Institute for International Economic Policy Working Paper Series
Elliott School of International Affairs
The George Washington University

Personal Privacy of HMDA in a World of Big Data

IIEP-WP-2017-21

Anthony Yezer
George Washington University

October 2017

Institute for International Economic Policy
1957 E St. NW, Suite 502
Voice: (202) 994-5320
Fax: (202) 994-5477
Email: **iiep@gwu.edu**
Web: **www.gwu.edu/~iiep**

Personal Privacy of HMDA in a World of Big Data

September 29, 2017

Anthony M. Yezer[+]

Abstract When the Home Mortgage Disclosure Act was passed in 1975, it required selected depository institutions to report limited data from mortgage applications. This was collected and processed by the Federal Reserve Board in accordance with Regulation C. A subset of the reported information was then disclosed to the public. At the time, it was difficult to determine the identity of individual respondents in HMDA data. Since that time four things have changed. First, reporting requirements have been expanded to an increasing range of lenders. Second, the personal information reported and revealed has expanded. Third, over 30% of home purchases do not involve a HMDA reported mortgage and mortgage lending is increasingly internet based. Fourth, modern computing and big data techniques now allow the HMDA data releases to be matched with the names of individual borrowers in a fashion that violates standards for privacy established by the U.S. Bureau of the Census and appears to violate privacy standards of HMDA itself. Lack of privacy is particularly a problem for minority borrowers for whom the "risk" of re-identification is a virtual certainty.

The Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank) amended HMDA to require collection and reporting of additional information. It assigned responsibility for implementation to the Consumer Financial Protection Bureau (CFPB) and authorized it to require reporting even more data but did not reduce privacy protections in HMDA. Under its October, 2015 rule implementing these provisions, the CFPB requires lenders to report all of the new variables specified in Dodd-Frank including credit score and property value. The CFPB has added several variables including interest rate, points, debt-to-income, and loan-to-value ratios. This paper discusses both the risks to consumers' privacy under the "old" HMDA reporting rules and the how these risks increase if the "new" expanded list of variables were released. It also considers the conflicts between both old and new HMDA data disclosures and traditional and legal restrictions that protect consumer privacy. Finally, it notes how current big data techniques provide information on credit flows into housing markets that make HMDA obsolete, misleading, and redundant.

[+]Professor of Economics, Department of Economics, George Washington University contact at yezer@gwu.edu. Financial support for this research was provided by the Mortgage Bankers Association. The analysis and conclusions in this paper are the responsibility of the author and do not necessarily reflect views of the Mortgage Bankers Association.

Personal Privacy of HMDA in a World of Big Data

Executive Summary

Since its passage in 1975, HMDA has been expanded to include an increasing range of lenders. Reporting requirements expanded the list of variables collected from lenders on the characteristics of borrowers and mortgage loans. For example, in 1989, the Financial Institutions Reform, Recovery and Enforcement Act required creditors to collect race, sex, and income data. In 2002, Regulation C modified the collection of race and ethnicity information. This Government Monitoring Information (GMI) has been increasingly challenging to collect in a world of internet lending. Lenders, of course are asked to collect this information but prohibited from using it. This formed what will be termed the "old" set of HMDA data which was being collected and reported in 2010.

The Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank) of 2010 moved responsibility for design of HMDA reporting away from the Board of Governors of the Federal Reserve (FED) and vested it in the new Consumer Financial Protection Bureau (CFPB). Furthermore, Dodd-Frank modified the characteristics of lenders required to report HMDA data, added variables to the list of information collected, and empowered the CFPB to make further changes to the scope of HMDA coverage. This has given rise to what will be termed the "new" HMDA data set.

Under its October, 2015 rule implementing changes from the old to the new HMDA data, the CFPB requires lenders to collect and report all of the new data specified in Dodd-Frank and several additional variables including borrower's credit score, debt-to-income, and loan-to-value ratios. By January 1, 2018, lenders are required to devise methods to collect all of the added variables for applications and endorsed loans for purchase mortgages, refinances, and a variety of other credit transactions collateralized by residential property. This information is to be reported to the CFPB by March, 2019 and a subset of the variables on these 2018 applications and loans will be disclosed as the new public HMDA data in 2019.

In the past, virtually all of the old HMDA variables collected except unique loan identifiers and application date have been released to the public in easily downloaded datasets. The identity of each borrower is not revealed. However, due to computational advances and the ability to scrap public data recording property ownership including liens, it has become increasingly possible to either identify the mortgage transactions of individuals whose identity and property address are known or to re-identify HMDA data by determining the name(s) and address of the borrower(s).

The first purpose of this research is to determine the extent to which borrowers can be identified or re-identified using the old public HMDA data and the privacy concerns raised by the information disclosed. The method(s) of accomplishing the match between individual identifies and individual HMDA data records are explored. Second, the effects of additional new HMDA variables, both those identified under Dodd-Frank and the variables added by the CFPB, on identification and re-identification rates are explored. Third, privacy of both old and new HMDA data is compared to privacy standards applied to other government mortgage data disclosed to the public. Fourth, privacy standards imposed on personal data collected and released to the public by the U.S. Bureau of the Census (USCB) is be contrasted with privacy of HMDA data. Specific implications of imposing current USCB standards to HMDA disclosures are developed. Finally suggestions for limits on HMDA data disclosures designed to preserve privacy are provided along with suggestions for alternatives to the current use of HMDA data that would provide better information on the state of mortgage lending to neighborhoods in the U.S. – i.e. reliance on big data.

Short responses to each of these elements of the research are enumerated below.

1. There is very little protection of consumer privacy in old HMDA data disclosures. Assuming that HMDA data are correct, virtually all borrower(s) can be identified provided that they used a lender reporting to HMDA. Re-identification, i.e. attaching a borrower name and property address to HMDA data that was correctly recorded, can be achieved in over 80% of all cases. This is not a new development. Published economic research documented high rates of matching old HMDA data to borrower names over twenty years ago when less information was disclosed and big data scraping and matching techniques were just being developed. The current practice of disclosing high rate loans allows the personal identification of the names of those borrowers who are not creditworthy at lower interest rates. Re-identification is easiest for minority borrowers and those not using the largest lenders.

2. The additional variables required to be reported under Dodd-Frank in new HMDA, both under the statute and pursuant to the Bureau's discretionary authority, would raise the probability of identifying or re-identifying borrowers whose data is correct to virtually 100%. The combination of loan amount, house value, lender name and census tract is sufficient to identify virtually every mortgage transaction.

3. The data points that have been added under the CFPB's discretion would, independent of those required under Dodd-Frank, raise the ease of identification and re-identification to a virtually certainty. Furthermore, providing additional detail on the cost of credit, in the form of interest rate and discount points paid,

along with the loan-to-value and debt-to-income ratios allows imputation of the approximate credit score of the borrower. These two ratios, along with the borrowers credit score determine the cost of credit. Knowledge of any three numbers gives informed individuals the ability to impute the third.

4. The value of the new HMDA data to identity thieves and individuals engaged in financial fraud in targeting the financially vulnerable would be considerable. For example, the new HMDA would expose the age and financial condition of elderly households who had just endorsed reverse mortgages. To the extent that households believe that the details of their financial condition reported to lenders are private, those seeking to prey on these individuals can gain their confidence by appearing to detailed knowledge of loan terms and their financial condition. One click on an attachment to what appears to be a legitimate inquiry from their lender can expose these individuals to a variety of cyber crimes.

5. Identification and re-identification rates vary substantially among mortgage data sets that are released to the public. Borrower identification protection is generally achieved by suppressing information on location of the real estate collateral or detailed demographic characteristics of the borrower. Another data set providing the same low level of privacy as old HMDA data is the Federal Home Loan Bank Board Purchased Mortgage File which only covers a small number of loans annually (about 35,000). Some of the GSE Enterprise public use datasets also have high re-identification rates while in other cases there has been significant masking to protect borrower identity. The comments on privacy issues in HMDA made here also extend to these other data sets because they are supposed to be subject to the same standards of anonymity provided under HMDA.

6. The decision to require reporting credit scores in the new HMDA data appears to be in conflict with the privacy provisions of the Fair Credit Reporting Act (FCRA). In addition to its general privacy protections, the FCRA includes specific opt out provisions designed to protect the individual privacy.

7. The lack of consumer privacy protection in old HMDA data contrasts sharply with the anonymity in data released by the (USCB) which has an elaborate research program to identify privacy issues and suppress variables that could be used to identify respondents. Old HMDA data disclosures also appear inconsistent with the Right to Financial Privacy Act (RFPA) of 1978, and even to the privacy considerations described in HMDA, 12 U.S.C. 2803(j) which require modification, often termed "masking", of itemized information, for the purpose of protecting the

privacy interests of the mortgage applicants or mortgagors, in data that is available to the public.

8. For fundamental recommendations for both the old and new HMDA data arise naturally from the analysis performed here.

   a. First, information that can be used to identify individual creditworthiness of borrowers should be suppressed, perhaps by only disclosing information on credit score and interest rates into very broad intervals.
   b. Second, the privacy criteria of the USCB should be applied to HMDA data releases consistent with the requirement for modification to provide consumer privacy protection in the current Act. A semblance of this level of concern with privacy is evidenced in the masking of detailed data in one of the GSE private enterprise loan level datasets. This would mean disclosing far less geographic detail and suppressing lender identification except in areas where a lender was making a substantial number of loans. Demographic detail on borrowers, which is problematic at best in a world of electronic lending, would also need to be aggregated. The advantage of preserving privacy using USCB criteria is that the expertise to make disclosure decisions that preserve privacy has already been developed and applied to a number of data sets including some, like the Survey of Consumer Finances, that contain information similar to that collected under HMDA.
   c. Third, abandon HMDA altogether in favor of the data commercially available from property records. This has the advantage of covering all housing finance in neighborhoods without invading the personal privacy of borrowers. This would serve the original purpose of HMDA far better and produce a substantial financial saving to the government and the banking system.
   d. Fourth, inform each applicant of the lack of privacy in both the old and particularly the new HMDA so that they are fully aware of the privacy implications of borrowing from a given lender and allow them to opt out of having their personal financial information disclosed to the public.

Virtually all of the findings in this report are taken from previously published work. Some updated demonstrations have been performed in order to illustrate the mechanisms used to achieve identification or re-identification of borrower identity using old HMDA data. However, the lack of consumer privacy protection in HMDA has been well known for almost 20 years. It is disturbing that the CFBP website fails to reveal this fact to the public and gives the impression that data on individual financial condition taken from the mortgage application and revealed under HMDA remains private. Old HMDA data identifies borrowers who have low creditworthiness by indicating that they paid high interest rates. New HMDA data would reveal all aspects of borrower credit worthiness in a fashion that seems completely inconsistent with FCRA and with privacy protections routinely applied to other government data by the (USCB).

