Refinancing Frictions, Mortgage Pricing and Redistribution*

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

December 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 with heterogeneous borrowers and use it to 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.

Keywords: Refinancing, Inequality, Mortgages, Equilibrium, Interest Rates

*We would like to thank Andreas Fuster, Morris Davis, David Zhang, Lu Liu and Kris Gerardi (all of them discussants), Fernando Alvarez, You Suk Kim, and James Vickery (for helpful discussions) and seminar participants at USC, UIUC, Northwestern University, UCLA, Copenhagen Business School, the Federal Reserve Board, UIC, Princeton, Baruch College, McGill, Cleveland Fed, Dallas Fed, Philadelphia Fed, NYU Stern, University of Chicago Booth, BU Finance, Society of Economic Dynamics, Chicago Fed Housing and Racial Bias Workshop, Rio FGV, TAU, Adam Smith Workshop and UCLA/SF Fed conference on Housing, Financial Markets, and Monetary Policy (for helpful feedback). Fabrice Tourre acknowledges financial support from the Bert and Sandra Wasserman endowment.

†Duke University and NBER; david.berger@duke.edu
‡Northwestern University and NBER; milbradt@northwestern.edu
§Baruch College; fabrice.tourre@gmail.com
¶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 the ability for lenders 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 these effects on inequality 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 permanent "ex-ante" 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" frictions (which generate time-dependent inaction) and for fixed costs of refinancing (which generate state-dependent inaction) and solve for the optimal behavior of borrowers.

In the second step, we embed this 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, and we characterize the relative role of different frictions for equilibrium mortgage rates. We show that equilibrium mortgage pricing is barely affected by fixed costs, but is highly sensitive to the degree of borrower inattention. Indeed, fixed 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 changes lender profits. In contrast, inattention reduces refinancing even for borrowers with large gaps, substantially affecting lender profits and thus mortgage pricing.

In the third step, we introduce heterogeneous refinancing frictions across borrowers into this equilibrium environment. This environment with heterogeneos borrowers allows us to explore the equilibrium 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. The pooling environment adds the cross-sectional distribution over coupons and attention rates as a (infinite-dimensional) state variable in both the borrower and investor problems.

In order to increase tractability without compromising our results, we make two simplifying assumptions. 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 broadly consistent with empirical evidence, and substantially simplifies borrowers' decisions. Second, we assume that investors exhibit a simple form of bounded rationality: they value mortgages using the average attention distribution of those refinancing at the current mortgage rate rather than using the entire history of rates to infer the current attention distribution.[3] These two assumptions simplify the calculation of the pooling

[3]The quantitative importance of this assumption cannot be fully evaluated without solving the

equilibrium 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 the data. We take US treasury yields from 2005 to 2017 together with estimates of intermediation costs from the literature and calculate the time-series of equilibrium mortgage rates as well as mortgage coupons for borrowers with different attention rates implied by the model. These align well with the data, giving us confidence in the model's predictions.

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 at origination. 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. 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 100 bps below the slowest borrowers. While interesting, this measure of ex-post inequality arising in the current rate environment does not require an equilibrium model. Our other results rely crucially on our model's counterfactual equilibrium analysis.

First, we compute the mortgage coupons that fast and slow borrowers would pay in a counterfactual "separating" equilibrium.[4] We find that on average the fastest borrowers pay 125 bps higher coupons in the separating equilibrium than they do in the pooling equilibrium. This 125 bps change in coupons is a measure of the cross-subsidies received by fast borrowers from slow borrowers through pooling. Notably,

true (intractable) pricing problem. However, we provide evidence that this simplifying assumption likely has little quantitative effect on our conclusions.

[4]The relevance of this counterfactual depends on whether borrower speed is observable and thus potentially priced ex-ante. Evidence in Gerardi, Willen, and Zhang (2020) as well as our own empirical analysis (see Online Appendix C) suggest that ex-ante observable heterogeneity is indeed substantial.

this ex-ante cross-subsidy at origination is even larger than the 100 bps ex-post coupon inequality that emerges after origination. The fact that ex-ante cross-subsidies are even larger than the ex-post differences typically studied by the past literature demonstrates the importance of including equilibrium forces when measuring redistribution in the existing mortgage market.

We next move away from studying the existing mortgage market and use our model to analyze the equilibrium effects of alternative contract designs and mortgage market trends. First, we investigate the automatically refinancing mortgage, which refinances automatically with no active borrower intervention when rates decline. This makes borrowers infinitely fast and eliminates cross-subsidies and inequality across borrower types. This contract leads to much more refinancing for slow borrowers but also to an increase in mortgage rates at origination of about 110 bps that offsets some of this benefit. Over the loan life, automatically refinancing mortgages reduce average coupons by 70 bps for the slowest borrowers, but this decrease is substantially smaller than the 120 bps reduction that would arise without the equilibrium rate increase at origination. Automatically refinancing mortgages yield individual time-paths of mortgage coupons that decline more rapidly than traditional mortgages but that start from higher initial values, which could have undesirable consequences for housing markets: higher initial rates may force borrowers that are at debt-to-income (DTI) limits to downsize their purchases or exclude them 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. Our calculations imply that the increase in interest rates arising from a move to automatically refinancing mortgages would force almost 20% of borrowers to select smaller homes.

Second, we study the effect of alternative mortgage contracts which prevent refinancing during some initial "lockup" period which lasts for the first few years of the mortgage.[5] It might seem that introducing a constraint on refinancing would only hurt borrowers, but we show that this is not the case in equilibrium. Since these contracts only limit refinancing for some temporary period, they do not permanently lock borrowers into high rates. However, they do reduce the ability to engage in repeated "churning" of mortgages through refinancing. Eliminating this churning reduces the dead-weight costs associated with mortgage origination, and

[5]Small prepayment penalties likely have similar implications but are more complicated to analyze.

these savings are ultimately passed through into lower equilibrium mortgage rates at origination. This benefits all borrowers. Of course, there are heterogeneous effects of these contracts on different borrowers, since the initial lockup period mostly affects fast borrowers. On net in the pooling equilibrium, this alternative contract results in a substantial reduction in average coupons for slow borrowers and a small increase in average coupons for fast borrowers. Imposing a small constraint on rapid mortgage churning thus reduces inequality and actually helps many borrowers in equilibrium.

Third, we explore the effects of information disclosure. Byrne et al. (2023) run on experiment in Ireland and find that sending letters to borrowers with various reminders can lead to substantial increases in refinancing. While they study a single bank, we analyze the equilibrium effects of similarly increasing attention for the economy as a whole and find that it would lead mortgage rates to rise by about 30 bps. Interestingly, their information treatment induces a change in refinancing similar in magnitude to that observed for loans originated with fintech and other nonbank lenders in the US. This suggests that the rise of fintech lending in the US may have lead to a non-trivial increase in refinancing and thus equilibrium mortgage rates.

While our paper focuses on mortgage markets, our modeling framework can be used in many other settings. Applicable environments share the following features: on one side of the market, ex ante heterogeneous agents make dynamic discrete choices about entering a long-term, non–state-contingent contract subject to some frictions. The other side of the market is competitive but cannot, for informational or legal reasons, price-discriminate. We provide particular illustrations of how this framework can be applied to labor markets and to the small business credit market. 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. 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.

2 Related literature

A growing literature provides evidence that borrowers fail to refinance their mortgages optimally. Keys, Pope, and Pope (2016) argue that approximately 20% of unconstrained US borrowers who would benefit financially from refinancing fail to do so, and they provide some survey evidence supporting inattention and behavioral explanations. Agarwal, Rosen, and Yao (2016) provide empirical evidence that US borrowers 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. Byrne et al. (2023) provide further evidence for the importance of information frictions using an RCT to show that information treatments can affect refinancing for Irish borrowers.

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 borrower 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.

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 borrowers 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.

Lastly, 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 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 Borrowers' refinancing behavior

In this section, we present a model of mortgage refinancing decisions. Given our focus on the US mortgage market, we study fixed-rate mortgage contracts that can be refinanced at any time. We consider borrowers 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. A continuum of risk-neutral, long-lived borrowers of measure 1 discount flow utility at rate ρ. Each borrower has a long-term fixed-rate prepayable mortgage with coupon rate c_t and constant unit balance.[6] Let m_t be the prevailing mortgage interest rate, i.e., the rate that a borrower can lock in by refinancing at time t. Refinancing is hindered by two different frictions. First, borrowers 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

[6] Although our model abstracts from loan size, we estimate the model weighting observations by loan size. This allows us to capture relationships between loan size and prepayment which are relevant for mortgage pricing without explicitly introducing this heterogeneity into the model. Allowing for persistent cross-sectional heterogeneity in loan balances and attention in the model would complicate the notation but would not change the conclusions.

refinancing. In addition to refinancing, borrowers move at intensity ν and must reset their mortgage coupon to the prevailing mortgage rate when doing so.[7]

We focus on a Markovian environment and assume that aggregate uncertainty is summarized by a latent state vector x_t, where x_t is a possibly multidimensional, time-homogeneous Itô process with drift $\mu(x)$, diffusion $\sigma(x)$ and infinitesimal generator $\mathcal{L}$.[8] 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 borrower-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.[9] The attention parameter χ should be viewed as a stand-in for various nonmonetary frictions. Some borrowers, for example, cannot refinance even if it is beneficial, because they have insufficient home equity or income (see Beraja et al. 2019). Other borrowers have low financial literacy and might only partially understand the mechanics of refinancing a mortgage. Thus, while we refer to borrowers' *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* 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.

[7] 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.

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

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

3.3 Borrower optimal behavior

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

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

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, $N_t^{(\chi)}$ and $N_t^{(\nu)}$ are counting processes with respective jump intensities χ and ν, $c_t^{(a)}$ is the coupon rate on the mortgage for a borrower 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 borrower pays attention, the borrower choice $a_t = 1$ represents a decision to refinance, while $a_t = 0$ means that the borrower chooses to keep its existing mortgage. V captures all mortgage liabilities—including both the *current* mortgage (with coupon c) and 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 borrower at time t. In Online Appendix A.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*

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

The optimal refinancing choice satisfies

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

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

$$V(x,m(x)) + \psi = V(x,m(x)+\theta(x)). \tag{4}$$

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 borrower refinances optimally by following a state-dependent rate-gap cutoff $\theta(x)$ when it pays attention to mortgage rates.

Two special cases deserve particular attention. First, consider an environment where borrowers bear no up-front closing costs. In this case, borrowers optimally refinance if their contractual coupon is above the mortgage market rate when they pay attention. 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 $a^*(x, c) = \mathbb{1}_{\{c \geq m(x)\}}$.*

Second, consider the case where the mortgage rate is a Brownian motion. This special environment is analyzed and solved analytically in Berger et al. (2023), who show that the rate gap threshold in that case is (a) independent of the state, i.e. $\theta(x) = \theta$ and is (b) an increasing function of the attention rate χ. That is, for a given up-front closing cost, borrowers optimally choose to refinance at smaller rate gaps if they only pay attention to rates sporadically. This implies that inattention frictions reduce the importance of up-front closing costs for refinancing decisions. This is one of three quantitative arguments we will rely on to justify abstracting from these up-front costs in our full model with heterogeneity.

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).[10] Thus, originators collect revenues $\psi + \pi$ per mortgage originated. With perfect competition, this revenue must equal marginal origination costs (what Fuster, Lo, and Willen 2017 refer to as the *price of intermediation*).[11]

[10] 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. See Fuster et al. (2013) for a detailed description of mortgage lenders' costs of origination.

[11] The Mortgage Bankers Association reports an average profit per loan origination of 52bps for the period 2008-2022 (https://themreport.com/featured/11-23-2022/mba-imb-report). This represents

We initially focus on borrowers that are ex ante homogeneous in their attention rate χ. As we discuss in more detail in Section 5, 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.

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:

$$P(x, c; \chi) := \mathbb{E}_x \left[\int_0^\tau e^{-\int_0^t r(x_s)ds} (c - f)\, dt + e^{-\int_0^\tau r(x_s)ds} \right], \tag{5}$$

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. Thus, they need to generate a gain on sale π at the time of loan origination to recoup their costs. This yields the equilibrium condition

$$P(x, m(x); \chi) = 1 + \pi. \tag{6}$$

We can now 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 (5) and (iv) a mortgage rate function $m(x)$ that satisfies (6).*

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.*

accounting profits. Loan originators bear a variety of risks that cannot be hedged, and they need a minimum amount of equity capital in order to operate. Thus, we view these accounting profits as returns on capital, implying that *economic* profits in that business are small.

Since the definition of P in (5) implicitly depends on a mortgage rate function $m(x)$ (via the prepayment time τ), and since the equilibrium condition (6) 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 default or dynamic corporate debt literature.[12] In these environments, the existence and uniqueness of the equilibrium frequently depend on various assumptions. In the context of mortgage prepayments, the special case without up-front closing costs allows us to derive the following sharp result (see Online Appendix B.1).

Proposition 2. *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;

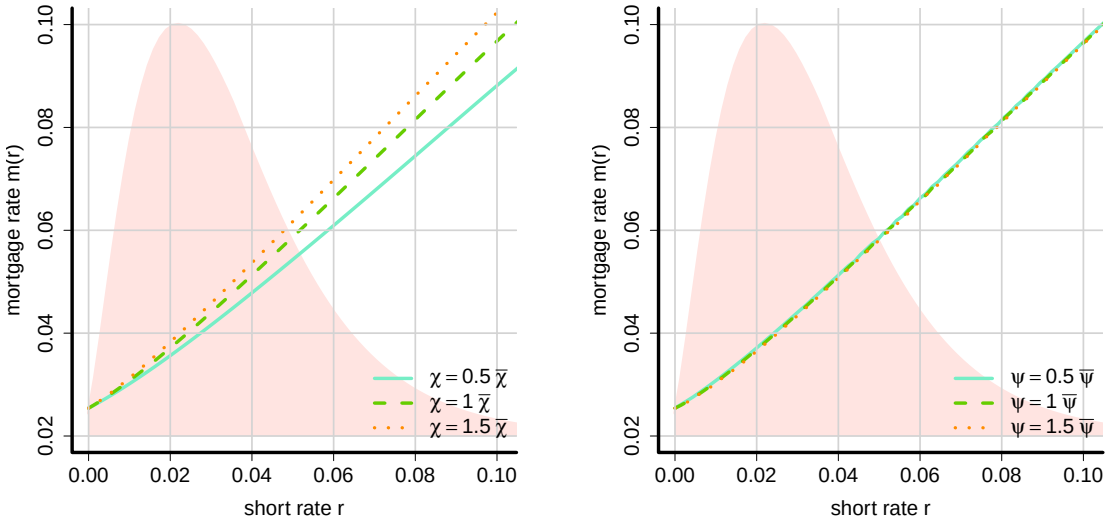
ii. If the gain on sale $\pi > 0$, and if x is uni-dimensional, there exists a unique monotone MPE.

Borrowers optimize over their refinancing decisions, taking mortgage rates as given. Investors price mortgages competitively, taking borrowers' refinancing behavior as given. Proposition 2 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.In this environment, we can also derive the following comparative static result.

Proposition 3. *Under the assumptions of Proposition 2 for which a unique MPE exists, the mortgage market interest rate $m(\cdot)$ is increasing in the attention rate χ.*

Proposition 3 (see Online Appendix B.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 1 illustrates how the equilibrium mortgage rate function changes as we move attention from 50% of the average value estimated in the data (Section 5.2) to 150% of this average value. When χ increases from 50% to 150% of the average value $\bar{\chi}$ in the data (Section 5.2), $m(r)$ becomes steeper and the resulting ergodic average

Figure 1: 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 $\bar{\chi} = 30\%$ (see Section 5.2). All other parameters are given in Table 1. The ergodic distribution of r is shown in the shaded pink area.

equilibrium mortgage rate increases by 50 bps.

We can compare these equilibrium effects of attention to those implied by up-front fixed costs. The right-hand side of Figure 1 illustrates how the the equilibrium mortgage rate function changes as we move up-front fixed costs from 50% to 150% of average costs $\bar{\psi}$ in the data.[13] Equilibrium mortgage rates decrease on average by 6 bps when going from one extreme to the other. Figure 1 shows that equilibrium mortgage rates have low sensitivity to up-front closing costs. Why do fixed costs have little effect on mortgage pricing? Intuitively, fixed costs induce inaction only for borrowers with small rate gaps. Whether these mortgages with small rate gaps get refinanced or not in turn has little effect on investor profits since refinancing a small gap has little effect on the resulting coupon. The low sensitivity of investors' profits to up-front fixed costs then translates into a low of sensitivity of the equilibrium mortgage rate to ψ. We develop this intuition more formally in the following analytical special case (see Online Appendix B.3 for detail):

[12] 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.

