fed. See Boyarchenko, Fuster, and Lucca (2019) for an extensive discussion of the option-adjusted spread in the agency MBS market.

[34]For the model implied average coupon, we use the initial distribution of coupons as of May 2005, as well as the group assignment obtained with our clustering algorithm. We drop households that cannot be assigned, for instance since they never have a single month with $gap > \theta$. We then use

except for an under-estimation of the average coupon from 2006 to 2009. This occurs because during that period, new households enter our sample with higher coupons than our model predicts, leading to an upward trend in empirical average coupons.

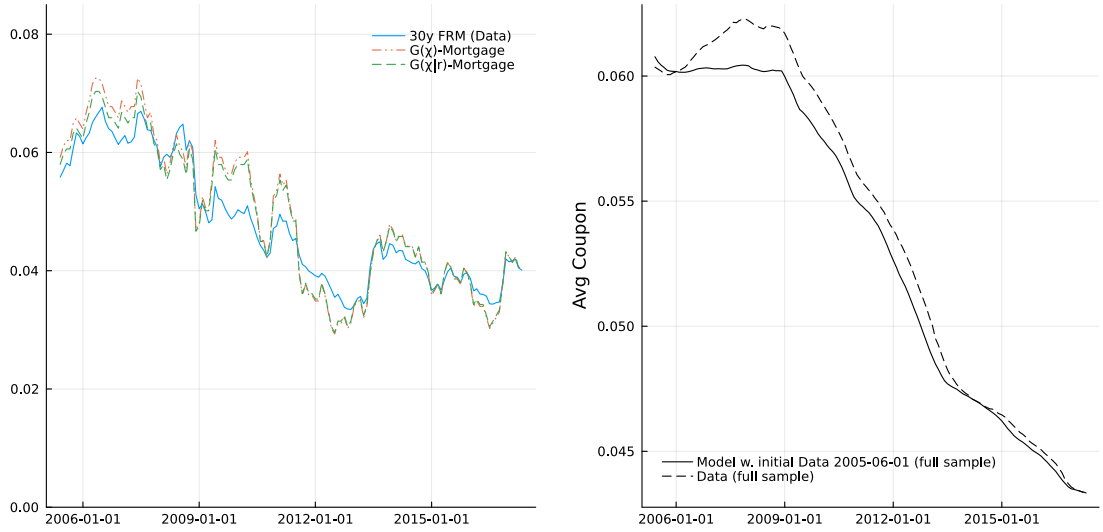

Figure 4: **Mortgage rate and average coupon time series 2005-2020.** Left panel: blue line represents Freddie Mac's 30-year PMMS rate, while the red dot-dashed and green dashed lines represent the model-implied mortgage rates for unconditional and conditional pricing, respectively. Right panel: solid black line shows the model-implied average outstanding coupon built from (a) the empirical coupon distribution in May 2005 and (b) the realized path of mortgage rates, while the dashed line shows the average coupon in the data.

7.3 Mortgage rates and redistribution

We next study the quantitative impact of attention heterogeneity on mortgage rates and its redistributive consequences.

The left panel of Figure 5 shows the mortgage rate $m(r, G)$ in the approximate pooling MPE and the counterfactual rate $m(r, \bar{\chi}_G)$ in the hypothetical environment where households are homogeneous with attention rate $\bar{\chi}_G$. Both mortgage functions are increasing in the short rate, with identical values at $r = 0$ but with $m(\cdot, \bar{\chi}_G)$ otherwise lying above $m(\cdot, G)$, as proved in Corollary 2. The ergodic average difference

the model to predict prepayment rates, based on the realized mortgage rate as the input into the model, thus purely isolating the prediction to the quantity side of the model.

in mortgage rates is 77 bps p.a., highlighting the nontrivial impact of cross-sectional heterogeneity in attention on the *level* of mortgage market rates.

In the same figure, we plot the corresponding separating MPE mortgage rates $m(r, \chi)$ for the slowest ($\chi = \chi_1 = 0\%$ p.a.) and fastest ($\chi = \chi_5 = 221\%$ p.a.) groups of households according to our clustering algorithm. The separating MPE mortgage rates for intermediate attention levels will be located in between these two curves, since these mortgage rates are increasing in attention (see Proposition 4). Fast households benefit tremendously from the pooling environment; with price discrimination based on type, they would face mortgage rates with an ergodic average that is 274 bps higher than that of the slowest households and 213 bps higher than those that they face in the approximate pooling MPE.