The original problem that HMDA was to address concerns regarding the availability and flow of investment funds in local housing markets. Today HMDA does that job very poorly because it misses a substantial fraction of housing purchases. Based on Core Logic data from property transfer records, cash only sales rose from 23% in 2001 to a high of 43% in 2012 and currently are above 30% of all sales. Another fraction were financed by transactions, including seller financing, not reported under HMDA. Home purchases not reported in HMDA not only vary over time they vary by location. In some neighborhoods the percentage of transactions not covered by HMDA is much higher. Thus the current HMDA sampling frame of reporting institutions creates a biased view of housing finance and the degree of bias varies across communities. Removing privacy protections in HMDA will only provide greater incentive to avoid purchasing through channels that report making the underreporting problem even worse. As an alternative, commercially available datasets provide information on all property transactions and all financing in neighborhoods rather than the selected, incomplete, and misleading sample covered in HMDA data. These data sets protect the financial privacy of individual homeowners. Increasingly these data sets have become the basis for research in finance and economics.

The reason for including “in the world of big data” in the title to this report is that HMDA has become obsolete as a database for monitoring real estate finance in neighborhoods in the current world of big data because so many transactions do not involve mortgages originated by covered lenders. Expanding HMDA data by invading the privacy of homeowners does nothing to solve the problem of biased coverage and encourages a flight to non-reported financing. In 2015 HMDA data there are 3.66 million home purchase transactions but the Federal Reserve Bank of St Louis FRED database reports 5.4 million existing and another half million new home purchases. If the government is really concerned about monitoring the flow of investment capital into all residential real estate transactions rather than a modest fraction, then both old and new HMDA provide biased and misleading evidence on the number and nature of real estate transactions today. Expanding HMDA data to further invade the privacy of homeowners will only increase flight to alternative sources of financing where privacy is preserved and increase the inadequacy of HMDA data to characterize the flow of finance into neighborhood housing.

Personal Privacy of HMDA in a World of Big Data

I. Introduction

Since its passage in 1975, the Home Mortgage Disclosure Act (HMDA) has required the reporting, collection and dissemination of information on individual mortgage applications, whether they result in endorsed mortgages or not, at covered lenders. Virtually all the data collected has been disclosed to the public. Over time the coverage of lenders, data collected, and public disclosures have expanded. For example, the Financial Institutions Reform, Recovery and Enforcement Act of 1989 required creditors to collect race, sex, and income data. In 2002, Regulation C modified the collection of race and ethnicity information. This Government Monitoring Information (GMI) has been increasingly challenging to collect in a world of internet lending. Lenders are asked to collect this information but prohibited from using it.

The Dodd Frank Wall Street Reform and Consumer Protection Act of 2010 (Dodd-Frank Act) moved responsibility for determining the data collected and disclosure policy from the Board of Governors of the Federal Reserve (FED) to the new Bureau of Consumer Financial Protection (CFPB) and required the collection of data on more characteristics of the applicant and mortgage. The CFPB has the authority to expand data collection and dissemination beyond variables specifically noted in the Dodd-Frank Act. Privacy concerns in HMDA data from the pre and post Dodd-Frank eras will be discussed here. To avoid confusion, the data collected and disclosed in 2016 will be termed the "old" HMDA data and that data augmented by the variables either mandated in Dodd-Frank or proposed by the CFPB for collection beginning January 1, 2018 will be known as the "new" HMDA data.

While the coverage of HMDA has expanded, both in terms of institutions reporting and information extracted from lenders, computer technology has substantially changed the ability of researchers, hackers, and identity thieves to uncover the identify of individual HMDA borrowers. This has prompted research by the U.S. Census Bureau (USCB) into privacy protections for individuals covered in government data sets. One purpose of this report is to document these developments and their implications for the lack of privacy in HMDA data currently disclosed to the public. Another development in the world of big data is that commercial vendors have developed techniques for assembling big databases on all residential real estate transactions (and commercial transactions as well). Compared to these datasets, HMDA is seriously incomplete in its coverage of residential real estate transactions and provides a misleading view of developments in housing finance.

As directed by the Dodd-Frank Act, the CFPB has promulgated regulations that alter the organizations reporting HMDA data, and greatly expand the amount of data to be reported. The new HMDA data is to be collected during 2018 and reported by March, 2019. Preparations to accomplish the collection and reporting of new HMDA data are underway at this time. What is not clear at this time is the portion of the data collected that will be disclosed. This report provides information on the privacy concerns related to both the old HMDA data and to the possibility that further expansion of the information taken from lenders and revealed to the public will erode privacy and violate important consumer protections.

Privacy concerns regarding collection and disclosure of data involve two elements. First is the intrusive nature of the data itself. As this report will make clear, there is already substantial publicly available personal information on borrowers, lenders, the terms of the mortgage contracts and the real property collateral that connects them. However, thus far privacy concerns have limited the form in which this data is disclosed to protect the identity of individuals and also limit indicators of creditworthiness. The old HMDA disclosures are an exception to this privacy protection because individuals are readily identified or re-identified. The main protection against abuse in the old HMDA data is that only very limited indicators of creditworthiness are being revealed.

While it is true that old HMDA disclosures do not contain the names of borrowers, this report will demonstrate that, except for white non-Hispanic borrowers purchasing or refinancing with the largest lenders in suburban areas or those providing false information to HMDA, it is a relatively simple matter to match individual names with their records in HMDA. Given that old HMDA data includes loan amount and indicators of mortgage cost, many borrowers might regard this as an unwarranted intrusion if they were truthfully told that their names could be linked with the information currently disclosed under HMDA. Unfortunately, the discussion of privacy in old HMDA data on the CFPB website gives the false impression that this type of borrower identification is not easily accomplished.

The remainder of this report asks and answers a series of questions.

First, what does past research tell us about identification and re-identification risk in old HMDA data? Surprisingly, published research has established methods for linking names and addresses of individuals to their mortgage transactions disclosed to the public in HMDA data. The process used for achieving these matches is discussed and illustrated. This research in which the author of this report was an early participant, demonstrates conclusively that more that 80 percent of borrowers can be re-identified in old HMDA data disclosures and that the re-identification rate is particularly high for minority borrowers. Under new HMDA these rates would rise to nearly 100 percent.

Second, how would expanding new HMDA data aid in this linking process and what threats to individual confidentially are included in the additional variables being collected from lenders? The potential for harm to individual privacy and security are substantial and protections offered under the Fair Credit Reporting Act (FCRA) and even HMDA itself appear to be ignored.

Third, what is the relation between old and possible new disclosures of HMDA data and information available from other data sources? Would new HMDA disclosures reveal information about borrowers that is routinely suppressed based on privacy concerns in other government data?

Fourth, what principles should guide those seeking to protect borrower confidentiality as the debate over old and new HMDA data disclosures goes forward and is HMDA reliable compared to other sources of information on housing sales and financing? Fortunately the principles for data disclosure to preserve privacy have been developed by the U.S. Census Bureau (USCB). They have been applied to some mortgage data sets that are disclosed to the public and are widely used in preserving anonymity in other government data sets. The central recommendation is that, if HMDA is to be continued as a database, disclosures be given the same careful treatment to avoid identification and re-identification of any households as the USCB uses to protect privacy in other data sets. Alternatively, mortgage applicants should be told that there is no privacy in HMDA data disclosures and that they have the ability to opt out of having any information regarding their income, assets, or credit score collected as part of HMDA. Additionally, HMDA data are incomplete and misleading as a guide to the financing of housing compared to commercially available data sets that are far more comprehensive in their coverage of transactions. Simply put, HMDA data provide a biased and incomplete view of financing of residential property in the U.S. and alternative commercial data sets are available that monitor credit and investment in housing far better than HMDA.

What does past research tell us about identification and re-identification risk using old HMDA data?

As will be documented below, the ability to re-identify borrowers in old (pre-Dodd-Frank) HMDA data has been established in the research literature. Furthermore the logic behind the ease of re-identification for all groups, except white non-Hispanic individuals borrowing from the largest lenders in suburban areas where there are high rates of owner occupancy and turnover, is easily understood. Finally, an exercise in re-identification is performed using 2015 HMDA data to illustrate the ease of re-identification without using complex programming or statistical methods.

What is the current CFPB position on privacy of old HMDA data disclosures?

Given its charge to protect consumers, it is logical to begin with the position of the CFPB on consumer privacy protections in old HMDA data. The CFPB takes the public position that individual privacy is preserved. Specifically, the transcript of a video titled "About HMDA," which is designed to acquaint the public with privacy in old HMDA disclosures states:

> "And finally, there's information about the property itself. You can see the type of property and whether the owner intends to live there. Instead of disclosing the

> address, lenders disclose the census tract, which is the part of a community where the property is located. Census tracts vary in size, but on average about 4,000 people live in a census tract. This provides enough information about the location to be useful, but still provides protections for individual privacy." CFPB Website http://www.consumerfinance.gov/data-research/hmda/learn-more#transcript

It is not clear whether this statement refers to identification of individual borrowers, which is likely the prime concern of homeowners, or re-identification of all HMDA data records. In either case, the statement suggests that identification and/or re-identification is unlikely because census tracts are large. While 4,000 people or perhaps 1,800 housing units may appear to be a large number, this report reveals that re-identification is relatively straightforward and that, for the vast majority of households purchasing or refinancing homes using mortgages supplied by reporting lenders, there is no consumer privacy protection in old HMDA data released to the public. Furthermore, the degree of privacy protection afforded minorities in old HMDA data is far lower than that enjoyed by non-Hispanic white borrowers.

What does academic research show about re-identification using old HMDA data?

Re-identification, adding the names of the borrowers to old HMDA data records, has become possible due to modern computer technology that has automated the linking process. First, information on property transactions that is needed to record ownership and deeds is now available online. This data is searchable and scrapable. It contains names of mortgagees, mortgagors, and even trustees along with the loan amount, property address, and, sometimes, terms of the debt instrument. However, it does not contain information on the creditworthiness, income, credit score, etc. of the individual. Second, property tax assessment, and estimates of market value used in mortgage underwriting rely on automated appraisals. This data includes property address and physical characteristics of the housing unit. A number of commercial vendors have scraped or purchased this information and assembled it into datasets that contain names of owners and information on neighborhood characteristics, tax liabilities, assessments, etc. while preserving the financial privacy of the homeowner. For purposes of monitoring mortgage flows and real estate investment into neighborhoods, this data is more valuable than HMDA because it contains a complete record of all property transfers. It is routinely used in economic research today. HMDA does not cover cash purchases, sellers taking back mortgages, and credit extended by non-traditional lenders. As a result it presents an obsolete and misleading impression of the sources and consequences of lending and investment activity, particularly in low income neighborhoods.

In the 1970's and 1980's digital property records and the computer technology needed to scrap and process them were not available. Only in the 1990's did this technology emerge. The author of this report was the co-principal investigator on a large project sponsored by the Department of Housing and Urban Development in the mid 1990's that was designed to aid FHA in finding areas where its market share could be increased by lending to the underserved. In order to locate home buyers not served by FHA or conventional lenders, it was necessary to match HMDA data to the property records available at that time so that individuals using what was termed “brand X” financing could be identified. These individuals were then evaluated and possible targets for FHA financing were identified. Given the data quality and linking technology available at the time, the re-identification rate was just over *50%.*

Subsequently, Pennington-Cross and Nichols used data from this research project to analyze the choice of FHA versus conventional mortgages.[1] HMDA data was needed to identify the ethnicity and race of the borrower. They report a *52%* success rate in linking conventional loans to old HMDA data. The FHA match rate to HMDA was much higher because they had access to the full FHA loan file. The point here is that, even using the quality of data and programming techniques available in 1996 (when that matching was done), they were able to re-identify over 50% of conventional mortgages in HMDA data.

