PRELIMINARY AND INCOMPLETE

Bank Specialization in Sparse Markets*

Anne Duquerroy
Banque de France

Clément Mazet-Sonilhac
Bocconi University

Jean-Stéphane Mésonnier
SciencesPo, Banque de France

Daniel Paravisini
LSE, CEPR

Niels Wagner
London School of Economics

November 13, 2026

Click here for the latest version

Abstract

Competition authorities increasingly narrow credit markets by borrowers' characteristics, on the premise that small firms may need lenders with specific expertise. Yet narrowing the market definition makes data sparse, which mechanically inflates the standard measures of specialization and concentration. We measure a branch's specialization by the percentile of its portfolio share under random matching, a counterfactual without specialization, so that sparsity is not mistaken for specialization. In French credit data, branches specialize well beyond random matching in every market definition. Branch closures, which move borrowers to another branch of the same bank, show that a more specialized branch shifts credit away from the firms that later default and towards those that do not. We conclude that specialized lenders gain informational advantages through better screening in narrowly defined markets.

JEL classification: G21
Keywords: bank specialization, screening, sparse markets, SMEs, branch closures.

*Anne Duquerroy: Banque de France (e-mail: anne.duquerroy@banque-france.fr), Clément Mazet-Sonilhac: Bocconi University – Department of Finance (e-mail: clement.mazetsonilhac@unibocconi.it), Daniel Paravisini: LSE, CEPR (e-mail: D.Paravisini@lse.ac.uk), J.-S. Mésonnier: Banque de France, SciencesPo (e-mail: jean-stephane.mesonnier@banque-france.fr), Niels Wagner: LSE (e-mail: n.k.wagner@lse.ac.uk).

We thank Philippe Aghion, Antonin Bergeaud, Thomas Chaney, Hans Degryse, Michael Goedde-Menke, Mathias Le, Filippo De Marco, Nicola Limodio, Melina Papoutsi, Daniel Streitz, and participants in the seminars at LSE, Sciences Po, Bocconi University, OFCE, 2020 EFA Annual Congress and the 2020 Annual Workshop of the Eurosystem's Research Cluster on Financial Stability for their comments. The views expressed herein are those of the authors and do not necessarily reflect those of the Banque de France or the Eurosystem.

1 Introduction

> *"Small businessmen especially are, as a practical matter, confined to their locality for the satisfaction of their credit needs."*
>
> — *United States v. Philadelphia National Bank*, 374 U.S. 321, 369 (1963)

> *"Small businesses ... may need specific kinds of credit products or institutional expertise."*
>
> — U.S. Department of Justice, *Banking Addendum to 2023 Merger Guidelines* (2024)

How narrowly to draw the boundaries of credit markets is an open question for competition policy and banking supervision.[1] Competition authorities and researchers, who long drew these boundaries by geography alone, have increasingly added borrowers' characteristics, such as their size and economic activity, to the dimensions along which markets are partitioned.[2] An unintended consequence of this evolution is that data become sparse: every borrower characteristic added to the definition splits each lender's borrowers across more markets, which reduces the number of borrowers per lender in each market. Even natural definitions make data sparse. In the French credit register, crossing commuting zones with two-digit industries leaves fewer borrowing firms than bank branches in most markets.[3]

[1] As of 2025, the Federal Reserve screens bank mergers by the concentration of deposits in banking markets made of groups of counties (Board of Governors of the Federal Reserve System, 2025), and in 2026 the Federal Deposit Insurance Corporation proposed to adopt the same markets (Federal Deposit Insurance Corporation, 2026). Bank supervisors require banks to identify and limit concentrations of credit to borrowers "in the same economic sector or geographic region", and leave each bank to define them (Basel Committee on Banking Supervision, 2023, SRP32.9–32.10).

[2] The U.S. Supreme Court defined banking markets by geography in *United States v. Philadelphia National Bank* (1963). The U.S. bank merger guidelines of 1995 added a second screen, on commercial lending, to capture competition for loans to small and medium-sized businesses (U.S. Department of Justice and federal banking agencies, 1995). In 2024, the U.S. Department of Justice withdrew from those guidelines and stated that it would consider whether a merger harms any distinct group of customers (U.S. Department of Justice, 2024, Section 1); the merger guidelines of 2023 allow such groups to be defined by observable characteristics of the customers (U.S. Department of Justice and Federal Trade Commission, 2023, Section 4.3.D.1). In 2009, the French competition authority assessed a bank merger by examining lending to firms département by département, and treated credit to farmers as a market of its own (Autorité de la concurrence, 2009). In research, whether credit markets are still local is itself a question (Petersen and Rajan, 2002; Nguyen, 2019), and the concentration of deposits in a county or metropolitan area measures the competition that shapes credit to young firms in Petersen and Rajan (1995) and the transmission of monetary policy through deposit rates in Drechsler et al. (2017). Research has also distinguished borrowers by their activity (Paravisini et al., 2023; Blickle et al., 2026); we review this work at the end of this introduction.

[3] In the data of Section 3.1, commuting zones are the urban areas of the French statistical institute, and firms are grouped into 78 two-digit industries. The median urban area has 6 bank branches, and the

This is a problem because sparse data produce statistical patterns that can be mistaken for the economic phenomena under study. For example, drawing credit markets more narrowly by industry mechanically raises the standard measures of lender concentration and specialization, such as the Herfindahl index and portfolio shares, which makes these measures unreliable.[4] The problem is compounded when a few large borrowers hold much of a market's credit (the market is granular, Gabaix, 2011), because granularity further reduces the effective number of borrowers in a market.[5] Sparsity thus makes it difficult to tell whether lenders specialize in narrow markets at all, let alone whether their specialization signals lending advantages or market power.

In this paper we propose a statistical approach to measure specialization when narrowly defined markets create a sparse data problem. We combine it with a branch closure research design using branch-level data for France to answer two questions. First, do bank branches specialize in narrow markets, defined jointly by geography and industry, beyond the statistical artifact of sparsity? As in Paravisini et al. (2023), we define a lender to be specialized in an industry if its loan portfolio is tilted towards that industry. We depart from them in how we measure it: we compare the observed tilt with the tilts produced by a random counterfactual without specialization. Second, does a branch's observed spe-

median market formed by an urban area and an industry has 4 borrowing firms. In 72% of these markets, there are fewer borrowing firms than branches. In 90% of the pairs formed by a branch and an industry of its market, the branch lends to at most one firm of the industry.

[4] As markets are drawn more narrowly, each holds fewer lending relationships (all the loans of a firm at one branch), and the lowest value the Herfindahl index can take rises. With n relationships in a market, at most n lenders hold its credit, so the index is at least $10{,}000/n$ points, whatever the lenders do. In the French credit data of Section 3.1, the median market of an urban unit and an industry holds 18 relationships with the 9 broad industries and 2 with three-digit industries, so its floor rises from 556 to 5,000 points, far above the 1,800 points at which the 2023 U.S. merger guidelines consider a market highly concentrated (U.S. Department of Justice and Federal Trade Commission, 2023, Section 2.1). Unequal relationship sizes raise the floor further. The index is lowest when every relationship is with a different lender, and even then a single relationship holding 43% of the market's credit puts it above 1,800. In the same data, this lowest possible value exceeds 1,800 in 39% of the urban-unit markets with broad industries and in 82% with three-digit industries. In Section 3.1 and Appendix C we measure how much of the concentration and specialization of French credit markets is mechanical.

[5] The effective number of borrowers is the number of equal-sized borrowers that would give the market the same Herfindahl index. When firm sizes follow Zipf's law, the Herfindahl index of N firms falls only like $1/(\log N)^2$, rather than $1/N$ (Gabaix, 2011, Proposition 2), so their effective number grows only like $(\log N)^2$. Granularity also affects share-based measures in other fields: the index of geographic concentration of Ellison and Glaeser (1997) corrects for it through the Herfindahl index of plant sizes, and Gaubert and Itskhoki (2021) show that sectoral export shares overstate a country's fundamental comparative advantage in its most export-intensive sectors.

cialization signal a screening advantage in lending to small and medium-sized enterprises (SMEs)?[6] We infer it from how much new credit the borrowers of a branch receive, and which of them receive it, when the specialization of their lending branch changes. We exploit the bulk transfer of loan accounts in restructuring branch closures as a plausibly exogenous source of variation in the specialization of a firm's lending branch. In these closures, unrelated to mergers, a French bank transfers the loan accounts of the closing branch to another of its branches, which reshuffles the match between firms' industries and the industry specialization of their branches. Better screening predicts that a more specialized branch shifts its credit away from the firms that turn out to be bad risks and towards those that turn out to be good risks.

To distinguish true specialization from the one that emerges mechanically in sparse markets, we adopt the random benchmark framework that has been used in economics to study stock market forecasting (Cowles, 1933), the geographic agglomeration of industries (Ellison and Glaeser, 1997), and the extensive margin of international trade (Armenter and Koren, 2014).[7] The framework compares an observed statistic, such as the share of a branch's lending that goes to an industry, with the distribution of the same statistic that would arise if firms and branches were matched at random within each geographic market. Random matching is particularly well suited to the analysis of credit market specialization because it captures the economics of the null hypothesis of no specialization. On the supply side, random matching gives each branch, in expectation, the market's portfolio: the natural benchmark when no branch has a lending advantage in any industry.

[6] A lender has a *screening advantage* in an industry when it is better than other lenders at telling the firms of the industry that will repay from those that will not. Our null hypothesis is that specialization signals no lending advantage: in a competitive market, word of mouth among borrowers, herding, or differences in how lenders market their credit can tilt portfolios without any lending advantage. A screening advantage can also give a lender market power, its ability to worsen the terms of its credit without losing its borrowers, because lenders that know less face a winner's curse (Section 4). SMEs are firms with fewer than 250 employees and either turnover below EUR 50 million or total assets below EUR 43 million. We exclude micro-enterprises, financial and real estate firms, public entities and non-resident firms (Section 4.2).

[7] In international trade, Armenter and Koren (2014) show that the pattern of zero trade flows and the frequency of multiproduct exporters in disaggregated data, facts used to test theories of the extensive margin of trade, are reproduced by random assignment alone. In economic geography, Ellison and Glaeser (1997) show that an industry with few plants looks geographically concentrated even when its plants choose their locations at random. The closest problem to ours arises in the measurement of segregation: when units are small, random allocation of workers to firms alone produces large segregation indices, and Carrington and Troske (1997) measure segregation against random allocation instead of an even allocation.

It also gives the dispersion around that portfolio, so that the large shares produced by sparse markets are not mistaken for specialization. On the demand side, random matching captures the notion that firms do not value the lender's expertise in their industry: for a firm, any lender in its market is as good as any other.

Our measure of specialization, p, applies this comparison to the share of a branch's lending that goes to an industry. It is the probability that random matching would give the branch a smaller share of its lending in the industry than it actually has. That is, p is the percentile of the branch's share in the distribution of shares under random matching: the random-matching percentile, for short.[8] A value of p close to 1 means that random matching would rarely give the branch a share as large as the one it has.

We construct p to have four properties. First, it is specific to the bank branch: firms borrow from branches, and branches of the same bank can specialize in different industries. Second, it does not mistake for specialization the large portfolio shares produced by sparse data without specialization, a branch's size, or the size of an industry's loans. Third, its value for a branch that is not specialized is the same in every market, industry, market definition and period, so that values can be compared across them. Fourth, it leaves out the firm's own loan, so it takes one value for each firm that borrows from the branch. Without leaving it out, a firm's borrowing would mechanically raise the measured specialization of its lender, and any regression of the firm's outcomes on that specialization would be biased

[8]Formally, a relationship r is all the loans of a firm in industry j at branch b in a quarter. Let S_{bj}^{-r} be the share of industry j in the branch's lending to its other borrowers, and $\tilde{S}_{bj}^{-r}$ the same share under random matching, which reassigns every relationship of the market to one of its branches, keeping its amount and industry. Then $p_r = \Pr(\tilde{S}_{bj}^{-r} < S_{bj}^{-r}) + \frac{1}{2}\Pr(\tilde{S}_{bj}^{-r} = S_{bj}^{-r})$, over the draws in which the branch receives at least one relationship other than r. Together with leaving r out, counting ties as one half and keeping the draws in which the branch lends nothing to industry j make the expected p_r exactly $1/2$ for a branch that is not specialized, however sparse the market. The benchmark is exact given the assignment probabilities; we estimate them from the same data, which moves it by a few hundredths for the largest branch of each market. A geographic market holds the branches located in it and all their relationships, wherever their borrowers are, so, under random matching, a firm's alternatives are the branches located where its lender is. A relationship goes to a branch with a probability proportional to the branch's lending, tilted towards branches whose size class holds more than its share of lending in the relationship's size class. We measure these matching rates in the branch's département rather than in each market, because in a small market a dominant branch forms a size class of its own and would receive back the relationships it actually holds. The cost is that each branch's expected lending equals its actual lending exactly when the market is a département, and approximately in other markets. We compute p separately for each market and quarter. We define p in Section 2.4 and construct it step by step in Appendix A.

by construction.[9]

We made several explicit choices so that p has these properties. For example, we define geographic markets by the location of the branch, so that each branch's whole portfolio lies in one market. Random matching gives each branch, in expectation, the lending it actually has, so that its share is compared with the shares of a branch of the same size. It also gives large loans mostly to large branches, as in the data, so that a large branch does not look specialized in industries of large loans. We count ties (draws in which random matching gives the branch exactly the share it has) as one half. Together with leaving the firm's own loan out, this is needed for p to equal $1/2$ on average for a branch that is not specialized, in every market and period. The approach is flexible, and these choices can, and should, be adjusted to the economic question at hand.

We apply the method to France, whose credit register records the branch that lends to each firm. We measure specialization under twenty market definitions, from France as a whole to urban units and from 9 broad to 241 three-digit industries.[10] The specialization shown by portfolio shares in narrow markets is largely reproduced by random matching, which contains no specialization: when we split industries from broad to three-digit, random matching reproduces at least 90% of the rise in the shares of the branches with top-quartile portfolio shares.[11] Yet, measured by p, bank branches specialize in narrow markets well beyond what random matching produces: median p is between 0.84 and 0.87 with broad industries, against $p = 0.5$ for a branch that is not specialized. Drawing markets more narrowly also uncovers real differences in specialization. Neighbouring

[9] Leaving out the firm's own loan is also needed for the third property: without it, even a branch that is not specialized would look specialized on average, the more so the sparser the market (Section 2.3).

[10] An urban unit of the French statistical institute is a continuously built-up area of at least 2,000 inhabitants, similar in concept to an Urban Area of the US Census Bureau (U.S. Census Bureau, 2022). The urban areas of the French statistical institute, despite the name, are commuting zones, larger than urban units (Section 3.1).

[11] We select these branches with the top-quartile rule of Paravisini et al. (2023), which classifies a bank as specialized in an activity when its portfolio share in that activity is in the top quartile of all banks' shares. We measure their specialization by their relative portfolio share, their portfolio share divided by the industry's share of the market's lending (Section 3.1). Random matching reproduces 94% of the rise in urban units and 97% in urban areas, and it produces a rise 1.5 to 1.8 times as large as the observed one in départements, régions and France as a single market. Industries are those of the French classification, NAF rev. 2, which coincides with the European classification, NACE Rev. 2, down to four digits: the 9 broad industries are its coarsest aggregation (A10) and the 241 three-digit industries its groups; between them lie 17 industries and 78 two-digit industries.

towns differ in how specialized their branches are about as much as commuting zones differ from one another, so specialization is local.[12] Because p is specific to the branch, it also shows that branches of the same location specialize in different narrow industries of the same broad industry, a division that averages over the location hide and that the branch closures exploit.[13]

Since sparse data can also hide specialization, we pair p with a diagnostic tool: a power analysis that tells how much specialization p can detect in each market, given its sparsity. It tells, for example, whether a value of p close to $1/2$ means that the branch is not specialized or that the market is too sparse to tell. We construct it by comparing a branch's lending to an industry under random matching and under specialized matching, a second random counterfactual that doubles the branch's expected portfolio share in the industry. As expected, we find that p reveals less as industries are split more finely, which makes markets sparser.[14] We also find that p reveals less specialization in the broadest geographic market (France), because credit is granular: the largest 0.1% of relationships hold 30% of all bank credit to firms.[15]

We use branch closures to assess whether measured specialization signals a screening advantage. Branch closures induce a plausibly exogenous reshuffling of the match between

[12]Towns and commuting zones are the urban units and urban areas of the French statistical institute. Without specialization, the mean p of a market is $1/2$ in expectation, but it is an average over the market's relationships, so it still varies around $1/2$, and more in smaller markets. We measure this variation by computing p on an allocation of the relationships drawn by random matching and treated as data. In the French credit data of Section 3.1, 64 to 88% of the variance of the mean p across the narrower geographic markets within a broader one exceeds the variance on this allocation, against 16 to 27% across the narrower industries within a broader one. Net of market size, the standard deviation of the mean p beyond random matching is 0.057 both across the towns of a commuting zone and across commuting zones (Section 3.2.3). The two comparisons are of a different nature. The narrower geographic markets within a broader one contain different branches, and the p of each is computed against the random matching of that market. The narrower industries within a broader one contain relationships of the same branches, measured against the same random matching (Section 2.3).

