Market Power in Small Business Lending: A Two-Dimensional Bunching Approach*

Natalie Cox†, Ernest Liu‡, and Daniel Morrison§

August 10, 2021

Abstract

Do government-funded guarantees and interest rate caps primarily benefit borrowers or lenders under imperfect competition? We study how bank concentration impacts the effectiveness of these policy interventions in the small business loan market. Using data from the Small Business Administration's (SBA) Express Loan Program, we estimate a tractable model of bank competition with endogenous interest rates, loan size, and take-up. We introduce a novel methodology that exploits loan "bunching" in the two-dimensional contract space of loan size and interest rates, utilizing a discontinuity in the SBA's interest rate cap. In concentrated markets, we find that a modest decrease in the cap would increase borrower surplus by up to 1.5%, despite the rationing of some loans. In concentrated markets with a 50% loan guarantee, each government dollar spent raises borrower surplus by \$0.64, boosts lender surplus by \$0.34, and generates \$0.02 of deadweight loss.

*We thank Matteo Benetton, Vivek Bhattacharya, Olivier Darmouni, Daniel Green, Yiming Ma, Gregor Matvos, Amit Seru, David Sraer, Kairong Xiao, Anthony Zhang, our discussants Mark Egan and Farzad Saidi, and seminar and conference participants at UNC, Princeton, Stanford SITE, Stern, the CFPB, the Federal Reserve Board, UT Austin Finance, Berkeley Haas, and Columbia for helpful comments and discussions. We are grateful to Brian Headd at the SBA for helpful discussions and comments about institutional details. Christian Kontz provided excellent research assistance.

†Princeton University. Email: nbachas@princeton.edu.

‡Princeton University. Email: ernestliu@princeton.edu.

§Princeton University. Email: dm31@princeton.edu

1 Introduction

Bank lending is an important financing channel for young and small firms and is therefore critically important for the aggregate economy (Kaplan and Zingales (1997); Adelino, Ma and Robinson (2017)). In many countries, governments stimulate lending to small businesses through loan guarantees (Lelarge, Sraer and Thesmar (2010)) and by imposing interest rate caps (Maimbo and Gallegos (2014)). Yet, reliance on geographic proximity for small business lending (Petersen and Rajan (1994); Nguyen (2019)) can give banks substantial market power (Drechsler, Savov and Schnabl (2017)) and potentially cause under-provision of credit. How does bank market power (Egan, Hortaçsu and Matvos (2017); Carlson, Correia and Luck (2019); Benetton (2018)) impact the pass-through of indirect loan guarantees to borrowers? Do interest rate caps benefit small business borrowers, and is there room for better policy? Despite broad academic and policy interest, these remain open questions.

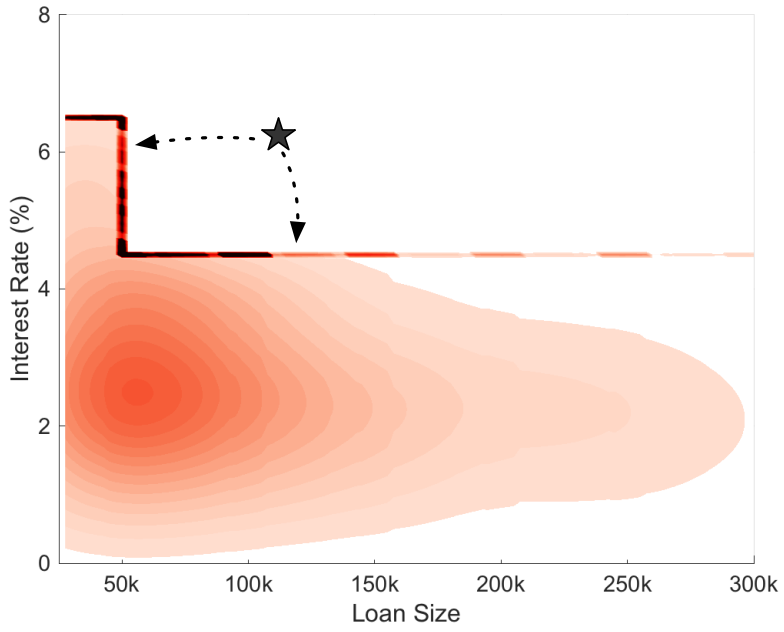
In this paper, we develop a novel two-dimensional bunching estimator to quantify banks' market power and evaluate the effectiveness of policy interventions, utilizing data on loans made through the federal Small Business Administration (SBA). The SBA partially guarantees loans made by commercial lenders to in-need small businesses that are otherwise rejected by all other sources of external financing sources (Brown and Earle (2017); Granja, Leuz and Rajan (2018)). We study loans that are subject to an interest rate cap that decreases discretely for loans larger than $50,000. This "notch" in the interest rate cap imposes a size-dependent constraint on the set of loans that can be offered and generates excess mass in the distribution of loan contracts along the cap and at the discontinuity.

The distribution of excess mass—which encodes how banks react to the policy discontinuity—is indicative of bank's market power. We write down and estimate a tractable model that links the observed distributional distortions to parameters governing imperfect competition. We find quantitatively substantial market power and inefficiencies: depending on the level of market power, banks capture 27–36% of surplus in laissez-faire[1] lending relationships and a similar percentage of the additional surplus created by loan guarantees. We study a wide range of counterfactual policies and compare welfare under each scenario.

The intuition for why the distribution of loans is informative of banks' market power can be understood graphically through Figure 1, which is a density plot of loan contracts seen in the data, with darker shades indicating greater density of loans. The SBA requires

[1] Throughout the paper, we use the term "laissez-faire" to refer to counterfactual environments without interest rate caps or government guarantees. We use the term "unconstrained" to describe environments without interest rate caps.

Figure 1: Loan Size-Dependent Interest Rate Cap and the Empirical Distribution of Contracts



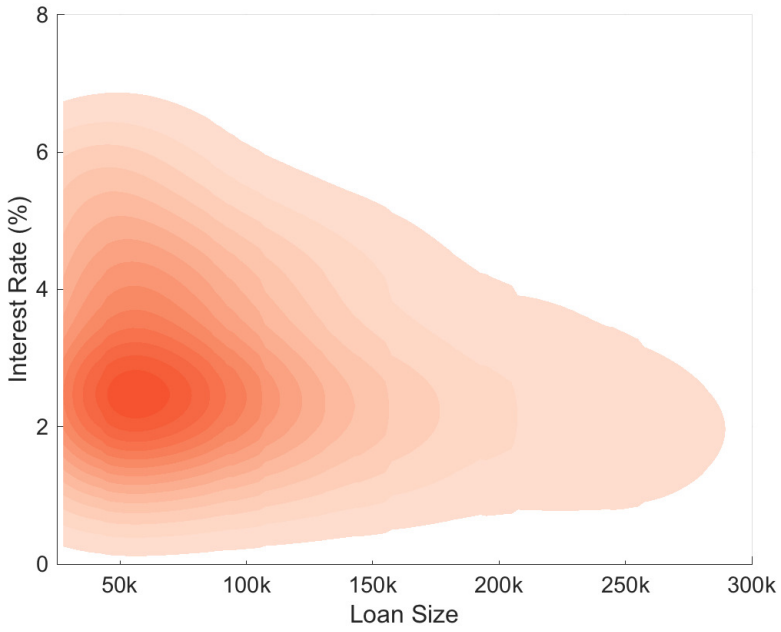
that loans smaller or equal to \$50,000 are capped at a rate of 6.5% above a reference rate (the "prime rate"), while loans larger than \$50,000 are limited to the prime rate plus 4.5%. Banks compete on two-dimensional loan contracts—loan size and interest rates; hence, banks respond in both dimensions to the interest rate cap. Consider a borrower who, in the absence of the rate cap, would have been offered contract ($\star$), which is infeasible under the rate cap. With the rate cap imposed, multiple contracts along the rate cap could plausibly be offered to the borrower. Specifically, banks could lower the interest rate and stay unconstrained on loan size, or they could scale back loan size and, in exchange, charge a relatively higher interest rate. These options—marked by the dotted arrows in the figure—result in different expected profits for banks.

How a profit-maximizing bank responds to the size-dependent interest rate cap depends on its market power. To see this, note the profit from a loan is the product between the loan size and the profit margin per dollar lent. For contract ($\star$), both adjustment margins—reducing the interest rate to the lower cap (prime rate plus 4.5%) or reducing the loan size down to \$50,000—lead to lower profits; yet, a bank with no market power will always choose to scale back loan size to avoid lowering interest rate. This is because if the loan ($\star$) were offered by a competitive bank, then the interest rate must fully reflect lending cost, and a rate any lower could only lead to losses for the bank. On the other hand, a bank with

sufficient market power may prefer to lower rates to avoid decreasing loan size, that is, using smaller profit margins in exchange for relatively larger lending volume under the rate cap restrictions. A bank's conduct under the size-dependent interest rate cap is therefore highly informative of its market power.

To operationalize the strategy, we build a tractable and parsimonious model of imperfect bank competition. A finite number of banks compete for borrowers by offering loan contracts that specify both the interest rate and the loan size. Banks are differentiated vertically—by heterogeneous lending costs—and horizontally—by borrowers' idiosyncratic taste shocks for banking services. Borrower taste and bank cost heterogeneity, along with the finiteness of competing banks, grant banks market power. The model generates a mapping from bank concentration to lending outcomes with and without size-dependent interest rate caps, thereby enabling us to use the observed loan size under the rate cap to recover model parameters that govern market power and lending surplus. The model yields analytic solutions and is amenable to normative policy analysis.

Figure 2: Counterfactual Distribution of Contracts



Equipped with our model, we compare the empirical distribution of loan contracts under the interest rate cap—as in Figure 1—to a counterfactual distribution of contracts without the rate cap—as in Figure 2—to form moment conditions that identify model parameters that govern market power and lending surplus. We estimate the counterfactual contract

distribution by assuming that the distribution of contracts strictly below the policy cap is unaffected by the policy, and we extrapolate the distribution above the cap using the empirical distribution below.

Using our estimates, we analyze the indirect loan guarantee program's impact on borrower, lender, and total surplus. We find that an indirect guarantee expands both borrower surplus (BS) and lender surplus (LS) by an equal factor; for example, the 50% guarantee we observe in the data scales both BS and LS by 2.6%. This is a relatively small percentage increase for either party because of the low default rates observed in the data. Even a 90% guarantee raises BS and LS by only 4.8%. From a levels perspective, in a concentrated market with only 2 symmetric banks, only \$0.63 of each dollar spent by the government reaches the borrower. The remaining subsidy is captured either by the lender (\$0.35) or becomes deadweight loss. The lenders' capture is alleviated only slightly, to \$0.27, as the number of banks in a county increases to seven. We compare the guarantee to other policies commonly used to address imperfect competition. A prime plus 4% interest rate cap, for instance, causes just over 1% of loans to be rationed but boosts average borrower surplus by over 2% conditional on the loan being made; the net effect inclusive of rationing is a 1.3% increase in BS and nearly a 6% decline in LS.

This paper contributes to three distinct branches of literature. First, our novel empirical methodology extends the "bunching" literature in public finance that uses kinks and notches to identify key elasticities (Kleven (2016); Best and Kleven (2018); DeFusco and Paciorek (2017); Cengiz, Dube, Lindner and Zipperer (2019); Antill (2020); Gelber, Jones and Sacks (2020)). Broadly speaking, this approach uses discontinuities in economic agents' choice set and the consequent distortions in outcome to infer structural parameters that govern economic behaviors. Existing papers study one-dimensional bunching (i.e., distortions in a single choice variable) that arise from a single decision-maker's optimization problem. Our key methodological contribution to this literature is twofold: we extend the bunching approach to settings where 1) each decision maker has multiple choice variables, and 2) multiple agents strategically interact, where we study the Nash equilibrium of such interaction. In our application, multiple banks compete with each other by choosing both loan size and the interest rate. Policy discontinuities therefore create distortions over a two-dimensional distribution of loan characteristics, and we uncover banks' market power by selecting appropriate moment conditions and interpreting the empirical distribution of loans offered by each bank as the outcome of a distorted Nash equilibrium induced by SBA's policy interventions.

Second, we add to the literature that studies market power in consumer credit markets

and associated policy interventions. The empirical literature has studied a host of markets, often using reduced-form methods to estimate increased competition's causal impact on separate outcomes, such as risk taking (Jiang, Levine and Lin (2017)), financial stability (Jayaratne and Strahan (1998)), and economic growth (Carlson et al. (2019)). We take a semi-structural approach[2] that allows us to measure the concurrent impact of bank concentration on interdependent loan terms, namely loan size and interest rate. Modeling this joint problem provides a more holistic understanding of how concentration and policy interventions distort loan contracts in several dimensions, and, in turn, impact borrower surplus. We use this model to study indirect government guarantees. Although papers have estimated the impact of guarantees on broader economic outcomes (Brown and Earle (2017); Lelarge et al. (2010); Gale (1991)), we measure their efficiency from a public finance perspective and quantify the benefit they bring to borrowers and lenders.

Third, we provide a nuanced answer to the question of whether interest rate caps, used alone or in conjunction with loan guarantees, help or hurt borrowers. Many countries, including the United States, place restrictions on maximum interest rates (Maimbo and Gallegos (2014)). Policy makers often argue that these caps protect borrowers from lenders who utilize their market power to charge excessively high rates. On the other side of this debate, much of the academic literature contends that rate caps substantially limit access to credit, leading to decreased borrower surplus (Benmelech and Moskowitz (2010); Zinman (2010); Rigbi (2013); Melzer and Schroeder (2017); Cuesta and Sepúlveda (2019)). Our structural approach incorporates both perspectives and evaluates the impact of hypothetical policies that vary the maximum interest rate charged in small business lending. We find that a carefully chosen rate cap can provide modest benefits to borrowers relative to a laissez-faire policy, especially in highly concentrated markets in which banks charge high markups; when used in conjunction with an indirect guarantee, the interest rate cap can also increase the portion of the subsidy that is passed through to borrowers.

We begin with a description of the empirical setting, data, and relevant policy variation in Section 2. We provide an exposition of the model in Section 3. Section 4 discusses the identification strategy and Section 5 summarizes the results. We conduct a counterfactual policy analysis in Section 6. Section 7 concludes.

[2]For similar structural analyses of financial market power see Nelson (2018), Benetton (2018), Egan et al. (2017), Crawford, Pavanini and Schivardi (2018), Bhattacharya, Illanes and Padi (2019), and Cuesta and Sepúlveda (2019).

2 Empirical Setting: SBA Express Loan Program

We analyze small business loans made through the Small Business Administration (SBA) Express lending program in 2008–2017. In this section, we describe the SBA guaranteed lending program and provide some descriptive statistics of the data. In the next two sections, we create a model and discuss the identification strategy that allows us to estimate the model's parameters using the empirical policy variation.

The SBA is an independent federal government agency. It provides commercial lenders with a partial indirect guarantee—the SBA pays the lender a percentage (50% for the Express lending program we study) of the unrecovered principal in case of default—on loans made to participating small businesses. Thousands of commercial lenders across the country participate in the program, and offer partially guaranteed SBA loans to clients who qualify. During the COVID-19 crisis, the SBA expanded as it administered loans made through the Paycheck Protection Program. By August 2020, year-to-date SBA loan volume was more than $500 billion, with 5.2 million loans made through a participating network of 5,460 private lenders.

Lenders are charged a fixed fee (1%–3% of loan principal depending on the year) to the SBA in return for a guarantee that the SBA will reimburse a certain percentage of loan principal in the event of default. In most cases, lenders collect this guarantee fee along with other fixed "closing cost" fees, including legal fees, servicing fees, and filing fees from the borrower when the contract is signed. Loans made through the SBA guarantee program are subject to specific rules and regulations, including the interest rate cap studied here.

The coverage, granularity, and policy variation contained within this dataset makes it the ideal laboratory to study market concentration. The dataset contains contract-level information on loan terms (interest rate, size) and repayment outcomes, borrower identity and characteristics, and bank identity. We know the location and date of both borrowers and banks, which allows us to generate measures of market concentration both cross-sectionally and over time. The SBA Express lending program appeals to borrowers due to its expedited approval process, which allows borrowers to receive funds far faster than through other lending programs. This feature creates a clearly defined market of banks (regional SBA lenders) for that particular borrower; furthermore, borrowers who receive funds through the Express lending program are prohibited from "topping up" their SBA loans with additional sources of credit. For these borrowers, the only relevant lenders are those we observe participating in the SBA program.

Table 1: Summary Statistics for the SBA Express Loan Program Data

	Mean	Std. dev.	Median
Loan Size (in th.)	90.7	63.1	69.9
Interest Rate (%)	2.99	1.38	2.75
Inverse HHI	4.54	2.07	4.36
Maturity (in month)	79	35	84
% at Cap	10.89	-	-
% Charge-off (5 year)	3.23	-	-
Number of Loans	122,569	-	-
Number of Counties	1,806	-	-

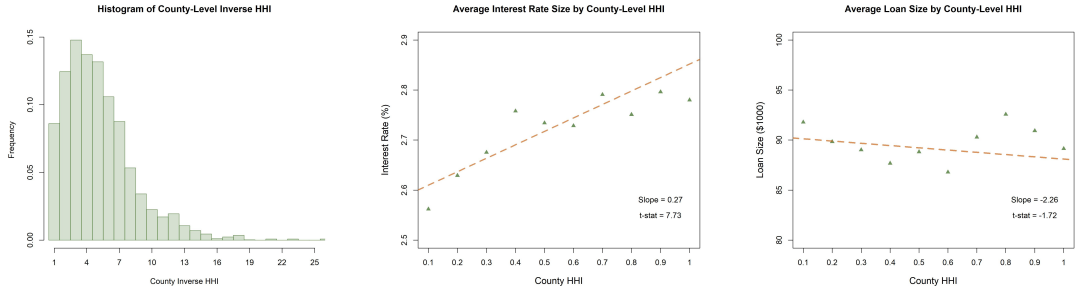
This table displays summary statistics for loans used in our estimation sample from the SBA Express Loan Program, 2008--2017. Average interest rate is expressed in percentage points net of the prime rate, and is captured when the loan is first made. Loan size is expressed in dollar units. The charge-off rate is calculated using a dummy variable for whether a loan charged off during its first five years. Since data are available only through 2017, we only analyze loans created in 2012 or earlier for this statistic. % at Cap is calculated using a dummy variable that indicates whether a loan's terms place it directly on the SBA's notched interest rate cap. Maturity is expressed in months.

Table 1 presents summary statistics for our Express loan sample, which includes 122,569 loans made under the SBA Express program between 2008 and 2018. On average, these loans are for $90,652, and have a maturity of 6.5 years. Interest rates for SBA Express loans can be fixed or variable and are tied to base rates, with the maximum allowable interest rate ranging from 4.5% to 6.5% above the base rate, depending on loan size.[3] The average interest rate in our sample is 2.99% above the base rate, which is well above typical rates for corporate loans.