Since publication of the Pennington-Cross and Nichols paper, academic research in economics has continued to re-identify HMDA borrowers in order to match them with other data sets. Sorenson matches HMDA data to local assessment and property data to study the foreclosure process.[2] Bocian, et. al. perform a massive *27* million loan match of HMDA data to available data from loan processors.[3] The authors note that unique matching is more difficult for prime loans, loans not originated by brokers, government loans, and loans in boom areas. Put another way factors that raise the volume of lending by a given lender in a particular census tract make unique linking more difficult. Laderman and Reid undertake a similar match between HMDA data and loan files for California.[4] While these matches use loan files rather than property records, the matching algorithms use the same information available publically in property records, i.e. lender identification number, property location, year of endorsement, and loan purpose (purchase or refinance). Unfortunately precise accounts of the quality of the matching process are not provided. However, the validity of these published studies is based on the presumption that the matching process was precise and successful.

[1] See, Pennington-Cross, Anthony, and Joseph Nichols, 2000. Credit History and the FHA-Conventional Choice, *Real Estate Economics*, Vol 28 (2), 307-336.

[2] David J. Sorenson, 2015. Loan Characteristics, Borrower Traits, and Home Mortgage Foreclosures: The Case of Sioux Falls, South Dakota, *Journal of Regional Analysis and Policy,* Vol 45, No 2, 163-172.

[3] Debbie Gruenstein Bocian, Wei Li, Carolina Reid, and Roberto G. Quercia, 2011. Lost Ground, 2011: Disparities in Mortgage Lending and Foreclosures, Center for Responsible Lending. This match was done using location and mortgage amount. Rather than match all loans uniquely, a probabilistic match was performed because loans not uniquely matched are not missing at random.

[4] Elizabeth Laderman and Carolina Reid, 2008. Lending in Low- and Moderate-Income Neighborhoods in California: The Performance of CRA Lending During the Subprime Meltdown, Working Paper 2008-05, Federal Reserve Bank of San Francisco.

A significant body of literature on the foreclosure process during the 2004 to 2008 period of high foreclosure activity has matched HMDA data to local property records in exactly the fashion described in this report. Coulton, Chan, Schramm, and Mikelbank match property records from the Cuyahoga County Recorder with HMDA data for several years and report an overall *68%* re-identification rate.[5]

Gerardi and Willen report research results using matched HMDA data and property records in Massachusetts.[6] The matching performed in this paper is notable for a number of reasons. First, HMDA data for several years were re-identified by matching with property records. Second, the matching was done both from property records to HMDA data and also run in the opposite direction. The match rate from property records to HMDA was *60%,* while the re-identification rate from HMDA to property records was *70%* in 1998 and rose steadily to *75%* in 2001. What accounts for the difference? Property records include all transactions involving liens against real estate. These differences in match percentage reflect the fact that HMDA does not cover all debt finance of housing purchases or any cash sales. This illustrates the inadequacy of HMDA for its intended purpose, i.e. studying the flow of financial resources into local housing markets. Both current HMDA and its proposed extension, completely miss a substantial portion of the market. The asymmetry in matching rates implies that about *15%* of the mortgage finance in the property records did not appear in HMDA, or that HMDA only includes *85%* of mortgage finance, and naturally a much smaller percentage of all home purchases because cash purchases involve no liens. The *70* to *75%* HMDA match rate was achieved using an algorithm that was developed internally by the staff of the Supervision, Regulation, and Credit Unit of the Federal Reserve Bank of Boston.

In all the examples described above, the re-identification rates were achieved using computer algorithms that matched based on a limited range of information because the research purpose was to assemble large data sets that included HMDA information to supplement other data. The goal was not to maximize re-identification rates and certainly not to identify any single individual. Presumably the precise property address and mortgagor names were removed from the data files in a process called depersonalization as part of the merging algorithm.[7] Clearly identity thieves or others whose purpose was invasion of privacy rather than economic

[5] Claudia Coulton, Tsui Chan, Michael Schramm, and Kristen Mikelbank, 2008. Pathways to Foreclosure: A Longitudinal Study of Mortgage Loans, Cleveland and Cuyahoga County, 2005-2008, Center on Urban Poverty and Community Development, Mandel School of Applied Social Sciences, Case Western Reserve University, Cleveland, Ohio.

[6] Kristopher S. Gerardi, and Paul S. Willen, 2008. Subprime Mortgages, Foreclosures, and Urban Neighborhoods, Public Policy Discussion Paper No. 08-06, Federal Reserve Bank of Boston.

[7] In most cases the publications mention that the matched HDMA records were depersonalized as part of the merger algorithm so that privacy was protected. Accordingly there is no intent to suggest that these researchers, or the organizations that employ or finance them (including Federal Reserve Banks), have any intent to invade the privacy of mortgage borrowers. However, they could have attached names and property addresses to the merged files if they were less scrupulous.

research are not so scrupulous and they do not publish their work in academic papers where the matching rates can be revealed. Furthermore, more precise algorithms can raise match rates. For example, there are ways to determine that "2nd St," "2nd Street", "Second St", "Secnd Street", all indicate the same street location. Similar typographical errors in recording Zip codes and census tract numbers can be identified and remedied if the algorithm is sufficiently sophisticated. It appears that none of the match rates reported above for the academic studies were developed by algorithms that would pass as top of the line today, and, of course, future matching capabilities should improve along with the underlying quality of recorder of deeds records.

This research literature, particularly the merge evidence documented by the Federal Reserve Bank of Boston working paper, demonstrates that old HMDA data disclosed to the general public can be merged with property records to achieve a re-identification rate of 75%. This merge was achieved using only census tract, lender id, loan amount and date, and the mortgage purpose (purchase or refinance).

Higher match rates could have been achieved with a more complex program that considered additional variables. As noted below, there are algorithms for matching the borrower(s) surnames to ethnic or racial identities. As discussed below, the CFPB itself has claimed substantial precision for these methods. Use of such an algorithm would allow further matching to the racial and ethnicity variables disclosed in HMDA data. Because racial and ethnic minorities comprise a small percentage of borrowers in HMDA data, the potential to raise the match rate from *75%* to *100%* for these groups is significant for all HMDA records where the race and ethnicity fields are filled in and the responses are accurate.[8] These arguments are discussed in more detail below. The academic studies reviewed above did not consider refinancing or home equity loans but the status of all liens, first, second, home equity, are filed with the recorder of deeds records and separate matching by seniority of collateral and loan type is possible. Finally, the Federal Reserve, in 2002, amended Regulation C and added information on loans with high rate spreads. In areas where online property records include information on interest rates, this could be used to raise match rates. Of course this also tends to impact minority and low-income borrowers who, research has demonstrated, are more likely to obtain loans with high rate spreads.

In sum, the *75%* re-identification rate arrived at in this literature should be viewed as a lower bound achieved using only a fraction of the variables available in old HMDA data. Furthermore, the average re-identification rate underestimates the match rate for racial and ethnic minorities whose identification can be augmented using the race and ethnic identifiers in HMDA and the fact that they are more likely to obtain high cost credit.

The current CFPB website statement on the protection of consumer privacy in old HMDA data is not consistent with available academic studies that have matched available loan

[8] Gerardi, et. al. note that "there were many instances in which the race of the household taking the mortgage was not determined in the HDDA data." Pg. 10.

level information to HMDA files. This is even more remarkable given that some of these studies were performed at regional Federal Reserve Banks. In view of the current academic literature, a more factual public statement regarding privacy in old HMDA data would be that, for borrowers getting loans from lenders reporting to HMDA, the assumption should be that they have no consumer privacy protection. All of their personal information disclosed by HMDA can be linked to their names and property addresses. Under these circumstances, consumer privacy protection would be enhanced if applicants were first informed of the lack of privacy in HMDA data and then given the option of opting out of having their information included in HMDA data collection. Such opt-out provisions were included in the FCRA.

Sample Re-identification Exercise Using old HMDA data (from 2014)

The CFPB is correct in its discussion of privacy of old HMDA data in stating that census tracts generally have a population of about *4,000*. Given that the smallest geographic unit used to identify the location of a property secured by a mortgage included in HMDA is a census tract, the *4,000* number appears to provide substantial immunity against re-identification of individuals whose mortgages are recorded in HMDA.

The purpose of this section is to illustrate how the re-identification process works in an environment where it is particularly challenging. The case example is Montgomery County, Maryland where a population of *1,040,116* is spread over *233* census tracts yielding an average of *4,464* persons per tract. Population per tract tends to be larger in growing areas because of the lag in splitting tracts in response to population increase. Larger population should provide greater anonymity.

Average household size, *2.75*, is not large. This is important because mortgages are associated with housing units and *4,464* persons at *2.75* persons per housing unit implies an average of *1,623* housing units per tract. Turnover of housing units is *14.2%* per year so that the potential for mortgage activity is large and *66.6%* of the housing stock is owner occupied implying substantial potential to have transactions involving owners rather than investors. All of these factors mark the county as a place where individual census tracts are likely to generate large numbers of home purchase mortgages per year compared to other locations.

Finally there is the question of anonymity for minority households. Montgomery County has only *45.2%* white, non-Hispanic households*, 19%* Hispanic, *15.4%* Asian/Pacific Island and *19.1%* black households. This means that minority households are less likely to be uniquely identifiable in the county because there is substantial diversity. Put another way, the incremental value of information on ethnicity and race provided in HMDA is less likely to be useful in identifying individual mortgagors in Montgomery County than in other, less diverse, parts of the United States.

Overall the problem of re-identifying individuals whose home purchase mortgages are included in the 2014 HMDA data for Montgomery County, Maryland is likely more challenging that for most other U.S. counties. The first point about re-identification of these individuals is that, although there is an average of *1,623* housing units per census tract, there are only *10,015* purchase mortgages recorded in the 2014 data. Spread over *233* census tracts, this implies an average of only 43 mortgages per census tract per year. Given that observations can be identified, within tracts, by the loan amount and lender, this average is small enough so that unique re-identification should easily be possible. However, the actual distribution of mortgage activity is very uneven. Specifically *60* tracts have fewer than *30* observations, *74* have at least *30* but less than *50*, and *9* have more than *100*, re-identification is trivial for all observations except in the *9* tracts where observations are concentrated. Put another way, the probability that two mortgages have the same lender (there are *324* different lenders identified in the data), and the same loan amount) in census tracts with fewer than *100* observations is remote. Indeed a check of census tracts with observations between *80* and *100* revealed no cases in which the combination of census tract, lender, and loan amount failed to identify uniquely a single HMDA observation. Given that lender, census tract, and loan amount can be identified in local property records, matching these observations is relatively easy except for cases where there are errors in the data sets being merged.