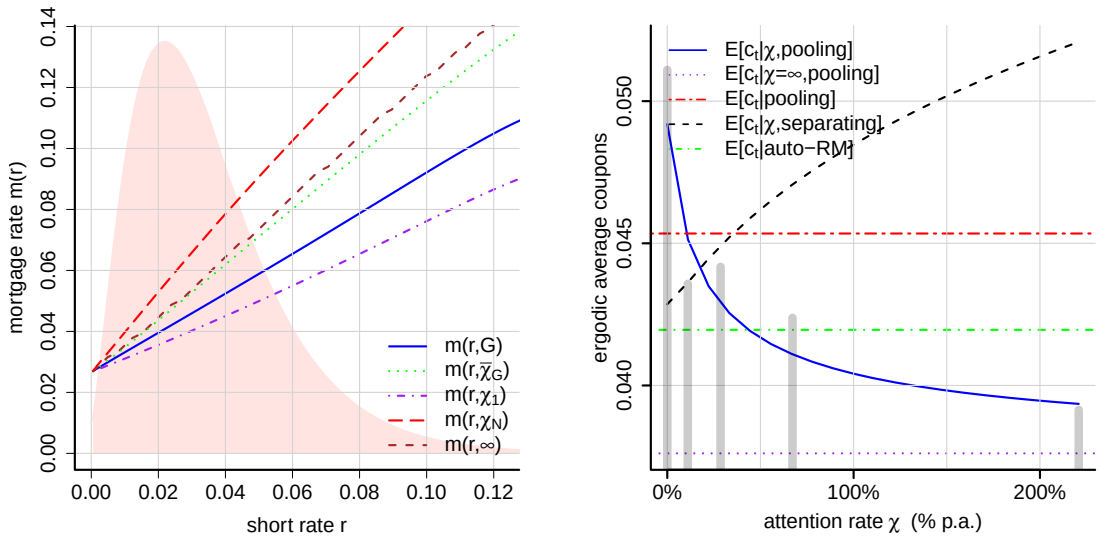

Figure 5: **Equilibrium mortgage rates and ergodic coupons.** The left panel shows equilibrium mortgage rates for (i) the approximate pooling MPE $m(r, G)$ (solid blue), (ii) the MPE with homogeneous households with attention $\bar{\chi}_G = 55\%$, $m(r, \bar{\chi}_G)$ (dotted green), (iii) the separating MPE $m(r, \chi)$ for type $\chi = \chi_1$ (dot-dashed purple) and $\chi = \chi_5$ (dashed red), and (iv) the auto-RM equilibrium $m(r, \infty)$ (dashed brown). The right panel shows the ergodic coupons as a function of attention χ for (i) the approximate pooling MPE (solid blue) and its cross-sectional average (double-dashed red), (ii) the separating MPE (dashed black), and (iii) the auto-RM equilibrium (dot-dashed green).

As noted before, fast (slow) households benefit from (are hurt by) cross-subsidies

in connection not only with a given mortgage but also with all future refinancing transactions. To capture the full scope of this redistribution, we plot in the right panel of Figure 5 the ergodic average coupon $\mathbb{E}[c_t|\chi, \text{pooling}]$ as a function of attention χ in the approximate pooling MPE and its corresponding population cross-sectional average (under distribution H) $\mathbb{E}[c_t|\text{pooling}]$. The ergodic average coupons in the approximate pooling MPE decrease as χ increases, with the fastest group paying on average 99 bps less than the slowest group and 60 bps less than the "average" household. Importantly, this difference stems purely from the ex post difference in household refinancing rates rather than from equilibrium forces.

To capture the distributional effects stemming from the pooling—rather than separating—equilibrium, we also plot the ergodic average coupon $\mathbb{E}[c_t|\chi, \text{separating}]$ as a function of attention χ in the separating MPE. The ergodic coupon is upward sloping as a function of attention χ, mainly due to the need for lenders to sell newly issued mortgages at a premium π to recoup their origination costs—and thus the need to charge higher rates to generate such gains on sale.[35] The degree of redistribution arising from the equilibrium effects is, in magnitude, as important as that arising from the ex post differences in attention; for instance, while the ergodic average coupon paid by the fastest households in the approximate pooling MPE is 60 bps below that paid by the "average" household in that equilibrium, it is also 127 bps lower than what such household would pay on average in the separating MPE.