Although the SBA lending market is heavily regulated, the data nevertheless reveal strong suggestive evidence of imperfect competition. We calculate the Herfendahl-Hirschman Index (HHI) based on the dollar volume lending share (s_{kct}) of each bank k within a given county c and year t: $HHI_{c,t} = \sum_{k=1}^{K} s_{kct}^2$. The HHI index is a summary statistics for market concentration and it ranges between zero and one. Higher HHI indicates greater market concentration. The index is equal to 1 when a single bank holds the entire market, whereas smaller values signal less market concentration. In a market in which banks have equal market shares, the inverse of the HHI is simply the number of banks in the market. Figure 3, which plots the distribution of the inverse HHI across county-years, suggests that many markets are nearly monopolistic and only a small minority feature significant lender competition. The inverse

[3]These base rates are the prime rate, the LIBOR, and the PEG, which can fluctuate based on market conditions. For variable rates, the base rate used for computing interest rates is the lender's choice, provided that the maximum interest rate the borrower is charged still does not exceed the prime rate plus 4.5% to 6.5%.

Figure 3: Distribution of County HHI and Observed Average Interest Rate and Loan Size by Market Concentration



The far-left figure plots the distribution of inverse HHI over all county-year observations in our data. We calculate an *inverse* Herfendahl-Hirschman Index (HHI) based on the dollar volume lending share (s_{ict}), of each bank within a given county-year. A value of 1 means that a single bank holds the entire market share, whereas larger values signal less market concentration. The majority of counties in a given year in our dataset are dominated by fewer than four lenders. The center figure plots the average interest rate charged in markets with differing levels of concentration, as measured by the number of competing banks within a county. This plot includes all loans strictly below the interest rate cap. The interest rate measure controls for loan maturity, log size, business NAICS category, time fixed effects, ex-post performance, and bank brand. The plot suggests that higher interest rates are charged in more-concentrated markets. Default (i.e. charge-off) rates do not increase in more-concentrated markets; if anything, loans in more-concentrated markets are less costly for lenders. Therefore risk-related costs cannot explain the downward sloping relationship between competition and interest rates. The right-hand figure plots the average loan size in each of these markets, which is relatively flat.

HHI of a median county-year is 3.35.

We also observe the impact of market concentration on loan pricing. Figure 3 documents a strong positive relationship between the average initial interest rate charged on observationally identical loans[4] within a county and the HHI of that county. This is not driven by changes in borrower risk across markets—we correlate ex-post measures of default with market HHI and reject a positive relationship. The right-hand panel plots average loan size across the same measure of market concentration and documents a flat relationship.

Although the relationship between HHI and interest rates shown above motivates an analysis of market power, it remains suggestive; identifying the relevant demand and supply parameters from our model requires an exogenous shift or shock to lenders' maximization problem. Loans made through the SBA Express program are subject to specific SBA rules and regulations that provide this identifying policy variation. Specifically, they face an interest rate cap that is dependent on loan size—loans smaller or equal to $50,000 are capped at prime plus 6.5%, while loans larger than 50,000 are limited to prime plus 4.5%. This "notch" in the interest rate cap imposes a size-dependent constraint on banks' pricing problem and generates excess loan density along the interest rate cap, both empirically and in our model. In total, 11% of loans bunch to the interest rate cap. Because SBA regulations do

[4]We control for bank brand (i.e. West America, Chase, etc), borrower business NAICS code, loan maturity, loan size, ex-post loan performance, and time fixed effects.

not allow lenders to originate multiple loans to the same borrower at the same time, lenders cannot "piggyback" loans to take advantage of the notch. In Section 4, we discuss how we exploit this bunching to estimate the structural parameters in our model.

3 Model

We build a tractable model of bank competition with endogenous interest rates, loan size, and take-up. The model is simple yet sufficiently rich to generate empirical predictions of how loan contracts respond to policy.

3.1 Setup

Consider a market with finite K banks and a continuum of borrowers of finite measure. Both parties are risk-neutral. Let k index for banks and i index for borrowers.

Investment Technology Each borrower i has a stochastic investment technology that produces output as a function of investment size L:

$$f_i(L) = \begin{cases} z_i L^\alpha & \text{(succeeds) with probability } p_i, \\ \delta_i L & \text{(fails) with probability } (1-p_i). \end{cases}$$

With probability p_i, the investment succeeds and generates output $z_i L^\alpha$. The term z_i is a productivity shifter, and the parameter α captures the concavity of the production function. With probability $(1-p_i)$ the investment fails, and only $\delta_i < 1$ fraction of investment can be recovered. Each borrower i can be summarized by its characteristic (z_i, δ_i, p_i).

Loan Contracts Borrowers may obtain investments from bank loans. A loan contract is a duplet of interest rate and loan size, (r, L). If the contract (r, L) offered by bank k is accepted by borrower i, it generates contractual value $v_i(r, L)$ to borrower i and expected profit $\pi_{ik}(r, L)$ for bank k:

$$v_i(r, L) \equiv p_i(z_i L^\alpha - (1+r)L) \tag{1}$$

$$\pi_{ik}(r, L) \equiv (p_i(1+r) + (1-p_i)\delta_i - c_k)L \tag{2}$$

Note that loan contracts are similarly to debt: the lender captures the investment payoff up to the specified repayment $(1+r)L$, and the borrower is the residual claimant. When the project fails, the borrower gets paid zero and the bank gets paid $\delta_i L$. When the project

succeeds, borrower gets paid $z_i L^\alpha - (1+r)L$ and the bank gets paid $(1+r)L$. The term c_k represents the opportunity cost of funds to bank k.

The expected utility that borrower i obtains from selecting contract (r, L) from bank k is

$$u_{ik}(r, L) \equiv \xi_{ik} \times v_i(r, L). \tag{3}$$

The term $\xi_{ik} \geq 0$ is a random taste shock and is i.i.d. across borrowers and banks. We refer to $v_i(r, L)$ as the *contractual* value, and $u_{ik}(r, L)$ as the *expected utility,* of loan (r, L) to borrower i. The taste shock ξ_{ik} represents idiosyncratic heterogeneity, such as borrowers' differential preferences for the services provided by differentiated banks.

Bank Competition Banks $k = 1, \ldots, K$ compete for borrowers by simultaneously offering contracts. Banks are differentiated both vertically—due to cost heterogeneity c_k—and horizontally—due to idiosyncratic taste shocks ξ_{ik}. Each bank k offers one contract (r_{ik}, L_{ik}) to each borrower i.[5] We assume each borrower can always walk away from the investment opportunity if loan terms are too unattractive; that is, every borrower has an outside option of zero utility. Given the set of contracts offered by competing banks, borrowers accept the contract that generates the highest and non-negative expected utility. The probability that borrower i chooses the contract offered by bank k is

$$q_{ik} \equiv Pr(i \text{ chooses } k) = Pr\left(u_{ik} \geq \max\left\{0, \max_{k'} u_{ik'}\right\}\right). \tag{4}$$

The randomness in the borrower's choice of contract originates from idiosyncratic taste shocks. Note that q_{ik} increases in the contractual utility v_{ik} offered by bank k and decreases in $v_{ik'}$ for all $k' \neq k$. When competing for borrowers, banks observe the borrower's production technology $f_i(L)$ but do not observe the idiosyncratic shocks. Each bank k offers the contract that maximizes expected profit:

$$(r_{ik}^*, L_{ik}^*) \equiv \arg\max_{r_{ik}, L_{ik}} q_{ik} \times \pi_{ik}. \tag{5}$$

Distribution of Taste Shock For tractability, we assume the idiosyncratic taste shocks ξ_{ik} are drawn from a Fréchet distribution, with CDF $G(\xi; \sigma) = e^{-(\gamma\xi)^{-\sigma}}$, where $\gamma \equiv \Gamma(1 - 1/\sigma)$ is a normalizing constant and Γ is the Gamma function (Johnson and Kotz (1970)). This distributional assumption enables us to analytically solve for equilibrium loan contracts as

[5] Because banks observe borrower's productivity z_i, recovery rate δ_i, and probability of succcess p_i, it is without loss of generality to specify that banks offer a single (optimal) contract.

a function of the market structure.

$\sigma > 0$ is the key parameter that captures the substitutability of loans across banks and relates inversely to the variance of borrowers' idiosyncratic taste shocks. Banks are more substitutable when σ is high. As we show below, in the limit as $\sigma \to \infty$, banks become perfect substitutes. Conversely, as $\sigma \to 0$, the banking choice becomes entirely idiosyncratic and is driven by ξ_{ik}; consequently, the choice probability for any given bank becomes independent of contractual utilities $\{v_{ik'}\}$.

Under the distributional assumption, the choice probability for any given bank becomes

$$q_{ik}\left(\{v_{ik'}\}_{k'=1}^{K}\right) = \frac{\max\{0, v_{ik}^{\sigma}\}}{\sum_{k'=1}^{K}\max\{0, v_{ik'}^{\sigma}\}}. \tag{6}$$

Let $\epsilon_{ik} \equiv \partial \ln q_{ik} / \partial \ln v_{ik} \geq 0$ denote the elasticity of the choice probability q_{ik} (that borrower i chooses bank k) with respect to the contractual utility v_{ik}, holding contracts offered by all other banks constant. We refer to ϵ_{ik} simply as the "choice elasticity", and under the distributional assumption,

$$\epsilon_{ik} = \sigma\left(1 - q_{ik}\right). \tag{7}$$

Ceteris paribus, the choice elasticity ϵ_{ik} is greater when banks are more substitutable (higher σ). The expected utility of borrower i is

$$EU_i \equiv \mathbb{E}\left[\max_k \xi_{ik} v_{ik}\right] = \left(\sum_{k=1}^{K}\max\{0, v_{ik}^{\sigma}\}\right)^{\frac{1}{\sigma}}.$$

Definition 1. The unconstrained equilibrium is the set of contracts $\{(r_{ik}^*, L_{ik}^*)\}$ that solves the profit maximization problem (5) for each bank k and each borrower i.

We use the term "unconstrained" to refer to equilibrium contracts in a policy environment featuring no interest rate caps. We reserve the term "laissez-faire" for environments that also feature no government guarantee on loans.

Default happens in the model when the output is below the required loan repayment ($f_i(L) < (1+r)L$). Importantly, default is always involuntary: the borrower repays as much as the output allows, and there is no strategic decision regarding default. Another feature of the model is that default generates no deadweight loss and simply represents a transfer between the borrower and the lender under the contingency that output is low. This can be seen by noting that the sum of bank profit and the contractual value to the borrower

is a function of only loan size and is invariant to the interest rate:

$$v_i(r, L) + \pi_{ik}(r, L) = \mathbb{E}_i[f(L)] - c_k L.$$

This simplification enables us to abstract away from inefficient default and focus on market power as the only source of potential inefficiency in the model.

Each bank's profit maximization problem can be written as

$$\max_{r,L} \underbrace{[p_i(1+r) + (1-p_i)\delta_i - c_k] L}_{\substack{\text{expected profit conditioninal} \\ \text{on contract being accepted}}} \times \underbrace{\frac{\max\{0, v_{ik}^{\sigma}\}}{\sum_{k'=1}^{K} \max\{0, v_{ik'}^{\sigma}\}}}_{\text{choice probability}} \quad \text{s.t. } v_{ik} = p_i(z_i L^{\alpha} - (1+r)L). \tag{8}$$

We define bank k's profit margin as

$$\mu_{ik}(r, L) \equiv \frac{p_i(1+r_i)}{c_k - (1-p_i)\delta_i}, \tag{9}$$

which is the ratio between expected bank profit and effective marginal cost, conditioning on the loan being accepted.

Proposition 1. *The unconstrained equilibrium has the following features.*

1. *Every borrower chooses bank k with the same probability: $q_{ik} = s_k$ for all i, where s_k is the market share of bank k.*

2. *Loan terms satisfy*

$$L_{ik} = \left(\frac{\alpha p_i z_i}{c_k - (1-p_i)\delta_i}\right)^{\frac{1}{1-\alpha}}, \tag{10}$$

$$\mu_{ik} = \frac{1 + \alpha\sigma(1-s_k)}{\alpha + \alpha\sigma(1-s_k)} = 1 + \frac{1-\alpha}{\alpha} \cdot \frac{1}{1+\sigma(1-s_k)}. \tag{11}$$

   The interest rate can be recovered from the profit margin μ_{ik} according to equation (9): (i.e., $(1+r_i) = \mu_{ik}\frac{c_k - (1-p_i)\delta_i}{p_i}$).

3. *Contractual value of every loan contract is positive: $v_{ik} > 0$ for all i, k.*

4. *Let $HHI \equiv \sum_{k=1}^{K} s_k^2$ denote the Herfendahl index in the lending market. Then the average profit margin ($\mu = \int \mu_{ik} s_k dF_i$) in the market can be written as*

$$\mu \approx \frac{1-\alpha}{\alpha(1+\sigma)} + \frac{(1-\alpha)}{\alpha(1+\sigma)} \frac{\sigma}{(1+\sigma)} \times HHI,$$

where the approximation error is $o\left(\max_k (s_k)^2\right)$*, i.e., second-order in the market share of the largest bank.*

The first part of the proposition states that each bank's market share captures the choice probability for any borrower in the market. Note that market share s_k is itself an endogenous outcome of bank competition. Banks may differ in their equilibrium market share due to vertical differentiation, i.e., heterogeneity in their funding cost, c_k. Banks with lower funding costs have higher market share. When all banks have identical funding costs, they also have the same market share: $s_k = 1/K$ for all k, where K is the total number of banks.

The second part of Proposition 1 characterizes equilibrium loan terms. Equation (10) implies that equilibrium loan size is efficient as it equals the loan size that maximizes total surplus: $\max_L p_i z_i L^\alpha + (1 - p_i)\delta_i L - c_k L$. To understand this, note that, because defaults happen only involuntarily, interest payments serve as a linear transfer from each borrower to the lender. Hence, the two loan characteristics serve distinct roles in equilibrium: loan size is always chosen to maximizes total surplus, whereas interest rate pins down the division of surplus between the lender and the borrower.

Equation (11) solves for the equilibrium profit margin and thus the interest rate. To understand this, note the total surplus of each loan in the unconstrained equilibrium, $\max_L p_i z_i L^\alpha - (c_k - (1 - p_i)\delta_i) L$, depends on the concavity of the borrower's production technology, α. Ceteris paribus, when the production technology is more concave—lower α—investment generates more surplus per unit cost of lending. The profit margin therefore depends on α. The remaining term $\frac{1}{1+\sigma(1-s_k)}$ captures the fraction of surplus accrued to the bank. The bank's share of surplus relates to its market power; according to equation (7), the term $\sigma(1 - s_k)$ in the numerator is exactly equal to the choice elasticity ϵ_{ik} of borrower i with respect to the contractual value offered by bank k. Banks have higher profit margins when they are less substitutable (lower σ) or have greater market shares (higher s_k).

The last part of Proposition 1 shows that in equilibrium, the average profit margin in a given market is approximately linear in the HHI index for bank loans. Because the HHI index is equal to $1/K$ when all K banks are symmetric, HHI can therefore also be seen as inversely related to the effective number of banks operating in the market.

In summary, the proposition implies that, holding borrower characteristics constant, profit margins and interest rates are higher in more-concentrated markets, that is, when there are fewer banks or when banks have more asymmetric lending costs, but market concentration should not have predictive power over loan size. That across U.S. counties, interest rates

are increasing in HHI but loan size is flat, as shown in Figure 3, serves as an external validation of the model. The proposition also implies that profit margins and interest rates are higher when banks are less substitutable (lower σ). These results are intuitive: when there are fewer competing banks or when banks are less substitutable, demand for loans from a specific bank should become more inelastic, as a marginal increase in the interest rate—and the consequent reduction in contractual utility—should lead to a smaller outflow of potential borrowers. Consequently, competition is weaker, and banks offer loan terms that are less favorable to borrowers.

3.2 Banks' Response to Policy Interventions

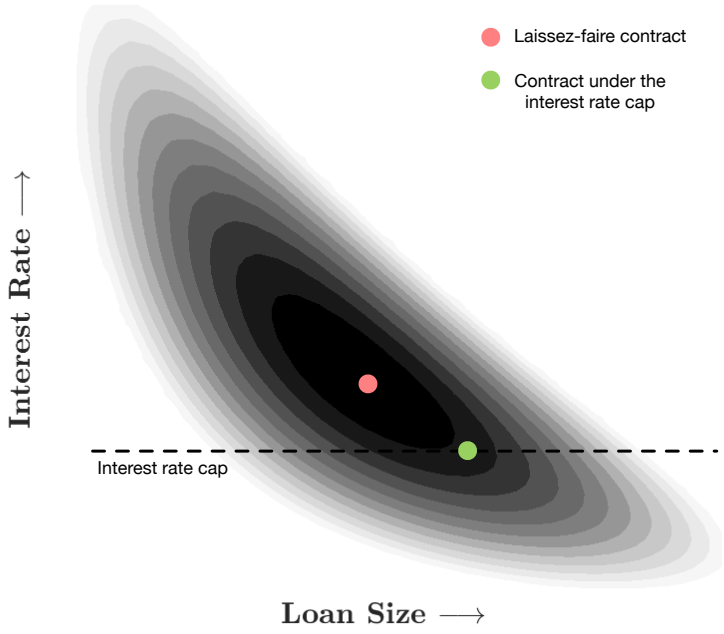
We now analyze how banks respond to constraints imposed by government policies in the contract space. We conduct this analysis for two reasons. First, when we estimate the model in the data, our identification strategy exploits banks' responses to constraints in the contract space in order to recover model primitives. Second, interest rate caps are common policy tools; this section guides our analysis of these policies as we perform counterfactuals in Section 7.

We first analyze how banks respond to simple, flat constraints on interest rate and loan size. We then analyze how banks respond to interest rate caps that vary with loan size.

Interest Rate and Loan Size Are Strategic Substitutes Because contracts are two-dimensional, banks have two levers to extract profit from borrowers: the interest rate and the loan size. In equilibrium, a bank sets contractual terms to balance the trade-off between extracting profits π_{ik} and leaving surplus to the borrower to raise the probability of loan take-up q_{ik}. Imposing any binding constraints on one of the choice variables r and L will intuitively cause banks to respond over the other choice variable as well. More importantly, an increase in either the interest rate or the loan size leads to higher profits π_{ik} and lower take-up q_{ik}, the two choice variables are strategic substitutes, meaning imposing a binding interest rate cap leads banks to overlend, as loan size becomes larger than what is efficient; likewise, imposing a binding loan size cap leads banks to charge higher interest rates than what would have prevailed absent the constraints.

To demonstrate this, Figure 4 shows a contour plot of a bank's isoprofit curve as a function of the two choice variables r and L. Darker shades indicate higher profits. Because banks' maximization problem is concave, the profit function is single peaked: the pink dot indicates the contract that would be offered if no policy constraints were imposed. Next, if an interest rate cap (dotted horizontal line) is imposed so that the contract in pink is no longer feasible,

Figure 4: Contour plot of bank's profit as a function of contractual terms



which contract does the bank offer? The bank would choose, among all feasible contracts, the one with the highest profit—the darkest spot—in the constrained set. Because the decline in profits is least steep in the direction of higher L, the bank conforms to the interest rate cap and sets a larger loan size, as indicated by the green dot. Likewise, a binding loan size cap induces the bank to raise the interest rate.