Insight into the ease of re-identification can also be gained from tabulating HMDA responses by lender. The *10,015* observations are spread over *324* lenders for an average of *31* per lender. For a lender with only 31 loans endorsed in 233 census tracts, cases in which there was more than one loan per tract would be exceptional. These few cases could easily be identified by differences in the loan amount and matched to names in local property records. However, mortgage activity is even more unequally distributed across lenders than it is across census tracts. For example, *177* lenders have fewer than *10* mortgage originations in the entire county. Another *71* have at least *10* but fewer than *50* mortgages for 2014. Unique identification for these cases using census tract location plus loan amount is easily achieved. Furthermore, it is not clear what purpose is served by reporting data on lenders whose level of activity in a given area is so low. It is hard to imagine a statistical inference relating to mortgage location that could be informed by a data set consisting of fewer than *50* mortgages spread over *233* census tracts given that a minimum of *233 – 50 = 183* of the tracts would show zero activity. Indeed, other than re-identification of these borrowers, it is difficult to see a reason for disclosing the details of their mortgage activity at the tract level. The county would be a more appropriate geographic unit although re-identification of lenders with fewer than perhaps *30* mortgages per county could be accomplished by using loan amount alone.

However, *30* lenders have more than *100* mortgages in the county. These thirty cases present a more significant re-identification challenge. By far the greatest re-identification task is associated with the lender with *636* mortgages in the county for 2014 because the next largest lender endorsed *443* mortgages and others of the 30 top were all under *330*.

Having isolated the *636* HMDA observations from a single lender that are the most difficult to re-identify, the next task in this effort is to illustrate the ease or difficulty of re-identification to determine how many of these cases cannot be matched to names in local property records.

This high-volume lender had mortgages in *183* of the *233* census tracts. In *156* tracts volume was *5* or fewer, in *24* it was *6* to *10*, and in only three tracts was the volume of lending greater than *10*. Clearly these three highest volume tracts create the greatest potential for preserving borrower privacy and merit more detailed examination for possible cases where re-identification is not possible. After lender and census tract are used to identify borrowers, the third piece of identifying information in HMDA is loan amount which is rounded to the nearest thousand. The numbers of borrowers for this largest lender in the three census tracts with over *10* mortgages were *25, 15,* and *13* respectively. These were all tabulated by loan amount, requiring a margin of ±*$2,000* to separate mortgage amounts to deal with the possibility that rounding error made identical mortgages appear different. A total of four cases where loan amount was identical (given the ±*$2,000* margin) for *2* borrowers were found, two each in the tracts with *25* and *15* borrowers and none in the tract with *13* borrowers. In no case was loan amount similar for *3* or more borrowers. These *4* cases can then be examined for differences in responses to the ethnicity and race questions because this is the last resource to establish borrower identity when lender, census tract, and loan amount are not jointly sufficient. In one case, there was a single white Hispanic borrower paired with a white non-Hispanic couple and another could be distinguished because there was no co-applicant. However, two pairs could not be separated because, in each case, the race and ethnicity information was not supplied. The failure to supply information prevented re-identification.

In this most challenging census tract with *25* borrowers from a single lender with *2* cases of identical loan amounts, matching with information from county property records succeeded in uniquely identifying the names of *18* of the borrowers. This resulted in *a 72%* re-identification rate under the most difficult of circumstances. Obviously cases with fewer than five borrowers per tract should have a *100%* re-identification rate barring data anomalies. In sum, this step-by-step analysis has shown that, even in the "worst case" considering a large lender and picking a census tract where that lender was most active, the re-identification rate was *72%.* Simply put, there is very little consumer privacy protection in the old HMDA data being disclosed to the public.

Identification of homeowners in HMDA data is easy

Identification is the inverse of re-identification. Starting from a property record giving borrower(s) and lenders names, a specific property address, date of the mortgage (actually all liens are identified separately in the record), along with house value and a variety of other information not used in the matching process, the problem is to find the matching transaction in HMDA data. The probability of identification depends initially on the probability of using a lender required to report to under HMDA regulations. Clearly, individuals who do not use a reporting lender cannot be identified. This, of course, is why HMDA data should never be used to measure credit access in neighborhoods. As noted above, in some neighborhoods, HMDA includes less than have of the home purchase transactions.

Generally identification of the transaction in HMDA data conditional on the loan coming from a lender reporting to HMDA is even easier than the re-identification problem discussed above. Individual homeowners have virtually no privacy in current HMDA data. All that is needed is the location of a *1-4* family home. This address can be taken to the cadaster, i.e. property records, maintained by each local jurisdiction. That record, which is necessary to record property title and liens, has the name(s) of the owners, and information on the mortgagee, the date of closing, the initial loan amount, loan purpose (purchase or refinance), lien type, and the sales price for purchases. Indeed the entire note and deed of trust are often scanned into the online files. This means that the note rate, payment schedule, and signatures of the mortgagor and trustee are available.

Even without this extra information, matching to the mortgagor's HMDA data record is easily accomplished. Simply select the appropriate year of HMDA data, match the property address to the local census tract map and look up the HMDA identification number of the mortgagee. Then using the appropriate year of the online HMDA data, the match is accomplished by going to the census tract, sorting on lender identification number to get the correct lender, and then matching the loan amount from the property records with the loan amount in HMDA. In rare cases, see discussion below, there may be two loans from the same lender in the same year and census tract for the same loan amount. In these exceptional cases, final identification can be made if the race and ethnicity of the borrower can be deduced from the borrower name or names in the property record. Obviously minorities are more easily identified because most borrowers in HMDA data are non-Hispanic whites. Another possibility is using the note rate to identify high rate spread mortgages which are separately indicated in HMDA data since the FED's 2002 amendments to Regulation C.

In sum, it is relatively easy to match particular homeowners to HMDA data unless they are non-Hispanic whites who borrow from a lender that has a large number of mortgages in the same census tract that are endorsed in the same year. In such cases there may be more than one record with the same combination of lender, census tract, loan amount, and borrower race and ethnicity. Other property owners are easily matched to HMDA records provided that they secure financing from an entity required to report under HMDA.

How does lender size influence identification and re-identification in HMDA data?

The volume of loan activity by a lender has a dramatic effect on the ease of identifying borrowers. The reasoning behind the inverse relation between lender volume and identification or re-identification probability is easily understood. Property records contain the names of the borrowers, and lender along with the property address which is easily converted into a census tract location. HMDA includes both census tract and lender identification. There are *73,000* census tracts in the US with population ranging from *1,500* to *8,000* and averaging *4,000* persons or slightly under *2,000* households. To put this in perspective, given that there are *3,100* counties (or equivalents) in the U.S., there are *24* census tracts in the average county. Furthermore, mortgage originations can be disaggregated into purchases versus refinances for the purpose of re-identification because the loan purpose is evident in property transaction data. Purchase loans are first liens on *1-4* family properties that are not refinances.

Given that census tracts are small, it is obvious that small lenders, even if their market area only spans a few counties, have a very small number of mortgages per census tract. This makes re-identification very easy for borrowers patronizing these lenders and the problem is that most lenders reporting under HMDA are small. In 2015, *6,913* institutions were required to report HMDA data. The *25* largest institutions accounted for approximately one-third of all loan originations and more than half of loan purchases. At the other end of the scale, fewer than *100* originations were reported by almost half, *3,071*, of the reporting institutions.[9]

Clearly, even if institutions with very small mortgage operations have market areas that extend over only a few counties, the annual number of originations per census tract is so small that re-identification is virtually certain based only on lender name and census tract location. An institution originating *80* mortgages, half of which are purchase mortgages, over the average *4* county area, is only originating an average of *0.4* purchase mortgages per census tract. Clearly, even without loan amount and ethnicity to use in re-identification, the problem of matching mortgages originated by these small institutions to names of borrowers in property records is trivial.[10]

This does not imply that re-identification rates automatically fall as the size of the institution grows because the spatial coverage of its market area also rises and lenders often specialize by loan type. Consider Nationstar Mortgage, the *14th* largest institution in terms of total originations. Because it specializes in refinancing, it originated only *1,000* purchase mortgages in 2015. This loan volume spread over even a small fraction of the *73,000* census tracts in the U.S. would be easy to re-identify. The *49,000* mortgages refinanced by Nationstar in 2015 would be more of a challenge to re-identify, particularly for white borrowers in suburban areas where housing is generally owner-occupied and refinancing volume was high. Thus both

[9] Tabulations of HMDA data for 2015 referred to in this section are taken from Neil Bhutta and Daniel R. Ringo, "Residential Mortgage Lending from 2004 to 2015: Evidence from the Home Mortgage Disclosure Act Data," *Federal Reserve Bulletin,* Vol 102, No 8, November, 2016.

[10] It is also possible to take the position that the relevance of these 3,071 institutions to any national need to understand mortgage flows would be difficult to demonstrate logically given that they account for less than 2.5% of all mortgage originations. The national significance of this lending for mortgage flows is trivial compared to other sources of mortgage finance, such as owner financing, cash purchases, etc, that are not covered in HMDA data.

the size of the lender and its degree of specialization in either home purchase or refinancing are important determinants of the ease of re-identification.[11]

At what point does lender volume become so large that it is difficult to re-identify borrowers? As noted above, the first consideration is that volume must be disaggregated into purchase versus refinancing (or home equity, etc.) because liens filed with the recorder of deeds allow identification of purchase versus refinancing activity. Then, given the volume of activity in each of these categories, the next question is the geographic variation in coverage of the lender. SunTrust Mortgage Company is the *23rd* largest lender by total originations in the 2015 HMDA report. Its activity was balanced between purchase mortgages (*15,000*) and refinancing (*22,000*). These numbers seem large enough to offer a degree of anonymity to most borrowers. But for a national lender with activity spread over *73,000* census tracts, it is evident that most borrowers must be located in tracts with fewer than *5* SunTrust mortgages. For minorities, these borrowers can be identified by race and ethnicity and, for white borrowers, the identifying information added by mortgage amount usually provides a unique match that achieves re-identification. Thus, even among the top *25* mortgage lenders reporting under current HMDA, it is relatively easy to re-identify borrowers if the lender serves a national market and the information provided in HMDA is accurate.

The most challenging scenario for re-identification would be a lender like Wells Fargo that topped the 2015 list as the lender with the most reported HMDA loan originations balanced nicely between purchase mortgages (*156,000*) and refinanced loans (*186,000*). With an average of between *2* and *2.3* borrowers per census tract, Wells has some cases in which there are as many as *20* or even *25* purchase mortgages in a given census tract. For this reason, the special case of re-identification in such tracts is illustrated in a practical example on page 14 of this report.

This is a good case to use to illustrate the differential effect of borrower race and ethnicity on re-identification. Publically available property records have Wells Fargo as the lender, the borrower's name, and loan amount. Using techniques adopted by the CFPB and discussed in the next subsection, it is possible to determine if the borrower is Asian, black or African American, Hispanic white, white non-Hispanic or other minority such as Native American based on the first and last name. Race and ethnicity information are also included in current HMDA data.