The differences in the two panels of Figure 5 are substantial; however, as a household's attention rate is not observable, these differences do not speak directly to the nature of the redistribution among households with different *observable* characteristics. In Online Appendix 5.1, we address this question by computing the cross-sectional correlation between (a) household i's assigned attention rate $\chi_{\alpha(i)}$ and (b) various covariates X_i obtained either at the household, or zip-code level. Lower-income, lower-FICO, and younger households tend to be less attentive and will thus make on average greater mortgage interest payments than higher-income, higher-FICO and older households. Our computations thus suggest that the cross-subsidies we document are regressive, and go against the progressive subsidies embedded in the credit guarantee scheme provided by the GSEs.[36]

[35] Absent the gain on sale (when $\pi = 0$), the ergodic average coupon in the separating MPE is relatively insensitive to the level of attention; the only source of sensitivity stems from discount rate effects.

[36] The degree of progressivity of such scheme has evolved over time. Before the advent of LLPA

7.4 Evaluating policy proposals via model counterfactuals

In this section, we quantitatively evaluate the policy proposals and counterfactuals discussed in Section 5. We first focus on automatically refinancing mortgages and then turn to the impact of the rise of nonbank and fintech mortgage lenders on market interest rates.

7.4.1 Automatically refinancing mortgages

As described in Section 5.1, consider the introduction of an auto-RM, i.e., a mortgage whose coupon automatically resets to the prevailing market rate if that rate is below the current mortgage coupon. Figure 5 illustrates the equilibrium rate of an auto-RM. As stated in Proposition 9, the auto-RM rate (net of fees) is always above the short rate, i.e., $m(r,\infty) - f \geq r$. The auto-RM rate is also systematically higher than the mortgage rate in the approximate pooling MPE, i.e., $m(r,\infty) \geq m(r,G)$; the ergodic average difference between these two initial rates is 91 bps, highlighting the substantial increase in initial rates with a move from the traditional mortgage to this new financial instrument. To assess the potential effect of such an increase in equilibrium mortgage interest rates on households' housing and mortgage choice, we plot in Online Appendix 3.3 the debt-to-income (DTI) distribution at origination observed in our SFLP data vs. the counterfactual DTI distribution that would be prevalent in a world where households had access only to the auto-RM product. Focusing on the 43% DTI cutoff—the limit below which mortgages, until 2021, satisfied the "qualified mortgage" definition of the Consumer Financial Protection Bureau—approximately 16% of borrowers would be pushed above the DTI cutoff, potentially forcing them to downsize their house or increase their down payment upon purchase.

The right panel of Figure 5 shows that the ergodic average coupon in the auto-RM equilibrium is lower than the ergodic average coupon paid on average by households in the approximate pooling MPE—even though initial rates are higher, i.e., $m(r,\infty) \geq m(r,G)$. In a partial equilibrium setting, in which we would hold $m(r,G)$ constant, the ergodic average coupon for households able to entirely overcome their

price adjustments, G-fees were uniform across borrower types, implying that high credit risk borrowers were receiving a subsidy from low credit risk borrowers. Since the implementations of LLPA price adjustments, to the extent that these adjustments exactly correspond to the differential in expected credit losses of borrowers as a function of their FICO score or LTV ratio, most of the cross-subsidies embedded in the credit guarantee scheme have disappeared.

inattention friction would be $\mathbb{E}[c_t|\chi = \infty, \text{pooling}] = 3.5\%$ (purple dotted line)—i.e., 57 bps lower than when mortgage rates adjust to the equilibrium auto-RM rate. This calculation highlights the need to factor in the equilibrium response of mortgage rates in considering alternative contract designs: while some of these designs might appear to be substantially beneficial to households when prices are held fixed, their effect can be considerably dampened by equilibrium responses. In the approximate pooling MPE, the top two fastest groups have an ergodic average coupon below that in the auto-RM equilibrium. Thus, the faster households are hurt by the introduction of the auto-RM given our unraveling argument from Proposition 10, while the slower households benefit.