How to interpret the prediction that, in the presence of a binding interest rate cap, the bank offers a larger loan relative to the unconstrained equilibrium? Note that the borrower's investment technology—with expected output $p_i z_i L^\alpha + (1 - p_i)\,\delta_i L$—always benefits from a larger loan. The prediction is therefore not that the bank offers more than what the borrower asks on the loan application, which may exceed the unconstrained loan offer; instead, the theory predicts that the bank approves and offers a greater loan under a binding interest rate cap than it otherwise would in the absence of the cap.

We now formalize these predictions and provide analytic solutions to the contractual response under various policy constraints. For notational simplicity, we assume all K banks are symmetric, and we characterize how contracts change in response to policy constraints. We drop the subscript k whenever it is unambiguous.

Simple Interest Rate Caps

Proposition 2. *Consider a bank's profit maximization problem (8) under additional constraints. Let (r_i^*, L_i^*) represent the unconstrained contract.*

1. *Consider the constraint $r_i \leq \bar{r}$. If $p_i(1+\bar{r}-\delta_i) < c_k - \delta_i$, then the loan will be rationed as the bank is no longer able to recover lending cost under the constraint. Otherwise, the equilibrium contract is*

$$(r_i, L_i) = \left(\min\{\bar{r}, r_i^*\}, L_i^* \times \max\left\{1, \left(\frac{1+r_i^*}{1+\bar{r}}\right)^{\frac{1}{1-\alpha}}\right\}\right).$$

2. *The equilibrium contract under the constraint $L_i \leq \bar{L}$ satisfies*

$$(r_i, L_i) = \left(\max\left\{(1+r_i^*)\frac{(L_i^*/\bar{L})^{1-\alpha} + \alpha\sigma(1-1/K)}{1+\alpha\sigma(1-1/K)} - 1, r_i^*\right\}, \min\{\bar{L}, L_i^*\}\right).$$

Proof: See Appendix A. This formulation allows us to express the constrained contract (r_i, L_i) in terms of the unconstrained contract (r_i^*, L_i^*) and the market parameters $\alpha, \sigma,$ and K, without referencing the borrower and lender-specific parameters p_i, δ_i, z_i, and c_k. This property is the result of the fact that the unconstrained loan terms (r_i^*, L_i^*) incorporate the relevant information from these parameters.

Loan Size–Dependent Interest Rate Caps Now consider size-dependent interest rate caps, that is., an interest ceiling $\bar{r}^H$ for loans sizes below $\bar{L}$ and ceiling $\bar{r}^L < \bar{r}^H$ for loan sizes above $\bar{L}$. We continue to use (r_i^*, L_i^*) to represent the unconstrained contract and use (r_i, L_i) to represent the equilibrium contract under the policy.

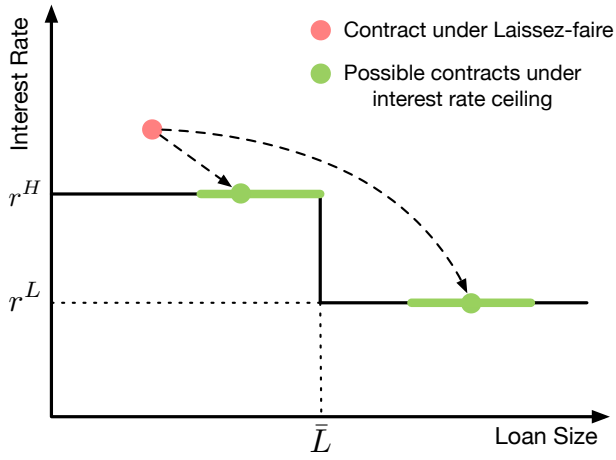
Because the two choice variables r and L are strategic substitutes, we can intuitively categorize each bank's response to the size-dependent interest rate cap into three scenarios, depending on borrower i's characteristics and the policy environment $\left(\bar{r}^L, \bar{r}^H, \bar{L}\right)$:

A. $r_i^* < \bar{r}^L$, or ($r_i^* < \bar{r}^H$ and $L_i^* < \bar{L}$): for these borrowers, the interest rate ceilings do not bind.

B. $r_i^* > \bar{r}^H$ and $L_i^* < \bar{L}$: for these borrowers, equilibrium loan terms have two possibilities other than being rationed:

   (a) $L_i \leq \bar{L}$ and $r_i = \bar{r}^H$;

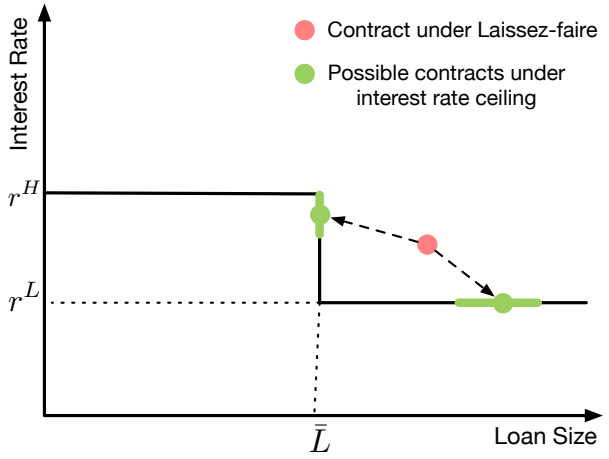
   (b) $L_i > \bar{L}$ and $r_i = \bar{r}^L$.



C. $r_i^* > \bar{r}^L$ and $L_i^* \geq \bar{L}$: for these borrowers, equilibrium loan terms have two possibilities other than being rationed:

   (a) $L_i > \bar{L}$ and $r_i = \bar{r}^L$;

   (b) $L_i = \bar{L}$ and $r_i \in (\bar{r}^L, \bar{r}^H]$.



The next proposition formally characterizes the equilibrium contract.

Proposition 3. *Suppose the unconstrained contract* (r_i^*, L_i^*) *is infeasible under the policy environment with rate cap* $\bar{r}^H$ *for* $L < \bar{L}$ *and* $\bar{r}^L$ *for* $L > \bar{L}$. *Let* $\left(r_i^L, L_i^L\right) \equiv \left(\bar{r}_i^L, L_i^* \left(\frac{1+r_i^*}{1+\bar{r}^L}\right)^{\frac{1}{1-\alpha}}\right)$, *and let*

$$\left(r_i^H, L_i^H\right) \equiv \begin{cases} \left(\bar{r}^H, \min\left\{\bar{L}, L_i^* \left(\frac{1+r_i^*}{1+\bar{r}^H}\right)^{\frac{1}{1-\alpha}}\right\}\right) & \text{if } L_i^* < \bar{L} \\ \left(\min\left\{\bar{r}^H, (1+r_i^*)\frac{(L_i^*/\bar{L})^{1-\alpha}+\alpha\sigma(1-\frac{1}{K})}{1+\alpha\sigma(1-\frac{1}{K})} - 1\right\}, \bar{L}\right) & \text{if } L_i^* \geq \bar{L}. \end{cases}$$

The loan will be rationed if $p_i\left(1+\bar{r}^H-\delta_i\right) < c_k - \delta_i$. *Otherwise, the equilibrium contract is*

$\left(r_i^H, L_i^H\right)$ *if*

$$L_i^H(p_i(1+r_i^H-\delta_i)-(c_k-\delta_i))q_{ik}(r_i^H, L_i^H) \geq L_i^L\left(p_i\left(1+r_i^L-\delta_i\right)-(c_k-\delta_i)\right)q_{ik}(r_i^L, L_i^L) \quad (12)$$

and is $\left(r_i^L, L_i^L\right)$ *if the inequality (12) fails to hold.*

For borrowers whose unconstrained contract is infeasible under the interest rate ceilings, the bank offers either smaller loans with higher interest rates, $\left(r_i^H, L_i^H\right)$, or larger loans with lower interest rates, $\left(r_i^L, L_i^L\right)$. The inequality (12) states that a bank selects the option that is most profitable. What that option is will depend on the borrower choice probability, q_{ik}, which in turn depends on the contracts offered by all banks in the market. We solve for a symmetric Nash equilibrium in which all banks choose between offering $\left(r_i^L, L_i^L\right)$ and $\left(r_i^H, L_i^H\right)$. The proof for Proposition 3 is given in Appendix A.

Indirect Loan Guarantees With a λ percent indirect loan guarantee, the SBA promises to pay the lender λ percent of the unpaid principle of the loan if the borrower defaults. The lender's profit maximization problem now becomes:

$$\max_{r,L}\left[p_i\left(1+r\right)+\left(1-p_i\right)\left(\delta_i+\lambda(1-\delta_i)\right)-c_k\right]L\cdot q_{ik}\left(\{v_{ik'}\}_{k'=1}^K\right) \quad (13)$$

We can define the recovery rate inclusive of the indirect guarantee as $\delta_i^G = \delta_i+\lambda(1-\delta_i)$. δ_i^G can be interpreded as the fraction of principal that lenders recover either from the borrower or from the SBA in the case of default. Making this substitution in 13 brings us back to 8; thus replacing δ_i with δ_i^G preserves the validity of all equations in the previous sections.

4 Two-Dimensional Bunching: Methodology

We use the empirical distribution of contractual terms under a loan size–dependent interest rate cap policy to identify model parameters. Our identification strategy builds on and extends the "bunching" literature that uses kinks and notches to identify key elasticities (Best and Kleven (2018); Kleven (2016)). Broadly speaking, this approach exploits discontinuities in economic agents' choice sets and the consequent distortions in the equilibrium outcome distribution to infer structural parameters that govern economic behaviors. The bunching approach requires two steps: 1) recover a counterfactual distribution $H^0(\cdot)$ of equilibrium outcome absent the policy discontinuity; 2) use the difference between the counterfactual distribution and the observed, equilibrium distribution $H^P(\cdot)$ under policy

to infer parameters.

Our methodological contribution advances the bunching approach to a setting with a multi-dimensional behavioral response. In our setting, loan contracts are two-dimensional, involving both the interest rate and the loan size. As we have shown, imposing policy constraints on either choice variable would distort the loan contract on the other variable. The policy variation we exploit therefore naturally calls for a two-dimensional bunching approach to inference, and in this section we outline how we can identify the two key model parameters—σ and α—from the empirically observed distribution of loan contracts.

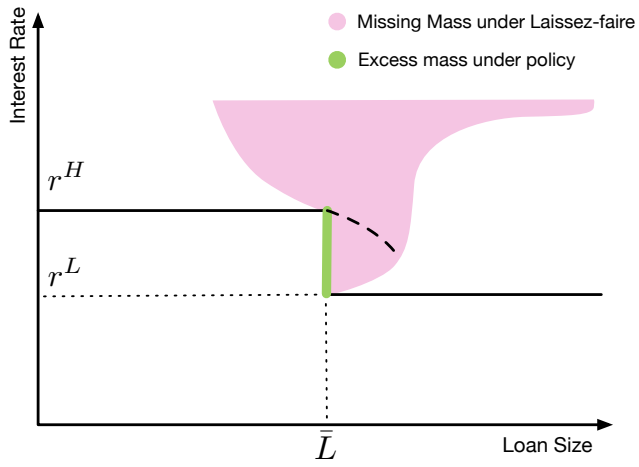
First, we recover the two-dimensional joint distribution $H^0(r, L)$ of unconstrained contracts from the distribution of contracts $H^P(r, L)$ observed in the data. We start from the subsample of loans with interest rates strictly below the cap, $S \equiv \left\{r, L | r < \bar{r}^L \text{ or } \left(r < \bar{r}^H \text{ and } L < \bar{L}\right)\right\}$, and we recover $H^0(r, L)$ from the conditional distribution of loans $H^P(r, L|S)$. This strategy is motivated by the fact that if it were optimal for banks to offer a contract in set S when unconstrained, then such a contract is still optimal and available even in the presence of the policy cap. Moreover, because each bank's profit maximization problem is concave, the policy cap does not move any unconstrained contract strictly outside of the set S to the interior region strictly below the policy cap. Therefore, the distribution of contracts in set S under the policy cap coincides with the conditional distribution in an unconstrained environment: $H^P(r, L|S) = H^0(r, L|S)$. Under the assumption that $H^0(\cdot)$ is analytic over its domain—a standard assumption in the bunching literature—we then recover $H^0(\cdot)$ over the entire two-dimensional domain by extrapolating from the conditional distribution $H^P(r, L|S)$. Section 5.1 provides a detailed discussion of the statistical procedure that recovers $H^0(r, L)$ from $H^P(r, L|S)$.

Once we have the counterfactual and empirical distribution of contracts H^0 and H^P, we then take the difference between the two and define $D \equiv H^P - H^0$. Conceptually, we could do this market by market generating a collection of Ds across markets. We refer to D simply as the "distortion", as it represents the distortion in the distribution of contracts due to the interest rate cap. As an example, D for a hypothetical market is visualized in Figure 5. Green denotes the regions in which there is "excess mass"—i.e. the observed joint distribution has more mass than the predicted counterfactual distribution. The excess mass is concentrated along the interest rate cap, where banks have "bunched" loans that would have otherwise existed *above* the cap. Pink marks regions in which there is "missing mass", where the observed distribution has less mass than the predicted counterfactual distribution. Because banks are not allowed to make loans above the cap, this missing mass is concentrated

in the region above the cap. In principle, the two distributions should be identical strictly below the cap; any difference therein is due to imperfect fit in our estimation procedure.

Identification We now discuss how to identify model parameters α and σ based on the collection of distortions D across markets. We form moment conditions guided by Proposition 3, which describes how banks would offer contracts in the presence of a size-dependent interest rate cap. The proposition clarifies which loans offered in an unconstrained environment, (L^*, r^*), would be distorted under a size–dependent interest rate cap, and, importantly, *where* they would be relocated, as functions of α and σ. In other words, the proposition explains how the distortions we observe in D relate to the model parameters, directly linking the observables, L, r, and K, to the structural parameters. As we show below, this approach enables us to be completely agnostic about borrower heterogeneity: we allow for an arbitrary distribution of borrower characteristics (z_i, p_i, δ_i), and the distribution could vary arbitrarily across markets.

Figure 5: Excess and Missing Mass Regions Used for Identification



Specifically, we use two moment conditions per market for identification. These conditions equate the excess mass at the notch (where $L = \$50,000$) to specific regions of missing mass, as illustrated in Figure 5. The notch is tinted green to represent excess mass, and the purple region represents missing mass. There are two purple regions separated by a dotted curve. Our first moment condition equates the excess mass at the point $\left(\bar{r}^H, \bar{L}\right)$ to the missing mass in the upper purple region, which contains the set of unconstrained contracts that would relocate to $\left(\bar{r}^H, \bar{L}\right)$ under the policy cap. Analogously, our second moment condition equates the missing mass in the lower purple region to the excess mass along the notch, but excluding the end points (i.e. excess mass over the set $B \equiv \left\{(r, L) \,\middle|\, r \in (\bar{r}^L, \bar{r}^H), L = \bar{L}\right\}$).

The boundaries that define the two purple regions vary continuously with model parameters α and σ. Under the true structural parameters that generate the data, the excess mass in green should be exactly equal to the missing mass in purple.

The intuition for selecting these two moments is as follows. Recall that α and σ respectively capture: 1) how lending surplus relates to loan size —more concave technologies generate more surplus per unit cost of lending —and 2) how each bank's market power varies with its market share. First consider an unconstrained contract with $r_i^* > \bar{r}^H$ and $L_i^* < \bar{L}$. If the loan is not rationed under the policy cap, Proposition 3 shows that the equilibrium contract will carry interest rate $r_i = \bar{r}^H$ and loan size $L_i = L_i^* \left(\frac{1+r_i^*}{1+\bar{r}^H}\right)^{\frac{1}{1-\alpha}} < \bar{L}$ if r_i^* and L_i^* are relatively small. For this contract, the rate cap $\bar{r}^H$ binds, but the loan size remains locally unconstrained ($L_i < \bar{L}$), in which case the equilibrium loan size depends only on the structural parameter α and not σ. Intuitively, because the interest rate cap binds, the bank's market power becomes irrelevant in choosing loan terms, and the parameter α governs the distortion in equilibrium loan size because it captures the elasticity of investment output to L. Consequently, the shape of the left boundary to the upper triangle—defined by the set $\left\{(r, L) \,|L \left(\frac{1+r}{1+\bar{r}^H}\right)^{\frac{1}{1-\alpha}} = \bar{L}\right\}$—is entirely pinned down by α and not σ.

Second, consider an unconstrained contract with $r_i^* \in \left(\bar{r}^L, \bar{r}^H\right)$ and $L_i^* > \bar{L}$. If the loan is not rationed, Proposition 3 shows that the corresponding equilibrium contract must either scale back loan size to $L_i = \bar{L}$ and charge a relatively higher interest rate $r_i \in [r_i^*, \bar{r}^H]$ or conform to the lower rate cap $r_i = \bar{r}^L$ and remain unconstrained on loan size $L_i > L_i^*$. For a given unconstrained contract, which of the two scenarios prevails in equilibrium depends on whether the contract falls to the left or to the right of the lower purple region's lower boundary. That boundary's shape in turn depends on banks' market power. Intuitively, when market power is low, the profit margin underlying the unconstrained contract is also low; hence, conforming to the lower rate cap $\bar{r}^L$ represents a disproportionately large decline in the profit margin and is relatively unattractive. In this case, banks are more likely to scale back loan size to maintain a larger profit margin. An extreme case is perfect competition and no market power. As $\sigma \to \infty$ banks become perfect substitutes. The unconstrained rate, r_i^*, already fully reflects marginal lending cost; therefore conforming to the lower rate cap $\bar{r}^L$ would generate losses to the banks. Banks' only remaining choice is to scale back loan size and offer $r_i \geq r_i^*$, $L_i = \bar{L}$. Conversely, when the unconstrained contract has high profit margin, conforming to the lower rate cap implies a relatively small decline in profit margin, and banks are more likely to find this option attractive relative to distorting loan size.

The discussion above illustrates that the lower purple region's lower boundary depends on banks' market power and the choice elasticity. Because the choice elasticity is equal to $\sigma(1-1/K)$, we first use the moment condition to identify the choice elasticity within each market (i.e., for a given K), and we then utilize the variation in choice elasticity across markets with different numbers of banks and HHIs to recover the structural parameter σ.

Moment Conditions We now formalize the identification strategy into moment conditions. We analytically formulate the boundaries of both purple regions in Figure 5 as continuous functions of α and σ. These parameters are then empirically pinned down when the boundaries are chosen such that the missing mass in each purple region matches exactly the excess mass in the corresponding part of the notch.