Assume that a single tract has *30* Wells Fargo purchase mortgages in a given year and further assume that CFPB techniques are successful in identifying race and ethnicity based on name(s) of the borrowers in the property record.[12] Thirty is a very large number but will help to

[11] This discussion is easily extended to home equity lines or credit and second trusts because these are also shown as liens in the property records. For simplicity the discussion focuses on new purchase mortgages.

[12] The claim that it is possible to identify race and ethnicity based on names alone is controversial. While this paper takes no position on the issue, the CFPB has a position, that such matching is possible. Of course the logic of the

illustrate the argument regarding differential ease of re-identification. Based on HMDA averages the expectation is that *5.3%* or *1.6* of these borrowers are Asian, *5.5%* or *1.6* are black, *8.3%* or *2.5* are Hispanic white, *68.1%* or *20.4* are white non-Hispanic, and only *0.8%* are other minorities. Average responses for borrower and co-borrower that cut across groups total *3.5%.* Finally *8%* of responses or *2.4* out of *30* are missing. Thus, even when there are a total of *30* purchase mortgages reported by Wells Fargo in a single census tract, on average only *2* or *3* will be Asian, black, or Hispanic borrowers and there will be another *2* or *3* cases where ethnicity is missing. Separating these cases to achieve unique re-identification by mortgage amount is not difficult. When there are only *5* or *6* cases (perhaps *3* black and *3* missing) whose race-ethnicity is either identical or missing, these can be readily identified by matching loan amounts in the public record with the amounts recorded (and rounded to the nearest thousand) in HMDA. This is not difficult even in the extreme case where a single lender has *30* purchase mortgage originations in a single census tract in a given year. Furthermore, the use of borrower and co-borrower gender could further simplify the matching process.

However, it may be difficult to separately identify all of the *20*+ white, non-Hispanic borrowers and *2.4* missing for a total of *23* who purchased homes in this single census tract using Wells Fargo unless their mortgage amounts are very dispersed. This illustrates the rational for the conclusion in this report that there are two ways to avoid re-identification in HMDA data. One is to be a white, non-Hispanic borrower using one of the *10* largest lenders and purchase in a suburban location where a substantial fraction of the housing is owner occupied and there is relatively high turnover. The other is to use a large lender in such a suburb and to fail to reveal your race or ethnicity or to provide false information.

Why are identification and re-identification easier for minorities?

The previous section gave an example of the differential ability to re-identify minority borrowers in the hypothetical case where a single lender made *30* purchase or *30* refinances in a single census tract. This section generalizes that discussion.

For any data that contains information on race and ethnicity, the possibility of differential effects on privacy based on race and ethnicity arises. HMDA has a full set of racial and ethnic identifiers. Initially lenders were not required to collect information on race and ethnicity. From 1991 to 2001, the fraction of loan reports lacking this information grew from *8% to 30%* as loans were increasingly made without face to face contact and lenders were not inclined to ask applicants for this type of information. Furthermore neither, race or ethnicity appears in credit reports or other sources used by lenders. In 2003, lenders were required to ask for race, ethnicity, and gender when taking telephone applications. This information is also solicited on internet applications. Of course it need not be given and there is no way in which the lender can validate responses.

CFPB position is that this makes re-identification of HMDA data much more likely than if names cannot be matched to the race and ethnicity reported in HMDA.

Old HMDA data includes separate information on ethnicity as Hispanic. Therefore, any race, white, black, Asian/pacific islander, American Indian/Alaskan native, or other, may be combined with Hispanic ethnicity. Differences in size of these groups among those reporting race and ethnicity in HMDA are huge. For example, white-non-Hispanic is *83%* of the total Hispanic respondent population. Black-non-Hispanic, Asia/Pacific Islands are all in the *4-6%* range. American Indian/Alaska Native and multiracial or other are both < *1%* of HMDA. Put another way, for non-Hispanic whites, disclosure of race-ethnicity does almost nothing to aid the process of matching HMDA data to other sources because it only narrows the matching pool from *100%* to *83%.* But for other groups, the presence of ethnicity and race reduces failure to achieve potential matches from *100%* to *6%* or less, i.e. a *94%* reduction in the difficulty of re-identification. Matching HMDA data to any other data set with information on race and ethnicity is more than ten times easier for minority groups than for non-Hispanic whites.

Information on ethnicity and race in HMDA data raises the re-identification rate with other datasets, particularly because the local recorder of deeds and property tax records in which the surname(s) of the owner(s) are present along with the property location generally has no information on ethnicity and race. Differential re-identification of minorities can be achieved in cases where ethnicity and race can be determined using a combination of surname(s) and property location. Property location is a potentially valuable source of identification for groups that tend to be spatially segregated.

The USCB maintains tabulations of the relation between specific surnames and ethnic characteristics of the population. Recently, the CFPB produced its own estimates of the precision with which information on surname(s) and property location may be combined to infer race and ethnicity in mortgage application data.[13] Rather than use surname alone, the combination of surname and residential location is used to infer race and ethnicity. The CFPB experimented with a sample of mortgages supplied by lenders where surnames, property location, and race/ethnicity were supplied. About 12% of the sample was dropped because names were not on the Census list (mainly) and/or property address locations could not be geocoded. Table 3 of the report is reproduced below.

[13] The statistical matching technique, known as Bayesian Improved Surname Geocoding (BISG), is discussed in "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment," CBPB, Summer, 2014.

Table 3: Correlations Between Proxy Probability and Reported Race and Ethnicity[14]

Proxy Method	Hispanic	White	Black	Asian- Pacific Is	American Indian Alaska Native	Multiracial
BSIG	0.81	0.77	0.70	0.83	0.06	0.05
Surname only	0.78	0.66	0.40	0.81	0.03	0.05
Location only	0.45	0.54	0.58	0.38	0.05	0.03

The results reproduced above from Table 3 indicate the CFPB position that using surname and location together, it is very likely that Asian/Pacific Islanders, Hispanic, white, and then black borrowers, in that order, can be successfully identified. American Indian/Alaska Native and multiracial borrowers are not easily identified using this information. It is important to recognize that this is a CFPB claim and decisions on the privacy implications of either old or new HMDA data releases by the CFPB should be consistent with its position on the ability to identify race and ethnicity from surnames.

Using the CFPB matching claims from Table 3, it is possible to show how this information on ethnicity and race is useful in establishing re-identification in HMDA data. This information can differentiate cases in which there are two or more HMDA observations, identical in all other aspects used in matching (census tract, lender identification number, and loan amount) of the borrower used in the identification process.

Consider the effect of the ability to identify ethnicity or race based on the surname and location from a local recorder of deeds. If an individual is a member of a group reflecting a fraction θ of the population which is disclosed in HMDA data, then the probability that the additional ethnicity or race information fails to differentiate them from another similar case with which they are paired is equal to θ^2 or the probability that both HMDA respondents are of the same race. Based on Table 3, for white borrowers in HMDA data $\theta = 0.83$. Therefore $\theta^2 = 0.68$ which reflects a substantial chance of not being re-identified based on the ability to detect race from surname and location. However, for other groups the role of surname and location identification is very likely to provide the information needed to differentiate in such cases. Comparable calculations are: Hispanic borrowers, $\theta^2 = 0.003$, black borrowers, $\theta^2 = 0.004$, Asian and Pacific Islanders, $\theta^2 = 0.002$. Accordingly it appears the ability to identify ethnicity and race from the data files containing local property records makes re-identification far more likely for minorities, other than American Indian/Alaska Native and multiracial where name recognition is poor, than it does for white borrowers.

[14] See, "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment," CBPB, Summer, 2014, page 14.

Indeed the difference in the probability that revealing ethnicity and race in HMDA data allows re-identification in cases where other information is ambiguous is over 100 times greater for these minority borrowers than it is for white borrowers. Put another way, because of their large numbers, white borrowers are more difficult to re-identify even if their race is revealed in HMDA data. These differences will be apparent in the re-identification exercises conducted later in this report.

II. How would disclosure of the new HMDA data influence the ability to re-identify individuals and the consequences for their privacy?

The probabilities of identification and re-identification are already very high using the old HMDA data set. Some of the additional data fields that would be reported to the CFPB under the new rule would raise the probability of re-identification to a near certainty provided that the data are recorded accurately. Table 1 below lists the current variables collected old HMDA and released to the public, additional variables identified as required in Dodd-Frank, and variables the CFPB has added under its discretionary authority that could be part of a new HMDA release. The table separately identifies variables useful in identification or re-identification because they can be matched to information in property records. It also notes variables that reduce privacy because they are not generally available elsewhere. Particular attention is given to variables that reveal the creditworthiness of borrowers because personal financial information has specially legislated privacy protections under FCRA and the Right to Financial Privacy Act (RFPA). Indeed, given the privacy accorded under FCRA and RFPA, borrowers should have a reasonable expectation that information which is provided to a bank in connection with a loan application will not be revealed under HMDA.

As established previously, using current HMDA disclosures on loan amount, lender, date, loan type and purpose, and census tract, rates of identification and re-identification well above *75%* are possible. Adding current disclosures of race, ethnicity, gender and number of borrowers raises this probability significantly for anyone sophisticated enough to use name recognition software to match borrower names in the property record to their likely demographic characteristics. The extra racial and ethnic disaggregation in the variables added pursuant to Dodd-Frank make identification and re-identification for minority groups a virtual certainty. How many females of Japanese heritage get mortgages in a given amount and year from the same lender in the same census tract? Likely the answer never exceeds one and hence that one borrower can be identified or re-identified easily.

TABLE 1

HMDA Data Variables: Currently Disclosed And Possible Additions

Variables Currently Disclosed by Category

Lender Identifier	Preapproval?	Approved?	**Ethnicity**	**HOEPA****
Application Date	Construction Type	Date of Action	**Race**	**Rate Spread****
Loan Type	Owner Occupied?	**State**	**Gender**	**Lien Status**
Loan Purpose	**Loan Amount***	**County**	Income**	Denial Reason
		Census Tract	Purchaser	

Additional Variables Identified in Dodd-Frank

Property Address	**Points**	Prepayment Penalty	**Ethnic Detail***	Mortgage Loan Originator ID*
Age*	**Fees**	**Introductory Rate Period**	**Racial Detail***	
Credit Score**	**Property Value**		Application Channel	
Loan Term		Amortization Rate		

Variables Added by the CFPB

Origination Charge	Debt/Income**	**Manufactured Home Details**	**Total Units**	**Reverse Mort****
Discount Points*	**Combined LTV****		**Affordable***	**Open-End Credit***
Lender Credits			Automated Underwriting Details	Business or Commercial Purpose
Interest Rate**				

Key:

BOLD variables are useful in identification or re-identification because it is available from public records and can be used in matching. Race, ethnicity, and gender can be inferred by analyzing names.
* indicates a variable that provides private information on the borrower or lender
** indicates a variable that reveals private information on the creditworthiness of the borrower

Information on the house value, whether revealed directly in the list of variables explicitly included by Dodd-Frank or obtained as the quotient of loan amount and loan-to-value ratio (LTV) can be matched to property or appraisal records and improve re-identification rates. Naturally property address matches property records perfectly. Equally obvious is the fact that including an identification number for the originator allows that individual to be identified. In some jurisdictions property records include detailed information on the mortgage instrument, including points, note rate, etc. that could be matched to these components of the new HMDA data being collected and considered for addition to HMDA disclosures. Reverse mortgages uniquely identify the elderly and should be quite valuable to those looking for potential targets for fraudulent schemes. Knowing that credit was extended for special purposes, such as mobile homes (particularly when land is encumbered) or for affordable housing, allows further assurance that the re-identification matching process has been successful.