7.4.2 The rise of nonbank mortgage lending

The mortgage industry has witnessed a structural shift over the past 20 years, with nonbank lenders responsible for a growing share of mortgage origination at the expense of traditional banks. Using SFLP data and the bank vs. nonbank classification of Buchak et al. (2018), we plot the share of mortgage origination by banks, nonbanks and "others" over time in Figure 6 (left panel).[37] While banks had a market share of conforming mortgage originations greater than 75% in 2000, that share has consistently declined. At the same time, nonbank originators' volumes have increased from less than 5% of total originations to more than 30% as of the end of 2021.

Using the SFLP data, we study the differential prepayment propensity for mortgages originated by banks vs. nonbanks as a function of rate gaps. We estimate the linear probability model

$$prepay_{i,j,t} = \mathbb{1}_{bank}\beta_{gapbin,bank}\mathbb{1}(gapbin)_{j,t} + \mathbb{1}_{non-bank}\beta_{gapbin,non-bank}\mathbb{1}(gapbin)_{j,t} + \beta_X X_{i,j,t} + \epsilon_{i,j,t}$$

for borrower i with mortgage contract j at time t, where X is a vector of controls. Figure 6 shows the point estimates for $\beta_{gapbin,bank}$ and $\beta_{gapbin,non-bank}$ in our fully saturated specification, and where the bins used are constructed in 50 bps intervals. The "S-curves" estimated—and, in particular, the difference in levels between the negative and positive rate gaps—directly give us the average attention rate for

[37] The identity of the originator (or "seller") is not available in the CRISM dataset. The SFLP data do not include the identity of the seller for each loan; instead, each month, sellers whose combined at-issuance unpaid principal balance is less than 1% of total issuances are classified as "others".

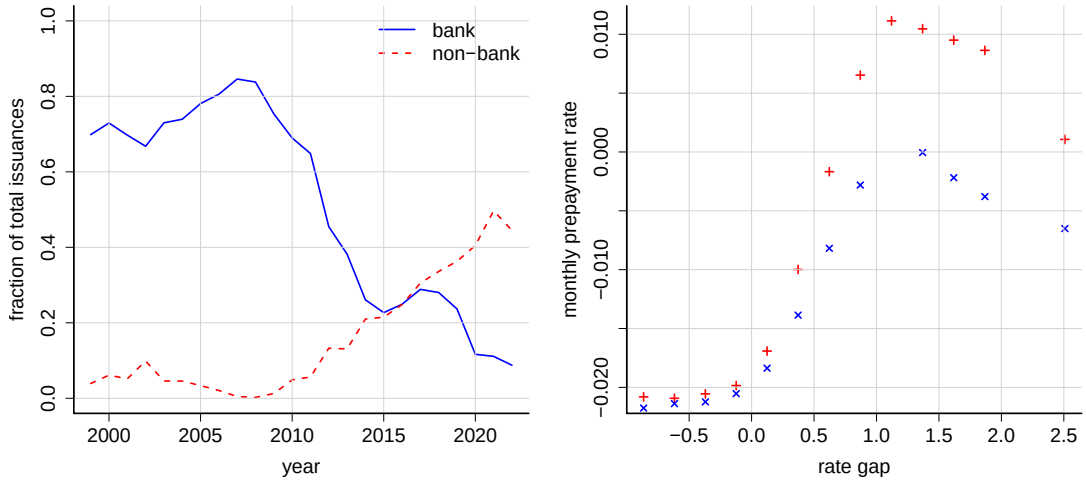

Figure 6: **Rise in nonbank lending.** The left panel shows the fraction of new mortgages classified as originated by "banks", "nonbanks" and "unknown" originators. The right panel shows our nonparametric estimate of the impact of the rate gap on prepayment rates in the SFLP loan sample.

bank- and nonbank-originated mortgage borrowers. On average, borrowers with bank-originated mortgages tend to be 100 bps per month less attentive than borrowers with nonbank-originated mortgages—a substantial difference in behavior. To reduce the risk that households borrowing from nonbank mortgage lenders are systematically different from those borrowing from banks, we saturate our regression with a battery of contract- and household-specific controls within the vector X.[38] Our results are consistent with those in Fuster, Lucca, and Vickery (2022), who conclude that faster prepayment speeds on fintech-originated mortgages stem from higher refinancing propensities rather than from selection of borrowers into fintech loans.