First, let $r^H = (1+r^*)\frac{\left(\frac{L^*}{L}\right)^{1-\alpha}+\alpha\sigma(1-1/K)}{1+\alpha\sigma(1-1/K)} - 1$, and let $L^L = L^* \left(\frac{1+\bar{r}^L}{1+r^*}\right)^{\frac{1}{\alpha-1}}$. Using Proposition 3, we can formally define the unconstrained equilibrium contracts $(r^*, L^*) \in S_K$ that bunch into the region $B \equiv \left\{(r,L) \middle| r \in (\bar{r}^L, \bar{r}^H), L = \bar{L}\right\}$ under the size-dependent interest rate cap as $S_K \equiv S_K^1 \cap S_K^2 \cap S_K^3$, where S_K^1, S_K^2, and S_K^3 are defined below. The set S_K corresponds to the lower purple region in Figure 5.

$$
\begin{aligned}
S_K^1 \equiv \Bigg\{ (r_i, L_i) \Bigg| & \bar{L}\left((1+r^H-(1+r_i)\frac{\alpha+\alpha\sigma(1-1/K)}{1+\alpha\sigma(1-1/K)}\right) q_{ik}(r^H, \bar{L}) \\
& > L^L\left((1+\bar{r}^L)-(1+r_i)\cdot\frac{\alpha+\alpha\sigma(1-1/K)}{1+\alpha\sigma(1-1/K)}\right)\cdot q_{ik}(\bar{r}^L, L^L)\Bigg\},
\end{aligned}
$$

$$
S_K^2 \equiv \left\{ (r_i, L_i) \middle| r^H < \bar{r}^H \right\}, \qquad S_K^3 \equiv \left\{ (r_i, L_i) \middle| L_i > \bar{L} \right\}.
$$

Intuitively, S_K^1 picks out the laissez-fare contracts that scale back to $\bar{L}$ (instead of overlending at an interest rate $\bar{r}^L$), and S_K^2 picks out unconstrained contracts that charge strictly less than $\bar{r}^H$ under the policy intervention. The borrower choice probabilities, q_{ik}, depend on the contracts offered by all banks in the market. In Appendix A, we find a symmetric (mixed-strategy) Nash equilibrium in which all banks offer the same contracts with the same probabilities. That the excess mass in the region defined by B is equal to the missing mass in the region defined by S_K is our first moment condition for all markets with K banks.

To formalize the second moment condition, let

$$R_K^1 \equiv \left\{ (r_i, L_i) \middle| L_i \geq \bar{L}, r^H \geq \bar{r}^H, r_i > \bar{r}^L \right\},$$

$$R_K^2 \equiv \left\{ (r_i, L_i) \middle| r_i \geq \bar{r}^H, L^L \geq \bar{L} \geq L_i \right\},$$

$$\begin{aligned} R_K^3 \equiv \Bigg\{ (r, L) \Bigg| & \bar{L} \left(1 + \bar{r}^H - (1 + r_i) \frac{\alpha + \alpha\sigma(1 - 1/K)}{1 + \alpha\sigma\left(1 - 1/K\right)} \right) q_{ik}(\bar{r}^H, \bar{L}) \\ & \geq L^L \left(1 + \bar{r}^L - (1 + r_i) \frac{\alpha + \alpha\sigma(1 - 1/K)}{1 + \alpha\sigma\left(1 - 1/K\right)} \right) q_{ik}(\bar{r}^L, L^L) \Bigg\}, \end{aligned}$$

$$R_K^4 \equiv \left\{ (r, L) \middle| (1 + \bar{r}^H) \geq (1 + r_i) \left(\frac{\alpha + \alpha\sigma(1 - 1/K)}{1 + \alpha\sigma(1 - 1/K)} \right) \right\}.$$

Here, R_K^1 identifies contracts right of the notch that move to the point $\left(\bar{r}^H, \bar{L}\right)$, as opposed to along the line segment $\left((\bar{r}^L, \bar{r}^H), \bar{L}\right)$. R_K^2 finds contracts left of the notch that move to the point $\left(\bar{r}^H, \bar{L}\right)$, as opposed to along the line segment $\left(\bar{r}^H, (0, \bar{L})\right)$. R_K^3 picks out contracts that scale back to $\bar{L}$ (instead of over-lending at an interest rate $\bar{r}^L$). Finally, R_K^4 specifies the loans that are not rationed due to the rate cap. The intersection $R_K \equiv (R_K^1 \cup R_K^2) \cap R_K^3 \cap R_K^4$ corresponds to the upper purple region, the unconstrained contracts that will bunch back to the point $\left(\bar{r}^H, \bar{L}\right)$.

For each market structure K, we generate the following moment condition:

$$\begin{cases} \iint_{(r,L) \in B} dD\left(r, L\right) + \iint_{(r,L) \in S_K} dD\left(r, L\right) = 0 \\ D\left(\bar{r}^H, \bar{L}\right) + \iint_{(r,L) \in R_K} dD\left(r, L\right) = 0. \end{cases}$$

The equations defining S_K and R_K show how the borders of the missing mass regions depend on both the parameters of the model and the level of market concentration. Figure 18 demonstrates how the missing mass region's boundaries vary with the parameters σ and α, and the number of banks in the market, K. The empirical shape and size of the missing mass regions allow us to identify and estimate the model parameters, both within and across markets with different concentrations. In general, a larger missing mass region is associated with higher values of σ and α. Intuitively, this is because for a given loan size, higher

values of σ and α are associated with the bank being able to charge a lower mark-up in the unconstrained environment. Under the notched rate cap, this thin profit margin forces banks to "scale back" (i.e. decrease L and potentially increase r) a larger portion of loans rather than pushing them out to the lower cap (hence increasing L and decreasing r). A similar phenomenon occurs as K increases—with more banks in the market, banks' profit margins are smaller. This means more loans are forced to scale back under the rate cap, creating a larger missing mass region. Within a given market, varying α and σ will not only increase the size of the missing mass region, but also change its shape. Therefore, identification of the parameters comes from *both* the absolute area and the shape of the missing mass region.

The parameters α and σ change the size of the missing mass region over which we integrate during estimation. We search for the values of α and σ that minimize the difference in excess and missing mass for each moment and market. Intuitively, the two moment conditions for a single market are sufficient to recover α and the choice elasticity of that market. Additionally, we exploit cross-market variation to recover σ, which governs how the choice elasticity varies with market structure. The model is therefore over identified.

5 Estimation and Results

Here we describe the empirical procedure for estimating the counterfactual distribution and the model parameters using the set of indifference conditions that equate the excess and missing mass in our two-dimensional setting.

5.1 Estimating the Counterfactual Joint Distribution

Estimating the excess mass requires that we compare the observed distribution of contracts $H^P(L, r)$ to the counterfactual distribution $H^0(L, r)$ that would exist in the absence of a notch. Therefore, in this section our goal is to generate a reasonable estimate of the joint distribution of loan amounts and interest rates in a hypothetical world in which the Small Business Administration did not impose a size-dependent interest rate cap. This is a nontrivial problem as we only observe loans created in an environment subject to this rate cap.

Following the identification argument in Section 4, we restrict our sample to the subset of contracts that have interest rates below the interest rate cap. Within this subsample, we estimate the joint distribution of loan size and interest rate, allowing for a flexible correlation structure between the two variables.

The distribution of both interest rates and loan size features pronounced "round number bunching" at familiar basis point or dollar multiples, which generates empirical challenges. This forces us to take a more nonparametric approach: we first fit a flexible polynomial with round number dummies to the marginal distribution of r, accounting for the fact that we observe only the truncated distribution of contracts with interest rates strictly below the lower rate cap ($r|r < \bar{r} \equiv 4.5\%$). We then estimate the distribution of loan size *conditional* on interest rate. Using the estimated parameters, we predict the distribution of contracts, $\left(\hat{r}, \hat{L}\right)$, for $r \geq \bar{r}$.

Below is a more detailed description of the estimation procedure, which composed of three steps:

1. Derive an estimate for the marginal probability function $H_r^0 = P(R = r)$.
2. Estimate the conditional probability function $H_{L|r}^0 = P(L = l|R = r)$.
3. Combine the output from steps 1 and 2 to obtain the joint distribution $H^0 = P(L = l, R = r)$.

Step 1. Estimate Marginal Density Function, $P(R = r)$

We initially focus on estimating the marginal distribution of interest rates $P(R = r)$ using the observed set of contracts with interest rates strictly below the rate cap $P(R < \bar{r})$. The key empirical assumption behind the strategy is that these contracts strictly below the lower rate cap $\bar{r} \equiv 4.5\%$ were not altered by the interest rate cap—which holds true if our model mirrors the true data-generating process—and thus an identical set of loan contracts would have existed in the counterfactual world without the rate caps.

The distribution of observed contracts displays distinct round number bunching at predictable intervals and is also truncated at the notch. Figure 16 in the appendix contains the histogram and CDF of observed loans. In both figures, we see significant spikes occurring at integer interest rates, as well as at multiples of 50 basis points and 25 basis points.

Using the observed data, we discretize r into bins of 1 basis point and fit the following model using nonlinear least squares:

$$P(R \leq r) = \frac{e^{\eta(r)}}{1 + e^{\eta(r)}}$$

where $\eta(r)$ is a polynomial in r with integer dummies:

$$\eta = \mathbb{P}(r) + \delta_1 \lfloor r/0.01 \rfloor + \delta_2 \lfloor r/0.005 \rfloor + \delta_3 \lfloor r/0.0025 \rfloor$$

Here, $\mathbb{P}(r)$ is a polynomial of r with a finite degree, $\lfloor \cdot \rfloor$ is the floor function, and the terms δ_1, δ_2, and δ_3 measure the discontinuous jump in the linear predictor when r reaches, respectively, a round integer interest rate (δ_1), a multiple of 50 basis points (δ_2), and a multiple of 25 basis points (δ_3). We vary the polynomial's degree over different specifications. Using the estimated coefficients, we then recover the CDF $P(R \leq r)$ for $r \geq \bar{r}$ by imposing the assumption that the relationship $\eta(r)$ estimated using $r < \bar{r}$ is accurate and holds for all r. Formally, this requires the true relationship $\eta(r)$ to be globally analytic over the support of r, which is the standard assumption underlying all bunching estimators. Figure 17 in the appendix overlays the estimated CDF from the model with the unconditional CDF from the data.

Step 2. Estimate Conditional Density Function $P(L = l|R = r)$

In this next step, we estimate the conditional density of loan size given interest rates, $P(L = l|R = r)$, using the sample $P(R < \bar{r})$. We again discretize L into bins, each \$2,500 wide, and fit

$$P(L \leq l|R = r) = \frac{e^{\chi(l,r)}}{1 + e^{\chi(l,r)}},$$

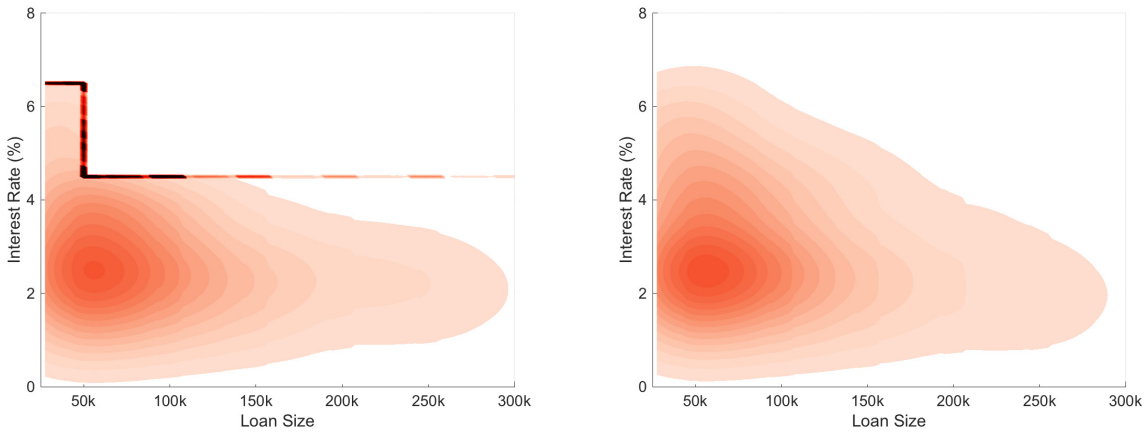
where $\chi(l, r)$ is a polynomial in r and $\log l$ with integer dummies:[6]

$$\chi(l, r) = \mathbb{P}(r, \log l) + \delta_1 \lfloor l/5 \rfloor + \delta_1 \lfloor l/10 \rfloor + \delta_1 \lfloor l/25 \rfloor + \delta_1 \lfloor l/50 \rfloor + \delta_1 \lfloor r/100 \rfloor$$

where l is the loan size in thousands; $\mathbb{P}(r, \log l)$ is a polynomial with finite degrees in r, $\log l$, and their interactions; and δ's captures the discrete change in the CDF around prominent integer levels of loan size. We then follow the estimation procedure described in step one, using NLLS to estimate the parameters of $\chi(l, r)$ from the sample of loans we observe with $r < \bar{r}$ and then employ the estimated χ function to recover the CDF over the entire support of (l, r).

[6] While the maximum loan size for the Express program is \$350,000, we only include loans between \$25,000 and \$300,000 in our estimation procedure, because some irregular endpoint "bunching" occurs at \$350,000, that is distinct from the rest of the distribution. Similarly, a collateral threshold occurs at \$25,000, which also generates idiosyncratic bunching.

Figure 6: Observed and Counterfactual Distribution of Loan Contracts (Averaged Across Markets)



This figure plots the observed and counterfactual densities, pooling over all markets and loans. Here the excess mass (observed - counterfactual) at the threshold, \$50,000, is pronounced and equal to 3 percentage points. The counterfactual distribution also displays some missing mass to the right of the threshold, where loan contracts would have been located in the absence of the discontinuity.

Step 3. Create $H^0(L, r)$

We create the joint predicted distribution $H^0(L, r) = P(L = l, R = r)$ by multiplying the marginal and conditional distributions estimated in steps 1 and 2. To ensure $H^0(L, r)$ is well behaved, we rescale the distribution to ensure that the mass of loans strictly below the rate cap integrates to the same number under H^0 and under H^P conditional on every r or on every l.

Figure 6 plots the observed loan distribution and the predicted counterfactual distribution, pooling over all markets and loans. In the observed distribution, the excess mass at the threshold, \$50,000, and along the interest rate cap is pronounced. The predicted counterfactual distribution spreads this excess mass throughout the region where loan contracts would have been located in the absence of the discontinuity, both above and to the right of the threshold.

5.2 Estimating Parameters

For each market k we calculate the observed empirical joint probability density, $\hat{H}_k^P$ over a two-dimensional grid, with grid points defined by the intervals $L = [25,000 : 2,500 : 300,000]$ and $r = [0 : .0001 : .8]$. The visible bunching in the loan size distribution at round number multiples requires that we use this discrete, rather than a continuous, approach. Using the

method described above, we predict the counterfactual density, $\hat{H}_k^0$, over this same domain and calculate the difference between the two as $\hat{D}_k = \hat{H}_k^P - \hat{H}_k^0$.

Using $\hat{D}_k$, we calculate the empirical analogues to our theoretical moment conditions. Our estimation routine then chooses $(\hat{\alpha},\hat{\sigma}) = \arg\min R(\alpha, \sigma)$, where

$$R(\alpha, \sigma) = \sum_k \left[\left(\hat{E}_{k,1} + \hat{M}_{k,1} \right)^2 + \left(\hat{E}_{k,2} + \hat{M}_{k,2} \right)^2 \right],$$

where $\hat{E}_{k,i}$ is the excess mass for the i-th moment condition in market k and $\hat{M}_{k,i}$ is the corresponding missing mass. The routine uses a Nelder-Mead algorithm to find the minimizers, $\hat{\alpha}$ and $\hat{\sigma}$, of the objective function.

To generate standard errors for these estimates, we resample the rows of our original dataset with replacement $B = 100$ times. For each of our B simulated datasets, we repeat the procedure outlined in this section to generate new estimates, $\{\alpha_b^*, \sigma_b^*\}_{b=1}^B$. We then compute the sample standard errors of our bootstrapped estimates.

5.3 Implementation and Results

Our main specification splits the data into three equally sized groups based on inverse HHI. The most concentrated market group has an average K of 2.28, while the least concentrated group has an average K of 6.98. In other specifications we vary the number of market concentration bins created.

We construct $\hat{H}_k^0$, and $\hat{D}_k$ for $k = 1$ through 3 using the procedure outlined in Section 5.1. Figure 7 plots the distribution of $\hat{D}_k$ across loan size and interest rate for the various groups; pronounced excess mass (in green) occurs along the border of the interest rate cap, where unconstrained contracts with higher interest rates have been forced to bunch. Missing mass (in pink) is concentrated above the cap and to the right of the notch. Visually, the missing mass shifts down as K increases, implying that markups are lower in more-competitive markets.

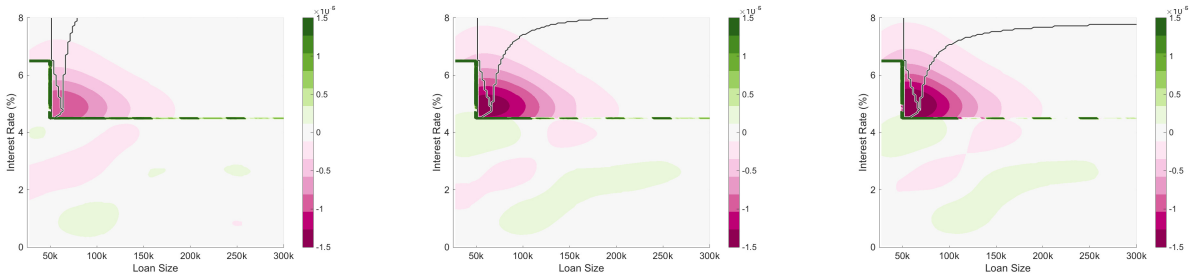
Using the various $\hat{D}_k$, we then select the set of parameters that minimizes the difference between observed excess and missing mass in all markets. Figure 7 overlays the missing mass region's boundary, estimated by these parameters, on top of the calculated difference in distributions.

Table 2: Parameter Estimates

	Main Spec.		Robust(1)		Robust(2)		Robust(3)	
	α	σ	α	σ	α	σ	α	σ
Estimate	0.931	3.225	0.917	3.189	0.933	3.108	0.909	4.365
	(0.011)	(0.472)	(0.012)	(0.696)	(0.007)	(1.39)	(0.013)	(0.769)
# Markets	3		3		6		6	
Polynomial Deg.	1		3		1		3	

This table reports the estimated parameter values for our main specification (column (1)) and three robustness specifications. Standard errors are provided below the point estimates in parentheses. We use a Nelder-Mead algorithm to find the set of parameters that minimizes the difference between observed excess and missing mass in both markets. Standard errors are based on 100 bootstrap simulations. For our robustness estimates we vary the number of K bins used to categorize market concentration, as well as the degree of the polynomial used in the counterfactual estimation. The point estimates in all speification are similar. The estimated parameters in our main specification imply that the most concentrated market (i.e. counties with 2 banks), lenders capture 36% of surplus.