[13] Zhang (2022) estimates average origination costs of around 4%, of which 80% are financed via higher rates and 20% via upfront closing costs— thus, we set $\bar{\psi} = 0.8\%$.

Proposition 4. *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) near $\psi = 0$. 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 $m_1(x) = 0$.

Proposition 4 says that when $\pi = 0$, small up-front closing costs have no impact (to the first order) on the equilibrium mortgage rate.

These results show that in the environment with homogeneous attention, up-front closing costs have little effect on equilibrium mortgage rates. Furthermore, Berger et al. (2023) shows that inattention frictions reduce the importance of fixed costs for optimal refinancing behavior, as previously discussed in Section 3.3. Finally, even if we set aside these two theoretical arguments for the limited role of up-front costs, it is important to note that only a small fraction of origination costs actually are paid up front in practice: as documented in Zhang (2022), 80% of origination costs are rolled into a higher mortgage coupon rather than paid up front.

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 thus far to the heterogeneous environment we explore next. If borrowers are heterogeneous in their attention rate but investors can screen on χ, then mortgage prices and mortgage market interest rates are type specific, i.e., $m(x, \chi)$, with each type's mortgage price determined by equation (5), and mortgage market interest rates determined by the break-even condition (6) just as in the homogeneous case. Thus, we will sometimes refer to the MPE in the homogeneous case as the *separating MPE*. When there is heterogeneity and 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 function $H(\chi)$ and associated density $h(\chi)$. 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 4.5.

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 B.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 (5). 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

$$g_t(\chi, m) = \frac{\int_c \left(\nu + \chi \mathbb{1}_{\{c>m\}}\right) f_t(c, \chi)\, dc}{\int_\chi \int_c \left(\nu + \chi \mathbb{1}_{\{c>m\}}\right) f_t(c, \chi)\, dc d\chi}, \tag{7}$$

with corresponding cumulative distribution function $G_t(\chi, m)$ for an offered rate m. 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 current rates are low enough that everyone's refinancing option is in the money at time t, e.g., $m = 0$. The origination distribution g_t is then given by

$$g_t(\chi, 0) = \frac{(\nu + \chi)h(\chi)}{\int_y (\nu + y)h(y)dy} = \left(\frac{\nu + \chi}{\nu + \bar{\chi}_H}\right) h(\chi), \tag{8}$$

where $\bar{\chi}_H := \mathbb{E}^H[\chi]$ is the average degree of attention in the population. Thus, in this case, g_t over-represents high-χ types relative to the population distribution h. Conversely, when no one's refinancing option is in the money at time t, e.g., $m = \infty$, the origination distribution g_t then coincides with the population distribution,

$$g_t(\chi, \infty) = h(\chi). \tag{9}$$

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

$$\mathbb{E}^{G_t(\chi, m(S_t))}\left[P\left(S_t, m(S_t); \chi\right)\right] := \int_\chi P\left(S_t, m(S_t); \chi\right) dG_t\left(\chi, m(S_t)\right) = 1 + \pi, \tag{10}$$

subject to g_t given by (7) 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 an exact pooling Markov perfect equilibrium of this economy as follows.

Definition 3. *An exact pooling MPE is defined as (i) a refinancing policy satisfying (3), (ii) a shadow pricing function P defined via (5), (iii) a joint density f_t with evolution consistent with borrowers' refinancing decisions, (iv) a mortgage rate function $m(S_t)$ that satisfies (10), with (v) an origination distribution G that satisfies (7).*

This exact 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 exact 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 $G(\chi)$ or (ii) a state-dependent function $G(\chi|x)$.*

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.[14] 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 D.2, we compute the pricing errors made by investors and show that they are 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.

[14] Our approach resembles Krusell and Smith (1998) in that we solve for a fixed point of the approximate distribution G. That is, given a conjectured constant (or state-dependent) G, the implied pricing and resulting refinancing behavior indeed delivers the conjectured G on average.

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

$$\bar{P}_G(x,c) := \mathbb{E}^G[P(x,c;\chi)]. \tag{11}$$

Under Assumption 2, the market equilibrium condition is given by

$$\bar{P}_G(x,m(x)) = 1+\pi. \tag{12}$$

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

$$g(\chi) = \frac{h(\chi)\int_x \left[\left(\nu + \chi\int_{c\geq m(x)} f_\infty(c|x,\chi)dc\right)\right] f_\infty(x)dx}{\int_\chi h(\chi)\int_x \left(\nu + \chi\int_{c\geq m(x)} f_\infty(c|x,\chi)dc\right) f_\infty(x)d\chi dx}, \tag{13}$$

while the conditional origination distribution is given by

$$g(\chi|x) = \frac{h(\chi)\left(\nu + \chi\int_{c\geq m(x)} f_\infty(c|x,\chi)dc\right)}{\int_\chi h(\chi)\left(\nu + \chi\int_{c\geq m(x)} f_\infty(c|x,\chi)dc\right)d\chi}. \tag{14}$$

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 (5), (iii) an ergodic joint density $f_\infty(x,c,\chi)$ and its corresponding ergodic marginal density over refinancers g satisfying either consistency condition (13) (in the unconditional case) or (14) (in the conditional case), (iv) a newly originated pool pricing function $\bar{P}_G$ defined via (11), and (v) the break-even condition (12).*

For the remainder of the paper, we focus on this approximate pooling MPE and so we henceforth refer to it as the pooling MPE for short. The separating MPE and the 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 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 5. *Existence and Uniqueness of Equilibrium: Let x be uni-dimensional and $r(\cdot)$ be monotone increasing. Define a candidate mortgage rate*

$$m(x;G) := f + \frac{1+\pi-\mathbb{E}^G\left[PO(x;\chi)\right]}{\mathbb{E}^G\left[IO(x;\chi)\right]}, \tag{15}$$

where G is a distribution defined in (13) (unconditional) or (14) (conditional), where

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

represent the prices of newly-issued "interest-only" (thus the acronym "IO") and "principal-only" (thus the acronym "PO") MBS, 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, then there exists a unique monotone pooling MPE of this economy and it has equilibrium mortgage rate $m(x;G)$.

See Online Appendix B.5. We show there that if a monotone equilibrium exists, the mortgage rate must satisfy (15). Conversely, if the object m, defined in (15), is monotone increasing in x, then a 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 B.6) allows us to be more specific about the impact of cross-sectional heterogeneity on mortgage rates in the case of the unconditional pooling MPE.

Proposition 6. *In an unconditional pooling MPE, the pool price $\bar{P}_G$ satisfies*

$$\bar{P}_G(x,c) = P(x,c;\bar{\chi}_G) - \mathbb{E}_x\left[\int_0^\tau e^{-\int_0^t r(x_s)ds} \mathbb{1}_{\{m(x_t)\leq c\}} \mathbb{C}\mathrm{ov}^G\left(\chi, P(x_t,c;\chi)\right) dt\right], \tag{16}$$

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

Proposition 6 means that the pool market price $\bar{P}_G$ behaves *as if* it were made up of homogeneous borrowers with attention $\bar{\chi}_G$, but with an additional adjustment term 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 an unconditional pooling MPE, if the average (conditional on a positive rate gap) 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 pooling MPE, we find that the correction term in equation (16) 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 losing money with faster borrowers. Since the average mortgage life 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. In the quantitative evaluation of our model in Section 6, we find the effect of heterogeneity to be significant: mortgage rates are on average 110 bps lower than they would be if households were homogeneous with attention rate $\bar{\chi}_G$.

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 7. *In a monotone 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 7 (proven in Online Appendix B.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 $m(\underline{x})$, they will never refinance by choice. Since refinancing behavior is then independent of attention, this means that the associated lowest-coupon mortgage has a shadow price that also 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 now consider the distributional effects across borrowers of the pooling equilibrium. Let the equilibrium mortgage rate at origination be $m(x, G)$ in the pooling MPE and let it be $m(x, \chi)$ in the separating MPE for type-χ borrowers.

In a pooling environment, fast borrowers face lower and slow borrowers face higher mortgage rates than they would in a separating equilibrium. Since investors break even on average, this means that pooling necessarily leads to cross-subsidies between borrowers. One simple measure of the extent of redistribution is $m(x, \chi) - m(x, G)$, i.e., the difference in mortgage rates at origination that a type-χ borrower faces in the separating vs. pooling MPEs. If this difference is positive, then type-χ borrowers benefit from cross-subsidies from other borrowers. If it is negative, then they are instead cross-subsidizing other borrowers.

However, this static difference in mortgage rates at origination provides only a partial picture of cross-subsidies over time when attention is a permanent borrower attribute. Fast types benefit from pooling and slow types are hurt by pooling not just in their current mortgage but every time they refinance in the future. As an alternative measure of redistribution, we thus consider $\mathbb{E}[c_t|\chi, \text{pooling}]$, the ergodic average coupon paid by type-χ borrowers in the pooling MPE relative to the average coupon of all borrowers in the pooling MPE $\mathbb{E}[c_t|\text{pooling}]$. This calculation takes into account not only the subsidies/taxes obtained by a borrower for a given mortgage but also those obtained on average for all future mortgages. Importantly, since these comparisons are made within the pooling equilibrium, differences across borrowers arise solely from ex post differences in refinancing and do not reflect any ex ante cross-subsidies arising from equilibrium forces. To factor equilibrium effects into this dynamic measure of redistribution, one needs to instead compare $\mathbb{E}[c_t|\chi, \text{pooling}]$ to $\mathbb{E}[c_t|\chi, \text{separating}]$—i.e., the ergodic average coupons paid by borrower types in the separating MPE.

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

4.4 Other drivers of mortgage prepayment

To maintain tractability, our model is necessarily stylized and abstracts from various drivers of prepayment that are unrelated to rate incentives but that can influence mortgage rates: seasonality (prepayment rates tend to be lower in winter months), time-series variation in the markup π (due to capacity constraints faced by originators during periods of high refinancing activity), "aggregate" attention rate changes (for instance related to the "media effect" when rates fall abruptly), changes in fees f (related to the slow rise in average guarantee fees charged by the mortgage agencies), and volatility (mortgage rates co-move with various measures of option-implied rate volatilities). While some of these features of the data could be captured in our modeling framework (via the introduction of parameter state-dependence, i.e. $\pi(x)$, $f(x)$ or $\chi(x)$ for example), these should have little interaction with the cross-subsidies that we focus on in our analysis, and we thus leave the study of their effects on equilibrium mortgage rates for future research.

4.5 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 conforming mortgage market we study in our empirical applications.

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.[15]

Second, US federal law protects people from discrimination based on certain characteristics — so called "protected classes".[16] Lenders would thus be taking legal risks if they were to price mortgages differentially based on the perceived inattention of borrowers, and if inattention was correlated with protected classes.[17]

Third, beyond these institutional arguments, we find that in our micro data (see Section 5) origination month, FICO score, and LTV explain 95% of the cross-sectional variation in mortgage coupons. This provides some empirical evidence that mortgage originators indeed do not price mortgages based on other borrower characteristics.[18]

5 Borrower attention in mortgage prepayment data

In this section we estimate attention rates for the population of US mortgage borrowers using two separate datasets: (1) a monthly *borrower*-level panel from Equifax Credit Risk Insight Servicing McDash (CRISM), 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.[19]

We focus on borrowers whose mortgages were sold into either a Fannie Mae or Freddie Mac agency MBS pool since we just argued that the pooling equilibrium is relevant for this segment of the US mortgage market.[20] Online Appendix C.1 provides

[15] 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. Huh and Kim (2023) estimate that 80% of mortgages included in Fannie Mae or Freddie Mac MBS are sold via the TBA market, with the remaining 20% sold via the "spec pool" market.

[16] These characteristics are age, race, national origin, religious beliefs, gender, disability, pregnancy, and veteran status.

[17] For instance, one interpretation of the results in Gerardi, Willen, and Zhang (2020) would be that Black borrowers exhibit more inattention than non-Hispanic white borrowers.

[18] See also Hurst et al. (2016) for evidence on the lack of spatial heterogeneity in mortgage rates.

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

[20] In our baseline estimation, we include all households whose mortgages are included in a Fannie Mae or Freddie Mac MBS. In a robustness check in Online Appendix C.3.3, since Huh and Kim (2023) argue that most loans with initial balances below $150,000 are sold via the "spec pool"

full details on both data sets as well as on our sample construction.

5.1 The data rejects homogeneous borrower attention

We begin our empirical analysis by showing that the data reject a homogeneous distribution of borrower attention.[21] We start by estimating a statistical model in which all N_h borrowers in the CRISM data share a common prepayment intensity $\nu + \chi \mathbb{1}_{\{gap>\theta\}}$.[22] Our MLE delivers point estimates of $\hat{\nu} = 0.061$ (translating to a monthly probability $\hat{p}_\nu = 0.5\%$ with an s.e. of 0.0036%) and $\hat{\chi} = 0.105$ (with a monthly probability $\hat{p}_\chi = 1.4\%$ with an s.e. of 0.0048%).

We then compare the cross-borrower distribution of prepayment implied by this statistical model to the distribution in the data. For each borrower $i \leq N_h$ in CRISM, we compute an empirical average prepayment rate $\hat{p}_i := s_i/t_i$, where s_i denotes borrower i's number of prepayment events and t_i represents how many periods they are observed for. We can then compare the empirical cross-sectional distribution of prepayment rates $\{\hat{p}_i\}_{i\leq N_h}$ to the theoretical distribution $\{p_i\}_{i\leq N_h}$ of prepayment rates that would arise if borrowers were homogeneous w.r.t. their prepayment intensities and borrower 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 borrowers, with a KS statistic of 24.62.[23]

market and are thus not part of the pooling equilibrium, we restrict our sample further, to include only borrowers whose initial loan balance exceeds $150,000.

[21] The empirical counterpart to borrower attention rate in the model is the *strategic prepayment intensity*, which we define as the intensity of any prepayment that is affected by the rate gap. This includes pure refi but also some cashout and moving activity since a non-parametric regression of prepayments on rate gaps reveals that all of these components are affected by rate gaps.

[22] 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.

[23] The KS test is conducted as follows: the length of each household i's number of trials (periods with a gap above θ and below θ) and success (number of prepay events with gap above θ and below θ) is taken as given. Then, assuming a binomial distribution with estimated parameters $\hat{p}_\nu$ for gaps below θ and with $\hat{p}_\nu + \hat{p}_\chi$ for gaps above θ, we simulate the theoretical distribution of s_i^{sim}/t_i by simulating the success s_i^{sim} for each household i given t_i. We then compare this distribution via the KS test to the data distribution of s_i^{data}/t_i, i.e., using the empirical measure of success s_i^{data}.

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

While it is straightforward to reject the null-hypothesis of homogeneity, estimating the degree of heterogeneity in the data requires some mild additional structure. In order to estimate the cross-sectional attention heterogeneity in our CRISM data, we use a clustering algorithm and assume that each borrower belongs to one of $N \ll N_h$ homogeneous groups.[24] 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 $\alpha : \{1,\ldots,N_h\} \to \{1,\ldots,N\}$ denotes a group assignment function, the log-likelihood of a prepayment observation y_{it} for borrower 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 is then maximized over (a) the parameters $(\nu, \chi_1, \ldots, \chi_N)$ and (b) the assignment function α. In order to account for the fact that high-balance borrowers matter more for investor profits and also tend to prepay faster than low-balance borrowers, we weight each borrower by their average loan balance. Thus, our procedure delivers a size-weighted attention distribution H.

The left panel of Figure 2 (blue bars) displays our estimated attention distribution $H(\chi)$. 45.2% of borrowers in our sample are almost never paying attention, while 6.1% are instead "hyper-attentive", with an estimated intensity of 225% p.a. The remainder of borrowers have attention rates between these two extremes and fall into the other three groups. The resulting average attention rate is $\bar{\chi}_H = 30.4\%$ p.a., yielding an average 2.5% monthly attention probability. For comparison, Andersen et al. (2020) estimate that 84% to 92% of borrowers are "asleep" in any given quarter, corresponding to attention rates between 33% and 70% p.a.