[13]Net of the size of the market's cell, the standard deviation beyond random matching of a branch's p across the three-digit industries of a broad industry is 0.12 in urban units and urban areas, against about 0.06 for the averages of the location (Section 3.2.3).

[14]In the French credit data of Section 3.1, p ranks the branch under specialized matching above the same branch under random matching with probability 0.95 in the median relationship with broad industries, and 0.82 with three-digit industries, in urban units. In France as a single market, these probabilities are 0.89 and 0.77 (Section 3.2.2).

[15]Setting every relationship amount to one closes the gap between France and urban units (Section 3.2.2). The estimated tail exponent of relationship sizes is between 0.97 and 1.00 (Appendix B.3). In theory, with a power-law tail of exponent below two, the effective number of relationships received by a branch of fixed size falls to zero as the market grows (Gabaix, 2011; Proposition 3).

a firm's industry and the industry in which its lending branch specializes. We follow the borrowers of each closing branch by industry, and the explanatory variable of interest is the change in specialization experienced by each such cohort: the p of the branch that receives most of the closing branch's borrowers, in the cohort's industry, minus the p of the closing branch, Δp henceforth.[16] Comparing the cohorts of the same closing branch removes everything common to its borrowers, including the closure itself. The identifying assumption is that, among the borrowers of the same closing branch, the industries hit hardest by the shocks behind a closure are not those in which the new branch is less specialized, and the evidence corroborates it: before the closure is announced, the outcomes of the cohorts that will move to a more specialized branch evolve like those of the other cohorts of their closing branch. After the announcement, moves to a more and to a less specialized branch affect the outcomes by similar amounts per unit of p.

The closures show that specialization signals a screening advantage. When the borrowers of a closing branch move to a branch more specialized in their industry, the bank shifts its credit away from the firms that will default within a year and towards the others. With 78 two-digit industries, a cohort whose new branch is more specialized by one standard deviation of Δp within its closure receives about a fifth more new credit than the average, and less of it goes to future defaulters. The screening advantage shows in narrow markets: in every geography, down to urban units, and with industries as narrow as the 78 two-digit ones. With 241 three-digit industries, lending to future defaulters no longer falls. Firms with several lenders also shift part of their borrowing towards the reassigned bank when its new branch is more specialized in their industry.

We also ask how these conclusions change when each element of our approach is replaced by an alternative. First, measuring specialization with p rather than with indicators built from portfolio shares changes the closure results materially, in the way a reduction in measurement error would. With the top-quartile rule of Paravisini et al. (2023) or the

[16] Both p are those of the two branches' shares of lending in the cohort's industry in the quarter before the closure is announced, so the transferred accounts enter neither. Where the absorbing branch lends in the market but not to the industry, its p is that of an empty cell; where it has no lending in that quarter, we set its p to 1/2. The closing branch's share contains the cohort's own credit (Section 4.1).

favorite-industry indicator of Blickle et al. (2026), few estimates are significant across the 16 market definitions, and a researcher using them would find little evidence that specialization signals a screening advantage. Such indicators flag branches as specialized in every market, including markets in which large portfolio shares arise only because data are sparse. A change in either indicator between the closing and the absorbing branch is then a noisy measure of the change in specialization, which biases its coefficient towards zero.

A growing body of work shows that banks specialize: their portfolios tilt towards particular industries, export markets, locations or types of collateral, compared with the portfolio of their market.[17] Most of this work reads the tilt as evidence of a lending advantage. A related literature asks a simpler question: how concentrated a bank's portfolio is across sectors, measured mostly by the Herfindahl index, with no benchmark for how concentrated the portfolio would be anyway.[18] Specialization adds a benchmark, the portfolio of the market. We replace it with the distribution of the branch's portfolios under random matching in its market, which does not mistake the large shares of sparse markets for specialization.[19]

We also test what kind of lending advantage specialization signals. Doing so requires variation in the specialization of the lender that screens a given set of borrowers, since

[17]By industry: Blickle et al. (2026); Giometti and Pietrosanti (2022); De Jonghe et al. (2020); Degryse et al. (2024); Simoens and Tamburrini (2027); Casado and Martínez-Miera (2025). By export market: Paravisini et al. (2023). By location: Dursun-de Neef (2023); Casado and Martínez-Miera (2024, 2025). By type of collateral: Gopal (2021). Most compare the bank's portfolio share in the segment with the market's, directly (the excess or relative share of Blickle et al., 2026) or through segment fixed effects; Paravisini et al. (2023) rank portfolio shares across banks. Specialized lenders have lower default rates in their segments (Blickle et al., 2026; Casado and Martínez-Miera, 2025), predict defaults better (Simoens and Tamburrini, 2027), lend more to opaque firms (Blickle et al., 2026), impose less restrictive covenants (Giometti and Pietrosanti, 2022), protect credit to their sectors of specialization after a funding shock (De Jonghe et al., 2020), and, after 2008, cut credit less to borrowers whose collateral matches theirs (Gopal, 2021). Other papers measure lenders' market shares in a segment rather than the tilt of their portfolios (Niepmann and Schmidt-Eisenlohr, 2017; Saidi and Streitz, 2021).

[18]Acharya et al. (2006) and Tabak et al. (2011) relate banks' returns and risk to the Herfindahl index of their loan portfolios across sectors. Federico et al. (2025) show that banks whose portfolios are concentrated in sectors exposed to Chinese import competition cut credit to all their borrowers. Some papers measure specialization itself by concentration, across locations (Loutskina and Strahan, 2011) or industries (Di and Pattison, 2023). Part of this literature adds a benchmark: Böve et al. (2010) and Tabak et al. (2011) measure the distance of a portfolio from a benchmark portfolio, and Beck et al. (2022) the distance of a bank's sectoral exposures from those of other banks.

[19]Blickle et al. (2026, Internet Appendix I) allocate loans to banks at random, given bank and loan size, to check that random lending would not produce the specialization they measure; Gopal (2021) compares the collateral match of observed borrower–lender pairs with that of all possible pairs.

firms and lenders choose each other. Paravisini et al. (2023) come closest: they already separate lending advantages from the difficulty of replacing specialized credit. They infer the first from how firms' borrowing responds when the firms expand exports to a country, and the second from how firms' exports respond to shocks to their banks' credit supply, and they note that their methodology leaves the source of the advantage unknown. We instead use the variation in specialization induced by branch closures, which change the branch that serves a given set of borrowers within the same bank, and observe to which of them the new branch lends. This lets us test whether specialization signals a screening advantage.[20]

2 Measuring specialization in sparse markets

We ask whether a bank branch lends to an industry more than it would if the market's borrowers were assigned to its branches at random. This is a question about specialization, the tilt of the branch's portfolio towards the industry, and not about what the tilt signals: a lending advantage in the industry, market power over its firms, or neither. In Section 4 we ask whether it signals a screening advantage. In this section we explain why portfolio shares, the usual measure, cannot answer this question in sparse markets, and we construct our measure, p. We use one example throughout the paper. In Section 2.5 we show how to measure, for each relationship, how much specialization p can detect. We give the full procedure for p in Appendix A, and the details of the diagnostics in Appendix B.

The method deliberately does not classify branches as specialized or not. Any cutoff imposes one view of how much specialization matters. In sparse markets, a cutoff on portfolio shares, such as the top quartile of Paravisini et al. (2023), also picks up mostly large shares that would arise even if borrowers were assigned to branches at random (Section 2.2; we show this in French data in Sections 3.1 and 3.2.4). Instead, the method delivers two objects for every relationship. The first is p, the probability that random matching would

[20] Degryse et al. (2024) use bank mergers as a source of variation in bank specialization, with corporate innovation as the outcome.

give the branch a lower share of its lending in the firm's industry than the share it has, which has the same meaning in every market. Because p places the branch's actual share among the shares produced by random matching, we call it the *random-matching percentile* for short. The second is a set of diagnostics that say how much specialization p can reveal there, summarized by the probability that p ranks a branch that expects to lend to the industry twice what random matching would give it above the same branch under random matching. Together they leave the interpretation to the researcher or policymaker, who can apply their own threshold, their own prior on how specialized branches are, or their own tolerance for sparse markets, and see how conclusions change.

2.1 A running example

A geographic market m contains a set of bank branches. The unit of observation is a *lending relationship* $r = (f, b)$ between a firm f and a branch b of the market in a quarter, and $L_r > 0$ is the amount of credit in it (if a firm has several loans with the same branch, they form one relationship). Each firm belongs to an industry, and $j(r)$ denotes the industry of the firm in relationship r.

Our example market has three branches and ten relationships with firms in three industries, Services, Manufacturing and Real Estate. Panel A of Table 1 lists them. Branch 1 is small: it lends 2 to two firms. Branches 2 and 3 each lend 7, to four firms. Total lending in the market is 16.

2.2 Portfolio shares in sparse markets

Let V_b be the total lending of branch b in the market and A_{bj} its lending to firms in industry j. The *portfolio share* of branch b in industry j, and the *market share* of industry j, are

$$S_{bj} = \frac{A_{bj}}{V_b}, \qquad s_j = \frac{\sum_b A_{bj}}{\sum_b V_b}. \tag{1}$$

Table 1: The running example: relationships and portfolio shares (hypothetical example)

Panel A: The relationships

	Branch 1		Branch 2				Branch 3			
Relationship r	1	2	3	4	5	6	7	8	9	10
Industry	S	M	S	S	R	M	S	S	R	M
Amount L_r	1	1	2	1	3	1	1	3	1	2

Panel B: Lending and portfolio shares

	Branch 1		Branch 2		Branch 3		Market	
	Lending	Share	Lending	Share	Lending	Share	Lending	Share
Services	1	1/2	3	3/7	4	4/7	8	1/2
Manufacturing	1	1/2	1	1/7	2	2/7	4	1/4
Real Estate	0	0	3	3/7	1	1/7	4	1/4
Total	2	1	7	1	7	1	16	1
Relationships	2		4		4		10	
Lending share π_b	1/8		7/16		7/16		1	

Notes: A hypothetical market with three branches and ten lending relationships, listed in Panel A by branch (S, M, R: Services, Manufacturing, Real Estate). Panel B, Lending: A_{bj} for each branch and $\sum_b A_{bj}$ for the market. Share: portfolio share $S_{bj} = A_{bj}/V_b$ for each branch and market share s_j for the market, equation (1). $\pi_b = V_b / \sum_{b'} V_{b'}$ is the branch's share of market lending.

A branch whose portfolio share in an industry is high relative to the market share, or relative to the shares of other branches in the same market, is commonly considered specialized in that industry. We call the ratio of the two the *relative portfolio share* of branch b in industry j:

$$\frac{S_{bj}}{s_j}. \tag{2}$$

It exceeds one when the branch lends a larger fraction of its portfolio to the industry than the market as a whole does, and it is the lending counterpart of the index of revealed comparative advantage of Balassa (1965), a country's share in the exports of a commodity relative to its share in all exports. Paravisini et al. (2023) propose this ratio as the ideal index of a bank's specialization in an activity (their equation 1), but note that "the cardinal value" of a portfolio share "does not capture specialization, per se." They classify a bank as specialized when its portfolio share in an activity is in the top quartile of the distribution of all banks' shares in that activity. Panel B of Table 1 reports these shares for the example.

Branch 3 lends 4/7 of its portfolio to Services, against 1/2 for the market as a whole, a relative portfolio share of 8/7; Branch 1 lends half of its portfolio to Manufacturing, twice the market share of 1/4 (a relative portfolio share of 2); and Branch 2 lends 3/7 of its portfolio to Real Estate, against 1/4 (a relative portfolio share of 12/7).

Portfolio shares, and the relative portfolio share built from them, have two flaws as measures of specialization. *First, sparse markets produce large and uneven shares without specialization.* A market is sparse when it contains few relationships relative to the number of branch–industry cells they are spread over. Branch 1's relative portfolio share of 2 in Manufacturing comes from a single loan of 1. With two relationships, its share in a given industry can only be 0, 1/2 or 1, so its relative portfolio share in Manufacturing can only be 0, 2 or 4: a large deviation from one is the rule, not the exception, whatever the branch's lending policy. Finer industry definitions make this worse, because they spread the same relationships over more cells: splitting an industry in two leaves fewer relationships in each part, and a single loan then weighs more in each branch's share. Finer partitions of the market therefore mechanically produce more apparent specialization. The problem is the one described by Armenter and Koren (2014) in trade data: when observations are few relative to the categories they are spread over, random assignment alone reproduces patterns such as the prevalence of zero trade flows, which then say little about the economics behind them.

Second, neither shares nor relative portfolio shares are comparable across markets. A share tells us how a branch ranks among the branches of its market, but not how unusual that rank is. The relative portfolio share seems to do better: it is centered at one, so it appears to measure, on a common scale, how much more a branch lends to an industry than the market as a whole. The impression is false, for two reasons. First, the range of the relative portfolio share depends on the size of the industry: since a share cannot exceed one, $S_{bj}/s_j \leq 1/s_j$. Branch 1's relative portfolio share of 2 in Manufacturing, an industry with a market share of 1/4, is half of the maximum of 4. The same relative portfolio share in Services, with a market share of 1/2, would mean that the branch lends only to Services. Second, the same relative portfolio share carries more or less information

depending on how many relationships it rests on. To see this, consider the *h-replica* of the example: the market in which every relationship of Table 1 appears h times, held by the same branch. All portfolio shares, market shares and relative portfolio shares are the same as in the original market. Yet Branch 3's relative portfolio share of 8/7 in Services means much more in the 10-replica, where it rests on 40 relationships, than in the original market, where it rests on four. Neither shares nor the relative portfolio share can tell the two apart, so neither can be compared across markets of different size, across industries of different size, or across market definitions.

2.3 The counterfactual: random matching

The measure we propose compares a branch's portfolio share with its shares when firms are matched to branches at random. We treat relationships as balls and the branches of the market as bins.

Random matching. In the counterfactual, every relationship of the market is assigned to a branch at random, independently of the others. It keeps its amount L_r and its industry. The assignment probabilities hold the structure of the market fixed but ignore the borrower's industry. In the simplest version, relationship r goes to branch b with probability equal to the branch's share of market lending,

$$\pi_b = \frac{V_b}{\sum_{b'} V_{b'}}, \tag{3}$$

so that each branch receives, on average, as much lending as it actually does. In the example, π_b is 1/8 for Branch 1 and 7/16 for Branches 2 and 3 (Table 1). Because the assignment ignores industries, a branch's counterfactual portfolio share in an industry is, on average, close to the market share of that industry, and its dispersion around that value is produced by random matching alone. The logic is that of the "dartboard" approach of Ellison and Glaeser (1997), who measure the geographic concentration of an industry against the concentration that would arise if its plants chose their locations at random, so

that their index is zero in expectation without agglomeration forces and comparable across industries with different plant sizes and across levels of geographic aggregation.

Geography and industry. Geography and industry enter the counterfactual differently. A geographic market is defined by the location of the branch: it contains the branches located in the geographic unit and all their relationships, irrespective of the location of the borrowers (Appendix A, Step 2). Random matching reassigns every relationship of the market to one of these branches, and each relationship retains its industry, so that the relationships of all industries are reassigned jointly. We adopt this definition because the object we measure is the specialization of a branch, which is a property of the branch's entire portfolio.

We assume that a lending advantage in an industry is a property of the branch: it applies to every firm of the industry that borrows from the branch, wherever the firm is located. The closures of Section 4 also change the branch, not the location of the firm. Defining markets by the location of the branch keeps each branch's entire portfolio within a single market, so that each branch has a single value of p in each industry, measured against the portfolios of the other branches located in the same geographic unit. Industries define the measured share rather than the boundary of the counterfactual, because a branch's share of lending in one industry depends on its lending to all industries.