Figure 7: Difference in counterfactual and observed (L, r) distribution and estimated boundaries of missing mass regions for large and small K

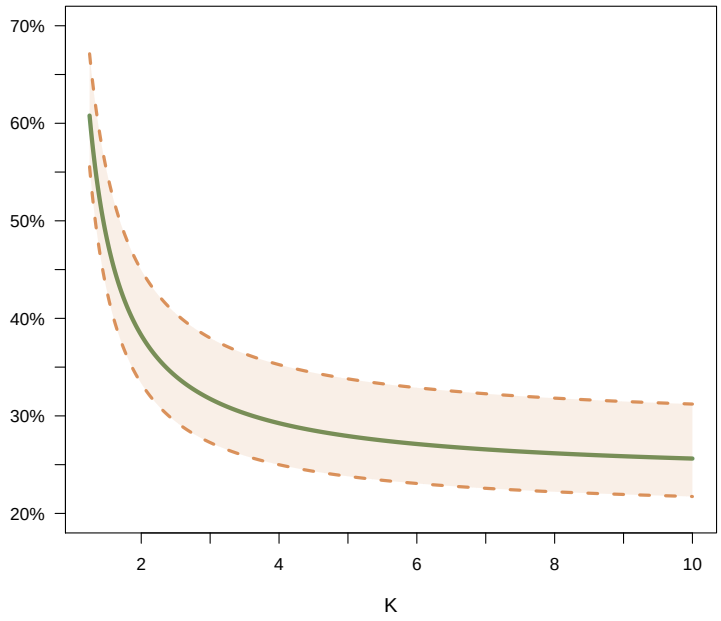


This figure plots the difference in density between the observed loan distribution and the counterfactual distribution after dividing the data into tertiles of inverse HHI, K. Excess mass,shown in green, is concentrated along the interest rate threshold. Missing mass, plotted in pink, indicates where loan contracts would have been located in the absence of the cap. The differences in density below and above the interest rate cap are separately smoothed using a Gaussian low-pass filter. As predicted by the model, the missing mass is concentrated primarily above the cap and to the right of the notch. In gray, we overlay the boundaries of the missing mass region, which is determined by the estimated parameters as well as K.

Using our parameter estimates, we can identify the fraction of total surplus captured by the lender. For example, our results imply that in the most concentrated market (i.e. counties with 2 banks), lenders capture 36% of surplus. Figure 8 graphs the estimated relationship, as described in Section 3, between market concentration and the fraction of surplus captured by lenders in a world without interest rate caps.

Figure 9 then represents the corresponding share of surplus that would be captured by lenders in each U.S. county. As the figure shows, small business lending is more competitive along the coasts, especially in the Northeastern states and in California and Washington.

Figure 8: Fraction of Surplus captured by Lenders vs. Inverse HHI



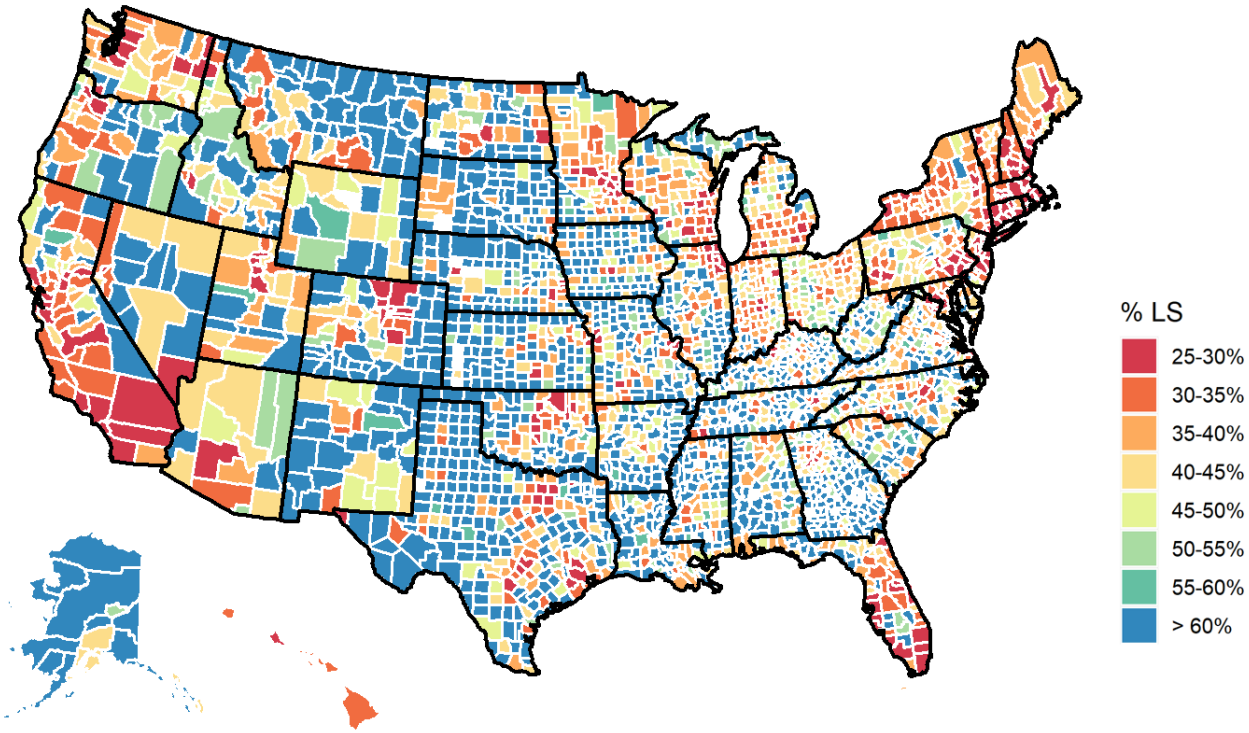
The figure plots the fraction of total surplus that is captured by lenders as a function of market concentration (inverse-HHI) in the absence of a rate cap. Dotted lines represent a 90% confidence interval for our estimate.

Counties have warmer colors on the map, signaling that lenders capture less than half of the total surplus in lending. In contrast, market power is greater in the Midwest and in the South, with a majority of counties colored in blue; this pattern is driven by the fact that there are often few lenders operating in counties within these regions.

While our model and estimation procedure currently abstracts away from moral hazard, in Appendix C we calibrate the impact moral hazard would have on our identification strategy and parameter estimates. We find that given the low default and generous guarantee rate in the data, moral hazard, if it exists, would have only a negligible effect on lenders' decisions in the presence of the interest rate cap and thus our estimates of α and σ.

We also take note that our estimates imply that only counterfactual loans in a duopoly with unconstrained interest rates above 9.52% (8.64% for inverse HHI $= 6$) would have been rationed under the interest rate caps imposed by the SBA policy. Our fitted counterfactual distribution of interest rates, generated using observed loans below the cap with the nonparametric fit described in section 5.1, predicts no loans in this region.

Figure 9: Share of surplus captured by lenders, average of each U.S. county



6 Evaluation of SBA Program and Counterfactual Policies

Our counterfactual policy analysis uses the predicted contract distribution and estimated parameters $(\hat{\alpha}, \hat{\sigma})$ found in Section 5 as a starting point to compute the impact of several policies commonly implemented to address market power. We measure how a government loan guarantee program, a uniform interest cap, and an increase in bank competition change the distribution of contracts, and consequently, the size and division of surplus between borrowers and lenders. We also analyze the welfare impact of the existing policy — the "notched" interest rate cap — both with and without the loan guarantee.

6.1 Theoretical Results

In Section 5, we predicted the distribution of contracts, (r^G, L^G), that would have arisen in a policy environment that featured a 50% loan guarantee program without any interest rate caps. In this subsection, we first show how we can use this to compute the laissez-faire distribution of contracts, (r^{LF}, L^{LF}), that would have appeared without either an interest rate cap or a loan guarantee. Using this, we calculate the hypothetical distribution of

contracts under a wide variety of commonly pursued government policies.

Proposition 4. *Let λ be the fraction of the unpaid loan balance guaranteed by the government in the case of default, and let $\delta_i^G = \delta_i + \lambda(1-\delta_i)$ denote the effective rate of recovery on defaulted debt. Let $R = 1 + r$. The laissez-faire contract can be written as the following function of the loan terms under a guarantee program:*

$$L_i^{LF} = L_i^G \left(\frac{\tilde{c}_{ik}^G}{\tilde{c}_{ik}} \right)^{\frac{1}{1-\alpha}}, \tag{14}$$

$$R_i^{LF} = R_i^G \frac{\tilde{c}_{ik}}{\tilde{c}_{ik}^G}, \tag{15}$$

where $\tilde{c}_{ik} = \frac{c_k - (1-p_i)\delta_i}{p_i}$ and $\tilde{c}_{ik}^G = \frac{c_k - (1-p_i)\delta_i^G}{p_i}$.

The proof is displayed in A. The intuition for Proposition 4 is as follows. The lender selects the loan size to maximize the sum of the borrower and lender surplus. For a given contract size and interest rate, a loan guarantee leads to a negative government surplus and to a higher expected lender surplus; however, the former does not factor into the equilibrium loan size. Therefore, when compared with L_i^G, the laissez-faire setting features a reduced loan size. Additionally, the interest rate increases in the laissez-faire setting, as the lender must be compensated more when the project is successful to cover the additional losses incurred when the project fails.

Using the laissez-faire distribution of contracts, we can directly apply the results in Section 3 to compute the distribution of contracts under a uniform interest rate cap, (R_i^{UC}, L_i^{UC}), and under a "notched" interest rate cap, (R_i^{NC}, L_i^{NC}).

The last counterfactual policy we analyze is one in which the government incentivizes an increase in market competition. Let γ denote the rise in inverse-HHI, so that $(1+\gamma) = \frac{K'}{K}$, where K' denotes the new inverse-HHI. The equilibrium loan size under this policy, L^{IC}, is identical to L^{LF}, as the same loan size continues to maximize the sum of borrower and lender surplus. The equilibrium interest rate will decrease, reflecting the lower lender markup associated with a drop in market share. Specifically, we have that $\frac{R_i^{IC}}{R_i^{LF}} = \frac{\mu^{IC}}{\mu}$, where $\mu^{IC} = \frac{1+\alpha\sigma(1-\frac{1}{K'})}{\alpha+\alpha\sigma(1-\frac{1}{K'})}$.

To evaluate the different policies, we compute the expected borrower surplus, lender surplus, and government surplus for each loan that is not rationed using the following formulas:

$$BS_i = p_i(z_i L_i^{\alpha} - (1+r_i))$$

$$LS_{ik} = L_i(p_i(1+r_i) + (1-p_i)(\delta_i + \lambda(1-\delta_i)) - c_k)$$

$$GS_i = -(1-p_i)\lambda(1-\delta_i)$$

Loans that become rationed contribute zero surplus. Using our prediction for the probability distribution of (R_i, L_i) in the different counterfactual policy environments, we can combine the contract-specific expected surpluses into economy-wide aggregates.

6.2 Recovering Borrower and Bank Characteristics

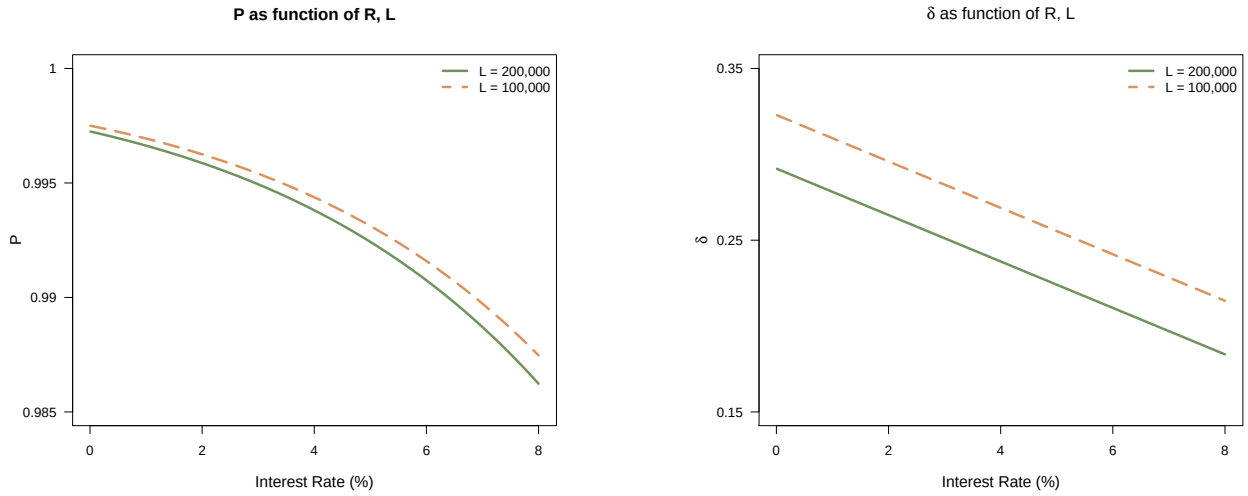
Our identification strategy for the two key model parameters (α and σ) leverages on the bunching of loans along the size-dependent interest rate cap, as detailed in Section 5, and our strategy does not require us to know the characteristics of individual borrowers (probability of default, p_i, recovery rate δ_i, and productivity z_i) and banks (opportunity cost of funds c_k) that give rise to the loan terms we observe in the data.

In this section, we evaluate policy counterfactuals, and doing so requires us to recover these characteristics from the observed loan terms and the level of bank competition in the market. We perform this estimation using information on loan repayment, while maintaining full flexibility in the correlation structure between these parameters. We analyze the loan-level repayment data for contracts lying strictly below the interest rate cap. We split U.S. counties into three equally sized bins based on the inverse-HHI, and within each bin, we use a logistic regression to predict loan charge-off in the first five years of the loan using loan terms (r_i, L_i) as covariates. We then compute the repayment probability p_i by annualizing the estimated probability of full repayment from the logistic model. Next, we need an estimate of the recovery rate, δ_i. To generate this, we again split U.S. counties into three equally sized bins and filter to all loans below the interest rate cap that charged off. We then regress the fraction of the defaulted balance recovered from the borrower (ignoring the loan guarantee money received from the government) on the loan terms (r_i, L_i). Using the fitted coefficients, we estimate δ_i for all loans. Figure 10 displays how the estimated parameters depend on the loan terms for a given level of market competition. As expected, the probability of full repayment and the recovery rate decrease in the interest rate of the loan. For the same interest rate, larger loans have lower default risk and higher expected recovery rates.

To compute c_k, we plug the estimates of $\alpha, \sigma, p_i, \delta_i$, and the observed R_i^G into the equation derived in Proposition 1 for the equilibrium interest rate. Solving that equation for c_k, we get:

$$c_k = p_i R_i^G \left(\frac{\alpha + \alpha\sigma(1-\frac{1}{K})}{1+\alpha\sigma(1-\frac{1}{K})}\right) + (1-p_i)(\delta_i + \lambda(1-\delta_i))$$

Figure 10: p, δ as functions of (r, L) for smallest K group



These plots show the annualized full repayment rate (left) and recovery rate (right) as functions of the interest rate for two different loan sizes. These plots are created using data from the most concentrated market (smallest inverse-HHI)

Finally, we use the equation for the equilibrium loan size derived in Proposition 1 to solve for z_i.

$$z_i = \frac{c_k - (1 - p_i)(\delta_i + \lambda(1 - \delta_i))}{p_i \alpha} (L_i^G)^{\frac{1}{1-\alpha}}$$

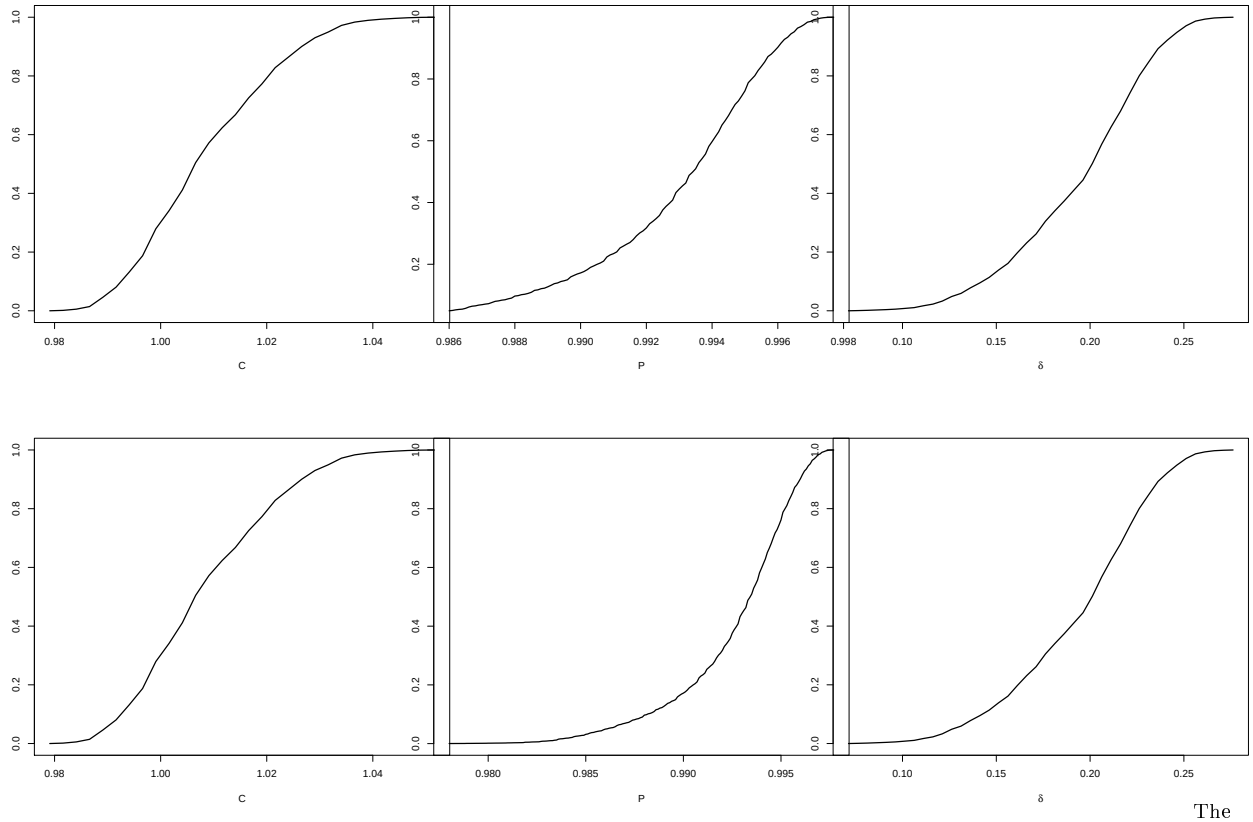
Employing these loan-specific parameter estimates, we can evaluate the government policies as described in the previous subsection.

6.3 Empirical Results

For every laissez-faire contract (r^*, L^*), we compute the counterfactual value of r and L under the following policy interventions: 1) a uniform interest rate cap of 5.5%; 2) the "notched" interest rate cap found in the SBA policy, without the loan guarantee; 3) a guarantee-based subsidy to the lender that reimburses losses at a 50% rate; 4) a 20% increase in market competition; and 5) the SBA policy, which features a "notched" interest rate cap and a 50% loan guarantee. Figure 12 shows the resulting loan-interest rate distributions in a laissez-faire environment and under each scenario.