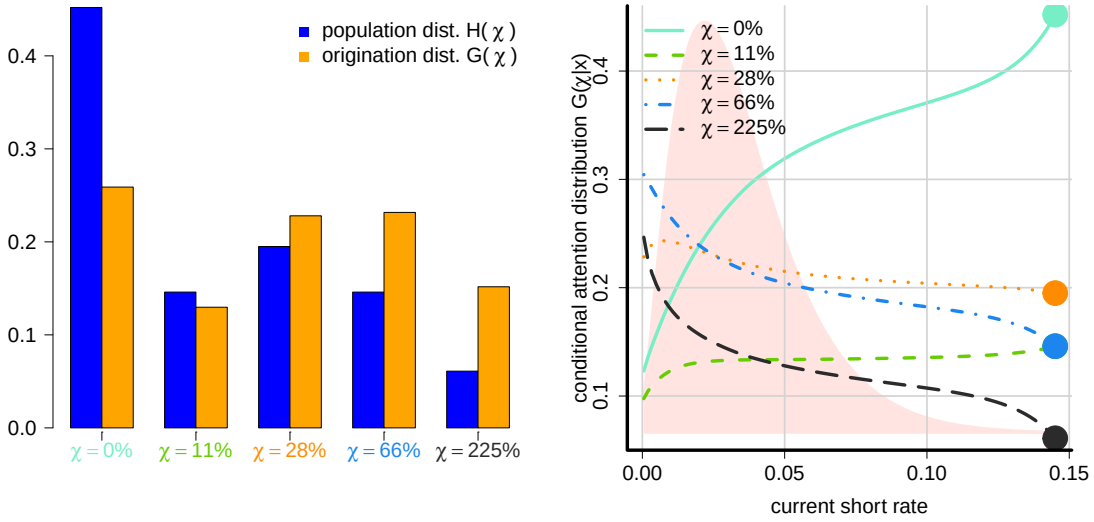
We then recover both the ergodic unconditional distribution of refinancers $G(\chi)$ and the conditional distribution $G(\chi|x)$ in our pooling MPE using the estimated distribution H, an estimated short rate process and other model parameters described

[24] We choose $N = 5$ and use $\theta = 0.25\%$. We show robustness to different choices of parameters in Online Appendix C.3: we provide the complete baseline results with gap threshold $\theta = 0.25\%$ in Online Appendix C.3.1, present robustness results for $\theta \in \{0\%, 0.5\%, 1\%\}$ in Online Appendix C.3.2, and present results for a subsample that additionally imposes an Original Loan Amount of at least \$150k in Online Appendix C.3.3.

in Section 6.[25]

The left panel of Figure 2 (orange bars) shows the ergodic unconditional origination distribution $G(\chi)$ that we will use for most of our quantitative analysis in Section 6. The right panel shows the *conditional* distribution $G(\chi|r)$, which we use for some robustness analysis. The origination distribution over-represents high-χ types and under-represent low-χ types relative to the population distribution H, as discussed in Section 4.2.1. The right panel shows that this force is especially strong when interest rates are low. Overall, the ergodic average attention rate of refinancers is $\bar{\chi}_G = 57.2\%$, substantially greater than the population average $\bar{\chi}_H = 30.4\%$.

Figure 2: Ergodic origination distribution $G(\cdot)$ implied by $H(\cdot)$



The left panel blue bars show the unconditional population distribution H and the orange bars show the origination distribution G. The estimation focuses on borrowers 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).

6 Quantifying Redistribution

We now use our model of mortgage rate determination quantify cross-subsidies in the US agency mortgage market.

[25] We assume a unidimensional interest rate process $r(x) = x$ and verify that the resulting pooling MPE is monotone and thus unique for both the unconditional and conditional case. Our numerical derivation of G closely follows from the proof of Proposition 5.

6.1 Estimation/calibration of remaining model parameters

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 the 3-month treasury rate as the relevant short rate and estimate the parameters of our term structure model via MLE using data from 1971 to 2021.[26]

We use the cross-sectional attention distribution together with the unconditional prepayment intensity of 3.7% p.a. that we estimated in Section 5.2. The parameter ν can be interpreted as the sum of unconditional prepayment and maturity intensities. Since our empirical work focuses on 30-year mortgages, we assume a maturity intensity of 3.3% p.a. Thus, we set $\nu = 7.1\%$.

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).[27] Finally, since we assume no closing costs borne by the borrower ($\psi = 0$), 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.

Table 1: Parameter values

Parameter	Value	Interpretation
μ	0.035	Long-run short rate mean
κ	0.13	Mean reversion coefficient
σ	0.06	Volatility
ν	0.071	Total unconditional prepay rate
f	0.0045	Ongoing portion of G-fees
π	0.0368	Gain on sale

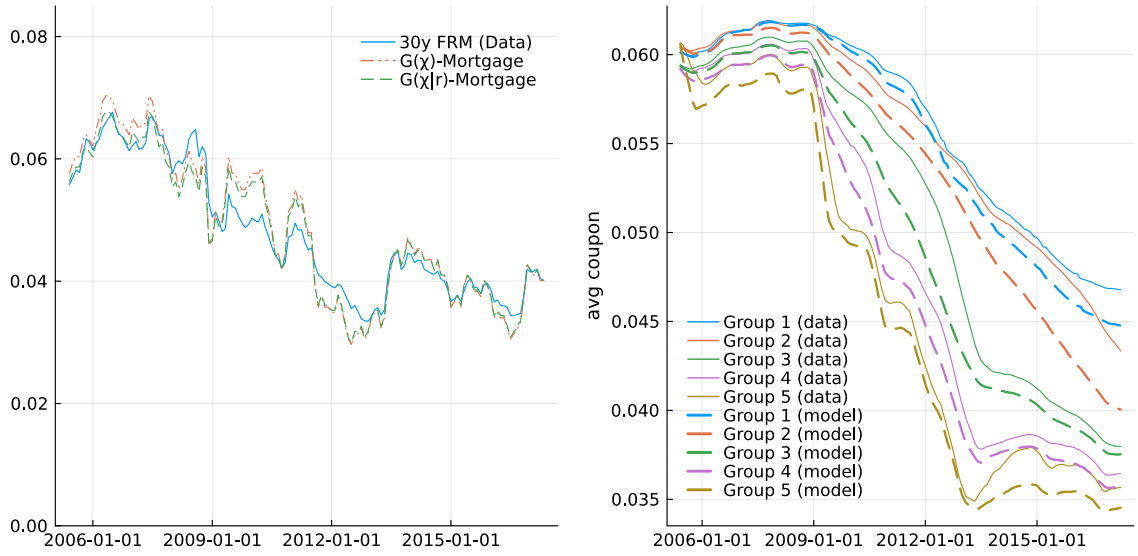
[26] In particular, we set $r(x) = x$, $\mu(x) = \kappa(\mu - x)$ and $\sigma(x) = \sigma\sqrt{x}$. The parameters to estimate are thus the long-run mean μ, the speed of mean reversion κ, and the volatility parameter σ.

[27] In general this wedge reflects both G-fees and the 25 bps servicing fee paid to mortgage servicers. We do not include the servicing fee in f since this fee is usually sold off by the originator and is thus already captured in the gain on sale π. See Fuster, Lo, and Willen (2017).

6.2 Model validation

We first compare our model-implied mortgage rate function $m(x, G)$ with its data counterpart. Since we estimated the attention distribution using a sample of borrowers observed between June 2005 and May 2017, we use this time period for comparison. Given our one-factor term structure model of interest rates, the 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 $m_t = m(x_t, G)$.[28]

Figure 3: Mortgage rate and average coupon outstanding time series 2005-2020.



Left panel: blue line is the 30-year FRM rate from Freddie Mac. The red dot-dashed and green dashed lines are the model-implied mortgage rates under unconditional and conditional pricing, respectively. Right panel: the solid lines show the data implied average outstanding coupon rates by group for all assigned households, while the dashed lines shows the model implied average outstanding coupon rates starting from the observed cross-sectional distribution 2005-06 and adjusting for monthly entry and exit of households in the panel. See Online Appendix C.1 for detail.

The left panel of Figure 3 compares the time series of the actual 30-year mortgage rate from the Freddie Mac PMMS to model-implied mortgage rates. We compute

[28] We use Libor swap rates since agency MBSs trade at an option-adjusted spread (equal to zero in our model) relative to the Libor swap curve. 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. See Boyarchenko, Fuster, and Lucca (2019) for an extensive discussion of the option-adjusted spread in the agency MBS market.

model implied mortgage rates under the two different forms of investor beliefs discussed before, but Figure 3 shows that this makes little difference for pricing.[29] The model yields a good fit to actual mortgage rates given our simple one-factor model of interest rates: the average level of model-implied mortgage rates matches the data, and model and data series move closely together over time.

The right panel of Figure 3, shows the weighted average outstanding coupon in the data and model by group.[30] The fit is once again good, considering the simplicity of the model. In particular, the model matches the substantial fanning out of coupons across groups over time and broadly matches the time-series patterns by group.

6.3 Mortgage rates and redistribution

We next study the quantitative impact of attention heterogeneity on mortgage rates and redistribution. The left panel of Figure 4 compares the mortgage rate $m(r, G)$ in the pooling MPE (in solid blue) to mortgage rates in a counterfactual separating equilibrium. Specifically, we plot the separating MPE mortgage rates $m(r, \chi)$ for the slowest (dot-dashed purple) and fastest (dashed red) borrowers identified in Section 5.2.[31] This figure shows that the fastest borrowers pay much lower and the slowest borrowers pay much higher rates at origination in the current pooling equilibrium than they would in a counterfactual separating equilibrium. For example, in the separating equilibrium average mortgage rates at origination are 290 bps higher for the fastest borrowers than for the slowest borrowers and are 230 bps higher than in the pooling MPE. These differences imply that fast borrowers receive significant cross-subsidies from slow borrowers at mortgage origination.

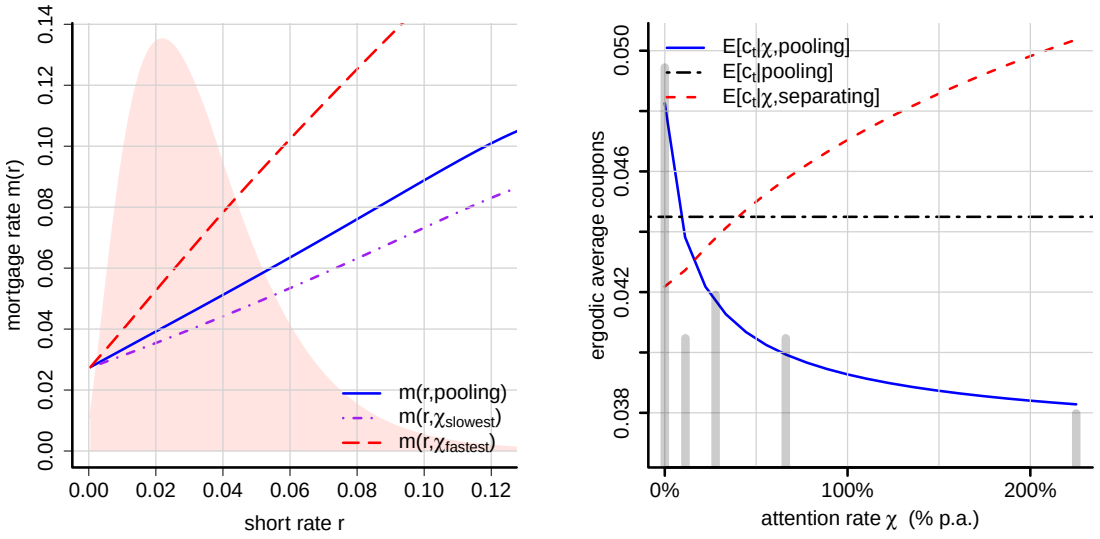
However, these comparisons capture only the cross-subsidies arising from the pooling equilibrium at the moment a mortgage is originated. They miss further dynamic redistribution arising from persistent differences in refinancing over time. To capture this dynamic redistribution, we plot in the right panel of Figure 4 the ergodic average coupons realized by borrowers over time rather than mortgage rates at origination.

[29] The fact that moving from unconditional to conditional beliefs has little pricing impact supports our claim that fully rational beliefs would complicate the model but likely not change conclusions.

[30] We impute coupons assuming an initial gap of zero at origination. Furthermore, we adjust the distribution of types in the model by the monthly weighted flows of households entering and exiting the panel. See Online Appendix C.1 for more detail.

[31] The separating MPE mortgage rates for intermediate attention levels are located between these two curves, since mortgage rates are increasing in attention (see Proposition 3).

Figure 4: Equilibrium mortgage rates and ergodic coupons.



Left panel shows equilibrium mortgage rates for (i) the pooling MPE $m(r, G)$ (solid blue), and (ii) the separating MPE $m(r, \chi)$ for type $\chi = \chi_1$ (dot-dashed purple) and $\chi = \chi_5$ (dashed red). Right panel shows the ergodic coupons as a function of attention χ for (i) the pooling MPE (solid blue) and its cross-sectional average (double-dashed black), and (ii) the separating MPE (dashed red).

The dot-dashed black line $\mathbb{E}[c_t|\text{pooling}]$ shows the ergodic average coupon for the population as a whole in the current pooling equilibrium. The solid blue line $\mathbb{E}[c_t|\chi, \text{pooling}]$ shows how ergodic average coupons vary by borrower type in this same equilibrium. The difference between blue and black lines thus captures redistribution arising in partial equilibrium ("PE") from heterogeneous attention, since it measures coupons across borrowers *within* the current pooling equilibrium where everyone pays the same rate at origination.

To capture distributional effects from equilibrium forces, we plot in dash red the ergodic average coupon $\mathbb{E}[c_t|\chi, \text{separating}]$ as a function of type χ in the counterfactual separating equilibrium. The ergodic coupon rises with attention χ in the separating equilibrium, mainly because lenders must sell newly issued mortgages for fast borrowers at a premium π to recoup their origination costs. The difference between black and red lines captures redistribution arising from equilibrium forces ("GE").

The difference between blue and red lines thus captures the total amount of redistribution arising from heterogeneous attention (PE + GE). Table 2 collects these effects under unconditional pricing for each borrow type and shows that redistribution arising through equilibrium changes in mortgage rates at origination (GE) is

of similarly large quantitative importance to that arising from ex post differences in refinancing (PE).[32] For example, the fastest borrowers pay 62 bps p.a. less than the average borrower in the pooling equilibrium, and they pay 121 bps less than they would pay in a counterfactual separating equilibrium.

Table 2: Redistribution by Borrower Type

Group	PE (bps p.a.)	GE (bps p.a.)	total (bps p.a.)
1 (slowest)	38	23	60
2	-7	17	10
3	-28	7	-22
4	-46	-13	-58
5 (fastest)	-62	-59	-121

Notes: This table shows how mortgage coupons vary with attention when fixing prices (PE) as well as compared to an alternative separating equilibrium (GE) under unconditional pricing. Specifically, the PE column computes $\mathbb{E}[c_t|\chi, \text{pooling}]$-$\mathbb{E}[c_t|\text{pooling}]$ and the GE column computes $\mathbb{E}[c_t|\text{pooling}]$ - $\mathbb{E}[c_t|\chi, \text{separating}]$ for each borrower type χ. A negative value indicates that a borrower receives subsidies and a positive value indicates that a borrower is taxed in the pooling equilibrium.

Redistribution across borrowers with different attention rates is substantial. How does this redistribution correlate with other directly observable characteristics? In Online Appendix D.1, we explore this by computing the cross-sectional correlation between (a) borrower i's assigned attention rate $\chi_{\alpha(i)}$ and (b) various borrower and zip-code level covariates. We find that lower-income, lower-FICO, and younger borrowers tend to be less attentive and this means they end up paying higher mortgage coupons on average than higher-income, higher-FICO and older borrowers who refinance more often. This suggests that the cross-subsidies we quantify are regressive, and go against the progressive subsidies embedded in the credit guarantee scheme provided by the GSEs.[33]

7 Policy evaluations and counterfactuals

In this section, we use our estimated model to evaluate the redistributive consequences in equilibrium of several alternative mortgage contract designs as well as effects of

[32] Equilibrium redistribution results for conditional pricing can be found in Online Appendix C.3.1.

[33] This progressivity has evolved over time. Before the advent of LLPA price adjustments, G-fees did not vary with credit scores, implying that high credit risk borrowers received subsidies from low credit risk borrowers. The implementation of LLPA price adjustments reduced these cross-subsidies.

information disclosures and trends in fintech lending.