This definition has two implications. First, under random matching, a firm's alternatives are the branches located in the same geographic unit as its lender, which need not be the branches closest to the firm. Second, finer partitions of the two dimensions change the counterfactual differently. A finer geographic partition reduces the set of branches among which relationships are reassigned and leaves each branch's portfolio unchanged. A finer industry partition leaves the set of branches and the reassignment unchanged and divides each branch's portfolio into more industry cells. A researcher with a different question could consider a different definition of the counterfactual. For example, a study of the alternatives available to firms could define geographic markets by the location of the borrower, and reassign the borrowers located in each geographic unit among the branches that

lend to them, holding fixed each branch's lending to these borrowers.

In the data, large branches tend to hold large relationships. Our counterfactual reproduces this pattern, which we call *size matching*: the probability that relationship r is assigned to branch b, $\pi_b(r)$, also depends on how often branches of b's size hold relationships of r's size. Measured within the market itself, this pattern leaves the structure of the market intact: each branch still receives, on average, as much lending as it actually does. In sparse markets, however, the largest branch is often the only branch of its size, and the counterfactual would then return to it the relationships it actually holds. We therefore measure the pattern in b's *département*, one of the 96 administrative divisions of metropolitan France, which pools a median of 152 branches, so that branches of every size have many counterparts (Appendix A, Steps 4 and 5).[21] The cost is that, in markets other than départements, a branch's expected lending under random matching is no longer exactly its actual lending. In Appendix A we measure this distortion and describe how it could be removed. We leave size matching out of the example, where every relationship is assigned with (3).

Leaving one out. We measure specialization separately for each relationship r of branch b, in the industry $j = j(r)$ of its borrower, and leave r itself out of the branch's portfolio. The observed *leave-one-out* share is

$$S_{bj}^{-r} = \frac{A_{bj} - L_r}{V_b - L_r}, \tag{4}$$

the share of industry j in the branch's lending to its *other* borrowers. The counterfactual share, which we denote $\tilde{S}_{bj}^{-r}$, is computed in the same way: it is branch b's share in industry j among the relationships other than r that random matching assigns to it.

Leaving r out serves two purposes. First, it is required for the benchmark of p. If the observed share included r, the branch would be compared with counterfactual portfolios

[21] Of the geographic partitions we use (Section 3.1), the départements are the finest that cover the whole territory, since urban units and urban areas (built-up areas and commuting zones) leave out the municipalities outside them.

that contain r only with probability $\pi_b(r)$: relationship r would raise the observed share with certainty and the counterfactual share only in some draws, so that the observed share of a branch that is not specialized would exceed its counterfactual share in distribution, and the percentile p defined in Section 2.4 would exceed 1/2 on average.[22] The distortion is largest in sparse markets, where a single relationship can make up a large part of a branch's lending in an industry. Second, leaving r out removes a mechanical correlation between the measure and the firm's own borrowing. Without it, a larger loan of the firm would raise the measured specialization of its own lender, and any regression of the firm's outcomes on that specialization would be biased by construction. Paravisini et al. (2023) leave the firm out of its bank's share when they relate the firm's borrowing to its bank's specialization (their equation 10), to avoid a spurious correlation between the two. In the example, the leave-one-out share of Branch 3 in Services is $3/6 = 1/2$ for its relationship of amount 1 (relationship 7) and $1/4$ for its relationship of amount 3 (relationship 8): in each case, the other Services loan is all that remains of the branch's Services lending.

A counterfactual share is defined only when random matching assigns branch b at least one relationship other than r, that is, when $\tilde{V}_b^{-r} > 0$, where $\tilde{V}_b^{-r}$ is the branch's counterfactual lending excluding r. The same holds for the observed share, which requires the branch to have borrowers other than firm f. Draws in which the branch receives other relationships, but none in industry j, are kept: its share in industry j is then zero, and the observed share can be zero too.

2.4 The random-matching percentile p

The specialization measure of relationship r is the probability that random matching gives branch b a lower share in industry $j = j(r)$ than the share it actually has, where an equal

[22] For a branch that is not specialized, the observed share including r is distributed as the counterfactual share given that r is assigned to the branch. Adding a relationship of industry j raises the share, since $(A + L_r)/(V + L_r) \geq A/V$, so in every draw the share with r is at least the share without it. The observed share therefore first-order stochastically dominates the counterfactual share, which mixes draws with and without r, and the expected value of p computed without leaving r out is at least 1/2, strictly so when $\pi_b(r) < 1$ and adding r changes the share.

share counts one half:

$$p_r = \Pr\left(\tilde{S}_{bj}^{-r} < S_{bj}^{-r} \;\middle|\; \tilde{V}_b^{-r} > 0\right) + \frac{1}{2}\Pr\left(\tilde{S}_{bj}^{-r} = S_{bj}^{-r} \;\middle|\; \tilde{V}_b^{-r} > 0\right). \tag{5}$$

In other words, p_r is the percentile of the branch's observed leave-one-out share within its counterfactual distribution, with ties split evenly between the draws below and above: its random-matching percentile. Since the observed and the counterfactual shares are compared with each other, dividing both by the same market share leaves p unchanged: p is also the percentile of the branch's relative portfolio share (computed, like the share, without relationship r) within the distribution of the relative portfolio share under random matching. Unlike the relative portfolio share itself, this percentile has the same meaning in every market. In the data we estimate these probabilities with 1,000 random draws (Appendix A, Steps 6–8). In the example, all amounts are integers, and we compute them exactly.

Two features of definition (5) adjust p for sparsity. The first is leaving relationship r out of both the observed and the counterfactual shares. Without it, a branch that is not specialized would have an expected p of at least 1/2, and more so the larger the loan relative to the branch's lending in the industry, that is, the sparser the cell (we prove this in Section 2.3, and measure the size of the bias in French data in Section 3.2.4). The second is the treatment of ties. In a sparse cell many draws give the branch the same share, often zero. Keeping the draws with a zero share and counting ties as one half give p a common benchmark. For a branch that is not specialized, the observed share is just one more outcome of random matching, equally likely to fall above or below any counterfactual draw, so the expected value of p is exactly 1/2, in every market and at every industry level. Without them, the benchmark falls well below 1/2 in sparse markets, and falls further the sparser the market (Appendix A, after Step 8). The benchmark is exact given the assignment probabilities. In the data, these are estimated from the same relationships, through the branches' lending volumes and, with size matching, the matching rates, and this can move the benchmark slightly: in simulations on the structure of French

markets, by a few hundredths for the largest branch of each market (Appendix A, footnote to Step 4).

Figure 1 shows the counterfactual distribution for relationship 7, Branch 3's Services loan of amount 1. In the original market (left panel), Branch 3 can expect to receive only a few of the other Services relationships, so its counterfactual share takes only about forty distinct values, spread from 0 to 1. Random matching gives it a share below its observed leave-one-out share of 1/2 (3 of its 6 units of lending once relationship 7 is left out) in 49% of the draws (including 9.5% in which it receives no other Services relationship), exactly 1/2 in 13%, and above 1/2 in 37%, so $p_7 = 0.49 + 0.13/2 = 0.56$. In the 10-replica (right panel), the same branch has the same portfolio shares, but the counterfactual distribution is concentrated around the market share. The observed leave-one-out share is now $39/69 = 0.57$: leaving out a relationship of amount 1 lowers the share less when the branch lends ten times more (with relationship 7, the share is 4/7 in both markets). It exceeds the counterfactual share in most draws: $p_7 = 0.86$.

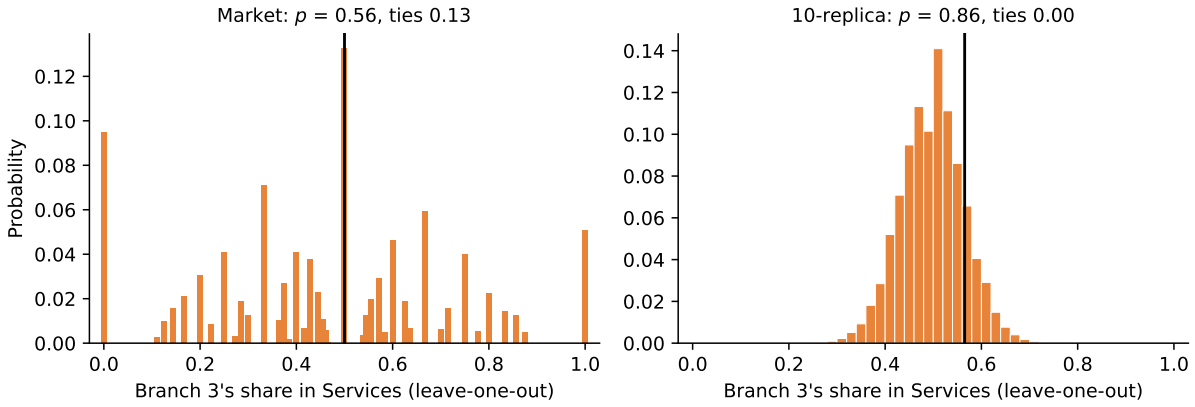


Figure 1: Counterfactual distribution of a portfolio share in the running example (hypothetical example)

Notes: Relationship 7 of Table 2 (Branch 3, Services, amount 1). Exact distribution, under random matching with probabilities (3), of Branch 3's leave-one-out share in Services, over the draws in which Branch 3 receives at least one relationship other than relationship 7 (zero shares included). Left: the market of Table 1. Right: its 10-replica (every relationship appears ten times), shown as a histogram with bins of width 0.02. The vertical line is the observed leave-one-out share. "Ties": probability that the counterfactual share equals the observed share.

Table 2 reports p for every relationship, in the original market and in its 10-replica. Four properties of p stand out.

Table 2: The measure p in the running example (hypothetical example)

					Market				10-replica	
r	Branch	Industry	L_r	S_{bj}/s_j	S_{bj}^{-r}	$P(A>0)$	Ties	p_r	S_{bj}^{-r}	p_r
1	1	Services	1	1	0.00	0.59	0.41	0.20	0.47	0.44
2	1	Manufacturing	1	2	0.00	0.34	0.66	0.33	0.47	0.95
3	2	Services	2	6/7	0.20	0.90	0.05	0.19	0.41	0.10
4	2	Services	1	6/7	0.33	0.90	0.07	0.29	0.42	0.12
5	2	Real Estate	3	12/7	0.00	0.44	0.56	0.28	0.40	1.00
6	2	Manufacturing	1	4/7	0.00	0.69	0.31	0.16	0.13	0.01
7	3	Services	1	8/7	0.50	0.90	0.13	0.56	0.57	0.86
8	3	Services	3	8/7	0.25	0.90	0.07	0.26	0.55	0.83
9	3	Real Estate	1	4/7	0.00	0.44	0.56	0.28	0.13	0.03
10	3	Manufacturing	2	8/7	0.00	0.69	0.31	0.16	0.26	0.69

Notes: One row per relationship r of the market of Table 1. L_r: amount. S_{bj}/s_j: relative portfolio share of the branch in the industry of the relationship, equation (2), over its whole portfolio; it is the same in the market and in every replica. S_{bj}^{-r}: observed leave-one-out share of the branch in the industry of the relationship, equation (4). $P(A>0)$: probability that random matching gives the branch positive lending to that industry from borrowers other than r, among the draws in which it receives some relationship other than r. Ties: probability, in the same draws, that the counterfactual share equals the observed share. p_r: equation (5), computed exactly. 10-replica: the same market with every relationship appearing ten times; values are the same for all ten copies.

p is cardinal. It is a probability, measured on the same scale in every market, with the same benchmark, $p = 0.5$, for a branch that is not specialized: a value of 0.86 means that random matching produces a lower share in 86% of cases (counting ties as one half), whatever the size of the market, the number of branches, or the industry definition. It can therefore be compared across markets and across partitions of the same data, which portfolio shares and the relative portfolio share cannot. What is common across markets is this probability, not the degree of specialization reflected by a given value: how far p moves away from 1/2 for a given degree of specialization depends on how much information the market contains, as the next property shows (we measure it in Section 2.5).

p accounts for how much information the market contains. The portfolio shares and the relative portfolio share of the original market and its replicas are the same, but p is not. In the original market, every value of p lies between 0.16 and 0.56: with so few relationships, no portfolio share is far from what random matching could produce. For relationship 7 (Branch 3 in Services, with a relative portfolio share of 8/7 in every replica), p rises from 0.56 in the original market to 0.64, 0.76, 0.86 and 0.94 in the 2-,

5-, 10- and 20-replicas. As the market becomes thicker, the shares produced by random matching concentrate around the share it gives the branch on average, so a given deviation from that share becomes ever less likely under random matching, and p tends to one or to zero according to the sign of the deviation. Let μ_{bj} denote this benchmark, the branch's expected lending to industry j under random matching divided by its expected total lending, $\mu_{bj} = \sum_{r:\, j(r)=j} \pi_b(r) L_r \Big/ \sum_r \pi_b(r) L_r$, where the sums run over all relationships of the market. Without size matching, as in the example, μ_{bj} is the market share s_j, and the deviation has the sign of $S_{bj}/s_j - 1$.

Proposition 1 (Thick markets)**.** *Consider the h-replicas of a market, with the same assignment probabilities $\pi_b(r)$ in every replica, and a relationship r of branch b in industry j. As $h \to \infty$, $p_r \to 1$ if $S_{bj} > \mu_{bj}$, and $p_r \to 0$ if $S_{bj} < \mu_{bj}$.*

The proof (Appendix B.5) rests on two facts. By the law of large numbers, the branch's counterfactual share concentrates around μ_{bj}, while its observed leave-one-out share converges to S_{bj}, which replication leaves unchanged. In the 10-replica, Branch 3 in Services, Branch 1 in Manufacturing and Branch 2 in Real Estate, whose relative portfolio shares exceed one (8/7, 2 and 12/7), have p of 0.86, 0.95 and 1.00, whereas Branch 2 in Services and in Manufacturing and Branch 3 in Real Estate, whose relative portfolio shares are below one (6/7, 4/7 and 4/7), have p of 0.10–0.12, 0.01 and 0.03. The convergence need not be monotone. Branch 3's relative portfolio share in Manufacturing is 8/7, yet p_{10} is 0.16 in the original market, below 0.5, before rising to 0.69 in the 10-replica and 0.93 in the 50-replica (Table 3). Proposition 1 says nothing about the knife-edge case $S_{bj} = \mu_{bj}$, such as Branch 1 in Services, whose relative portfolio share is exactly one, where p_1 approaches 0.5 in the example (0.47 in the 50-replica).

p is specific to the branch and to the relationship. The counterfactual distribution depends on the branch's lending volume and, with size matching, on the sizes of the relationships it can receive. A given relative portfolio share can be unremarkable for a small branch, whose counterfactual shares are widely dispersed, and exceptional for a large one. Because of the leave-one-out construction, two relationships of the same branch and

industry can also have different values of p when their amounts differ (0.56 and 0.26 for relationships 7 and 8 in the original market); the difference vanishes as the market grows (0.86 and 0.83 in the 10-replica).

p does not mistake large shares in sparse cells for specialization. A large relative portfolio share that rests on a single relationship does not raise p. In the original market, Branch 1's relative portfolio share in Manufacturing is 2, but its only Manufacturing relationship is relationship 2 itself, so its leave-one-out share is zero. Random matching would also leave Branch 1 without other Manufacturing borrowers in 66% of the draws. Having none is therefore unremarkable, and $p_2 = 0.66/2 = 0.33$: the relative portfolio share suggests strong specialization, p does not. Six of the ten relationships of the original market have a leave-one-out share of zero. For each of them, p is half the probability that random matching also gives the branch no lending in the industry: between 0.16 and 0.33. It would be close to zero only if random matching almost always gave the branch some lending in the industry. Finer industry definitions make such cases more common, since they leave more branches with no other borrower in the firm's industry, both in the data and in the counterfactual. In the 10-replica, Branch 1's relative portfolio share of 2 in Manufacturing rests on ten relationships, and $p_2 = 0.95$.

