guarantees. Consequently, these mortgages are significantly more difficult to securitize, and the vast majority are retained by the originators.

II.B Description of Datasets

Our paper brings together a number of datasets which we describe below:

HMDA: Mortgage-level application data is the main source for market shares across lender and product types. The Home Mortgage Disclosure Act (HMDA) collects the vast majority of mortgage applications in the United States, along with their approval status. In addition to the application outcome, the dataset includes loan type, purpose, amount, year of origination, and location information down to the applicant's census tract. It further contains demographic information on the applicant, including race and income. Important for this analysis, it includes the originator's identity, which we link manually across years. Finally, it documents whether the originator sells the loan to a third party, and if so, whether the loan purchaser is a GSE. An important caveat with the sales data is that if the originator retains the loan through the end of the calendar year and sells it in the subsequent year, it is recorded in HMDA as a non-sale. We use data beginning in 2010 and ending in 2016.

Fannie Mae and Freddie Mac Single-Family Loan Origination Data: These datasets, provided both by Fannie Mae and Freddie Mac, contain origination data from the GSEs' thirty-year, fully amortizing, full-documentation, single-family, conforming fixed-rate mortgage purchases.[4] The loan-level data contain information on the loan, property, and borrower, including loan size, interest rate, loan purpose, property location, borrower credit score, loan-to-value ratio, and, importantly, the identity of the lender that sold the loan to the GSE. We use these data to calculate average interest rates by lender type and market.

Black Knight McDash Loan-Level Mortgage Performance Dataset: Black Knight is a private company that provides a comprehensive, dynamic loan-level dataset on mortgages, including loans serviced by the ten largest U.S. mortgage servicers, accounting for approximately 75% of all mortgages in the U.S. as of year-end 2010 (data vendor estimate). Importantly for our purpose, Black Knight includes information on both jumbo and GSE loans and includes loans retained on banks' balance sheets. Much like the Fannie Mae and Freddie Mac data, Black Knight McDash data contain interest rates and a large number of borrower- and loan-specific characteristics, including FICO score at origination, loan-to-value ratio, five-digit zip code of origination, loan purpose, and whether the loan is fixed or adjustable-rate. The Black Knight McDash data also include dynamic data on monthly payments, mortgage balances, and delinquency status.

BlackBox: BlackBox is a private company that provides a comprehensive, dynamic loan-level dataset with information about more than twenty million privately securitized subprime, Alt-A, and prime loans originated after 1999. These loans account for about 90% of all privately securitized mortgages

[4] The dataset does not include adjustable-rate mortgage loans, balloon loans, interest-only mortgages, mortgages with prepayment penalties, government-insured mortgage loans such as Federal Housing Authority loans, Home Affordable Refinance Program mortgage loans, Refi Plus™ mortgage loans, or nonstandard mortgage loans. The dataset also excludes loans that do not reflect current underwriting guidelines, such as loans with originating LTVs over 97% and mortgage loans subject to long-term standby commitments, those sold with lender recourse or subject to other third-party risk-sharing arrangements, or those acquired by Fannie Mae on a negotiated bulk basis.