7.1 Automatically refinancing mortgages

We begin by studying an automatically refinancing mortgage ("auto-RM"), as suggested by Campbell et al. (2011) and Keys, Pope, and Pope (2016). In an auto-RM, the borrower pays the minimum realized mortgage rate since its inception at time τ:

$$\underline{m}_t := \min_{\tau \leq s \leq t} \{m_s\}. \tag{17}$$

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 allows them to take advantage of rate reductions they would otherwise miss.[34] However, discussions around this proposal are usually cast in partial equilibrium and do not take into account the equilibrium response of mortgage rates. Our model can speak to this response. We first discuss the construction and some key theoretical properties of the auto-RM before studying it quantitatively.

7.1.1 Auto-RM theoretical results

We begin by making the following *smart-contract* assumption, which ensures the existence of an MPE in the auto-RM environment:

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

Under Assumption 3, a change in rates leads to a coupon reset just like under adjustable rate mortgages (ARM). Unlike an ARM, this adjustment 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 when borrowers move and take on a new mortgage at the then-current auto-RM rate, denoted (with an abuse of notation) $m(x, \infty)$.[35] We relegate all technical details to Online Appendix E.1.

We make three observations about our environment. First, the auto-RM is equivalent to an economic environment in which borrowers face no refinancing frictions,

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

[35] 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 use this notation for $\pi > 0$, even though no origination costs are incurred at rate reset under Assumption 3.

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.[36] Third, when starting from the same rate at origination, borrowers almost always pay less in an auto-RM than in the traditional ARM which adjusts to both rate decreases and increases.[37] This means that lenders must charge higher rates at origination for an auto-RM than for an ARM. We formalize this in Proposition 8, which we prove in Online Appendix E.2:

Proposition 8. *The auto-RM rate satisfies $m(x, \infty) \geq r(x) + f$ for all x.*

We now show that auto-RMs can lead to unraveling of traditional fixed-rate prepayable mortgages. Suppose that borrowers with heterogeneous χ all pool in a traditional mortgage and then consider the introduction of an auto-RM option. The slowest borrowers overpay for traditional mortgages in the pooling equilibrium. Thus, they find it beneficial to migrate to the auto-RM 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 mortgages increases, pushing those mortgage rates higher in equilibrium. The slowest remaining borrowers in the traditional mortgage now subsidize the fastest ones and so now they also leave, further raising effective attention in the traditional mortgage pool and pushing up traditional mortgage rates even more. Continuing this unraveling argument leads to:

Proposition 9. *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 eventually migrate to the auto-RM.*

What might prevent unraveling? First, some borrowers might not understand the value of the refinancing option embedded in the auto-RM. When faced with rates $m(r, G) < m(r, \infty)$, they might choose the cheaper initial rate of the traditional mortgage, even though it generates payments with higher expected net present value than the auto-RM. Second, this lower initial rate when paired with financial constraints

[36] In the in the auto-RM mortgage origination costs are incurred only when moving, and they are recovered by the originator through a sale of that auto-RM at market price $1 + \pi$.

[37] In our environment the ARM rate is simply $r(x) + f$.

might also lead borrowers to pick traditional mortgages even when they know 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. Finally, the auto-RM might not be the most desirable option if a borrower is risk averse, since this mortgage leads to coupons with lower expected value but higher variance.

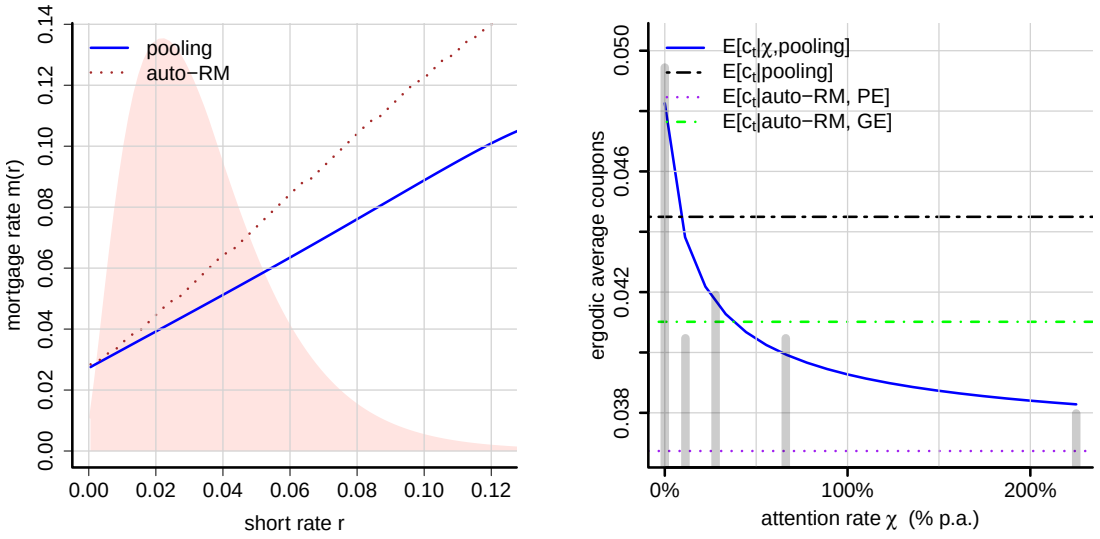
7.1.2 Auto-RM quantitative implications

Using model parameters from Section 6.1, we compute the equilibrium auto-RM mortgage rate at origination as a function of the short rate, and plot it in Figure 5. This figure shows that equilibrium mortgage rates in the auto-RM are systematically higher than in the pooling MPE, i.e., $m(r, \infty) \geq m(r, G)$. The ergodic average difference between these two rates is 110 bps, highlighting the substantial increase in mortgage rates at origination when moving from the traditional mortgage to this new financial instrument. Of course, the ultimate effect on coupons depends not just on rates at origination but also on refinancing over time.

The right panel of Figure 5 shows that ergodic average coupons in auto-RM mortgages (in green) are lower than in traditional mortgages (in black), despite the higher rates at origination. However, this equilibrium reduction in coupons from switching to auto-RM is much smaller than when holding mortgage prices fixed at $m(r, G)$ (in purple). This comparison highlights the need to factor in equilibrium responses when considering alternative contract designs: some alternatives that appear attractive when fixing prices may be less attractive after factoring in equilibrium responses. Indeed Figure 5 shows that the two fastest groups have a higher ergodic average coupon in the auto-RM equilibrium than in the traditional mortgage pooling MPE, so these borrowers are hurt by the introduction of the auto-RM. Slower borrowers still benefit from the auto-RM in equilibrium (blue vs. green) but by much less than when ignoring equilibrium effects (blue vs. purple).

In addition to implications for coupons over time, mortgage rates at origination matter directly for housing affordability, since lending constraints are based on characteristics at origination. To assess the effect of the equilibrium increase in mortgage interest rates at origination on borrowers' housing and mortgage choice, we plot in Online Appendix E.3 the debt-to-income (DTI) distribution at origination observed

Figure 5: Mortgage rates and ergodic coupons with auto-RM.



The left panel shows equilibrium mortgage rates for (i) the pooling MPE $m(r, G)$ (solid blue), and (ii) the auto-RM MPE $m(r, \infty)$ (dot brown). The right panel shows the ergodic coupons as a function of attention χ for (i) the pooling MPE (solid blue) and its cross-sectional average (double-dashed black), (ii) the auto-RM "PE" (dot purple) and (iii) the auto-RM "GE" (dot dash green) environments.

in SFLP data vs. the counterfactual DTI distribution from moving all borrowers to auto-RM mortgages. 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 18% of borrowers would be pushed above the DTI cutoff by a switch to auto-RM mortgages, potentially forcing them to downsize their house or increase their down payment upon purchase.

7.2 Mortgages with lockup periods

We next study mortgage contracts which prevent refinancing for some period of time after inception. It might seem that this "lockup" would only hurt borrowers, but we show that in equilibrium these contracts can both reduce mortgage inequality and improve efficiency by reducing origination costs induced by rapid refinancing.

To model this lockup period, we assume that a mortgage cannot be refinanced or prepaid for an exponentially distributed length of time with expected duration $1/\gamma$. Mortgage pricing now depends on the fraction of mortgages in the lockup period, but as we discuss in detail in Online Appendix E.4, it is straightforward to extend our

previous setup to this environment.

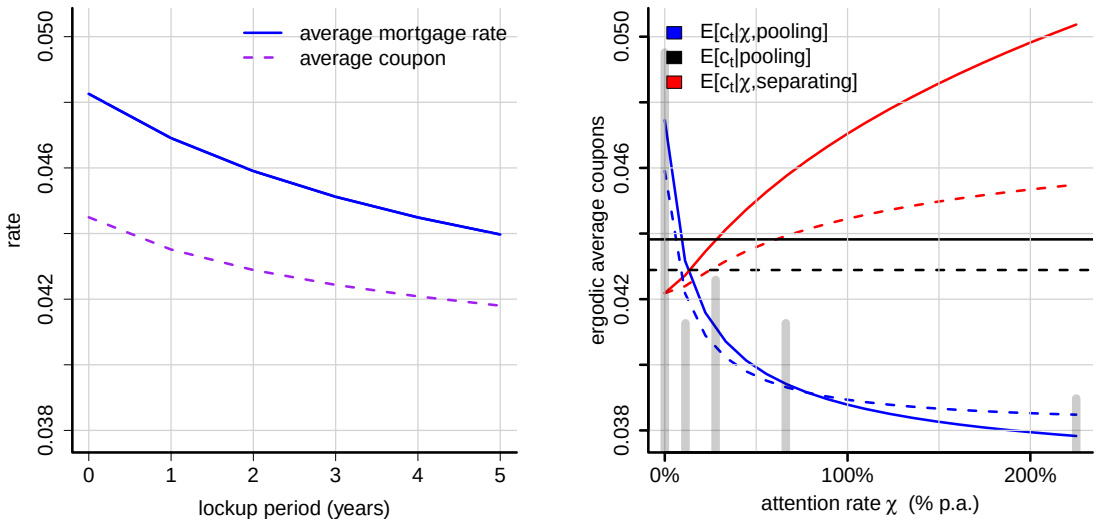
The left panel of Figure 6 shows that average mortgage interest rates at origination decline with the expected length of the lockup period and resulting coupons decline by almost the same amount. Rates at origination decline because expected origination costs decline and these savings are passed on to borrowers in the form of lower rates. The right panel of Figure 6 shows redistribution using a lockup period with expected length of $1/\gamma = 2$ years. The solid lines replicate the analysis of expected mortgage coupons under traditional mortgages from Figure 4. The dashed lines then repeat this analysis for the mortgages with lockup.

Comparing the solid and dashed blue lines shows that slow households benefit from a decrease in equilibrium rates caused by the lockup and reduce their average coupon payments by 16 bps. The fastest borrowers cannot utilize their refinancing option to the same extent, so although they benefit from lower mortgage rates at origination their average coupons nevertheless increase by a few bps. Finally, we note that this contract also has important implications for the strength of GE redistribution induced by pooling. This contract reduces refinancing costs, and this improved efficiency greatly reduces mortgage rates for fast borrowers in the counterfactual separating MPE. Since this contract reduces the need for price discrimination in the separating equilibrium, it also means that the pooling equilibrium provides a smaller relative benefit to fast households. That is, GE effects which compare the red to the black lines are much bigger for traditional mortgages (solid lines) than for mortgages with lockup (dashed lines).

7.3 Mandatory disclosure and attention shifters

It is also interesting to study the potential effects of mandatory information disclosure on equilibrium mortgage rates, since enhanced communication from lenders can improve borrowers' refinancing decisions. For example, Byrne et al. (2023) use an RCT to study the causal effect of various letters and reminders sent to Irish borrowers who could refinance their mortgage and generate savings. Over a 6-month period following this information disclosure, they find that borrowers' refinancing rates are 5.4 pp greater, corresponding to a 10.8 pp increase in (per annum) attention rates. Feeding a 10.8 pp increase in attention into our model leads ergodic average mortgage rates at origination to increase by approximately 35 bps.

Figure 6: Mortgage rates and ergodic coupons with lockup period.



The left panel shows equilibrium average mortgage rates (solid blue) and coupons (dash purple) as a function of the lockup period $1/\gamma$. The right panel shows the ergodic coupons as a function of attention χ for (i) the pooling MPE (blue) and its cross-sectional average (black) and (ii) the separating MPE (red). The solid (resp. dashed) lines correspond to the equilibrium with the standard mortgage contract (resp. the contract with a lockup period of $1/\gamma = 2$ years).

Attention rates can also be affected by trends in mortgage origination. For instance, the nonbank lender share of the conforming mortgage market has steadily increased over the past 20 years (Figure 7 left panel).[38] Using SFLP data, we show that there is differential prepayment as a function of rate gaps for mortgages originated by banks vs. nonbanks. We estimate the linear probability model

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

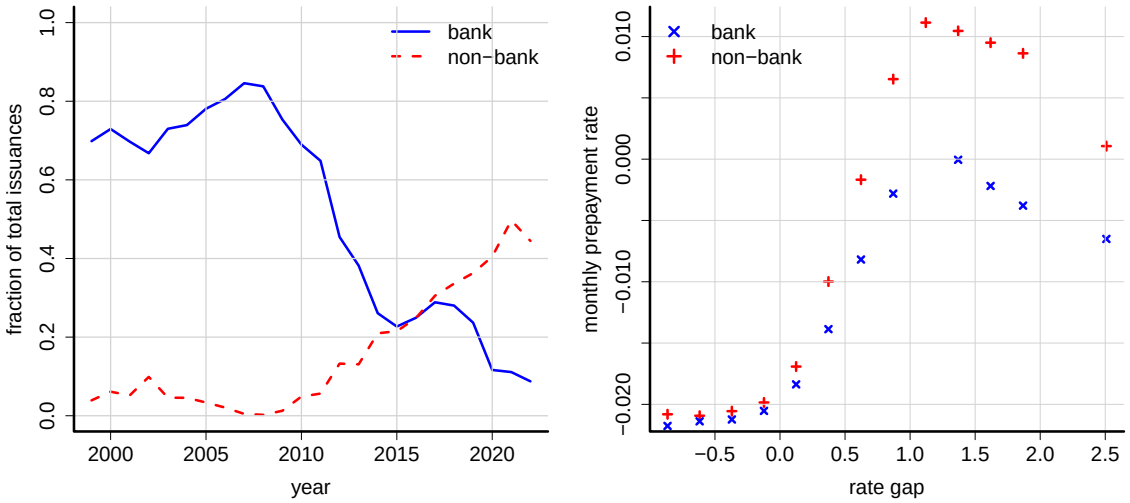
for borrower i with mortgage contract j at time t, where X is a vector of controls.[39] The right panel of Figure 7 shows point estimates for β^{bank}_{gapbin} and $\beta^{non-bank}_{gapbin}$ using 50

[38] Our computations use SFLP data and the bank vs. nonbank classification of Buchak et al. (2018). Sellers whose combined at-issuance unpaid principal balance is less than 1% of total issuances are classified as "others", so bank and nonbank origination shares do not sum up to 100%.

[39] Those controls are (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 borrower fixed effects in our regression. 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.

bps intervals for the gap bins. The difference in levels between negative and positive rate gaps directly give us the average attention rate for bank- and nonbank-originated mortgage borrowers. On average, borrowers with nonbank-originated mortgages are 100 bps per month more attentive than borrowers with bank-originated mortgages. This is a substantial difference. It is also quite similar to the increase in effective attention rate triggered by mandatory disclosure that we discussed above and so implies a similar increase in equilibrium mortgage rates when fed into our model.

Figure 7: Rise in nonbank lending.



The left panel shows the fraction of new mortgages classified as originated by "banks", "nonbanks" and "unknown" originators in the SFLP data. The right panel shows the point estimates for the rate gap bins for the bank (blue) and non-bank (red) originated mortgages in the SFLP sample.

8 Generalizing beyond mortgage markets

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. As illustrative examples, we discuss two additional settings in which our framework can be applied: the labor market and the small business credit market.

We develop these applications to show the generality of our framework but leave their precise quantitative and empirical analysis for future research.

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-offer rate* χ—the rate at which they receive offers from competing firms.[40] 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 they are hired.[41] 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 their 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 switching cost ψ.