Using our estimate that nonbank lenders increase households' effective attention rates by 12 pp pa, we can compute the extent to which the rise of nonbank mortgage lending puts upward pressure on mortgage rates. In a counterfactual approximate pooling MPE in which households' attention rates are uniformly increased by 12%, ergodic average mortgage rates would increase by approximately 35 bps. This coun-

[38] Those controls include (i) a fully nonparametric function of the borrower's original FICO score, (ii) a fully nonparametric function of the borrower's original combined LTV ratio, (iii) a first-time home buyer flag, and (iv) the borrower's original real income. The SFLP data do not contain a borrower ID, only a loan ID; we are thus unable to include household fixed effects in our regression.

terfactual thus highlights that the dynamics of the mortgage origination market, and in particular the aggressive nudging practices of nonbank lenders, can have a large and systematic effect on mortgage market rates.

8 General framework

While we focus on the US residential mortgage market, our modeling approach is more general and can be applied to other environments where economic agents are ex ante heterogeneous and make dynamic discrete choices about entering into or renewing a long-term (non–state-contingent) contract subject to some frictions and the other side of the market is competitive but cannot price-discriminate for informational or legal reasons. We discuss below two alternative economic settings in which our framework can be applied: the labor market and the small business credit market.

8.1 Wage setting in labor markets

Consider a model of wage determination with stochastic productivity, risk-averse workers and one-sided commitment by the firm, as in Harris and Holmstrom (1982). Each worker has productivity x_{it} that follows a time-homogeneous Itô process. For simplicity, assume that individual worker productivity shocks are purely idiosyncratic. Workers are heterogeneous in their *job-hunting rate* χ—the rate at which they seek offers from competing firms. With risk-neutral firms and risk-averse workers, the optimal labor contract is a fixed-wage contract, with a wage w that is an endogenous function $\mathcal{W}(x_{it})$ of the worker's productivity at the time t at which she is hired.[39] Workers stay in their job, earning their fixed wage, but might quit and move to another firm if and when they receive an outside offer. When a job offer is received at time τ, the worker compares the proposed wage $\mathcal{W}(x_{i\tau})$ to her current wage w and accepts the offer if the lifetime utility $V(x_{i\tau}, w)$ from staying in the current job is below the lifetime utility $V(x_{i\tau}, \mathcal{W}(x_{i\tau})) - \psi$ from moving, with ψ a switching cost.

Firms are risk neutral and competitive, with discount rate r. They value a worker

[39] See, for instance, Harris and Holmstrom (1982) or Berk, Stanton, and Zechner (2010) for a discussion on the optimal labor contract in settings with risk-averse workers and a risk-neutral firm.

with productivity x, wage w, and job-hunting rate χ according to

$$\Pi(x, w; \chi) = \mathbb{E}_x \left[\int_0^\tau e^{-rt} (x_t - w) \, dt \right], \tag{22}$$

where τ is the quit time of the type-χ worker. With idiosyncratic productivity shocks only, there exists a well-defined stationary density $f_\infty(x, w, \chi)$ over workers' productivity x, wage rate w and type χ,[40] and a corresponding stationary conditional type distribution of *job transitioners* $G(\chi|x)$. When pursuing a prospective employee, the firm acts competitively and offers a wage $\mathcal{W}(x)$ that satisfies

$$\mathbb{E}^G \left[\Pi(x, \mathcal{W}(x); \chi) \right] = 0. \tag{23}$$

This condition is the counterpart to (16) in the context of the mortgage market, and it pins down the equilibrium wage rate $\mathcal{W}$. The expectation in (23) encodes firms' inability to discriminate based on workers' type χ—due to either information asymmetry or discrimination laws covering protected class statuses that might be correlated with χ.[41] One can define a pooling MPE of this environment, in which (i) workers optimally switch firms subject to their search and job-hunting frictions, (ii) firms' profits satisfy (22), and (iii) the equilibrium wage rate satisfies (23).

This environment allows us to study equilibrium wages; it allows us to analyze the impact of workers' cross-sectional heterogeneity in job hunting rates on equilibrium wages and on the implicit cross-subsidies that aggressive at-work job hunters receive from loyal workers via the labor market. Rich micro data on job-to-job transitions and wages within industries can then be leveraged to discipline the model and discuss the quantitative implications of this cross-sectional heterogeneity, as well as policy counterfactuals.

8.2 Other applications

While we discuss in some detail a labor market application of the framework developed in this article, other environments lend themselves to such an analysis.