Although, the implications of the proposed additional HMDA variables for identification and re-identification are significant the effects on individual privacy are far more extreme. Current HMDA data discloses some personal financial characteristics such as income and high interest rate that are not easily available elsewhere. The additional HMDA is far more intrusive.

The consequences for privacy depend on whether these aspects of personal data are currently well protected so that matching to HMDA would reveal personal characteristics that households could reasonably them expect to remain private in the absence of disclosure in HMDA data. Effects on privacy depend not only on the particular data item but how it can be used in conjunction to other items to infer the condition of a borrower. Obviously revealing both the loan amount and the loan-to-value ratio allows computation of the sales price of the housing unit, or the appraised price if loan-to-value is based on the appraisal. These connections arise with mathematical certainty. But other connections are more subtle. The back-end-ratio is the ratio of total monthly debt payments to monthly income. If income and mortgage amount are disclosed along with the back-end-ratio, the borrower's non-mortgage debt can be approximated. The computation is made even more precise if the front-end-ratio and note rate are disclosed.

Perhaps the most important element of privacy in lending is the borrower's general level of creditworthiness. This is conventionally measured by credit score which is computed from items in the borrowers credit history file. Credit history data has been given substantial protection under FCRA because individuals do not wish their creditworthiness to be revealed. For example, there are state restrictions on the ability of employers to use credit history data in evaluating applicants and permission to access even limited information held at the credit repositories must be given explicitly. Clearly any proposal to reveal information on applicant credit scores would be inconsistent with the privacy protections given to this information in other transactions such as applying for credit, employment, etc.

In addition to credit score, other variables reveal important information on borrower creditworthiness because lender set terms of the mortgage transaction based on credit score and other elements of credit history. Two of the variables currently revealed in HMDA indicate borrowers whose creditworthiness was judged to be relatively low. For example the current HMDA disclosure that the rate spread to Treasury debt is unusually high or that it is high enough to equal Home Ownership and Equity Protection Act (HOEPA) threshold reveals a great deal about the creditworthiness of the borrower.

Collection of additional information on the note rate combined with other information such as the loan-to-value and payment-to-income ratios will allow educated observers to form a very accurate estimate of borrower creditworthiness even if credit score is suppressed. The note rate would provide information on the status of all borrowers. Most individuals who would object to disclosure of their credit history or score may not recognize the interest rates that they are paying on debt, combined with LTV and payment-to-income ratio, can be used to infer the credit score used to underwrite their mortgage application. The classification of the privacy implications of variables listed in Table 1, in the interest of consumer protection, keeps track of these indirect privacy considerations.

III. Would the release of the new HMDA data be an additional threat to consumer privacy, or are these data already available elsewhere?

To answer this question, it is instructive to review alternative sources of data on consumer finances, particularly insofar as it concerns mortgage market activity and/or creditworthiness. Such data is collected and released by local governments, private entities, and the national government. Indeed it is by matching local data with HMDA data that identification and re-identification are achieved. Thus the consumer protection required in collecting and releasing HMDA data should be considered in relation to policies adopted for these other data sets.

Local Recorder of Deeds, the Cadaster

Efficient operation of real estate markets requires the maintenance of a cadaster that records ownership of real property and liens encumbering that real estate. The cadaster must also include information on liens outstanding so that purchasers can secure clear title. Indeed, transfer of title to property is said to be “perfected” when it has been registered with the appropriate public authorities who maintain the cadaster. Traditionally, maintenance of the cadaster has been the province of an office with a title like “registrar of deeds.” This is usually an instrumentality of the local, county or municipal, government and may also maintain a register of licenses, etc. Given the economies associated with data processing, it is common to find that local registrars have combined their records to provide uniform state-wise coverage of real estate ownership and transactions.

The record of deeds literally includes copies of documents such as the deed and deed of trust for real estate transactions. These documents contain information on location and nature of the real property, the terms under which title has been transferred, and the detailed nature of any liens secured by the property. Not only the names but the signatures of the individuals involved in the transaction are provided. Signatures of the mortgagee(s) and the representatives of the mortgagor and others involved at closing can be observed. Much of this detail is not directly relevant to the issues discussed in this report but it is important to note that it is available, online, in a public record.

The deed includes information on property location and characteristics, the purchase price, and the names of the seller(s) and buyer(s). The deed of trust includes the name(s) of the mortgagor(s), mortgagee, trustee, and the amount of the loan as well as payments required to avoid default. Some information on the nature and characteristics of the mortgage is apparent. Many deeds of trust involve financing provided by individuals or entities that do not report to HMDA, including seller financing. Other liens on the property are also recorded, including second mortgages, tax liens, contractor's liens, and other types of debts that have a claim against the property. When mortgages are transferred, this generates a deed assignment document which is also recorded along with the release of the deed of trust when the mortgages is paid off or prepays (often at sale). Refinance transactions generate a deed of trust with the same information on loan amount and names. This is all essential information for buyers who wish to have clear title when they purchase a property. Given information on the sale price from the deed and the loan amount from the deed of trust, computation of the initial loan-to-value ratio follows easily.

The information in online data from local recorders of deeds is not standardized and there may be charges for access to the files. For this reason, a number of vendors have developed programs to access key elements of this information, such as the names of the mortgagor and mortgagee, sales price, and initial loan balance. This information is available for purchase. It plays an important role in allowing the re-identification of HMDA respondents.

Local tax and revenue office

In order to reduce the costs of administering local property taxes, particularly the cost of appraisals, local government tax and revenue offices maintain files of property characteristics, location, and recent sales price(s). These files include the names of the property owner(s), current appraised value, amounts of any tax liabilities, and usually the name of the mortgagee, including entities that do not report to HMDA. Owners can use this online data to compare their assessments and tax bills with comparable properties. It is a relatively simple matter to merge information from these files with deed recordation information discussed above because property address appears in both files. A number of firms have developed automated systems that perform this matching service and link information on the name of the mortgagor and mortgagee with sales price, loan amount, tax information including appraised value, and detailed property characteristics. Data from these firms is commonly used in economic research as well as for a number of private purposes by investors and developers.

Consider, for example, the very user-friendly website of the Washington, D.C. Office of Tax and Revenue at:

https://www.taxpayerservicecenter.com/RP_Search.jsp?search_type=Assessment

It is possible to search by lender name and retrieve information on every property in the District of Columbia with a lien from that lender. The pop-up table gives the lot location in the local cadaster, the street address, owner names, most recent sale price, date that the lien was recorded, property use code, and assessed value of the property. It is then a simple matter to match the property address to the census tract map, look up the lender identification number and match the observation to HMDA data for the appropriate year. It is also possible to search by property address, sale date, use code, etc.

Multiple listing services (MLS)

The traditional multiple listing service has been supplemented by a number of competitors that give very detailed information on property characteristics, even including pictures of the interior and exterior. This information, is usually entered by a realtor in order to facilitate sale of the property. Accordingly, substantial detail on the property, including address, interior space, lot size, rooms, year build, year of last renovation, etc. are provided. Photographs of the interior and exterior are supplied. The sales price and loan amount, but not the identity of the owner or lender, are also available.

Because sales price, date of sale, and address are included in the MLS data, it is easily matched at a rate approaching 100% with tax and loan data described above. Although the identity of the seller, buyer, and lender are not included in MLS data, they can be identified by matching with recorder of deeds and tax and loan data. Furthermore, a data set such as HMDA data on purchase mortgages that can be matched with recorder of deeds or tax and loan data can also be matched with MLS information supplied at the time the property was sold. As a general proposition, any of the data sets discussed here, once matched with tax and loan data, can be matched with every other dataset. The result is that all information in these data sources can be associated with the personal identify of the buyer and seller of residential real estate because property records also include the name(s) of the seller. However, MLS listings do not include personal information, and certainly not financial information about the seller or buyer.

GSE Single Family Loan-Level Datasets

The Housing and Economic Recovery Act of 2008 (HERA) required the GSE regulator, the Federal Housing Finance Agency (FHFA), to collect and publish individual loan data on mortgages purchased by Fannie Mae and Freddie Mac. The GSEs produce several data sets but two examples, discussed in some detail below, are most closely related to HMDA and to the re-identification issue are discussed here.

GSE Enterprise public use databases

While these data sets are very large, covering millions of mortgages endorsed since January, 1999, this is still a sample of all GSE purchases and the overall GSE market share of mortgages is under fifty percent. Nevertheless this is a very large data set and both GSEs are identified in current HMDA data so that mortgages eligible for inclusion in this data should also appear in HMDA data. Names of the mortgagor and mortgagee are excluded.

The key variables in this dataset that can be used to link it to HMDA and other data include the census tract in which the property is located, property type, occupancy (owner or not), unpaid balance at acquisition, lien status and borrower demographic characteristics. Curiously the year in which the mortgage was endorsed does not appear to be in the data although most acquisitions should be recent endorsements. Unpaid principal balance should be close to the loan amount in other data sets unless the loan was seasoned for some time before purchase. Purchase by Fannie Mae or Freddie Mac and the presence of an FHA or VA guarantee are also indicated and can be matched to HMDA. The key variable used for matching in the old HMDA data that is not present here is the identity of the lender.

Because, a full set of identifying information on the demographic characteristics of the borrower and co-borrower is included it is likely that only members of minority groups can be re-identified. In particular, members of minority groups living in census tracts where they are a very small minority of homebuyers are likely to be identified. In contrast, anonymity of white non-Hispanic borrowers is preserved except in census tracts where they constitute a small fraction of homebuyers. The absence of lender identification in this dataset makes re-identification much more difficult than current HMDA data but also makes the inequality in re-identification rates across different demographic groups much larger than in HMDA data.

Matching this data to HMDA is aided by the inclusion of borrower income. For mortgages endorsed in the same year that they are sold, this income should be identical with reported HMDA income which would, along with the proximity of loan amount to unpaid balance in the first year, make matching to HMDA data straightforward. For mortgages endorsed in previous years, the matching process would be slightly more difficult because income is adjusted slightly to account for local income growth rates.

Beyond income and demographic characteristics, the data includes age, and rate spread paid on the mortgage. Thus there is a small amount of information not currently included in HMDA records. It is not clear that disclosures in this dataset have been analyzed for re-identification problems. In contrast to the next GSE dataset, there is no explicit discussion of masking procedures.

GSE Single family loan level dataset

This dataset is designed to track the performance of a sample of the mortgages purchased by the GSEs. There is a record for the origination data file and this is updated with a record indicating recent loan performance. Names of the mortgagor and mortgagee are not given. The origination data file includes credit score, mortgage insurance, debt to income ratio, month of origination, loan to value ratio (LTV) and cumulative LTV (including second mortgage amounts and home equity line balances), interest rate, prepayment penalties, loan purpose, term, original unpaid balance, channel and the name of the entity selling the mortgage to the GSE provided volume from that seller is sufficiently large. There is very extensive information on the mortgage and the financial circumstances of the borrower. The monthly loan performance data file includes the current unpaid balance, delinquency indicators, current interest rate, and information on costs associated with non-performance of the loan including expenses associated with default, short sale, and/or foreclosure.