Firms are risk neutral and competitive, with discount rate r. They value a worker with productivity x, wage w, and outside offer rate χ according to

$$\Pi(x, w; \chi) = \mathbb{E}_x \left[\int_0^\tau e^{-rt} (x_t - w)\, dt \right], \tag{18}$$

where τ is the quit time of a type-χ worker. With only idiosyncratic productivity shocks, there exists a well-defined stationary density $f_\infty(x, w, \chi)$ over workers' productivity x, wage rate w and type χ, and a corresponding stationary conditional type distribution of *job changers* $G(\chi|x)$.[42] When pursuing a prospective employee, the firm acts competitively and offers a wage $\mathcal{W}(x)$ that satisfies

$$\mathbb{E}^G \left[\Pi(x, \mathcal{W}(x); \chi) \right] = 0. \tag{19}$$

This is the counterpart to (12) in the mortgage market context, and it pins down the

[40] There are many reasons the arrival rate of outside offers differs even for workers with identical productivity. Some workers solicit outside offers more aggressively. Others have constraints related to children or spousal employment that lead employers to view them as less "movable." Pooling is likely to arise in part because it is illegal to set wages based on many of these characteristics.

[41] 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.

[42] This statement assumes the equilibrium wage rate $\mathcal{W}$ is monotonically increasing in productivity, which can be verified ex post when computing the equilibrium numerically.

equilibrium wage $\mathcal{W}$. The expectation in (19) encodes firms' inability to discriminate based on type χ—due to either information asymmetry or anti-discrimination laws covering characteristics correlated with χ. A pooling MPE is then defined by (i) workers optimally switch firms subject to their search and job-hunting frictions, (ii) firms' profits satisfy (18), and (iii) the equilibrium wage rate satisfies (19).

This environment could thus be used to analyze the impact of workers' cross-sectional heterogeneity on equilibrium wages and on the implicit cross-subsidies that aggressive job hunters receive from loyal workers via the labor market. Using linked employer-employee data sets, it would be straightforward to measure job-to-job transitions and wages within industries to discipline the model and explore quantitative implications of this cross-sectional heterogeneity.

In Online Appendix F, we further apply our framework to the small business credit market by building a model 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 in the data. This application of our modeling framework leads to predictions about 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).

Many other applications could also be adapted to our framework. While assumptions and modeling choices must be tailored to the particular applications, many of our insights are likely to apply. Our framework's tractability and ability to support systematic analysis of counterfactuals makes it attractive more broadly.

9 Conclusion

In this paper, we study the equilibrium consequences of pooling ex ante heterogeneous agents into long-term, frictional contracts with a competitive sector that cannot price-discriminate based on type. We apply this 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 implies a large degree of cross-sectional heterogeneity in borrowers' attention rates, leading us to estimate significant cross-subsidies through pooling. Furthermore, we show that many alternative mortgage contracts would have important equilibrium implications that can offset some of their benefits.

Policy discussions regularly take 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 these counterfactuals.

References

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

Achdou, Yves, Francisco J Buera, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. 2014. "Partial differential equation models in macroeconomics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372 (2028).

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

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

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).

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

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

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).

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

———. 2023. "Optimal Mortgage Refinancing with Inattention." Tech. rep., NBER.

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

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

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru. 2018. "Fintech, regulatory arbitrage, and the rise of shadow banks." *Journal of financial economics* .

Byrne, Shane, Kenneth Devine, Michael King, Yvonne McCarthy, and Christopher Palmer. 2023. "The Last Mile of Monetary Policy: Inattention, Reminders, and the Refinancing Channel." Tech. rep., NBER.

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

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

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

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

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).

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).

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).

Fleming, Wendell H and Halil Mete Soner. 2006. *Controlled Markov processes and viscosity solutions*, vol. 25. Springer Science & Business Media.

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., NBER.

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

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).

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

Huh, Yesol and You Suk Kim. 2023. "Cheapest-to-deliver pricing, optimal MBS securitization, and market quality." 150 (1).

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

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).

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

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

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

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

Strulovici, Bruno and Martin Szydlowski. 2015. "On the smoothness of value functions and the existence of optimal strategies in diffusion models." *Journal of Economic Theory* 159.

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

Online Appendix

A Borrowers' refinancing behavior

A.1 Value function V

Proof of Proposition 1. First, the borrower decision problem can be recast as follows:

$$V(x,c) := \inf_{k\in\mathcal{K}} \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi dN_t^{(k)}\right)\right], \tag{A.1}$$
$$\text{s.t.} \qquad dc_t^{(k)} = \left(m(x_t) - c_{t-}^{(k)}\right)\left(dN_t^{(k)} + dN_t^{(\nu)}\right),$$

where $\mathcal{K}$ is a set of progressively measurable intensity processes $k = \{k_t\}_{t\geq 0}$ such that $k_t \in [0,\chi]$ at all times. Using this definition, we first show that V must be increasing in c. Take $c' > c$ and an arbitrary intensity policy $k \in \mathcal{K}$. The difference in payoffs for such intensity policy k is:

$$\mathbb{E}_{x,c'}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi dN_t^{(k)}\right)\right] - \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi dN_t^{(k)}\right)dt\right]$$
$$\geq \mathbb{E}_x\left[\int_0^{\tau} e^{-\rho t}(c'-c)dt\right] > 0,$$

where $\tau > 0$ a.s. is the first refinancing time under policy k. Taking the infimum over all admissible policies yields

$$V(x,c') = \inf_{k\in\mathcal{K}} \mathbb{E}_{x,c'}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi dN_t^{(k)}\right)dt\right]$$
$$\geq \inf_{k\in\mathcal{K}} \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi dN_t^{(k)}\right)dt\right] = V(x,c)$$

Thus V is increasing in c. Moreover, since $\chi < +\infty$, a reasoning by contradiction can show that V is in fact strictly increasing in c. Problem (A.1) is a standard stochastic control problem, for which standard results apply. For instance, for one-dimensional diffusions, and subject to some technical conditions on the operator $\mathcal{L}$, Strulovici and Szydlowski (2015)[43] provide for the value function V being twice continuously differentiable in x, and satisfying the following HJB equation:

$$(\rho+\nu)\,V(x,c) = c + \mathcal{L}V(x,c) + \nu V(x,m(x)) + \min_{k\in[0,\chi]}\left\{k\left(V(x,m(x)) + \psi - V(x,c)\right)\right\}$$

[43] See also Fleming and Soner (2006); this latter article is not limited to one-dimensional diffusions, but includes additional – and more restrictive – conditions on the operator $\mathcal{L}$.

The optimal Markov control is $k^*(x,c) = \chi \mathbb{1}_{\{V(x,m(x))+\psi \leq V(x,c)\}}$. Since V is strictly increasing in c, this optimal policy can be re-written $k^*(x,c) = \chi \mathbb{1}_{\{c-m(x)\geq \theta(x)\}}$, for a rate gap cutoff $\theta(x)$ that satisfies

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

$\theta(x)$ is well defined since V is continuous and strictly increasing in c. Reinjecting the optimal Markov control into the HJB equation satisfied by V yields the HJB equation in the main text. □

B Mortgage market equilibrium

B.1 MPE existence and uniqueness in homogeneous case

Proof of Proposition 2. Discounted debt prices must be martingales, thus

$$r(x)P(x,c;\chi) = c - f + \mathcal{L}P(x,c;\chi) + \left(\nu + \chi \mathbb{1}_{\{c-m(x)\geq\theta(x)\}}\right)(1 - P(x,c;\chi)). \quad \text{(B.1)}$$

The function P, solution of (B.1), is implicitly dependent on a mortgage rate function $m(x)$, via the decision rule $\theta(x)$, which comes out of the borrower refinancing problem. It thus means that the equilibrium mortgage rate, implicitly defined via $P(x, m(x); \chi) = 1 + \pi$, is the outcome of a potentially complex fixed-point problem. Our proof has two steps; we first tackle the case $\pi = 0$, and then generalize to the case $\pi > 0$. In both cases, we assume no upfront closing costs (i.e. $\psi = 0$), and we assume that $r_t \in [\underline{r}, \bar{r}]$, with $0 \leq \underline{r} < \bar{r} < +\infty$, and $\chi < +\infty$. In that environment without upfront closing costs paid by borrowers, the decision rule simplifies to $\theta(x) = 0$, in other words the optimal intensity solves $k^*(x,c) = \chi \mathbb{1}_{\{c\geq m(x)\}}$.

i. In this section, we restrict ourselves to the case where $\pi = 0$. To make further progress, we study the auxiliary problem

$$\tilde{P}(x,c;\chi) := \inf_{k\in\mathcal{K}} \mathbb{E}_x \left[\int_0^{+\infty} e^{-\int_0^t (r(x_s)+k_s+\nu)ds} (c - f + k_t + \nu)\, dt\right], \quad \text{(B.2)}$$

where $\mathcal{K}$ is defined in Section A.1. The function $\tilde{P}$ does not depend, directly or indirectly, on any equilibrium object; in other words, one can view $\tilde{P}$ as the solution to a single-agent stochastic control problem. Arguments similar to those developed in Section A.1 allow us to argue that $\tilde{P}$ is twice continuously differentiable in x, continuous and increasing in c, satisfying the HJB equation

$$(r(x) + \nu)\,\tilde{P}(x,c;\chi) = c - f + \nu + \mathcal{L}\tilde{P}(x,c;\chi) + \min_{k\in[0,\chi]} \left\{k\left(1 - \tilde{P}(x,c;\chi)\right)\right\}.$$

The optimal Markov control is $\tilde{k}(x,c) = \chi \mathbb{1}_{\left\{\tilde{P}(x,c;\chi)\geq 1\right\}}$. Since r_t is restricted to

be on $\mathbb{R}_+$, we must have $\tilde{P}(x, 0; \chi) < 1$. Similarly, since r_t is bounded above by $\bar{r}$, for c sufficiently high we must have $\tilde{P}(x, c; \chi) > 1$. Since $\tilde{P}$ is continuous and increasing in c, by the intermediate value theorem there must exist a unique real value $c = m(x)$ that satisfies

$$\tilde{P}(x, m(x); \chi) = 1 \tag{B.3}$$

Given this construction, and given that $\tilde{P}$ is monotone in c, the set of events $\{\tilde{P}(x_t, c; \chi) \geq 1\}$ is identical to the set of events $\{m(x_t) \leq c\}$. We can then verify that the auxiliary function $\tilde{P}$ is none other than the pricing function P, and the mortgage rate function $m(x)$ defined via (B.3) is unique and satisfies the equilibrium condition $P(x, m(x); \chi) = 1$.

ii. We now consider the case $\pi > 0$ — i.e. the case where mortgage origination triggers costs, borne by lenders and recouped via higher mortgage rates. In this section, we also assume that the latent state x is one-dimensional and $r(\cdot)$ is increasing. We will prove that there exists a unique monotone equilibrium in that case — i.e. a unique MPE in which the mortgage rate function is monotone increasing in x. Take an arbitrary x^*, and define $\tau_{x^*,\chi}$ as a stopping time with arrival intensity $\nu + \chi \mathbb{1}_{\{x_t \leq x^*\}}$.[44] As will be seen shortly, x^* represents the latent state that was prevalent the last time a borrower refinanced. Consider the interest-only "IO" and principal-only "PO" net present values, defined via

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

These objects represents, respectively, the valuation of an IO and a PO whenever the latent state variable is x, and whenever the prepayment time is driven by a point process with (time-varying) intensity $\nu + \chi \mathbb{1}_{\{x_t \leq x\}}$. Introduce the function m, defined via

$$m(x) := f + \frac{1 - PO(x; \chi)}{IO(x; \chi)} + \frac{\pi}{IO(x; \chi)}. \tag{B.4}$$

m is continuous in x. We argue that m is a monotone increasing function of x, and that a monotone equilibrium exists, in which $m(x)$ is the equilibrium mortgage market interest rate. Consider first the special case $\pi = 0$. In that case, we know from the previous section (i) that an equilibrium exists and is unique. Since the objective in problem (B.2) is decreasing in x, it must be the case that the function $\tilde{P}$ defined in (B.2) is decreasing in x, which must mean that the equilibrium mortgage rate, when $\pi = 0$, is monotone increasing in x. In that case, the mortgage rate function must correspond to that defined in

[44] Formally, if ω is a (unit mean) exponentially distributed random variable and if we introduce the compensator $\Lambda_t = \int_0^t (\nu + \chi \mathbb{1}_{\{x_s \leq x^*\}}) ds$, then the stopping time $\tau_{x^*,\chi}$ is the (random) time that satisfies $\Lambda_{\tau_{x^*,\chi}} = \omega$.

(B.4) (with $\pi = 0$)—this is the case since

$$\begin{aligned}\tilde{P}\left(x, m(x); \chi\right) = 1 &= \mathbb{E}_x\left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds}(m(x) - f)dt + e^{-\int_0^{\tau_{x,\chi}} r_s ds}\right] \\ &= (m(x) - f)IO(x;\chi) + PO(x;\chi),\end{aligned}$$

which directly implies (B.4) for $\pi = 0$. As $m(x)$ is increasing when $\pi = 0$, it must be the case that $\left(1 - PO(x;\chi)\right)/IO(x;\chi)$ is increasing in x. For $\pi > 0$, we additionally need to show that $1/IO(x;\chi)$ is increasing in x. To this end, note that for $x_1 < x_2$, we must always have

$$\mathbb{E}_{x_2}\left[\int_0^{\tau_{x_2,\chi}} e^{-\int_0^t r_s ds}dt\right] \leq \mathbb{E}_{x_1}\left[\int_0^{\tau_{x_2,\chi}} e^{-\int_0^t r_s ds}dt\right] \leq \mathbb{E}_{x_1}\left[\int_0^{\tau_{x_1,\chi}} e^{-\int_0^t r_s ds}dt\right].$$

The first inequality above stems from the fact that if the initial interest rate is $r(x_1)$, the full time path of future interest rates is below that which would be relevant if the initial interest rate was $r(x_2)$. The second inequality stems from the fact that, for a given starting level of the latent state x_1, we must have the stopping time inequality $\tau_{x_2,\chi} \leq \tau_{x_1,\chi}$ almost surely. In other words, $IO(x;\chi)$ must be decreasing in x. This allows us to conclude that m, defined in (B.4), is monotone increasing in x. Given this observation, we must have an equilibrium in which m is the mortgage rate, since m must satisfy

$$1 + \pi = \mathbb{E}_x\left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds}(m(x) - f)dt + e^{-\int_0^{\tau_{x,\chi}} r_s ds}\right]$$

That equilibrium is unique, since we showed its existence by construction. In other words, in any monotone equilibrium, it must be the case that the mortgage rate function satisfies (B.4).