[40] Our statement assumes that the equilibrium wage rate $\mathcal{W}$ is monotonically increasing in productivity; this property can be verified ex post, when our postulated equilibrium has been computed numerically.

[41] For example, marital status might predict mobility, but it is illegal to condition wages on marital status.

In Online Appendix 6, we build a model of the small business credit market, in which borrowers with partially observable and heterogeneous credit quality receive debt funding at fixed interest rate spreads from a competitive banking sector. With pooling, all borrowers entering into a credit agreement at a given point in time receive identical terms, creating cross-subsidies from good to bad credit quality firms. When credit conditions improve, borrowers refinance their debt at lower interest rate spreads, triggering a wave of loan prepayments, as is the case empirically. This application of our modeling framework leads to predictions for the degree of capital misallocation and the trajectory of interest rate spreads in the bank loan market, complementing previous studies of cross-subsidies in the presence of asymmetric information in banking (see Sharpe (1990) or Petersen and Rajan (1995)).

Other economic environments could be mapped into our model. While a given application might come with specific assumptions and modeling devices that are unique to that environment, our approach's tractability and ability to support systematic analysis of counterfactuals makes it an attractive framework. We leave the precise evaluation and analysis of these different settings for future research.

9 Conclusion

In this paper, we have studied the equilibrium consequences of pooling ex ante heterogeneous agents facing various frictions and making dynamic discrete choices about entering into or renewing a long-term contract with a competitive sector that cannot price-discriminate based on type. We have applied our theoretical framework to the US conforming mortgage market—an ideal laboratory in which mortgage lenders, for various institutional reasons, end up offering mortgages without type-specific pricing, creating cross-subsidies from slow borrowers to fast ones. Our micro data suggests a large degree of cross-sectional heterogeneity in households' attention rates, leading us to estimate significant cross-subsidies. Given that our measure of attention is correlated with income, the resulting redistribution is regressive, potentially to a much larger extent than that implied by the uniform credit guarantee scheme for agency mortgages. As policy discussions are regularly taking place in connection with a potential exit of the GSEs from conservatorship and the future of US housing finance, our paper provides a framework for exploring alternative mortgage market designs, taking into account the equilibrium effects of such counterfactuals.

References

Abel, Andrew B, Janice C Eberly, and Stavros Panageas. 2007. "Optimal inattention to the stock market." *American economic review* 97 (2):244–249.

Agarwal, Sumit, John C Driscoll, and David I Laibson. 2013. "Optimal mortgage refinancing: a closed-form solution." *Journal of Money, Credit and Banking* 45 (4):591–622.

Agarwal, Sumit, Richard J Rosen, and Vincent Yao. 2016. "Why do borrowers make mortgage refinancing mistakes?" *Management Science* 62 (12):3494–3509.

Amromin, Gene, Jennifer Huang, Clemens Sialm, and Edward Zhong. 2018. "Complex mortgages." *Review of Finance* 22 (6):1975–2007.

Andersen, Steffen, John Y Campbell, Kasper Meisner Nielsen, and Tarun Ramadorai. 2020. "Sources of inaction in household finance: Evidence from the Danish mortgage market." *American Economic Review* 110 (10):3184–3230.

Bach, Laurent, Laurent E Calvet, and Paolo Sodini. 2020. "Rich pickings? Risk, return, and skill in household wealth." *American Economic Review* 110 (9):2703–47.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu. 2011. "The distribution of wealth and fiscal policy in economies with finitely lived agents." *Econometrica* 79 (1):123–157.

Beraja, Martin, Andreas Fuster, Erik Hurst, and Joseph Vavra. 2019. "Regional heterogeneity and the refinancing channel of monetary policy." *The Quarterly Journal of Economics* 134 (1):109–183.

Berger, David, Konstantin Milbradt, Fabrice Tourre, and Joseph Vavra. 2021. "Mortgage prepayment and path-dependent effects of monetary policy." *American Economic Review* 111 (9):2829–78.

Berk, Jonathan B, Richard Stanton, and Josef Zechner. 2010. "Human capital, bankruptcy, and capital structure." *The Journal of Finance* 65 (3):891–926.

Boyarchenko, Nina, Andreas Fuster, and David O Lucca. 2019. "Understanding mortgage spreads." *The Review of Financial Studies* 32 (10):3799–3850.