More generally, when the market is sparse, the counterfactual distribution is lumpy, has many ties, or puts much of its mass at zero (the columns Ties and $P(A > 0)$ of Table 2), and these features pull p towards 1/2. Let τ_r be the probability of a tie, the probability that the counterfactual share equals the observed one. By (5), p_r is the probability of a lower counterfactual share plus $\tau_r/2$. The probability of a lower share is at least zero and at most $1 - \tau_r$, so p_r lies between $\tau_r/2$ and $1 - \tau_r/2$:

$$\left| p_r - \tfrac{1}{2} \right| \leq \tfrac{1}{2} (1 - \tau_r) . \tag{6}$$

The more likely random matching is to reproduce the observed share exactly, the closer p is to 1/2, whatever the branch's lending policy. A sparse market can also produce values of p far from 1/2, however: when the counterfactual share is zero in most draws but the

branch does have another borrower in the industry, p is close to one. In Section 2.5 we show that in sparse markets such values are rare even for a specialized branch, and we turn these observations into diagnostics.

Readers familiar with hypothesis tests will recognize the construction. Consider the hypothesis that branch b is matched to firms at random, and the alternative that it lends more to industry j than random matching would make it lend. Random matching gives the distribution of the leave-one-out share under the hypothesis, and the one-sided mid-p-value of the observed share, the mid-probability of Lancaster (1961), the probability of a higher counterfactual share plus half the probability of an equal one, is $1-p_r$. We use p differently from a test in three ways. First, the direction: a high p, not a low one, signals specialization, and a low p signals that the branch lends less to the industry than random matching would make it lend. Second, we do not test. No significance level separates specialized from non-specialized branches: we read p over its whole range against its benchmark of $1/2$, and we compare the distribution of p with its distribution under random matching rather than classify relationships one by one, so the large number of relationships raises no multiple-testing problem. Third, random matching is not a hypothesis that we expect to hold, but a benchmark without specialization, against which we measure specialization, as the dartboard of Ellison and Glaeser (1997) is for the geographic concentration of industries. The parallel also points to the limitation of p. A p-value mixes the size of an effect with the information in the data: a value in the middle of its range can mean that there is no effect, or that the data cannot reveal one. The same holds for p. In Section 2.5 we measure how much specialization p can reveal, the analogue of the power of a test: the AUC we define there, the expected value of p under specialized matching, a second random counterfactual in which the branch is specialized, is one minus the expected mid-p-value of the test under that alternative. It is also the power of the test averaged over all significance levels: the test that flags specialization at level α when $p_r > 1-\alpha$ detects it under specialized matching with probability $\Pr(p_r > 1-\alpha)$, and since $E[X] = \int_0^1 \Pr(X > t)\, dt$ for any X in $[0,1]$, the average of this probability over $\alpha \in [0,1]$ is the expected value of p_r for that branch. The AUC curves of Section 2.5 are therefore power curves, averaged over significance levels.

Sackrowitz and Samuel-Cahn (1999) propose the expected p-value under the alternative as a measure of the performance of a test that does not depend on a significance level, and show that, for continuous test statistics, it equals one minus the power averaged over a uniformly distributed level.

2.5 When is p informative?

In sparse markets the pull of p towards 1/2 (equation (6)) keeps the large shares also produced by random matching from passing for specialization, but it acts whether or not the branch is specialized, so a value close to 1/2 is ambiguous (Section 2.4): it can mean that the branch is not specialized, or that the cell is too sparse for specialization to show. We need, for each relationship, a measure of how much specialization p could have revealed. We build it as a power calculation that compares two random counterfactuals for the same branch, one without specialization and one with: how well p separates them, measured by the area under the ROC curve (AUC), the standard summary of how well a variable separates two groups. The details are in Appendix B; here we describe the approach and apply it to the running example.

Two counterfactuals. The first counterfactual is random matching, which contains no specialization. The second, which we call *specialized matching*, makes the branch specialized in industry j to a known degree. Relationships are still assigned at random, but the branch receives each relationship of industry j with a higher probability than under random matching, and each relationship of other industries with a lower probability, in such a way that its expected total lending is unchanged and its expected lending to industry j is $\kappa \geq 1$ times what random matching gives it (Appendix B, equation (B.3)). Specialization therefore changes the composition of the branch's portfolio, not its size. Because expected total lending is unchanged, κ is also a ratio of two shares of expected lending: the share of industry j in the branch's expected lending under specialized matching, over the same share under random matching, which is the benchmark μ_{bj} of Section 2.4 computed without relationship r. κ thus measures specialization against the random-matching benchmark,

not against the market share, and with expected rather than realized lending. In the example, where every relationship is assigned with the same probability (3), the benchmark is the market share of industry j among the other relationships, so κ is the relative portfolio share, equation (2), that the branch can expect. With size matching, the benchmark differs from the market share. At $\kappa = 1$ specialized matching is random matching. κ cannot exceed a maximum, $\kappa_{\max}$, reached when the branch receives every relationship of industry j or lends only to industry j, which we call full specialization (equation (B.4)). For Branch 3 in Services, the other relationships of the market lend 7 to Services out of 15, so the branch's relative portfolio share can be at most 15/7, the bound $1/s_j$ of Section 2.2 computed without relationship 7: $\kappa_{\max} = 2.14$. $\kappa_{\max}$ is small in large industries, and large for small branches in small industries (5.00 for Branch 1 in Manufacturing).

The AUC of p. Under specialized matching of degree κ, the observed share is a draw from the distribution of the branch's shares under that matching, and p is computed from it as in (5), against random matching. Write S^{κ}_{bj} for this leave-one-out share, drawn independently of the random-matching share $\tilde{S}^{-r}_{bj}$, both given that the branch receives some relationship other than r. By definition (5), the expected value of p is

$$\text{AUC}(\kappa) = E_{\kappa}\left[p_r\right] = \Pr\left(\tilde{S}^{-r}_{bj} < S^{\kappa}_{bj}\right) + \tfrac{1}{2}\Pr\left(\tilde{S}^{-r}_{bj} = S^{\kappa}_{bj}\right), \tag{7}$$

the probability that the branch has a higher share under specialized matching of degree κ than under random matching, with ties counted one half. This is the area under the receiver operating characteristic (ROC) curve, or AUC, of p used to tell the two counterfactuals apart.[23] At $\kappa = 1$ the two counterfactuals coincide and $\text{AUC}(1) = 1/2$, the benchmark of Section 2.4: the two shares have the same distribution, and p tells them apart no better

[23] Classify a branch as specialized when its p exceeds a threshold c. The true positive rate is the probability that the branch is classified as specialized under specialized matching, $\Pr(p > c)$ with its share drawn as S^{κ}_{bj}, and the false positive rate is the same probability under random matching, with its share drawn as $\tilde{S}^{-r}_{bj}$. As c falls from one to zero, the pair (false positive rate, true positive rate) traces the ROC curve from $(0, 0)$ to $(1, 1)$. Since p is an increasing function of the share over the possible values of both shares, the area under this curve, with the steps at tied values joined by straight segments, is the probability that the first share exceeds the second, ties counted one half: the right-hand side of (7).

than a coin toss. The AUC increases with κ, and its value at full specialization, the *ceiling* AUC($\kappa_{\max}$), is the highest expected p that the relationship can reach. In a thick market, the AUC rises quickly from 1/2 towards one: even mild specialization shows up in p. In a sparse market, it rises slowly, and its ceiling may remain far from one.

Summarizing the AUC. We summarize the AUC by its value at $\kappa = 2$, AUC(2): the probability that p ranks the branch under specialized matching with $\kappa = 2$, where it expects to lend to industry j twice what random matching would give it, above the same branch under random matching. Fixing the degree of specialization makes the statistic comparable across cells: the closer AUC(2) is to one, the better p separates the two counterfactuals, and at 1/2 it does not separate them at all. A doubling is a round degree of specialization, but how much of the room for specialization it uses depends on the cell, so we report it together with $2/\kappa_{\max}$. Because κ is a ratio of expected shares, $2/\kappa_{\max}$ is the expected share of the branch under specialized matching with $\kappa = 2$, as a fraction of the largest share it could have, when it lends only to industry j or receives every relationship of the industry. When $2/\kappa_{\max}$ is close to one, as for a branch that accounts for a large part of its market's lending or in an industry that accounts for a large part of the branch's expected lending, a doubling is nearly full specialization. When it is small, as in a small industry, a doubling uses little of the room for specialization. When $2/\kappa_{\max}$ exceeds one, no branch can lend twice its random-matching share and AUC(2) is not defined; the ceiling AUC($\kappa_{\max}$) then says how much specialization p could reveal at most.

In the example, these quantities can be computed exactly. In the data, they cost much more to compute than p itself. To compute p, one set of random reassignments of the market serves every relationship: each draw gives the counterfactual portfolio of every branch in every industry at once, and leaving a relationship out only subtracts it from its branch's portfolio. The diagnostics instead ask a hypothetical question about one cell: how high p would be if this branch were specialized to degree κ in this industry. Each cell has its own specialized matching, so nothing computed for one cell serves another. We compute the diagnostics by simulation for samples of relationships, and for every relationship with

an approximation for screening that is validated against the simulation (Appendix B) and that we use in Section 3.2.2. In Appendix B.3 we report what each computation costs in our data.

Why sparse cells reveal little. Write P_0 for the probability that random matching gives branch b positive lending to industry j from borrowers other than the firm in r, among the draws in which the branch receives some relationship other than r (the column $P(A > 0)$ of Table 2), and P_κ for the same probability under specialized matching of degree κ. With probability $1-P_\kappa$, specialized matching gives the branch no other borrower in industry j. Its leave-one-out share is then zero, and $p_r = \frac{1}{2}(1 - P_0)$ whatever its degree of specialization. The value of p can therefore separate the two counterfactuals only in the draws in which the branch has another borrower in the industry, and $\mathrm{AUC}(\kappa) \leq \frac{1}{2}(1 + P_\kappa)$ (Proposition 2, Appendix B.1). An AUC of 0.9, for instance, requires $P_\kappa \geq 0.8$, which a sparse cell reaches only when κ is large. Sparse markets thus keep the expected value of p near 1/2, not each of its values: when P_0 is small and the branch does have another borrower in the industry, p is close to one. In the running example, Branch 1 can expect only 0.25 other Manufacturing borrowers under random matching (relationship 2). Under specialized matching with $\kappa = 2$, it has another Manufacturing borrower with probability $P_2 = 0.61$, so AUC(2) is at most 0.81; its exact value is 0.65.

The running example. Figure 2 shows the AUC for two relationships, and Table 3 reports the diagnostics for all ten. In the original market, AUC(2) ranges from 0.65 to 0.95, but for nine of the ten relationships a doubling is close to full specialization ($2/\kappa_{\max}$ of 0.88 or 0.93): Branches 2 and 3 each account for a large part of the market's lending, and Services for half of it, so under specialized matching with $\kappa = 2$ the branch would lend almost only to the industry or receive almost all of it. A high AUC(2) then shows only that p would detect near-complete specialization. Branch 2's relative portfolio share in Real Estate, for instance, is 12/7, but its AUC(2) is 0.74 and its ceiling 0.81 (Figure 2, right panel). Branch 1 in Manufacturing is the exception: a small branch in a small industry, it

could be up to five times as specialized as random matching makes it ($2/\kappa_{\max} = 0.40$), but p reveals a doubling poorly (AUC(2) = 0.65). As the market becomes thicker, the AUC rises more steeply: in the 10-replica AUC(2) is 0.90 for relationship 2 and above 0.99 for every other relationship, and in the 50-replica it is above 0.99 for all.

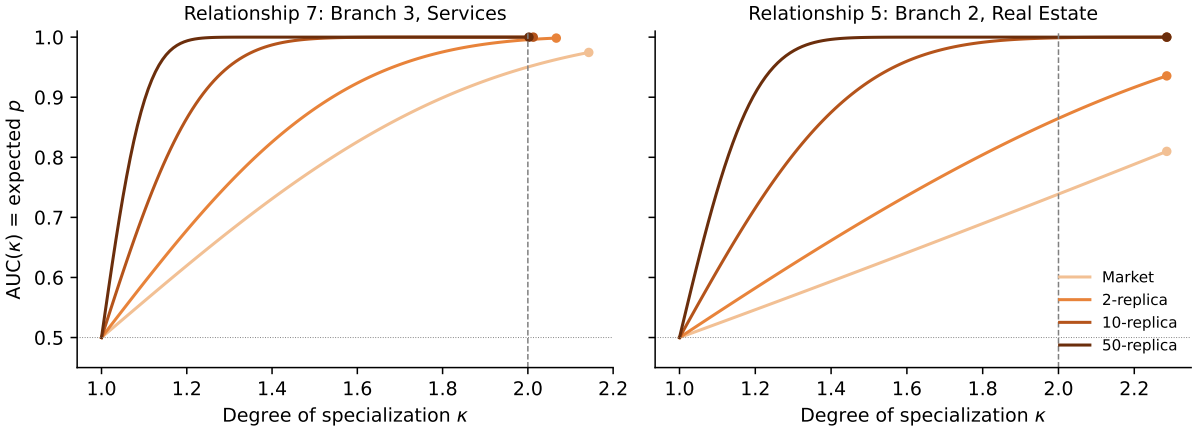


Figure 2: AUC of p as a function of specialization in the running example (hypothetical example)

Notes: AUC(κ), the expected value of p_r (equation (7)), under specialized matching of degree κ, from $\kappa = 1$ (random matching) to $\kappa_{\max}$ (dot). Left: relationship 7 of Table 2 (Branch 3, Services). Right: relationship 5 (Branch 2, Real Estate). One line for the market of Table 1 and one for each of its 2-, 10- and 50-replicas. The dashed vertical line marks $\kappa = 2$, where the height of each line is AUC(2). Exact computation, for random matching with probabilities (3) and the corresponding specialized matching.

The diagnostics give the reading of p that its value alone could not. Relationship 7 has $p = 0.56$ in the original market. Its AUC(2) is 0.95: a branch that expects to lend to Services twice what random matching gives it would, on average, have p of 0.95. But $2/\kappa_{\max} = 0.93$: in this market, such a branch lends almost only to Services. A value of 0.56 therefore rules out near-complete specialization, but says little about milder specialization: at $\kappa = 1.5$ the AUC is only 0.78 (Figure 2, left panel). In the 50-replica, relationship 1 (Branch 1, Services) has a similar value, $p = 0.47$, but there the AUC exceeds 0.99 already at $\kappa = 1.5$: a branch that expects to lend to Services 50% more than random matching gives it would almost always have a higher p, so a value of 0.47 is evidence that Branch 1 is not specialized in Services, consistent with its relative portfolio share of exactly one. The same branch has $p = 0.20$ in the original market, where its AUC(2) is 0.81.

The diagnostics are also specific to the branch. In the original market, Branches 1

Table 3: Sparsity diagnostics in the running example (hypothetical example)

			Market				10-replica		50-replica	
r	b	Industry	p_r	$2/\kappa_{\max}$	AUC(2)	AUC($\kappa_{\max}$)	p_r	AUC(2)	p_r	AUC(2)
1	1	S	0.20	0.93	0.81	0.85	0.44	1.00	0.47	1.00
2	1	M	0.33	0.40	0.65	0.93	0.95	0.90	1.00	1.00
3	2	S	0.19	0.88	0.92	0.97	0.10	1.00	0.01	1.00
4	2	S	0.29	0.93	0.95	0.97	0.12	1.00	0.01	1.00
5	2	R	0.28	0.88	0.74	0.81	1.00	1.00	1.00	1.00
6	2	M	0.16	0.88	0.82	0.88	0.01	1.00	0.00	1.00
7	3	S	0.56	0.93	0.95	0.97	0.86	1.00	0.99	1.00
8	3	S	0.26	0.88	0.90	0.95	0.83	1.00	0.99	1.00
9	3	R	0.28	0.88	0.78	0.87	0.03	1.00	0.00	1.00
10	3	M	0.16	0.88	0.80	0.86	0.69	1.00	0.93	1.00

Notes: One row per relationship r of Table 2; b: branch; S, M, R: Services, Manufacturing, Real Estate. p_r: equation (5). $\kappa_{\max}$: maximum degree of specialization; $2/\kappa_{\max}$: expected share of the branch under specialized matching with $\kappa = 2$, as a fraction of the largest share it could have. AUC(2): expected p_r under specialized matching with $\kappa = 2$, equation (7). AUC($\kappa_{\max}$): expected p_r under specialized matching at $\kappa_{\max}$ (the ceiling). 10- and 50-replica: the same market with every relationship appearing 10 and 50 times; values are the same for all copies. $\kappa_{\max}$ is the same as in the market or lower: in the replicas it lies between 2.00 and 2.29 for every relationship except relationship 2 (4.08 in the 10-replica and 4.02 in the 50-replica, against 5.00 in the market). In this example, $\kappa_{\max}$ is the smaller of $1/\pi_b$ and $1/s_j$ computed without relationship r. Leaving out one of k copies of a relationship matters less as k grows, so the second bound falls towards $1/s_j$ (2 for Services, 4 for Manufacturing and Real Estate), and $\kappa_{\max}$ falls where this bound binds; it stays at $1/\pi_b = 2.29$ for relationships 5, 6, 9 and 10. Exact computation.

and 3 both lend to Services, and for both a doubling is close to full specialization ($2/\kappa_{\max} = 0.93$), but AUC(2) is 0.81 for the small Branch 1 (relationship 1) and 0.95 for Branch 3 (relationship 7): with fewer relationships, the small branch's counterfactual share is more dispersed, and the same degree of specialization stands out less.