□

B.2 Comparative statics

Proof of Proposition 3. Consider first the case $\pi = 0$. Since $P = \tilde{P}$ can be defined via equation (B.2), it must be the case that P is decreasing in χ. Thus, the mortgage rate $m(x)$, defined implicitly via (B.3), is increasing in χ, whenever $\pi = 0$. Consider then the case where $\pi > 0$, and where the latent state x is one-dimensional and $r(\cdot)$ is increasing. Given our conclusion for the case $\pi = 0$, it must be the case that $\left(1 - PO(x;\chi)\right)/IO(x;\chi)$ is increasing in χ. Define (with a slight abuse of notation) $IO(x, x^*;\chi) := \mathbb{E}_x\left[\int_0^{\tau_{x^*,\chi}} e^{-\int_0^t r_s ds}dt\right]$, which solves the PDE

$$\left(r(x) + \nu + \chi \mathbb{1}_{\{x \leq x^*\}}\right) IO(x, x^*;\chi) = 1 + \mathcal{L}IO(x, x^*;\chi)$$

Differentiate this equation w.r.t. χ to obtain

$$\left(r(x)+\nu+\chi\mathbb{1}_{\{x\leq x^*\}}\right)\partial_\chi IO(x,x^*;\chi)=-\mathbb{1}_{\{x\leq x^*\}}IO(x,x^*;\chi)+\mathcal{L}\partial_\chi IO(x,x^*;\chi)$$

Thus, $\partial_\chi IO(x,x^*;\chi)$ admits the integral representation

$$\partial_\chi IO(x,x^*;\chi)=-\mathbb{E}_x\left[\int_0^{\tau_{x^*,\chi}}e^{-\int_0^t r_s ds}\mathbb{1}_{\{x_t\leq x^*\}}IO(x_t,x^*;\chi)dt\right]<0$$

Thus, $IO(x;\chi)$ is monotone decreasing in χ. This must mean that the mortgage rate function, defined via (B.4), is increasing in χ, whenever $\pi>0$. □

B.3 Small fixed costs

Proof of Proposition 4. Suppose an environment where upfront closing costs ψ are small, and where the gain on sale is $\pi=0$. Postulate an MPE in which mortgage prices P, borrowers' optimal rate gap threshold θ, and equilibrium mortgage rates m can be written:

$$\begin{aligned}P(x,c)&\underset{\psi\to 0}{=}P_0(x,c)+\psi P_1(x,c)+o(\psi)\\ \theta(x)&\underset{\psi\to 0}{=}\theta_0(x)+\psi\theta_1(x)+o(\psi)\\ m(x)&\underset{\psi\to 0}{=}m_0(x)+\psi m_1(x)+o(\psi),\end{aligned}$$

where P_0,θ_0,m_0 are respectively mortgage prices, borrowers' optimal rate gap threshold, and equilibrium mortgage rates in the MPE where $\psi=0$. We know from our previous analysis that $\theta_0=0$. Consider the PDE satisfied by mortgage prices:

$$\left(r+\nu+\chi 1_{\{c-m(x)\geq\theta(x)\}}\right)P(x,c)=c-f+\nu+\chi 1_{\{c-m(x)\geq\theta(x)\}}+\mathcal{L}P(x,c)$$

Let $\epsilon_1(x):=m_1(x)+\theta_1(x)$, we then have

$$\begin{aligned}1_{\{c-m(x)\geq\theta(x)\}}=1_{\{c-m_0(x)\geq\psi\epsilon_1(x)\}}&=1_{\{c-m_0(x)\geq 0\}}+1_{\{c-m_0(x)\geq\psi\epsilon_1(x)\}}-1_{\{c-m_0(x)\geq 0\}}\\ &=1_{\{c-m_0(x)\geq 0\}}-\frac{\epsilon_1(x)}{|\epsilon_1(x)|}1_{\{|c-m_0(x)|\in[0,\psi|\epsilon_1(x)|]\}}\end{aligned}$$

This allows us to write the zero-order expansion of the mortgage price as follows:

$$\left(r+\nu+\chi 1_{\{c-m_0(x)\geq 0\}}\right)P_0(x,c)=c-f+\nu+\chi 1_{\{c-m_0(x)\geq 0\}}+\mathcal{L}P_0(x,c)$$

For the first order expansion of the mortgage price, first note that we have, whenever c is in a neighbourhood of $m_0(x)$:

$$P_0(x,c) \underset{c\to m_0(x)}{=} 1 + (c - m_0(x))\,\partial_c P_0\left(x, m_0(x)\right) + o\left(|c - m_0(x)|\right),$$

where we have used the equilibrium condition $P_0\left(x, m_0(x)\right) = 1$. Note also that whenever $|c - m_0(x)| \in [0, \psi|\epsilon_1(x)|]$, there exists a $k_\psi(x,c) \in [0,1]$ s.t. $c = m_0(x) + \psi k_\psi(x,c)\epsilon_1(x)$. Whenever $|c - m_0(x)| \notin [0, \psi|\epsilon_1(x)|]$, set $k_\psi(x,c) = 0$. The first order correction term then satisfies:

$$\begin{aligned}\left(r + \nu + \chi 1_{\{c - m_0(x) \geq 0\}}\right) P_1(x,c) &= \frac{\chi\epsilon_1(x)}{|\epsilon_1(x)|} 1_{\{c - m_0(x) \in [0, \psi|\epsilon_1(x)|]\}} \frac{P_0(x,c) - 1}{\psi} + \mathcal{L}P_1(x,c) \\ &= \chi k_\psi(x,c)|\theta_1(x) + m_1(x)|\partial_c P_0\left(x, m_0(x)\right) + \mathcal{L}P_1(x,c)\end{aligned}$$

Finally, as $\psi \to 0$, $k_\psi(x,c)$ is a bounded function that is non-zero on an interval with measure proportional to ψ. In other words, the first order correction term P_1 satisfies

$$\left(r + \nu + \chi 1_{\{c - m_0(x) \geq 0\}}\right) P_1(x,c) = \mathcal{L}P_1(x,c)$$

Since the source term is identically zero, we conclude that $P_1(x,c) = 0$. Lastly, the break-even condition can be written:

$$P\left(x, m(x)\right) = 1 = P_0\left(x, m_0(x)\right) + \psi P_1\left(x, m_0(x)\right) + \psi m_1(x)\partial_c P_0\left(x, m_0(x)\right) + o(\psi)$$

Thus, we have

$$m_1(x) = -\frac{P_1\left(x, m_0(x)\right)}{\partial_c P_0\left(x, m_0(x)\right)}$$

Since $P_1(x,c) = 0$, we can conclude that the first order correction term $m_1(x) = 0$. In other words, whenever $\pi = 0$, fixed costs have no impact (at the first order) on the equilibrium mortgage rate function. This analysis is supported by our numerical computations.

□

B.4 Infinite dimensional problem with heterogeneity

In this section, we discuss the key mathematical equations characterizing the pooling MPE. As a reminder, $H(\chi)$ denotes the cumulative distribution over types (with associated density h), while F_t denotes the joint cumulative distribution over outstanding coupon rates c and types χ in the population at time t (with associated joint density

$f_t(c,\chi)$). Since types are a permanent borrower attribute, we must have

$$\int_c f_t(c,\chi)dc = h(\chi). \tag{B.5}$$

Consider then the density f_t. It evolves endogenously over time with idiosyncratic mortgage refinancing decisions, which, aggregated using a weak law of large numbers, lead to locally deterministic movements in f_t. The Kolmogorov Forward Equation ("KFE") that describes these changes is then, for $c \neq m(S)$:

$$df_t(c,\chi) = -\left(\nu + \chi \mathbb{1}_{\{c \geq m(S_t)\}}\right) f_t(c,\chi)dt, \qquad c \neq m(S). \tag{B.6}$$

The density f_t, between t and $t+dt$, looses mass at rate ν for $c < m(S_t)$, and at the higher rate rate $\nu + \chi$ for $c \geq m(S_t)$, as borrowers strategically refinance. This equation holds everywhere except at $c = m(S_t)$, a state at which refinancing and moving borrowers are being "reinjected"; the relevant equation in that case is

$$\lim_{c \uparrow m(S_t)} \partial_c f_t(c,\chi) - \lim_{c \downarrow m(S_t)} \partial_c f_t(c,\chi) = \nu h(\chi) + \chi \int_{m(S_t)}^{+\infty} f_t(c,\chi)dc. \tag{B.7}$$

The right-hand-side of this equation is the flux of borrowers exogenously moving at rate ν and the flux of type-χ borrowers refinancing in the time interval $[t, t+dt]$, while the left-hand-side is the kink in the density at $c = m(S)$ induced by the reinjection of such borrowers at that particular point of the state space.

Let $P(S,c;\chi)$ be the *shadow* price of a mortgage with coupon c, conditional on knowing that the related borrower has attention rate χ. The shadow price solves the following infinite dimensional Feynman-Kac equation, which takes into account (i) changes in the distribution f_t, and (ii) the behavior of type-χ borrowers:

$$\begin{aligned} r(x)P(S,c;\chi) = c + \mathcal{L}P(S,c;\chi) + \left(\nu + \chi \mathbb{1}_{\{c \geq m(S)\}}\right)[1 - P(S,c;\chi)] \\ + \int \mathcal{T}[f](c,\chi)\frac{\delta P}{\delta f(c,\chi)}dcd\chi, \end{aligned} \tag{B.8}$$

with $\delta P/\delta f$ the functional derivative of P w.r.t. f at (c,χ) and the operator $\mathcal{T}$ defined via

$$\mathcal{T}[f](c,\chi) = -\left(\nu + \chi \mathbb{1}_{\{c > m(S)\}}\right) f(c,\chi) \tag{B.9}$$

See Achdou et al. (2014) for another example of such infinite-dimensional PDE in the context of consumption-savings models in incomplete markets with aggregate shocks.

B.5 Approximate pooling MPE existence and uniqueness

Proof of Proposition 5. We establish the existence and uniqueness of the approximate pooling MPE using a method similar to Section B.1 for the case $\pi > 0$. To that effect, consider the dynamic system $(x_t, x^*_{t,\chi})$, where

$$\begin{aligned} dx_t &= \mu(x_t)dt + \sigma(x_t)dB_t \\ dx^*_{t,\chi} &= \left(x_t - x^*_{t-,\chi}\right) dN^\chi_t, \end{aligned}$$

where N^χ_t is a point process with arrival intensity $\nu + \chi \mathbb{1}_{\{x_t \leq x^*_{t-,\chi}\}}$. This dynamic system admits a generator $\mathcal{H}_\chi$ defined for any smooth function $\phi(x, x^*)$ via

$$\mathcal{H}_\chi \phi(x, x^*) = \mathcal{L}\phi(x, x^*) + \left(\nu + \chi \mathbb{1}_{\{x \leq x^*\}}\right) \left(\phi(x, x) - \phi(x, x^*)\right)$$

The eigen-function (associated with the eigen-value zero) of the adjoint of the operator $\mathcal{H}_\chi$ gives us the stationary density $f_\infty(x, x^*|\chi)$ of the dynamic system $(x_t, x^*_{t,\chi})$. Introduce then the distribution g, either the unconditional one defined via

$$g(\chi) = \frac{h(\chi) \int_x \left[\left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*|x, \chi)dc\right)\right] f_\infty(x)dx}{\int_\chi h(\chi) \int_x \left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*|x, \chi)dc\right) f_\infty(x)d\chi dx}, \tag{B.10}$$

or the conditional one defined via

$$g(\chi|x) = \frac{h(\chi) \left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*, x|\chi)dc\right)}{\int_\chi h(\chi) \left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*, x|\chi)dc\right) d\chi}. \tag{B.11}$$

Define the candidate mortgage rate $m(x; G)$ via

$$m(x; G) := f + \frac{1 + \pi - \mathbb{E}^G \left[PO(x; \chi)\right]}{\mathbb{E}^G \left[IO(x; \chi)\right]}, \tag{B.12}$$

If the function $m(x; G)$ is increasing in x, a monotone approximate pooling MPE must exist, and this equilibrium is unique amongst all monotone equilibria. Note that m must satisfy

$$1 + \pi = \mathbb{E}^G \left[\mathbb{E}_x \left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds}(m(x) - f)dt + e^{-\int_0^{\tau_{x,\chi}} r_s ds}\right]\right]$$

Consider then the price $\bar{P}_G\left(x, m(x^*)\right)$ of a mortgage with coupon $m(x^*)$,

$$\bar{P}_G(x, m(x^*)) := \mathbb{E}^G \left[\mathbb{E}_x \left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds}(m(x^*) - f)dt + e^{-\int_0^{\tau_{x,\chi}} r_s ds}\right]\right],$$

then clearly if m is increasing, $\bar{P}_G$ must be increasing in x^*, with $\bar{P}_G(x^*, m(x^*)) = 1+\pi$ – in other words the equilibrium conditions are satisfied. □

B.6 Integral representation of $\bar{P}_G$ for unconditional $G(\chi)$

Proof of Proposition 6. (B.1) holds for all χ, and thus, taking expectations w.r.t. the unconditional issuance type distribution $G(\chi)$, we have

$$r(x)\bar{P}_G(x,c) = c - f - \mathbb{1}_{\{m(x)\leq c\}}\mathbb{C}\text{ov}^G(\chi, P(x,c;\chi)) + \mathcal{L}\bar{P}_G(x,c) + \left(\nu + \bar{\chi}_G \mathbb{1}_{\{m(x)\leq c\}}\right)\left(1 - \bar{P}_G(x,c)\right) \quad \text{(B.13)}$$

One can then use Feynman-Kac to conclude that $\bar{P}_G$ admits the integral representation in Proposition 6. □

B.7 Invariance of lowest attainable mortgage rate

Proof of Proposition 7. Under the assumption that x is uni-dimensional and that $r(\cdot)$ is monotone increasing, call $\underline{x}$ the lowest bound for x. The monotone approximate pooling MPE implies m is increasing in x, and thus $m(\underline{x})$ must be the lowest attainable mortgage rate. Then we have

$$P(x, m(\underline{x}); \chi) = P(x, m(\underline{x}); \chi'), \qquad \forall \chi, \chi',$$

as χ only influences the refinancing channel, and whenever borrowers have locked in the lowest possible rate $c = m(\underline{x})$, we have $c \leq m(x_t)$ at all future date t regardless of type χ. Thus, from the break-even condition $\bar{P}(x, m(x)) = 1 + \pi$, we have

$$P(\underline{x}, m(\underline{x}); \chi) = 1 + \pi, \qquad \forall \chi.$$

$m(\underline{x})$ is thus invariant to the distribution G, and thus the distribution H. □

C Borrower attention in mortgage prepayment data

C.1 Data

We rely on information from Equifax Credit Risk Insight Servicing McDash ("CRISM"). This monthly-frequency data-set covers the period from May 2005 until June 2017. It contains unique borrower IDs, mortgage IDs, a prepayment indicator if a loan prepaid in a given month, the original coupon rate on the loan, its current principal balance, as well as the current FICO score of the related borrower. We build an indicator describing the type of prepayment (prepay, rate refinancing, cash-out refinancing or moves), and a measure of the current combined loan-to-value ratio

(thereafter, "CTLV") using house price data from Corelogic. For each borrower and each month, the effective mortgage market rate available to a borrower is simply assumed to be the FRM30 rate. Further, we impose based on reported origination date τ of a mortgage a "synthetic" coupon equal to the FRM30 rate that date, m_τ. Lastly, we backfill the sample for any loan that does enters our sample later than its origination date, knowing that it has not prepayed or defaulted until this entering date. This allows us to construct the "synthetic" rate gap – i.e. the difference between (i) the average mortgage coupon at the origination date and (ii) the current effective mortgage market rate.[45] Our data-set allows us to track a borrower and their different mortgages through time. It contains 20,094,230 loan-month-borrower observations, with 246,330 unique borrower IDs. We restrict CRISM to GSE loans of 30 year original maturity, and drop all loan-months with a principal balance of less than \$50k and with missing origination dates, while backfilling for those loans with origination dates. This leaves use with 131,652 unique borrower IDs, and 193,812 unique loans, for a total of 10,476,103 loan-months, of which 9,049,368 are original (non-backfilled observations). The mean principal balance amount in this restricted data-set is \$191,153. Via MLE, we allocate 115,493 borrower IDs to groups, with the remainder, 16,159 borrower IDs, un-allocated due to them never having a positive rate gap. The un-allocated borrowers account for 386,690 loan-months observations.

Our model assumes a constant distribution of attention in the population. In contrast, the data is unbalanced and the distribution of attention changes as borrowers enter and leave the panel. Thus, to maintain a fair comparison between model and data in time-series comparisons, when considering time-series outputs (e.g., monthly average coupon), we adjust the distribution of types in the model by the monthly weighted flows of households entering and existing the panel.

For some of our econometric work, we also leverage the single-family loan performance ("SFLP") data-set from Fannie-Mae. While CRISM allows us to track individual borrowers across loans, SFLP only allows us to track monthly mortgage performance data for a sample of conforming loans originated between January 2000 and December 2021. This means that the SFLP data cannot distinguish refinancing from other types of prepayment. However, this data is nevertheless useful since it contains covariates which are absent in CRISM — for instance the identity of the original lender and of the mortgage servicer.

[45] An alternative approach would be to use the reported coupon on each mortgage, and use borrower characteristics to predict which rate a borrower would be offered to construct the rate gap. Two issues arise: 1) we do not have dynamic characteristics for the backfilled portion of the loan, and more importantly, 2) some borrowers get uniformly better rates independent of dynamic characteristics. As this latter household "attribute" is assumed static, the "synthetic" rate gap is a consistent estimator of the actual rate gap estimtator facing a household when ignoring dynamic characteristics.

C.2 Estimating the attention distribution H

For our clustering algorithm, we fix N, the number of groups. Whether a borrower belongs to one group or another is determined via maximum likelihood — we note $\alpha : \{1, \dots, N_h\} \to \{1, \dots, N\}$ the group assignment function. It is easier to state this optimization in terms of attention probability per month p, rather than in terms of attention rate χ.