Ability to re-identify the borrower or merge with other data sets is constrained by the fact that there is no demographic information on the borrower(s) and geographic identification is limited to the first three digits of the five-digit postal code. This information appears to be omitted with a specific intent to preserve borrower anonymity in a manner that is not considered in old HMDA data. Even more important, the data documentation contains a specific section on "masking" which indicates that the disclosures have been checked for possible identification and re-identification. Restrictions on disclosure, i.e. masking", are designed to eliminate matching to other data an ultimately to borrower names. For example, names of low volume sellers of mortgages are aggregated in an "other" category. Presumably these masking decisions were made based on the type of statistical analysis that will be recommended at the end of this report but detailed documentation of these efforts was not found.

There is one exceptional case where matching with property records to re-identify borrowers may be possible. Property records include details cases of the foreclosure process from filing of an initial notice in cases of default to final transfer of title in foreclosure. In areas where foreclosure is uncommon, it might be possible to identify and re-identify the borrowers whose names and foreclosure process details are available in property records. Given that the discussion of masking, did not include foreclosure as a special case, perhaps the masking efforts are not adequate for properties that have gone through any part of the foreclosure process.

Federal Home Loan Bank purchased mortgage files

As required under HERA, the Federal Housing Finance Agency (FHFA) also supervises production of the Federal Home Loan Bank Purchased Mortgage File

(HLBPMF) of information on mortgages purchased by the Federal Home Loan Banks (FHLB). As opposed to comparable information from the GSEs, this data includes both the date of mortgage origination and acquisition along with the initial loan amount and unpaid balance. The full census tract code is disclosed. Names of the mortgagors are not included but both the name and location of the mortgagee are disclosed.

Analysis of this data set indicates that, with the possible exception of data errors, the ability to match mortgages in this file with those in HMDA data is approximately 100%. The combination of census tract, original loan amount, and borrower income is almost always sufficient for a match without considering borrower demographic characteristics. This means that re-identification of borrowers in HMDA implies that the borrowers in this data set are also re-identified. Indeed, re-identification in this data set is even easier than in HMDA because substantially more information is provided.

The HLBPMF contains many fields not in current HMDA although many have been proposed for addition. These include the borrower's front and back end monthly payment to income ratios, physical characteristics of the unit such as bedrooms and bathrooms, the interest rate, private mortgage insurance indicator, and some detailed characteristics of the mortgage instrument. There is also information on credit score divided into 5 ranges, < 620, 620 to < 660, 660 to < 700, 700 to < 760, and 760 or larger for both the borrower and co-borrower. While, the amount of information provided in the HLBPMF on each mortgage transaction is substantial, the total volume of mortgages in the dataset is not large. For 2014, there were 34,314 mortgages in the database. Accordingly the number of individuals potentially affected by the ease of re-identification and additional information disclosed is small. However, the lack of anonymity experienced by borrowers whose mortgages happen to appear in this data set is astounding. Conversely, individual identification in this data is not a privacy issue in that only a tiny fraction of the population is included in the data set and the probability of matching an individual or property to this data is tiny.

Credit reporting agency data

The credit reporting agencies (CRAs) maintain credit histories for individuals under the Fair Credit Reporting Act (FCRA) as supervised by the Federal Trade Commission. Credit history information is not disclosed to the public but it is used in the lending process. Most mortgagors have sufficient history so that their individual credit score can be computed. The credit score is a function of the information in the credit history which includes most credit transactions and payment history along with public actions such as bankruptcy or foreclosure. Credit score is important in the lending decision and some individual elements of credit history are also influential.

There are three reasons for covering the CRAs in this discussion. First, credit score is either disclosed or at least categorized in some to the data sets noted above. Second, credit score is being collected in the post Dodd-Frank HMDA data. Third, the FCRA provides explicit privacy protections to individuals and there is substantial concern regarding the dissemination, use, and accuracy of credit history. Much of this discussion is based on the author's personal experience in working with data provided by CRAs or on other applications of CRA data that have appeared in the academic literature and reveal its characteristics.

Credit history data, which is the basis for computing individual credit scores, includes the names of individual creditors, amounts owed, payment history, delinquencies, charge-offs, etc. With regard to mortgage credit, the names of mortgagees, amounts owed at origination, payment history, unpaid balance, and foreclosure actions are observed. The address of the real property collateral along with comparable information on any other real property owned by the consumer is part of the records kept by the CRAs. This information can be accessed by consumers and, under FCRA, they are entitled to challenge its accuracy. Some information on employment history and income is also present but this can be incomplete.

In recent years, several significant academic studies have used credit history data to study various aspects of mortgage market behavior. These studies have involved matching data held by the CRAs with information from the lender's loan file that would have been used to create HMDA data. The Federal Reserve System has recently created the CRISM database which links credit history information with Lender Processing Services loan servicing data so that loan performance can be linked with the evolution of credit history. When the CRAs perform this matching, the borrower(s) is identified because they have names and addresses in the credit history files. However, the matched data is returned to the researcher in depersonalized because of disclosure limits on FCRA discussed below. The important points for this study are that high match rates are achieved by the CRAs and that substantial privacy restrictions have been placed on the data.

Under FCRA, individuals must be told if the information in their credit report is used to take an adverse action against them in an application for credit, insurance, or employment. To enforce this provision, there must be control over the dissemination of the credit report because the name, address and telephone number of the CRAs that provided the report must be disclosed by the user of that information. Second, individuals have the right to view and challenge the report that is being disseminated and the credit score based on that report. Third, while CRAs sell listings of names and addresses with credit scores that meet certain criteria to firms making contingent credit offers, individuals are allowed to opt out of this and have their information withheld from such requests. Fourth, explicit permission to access the report is required for employment purposes. Fifth, in cases where the individual has applied for credit or insurance, other types of business transactions, and certain licenses, permission to access the credit report is assumed to be implicit in the application process. Lenders who have committed to make credit offers to individuals, may also view the score. Finally there are a number of special cases, such as court orders, in which credit history can be accessed by authorized individuals. Note that research on credit markets does not satisfy these criteria and this explains why only depersonalized credit history information is released by the CRAs for academic studies. Overall, FCRA provides individuals with assurance of a substantial degree of privacy and control over their credit history and credit scores computed using that history.

Summary of identification and re-identification in non-HMDA data disclosures

This review of other sources of information on mortgage transactions and personal finances demonstrates that identification and re-identification in non-HMDA data sources is uneven. There is one case, the small FHLBB purchased loan database, that has identification and re-identification probabilities as high as those in the current HMDA data release. The GSE single family loan level dataset is exceptional in that conspicuous attempts to mask the identity of the borrower have apparently been made. Presumably these are in reaction to the requirement in HERA that the privacy provisions of the HMDA Act be observed.

Local registrar of deeds and tax offices provide information where identification of individual owners and properties is easily accomplished. All liens are identified, including both the lender and amount, along with purchase prices and dates of endorsement. However, such disclosure is absolutely necessary to allow real estate markets to function and personal financial information on the borrower, beyond the note rate which can be observed in the case of some localities, is not part of the data record.

A review of data on GSE lending shows that identification and re-identification probabilities are reduced substantially below old HMDA by limiting the degree of geographic detail on the location of the mortgage. Given that the FCRA restricts disclosure and allows individuals to opt out of disclosure of even the general range in which their credit scores lie, it is extraordinary that any data set for which identification or re-identification was possible would reveal credit score or credit history information. The anonymity promised to individuals under FCRA is certainly inconsistent with current disclosures in the FHLBB purchased loan data based and proposals to expand new HMDA data. It seems pointless to have FCRA place substantial privacy protections on the CRAs if new HMDA data is going to allow commercial data processers to match individual borrowers with their credit score or other indicators of creditworthiness?

The enterprise public use data sets have a curious intermediate position when it comes to privacy. Because the lender's identity is concealed, they have more privacy than old HMDA data. However, because a full set of demographic identifiers for the gender, ethnicity, and race of both borrower and co-borrower are provided, groups that are a minority of homebuyers, or at least a small minority in some census tracts, are readily matched to old HMDA and property records. In contrast, white non-Hispanic households have a substantial degree of anonymity.

An entirely different level of concern with privacy is evident in the GSE single family loan level dataset which has substantial masking provisions including suppression of borrower demographic characteristics and limiting location to the first three digits of the Zip code. At a minimum it appears that this GSE dataset has been subject to professional analysis of the conditions for identification and re-identification and, as a result, the disclosures of a variety of variables are limited significantly compared to current HMDA data standards. The concern with privacy, and particularly with privacy for all gender, ethnic, and racial groups varies enormously from old HMDA through the GSE enterprise data, to the single family loan level dataset.

IV. What principles should be used by the CFPB to determine the form in which new HMDA data will be released?

There are already statutory guidelines that relate to privacy in data collected by the government before it is released to the public. The Federal Information Security Modernization Act of 2014 (FISMA) requires the use of "privacy impact assessments" (PIAs) for personally identifiable information (PII). These are to be reviewed and updated annually. They are available to the public. The current PIA for HMDA is discussed below. It does not deal with the privacy issues raised in this report. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) appears to have a privacy protection requirement that compels the USCB to avoid publication of data that can be identified or re-identified. In this connection, the USCB conducts and publishes research on ways to censor data available to the public to make it less vulnerable to modern computer re-identification techniques.

Given that data from bank records is being disclosed, the Right to Financial Privacy Act (RFPA) appears to be violated by current HMDA disclosures. Certainly the privacy protections in HMDA itself (12 U.S.C. 2803), which are also echoed in HERA, require modification (masking) to preserve privacy. This type of masking effort has not even been implemented in the current HMDA disclosures. Curiously, the GSE single family loan-level dataset has explicit masking procedures that prevent disclosures of information needed for identification or re-identification. This contrasts with the current approach to HMDA because the above discussion documents the high rates of re-identification reported in scholarly papers and even accomplished by the internal staff of bank regulators. Accordingly it is instructive to contrast treatment of PII in HMDA data with that in other government sponsored datasets where anonymity protections are in place.

Privacy protection and government sponsored data collection

Much of the data used in social science research is either collected directly by government agencies or regulated by them. The data collection often involves individual or household surveys or enumeration. Some is collected in response to taxation or participation in government programs, particularly transfer programs. Clearly there is no privacy in data originally collected by the government because the individual is the respondent. In some cases, such as the Census of Population, individuals are compelled to respond, although fines for non-response or false response are small and seldom enforced. However, the USCB produces a Privacy Impact Assessment (PIA) for each of these data sets. This includes a detailed review measures designed to preserve confidentially. These procedures often include a requirement that users sign a confidentiality agreement, receive special training, and report findings in a highly aggregated, and hence depersonalized, manner. The efforts to preserve privacy and to disclose those specific efforts are necessary because there is no privacy in the data originally collected by the government. Privacy is based on what is done by the governmental authority charged with preparing a data set for release to or use by the public.