Why a percentile. Specialized matching defines a degree of specialization, κ, and the data offer a direct estimator of it. Let A_{bj}^{-r} be the observed lending of branch b to industry j from borrowers other than the firm in relationship r, and let $\bar{A}_{bj}^{-r} = \sum_{r'} \pi_b(r') L_{r'}$, where the sum runs over the relationships of industry j in the market other than r, be its expected value under random matching. The ratio

$$\hat{\kappa}_r = \frac{A_{bj}^{-r}}{\bar{A}_{bj}^{-r}} \tag{8}$$

has expected value κ under specialized matching of degree κ, which multiplies the branch's expected lending to industry j by κ (Appendix B, equation (B.3)), and expected value one under random matching, in every market. It measures specialization by the excess of observed lending over its expectation under random matching, as the index of Ellison and Glaeser (1997) measures agglomeration by the excess of observed concentration over its expectation under random location. For a given relationship, $\hat{\kappa}_r$ and p_r are both increasing functions of the branch's observed lending to industry j, given its lending to other industries. They contain the same information about the branch, and differ in the scale on which they place it: $\hat{\kappa}_r$ divides the observed lending by its expected value, whereas p_r locates the observed share in the whole distribution produced by random matching for that branch. The two scales differ across branches and markets because the sampling noise does. Under random matching, the variance of $\hat{\kappa}_r$ is

$$\frac{\sum_{r'} \pi_b(r') \Big(1 - \pi_b(r')\Big) L_{r'}^2}{\Big(\sum_{r'} \pi_b(r') L_{r'}\Big)^2},$$

with the sums over the same relationships. It is at most the inverse of the effective number of relationships of Appendix B (equation (B.13)), and it is large when the branch's lending in industry j under random matching rests on few relationships, or on a few large ones. A value of $\hat{\kappa}_r$ well above one is then common without specialization when this variance is large, and rare when it is small. Read at face value, $\hat{\kappa}_r$ thus has the flaw of the relative portfolio share (Section 2.2): in sparse markets, large values arise without specialization. Correcting for this requires a description of the sampling noise of $\hat{\kappa}_r$, and when relationship sizes are heavy-tailed its variance is a poor one: much of the variance comes from a few large relationships, seldom assigned to the branch by random matching.[24] The percentile p_r instead reads the branch's share against the whole distribution produced by random matching, which is why its benchmark, and its distribution without specialization, are the same in every market, whatever the sizes of the relationships. The cost is that p_r is bounded. As a market becomes thicker, p_r tends to one for every branch whose share exceeds its benchmark, whatever the excess (Proposition 1), so in thick markets p_r does not distinguish a branch that lends twice its random-matching share from one that lends ten times as much, whereas $\hat{\kappa}_r$ does.[25] We therefore measure specialization with p, which we compare across markets and partitions in Section 3.2, and use $\hat{\kappa}$ as a measure of the degree of specialization only where its sampling noise is small (Section 4.1).

In Section 3.2.2 we compute these diagnostics for French credit markets.

3 Measuring specialization in practice: France

We now apply the random-matching percentile p of Section 2, the probability that random matching would give a branch a lower share of its lending in the firm's industry than the share it has, to the market for bank credit to firms in France. Nothing in the method

[24] In the French credit data of Section 3.1, in the third of relationships whose $\hat{\kappa}_r$ has the largest variance under random matching, the size class that holds the largest relationships of the industry accounts for a median of 30 to 47% of that variance across the 16 market definitions other than France as a single market, whereas the median probability that random matching assigns a given relationship of that class to the branch is between 0.001 and 0.003.

[25] In the French credit data, a quarter of relationships have p above 0.99 (Section 3.2.1), and among them $\hat{\kappa}_r$ ranges from 1.3 to 1.5 at the 10th percentile to 6.7 to 31 at the 90th, across the 16 market definitions.

is specific to France. It requires, for each period, the borrowing firms, their location and their industry, the amount of credit in each firm–branch relationship, and the identity and location of the branch that provides it. Any data of this kind will do. Unlike most credit data, the French credit register records the lending branch, and in Section 3.2.4 we use this feature to measure how much is lost when the branch must be inferred.

3.1 Data and market definitions

Credit. The relationships come from the French credit register maintained by the Banque de France, which records the credit granted by each bank to each firm above a reporting threshold of EUR 25,000. Credit is the sum of drawn loans, undrawn credit lines and guarantees. The register is monthly; we use end-of-quarter observations. For the vast majority of banks, credit is recorded at the level of the branch that grants it, and the register gives the postal code of both the firm and the branch. Two banks, La Banque Postale and a cooperative bank, report credit only at the level of the département; we exclude them. A relationship (Section 2.1) is a firm–branch pair in a quarter, with all the credit granted by the branch to the firm.

Descriptive statistics. We use the first quarter of 2016. Table 4 describes the data. They contain 1.82 million relationships between 1.36 million firms and 19,984 branches of 408 banks. Most firms borrow from a single branch: the mean number of relationships per firm is 1.3. The 19% of firms with more than one relationship hold 39% of all relationships. Amounts are highly skewed. The median relationship is EUR 117,000, the mean EUR 604,000, and the top percentile exceeds EUR 6 million; we show in Appendix B.3 that the tail is close to a power law with exponent one. Branches are small and unequal: the median branch has 39 relationships and EUR 6.7 million of credit, the mean branch 91 relationships and EUR 55 million. Real estate, which includes the property holding companies used by French households and firms to own buildings, accounts for 39% of relationships, followed by trade, transport and hotels (21%).

Table 4: The credit data, 2016Q1

Panel A: counts					
Relationships	1,815,416				
Firms	1,361,794				
Branches	19,984				
Banks	408				
Panel B: distributions					
	Mean	10%	50%	90%	99%
Credit per relationship (EUR thousand)	604	33	117	614	6,047
Relationships per firm	1.3	1.0	1.0	2.0	5.0
Relationships per branch	90.8	5.0	39.0	156.0	602.2
Credit per branch (EUR million)	54.8	0.9	6.7	40.6	690.5
Panel C: industries (NAF level 0)					
	Relationships (%)		Credit (%)		
Real estate	38.5		37.5		
Trade, transport, hotels	21.2		13.4		
Agriculture	10.6		3.7		
Manufacturing, mining, energy	8.4		11.0		
Construction	7.2		4.1		
Business services	6.4		12.0		
Public admin., education, health	4.6		15.6		
Other services	2.1		1.0		
Information, communication	1.0		1.8		

Notes: One observation per firm–branch relationship with positive credit, first quarter of 2016. Credit in EUR thousand. Panel C: shares of relationships and of credit by industry at NAF level 0 (the A10 aggregation, excluding finance).

Markets. A geographic market is a set of municipalities, and it contains every relationship of every branch located in it (Appendix A, Step 2). We use four partitions of metropolitan France defined by the national statistical institute, from finest to coarsest. An *urban unit* (unité urbaine) is a municipality or group of municipalities with a continuously built-up area of at least 2,000 inhabitants, where no building is more than 200 metres from the next (INSEE, 2020), similar in concept to an Urban Area of the US Census Bureau (U.S. Census Bureau, 2022). An *urban area* (aire urbaine) is a hub with more than 10,000 jobs together with the surrounding municipalities where at least 40% of employed residents work in the hub or in municipalities attracted by it, the French counterpart of a commuting zone (INSEE, 2025). The other two partitions are the départements (Section 2.3) and the 22 régions in place before 2016. We start from the coarsest market of all, metropolitan France as a single market, which is what a measure of specialization without a geographic dimension implicitly assumes, and serves as the benchmark for the finer partitions. Branches outside urban units and urban areas belong to no market in these two partitions, but they do belong to their département and région.

Industries. Firms are classified by the French industry classification (NAF rev. 2) at four levels, from coarse to fine: the A10 aggregation, sections, divisions and groups, with 9, 17, 78 and 241 industries in our data. Crossing the five geographic partitions (France as a whole and the four partitions above) with four industry levels gives twenty market definitions.

How sparse are French credit markets? Table 5 describes the market structure under the twenty definitions. The geographic partitions differ by orders of magnitude. France as a whole is a single market of 19,984 branches and 1.82 million relationships. The median région has 669 branches and about 55,500 relationships, the median département 152 branches and about 12,700 relationships, the median urban area 6 branches and 412 relationships, and the median urban unit 4 branches and 258 relationships. Industries fragment them further. Nationally, the median industry counts about 81,000 borrowing

firms at NAF level 0 and 518 at NAF level 3; in the median urban unit, 17 and 2.

Table 5: Market structure by market definition, 2016Q1

			Per market (median)			Firms per industry	Branch × industry		Relationships (%)	
Geography	Industry	Markets	Branches	Relationships	Industries	(median)	Relationships (median)	One relationship (%)	p undefined	Zero leave-one-out share
France	NAF 0	1	19,984	1,815,416	9	81279	4	22	0.0	1
	NAF 1				17	29609	3	30	0.0	2
	NAF 2				78	2091	2	47	0.0	6
	NAF 3				241	518	1	58	0.0	11
Regions	NAF 0	21	669	55,512	9	3111	4	22	0.0	1
	NAF 1				16	1358	3	30	0.0	2
	NAF 2				73	82	2	47	0.0	6
	NAF 3				212	27	1	58	0.1	11
Départements	NAF 0	94	152	12,729	9	728	4	22	0.0	1
	NAF 1				16	250	3	30	0.0	2
	NAF 2				69	22	2	47	0.1	6
	NAF 3				176	10	1	58	0.2	11
Urban areas	NAF 0	704	6	412	8	29	4	22	0.1	1
	NAF 1				14	13	3	30	0.1	2
	NAF 2				30	4	2	47	0.4	5
	NAF 3				57	3	1	57	1.0	10
Urban units	NAF 0	1,238	4	258	8	17	4	22	0.1	1
	NAF 1				13	8	3	30	0.2	2
	NAF 2				25	3	2	47	0.7	5
	NAF 3				44	2	1	58	1.5	9

Notes: Markets: geographic units of the branch with at least two branches (Appendix A, Step 2); urban units and urban areas exclude the pooled codes for municipalities outside them. Per market: medians across markets of the number of branches, of relationships and of industries with at least one relationship. Firms per industry: median, across market × industry cells, of the number of borrowing firms. Branch × industry: median number of relationships per non-empty branch × industry cell, and share of these cells with a single relationship. p undefined: share of relationships for which p is not defined (Appendix A, Step 8, cases 2 and 3). Zero leave-one-out share: share of relationships, among those with p defined, whose observed leave-one-out share is zero (the firm is the branch's only borrower in its industry).

A larger geography does not make the branch's own portfolio any thicker. The branch × industry cells are the same in every geographic partition: the median non-empty cell holds 4 relationships at NAF level 0 and 1 at NAF level 3, and the share of cells with a single relationship rises from 22% to 58%. What a larger geography changes is the counterfactual, which draws on more relationships and more branches. This is why sparsity cannot be escaped by choosing large markets: the observed share of a branch with few borrowers in an industry is equally coarse in its urban unit and in its région (we show in Section 3.2.2 that p does not reveal specialization more reliably in larger markets). As the industry partition becomes finer, the share of relationships whose firm is the branch's only borrower in its industry, so that its leave-one-out share is zero, rises from 1% to about 10%. The measure p is nevertheless defined for almost every relationship (at least 98.5% in every market definition), because the firm is rarely the only borrower of its industry in

the whole market.

Measures of specialization in sparse markets. Sparsity inflates the measures of specialization built from portfolio shares in the literature. We illustrate it with two of them. Paravisini et al. (2023) classify a bank as specialized in an activity when its portfolio share in that activity is in the top quartile of the distribution of all banks' shares in the same activity (their equation 2 and the quartile dummies that follow it). We apply the same rule to the branches of each market and industry, and measure how specialized the branches it selects appear: the mean relative portfolio share, equation (2), of the branches with a positive share at or above the 75th percentile. Blickle et al. (2026) define a bank's favorite industry as the one in which it "has overinvested most heavily," the industry with the largest excess share, its portfolio share minus the industry's share of all lending (their equation 1); we measure the share of a branch's portfolio in its favorite industry. For both statistics, Table D.1 in Appendix D reports the observed value and its value under the random matching of Appendix A, which contains no specialization at all: the typical (median) value over 100 draws. Panel A of Figure 3 plots the first statistic.

The branches selected by the top-quartile rule look far more specialized at fine industry levels: their mean relative portfolio share rises from 1.85 at NAF level 0 to 3.28 at NAF level 3 in urban units, and from 2.43 to 16.9 for France as a single market. Random matching, which contains no specialization, produces a rise about as large in urban units (from 1.63 to 2.97) and a larger one for France (from 2.21 to 28.2). The rise therefore carries no information about specialization: it is what classifying branches by their shares does in cells with few borrowers. Nor does the level. At NAF level 3, the observed value exceeds the typical random-matching value in about half of the cells of urban units and urban areas (49% and 47%), so in the typical local cell the rule cannot tell the data from random matching. In the larger markets it exceeds it in a minority of cells (37% of départements' cells, 30% of régions' and 16% for France). In these markets, lending to each industry is held by fewer branches than random matching would give it (5 to 12% fewer at NAF level 3), and the branches selected by the rule are larger (by 15 to 24% in median

lending), so fewer of them are small branches, the ones to which random matching gives large shares. Yet the same data are far from random matching by p (Section 3.2.1): the level of a share-based statistic does not order markets by how specialized their branches are. The favorite industry accounts for about two thirds of a branch's portfolio in every market definition (0.62 to 0.68), but random matching alone gives the typical branch a favorite industry worth between a quarter and two fifths of its lending (0.26 to 0.42). Only the difference between the two, not the level of the share, measures specialization. In both cases specialization is measured by the comparison with random matching, which is what p does relationship by relationship. The same problem affects any statistic built from shares, such as the concentration of lending (Appendix C).

3.2 Bank specialization in France

We compute p for every relationship of the first quarter of 2016, under the twenty market definitions of Section 3.1, following Appendix A, with 1,000 counterfactual draws per market.

3.2.1 The distribution of p

Panel B of Figure 3 shows the distribution of p under each market definition, next to its distribution under random matching (Table D.2 in Appendix D reports it in full). A value of p has the same meaning in every market and at every industry level: the probability, counting ties as one half, that random matching gives the branch a lower share in the firm's industry than the share it has. For a branch that is not specialized, its expected value is $1/2$ everywhere, given the assignment probabilities; estimating these from the same data can shift it by a few hundredths for the largest branch of a market (Section 2.4). How far p moves away from $1/2$ for a given degree of specialization, however, differs across cells: it depends on how much information the cell contains, which we measure in Section 3.2.2.