For each borrower i, we associate a binomial random variable where the number of "successes" is the number of observed prepayment events, and the number of "trials" is the number of observed monthly periods. Let s_i^+ and t_i^+ (resp. s_i^- and t_i^-) be the number of successes and trials when rate gaps satisfy $gap_{it} > \theta$ (resp. $gap_{it} \leq \theta$). Let y_{it} be an indicator of borrower i prepaying at time t, and define the gap-dependent monthly prepayment probabilities

$$p^+_{\alpha(i)} := 1 - \exp\left(-(\nu + \chi_{\alpha(i)})dt\right) \qquad p^-_{\alpha(i)} := 1 - \exp\left(-\nu dt\right)$$

Then, the log-likelihood of observing y_{it} is given by

$$\begin{aligned}\mathcal{L}(y_{it}) = y_{it} &\left(\mathbb{1}_{\{gap_{it}>\theta\}} \log p^+_{\alpha(i)} + \mathbb{1}_{\{gap_{it}\leq\theta\}} \log p^-_{\alpha(i)}\right) \\ &+ (1 - y_{it}) \left(\mathbb{1}_{\{gap_{it}>\theta\}} \log(1 - p^+_{\alpha(i)}) + \mathbb{1}_{\{gap_{it}\leq\theta\}}(1 - \log p^-_{\alpha(i)})\right)\end{aligned}$$

Let $t_i := t_i^- + t_i^+$ be the total number of months borrower i is in the sample. Then the log-likehood of observing a sequence $(y_{i1}, ..., y_{it_i})$ is given by

$$\begin{aligned}\mathcal{L}(y_{i1}, ..., y_{it_i}) &= \sum_{t=1}^{t_i} \left(y_{it}\mathbb{1}_{\{gap_{it}>\theta\}}\right) \log p^+_{\alpha(i)} + \sum_{t=1}^{t_i} \left(y_{it}\mathbb{1}_{\{gap_{it}\leq\theta\}}\right) \log p^-_{\alpha(i)} \\ &+ \sum_{t=1}^{t_i} \left((1 - y_{it})\mathbb{1}_{\{gap_{it}>\theta\}}\right) \log(1 - p^+_{\alpha(i)}) + \sum_{t=1}^{t_i} \left((1 - y_{it})\mathbb{1}_{\{gap_{it}\leq\theta\}}\right) (1 - \log p^-_{\alpha(i)}) \\ &= s_i^+ \log p^+_{\alpha(i)} + s_i^- \log p^-_{\alpha(i)} + (t_i^+ - s_i^+) \log(1 - p^+_{\alpha(i)}) + (t_i^- - s_i^-) \log(1 - p^-_{\alpha(i)})\end{aligned}$$

where we used the identities

$$\begin{aligned} s_i^+ &= \sum_{t=1}^{t_i} (1 - y_{it})\mathbb{1}_{\{gap_{it}>\theta\}}, & s_i^- &= \sum_{t=1}^{t_i} y_{it}\mathbb{1}_{\{gap_{it}\leq\theta\}} \\ t_i^+ &= \sum_{t=1}^{t_i} \mathbb{1}_{\{gap_{it}>\theta\}}, & t_i^- &= \sum_{t=1}^{t_i} \mathbb{1}_{\{gap_{it}\leq\theta\}} \end{aligned}$$

Of course, there are several sequences $(y_{i1}, ..., y_{it_i})$ that results in the same $(s_i^+, s_i^-, t_i^+, t_i^-)$. Thus, for the total log-likehoood we have to sum over all the appropriate permutation, which results in a factor similar to $\binom{a}{b}$, the binomial factor. However, we note that

this factor is independent of the choices of $p^+_{\alpha(i)}$ and $p^-_{\alpha(i)}$, the per-month probabilities, and thus does not affect our ML estimation. Given values $\{s_i^+, s_i^-, t_i^+, t_i^-\}_{i \leq N_h}$, we want to estimate the maximum likelihood that those observations were generated by $N+1$ prepayment probabilities $(p_0, ..., p_N)$, where p_0 represents the prepayment probability conditional on $gap \leq \theta$, while $p_1, ..., p_N$ represent N prepayment probabilities conditional on $gap > \theta$. Of course we need to insure that $p_k \geq p_0$, for $k \geq 1$. If $\mathbf{P} := \{(p_0, ..., p_N) : p_k \in [0,1] \, \forall k, p_k \geq p_0 \forall k \geq 1\}$, then the maximum log-likelihoood of the data, for a given $(p_0, ..., p_N) \in \mathbf{P}$, is

$$L(\mathbf{P}) = \sum_{i=1}^{N_h} \max_{p^+ \in \{p_1, ..., p_N\}} \left[s_i^+ \log p^+ + s_i^- \log p_0 \right.$$
$$\left. + (t_i^+ - s_i^+) \log(1 - p^+) + (t_i^- - s_i^-) \log(1 - p_0) \right\} \quad \text{(C.1)}$$

To get the MLE, we maximize $L(\mathbf{P})$ over the set $\mathbf{P}$.

C.3 Attention rate distribution

C.3.1 Full Baseline Estimation

Table C.1 reports our point estimate and standard errors for the discrete distribution H of attention types in our sample of borrowers, using our main group-based specification with $gap > 0.25\%$. It also shows the corresponding unconditional ergodic average origination distribution in our Approximate Pooling MPE. Table C.2 reports the corresponding ergodic redistribution results. Here, (vec) corresponds to unconditional pricing $m_t = m(x_t|G)$, while (mat) corresponds to conditional pricing $m_t = m(x_t|G(x_t))$. Thus, the columns marked (vec) simply replicate Table 2 in the main part of the paper.

C.3.2 Alternative specifiations for different refinancing thresholds

For alternative specifications, Table C.3 shows the MLE specification for $gap > 0\%$ and Table C.4 for $gap > 0.5\%$, and Table C.5 for $gap > 1\%$.

C.3.3 Alternative specifiations for different Original Loan Amounts

This section presents estimation (Table C.6) and redistribution results (Table C.7) when restricting the sample to loans with Original Loan Amount (OLA) of at least \$150,000 (previously there was no restriction), with $gap > 0.25\%$.

Group	$\hat{\chi}_i$ (per year)	$p(\hat{\chi}_i)$ (per month)	$H(\hat{\chi}_i)$	$G(\hat{\chi}_i)$	Std Error p gross
base	0.0382	0.0032	0.0	0.0	0.0
1 (slowest)	0.0	0.0	0.471	0.2686	NA
2	0.1342	0.0111	0.162	0.1496	0.0001
3	0.336	0.0276	0.1856	0.2291	0.0002
4	0.78	0.0629	0.1252	0.2089	0.0006
5 (fastest)	2.4897	0.1874	0.0562	0.1437	0.0025

Table C.1: Synthetic & Backfilled, threshold 25*bps*, OLA minimum 0k: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on borrowers and months with $gap > 0.25\%$, weighted by average loan amount.

Group	PE (vec) (bps p.a.)	GE (vec) (bps p.a.)	total (vec) (bps p.a.)	PE (mat) (bps p.a.)	GE (mat) (bps p.a.)	total (mat) (bps p.a.)
1 (slowest)	38	23	60	35	19	54
2	-7	17	10	-6	13	7
3	-28	7	-22	-26	3	-24
4	-46	-13	-58	-43	-17	-59
5 (fastest)	-62	-59	-121	-58	-63	-122

Table C.2: Synthetic & Backfilled, threshold 25*bps*, OLA minimum 0k: This table shows how mortgage coupons vary with attention when fixing prices (PE) as well as compared to an alternative separating equilibrium (GE) under unconditional (vec) and conditional (mat) pricing. Specifically, the PE column computes $\mathbb{E}[c_t|\chi, \text{pooling}]$- $\mathbb{E}[c_t|\text{pooling}]$ and the GE column computes $\mathbb{E}[c_t|\text{pooling}]$ - $\mathbb{E}[c_t|\chi, \text{separating}]$ for each borrower type χ. A negative value indicates that a borrower receives subsidies and a positive value indicates that a borrower is taxed in the pooling equilibrium.

Group	$\hat{\chi}_i$ (per year)	$p(\hat{\chi}_i)$ (per month)	$H(\hat{\chi}_i)$	$G(\hat{\chi}_i)$	Std Error p gross
base	0.0333	0.0028	0.0	0.0	0.0
1 (slowest)	0.0	0.0	0.4983	0.3015	NA
2	0.1098	0.0091	0.1294	0.1222	0.0001
3	0.2617	0.0216	0.1804	0.2223	0.0002
4	0.5691	0.0463	0.1389	0.2246	0.0004
5 (fastest)	1.7433	0.1352	0.053	0.1294	0.0018

Table C.3: Synthetic & Backfilled: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on borrowers and months with $gap > 0\%$, weighted by average loan amount.

Group	$\hat{\chi}_i$ (per year)	$p(\hat{\chi}_i)$ (per month)	$H(\hat{\chi}_i)$	$G(\hat{\chi}_i)$	Std Error p gross
base	0.0449	0.0037	0.0	0.0	0.0
1 (slowest)	0.0	0.0	0.4652	0.2559	NA
2	0.1637	0.0135	0.2017	0.1855	0.0001
3	0.4497	0.0368	0.1771	0.2266	0.0003
4	1.1398	0.0906	0.1066	0.1912	0.001
5 (fastest)	4.1002	0.2894	0.0494	0.1409	0.0041

Table C.4: Synthetic & Backfilled: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on borrowers and months with $gap > 0.5\%$, weighted by average loan amount.

Group	$\hat{\chi}_i$ (per year)	$p(\hat{\chi}_i)$ (per month)	$H(\hat{\chi}_i)$	$G(\hat{\chi}_i)$	Std Error p gross
base	0.0542	0.0045	0.0	0.0	0.0
1 (slowest)	0.0	0.0	0.4781	0.2508	NA
2	0.1889	0.0156	0.1827	0.1617	0.0002
3	0.5616	0.0457	0.1841	0.2337	0.0005
4	1.7274	0.1341	0.1112	0.2132	0.0016
5 (fastest)	7.4887	0.4642	0.0439	0.1406	0.0072

Table C.5: Synthetic & Backfilled: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on borrowers and months with $gap > 1\%$, weighted by average loan amount.

Group	$\hat{\chi}_i$ (per year)	$p(\hat{\chi}_i)$ (per month)	$H(\hat{\chi}_i)$	$G(\hat{\chi}_i)$	Std Error p gross
base	0.0381	0.0032	0.0	0.0	0.0
1 (slowest)	0.0	0.0	0.4412	0.2479	NA
2	0.1207	0.01	0.1493	0.132	0.0001
3	0.2947	0.0243	0.1938	0.2256	0.0002
4	0.6846	0.0555	0.15	0.2352	0.0005
5 (fastest)	2.2335	0.1698	0.0657	0.1593	0.0021

Table C.6: Synthetic & Backfilled, OLA minimum 150k: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on borrowers and months with $gap > 0.25\%$, weighted by average loan amount.

Group	PE (vec) (bps p.a.)	GE (vec) (bps p.a.)	total (vec) (bps p.a.)	PE (mat) (bps p.a.)	GE (mat) (bps p.a.)	total (mat) (bps p.a.)
1 (slowest)	39	24	63	37	20	56
2	-7	18	11	-6	14	7
3	-28	7	-22	-26	3	-24
4	-45	-13	-58	-42	-17	-59
5 (fastest)	-62	-58	-120	-58	-62	-120

Table C.7: Synthetic & Backfilled, threshold 25*bps*, OLA minimum 150k: This table shows how mortgage coupons vary with attention when fixing prices (PE) as well as compared to an alternative separating equilibrium (GE) under unconditional (vec) and conditional (mat) pricing. Specifically, the PE column computes $\mathbb{E}[c_t|\chi, \text{pooling}]$- $\mathbb{E}[c_t|\text{pooling}]$ and the GE column computes $\mathbb{E}[c_t|\text{pooling}]$ - $\mathbb{E}[c_t|\chi, \text{separating}]$ for each borrower type χ. A negative value indicates that a borrower receives subsidies and a positive value indicates that a borrower is taxed in the pooling equilibrium.

D Quantifying the redistribution

D.1 Attentions rates and covariates

In this section, we study the degree of redistribution amongst borrowers of different *observable* characteristics. To do this, we simply compute the cross-sectional correlation between observable characteristic X_i and borrower i's attention intensity $\chi_{\alpha(i)}$. We also look at a monotone transformation of this attention intensity — $1/(\nu+\chi_{\alpha(i)})$ — which is expressed in units of time. Some of the covariates are measured at the borrower level, and others are measured at the ZIP code level.[46] Table D.1 summarizes these correlations. Our results suggests that the cross-subsidies we are documenting are regressive — in the sense that lower income, smaller mortgage, and younger borrowers tend to be less attentive, and thus pay on average greater mortgage interest payments than higher income, larger mortgage and older borrowers.

covariate X_i	$\mathbb{C}\text{orrel}\left(\chi_{\alpha(i)}, X_i\right)$	$\mathbb{C}\text{orrel}\left(\frac{1}{\nu+\chi_{\alpha(i)}}, X_i\right)$
ficov5	0.1183	-0.2375
prin_bal_amt	0.102	-0.0992
CLTV	-0.1194	0.1425
multi_mortgage	0.0125	-0.2382
ho_rate	-0.0055	0.0037
less_hs	-0.028	0.0343
hs	-0.0084	0.0065
some_college	0.0047	-0.0053
bachelor_plus	0.0282	-0.0313
median_income	0.0547	-0.053
share_below_35	-0.0007	-0.0
median_age	0.0089	-0.0261
Population	-0.0499	0.1253
DensityPerSqMile	-0.0026	0.0145
white_share	0.0242	-0.0475

Table D.1: Synthetic & Backfilled: Correlation between attention and various covariates for baseline specification with $gap > 0.25\%$

[46] Borrower ZIP code level covariate is the average of the borrower's time series over the relevant zip code value.

D.2 Pricing errors

In this section we evaluate via simulation the pricing errors that arise from Assumption 2 — i.e. the assumption that investors assume either (a) a constant, or (b) a state-dependent origination distribution when pricing newly-issued mortgages. In order to compute investor pricing errors, we perform the following computations:

1. We simulate $T = 100$ million consecutive months for the interest rate process r_t, starting with $r_0 = \mathbb{E}[r_t]$;

2. Starting with an (arbitrary) distribution $f_0(c, \chi)$ over coupon and types[47], we compute, for our random path $\{r_t\}_{t\geq 0}$, the model-implied distribution f_t;

3. From f_t, we extract the origination distribution g_t as well as the prepayment flow mass $flow_t$.

4. Using the actual origination distribution G_t and the equilibrium mortgage rate $m(r_t, G)$ at time t, we derive the pricing error made by investors when originating mortgages at such time, i.e.

$$\mathbb{E}^{G_t}\left[P\left(r_t, m(r_t, G); \chi\right)\right] - (1 + \pi)$$

5. By construction, the flow-weighted mean expected pricing error converges to zero, since

$$\lim_{T\to\infty} \sum_t^T \frac{1}{\sum_s^T \mathbb{1}_{\{s\in\{\tau: r_\tau = r\}\}}} w_t \times \mathbb{E}^{G_t}\left[P\left(r_t, m(r_t, G(\chi|r_t)); \chi\right)\right] = 1 + \pi$$

   where

$$w_t := \frac{\mathbb{1}_{\{t\in\{\tau: r_\tau = r\}\}} flow_t}{\sum_s^T \mathbb{1}_{\{s\in\{\tau: r_\tau = r\}\}} flow_s}$$

6. Instead, the flow-weighted standard deviation of pricing errors converges to a state-dependent non-zero constant, depicted in Figure D.1.

The maximum conditional standard deviation is 90bps, and it is achieved at $r = 7.5\%$, while the standard deviation at the ergodic mean short rate is 77bps. The standard deviation vanishes at $r = \underline{r} = 0$: at this (lowest possible) value, as shown in to Proposition 8, the mortgage rate is invariant to the cross-sectional distribution. The standard deviation also vanishes at $r = \overline{r} = 14.5\%$, the upper bound on our state space, for similar reasons: at this level of interest rates, no borrower wants to strategically refinance, which means that the origination distribution must always be $G_t = H$. This standard deviation of pricing errors can be interpreted as mispricing

[47] In practice, we use the ergodic conditional distribution $f_\infty(c, \chi|r_0)$.

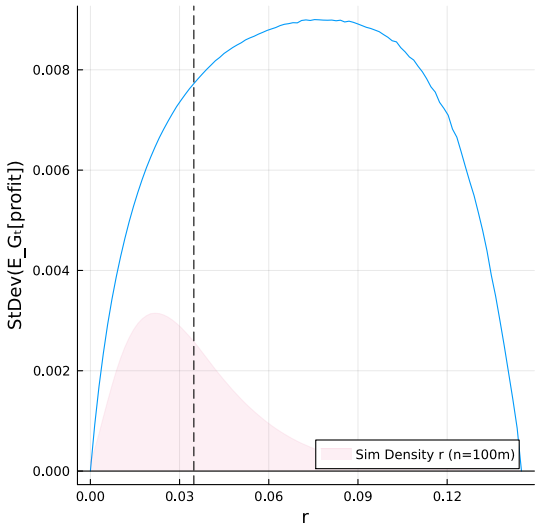


Figure D.1: **Pricing errors:** Standard deviation (solid blue line) of pricing error conditional on r for a random path of r_t that is $T = 100m$ months long, under $m(r, G)$ pricing, where G is the conditional (i.e., $G(\chi|x)$), ergodic origination density. The pink density is the scaled number of observations that the process spends at each interest rate state.

risk as it arises out of the simplified pricing assumption. Thus, while on average investors break even, they are bearing some mispricing risk.