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru. 2018. "Fintech, regulatory arbitrage, and the rise of shadow banks." *Journal of financial economics* 130 (3):453–483.

Calvo, Guillermo A. 1983. "Staggered prices in a utility-maximizing framework." *Journal of monetary Economics* 12 (3):383–398.

Campbell, John. 2006. "Household Finance." *Journal of Finance Studies* 61 (1):1553–1604.

Campbell, John Y, Howell E Jackson, Brigitte C Madrian, and Peter Tufano. 2011. "Consumer financial protection." *Journal of Economic Perspectives* 25 (1):91–114.

Chatterjee, Satyajit and Burcu Eyigungor. 2012. "Maturity, indebtedness, and default risk." *American Economic Review* 102 (6):2674–99.

Chernov, Mikhail, Brett R Dunn, and Francis A Longstaff. 2018. "Macroeconomic-driven prepayment risk and the valuation of mortgage-backed securities." *The Review of Financial Studies* 31 (3):1132–1183.

Cox, John C, Jonathan E Ingersoll Jr, and Stephen A Ross. 1985. "A theory of the term structure of interest rates." *Econometrica* .

DeMarzo, Peter M and Zhiguo He. 2021. "Leverage dynamics without commitment." *The Journal of Finance* 76 (3):1195–1250.

Fagereng, Andreas, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. 2016. "Heterogeneity in returns to wealth and the measurement of wealth inequality." *American Economic Review* 106 (5):651–55.

Fisher, Jack, Alessandro Gavazza, Lu Liu, Tarun Ramadorai, and Jagdish Tripathy. 2021. "Refinancing cross-subsidies in the mortgage market." *Bank of England Staff Working Paper* (948).

Fuster, Andreas, Laurie S Goodman, David O Lucca, Laurel Madar, Linsey Molloy, and Paul Willen. 2013. "The rising gap between primary and secondary mortgage rates." *Available at SSRN 2378439* .

Fuster, Andreas, Stephanie H Lo, and Paul S Willen. 2017. "The time-varying price of financial intermediation in the mortgage market." Tech. rep., National Bureau of Economic Research.

Fuster, Andreas, David O Lucca, and James I Vickery. 2022. "Mortgage-Backed Securities." .

Fuster, Andreas, Matthew Plosser, Philipp Schnabl, and James Vickery. 2019. "The role of technology in mortgage lending." *The Review of Financial Studies* 32 (5):1854–1899.

Gerardi, Kristopher, Paul Willen, and David Hao Zhang. 2020. "Mortgage Prepayment, Race, and Monetary Policy." Working paper.

Guren, Adam M., Arvind Krishnamurthy, and Timothy J. McQuade. 2021. "Mortgage Design in an Equilibrium Model of the Housing Market." *The Journal of Finance* 76 (1):113–168.

Harris, Milton and Bengt Holmstrom. 1982. "A Theory of Wage Dynamics." *The Review of Economic Studies* 49 (3):315–333.

Hurst, Erik, Benjamin J Keys, Amit Seru, and Joseph Vavra. 2016. "Regional redistribution through the US mortgage market." *American Economic Review* 106 (10):2982–3028.

Jiang, Erica Xuewei. 2019. "Financing competitors: Shadow banks' funding and mortgage market competition." *USC Marshall School of Business Research Paper Sponsored by iORB, No. Forthcoming* .

Keys, Benjamin J, Devin G Pope, and Jaren C Pope. 2016. "Failure to refinance." *Journal of Financial Economics* 122 (3):482–499.

Krusell, Per and Anthony A Smith, Jr. 1998. "Income and wealth heterogeneity in the macroeconomy." *Journal of political Economy* 106 (5):867–896.

Petersen, Mitchell A and Raghuram G Rajan. 1995. "The effect of credit market competition on lending relationships." *The Quarterly Journal of Economics* 110 (2):407–443.

Reis, Ricardo. 2006. "Inattentive consumers." *Journal of monetary Economics* 53 (8):1761–1800.

Sharpe, Steven A. 1990. "Asymmetric information, bank lending, and implicit contracts: A stylized model of customer relationships." *The journal of finance* 45 (4):1069–1087.

Zhang, David Hao. 2022. "Closing Costs, Refinancing, and Inefficiencies in the Mortgage Market." Working paper.