Three facts stand out. First, relationships are far from random matching. The mean of p is between 0.67 and 0.74, its median between 0.84 and 0.89, and a quarter of relation-

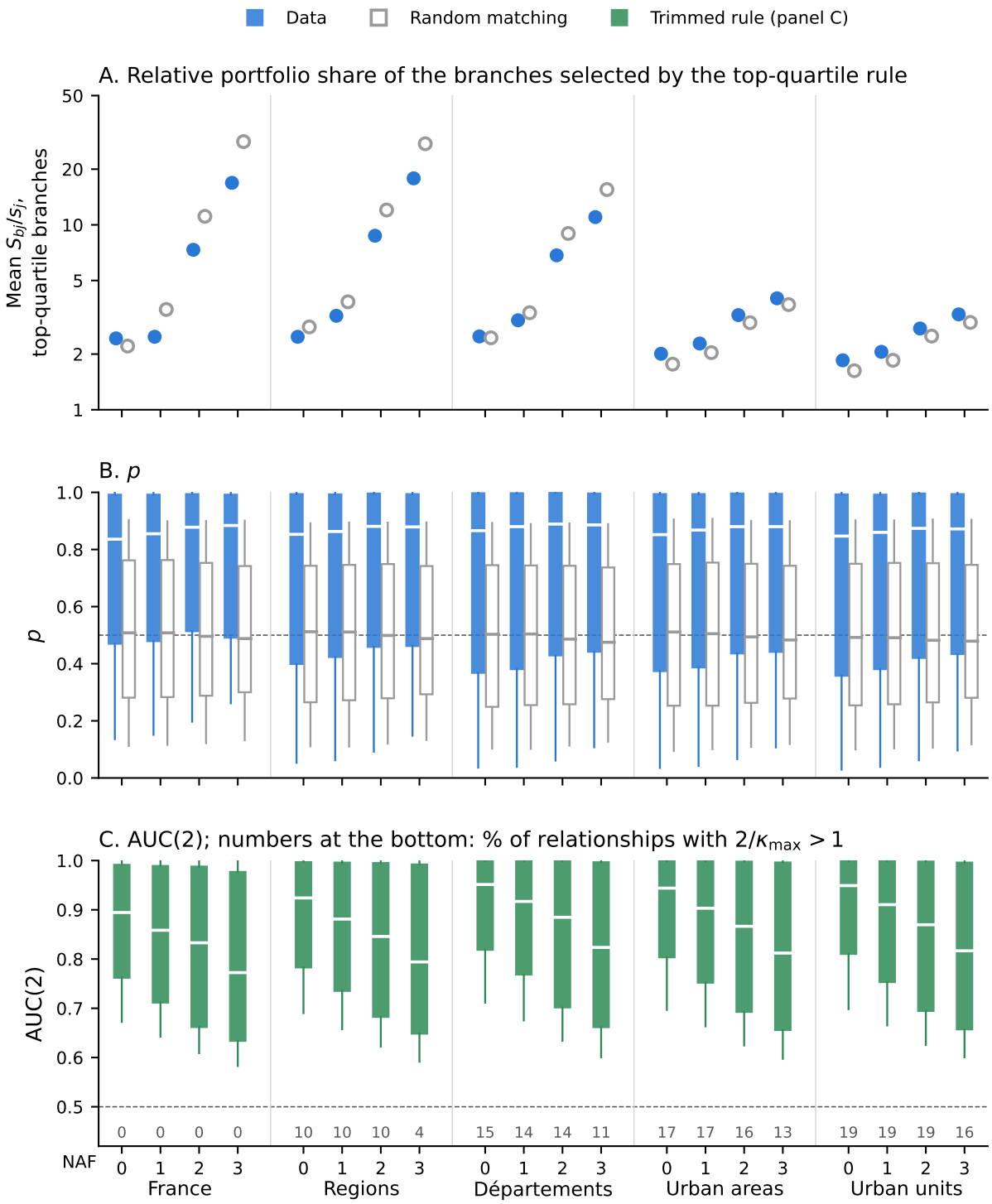


Figure 3: Relative portfolio share, p and the AUC of p by market definition, 2016Q1

Notes: Twenty market definitions: France as a single market, régions, départements, urban areas and urban units of the branch, each at NAF levels 0 to 3. Panel A: mean relative portfolio share, S_{bj}/s_j, of the branches selected by the top-quartile rule of Paravisini et al. (2023) (median across market × industry cells), observed (filled) and under random matching (open, median over 100 draws); log scale (Table D.1). Panel B: distribution of p across relationships, in the data (filled) and on one allocation of the relationships drawn by random matching and treated as data (open); boxes: 25th to 75th percentiles; line: median; whiskers: 10th and 90th percentiles (Table D.2). Panel C: distribution of AUC(2), the expected p under specialized matching with $\kappa = 2$, in which the branch expects to lend to the firm's industry twice what random matching gives it (Section 2.5), computed with the trimmed screening rule of Appendix B.3 on a fixed random sample of 1.2% of relationships, among those for which a doubling is feasible; the numbers above the boxes are the percentage of sampled relationships for which it is not ($2/\kappa_{max} > 1$) (Table D.3).

ships have p above 0.99: for these relationships, random matching almost never gives the branch a share in the firm's industry as high as the one it has. The distribution has a second, smaller mass near zero (the 10th percentile is between 0.03 and 0.26): relationships with branches that lend to the firm's industry much less than random matching would make them lend. Second, the largest branch of each market (France aside, which has a single largest branch) is less specialized than the others: the mean p of its relationships is between 0.56 and 0.60, against 0.69 to 0.73 for the other branches. This is not because p is less able to reveal the specialization of a large branch. On the diagnostics of Section 3.2.2, computed on a sample of relationships, the AUC of the largest branch at $\kappa = 1.5$ has a median between 0.99 and 1.00, against 0.67 to 0.81 for the other branches, and its p is as far from 0.5 on average (a mean $|p - 0.5|$ of 0.38 to 0.45, against 0.36 to 0.37). We use $\kappa = 1.5$ because a doubling is infeasible for many relationships of the largest branch (up to 45% in urban units), which already receives a large part of every industry's lending under random matching.

Third, and most important for the method, the distribution barely changes across market definitions. Moving from the coarsest to the finest industry level raises the mean of p by 0.03 to 0.04 in every geography, and the four partitions of France give almost the same distribution at a given industry level. France as a single market gives somewhat higher values (a mean of 0.70 to 0.74) and fewer values near zero.[26] Measures built from portfolio shares behave very differently: the branches selected by the top-quartile rule look much more specialized at NAF level 3 than at NAF level 0 (their mean relative portfolio share rises from 1.85 to 3.28 in urban units and from 2.43 to 16.9 for France; Figure 3, panel A). Because p has the same benchmark in every cell, the changes it shows across market definitions are not the mechanical inflation of shares in sparser cells. They are not pure changes in specialization either, because sparser cells also give p less power, which

[26] A national market compares each branch with the industry mix of the whole country, so a branch located among farms could look specialized in agriculture only because of where it is. We checked whether this explains the higher values by comparing the mean p at NAF level 0, industry by industry, in France as a single market and in urban areas. The national value is higher in agriculture (0.92 against 0.87), but the differences are small in every industry (at most 0.06) and of mixed sign (-0.02 in manufacturing), so the check neither supports nor rules out this explanation.

pulls it towards 1/2. Part of the small rise at fine industry levels comes from the bottom of the distribution: the 10th percentile rises from 0.03 to 0.10 in urban units, as more relationships fall in sparse cells, where p is pulled towards 0.5 (equation (6)). In the next subsection we measure how much of the sample this concerns.

3.2.2 How much specialization can p detect?

We explained in Section 2.5 why a value of p must be read together with diagnostics that measure how much specialization it could have revealed. Panel C of Figure 3 shows AUC(2) by market definition, and Table D.3 in Appendix D reports the diagnostics in full. The share of draws with positive lending, $P(A > 0)$, and the probability of a tie come from the same draws as p. AUC(2), $2/\kappa_{\max}$ and the ceiling AUC($\kappa_{\max}$) are computed with the screening rule of Appendix B.3, which closely reproduces the simulated diagnostics in our data (Table B.4), on a fixed random sample of 1.2% of relationships, the same in every market definition. We validate the rule on urban units and urban areas in Appendix B.3, and checked it also in the largest markets, on 100 relationships each in régions and in France as a single market, at NAF levels 0 and 3: the rule's AUC(2) differs from the simulated one by a median of -0.004 to 0.000, with 80% of the differences between -0.08 and $+0.01$, and the rank correlation is between 0.96 and 0.99. As in smaller markets, when the rule errs it understates how well p separates specialized branches.

Sparsity is mild at the coarsest industry level and severe at the finest. In the four partitions of France at NAF level 0, the branch receives other lending in the firm's industry in at least 90% of valid draws for about 95% of relationships, ties are rare (a mean probability of 0.004), and the median AUC(2) is between 0.92 and 0.95: p ranks the branch under specialized matching with $\kappa = 2$, where it expects to lend to the industry twice what random matching gives it, above the same branch under random matching more than nine times in ten. At NAF level 3, $P(A > 0)$ is below 0.9 for 27 to 32% of relationships and below one half for 14 to 17%, the median AUC(2) falls to between 0.79 and 0.82, and the lower quartile to about 0.65: for a quarter of relationships at this level, p ranks the branch under specialized matching with $\kappa = 2$ above the same branch under random matching less

than two times in three. The share of relationships with AUC(2) of at least 0.9 falls from 55 to 61% to 42 to 44%. Cells in which even full specialization would leave the expected p below 0.9 remain rare (a ceiling below 0.9 for at most 1.6% of relationships, in urban units at NAF level 3): in most sparse cells, p can reveal the highest degree of specialization the branch can reach, $\kappa_{\max}$, but often not a degree of $\kappa = 2$. The pattern does not depend on the choice of a doubling (Table D.4 in Appendix D): in urban units, from NAF level 0 to NAF level 3, the median AUC falls from 0.84 to 0.74 at $\kappa = 1.5$ and from 0.99 to 0.83 at $\kappa = 3$, a degree of specialization that is infeasible for 39 to 43% of relationships.

A doubling is not always possible. It is infeasible for no relationship in France as a single market, but for 4 to 10% in régions, 11 to 15% in départements, 13 to 17% in urban areas and 16 to 19% in urban units, at every industry level. Two kinds of cells account for these cases. In most of them, the firm's industry is more than half of the branch's expected lending under random matching (all the cases in régions and 61 to 69% in urban units), so the branch cannot double its share in it. Most of these (86 to 100%) are relationships in real estate (NAF group 68.2, renting and operating of own or leased real estate, at NAF level 3), an industry of large loans that makes up more than half of the expected lending of many branches. In the others, the branch can expect more than half of the industry's lending in its market (39 to 45% of the cases in urban units, none in régions), so it cannot receive twice that. This is the situation of the largest branch of a market: in urban units, a doubling is infeasible for 42 to 45% of the relationships of the largest branch, against 10 to 13% for the other branches. In both kinds of cells the ceiling AUC($\kappa_{\max}$) is the relevant diagnostic.

As industries become finer, the room for specialization grows: the median $\kappa_{\max}$ rises from 3.5 at NAF level 0 to 4.2 at NAF level 3 in urban units, and from 6.1 to 52 for France as a single market, because a branch could in principle devote its lending to a small industry many times over its random-matching share. A doubling therefore covers less of that room: the median $2/\kappa_{\max}$ falls from 0.58 to 0.48 in urban units and from 0.33 to 0.04 for France. Yet AUC(2) falls too: in fine industries a branch would have to be specialized well beyond a doubling, and far from the most it could be, for p to reveal it reliably.

Larger markets do not help, and the largest market of all is the hardest. In France as a single market, the median AUC(2) is 0.89 at NAF level 0 and 0.77 at NAF level 3, and its lower quartile 0.63, against 0.95, 0.82 and 0.66 in urban units. Two forces are at work. First, a larger geography leaves the branch's own portfolio unchanged (Section 3.1), so it cannot make the observed share any less coarse. Second, what makes the counterfactual share precise is the effective number of relationships received by the branch under random matching, the number of equal-sized relationships that would give its share the same dispersion (Appendix B.1, equation (B.13)). This number cannot exceed the expected number of relationships received by the branch in the industry (Appendix B.4). With credit amounts whose tail is close to a power law with exponent one (Appendix B.3), a larger market has a larger average amount, so a branch of a given size can expect fewer relationships under random matching, and its counterfactual share becomes sparser, not thicker (Proposition 3, part (iv)).[27] Computed on all relationships, the effective number falls even faster, because the largest relationships of a large market dominate the dispersion of the share (part (ii)), but the branch almost never receives these relationships, and p, a rank, barely reacts to them (Section 2.5, "Why a percentile").

Two checks support this explanation. On the same sampled relationships, we recompute the diagnostics after setting every amount to one, which removes the heterogeneity of amounts (and with it size matching) but keeps the number of relationships of every branch and industry. The median AUC(2) rises to 1.00 at NAF level 0 and to at least 0.998 at NAF level 3 in urban units, régions and France alike, so the gap between urban units and France in the median closes, from 0.05 at NAF level 0 and 0.04 at NAF level 3. Heterogeneous amounts, not the number of relationships, also account for most of the level of the diagnostics. Synthetic markets confirm the role of heavy tails in the level but only partly the role of market size. When a local market is pooled with up to 255 other, independent

[27] In our data the largest 0.1% of relationships hold 30% of all credit. At least one of them lies in 5% of urban units, 8% of urban areas, 59% of départements and every région. For the 32% of relationships whose urban unit holds none, the branch can expect 72% fewer relationships of the firm's industry at NAF level 3 in France as a single market than in its urban unit (a median ratio of 0.28, with assignment probabilities proportional to branch lending); for the others the median ratio is 1.04. Across relationships, the median expected number falls from 38.7 in urban units to 17.3 in France at NAF level 0, and from 10.2 to 5.6 at NAF level 3.

local markets, the expected number of relationships received by a branch under random matching falls from about 19 to 11 when amounts follow a Pareto distribution with tail exponent one, as in our data, and stays unchanged when amounts are equal or log-normal. With Pareto amounts the median AUC(2) of the same relationships is much lower (0.72 to 0.74, against 0.89 to 0.93 otherwise), but pooling lowers it only slightly, by about 0.02, whatever the distribution of amounts (simulated data). Coarser geographies therefore do not buy more power to detect specialization; finer industries cost a great deal of it.

These diagnostics change how the distribution of p (Section 3.2.1) should be read. A value of p near $1/2$ at NAF level 0 is, for most relationships, evidence that the branch is not specialized in the firm's industry: under specialized matching with $\kappa = 2$, its p would most likely have been much higher. At NAF level 3, for a large minority of relationships, it is consistent with no specialization and with a doubling alike. The data show the difference. Among the sampled relationships with AUC(2) of at least 0.9, p is decisive: 67 to 71% of them have p within 0.1 of zero or of one, in urban units and in régions at every industry level. Among those with a lower AUC(2), only 46 to 49% do, and the distribution of p is more concentrated around its center (standard deviation 0.29 to 0.31, against 0.37 to 0.38).

3.2.3 Specialization across French markets

Because p is cardinal, it can be averaged within a market and compared across markets, and across market definitions computed on the same data. We use this to ask two questions about partitioning credit markets more finely, by industry and by geography. First, does the level of specialization change when markets are partitioned more finely? Second, how much variation in specialization does a finer partition uncover, that is, how much do the smaller cells it creates differ from one another beyond what random matching produces in cells of their size? Both questions require a measure with the properties of p: a scale that does not depend on the partition, so that levels and dispersions can be compared across partitions, and a known value without specialization, $1/2$ in every cell, so that the dispersion produced by random matching can be simulated and set aside. Portfolio

shares have neither property: a share that ranks high in one market says nothing about how unusual it would be in another, and a market's average portfolio share in an industry mostly reflects the industry's size in the market. The relative portfolio share has a benchmark that can be simulated by random matching, but its range, $1/s_j$, grows as industries become finer, so its level and its dispersion grow with the partition whatever the specialization of branches (Appendix D, Table D.5, last column).

Figure 4 (panel A) maps the mean p of the relationships of each urban area at NAF level 2, on a color scale centered at the benchmark 1/2; Figures D.1 and D.2 in Appendix D map the other industry levels and geographies on the same scale. The maps, and the dispersion across markets measured below, count each market once, because the question is where specialization is: the mean p of a market describes that market, whether it holds a hundred relationships or a hundred thousand. Levels, as in Table D.2, are means across relationships, which are dominated by the largest markets (the largest quarter of urban areas holds 89% of the relationships of urban areas). The two differ because the mean p of a market rises with its number of relationships (a rank correlation of 0.37 across urban areas, 0.34 across urban units and 0.20 across départements): at NAF level 2, the median urban area has a mean p of 0.61, while the mean of p across the relationships of urban areas is 0.70.

Specialization is widespread: the mean p exceeds 0.5 in 94% of urban areas at NAF level 2, with a median across urban areas of 0.61. It is also uneven. It is highest in the Paris urban area (0.82) and in the North, and lower in the South. Across the régions, the mean ranges from 0.57 in Languedoc-Roussillon to 0.82 in Île-de-France; the next four régions are in the North (0.71 to 0.74) and the lowest four in the South (0.57 to 0.64).