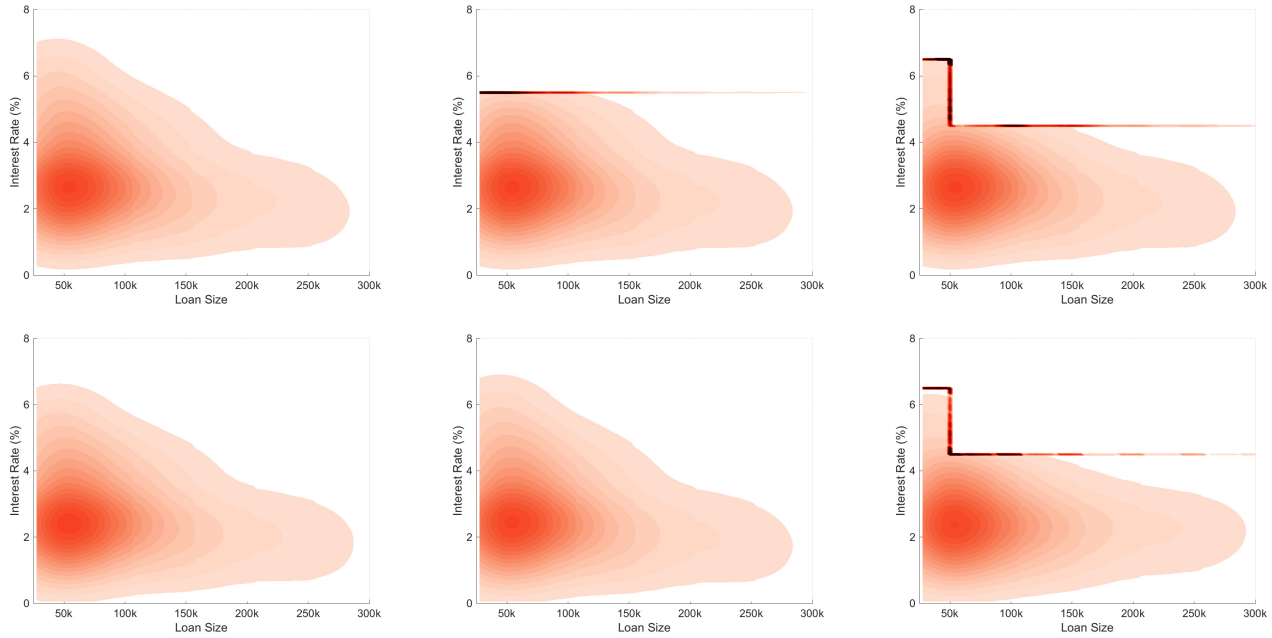
Table 3 reports these policies' impact on the average values of r and L in the distribution as well as on total surplus, lender surplus, and borrower surplus relative to the laissez-faire baseline. The table also reports the percentage of laissez-faire loans that are potentially

Figure 11: CDF of Model Parameters



The first row of figures plots the CDF of the bank's individual cost parameter, c_k (left panel), the borrower's annualized repayment probability, p_i (middle panel), and the expected recovery rate conditional on default, δ_i (right panel), for loans in counties with high market concentration (inverse HHI = 2.28). The second row plots these distributions for loans with low market concentration (inverse HHI = 6.98). For interpretation purposes, a c_k of 1 implies that a bank is indifferent between lending or not as long as the expected return from the loan is equivalent to the prime rate. It may seem surprising then to see a sizeable fraction of banks have a value c_k less than 1. However, banks receive other benefits in the form of closing costs and the possibility for late payment fees levied on borrowers when they offer loans to borrowers. Finally, we see that the banks operating in environments with less competition have lower costs, explaining their higher market share.

Figure 12: Distribution of (r, L) contracts under various counterfactual scenarios

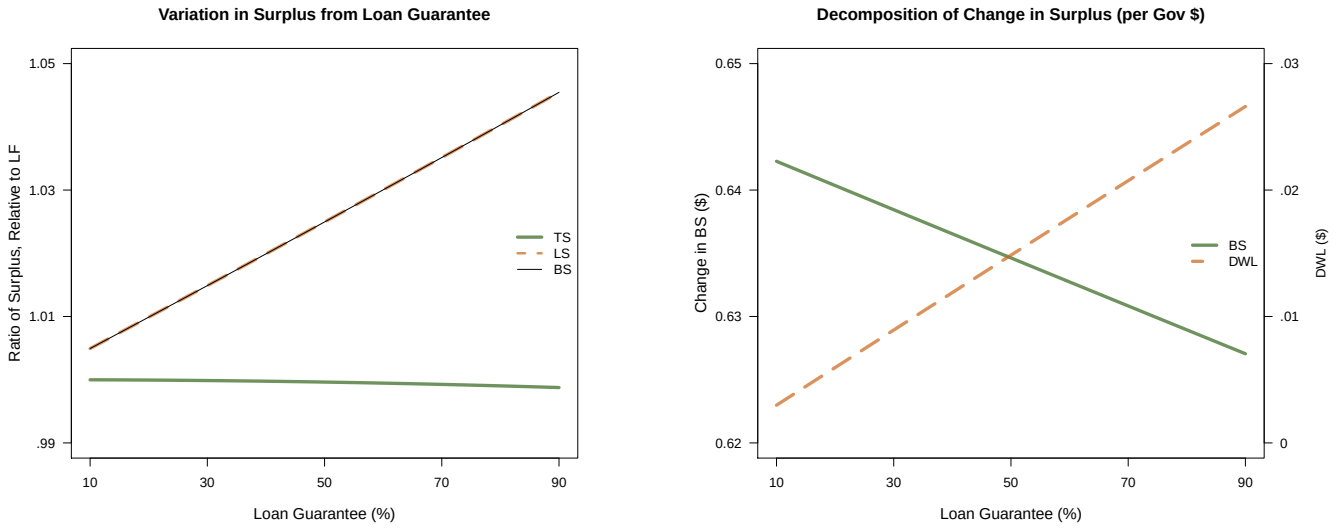


These six plots show the predicted distribution of r and L in concentrated markets (inverse HHI = 2.28) under the laissez-faire baseline (top left), a uniform interest rate cap of 5.5% (top center), the existing interest rate cap "notch" structure, without the guarantee (top right), a 50% guarantee with no interest rate cap (bottom left), a 20% increase in market competition (bottom center), and the existing SBA policy (bottom right).

rationed, or lost, under the counterfactual scenarios with interest rate cap. When interpreting the change in total surplus for the counterfactual policy of expanding market competition, we recognize that market competition is an endogenous outcome. Changing this outcome likely incurs costs, which we do not estimate in this analysis. Instead, we speak to how adjusting the level of market competition alters the split between borrower and lender surplus. We repeat the exercise for both a concentrated market (inverse HHI = 2.28) and a competitive market (inverse HHI = 6.98) to show the nonlinear policy response across markets of different sizes.

In a laissez-faire environment with concentrated markets, we find that lenders capture 36% of the total surplus, while the remaining 64% goes to borrowers. There is some reduction in total surplus due to the distortions induced in scenarios 2, 3, 4, and 6—loan size deviates from its efficient size under both the rate caps and the guarantee, generating inefficiencies. Total surplus remains constant when K increases (scenario 5), since increasing competition will only impact the division, but not the size, of surplus. While both interest rate cap policies boost borrower surplus for affected, but *non-rationed* borrowers, they negatively impact borrowers who are rationed. When the rate cap is set too low, this rationing can be so extensive that the net effect is less borrower surplus. The guarantee policy lowers costs

Figure 13: Changing the size of the loan guarantee



The plot on the left displays the ratio of borrower surplus, lender surplus, and total surplus in a concentrated market (inverse HHI = 2.28) with a loan guarantee to the corresponding surpluses in a laissez-faire environment. The plot on the right shows in green (on the left axis) the increase in borrower surplus per dollar spent through the government's guarantee subsidy. In orange (on the right axis), we see the amount of deadweight loss per dollar of government spending.

for lenders, which in turn both increases loan size and decreases interest rates. However, the cost of the guarantee subsidy must be born by the government and therefore reduces total surplus. Additionally, the benefits of this subsidy are split between the lender and the borrower, because the lender does not entirely "pass through" the reduction in costs to lower interest rates.

Next, we analyze the impacts of changing the policy variables in a continuous fashion. Figure 13 demonstrates how loan guarantee generosity influences surpluses. We see that borrower and lender surplus increase monotonically in the size of the guarantee; however, total surplus decreases slightly due to the costs imposed on the government. Furthermore, we see that less than \$0.65 of each government dollar spent flows to the borrower. This amount decreases in the generosity of the guarantee program, as the amount of deadweight loss as a ratio of government expenditure increases to over \$0.02 for larger guarantee programs. These results indicate that if the government's goal is to subsidize borrowers, then a loan guarantee program is relatively inefficient due to both the deadweight loss and due to the lenders capturing a significant share of the surplus.

Figure 14 shows the effects of changing the uniform rate cap threshold. In the left panel, we see that imposing a binding rate cap reduces lender and total surplus. For modestly sized rate caps, borrower surplus increases, as lower borrowing costs' positive impact dominates

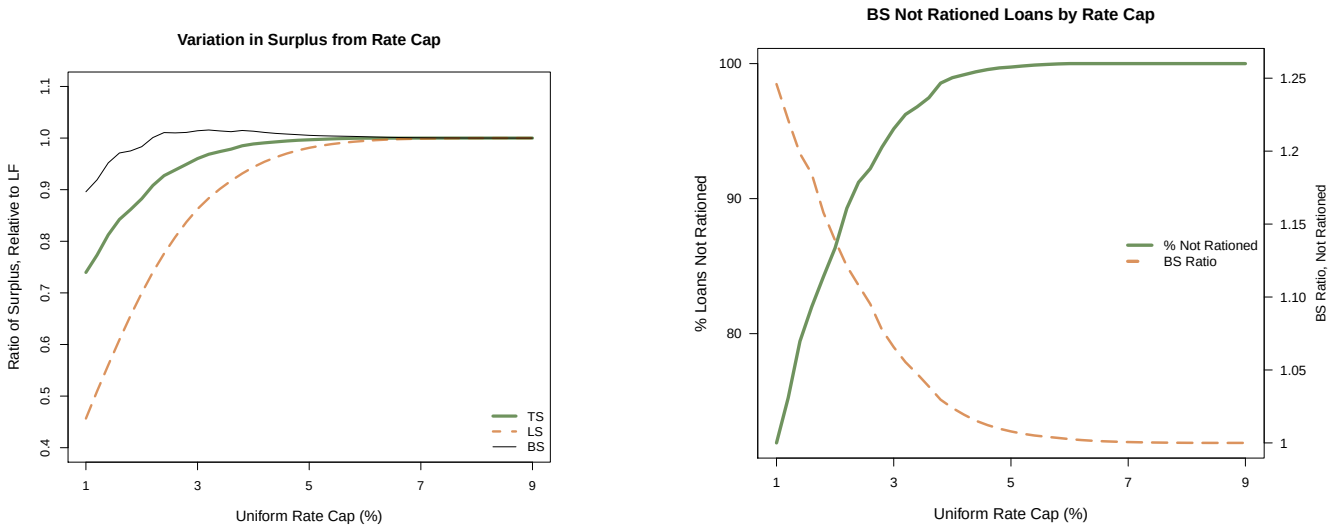
Table 3: Counterfactual Scenarios Calculated for Small and Large K Markets

inverse HHI = 2.28						
	LF	Cap	Notch	Guar	Inc. K	SBA Policy
AvgR	3.01	2.96	2.98	2.79	2.81	2.80
AvgL	88164.08	88601.57	88274.82	90635.43	88164.08	89202.66
BS/BS*	1.00	1.00	1.01	1.03	1.04	1.01
LS/LS*	1.00	0.99	0.98	1.03	0.92	1.01
GS	0.00	0.00	0.00	-173.14	0.00	-168.42
TS/TS*	1.00	1.00	1.00	1.00	1.00	0.98
BS/(BS+LS)	0.64	0.65	0.65	0.64	0.67	0.64
Rationed (%)	0.00	0.10	0.00	0.00	0.00	0.00

inverse HHI = 6.98						
	LF	Cap	Notch	Guar	IncK	SBA Policy
AvgR	3.46	3.33	3.39	3.15	3.42	3.11
AvgL	91721.77	92637.33	91465.98	95050.63	91721.77	93007.68
BS/BS*	1.00	1.00	1.01	1.03	1.01	1.01
LS/LS*	1.00	0.97	0.96	1.03	0.98	1.01
GS	0.00	0.00	0.00	-236.84	0.00	-225.19
TS/TS*	1.00	0.99	0.99	1.00	1.00	0.98
BS/(BS+LS)	0.73	0.74	0.74	0.73	0.74	0.73
Rationed (%)	0.00	0.88	0.16	0.00	0.00	0.00

This table reports how counterfactual policies impact average values of r and L in the distribution and total surplus, lender surplus, borrower surplus, and government surplus relative to the laissez-faire baseline. We analyze 1) the laissez-faire baseline; 2) a uniform interest rate cap of 5.5%; 3) the "notched" interest rate cap used by the SBA, but without the guarantee-based subsidy; 4) a 50% government loan guarantee; 5) an 20% rise in market competition; and 6) the policy used by the SBA, including the "notched" interest rate cap and a 50% guarantee-based subsidy to the lender. The table also reports the portion of laissez-faire loans that are rationed, or lost, under each counterfactual scenario. We repeat the exercise for both a concentrated market (inverse HHI = 2.28) and a competitive market (inverse HHI = 6.98) to show the nonlinear policy response across markets of different sizes.

Figure 14: Changing the size of a uniform rate cap

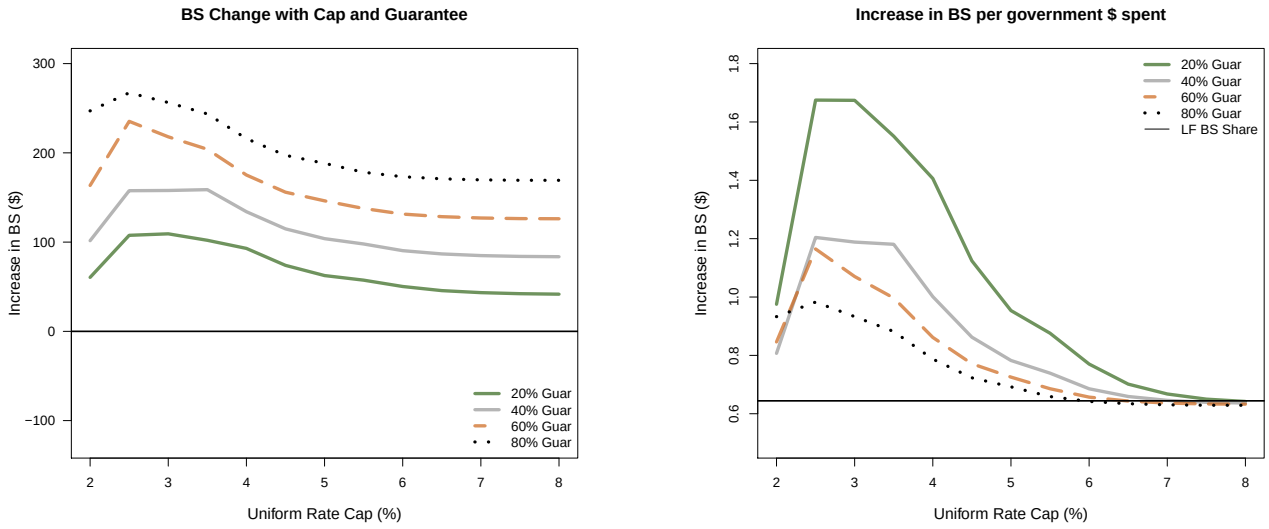


The plot on the left displays the ratio of borrower surplus, lender surplus, and total surplus in a concentrated market (inverse HHI = 2.28) under a uniform interest cap to the corresponding surpluses in a laissez-faire environment. The plot on the right shows in green (on the left axis) the percent of loans that are not rationed. Orange (using the right axis) shows the ratio of the borrower surplus under a rate cap to the laissez-faire surplus, focusing exclusively on loans that remain unrationed.

the effect of rationing; however, for stringent rate caps, borrower surplus diminshes due to substantial rationing. The right panel shows that borrowers, whose loans are not rationed, experience an increase in surplus due to lower borrowing costs. Nevertheless, with modestly sized rate caps, this benefit is small due to the tempering effect of inefficient loan sizes.

In Figure 15, we consider the impact of policies that combine a uniform rate cap and a loan guarantee. Because it decreases the effective cost of lending, a loan guarantee can reduce the rate of rationing in the presence of an interest rate cap. In the left panel, we show that policy makers can substantially increase borrower surplus through the combination of a loan guarantee and a uniform rate cap. As the size of the loan guarantee increases, rationing becomes less severe, and we see that there exists an interior solution for the rate cap's optimal size that maximizes borrower surplus for a given guarantee percentage. In the right panel, we show that coupling these two policies can be an efficient method for governments to subsidize borrowers. With a 20% loan guarantee and a uniform rate cap from 3%–4% percent, each dollar of government expenditure increases borrower surplus by \$1.6. This is more than a 200% increase in the pass-through rate compared with the SBA's current policy.

Figure 15: Loan guarantee and uniform rate cap



The left plot displays the increase in the average borrower surplus in a concentrated market (inverse HHI = 2.28) as a function of the size of the loan guarantee and the interest rate cap. On the right, we see the increase in borrower surplus per dollar spent by the government. Rate caps are evaluated in 50 basis point increments.

7 Conclusion

In this paper we propose a novel, two-dimensional bunching estimator for bank market power by exploiting the bunching of lending contracts given notches in regulatory interest rate caps. We methodologically contribute to the "bunching" literature (Kleven (2016)) by deriving a bunching estimator for settings in whih decision makers have multiple choice variables. We apply the estimator to loans made through the SBA, a federal agency that provides implicit loan guarantees. We find substantial market power in this setting: on average, banks capture 27–36% of surplus in laissez-faire lending relationships and a similar percentage of the additional value created by loan guarantees. We perform a wide range of policy counterfactuals and find the combination of a 20% loan guarantee and a uniform interest rate cap of 3–4% more than doubles the pass-through rate from government dollars to borrower surplus compared with the current SBA policy. Applications of our two-dimensional bunching estimator in other empirical contexts may be a fruitful avenue for future research.

References

Adelino, Manuel, Song Ma, and David Robinson, "Firm age, investment opportunities, and job creation," *The Journal of Finance*, 2017, *72* (3), 999–1038.

Antill, Sam, "Are Bankruptcy Professional Fees Excessively High?," *Working Paper*, 2020.

Benetton, Matteo, "Leverage regulation and market structure: An empirical model of the uk mortgage market," *Available at SSRN 3247956*, 2018.

Benmelech, Efraim and Tobias J Moskowitz, "The political economy of financial regulation: Evidence from US state usury laws in the 19th century," *The journal of finance*, 2010, *65* (3), 1029–1073.

Best, Michael Carlos and Henrik Jacobsen Kleven, "Housing market responses to transaction taxes: Evidence from notches and stimulus in the UK," *The Review of Economic Studies*, 2018, *85* (1), 157–193.

