Last, we show that banks themselves face an important margin of adjustment to balance sheet capacity shocks, the aforementioned balance sheet retention margin. Banks adjust to a decline in balance sheet capacity by decreasing lending on balance sheet and increasing originations of loans that are easier to securitize. In other words, banks' business models are adaptable. Banks, which are flush with capital, behave as standard balance sheet models of banking would suggest: they use their capital to extend loans, which they retain on their balance sheets. However, as banks' balance sheet capacity declines, they switch to originating mortgages, which they can sell, behaving more like shadow banks.

Motivated by this evidence, we build and estimate a workhorse structural model of the financial intermediation sector. It incorporates competition from shadow banks and banks' abilities to choose to lend on balance sheet and to sell loans. The model has several goals. First, we want to understand how the industrial organization of intermediation and the choice of bank business model determine who originates which types of mortgages, and at which interest rates. Second, we want to use the model to quantitatively analyze the consequences of capital requirements, access to secondary loan markets, and unconventional monetary policy on lending volume and pricing, bank stability, and the distribution of consumer surplus across rich and poor households. In the process, we quantify the importance of the shadow bank migration and balance sheet retention margins for policy analysis.

The supply side of the model is determined by competing banks and shadow banks, which offer differentiated mortgages in the jumbo and conforming markets. The central innovation is in modeling banks. As is common in banking models, banks can originate loans, which they keep on their balance sheet, and their balance sheet capacity is limited by their capitalization. We differ from standard models by allowing banks to adjust their business models on the balance sheet retention margin: banks can choose how many loans to retain versus sell. Shadow banks benefit from a lower regulatory burden and compete with banks in originating loans, but can only finance loans by selling them.

We model demand using a modified discrete-choice framework featuring rich heterogeneity (Berry et al. 1995; Nevo 2000). Consumers with heterogeneous preferences over price, quality, and mortgage size choose among a menu of mortgages offered by various types of originators. This heterogeneity is important to accommodate realistic consumer substitution patterns and, especially, to capture the redistributive consequences of policies (see Stroebel and Vavra (2019) and Wong (2018)). Because markets are segmented, policy interventions have redistributive consequences. For example, if capital requirements decrease the supply of on-balance-sheet lending, then this policy will likely be costlier for wealthier borrowers, who are more likely to take jumbo mortgages. We depart from discrete-choice models by also allowing consumers to *choose* their mortgage size and, consequently, decide whether they want a conforming or jumbo mortgage (see Benetton (2019) for an alternative way to model discrete-continuous choices in the mortgage market).

We estimate demand and supply separately. To identify standard demand parameters, we need to instrument for price endogeneity. We exploit geographic differences in financing cost of GSE-conforming mortgages, which arise through political economy considerations and are unrelated to mortgage demand (see Hurst et al. (2016) for extensive documentation of that fact). Second, we exploit bunching at the confirming-jumbo cutoff to help estimate consumer preferences for mortgage size. Intuitively, consumers who bunch at the conforming loan limit choose a smaller than ideal