Partitioning industries more finely barely changes the level of specialization. The median across urban areas of their mean p is between 0.59 and 0.61 at every industry level, and the correlation across urban areas between the mean p at NAF level 0 and at NAF levels 1, 2 and 3 is 0.99, 0.96 and 0.90 (Figure D.1). Across relationships, the mean p of urban areas rises from 0.67 at NAF level 0 to 0.71 at NAF level 3 (Table D.2), and part of this rise comes from relationships in sparse cells, whose p moves up towards 0.5: the 10th

percentile of p rises from 0.04 to 0.11. The relative portfolio share behaves differently. The ranking of urban areas by their mean relative portfolio share is as stable as their ranking by p (a correlation of 0.91 between NAF levels 0 and 3), but its level is not: from NAF level 0 to NAF level 3, the mean relative portfolio share of an urban area rises by a median of 0.02 in the smallest quarter of urban areas and 0.39 in the largest, while the median change in the mean p is at most 0.02 in every quarter. A map of the relative portfolio share would therefore show more specialization in the largest urban areas at fine industry levels, whereas the map of p does not.

Finer industries also uncover little variation beyond what random matching produces. Within the same coarser industry of an urban area, the mean p of the finer industries varies, but only 16 to 27% of that variance exceeds what random matching produces in cells of the same size (Figure 4, panel B). Net of cell size, the spread beyond random matching across the finer industries has a standard deviation of 0.053 to 0.075, against 0.097 to 0.123 between the coarser industries. The cells created by finer industries are sparse, and there the AUC of p rises little with specialization (Section 2.5), so this is the variation detectable with finer industries, not necessarily all the variation there is.

Partitioning geography more finely barely changes the level of specialization either: at NAF level 2, the mean p across relationships is between 0.70 and 0.73 in every geography (Table D.2). But finer geography uncovers much more variation than finer industries do (Figure 4, panel B). We compare the mean p of the smaller markets that lie within the same larger unit: the urban units of an urban area, the urban areas of a département and the départements of a région (each smaller market assigned to the larger unit that holds most of its lending). Part of this dispersion arises without specialization: the mean p of every market is then $1/2$ in expectation and departs from it only through sampling noise, which is larger in smaller markets. We measure this noise on an allocation of the relationships drawn by the random matching of Appendix A and treated as data, with p computed on it exactly as on the data. This allocation contains no specialization, so the dispersion of its market means is the dispersion produced by random matching. Across the urban units of the same urban area (605 urban units in the 72 urban areas that contain at least two), the standard

deviation of the mean p is 0.096, against 0.057 under random matching: 65% of the variance exceeds what random matching produces, against 16 to 27% for finer industries. Since the mean p of a market rises with its size, and an urban area joins a large core with smaller towns, we also remove the relation between a market's mean p and the log of its number of relationships. Net of market size, the spread beyond random matching across the urban units of an urban area has a standard deviation of 0.057, as large as the spread between urban areas (0.057), and the comparison is the same at every industry level (0.057 to 0.058 within, against 0.055 to 0.057 between). Urban areas differ as much within a département as départements do (0.053 against 0.051), and départements more within a région than régions do (0.051 against 0.042). Specialization is therefore local: neighbouring urban units of the same commuting zone differ about as much as commuting zones do, whereas the averages of the narrow industries within a broad one differ much less, at least by as much as p can detect in their sparse cells. The two comparisons differ in what they hold fixed (Section 2.3). The smaller geographic markets within a larger geographic unit, such as the urban units of an urban area, are distinct markets: each contains different branches, and the p of its relationships is computed against the random matching of that market. The finer industries within a coarser industry of an urban area belong to a single geographic market: their relationships are those of the same set of branches, and their p is computed against the same random matching. The dispersion across smaller geographic markets therefore measures how the specialization of branches differs across neighbouring locations, whereas the dispersion across finer industries measures how the average specialization of the branches of one location differs across related industries. Finer geographic partitions thus carry information about specialization, as long as the cells they create are not too sparse for p to reveal it (Section 3.2.2).

The comparison across finer industries is a comparison of averages over the branches of a location, and an average can hide how these branches divide an industry among them. Suppose that two branches of an urban area lend to the same coarser industry, that each is specialized in a different finer industry within it, and that most borrowers of each finer industry borrow from the branch specialized in it. The mean p of each finer industry is

then high, and the two means are similar, although each branch is specialized in one finer industry and not in the other. Because p is specific to the branch, we can repeat the comparison within branches. For each branch, we compare its p in the finer industries of the same coarser industry, and measure the dispersion of these values beyond what the random-matching allocation above produces; on that allocation, each relationship belongs to the branch to which random matching assigns it. As for markets, we remove the relation between p and the size of the cell, here the number of relationships of the market in the finer industry, which determines how much p can reveal. We do not use the branch's own number of relationships in the industry: a branch specialized in an industry has more of them, so this control would remove the specialization we measure.

The same branch differs across the finer industries of a coarser one much more than the averages of its location do. Net of the size of the market's cell, the standard deviation beyond random matching of a branch's p across the finer industries of a coarser one is between 0.10 and 0.14, against 0.05 to 0.08 for the averages of an urban area or an urban unit, at every industry level. For three-digit industries within broad ones, it is 0.12 for branches, against about 0.06 for the averages of their location. Most of the dispersion of a branch's p across finer industries is still sampling noise: only 15 to 25% of its variance exceeds what random matching produces, because a branch's cells in narrow industries hold few relationships. Branches therefore differ in the narrow industries in which they specialize, and averaging over the branches of a location hides most of this variation. A measure that is specific to the branch is needed to reveal it. It is also the variation used by the branch closures (Section 4.3.4): they compare two branches in the firm's own narrow industry.

The maps should be read with the diagnostics of Section 3.2.2: averages that include many cells where only extreme specialization is detectable are pulled towards 1/2.

Finer partitions, by industry or by geography, also make cells sparser, and specialization varies with how sparse a cell is. Figure 5 groups the market $\times$ industry cells of urban units by the number of firms that borrow in them (Figure D.3 in Appendix D shows the other geographic partitions), and compares each statistic with the same statistic on the

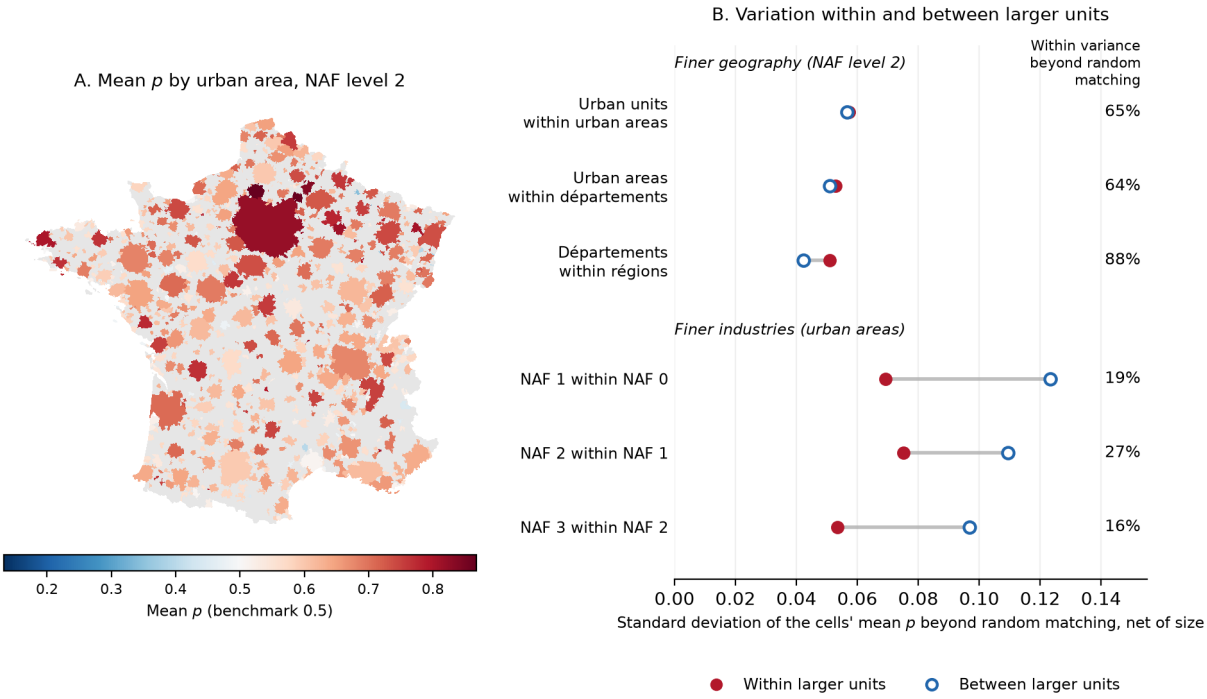


Figure 4: Where specialization is, and the variation uncovered by finer partitions, 2016Q1

Notes: Panel A: mean p of the relationships of the branches of each urban area, industries at NAF level 2; diverging color scale centered at 0.5, the expected value of p without specialization; grey: municipalities outside urban areas and urban areas with a single branch. Panel B: each row nests smaller cells in larger units, smaller markets in the larger geographic unit that holds most of their lending (industries at NAF level 2), or finer industries in the coarser industry of the same urban area. Dots: standard deviation beyond random matching (square root of the variance in the data minus the variance on an allocation drawn by random matching and treated as data) of the cells' mean p, within the larger units and between them, after removing the relation between a cell's mean p and the log of its number of relationships. Right: share of the variance of the cells' mean p within larger units that exceeds the variance under random matching. Values in Table D.5.

random-matching allocations described above, averaged over 5 allocations (dashed lines). It weights the observations in two ways, which answer different questions. Weighting each relationship of a cell equally (panel A) asks how specialized the lender of a typical borrower is. This is the weighting of Table D.2, and the benchmark of the mean p is 1/2 in every cell (Section 2.4). Weighting equally each branch that lends to the industry (panel B) asks how specialized a typical lender to the industry is. With this weighting the benchmark is below 0.5 (between 0.43 and 0.50), because a branch whose only borrower in the industry is the firm itself has a leave-one-out share of zero and a value of p below 0.5 (Section 2.4), and such branches are many.

Weighted by relationships, the mean p rises with the size of the cell, in urban units and in every other partition: from close to 0.5 in cells with two or three firms (0.48 to 0.57 across partitions) to 0.64 to 0.73 in cells with more than 200 firms. Weighted by lending branches, the excess over the benchmark does not rise. In cells with more than 200 firms it is between 0.01 and 0.02 at NAF level 0 and between 0.07 and 0.09 at NAF level 3. In départements and régions at NAF levels 0 and 1 it is close to zero or below in large cells (about 500 to 2,000 firms in départements, 2,000 to 6,000 in régions), and rises again in the largest ones. In thick cells, therefore, the branches that lend to an industry differ: the typical lender is close to random matching, but the branches that hold most of the industry's relationships lend to it more than random matching would. In sparse cells both statistics are close to their benchmarks (an excess of at most 0.07 for lending branches). This partly reflects how little p can move there: in urban units and urban areas, the mean probability of a tie is between 0.24 and 0.50 in cells with two or three firms, against at most 0.05 in cells with more than 200 firms, and ties pull p towards 1/2 (equation (6)).

3.2.4 Comparison with the measures used in the literature

Our approach departs from the measures of specialization used in the literature in three ways: the scale on which specialization is measured, the exclusion of the relationship from its own branch's portfolio, and the observation of the lending branch. We take each in turn.

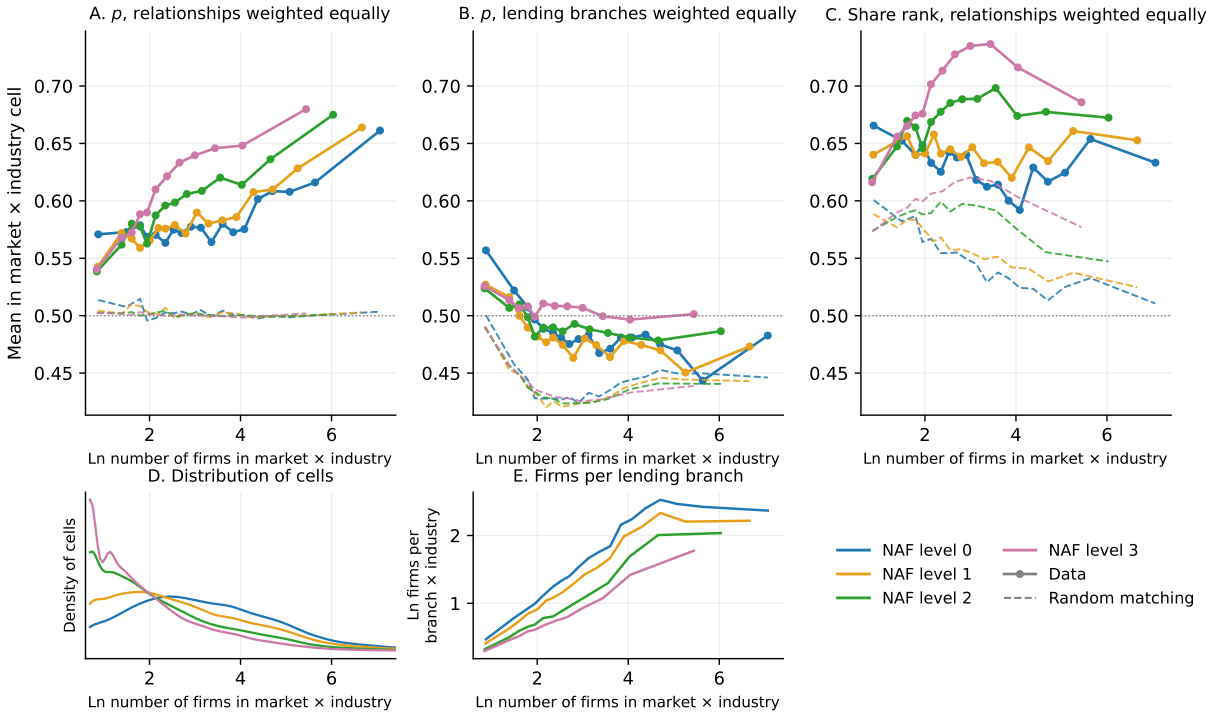


Figure 5: Specialization and the sparsity of markets, urban units, 2016Q1

Notes: Cells: urban unit × industry, at NAF levels 0 to 3, with at least two borrowing firms. Panel A: mean p over the relationships of the cell; panel B: mean over the branches that lend to the industry of the branch's mean p; panel C: mean share rank (the percentile of the branch's leave-one-out share among all branches of the market, Section 3.2.4) over the relationships of the cell. Solid lines: the data. Dashed lines: the same statistics on allocations of the relationships drawn by the random matching of Appendix A (Steps 5–6) and treated as data, averaged over 5 allocations, the benchmark without specialization. Means of the cells in 20 equal-sized bins of the log number of borrowing firms. Panel D: density of cells; panel E: mean log number of firms per branch × industry. The dotted line is 1/2.

Measures of specialization. We showed in Section 3.1 that the measures of specialization built from portfolio shares move with the industry partition for mechanical reasons. Here we ask whether, within a market definition, they at least order relationships as p does. We compute, for every relationship and on the same data as p, three such measures. All of them use the leave-one-out share S_{bj}^{-r} of equation (4), and the industry's share of market lending also without relationship r, so that they differ from p only in how they read the same shares. The *excess share*, the portfolio share minus the industry's share of lending, is the baseline measure of Blickle et al. (2026) (their equation 1). The *relative share*, the ratio of the two, is their second measure (their equation 2) and our relative portfolio share, equation (2); they note that it has large right tails with fine industries, which is the sparsity problem of Section 2.2. The *share rank* is the percentile of the branch's share among the shares of all branches of the market in the same industry (zero shares included, ties counting one half), a continuous version of the top-quartile rule of Paravisini et al. (2023). Paravisini, Rappoport and Schnabl define the relative portfolio share as the ideal index of specialization (their equation 1) but cannot compute it, because in their data credit is not observed by activity. They measure the portfolio share with a proxy, the debt-weighted exports of a bank's borrowers to each destination (their equation 2), and, since the cardinal value of this share does not capture specialization, they use its rank among banks.