Bhattacharya, Vivek, Gaston Illanes, and Manisha Padi, "Fiduciary duty and the market for financial advice," Technical Report, National Bureau of Economic Research 2019.

Brown, J David and John S Earle, "Finance and growth at the firm level: evidence from SBA loans," *The Journal of Finance*, 2017, *72* (3), 1039–1080.

Carlson, Mark A, Sergio Correia, and Stephan Luck, "The effects of banking competition on growth and financial stability: Evidence from the national banking era," *Available at SSRN 3202489*, 2019.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs*," *The Quarterly Journal of Economics*, 05 2019, *134* (3), 1405–1454.

Crawford, Gregory S., Nicola Pavanini, and Fabiano Schivardi, "Asymmetric Information and Imperfect Competition in Lending Markets," *American Economic Review*, July 2018, *108* (7), 1659–1701.

Cuesta, José Ignacio and Alberto Sepúlveda, "Price regulation in credit markets: A trade-off between consumer protection and credit access," *Available at SSRN 3282910*, 2019.

DeFusco, Anthony A and Andrew Paciorek, "The interest rate elasticity of mortgage demand: Evidence from bunching at the conforming loan limit," *American Economic Journal: Economic Policy*, 2017, *9* (1), 210–40.

Drechsler, Itamar, Alexi Savov, and Philipp Schnabl, "The deposits channel of monetary policy," *The Quarterly Journal of Economics*, 2017, *132* (4), 1819–1876.

Egan, Mark, Ali Hortaçsu, and Gregor Matvos, "Deposit competition and financial fragility: Evidence from the us banking sector," *American Economic Review*, 2017, *107* (1), 169–216.

Gale, William G, "Economic effects of federal credit programs," *The American Economic Review*, 1991, pp. 133–152.

Gelber, Alexander M, Damon Jones, and Daniel W Sacks, "Estimating Adjustment Frictions Using Nonlinear Budget Sets: Method and Evidence from the Earnings Test," *American Economic Journal: Applied Economics*, 2020, *12* (1), 1–31.

Granja, João, Christian Leuz, and Raghuram Rajan, "Going the extra mile: Distant lending and credit cycles," Technical Report, National Bureau of Economic Research 2018.

Jayaratne, Jith and Philip E Strahan, "Entry restrictions, industry evolution, and dynamic efficiency: Evidence from commercial banking," *The Journal of Law and Economics*, 1998, *41* (1), 239–274.

Jiang, Liangliang, Ross Levine, and Chen Lin, "Does competition affect bank risk?," Technical Report, National bureau of economic research 2017.

Kaplan, Steven N and Luigi Zingales, "Do investment-cash flow sensitivities provide useful measures of financing constraints?," *The quarterly journal of economics*, 1997, *112* (1), 169–215.

Kleven, Henrik Jacobsen, "Bunching," *Annual Review of Economics*, 2016, *8*, 435–464.

Lelarge, Claire, David Sraer, and David Thesmar, "Entrepreneurship and credit constraints: Evidence from a French loan guarantee program," in "International differences in entrepreneurship," University of Chicago Press, 2010, pp. 243–273.

Maimbo, Samuel Munzele and Claudia Alejandra Henriquez Gallegos, *Interest rate caps around the world: still popular, but a blunt instrument*, The World Bank, 2014.

Melzer, Brian and Aaron Schroeder, "Loan contracting in the presence of usury limits: Evidence from automobile lending," *Consumer Financial Protection Bureau Office of Research Working Paper*, 2017, (2017-02).

Nelson, Scott T, "Private information and price regulation in the us credit card market," *Unpublished Working Paper*, 2018.

Nguyen, Hoai-Luu Q, "Are credit markets still local? Evidence from bank branch closings," *American Economic Journal: Applied Economics*, 2019, *11* (1), 1–32.

Petersen, Mitchell A and Raghuram G Rajan, "The effect of credit market competition on lending relationships," Technical Report, National Bureau of Economic Research 1994.

Rigbi, Oren, "The effects of usury laws: Evidence from the online loan market," *Review of Economics and Statistics*, 2013, *95* (4), 1238–1248.

Saito, Kuniyoshi and Daisuke Tsuruta, "Information asymmetry in small and medium enterprise credit guarantee schemes: evidence from Japan," *Applied Economics*, 2018, *50* (22), 2469–2485.

Zinman, Jonathan, "Restricting consumer credit access: Household survey evidence on effects around the Oregon rate cap," *Journal of Banking and Finance*, 2010, *34* (3), 546–556.

A Proofs of Propositions

Proof of Proposition 1: The bank's maximization problem is:

$$\max_{r,L} \underbrace{\left[p_i\left(1+r\right)+\left(1-p_i\right)\delta_i - c_k\right]L}_{\substack{\text{expected profit conditioninal} \\ \text{on contract being accepted}}} \times \underbrace{\frac{v_{ik}^{\sigma}}{\sum_{k'=1}^{K} v_{ik'}^{\sigma}}}_{\text{choice probability}} \quad \text{s.t. } v_{ik} = p_i\left(z_i L^{\alpha} - \left(1+r\right)L\right)$$

Let $\tilde{c}_{ik} = \frac{c_k - (1-p_i)\delta_i}{p_i}$,and let $R = (1+r)$. The maximization problem simplifies to

$$\max_{R,L} \left(R - \tilde{c}_{ik}\right) L \frac{v_i^{\sigma}}{\sum_j v_{ij}^{\sigma}}$$

Taking the first-order condition with respect to R:

$$L^* \frac{v_i^{\sigma}}{\sum_j v_{ij}^{\sigma}} - (R^* - \tilde{c}_{ik}) L^* \frac{(p_i \sigma v_i^{\sigma-1} L^*)(\sum_j v_{ij}^{\sigma}) - v_i^{\sigma}(p_i \sigma v_i^{\sigma-1} L^*)}{(\sum_j v_{ij}^{\sigma})^2} = 0$$

Substituting the equilibrium condition $\frac{v_i^{\sigma}}{\sum_j v_{ij}^{\sigma}} = s_K$ and simplifying, we get:

$$L^* s_K - (R^* - \tilde{c}_{ik}) L^* (1 - s_K) \frac{p_i \sigma L^* s_K}{v_i} = 0$$

Substituting for v_i, and further simplifying we get:

$$z{L^*}^{\alpha} - R^* L^* = \sigma\left(1 - s_K\right)\left(R^* - \tilde{c}_{ik}\right) L^*$$

From Proposition 1, we know the equilibrium L^* is chosen to maximize total surplus. Thus L^*maximizes $p_i z_i L^{\alpha} + (1-p_i)\delta_i L - c_k L$. We therefore get:

$$L^* = \left(\frac{\alpha p_i z_i}{c_k - (1-p_i)\delta_i}\right)^{\frac{1}{1-\alpha}} = \left(\frac{\alpha z_i}{\tilde{c}_{ik}}\right)^{\frac{1}{1-\alpha}}$$

Substituting this into the first-order condition for R and simplifying, we get:

$$\mu_{ik} = \frac{R^*}{\tilde{c}_{ik}} = \frac{1+\alpha\sigma(1-s_K)}{\alpha+\alpha\sigma(1-s_K)} = 1 + \frac{1-\alpha}{\alpha}\cdot\frac{1}{1+\sigma(1-s_k)}$$

Proof of Proposition 2: Let $\tilde{c}_{ik} \equiv \frac{c_k-(1-p_i)\delta_i}{p_i}$. The lender's unconstrained profit maximization is given by:

$$\max_{R,L}\left(R-\tilde{c}_{ik}\right)L\frac{v_i^\sigma}{\sum_j v_{ij}^\sigma}$$

First-order condition (see proof of Proposition 1):

$$\{R\} \qquad z_i L^\alpha - RL = \sigma\left(1-1/K\right)\left(R-\tilde{c}_{ik}\right)L$$

$$\{L\} \qquad z_i L^\alpha - RL = \sigma\left(1-1/K\right)\left(RL-\alpha z_i L^\alpha\right)$$

The FOC wrt R implies that the lender's surplus relative to the borrower's surplus is $\frac{1}{\sigma(1-1/K)}$; in other words, the lender captures $\frac{1}{1+\sigma(1-1/K)}$ fraction of total surplus, and

$$\frac{1}{1+\sigma\left(1-1/K\right)}\left(\frac{z_i L^{\alpha-1}}{\tilde{c}_{ik}}-1\right) = \left(\frac{R}{\tilde{c}_{ik}}-1\right) \tag{16}$$

Using the FOC wrt L, we get

$$z_i L^{\alpha-1}\left(1+\alpha\sigma\left(1-1/K\right)\right) = R\left(1+\sigma\left(1-1/K\right)\right)$$

$$L = \left(\frac{R\left(1+\sigma\left(1-1/K\right)\right)}{z_i\left(1+\alpha\sigma\left(1-1/K\right)\right)}\right)^{\frac{1}{\alpha-1}} \tag{17}$$

Solving the unconstrained problem,

$$\tilde{c}_{ik}L = \alpha z_i L^\alpha \implies L^* = \left(\frac{\alpha z_i}{\tilde{c}_{ik}}\right)^{\frac{1}{1-\alpha}}$$

$$\frac{1}{1+\sigma\left(1-1/K\right)}\frac{1-\alpha}{\alpha} = \left(\frac{R}{\tilde{c}_{ik}}-1\right) \implies R^* = \left(\frac{1+\alpha\sigma\left(1-1/K\right)}{\alpha+\alpha\sigma\left(1-1/K\right)}\right)\tilde{c}_i$$

Now suppose R is given exogenously but L is optimally chosen; we want to express (17)

in terms of the unconstrained L^* and exogenous R.

$$\begin{aligned} L &= \left(\frac{R\left(1+\sigma\left(1-1/K\right)\right)}{z_i\left(1+\alpha\sigma\left(1-1/K\right)\right)}\right)^{\frac{1}{\alpha-1}} \\ \left(\left(L^*\right)^{1-\alpha}=\alpha z_i/\tilde{c}_{ik} \implies\right) &= \left(\frac{\alpha R\left(1+\sigma\left(1-1/K\right)\right)}{\left(L^*\right)^{1-\alpha}\tilde{c}_{ik}\left(1+\alpha\sigma\left(1-1/K\right)\right)}\right)^{\frac{1}{\alpha-1}} \\ \left(\tilde{c}_{ik}\left(1+\alpha\sigma\left(1-1/K\right)\right)=R^*\left(\alpha+\alpha\sigma\left(1-1/K\right)\right) \implies\right) &= \left(\frac{R}{\left(L^*\right)^{1-\alpha}R^*}\right)^{\frac{1}{\alpha-1}} \\ &= L^*\left(\frac{R}{R^*}\right)^{\frac{1}{\alpha-1}} \end{aligned}$$

Now suppose L is exogenously given but R is chosen optimally; we express (16) in terms of the unconstrained R^* and L.

$$\begin{aligned} \frac{R}{\tilde{c}_{ik}} &= \frac{1}{1+\sigma\left(1-1/K\right)}\left(\frac{z_i L^{\alpha-1}}{\tilde{c}_{ik}}+\sigma\left(1-1/K\right)\right) \\ \left(\left(L^*\right)^{1-\alpha}=\alpha z_i/\tilde{c}_{ik} \implies\right) &= \frac{1}{1+\sigma\left(1-1/K\right)}\left(\frac{1}{\alpha}\left(\frac{L^*}{L}\right)^{1-\alpha}+\sigma\left(1-1/K\right)\right) \end{aligned}$$

Using $R^*=\left(\frac{1+\alpha\sigma(1-1/k)}{\alpha+\alpha\sigma(1-1/k)}\right)\tilde{c}_{ik}$, we get

$$\begin{aligned} R &= R^*\frac{1}{1+\sigma\left(1-1/K\right)}\left(\frac{1}{\alpha}\left(\frac{L^*}{L}\right)^{1-\alpha}+\sigma\left(1-1/k\right)\right)\Bigg/\left(\frac{1+\alpha\sigma\left(1-1/K\right)}{\alpha+\alpha\sigma\left(1-1/K\right)}\right) \\ R &= R^*\frac{\left(\frac{L^*}{L}\right)^{1-\alpha}+\alpha\sigma\left(1-1/K\right)}{1+\alpha\sigma\left(1-1/K\right)} \end{aligned}$$

Proof of Proposition 3: Consider an unconstrained loan contract (R^*, L^*) that is infeasible under the notched rate cap. We separately consider two scenarios. First, imagine $R^* > \bar{R}^H$. This loan is rationed if the bank's expected profit from offering any feasible contract is negative. This is the case if:

$$(p_i\bar{R}^H+(1-p_i)\delta-c_k)L<0$$

Substituting $\tilde{c}_{ik}\equiv\frac{c_k-(1-p_i)\delta_i}{p_i}$, and recalling that $\frac{1+\alpha\sigma(1-1/K)}{\alpha+\alpha\sigma(1-1/K)}\tilde{c}_{ik}=R^*$, we get that infeasible loans become rationed if:

$$\bar{R}^H < R^* \frac{\alpha + \alpha\sigma(1 - 1/K)}{1 + \alpha\sigma(1 - 1/K)}$$

Assuming the loan is not rationed, by the convexity of the bank's optimization problem, the bank will offer either a contract at the upper rate cap, with $R = \bar{R}^H$or with $R = \bar{R}^L$. Label the latter $R^L$and the former R^H. From Proposition 2, the corresponding contracts will be:

$$(R^L, L^L) = \left(\bar{R}^L, L^* \left(\frac{\bar{R}^L}{R^*} \right)^{\frac{1}{\alpha - 1}} \right)$$

$$(R^H, L^H) = \left(\bar{R}^H, \min \left\{ \bar{L}, L^* \left(\frac{\bar{R}^H}{R^*} \right)^{\frac{1}{\alpha - 1}} \right\} \right)$$

Note that $L^* \left(\frac{\bar{R}^H}{R^*} \right)^{\frac{1}{\alpha - 1}}$could be larger than $\bar{L}$, in which case the maximum feasible loan size the bank could offer with interest rate $\bar{R}^H$is $\bar{L}$.

Second, imagine $\bar{R}^L < R^* < \bar{R}^H$,and $L^* > \bar{L}$. By convexity, this bank will either offer a contract along the "vertical part" of the rate cap (i.e. with $L = \bar{L}$) or along the lower rate cap (with $R = \bar{R}^L$). Again, Proposition 2 gives the contract under either scenario:

$$(R^L, L^L) = \left(\bar{R}^L, L^* \left(\frac{\bar{R}^L}{R^*} \right)^{\frac{1}{\alpha - 1}} \right)$$

$$(R^H, L^H) = \left(\min \left\{ \bar{R}^H, R^* \frac{\left(\frac{L^*}{\bar{L}} \right)^{1-\alpha} + \alpha\sigma \left(1 - 1/K \right)}{1 + \alpha\sigma \left(1 - 1/K \right)} \right\}, \bar{L} \right)$$

To choose between these two contracts, the bank computes its expected profit from each and offers the contract (R^H, L^H) if:

$$L^H \left(R^H - R^* \frac{\alpha + \alpha\sigma(1 - 1/K)}{1 + \alpha\sigma(1 - 1/K)} \right) q_{ik}(R^H, L^H) \geq L^L \left(R^L - R^* \frac{\alpha + \alpha\sigma(1 - 1/K)}{1 + \alpha\sigma(1 - 1/K)} \right) q_{ik}(R^L, L^L) \tag{18}$$

The choice probability, q_{ik}, depends on the contracts offered by all banks within the market. In particular:

$$q_{ik}(R_{ik}, L_{ik}) = \underbrace{\frac{v_{ik}^{\sigma}}{\sum_{k'=1}^{K} v_{ik'}^{\sigma}}}_{\text{choice probability}} \quad \text{s.t. } v_{ik} = p_i \left(z_i L_{ik}^{\alpha} - R_{ik} L_{ik} \right)$$

Note that we can back out a borrower's value of z_i using the equilibrium loan size:

$$z_i = \frac{L_i^{*1-\alpha} \tilde{c}_{ik}}{\alpha} = \frac{L_i^{*1-\alpha} R_i^*}{\alpha} \cdot \frac{1 + \alpha\sigma(1 - 1/K)}{\alpha + \alpha\sigma(1 - 1/K)}$$

Therefore, the probability a bank is chosen by a borrower is given by:

$$q_{ik}(R_{ik}, L_{ik}) = \frac{\left(\frac{L_i^{*1-\alpha} R_i^*}{\alpha} \cdot \frac{1+\alpha\sigma(1-1/K)}{\alpha+\alpha\sigma(1-1/K)} L_{ik}^{\alpha} - R_{ik}^{\alpha} L_{ik} \right)^{\sigma}}{\sum_{k'=1}^{K} \left(\frac{L_i^{*1-\alpha} R_i^*}{\alpha} \cdot \frac{1+\alpha\sigma(1-1/K)}{\alpha+\alpha\sigma(1-1/K)} L_{ik'}^{\alpha} - R_{ik'}^{\alpha} L_{ik'} \right)^{\sigma}}$$

We solve for a symmetric Nash equilibrium in which all banks pursue the same strategy by offering either (R^H, L^H) or (R^L, L^L). First, we check whether all banks offering (R^H, L^H) is a pure strategy Nash equilibrium. To do this, we assume that the other $(K-1)$ banks are offering (R^H, L^H), and then we evaluate equation (16). If satisfied, then (R^H, L^H) is a pure strategy Nash equilibrium. Next, we undertake the same procedure to test whether (R^L, L^L) is a pure strategy Nash equilibrium. This will be the case if the right-hand side of equation (16) is at least as big as the left-hand side.

There are three possibilities. First, if there is only one pure strategy Nash equilibrium, then all banks will offer that contract. Second, if both all banks offering (R^H, L^H) and all banks offering (R^L, L^L) are pure strategy Nash equilibriums, then we assume that banks will select the bank-optimal equilibrium out of these two options. Finally, if there is no pure strategy Nash equilibrium, then we solve for the symmetric mixed strategy equilibriums in which all banks offer (R^H, L^H) with probability λ and (R^L, L^L) with probability $1 - \lambda$. Numerical simulations confirm that this final scenario does not occur for realistic choices of our parameters.