E Policy evaluations and counterfactuals

E.1 Construction of the auto-RM

We first argue that the auto-RM market rate is a reference rate computed by looking at debt instruments traded in the market and prepayable at any time, with a call premium π. Indeed, note $P^*(x, c)$ the price of such a prepayable instrument with coupon c when the latent aggregate state is x:

$$P^*(x,c) := \inf_{\tau} \mathbb{E}_x \left[\int_0^{\tau \wedge \tau_\nu} e^{-\int_0^t r(x_s)ds} (c-f)\, ds + 1_{\{\tau \leq \tau_\nu\}}(1+\pi) e^{-\int_0^\tau r(x_s)ds} + 1_{\{\tau \geq \tau_\nu\}} e^{-\int_0^{\tau_\nu} r(x_s)ds} \right],$$

where τ_ν is a Poisson time with arrival rate ν. This optimal stopping problem is a free-boundary problem, with an endogeneous boundary $m^*(x)$ to be determined. The variational inequality, valid for any x, is

$$\max\left\{-\left(\nu + r(x)\right) P^*(x,c) + c - f + \nu + \mathcal{L}P^*(x,c), P^*(x,c) - (1+\pi)\right\} = 0$$

The HJB (i.e. the left hand side of the above inequality) holds in the continuation region $c \leq m^*(x)$, while the equality $P^*(x,c) = 1+\pi$ holds at the boundary of the continuation region $c = m^*(x)$, where $m^*(x)$ is the "auto-RM rate". The optimality condition for the stopping time τ takes the form of the smooth pasting condition $\partial_x P^*(x, m^*(x)) = 0$. The price function $P^*(x,c)$ satisfies $P^*(x,c) \leq 1+\pi$ for any coupon c and latent state x. P^* is increasing in c, and of course $P^*(x, m^*(x)) = 1+\pi$. Note then that P^* is the limit, as $\chi \to +\infty$, of the following problem

$$\begin{aligned}\hat{P}(x,c;\chi) :&= \inf_{k\in\mathcal{K}_\chi} \mathbb{E}_x \left[\int_0^{\tau_k \wedge \tau_\nu} e^{-\int_0^t r(x_s)ds} (c-f)\, ds \right. \\ &\qquad \left. + 1_{\{\tau_k \leq \tau_\nu\}} (1+\pi) e^{-\int_0^{\tau_k} r(x_s)ds} + 1_{\{\tau_k \geq \tau_\nu\}} e^{-\int_0^{\tau_\nu} r(x_s)ds} \right] \\ &= \inf_{k\in\mathcal{K}_\chi} \mathbb{E}_x \left[\int_0^{+\infty} e^{-\int_0^t (r(x_s)+\nu+k_s)ds} (c - f + \nu + k_t(1+\pi))\, ds \right],\end{aligned}$$

where $\mathcal{K}_\chi$ is the set of progressively measurable processes $\{k_t\}_{t\geq 0}$ so that $k_t \in [0,\chi]$ for all t, and τ_k, in the first equation, is a Poisson time with jump intensity k_t. The auto-RM rate is thus a reference rate that can be computed by looking at debt instruments traded in the market, and that are prepayable at any time at $1+\pi$. These prepayable debt instruments, when issued, have a price and market value of $1+\pi$, and a fair coupon equal to $m^*(x)$, the reference rate for the auto-RM.[48] Borrowers are then locked into that auto-RM instrument, pay the floating rate $m^*(x_t)$ at all times, up to the point where they move. At such time, they prepay the mortgage balance \$1, and are forced to refinance into a new mortgage. Upon taking a new mortgage, borrowers receive proceeds \$1 from lenders, but given that the loan pays a reference rate $m^*(x)$, the market value of such loan is equal to $1+\pi$, meaning that lenders can recoup their origination costs. Borrowers then pay the floating rate $m^*(x_t)$ until the time they move and sell their house. By construction, the reference rate $m^*(x_t)$ satisfies

$$m^*(x_t) = \inf_{t\geq s\geq 0} m^*(x_s)$$

E.2 Auto-RM vs. short rates

Proof of Proposition 8. We consider the case $\pi \geq 0$ — i.e. the case where mortgage origination costs are potentially incurred, and recouped by lenders via higher mortgage rates. As discussed in Online Appendix E.1, the price P^* of the auto-RM solves

$$P^*(x,c) = \inf_\tau \mathbb{E}_x \left[\int_0^\tau e^{-\int_0^t (r(x_s)+\nu)ds} (c-f+\nu)\, ds + (1+\pi) e^{-\int_0^\tau (r(x_s)+\nu)ds} \right].$$

[48] Given the nature of Brownian motions, these prepayable instruments have, at the time of issuance, zero duration.

Now assume for a second that there exists a latent state $\hat{x}$ so that $r(\hat{x}) > m(\hat{x}) - f$. Assume at time $t = 0$, $x_0 = \hat{x}$, and consider a stopping strategy $T = \inf\{t \geq 0 : r(x_t) = m(\hat{x}) - f\}$. Clearly, since $r(x_0) = r(\hat{x}) > m(\hat{x}) - f$ and since x has continuous sample path, $T > 0$ a.s. We then have the following set of inequalities

$$\begin{aligned} 1 + \pi = P^*(\hat{x}, m(\hat{x})) = \inf_{\tau} \quad & \mathbb{E}_{\hat{x}} \left[\int_0^{\tau} e^{-\int_0^t (r(x_s)+\nu)ds} (m(\hat{x}) - f + \nu)dt + (1+\pi)e^{-\int_0^{\tau}(r(x_s)+\nu)ds} \right] \\ & \leq \mathbb{E}_{\hat{x}} \left[\int_0^{T} e^{-\int_0^t (r(x_s)+\nu)ds} (m(\hat{x}) - f + \nu)dt + (1+\pi)e^{-\int_0^{T}(r(x_s)+\nu)ds} \right] \\ & < 1 + \pi, \end{aligned}$$

where the last inequality follows since for $t < T$, we must have $r(x_t) > m(\hat{x}) - f$. This is the contradiction we were looking for. □

E.3 Auto-RM impact on initial debt-to-income ratio

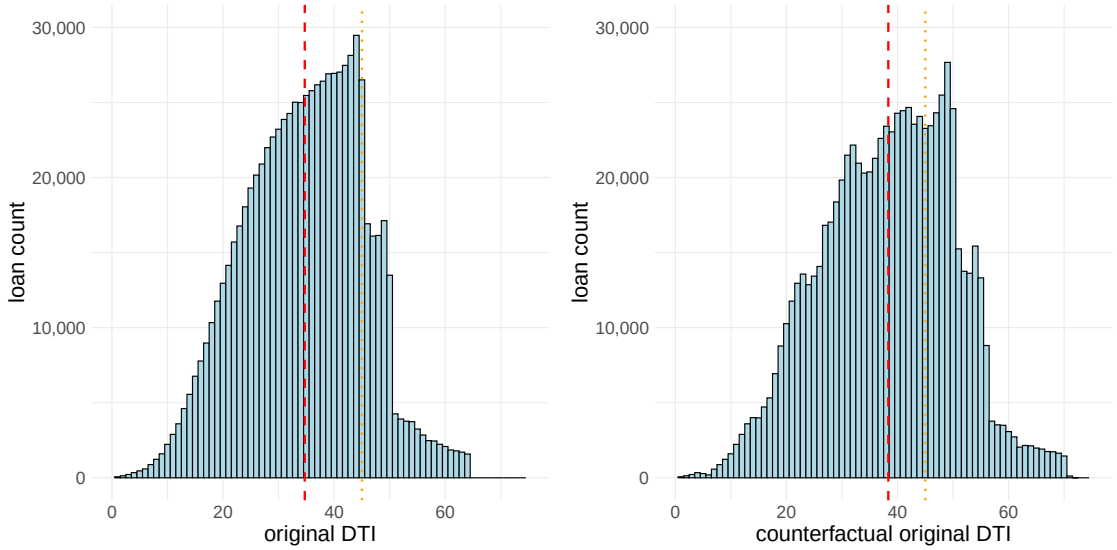


Figure E.1: **DTI distribution and counterfactual DTI distribution.** Left figure shows the DTI distribution in the SFLP data. Right figure shows the counterfactual DTI distribution if mortgage rates were higher than those actually realized, with a difference corresponding to the ergodic average difference between (a) mortgage rates in the approximate pooling MPE and (b) mortgage rates in the auto-RM equilibrium. Vertical red dashed lines indicate average DTI ratio, while orange dotted lines indicate the 43% DTI limit.

E.4 Mortgages with lockup periods

Consider the case where a mortgage cannot be refinanced or prepaid for an exponentially distributed length of time. Denote by $1/\gamma$ the expected duration of this lockup period. We continue to assume that households face no upfront closing costs when refinancing. Mortgage prices are now dependent on whether the mortgage is in its lockup period (state $i = 0$) or not (state $i = 1$). Let $P_i(S, c)$ be the price of a mortgage in state $i \in \{0, 1\}$. We have

$$P_0\left(S, c; \chi\right) := \mathbb{E}_S\left[\int_0^{\tau_\gamma} e^{-\int_0^t r(x_s)ds}\left(c - f\right)dt + e^{-\int_0^{\tau_\gamma} r(x_s)ds} P_1\left(S_{\tau_\gamma}, c; \chi\right)\right],$$

where τ_γ is an exponentially distributed time with parameter γ. Similarly, we have

$$P_1\left(S, c; \chi\right) := \mathbb{E}_S\left[\int_0^{\tau} e^{-\int_0^t r(x_s)ds}\left(c - f\right)dt + e^{-\int_0^{\tau} r(x_s)ds}\right],$$

where τ is the refinancing time (with intensity $\nu + \chi 1_{\{c \leq m(x_t)\}}$). The cross-sectional density of outstanding mortgages is now $f_{i,t}\left(c, \chi\right)$, i.e. it includes the indicator $i \in \{0, 1\}$ for whether a mortgage is currently in its lockup period or not. In the pooling MPE, this density satisfies equations similar to (B.6)-(B.7) — with the important adjustment that (1) borrowers in state $i = 0$ cannot refinance or move, and (2) borrowers transition from state $i = 0$ to state $i = 1$ at rate γ:

$$\begin{aligned}
df_{0,t}(c, \chi) &= -\gamma f_{0,t}(c, \chi)dt, \qquad c \neq m(S) \\
df_{1,t}(c, \chi) &= \left[-\left(\nu + \chi \mathbb{1}_{\{c \geq m(S_t)\}}\right) f_{1,t}(c, \chi) + \gamma f_{0,t}(c, \chi)\right] dt, \\
\lim_{c \uparrow m(S_t)} \partial_c f_{0,t}\left(c, \chi\right) - \lim_{c \downarrow m(S_t)} \partial_c f_{0,t}\left(c, \chi\right) &= \nu \int_{-\infty}^{+\infty} f_{1,t}(c, \chi)dc + \chi \int_{m(S_t)}^{+\infty} f_{1,t}(c, \chi)dc.
\end{aligned}$$

The type distribution of borrowers refinancing at time t is then

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

Finally, market equilibrium dictates that $\int_\chi P_0\left(S_t, m(S_t); \chi\right) dG_t(\chi) = 1 + \pi$. We can then use Assumption 2 in order to simplify substantially our analysis, and reduce the dimensionality of the state space to a 1 state variable problem, in a way similar to what was done in the main text.

F General framework

In this section, we show how to apply our modeling framework to the small business lending market. Consider a continuum of risk-neutral small firms of measure 1. Each

firm generates income normalized to 1, has debt with notional balance normalized to b[49], and technology that fails with intensity $\lambda(x_t, \chi)$, with $\lambda(\cdot, \cdot)$ a known positive function that is increasing in x_t and increasing in χ. x_t is an observable, aggregate variable that represents the state of the economy; it follows a diffusion with drift $\mu(x)$ and volatility coefficient $\sigma(x)$. χ is a firm-specific, time-invariant object that represents the intrinsic quality of the firm's project. Importantly, χ is not observable by the banking sector.[50] The firm quality distribution in the economy is H. When a firm's project fails, the firm ends up defaulting on its existing debt. At such time, a new firm immediately enters the economy, with the same quality, so as to preserve the distribution H over permanent quality heterogeneity.

The banking sector is risk-neutral and competitive; banks provide funding to small firms via loan contracts that mature at Poisson arrival rate ν and that carry an interest rate equal to the sum of (i) the risk-free rate r and (ii) a credit spread s_t, fixed and determined at the time the loan is originated, and meant to compensate banks for expected future credit losses. Firms have the option to refinance their bank debt early, subject to potential refinancing frictions—they must bear fixed debt issuance costs ψ, and only make decisions at discrete *points* in time, arriving with intensity α.

A firm currently financed with a loan at interest rate spread s has equity value

$$\begin{aligned} V(x,s) :&= \sup_{a \in \mathcal{A}} \mathbb{E}_{x,s}\left[\int_0^{\tau_\chi} e^{-rt}\left(1 - \left(r + s_t^{(a)}\right) b\right) dt - a_t \psi dN_t^{(\alpha)}\right], \\ \text{s.t.} \qquad ds_t^{(a)} &= \left(\mathcal{S}(x_t) - s_{t-}^{(a)}\right)\left(a_t dN_t^{(\alpha)} + dN_t^{(\nu)}\right), \end{aligned}$$

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, $\mathcal{S}(x_t)$ is the equilibrium credit spread charged by banks on new loans when the aggregate state of the economy is x_t, τ_χ is the firm's default time (with time-varying intensity $\lambda(x_t, \chi)$), $N_t^{(\nu)}$ (resp. $N_t^{(\alpha)}$) is a counting process for maturity events (resp. refinancing decisions).

Firms refinance whenever the economy is improving "sufficiently"; their decision depends on the spread s over the risk-free rate currently paid on their loan. Specifically, a firm optimally refinances when $s - \mathcal{S}(x) \geq \theta(x)$, where the state-dependent spread threshold $\theta(\cdot)$ satisfies

$$V(x, \mathcal{S}(x)) - \psi = V(x, \mathcal{S}(x) + \theta(x))$$

Banks are competitive when offering new loans to a new customer firm. The *shadow*

[49] The parameter b can thus be interpreted as the debt-to-income ratio of a given firm.

[50] This assumption can easily be relaxed, by assuming for example that investors can also rely upon a public and noisy signal of firm quality; in that case, the equilibrium would result in separation based on the public signal, and pooling for the private signal.

price of a given \$1 notional loan to a borrower with known quality χ is given by

$$P(x,s;\chi) = \mathbb{E}_x\left[\int_0^{\tau_\chi \wedge \tau_\theta} e^{-(r+\nu)t}\left(r+s+\nu\right)dt + \mathbb{1}_{\tau_\chi<\tau_\theta} e^{-(r+\nu)\tau_\chi}\rho + \mathbb{1}_{\tau_\chi>\tau_\theta} e^{-(r+\nu)\tau_\theta}\right],$$

where τ_χ (resp. τ_θ) is the firm's default time (resp. loan prepayment time), and ρ is the recovery rate realized by creditors upon a default. Banks break-even when a new loan is issued; they price loans under a (potentially state-dependent) firm quality distribution G, so that their no-profit loan origination condition can be written

$$\mathbb{E}^G\left[P(x,\mathcal{S}(x);\chi)\right] = 1$$

The origination distribution G is distinct from the distribution over firm quality H, with G skewed towards riskier firms, since (i) high risk firms will have a refinancing spread threshold θ that is lower than that of low risk firms, and (ii) riskier firms default at higher intensity, and are replaced by firms with identical quality that will immediately seek loan funding. In this model, low quality firms are subsidized by high quality firms, since they finance themselves at credit spreads more advantageous than if banks could observe the firm quality χ. An improvement in aggregate credit market conditions triggers a wave of loan refinancing events, consistent with the data. This model emphasizes the capital misallocation taking place in the banking sector due to the unobserved firm quality and the competitive banking sector.