Government sponsored data, even if originally collected in a form that excludes the identity of individuals and organizations, can be linked with publically available data that enumerates individual characteristics through the re-identification process. Linking across data sets is a serious concern. What appears to be confidential information provided to or collect under the auspices of government can be linked to individuals and the confidentiality violated. As long as the government maintains control of the underlying data, privacy can be preserved by restricting access to users who have specific, narrow research purpose, removing personally identifying information, and masking to eliminate re-identification.

It is apparent that a privacy standard, perhaps in order to comply with CIPSEA, has been imposed on many government data sets containing information on consumer finances and mortgage liabilities. For example, the public version of the Current Population Survey (CPS) contains substantial household detail but geographic identification is limited to state and location in the "principal city," other metropolitan area, non-metropolitan area or other portion of the state. Suppression of geographic detail preserves confidentiality. Similarly the public release of the American Housing Survey (AHS) with its information on mortgages, house prices and characteristics, only disaggregates location to the level of center city versus suburbs (for large cities) and does not identify lenders. Again, suppression of geographic detail preserves anonymity. Finally, the Survey of Consumer Finances (SCF) also contains mortgage information but does not identify the lender or the location of the respondent in the publically available data file. In all these data sets, lenders are not identified and geographic detail is restricted to prevent re-identification because identification is already limited due to the modest sample size in the survey.

Presumably limitations on disclosure of lenders and location in the CPS, AHS, and SCF are based on research on disclosure avoidance techniques. The USCB has an active research program which publishes studies on the general anonymity problem as well as commentaries on particular issues in specific government data releases. Preservation of anonymity is considered an issue in these databases. If the same analysis were conducted for HMDA it would clearly show that there is no privacy in the current data releases and, obviously adding more fields makes the problem even worse.

The current PIA for HMDA data

The sharp contrast between the level of disclosure in current HMDA data and that in the CPS, AHS, and SCF suggests very different standards for anonymity are guiding release of these datasets. This suggests that the privacy impact statements (PIAs) for HMDA be reviewed.

A PIA should include a comprehensive analysis of how personally identifiable information is collected, used, shared, and maintained. The purpose of a PIA is to demonstrate that those responsible for data collection, management, storage, and release have formally considered and implemented privacy protections throughout the development life cycle of a system or program. PIAs are required by the E-Government Act of 2002, which was enacted by Congress in order to improve the management and promotion of Federal electronic government services and processes. PIAs should be posted on the appropriate government website and reviewed annually. Under the act, agencies are to meet ethical and legal obligations to respondents to respect privacy and protect confidentiality. In the case with HMDA data releases unless there is a very narrow interpretation of confidentially which holds that the released data itself cannot contain names, addresses, telephone numbers or other direct and unique identifiers.

An examination of the PIA for the CFPB at:

http://files.consumerfinance.gov/f/2016_cfpb_privacy-impact-assessment-supervision-enforcement-and-fair-lending-data.pdf

and the PIA specific to HMDA from the Federal Reserve at:

https://www.federalreserve.gov/files/pia_hmda.pdf

reveals that the discussion of privacy in old HMDA data is very narrow. Specifically there are requirements that the original data be collected carefully, handled and processed properly and depersonalized before it is released to the public. However, it is clear that the current interpretation of the PIA does not include any concern for the possibility that the data can be easily matched with personalized data. It also does not consider whether the borrowers are being uniquely identified by a combination of lender, census tract, mortgage amount and gender - racial - ethnic characteristics. This narrow interpretation of the PIA requirement by the Federal Reserve Board and CFPB is like saying it is proper to release data with the latitude and longitude of the respondent as long as the property address and name are suppressed. It is a simple matter to match latitude and longitude to a property address and the address with the name of the owner or current resident just as it is simple to compare data taken from tax and property records and identify or re-identify individuals from old HMDA data.

This not the type of PIA needed for old and particularly for new HMDA data if privacy is to be preserved. Instead the same standards used to insure confidentially for respondents to the CPS, AHS, and SCF that all contain information on mortgage finance should be applied to a PIA for HMDA data.

General principle for preserving privacy in both old and new HMDA data

Based on the previous discussion, a proper PIA for old HMDA would indicate that current data disclosures violate privacy and confidentiality of borrowers. In addition, there should be special concern with elements of the data collected that reveal the creditworthiness of the borrower, such as not qualifying for lower cost mortgage credit. Protecting information on individual financial condition is a special concern in FCRA, RFPA, and in the very language of the HMDA Act.

One of the paradoxes of the HMDA data set is that the fields that discriminate most against privacy of minority borrowers are collected, not because the lender wants or needs them, but rather as "Government Monitoring Information." These are the very fields that elevate the identification and re-identification probabilities for minorities above those of non-Hispanic whites. Furthermore there is significant evidence that these fields are unreliable.[15] These fields are not disclosed in the GSE single family loan level dataset where masking is has been implemented to preserve privacy.

Overall, a review of privacy issues in HMDA compared to other data on mortgages collected and released by the government or subject to government supervision, suggests the following general principal for HMDA data disclosures:

[15] Gerardi and Willen, op. cit. note "There are many instances of ownership for which we were able to match multiple mortgages but for which the race variable was not consistent. We threw many of the ownerships out of the dataset, unless closer inspection revealed an obvious assignment. For example, if 4 out of 5 matched mortgages for a given ownership listed race a black, while the remaining mortgage listed a different race, we assumed the household was black. In addition, there were many instances in which the race of the household taking the mortgage was not determined in the HMDA data. Jason Dietrict, 2002, Mortgage applications with missing race data and the implications for Monitoring Fair Lending Compliance, *Journal of Housing Research*, Vol 13, No 1, pp 51-84, states that "Applications from Asians and Hispanics appear more likely to be missing race data than applications from whites. These findings suggest that the denial odds ratio statistics and statistically modeled estimates of racial effects, both used in fair lending examinations, may be biased."

The probability of identification and/or re-identification of borrowers in HMDA data available to the public, including freedom of information requests, should be as low as that for borrowers in the public version of the Current Population Survey, American Housing Survey, Survey of Consumer Finances and GSE single family loan level dataset. Special care be taken for information that reveals creditworthiness whether or not that is explicitly protected under the FCRA and the RFPA.

Some implications of the principles of privacy protection in HMDA data

When considering disclosure avoidance in HMDA data it is useful to employ the concept of *k*-anonymity. This means that the variables in the dataset should be cross-tabulated and that each cell in the cross-tabulation should have at least k > 1 individual observations. This concept has been applied to USCB data releases.[16]

In HMDA data this implies that for each census tract, each lender id, each loan amount, and each gender-racial-ethnicity borrower and co-borrower combination there should be at least $k > 1$ borrowers. Clearly this is not satisfied in old HMDA data because the previous discussion has established that the value $k = 1$, is most common, i.e. most borrowers are perfectly identified. The lowest possible value of *k* which provides any anonymity is *2*. It is useful to consider the degree of geographic aggregation needed to even achieve $k = 2$ level anonymity in current HMDA data with an illustrative example. The smallest geographic area would need to be large enough so that, for each lender identified, there were at least two observations where a female, Asian borrower with no co-borrower had a purchase mortgage with a loan amount of X thousand dollars (where X is any loan amount rounded to the nearest thousand). It is very likely that there are few lenders where the minimum geographic area is smaller than a state. For small and medium lenders, even the entire US would likely be too small a geographic areas because there would be some cases where only one female Asian borrower, with no co-borrowers from that lender had a purchase mortgage for X thousand dollars in a given year. This is the standard of analysis for *k*-anonymity would need to apply to every gender, race, ethnicity combination of borrower and co-borrower in order to eliminate the need for masking. HMDA data should be should be subject to the same type of analysis. This is the standard that the USCB has set.

If HMDA is to preserve even $k = 2$ anonymity, i.e. the lowest standard possible, it probably cannot disclose identities of any except the largest lenders in each category of mortgage. Proper masking would group lenders into large categories, perhaps by region

[16] There is a substantial literature on the topic of preserving anonymity in government data releases. An overview of USCB masking techniques is provided by Amy Lauger, Billy Wisniewski, and Laura McKenna, "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research", USCB Report, September 26, 2014. For a general literature view suggesting that k-anonymity may not be sufficient see John S. Davis II, and Osonde A. Osoba, "Privacy Preservation in the Age of Big Data: A Survey," RAND Working paper, WR-1161, September, 2016.

and type of institution. Even using aggregate gender - race - ethnicity groups such as non-Hispanic white, Hispanic, black, and other (note gender is masked), substantial masking of loan amount would be necessary. One way to do this would be to report average loan amount for the group in each cell or to disclose loan amount in broad ranges – perhaps intervals of *100,000* with top coding of large amounts.

The standard of *k*-anonymity to be adopted for HMDA data as well as the choice of variables to be aggregated, averaged, or reported as intervals to achieve that level of anonymity should be the object of research and choices about the future use and usefulness of the data. Research is necessary because the relation between different levels of disaggregation and preservation of *k*-anonymity involves tradeoffs that can only be determined by reviewing several years of HMDA data to insure stability over time in the levels of disaggregation. Choice is needed because the data will be useful for a narrow range of purposes given its aggregated nature.

One factor that may help this choice is recognition that HMDA is currently providing a biased and inaccurate view of housing credit flows into local areas. Commercially vended (big) data taken from property records includes credit flows that are missed by HMDA while preserving personal financial details of the homeowner. In 2015 HMDA reports 3.66 million purchase transactions while the Federal Reserve Bank of St. Louis FRED database indicates that there were almost 6 million purchases. In the modern world of shadow banking the divergence between HMDA and the reality of mortgage finance in local communities is likely to widen over time. Even identification of race and ethnicity is likely better using modern techniques to classify names in property transfer records than forced responses to HMDA data, particularly in cases of internet-based landing.

In sum, even old HMDA data disclosures violate basic privacy expectations of households that are applied in other data collection and dissemination efforts. It also appears some mortgage data currently disclosed as part of the GSE Enterprise Public Use Databases and Federal Home Loan Bank Board System also fails to preserve anonymity. Curiously, the GSE single family loan level dataset is masked to substantially lower re-identification risk. The same care in insuring that the privacy protections of HMDA (12 U.S.C. 2803), repeated in HERA, as well as FCRA and RFPA need to be applied to all banking sector data released by the CFPB and FHFA. This does not appear to be an accidental invasion of borrower privacy. The FHFA and CFPB should be well aware of these privacy problems. For example, they should know that re-identification is a virtual certainty for minority borrowers whose last names can be associated with a racial or ethnic group based on the lender name and census tract alone. The new HMDA data releases should be sufficiently masked to preserve at least *k=2* level anonymity. Alternatively the government may turn to commercial data sources that cover all purchase transactions, properly measure investment flows into neighborhoods, and preserve the financial privacy of individuals. Such a dataset would accomplish the fundamental goal of HMDA which is monitoring the flow of equity and debt finance into residential real estate across census tracts, and even smaller geographic divisions, of urban neighborhoods in the U.S.