P roof of Proposition 4: Let $\tilde{c}_{ik} \equiv \frac{c_k - (1-p_i)\delta_i}{p_i}$, and let $\tilde{c}_{ik}^G \equiv \frac{c_k - (1-p_i)\delta_i^G}{p_i}$. In Proposition 1, we showed that $L_i^{LF} = (\frac{\alpha z_i}{\tilde{c}_{ik}})^{\frac{1}{1-\alpha}}$, and that $R_i^{LF} = \tilde{c}_{ik} \left(\frac{1+\alpha\sigma(1-1/K)}{\alpha+\alpha\sigma(1-1/K)} \right)$. We now must compute the contract offered in a policy environment with a loan guarantee of size λ. The loan guarantee increases the fraction of the loan balance recovered by the lender in the case

of default from δ_i in a laissez-faire environment, to $\delta_i^G = \delta_i + \lambda(1-\delta_i)$. The new profit maximizing (R_i^G, L_i^G) is given by:

$$\max_{R,L} \left(p_i R + (1-p_i)\delta_i^G - c_k\right) L \frac{v_i^\sigma}{\sum_j v_{ij}^\sigma}$$

First-order condition:

$$\begin{aligned} \{R\} \qquad & zL^\alpha - RL = \sigma\left(1-1/K\right)\left(R - \tilde{c}_{ik}^G\right) L \\ \{L\} \qquad & zL^\alpha - RL = \sigma\left(1-1/K\right)\left(RL - \alpha z_i L^\alpha\right) \end{aligned}$$

These equations are idential to those written in the first part of the proof of Proposition 2, with $\tilde{c}_{ik}^G$ in place of $\tilde{c}_{ik}$. Repeating the steps from that proof, we get $L_i^G = (\frac{\alpha z_i}{\tilde{c}_{ik}^G})^{\frac{1}{1-\alpha}}$, and $R_i^{LF} = \tilde{c}_{ik}^G \left(\frac{1+\alpha\sigma(1-1/K)}{\alpha+\alpha\sigma(1-1/K)}\right)$. Taking the ratio of the optimal contracts under the two policy environments provides the desired results.

B SBA Express Loan Program

The SBA Loan Express program was established in 1995 (under the original name FA$TTRAK) and provides a 50% loan guarantee on loans up to $350,000. It is the second most popular SBA lending program, behind the 7(a) guarantee program.

The primary differences between the Express Loan Program and the SBA's flagship 7(a) loan program is in the maximum loan amounts, which are lower in the Express Loan Program, the prime interest rates, which are higher in the Express program, and the SBA review time, which is typically shorter for Express loans. The documentation necessary for the SBA Express loan is less burdensome compared to the standard SBA 7(a) loans, at the expense of higher interest rates.

There are two types of SBA Express loans. The first type of loans is for businesses that export goods, and the second type is for all other businesses. Lenders can approve a loan or line of credit up to $350,000 with an SBA Express loan, while an Export Express Loan can extend to $500,000. The SBA will respond to an Express Loan application within 36 hours, while the eligibility review for an Export Express Loan will take less than 24 hours.

Loan type and collateral determine the length of repayment. The (expected) life of the collateral is used to determine the repayment length: for example, using real estate for collateral generally leads to a longer repayment period than does using equipment for collateral. More specifically, the maximum SBA Express loan terms are up to 25 years for real estate term loans; up to ten years for leasehold improvement term loans; between 10 and 25 years for equipment, fixtures, or furniture term loans; up to ten years for inventory or working capital term loans; and up to seven years for revolving lines of credit.

C Moral Hazard's Impact on Identification and Parameter Estimates

While highlighting the consequences of imperfect competition on small business lending, our model abstracts away from asymmetric information. We model the probability of project failure, $(1 - p_i)$, as a deterministic, observable borrower characteristic. Moral hazard would instead imply that failure is an increasing function of loan size and interest rate—that is, $p_i'(L(1 + r)) < 0$.[7]

[7] This may occur via a liquidity channel, where an increase in loan size will increase monthly payments and thus the probability of default. It may also distort borrower incentives and lead to a higher default rate via the strategic default channel.

Our identification strategy depends on measuring distortions to (r, L) contracts while holding borrower characteristics constant. If movements in L and r simultaneously cause changes in the marginal cost of lending, this may bias our estimates. For example, banks may be less likely to enlarge constrained loans if doing so increases the probability of default, and thus costs. Banks may instead be more likely to scale back loan size to $\bar{L}$. This could extend the boundaries of our missing mass region, S_K, and inflate our estimates of α and σ.[8] Conversely, if increasing r has an even larger positive effect on default, banks may be *less* likely to scale back loans to $\bar{L}$, and this will bias our estimates downward.

To gauge the extent of this bias, we combine moral hazard estimates from the literature with our own estimates of default, costs, and changes in loan size. We estimate the implied change in lender surplus due to moral hazard for the marginal constrained loans that would otherwise exist on the border of the missing mass region. We compare this to the estimated *overall* lender surplus change that occurs when these loans are "constrained" by the interest rate cap and moved to $\bar{L}$. Moral hazard's impact on lender surplus is more than two orders of magnitude smaller than the typical impact of the cap. This is likely due to the very low default rates and generous loan guarantees observed in our data, as both drive down the benefits accrued by the bank when a loan becomes less risky.

In general, moral hazard estimates from the literature in other consumer credit markets, such as mortgages, credit cards, payday lending, and auto lending, focus on subprime, high-risk borrowers.[9] These estimates are not applicable to our setting, which has a very low average default rate as well as a government guarantee. We instead use IV estimates from a study of government-guaranteed small business loans (Saito and Tsuruta (2018)), that finds that a 1% increase in loan volume leads to a 0.1525% increase in the probability of default.[10]

Given this estimate, we calculate the predicted changes in default rates, costs, and, ultimately, lender surplus for loans that are scaled back to $\bar{L}$ from the border of the missing mass region S_K.[11] We compare the change in lender surplus that comes from scaling back loan size, holding default risk constant, to the change in lender surplus that comes from the moral hazard-induced change in default risk. The results are shown in Table 4, where we run

[8] See Figure 18 for a demonstration of how the missing mass regions' boundaries relate to the parameter values.

[9] For example, Noel and Ganong (2020) estimate that for struggling mortgage borrowers, principal reduction alone has no effect on the probability of default. However, a "one percent payment reduction reduces default rates by about one percent".

[10] This paper studies loans backed by both a 100% and 80% government guarantee. It finds a smaller moral hazard effect for loans made under the less generous guarantee. This suggests that moral hazard may have an even smaller effect in our setting, where the government guarantee is 50%.

[11] The details of the calculations can be found in the subsection below.

Table 4: Absolute and relative changes in lender surplus from loan size reduction and moral hazard

	$\%\Delta LS$, no MH	$\%\Delta LS$, from MH	Ratio
$\upsilon_{(1-p),L} = .05\%$	-11.09%	.01%	824
$\upsilon_{(1-p),L} = .15\%$	-11.09%	.04%	274
$\upsilon_{(1-p),L} = .25\%$	-11.09%	.07%	164

This table shows calculated changes in lender surplus in the constrained region for several moral hazard estimates. The first column calculates the change in lender surplus for constrained loans, without moral hazard. The second column calculates the change in lender surplus for constrained loans coming from the moral hazard effect. The third column shows the ratio of these two effects. The effect of moral hazard is negligible, especially in comparison to the effect of the scale-back in loan size. Each row repeats this exercise for a different etimate of moral hazard elasticity—i.e. the percentage change in default given a percentage change in loan size. The second row is our preferred specification, as it uses an empirical estimate from the literature. As moral hazard becomes more severe, the change in lender surplus from moral hazard increases slightly, but still remains very small.

the scenario for a range of moral hazard elasticities. Scaling back loan sizes in this region—which is the basis of our identification strategy—generates a substantial -11% change in lender surplus. In comparison, the additional change in lender surplus potentially stemming from moral hazard is only .01–.07%. This suggests that moral hazard, if it exists, would have only a negligible effect on lenders' decisions in the presence of the interest rate cap and thus on our estimates of α and σ.

It is important to note that the above exercise only considers moral hazard's effect on the loan size dimension, not the interest rate dimension. When loan size is scaled back in the region S_K, the interest rates on these contracts *increase.* Under a model of moral hazard in which $p_i'(L(1+r)) < 0$, this would elevate the probability of project failure and counteract the opposing effect of scaling back loan size. This would push the estimated effect of moral hazard on lender surplus in Table 4 even closer to zero.

Finally, the potential presence of moral hazard motivates our use of both within and across-market moments for identification. If moral hazard has similar consequences in markets of varying size K, then cross-market identification will "difference out" this common effect. The remaining variation in distortions across K will be caused by the competition mechanism that we highlight in our model.

Moral Hazard Calculation Details

Here we outline the details of the moral hazard calculation.

- Using the estimated missing mass boundaries, we first calculate the average percentage change in loan size for loans within the constrained region. This involves calculating the difference between the counterfactual loan size in a laissez-faire environment versus the constrained loan size at $\bar{L}$ for each contract.

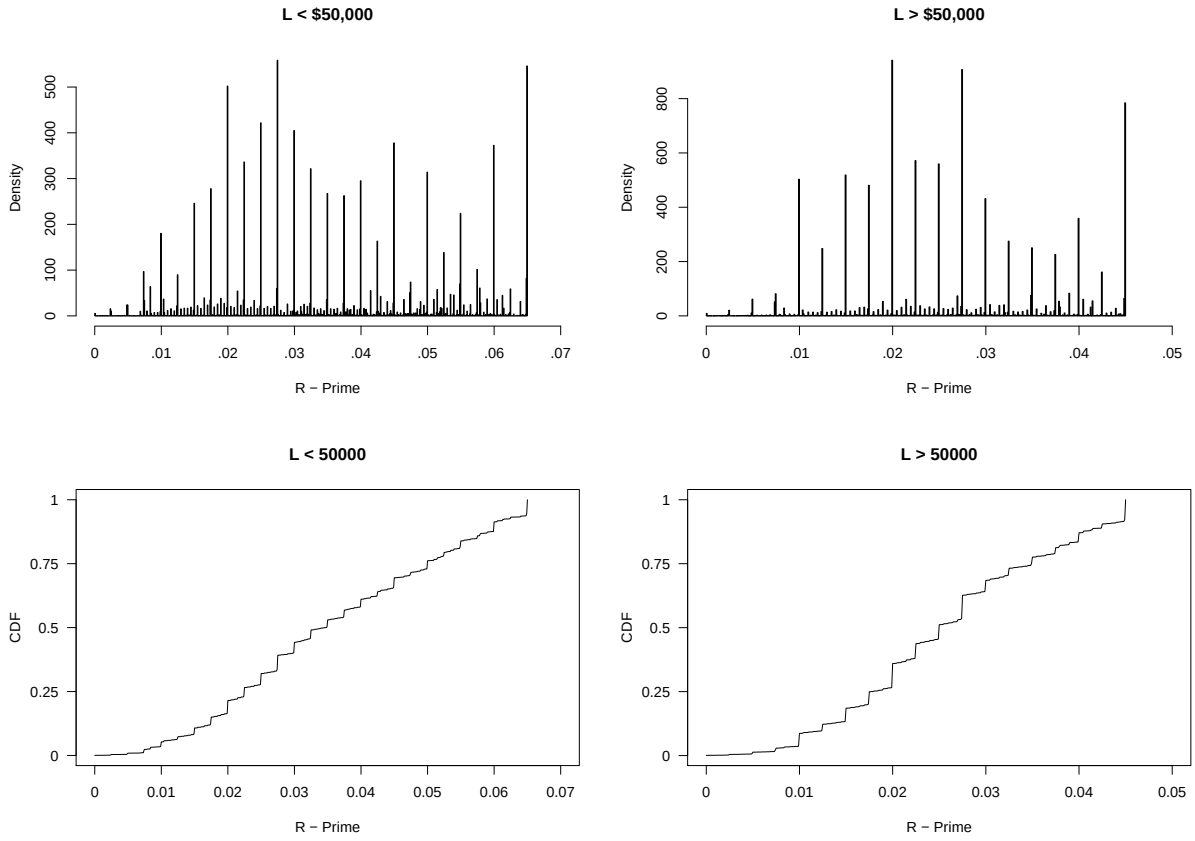
- Next, we multiply this percentage change in loan size by the elasticity of moral hazard. This gives us a predicted percentage change in default rates. Using observed default rates, $(1 - p_i)$, we then calculate the new predicted default rate, $(1 - p_i')$, for loans within the region after they become constrained at the notch.

- We calculate marginal costs for constrained loans both with and without moral hazard using the formula: $c_k = p_i R_i(\frac{\alpha+\alpha\sigma(1-\frac{1}{K})}{1+\alpha\sigma(1-\frac{1}{K})}) + (1 - p_i)(\delta_i + \lambda(1 - \delta_i))$. We adopt the same values for the recovery rate and guarantee rate as in the counterfactual section.

- Finally, we calculate lender surplus for loans in the missing mass region under three scenarios: 1) in an environment featuring unconstrained loan terms with observed default rates, 2) under the interest rate cap with the same observed default rates, and 3) under the interest rate cap with default rates adjusted for moral hazard. This uses the lender surplus equation: $LS_i = L_i(p_i(1 + r_i) + (1 - p_i)(\delta_i + \lambda(1 - \delta_i)) - c_k)$.

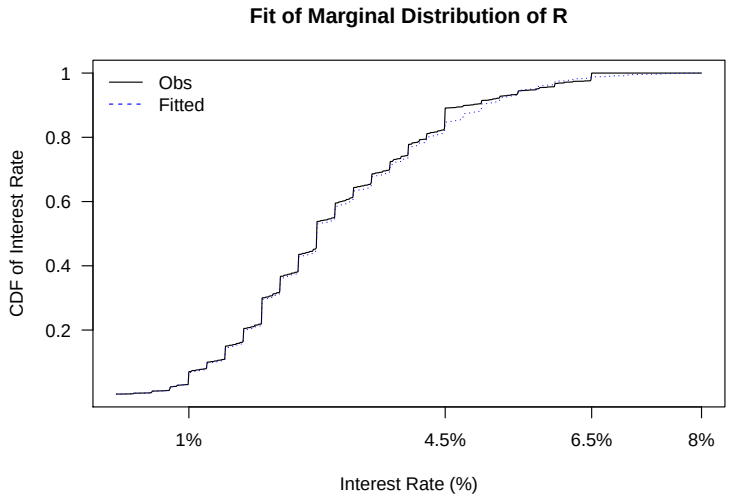
D Additional Figures

Figure 16: Observed Marginal Distribution of Interest Rates

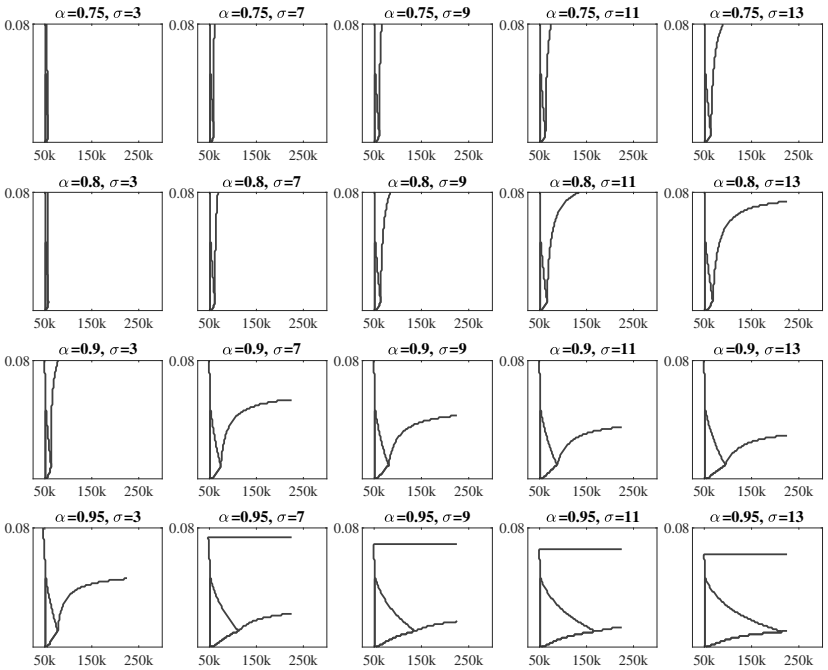


These plots demonstrate the extent of round number bunching in the observed distribution of interest rates. We observe substantial jumps in the density and CDF plots, whenever the interest rate is a multiple of 100 basis points, 50 basis points, or 25 basis points.

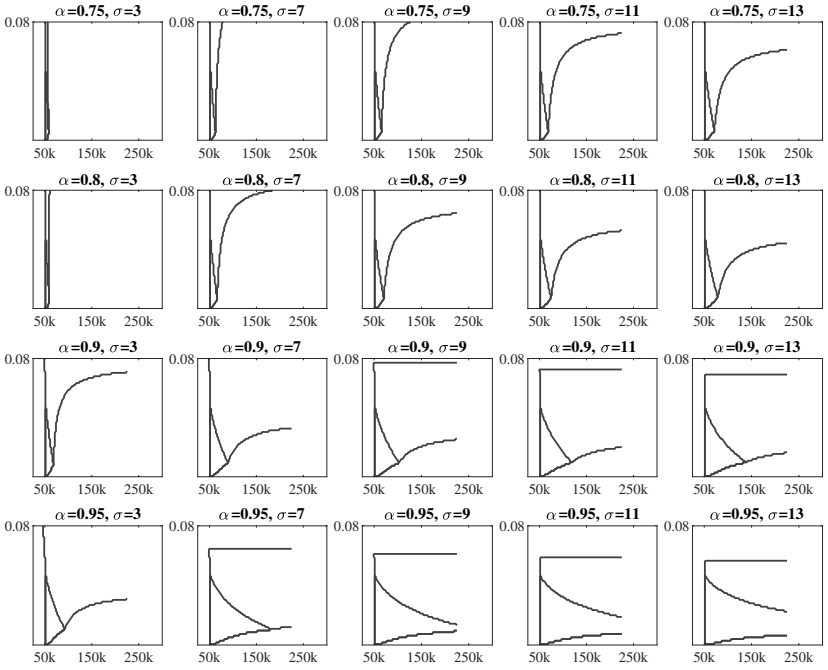
Figure 17: Observed vs. Estimated Marginal Distribution of Interest Rates



This figure plots the estimated (in red) and observed (in black) marginal CDF of interest rates. The estimated CDF is created by fitting the model $P(R \leq r) = \frac{e^{\eta}}{1+e^{\eta}}$ using nonlinear least squares where the linear predictor, η, is given by $\eta = \mathbb{P}(r) + \delta_1 \lfloor r/0.01 \rfloor + \delta_2 \lfloor r/0.005 \rfloor + \delta_3 \lfloor r/0.0025 \rfloor$. The floor function accounts for the visible "spikes" occurring in the distribution at integer interest rates and at multiples of 50 basis points and 25 basis points.$\mathbb{P}(r)$is a polynomial in r.



(a) K=2



(b) K=7

Figure 18: Variation in missing mass regions generated by changes in σ, α, and K.

These plots demonstrate how the boundaries of the missing mass regions vary with the parameters σ and α, as well as with the number of banks in the market, K. Further, these plots visualize how the empirical shape and size of the missing mass regions allows us to identify and estimate the model parameters, both within and across markets of different concentrations. In general, a larger missing mass region is associated with higher values of σ and α. Intuitively, this is because for a given loan size, higher values of σ and α are associated with the bank being able to charge a lower markup in the laissez-faire environment. Under the notched rate cap, this thin profit margin forces banks to "scale back" (i.e. decrease L and potentially increase r) a larger portion of loans rather than pushing them out to the lower cap (hence increasing L and decreasing r). A similar phenomenon occurs as K increases—with more banks in the market, the LF profit margin shrinks for a given loan. This means more loans are forced to scale back under the rate cap, creating a larger missing mass region.