The rank correlation of each measure with p varies little across the 20 market definitions (Table D.6 in Appendix D). The excess share and the relative share, which use the same leave-one-out shares as p and differ from it mainly in their scale, are closely related to it (rank correlations of 0.81 to 0.90 and 0.81 to 0.91). The share rank is not: its rank correlation with p is only 0.64 to 0.71.

Figure 5 shows where the difference comes from. It plots, for urban units, the mean share rank (panel C) and the mean p (panel A) of the relationships of a market $\times$ industry cell against the number of firms borrowing in the cell, together with the same means on allocations drawn by random matching, which contain no specialization (dashed lines). The number of firms in a cell varies with the size of the urban area and with the size of the industry, and finer industry levels shift cells towards fewer firms. Without specialization,

the mean of p is 0.5 in cells of every size and at every industry level (between 0.50 and 0.51). The share rank has no such benchmark. Under random matching it lies between 0.52 and 0.62, and it varies with the size of the cell and with the industry level, because relationships are, by construction, observed at branches that lend to the industry, and these tend to rank above the branches of the market that lend nothing to it. The same share rank can therefore be produced by random matching in one cell and be evidence of specialization in another. In the data, p is close to 0.5 in cells with two or three firms (0.54 to 0.57), as it should be when they contain little information, and moves away from 0.5 as cells grow (0.64 to 0.71 in cells with more than 200 firms). The share rank exceeds its benchmark at every cell size, by 0.04 to 0.13, but measuring by how much requires simulating the benchmark, which is what p does relationship by relationship.

Leaving the relationship out. Portfolio shares computed on a lender's whole portfolio, such as the bank-level shares of Blickle et al. (2026), include the firm's own loan. Leaving the relationship out is required for the benchmark of p, and it removes a mechanical correlation between the measure and the firm's own borrowing (Section 2.3). To measure the size of the bias that including the relationship introduces, we also compute p without leaving the relationship out: the observed share includes relationship r, and each counterfactual share includes it when random matching assigns it to the branch. Figure 6 plots the mean difference between p and this version, by size of the urban area. Including the firm's own loan raises p on average by 0.016 at NAF level 0 and by 0.061 at NAF level 3, and by more than 0.1 for 14% of relationships at NAF level 3. The bias is largest in small and mid-sized urban areas, and smallest in Paris (0.01 at NAF level 0). It grows with the fineness of the industry partition, so a measure that includes the firm's own loan, like portfolio shares, adds spurious specialization exactly where markets are sparsest.

Observing the lending branch. Most credit data record the bank that lends to a firm, but not the branch. A researcher who wants to measure specialization at the branch level, or in local markets, must then assign each relationship to a branch, and the natural choice

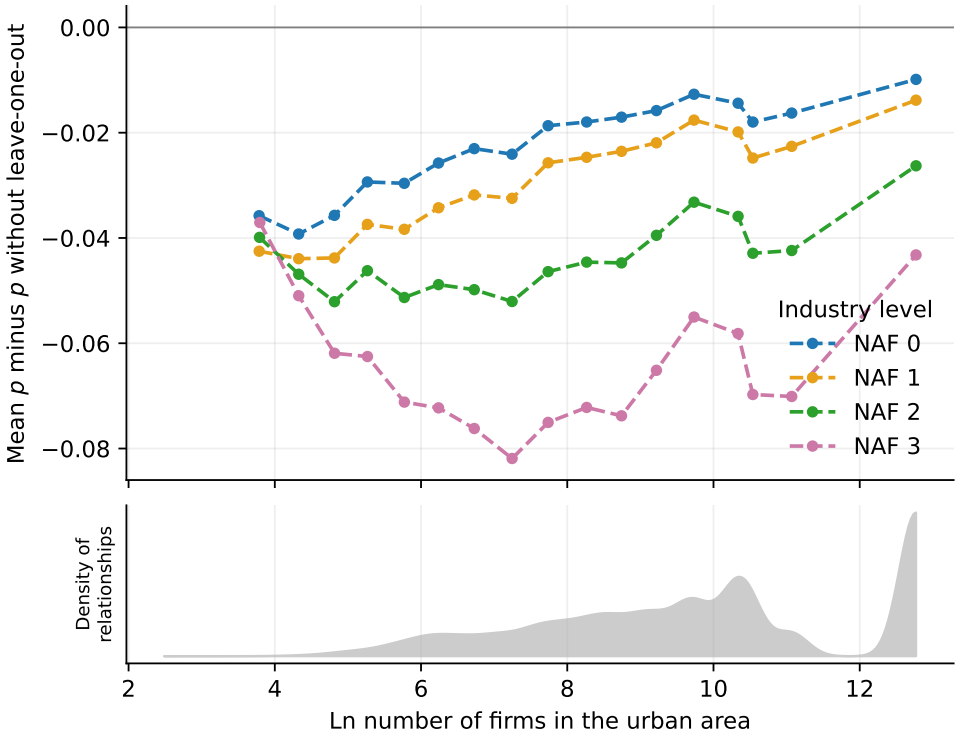


Figure 6: Leave-one-out bias correction, urban areas, 2016Q1

Notes: Mean difference between p (equation (5), leave-one-out) and p computed without leaving the relationship out (observed share A_{bj}/V_b; counterfactual share over all relationships assigned to the branch), in bins of width 0.5 of the log number of firms borrowing from branches of the urban area (bins with at least 200 relationships), by NAF level. Bottom: density of relationships.

is the branch of the bank closest to the firm. Because the French register records the actual branch, we can measure the error this assumption introduces.

For each relationship, we find the branch of the same bank that is closest to the firm. When several branches are equally close (for instance, in the firm's own municipality), the assumption does not determine the branch, and we pick one of them at random. We then ask four questions. How often is the actual branch not the closest one? How far is the actual branch from the assumed one? How often does the assumed branch lie in a different market, so that the relationship is counted in the wrong market? And how different is p when it is computed on the assumed matches rather than the actual ones? The last question measures the cost of the assumption for the object of this paper: reassigning relationships changes the lending of every branch involved, its portfolio shares, and its counterfactual distribution.

We locate firms and branches by urban unit: each firm and each branch is placed at the centroid of its urban unit, or of the part of its département outside urban units, so the branches of a bank in the same urban unit are tied and the distance between them is zero. This resolution overstates how often the assumption picks the wrong branch, through ties, and understates how often the actual branch is not the closest one, since it misses errors within an urban unit.

Table D.7 in Appendix D answers the first three questions. For 29% of relationships, the actual branch lies outside the closest urban unit in which the firm's bank has a branch, and for 54% of relationships of firms located outside urban units. Counting the ties broken at random, the assumption picks the wrong branch for 61% of relationships (49% of credit). When it does, the assumed branch lies in a different urban unit in 48% of cases and in a different urban area in 44%, so that the relationship is counted in the wrong market. Outside the urban units with the most branches, 78 to 89% of wrong assignments fall in another urban unit, at a median distance of 19 to 30 km from the actual branch; in the urban units with the most branches, most errors pick the wrong one of the bank's several branches in the same urban unit.

Table 6 answers the fourth. Computed on the assumed matches, p has a correlation

of only 0.50 to 0.52 with p on the actual matches (a rank correlation of 0.55 to 0.57), in urban units and urban areas and at every industry level, and it changes by more than 0.25 for 34 to 36% of relationships. It changes also for relationships whose branch the assumption gets right (a mean absolute change of 0.11 to 0.13, against 0.31 to 0.33 when it gets it wrong), because reassigning the other relationships of their branch and market changes the branch's portfolio and its counterfactual. On average, the assumption lowers p, from 0.67–0.71 to 0.61–0.64, as one would expect, since it assigns relationships by distance, whatever the firm's industry, as random matching does. Without the lending branch, much of the information in p about branch specialization is lost.

Table 6: p on actual and on assumed (closest-branch) matches, 2016Q1

			Mean p		Correlation		Mean $\lvert\Delta p\rvert$			
Markets	Industry	N	Actual	Assumed	Pearson	Rank	All	Branch wrong	Branch right	$\lvert\Delta p\rvert > 0.25$
Urban units	NAF 0	1,549,530	0.67	0.61	0.51	0.56	0.240	0.330	0.126	36%
	NAF 1	1,547,518	0.68	0.62	0.51	0.55	0.239	0.327	0.127	36%
	NAF 2	1,534,739	0.70	0.63	0.51	0.57	0.233	0.320	0.123	35%
	NAF 3	1,513,986	0.70	0.63	0.50	0.56	0.226	0.312	0.117	35%
Urban areas	NAF 0	1,561,292	0.67	0.61	0.52	0.57	0.234	0.322	0.120	36%
	NAF 1	1,560,250	0.68	0.62	0.52	0.56	0.234	0.319	0.121	36%
	NAF 2	1,552,530	0.70	0.63	0.51	0.57	0.228	0.312	0.117	35%
	NAF 3	1,539,318	0.71	0.64	0.51	0.57	0.221	0.305	0.112	34%

Notes: Relationships with p defined on both the actual and the assumed matches. Assumed matches: every relationship moved to its assumed branch (Table D.7) and to that branch's market, relationships of a firm that fall on the same branch merged, matching rates re-estimated, and p computed as in Appendix A. Correlation: Pearson and Spearman. Mean $\lvert\Delta p\rvert$: mean absolute difference between the two values of p, for all relationships and by whether the assumption picks the actual branch; last column: share of relationships whose p changes by more than 0.25.

4 The causal effect of branch specialization

We estimate the causal effect of a branch's specialization on the credit of its borrowers: how much new credit a set of firms receives, and which of them receive it, when their lending branch is replaced by one more, or less, specialized in their industry. From this effect we infer whether the tilt of a branch's portfolio towards an industry signals a screening advantage: whether the branch screens the industry's borrowers better than other lenders

do.[28]

A lending advantage in an industry is an advantage of a lender over other lenders in lending to the industry's firms. A lender has a screening advantage in an industry when the lender is better at telling the firms of the industry that will repay from those that will not. Specialization (a portfolio tilt towards an industry) does not imply any advantage. In a competitive market, word of mouth among borrowers, herding, or differences in how lenders market their credit can lead the firms of an industry to the same branch without any lending advantage. In such a market, the tilt signals no lending advantage: any branch screens the firms of an industry as well as any other. This is the null hypothesis of our tests.

The alternative hypothesis, that specialization signals better screening, has a unique prediction: a more specialized branch shifts its credit away from the firms that turn out to be bad risks and towards those that turn out to be good risks.[29]

Under the alternative, better screening also affects the market power of the specialized lender. Competing lenders that are not specialized face a winner's curse, because the firms that come to them include those turned away by the better-informed specialized lender. The specialized lender can therefore lend on worse terms without losing its borrowers. As a result, better screening has two opposing effects on the level of lending: it lowers the lender's expected cost of lending, because less of its credit goes to firms that default, which leads to more lending; and it gives the lender market power, which leads to higher markups and less lending. We do not observe loan prices, so we infer which force dominates from

[28] Paravisini et al. (2023) show that banks have lending advantages in the markets in which they specialize, and we follow their logic in asking what specialization signals. They posit that a lending advantage specific to a market makes credit from one bank difficult to substitute with credit from another, so that bank specialization can segment credit markets and increase lender market power, and they develop an empirical framework to evaluate this conjecture (p. 2050). They test its two parts separately. When firms expand exports to a country, they shift their borrowing towards banks specialized in that country (p. 2068), evidence that speaks "to the existence of market-specific lending advantages, but not to their importance for the real economy" (p. 2072). A shock to a bank's credit supply reduces firms' exports only to the countries in which the bank specializes, which "indicates that specialized bank credit is difficult to substitute, while nonspecialized credit is not" (p. 2052). They model specialization in reduced form, since "the source of lending advantages is unobserved by the econometrician" (p. 2059), and conclude that "while our methodology can identify the existence and importance of a lending advantage associated with an export destination, the specific source of the bank lending advantage is unknown" (p. 2080).

[29] An increase in credit together with a decline in default cannot come from a move along a supply curve (Fisman et al., 2017).

the quantity of lending. In both cases the specialized lender turns away the firms it can tell are bad risks, so the prediction of the previous paragraph holds whichever force dominates.

We use branch closures for the test. A closure transfers the accounts of the closing branch to other branches of the same bank, and so changes, for the borrowers of each industry, how specialized their lending branch is in their industry. In Section 4.1 we compare the cohorts of borrowers of different industries of the same closing branch, and ask whether their new credit, and the part of it that goes to firms that default within a year, change with the specialization of their new branch relative to the old one. Under the null hypothesis neither responds. Under the alternative, new lending does not fall, new lending to future defaulters does not rise, and at least one of them moves. We also estimate how firms that borrow from several banks distribute their borrowing across them when the specialization of one of their lenders changes, which shows whether their other lenders take up or give up the difference.

4.1 Research design

Reassignment after a branch closure. When a bank closes a branch for reasons other than a merger, it transfers the accounts of the closing branch to other branches of the same bank (Section 4.2 describes the setting). The borrowers of the closing branch keep their bank but change branch, and the new branch can be more or less specialized than the old one in each borrower's industry. We follow the borrowers of each closing branch by industry. The *industry cohort* cj consists of the firms in industry j that borrow from the closing branch c in the quarter before the closure is announced, and we follow these firms wherever they borrow afterwards. We date quarters by τ, relative to the quarter in which the branch closes ($\tau = 0$). The closure is announced six quarters earlier, in quarter $\tau = -6$, so the cohort is defined in quarter $\tau = -7$.

The change in specialization. The bank transfers the accounts of a closing branch to other branches of the same bank, and most of its borrowers end up at one of them. We call the branch that receives the largest number of the closing branch's borrowers its main

absorbing branch, $a(c)$. In the sample of our regressions, 86% of the firms of a cohort, holding 93% of its credit, move to another branch of the bank, and 90% of these, holding 88% of their credit, move to $a(c)$ (Table E.3 in Appendix E). We measure the change in the specialization of the cohort's lending branch by

$$\Delta p_{cj} = p_{a(c)j} - p_{cj}, \tag{9}$$

where p_{bj} is the random-matching percentile of branch b's share of its lending in industry j in quarter $\tau = -7$: the probability that random matching (Section 2.3) would give the branch a lower share of its lending in industry j than the share it has, with ties counting one half. Unlike the p of a relationship (Section 2.4), p_{bj} leaves no relationship out. It compares the branch's whole portfolio with the portfolios assigned to it by random matching, so it is 1/2 on average for a branch that is not specialized, as the p of a relationship is. We will also estimate the regressions at the firm level, with each firm's leave-one-out p.[30] We measure both terms in quarter -7, before any account is transferred, so that the portfolio of the closing branch does not enter $p_{a(c)j}$. Where $a(c)$ lends in the market but not to industry j, its share is zero and $p_{a(c)j}$ is half the probability that random matching gives it no lending to the industry (Section 2.4). Where $a(c)$ has no lending in that quarter, we assume that it is not specialized and set $p_{a(c)j} = 1/2$. $\Delta p_{cj} > 0$ means that the cohort moves to a branch more specialized in its industry than the closing branch, and $\Delta p_{cj} < 0$ that it moves to a less specialized one. We assign the main absorbing branch to every firm of the cohort, including the firms transferred to another branch and those that leave the bank, so that Δp_{cj} does not depend on where firms choose to borrow after the announcement: it is an intention-to-treat measure of the change in the specialization of the cohort's lending branch.

[30] Because the cohort is the closing branch's lending to industry j, p_{cj} contains the cohort's own credit in quarter -7, and leaving the cohort out would leave nothing of the branch's lending to the industry. The cohort regressions therefore keep the mechanical correlation between the measure and the borrowers' own credit, a correlation removed in firm-level regressions by leaving the firm's own relationship out (Section 2.3).

Regression. We measure each cohort's outcomes in two periods: before the announcement, from quarter -16 to quarter -7, and after it, from quarter -6 to quarter $+8$. We estimate

$$y_{cjt} = \beta \, \text{Post}_t \times \Delta p_{cj} + \alpha_{cj} + \alpha_{ct} + \varepsilon_{cjt}, \tag{10}$$

where t indexes the two periods, Post_t equals one after the announcement, α_{cj} is a cohort fixed effect and α_{ct} a closing branch $\times$ period fixed effect. Each cohort is weighted by its credit from the bank in quarter -7, and standard errors are clustered by closing branch. We report results for markets of départements and compare market definitions in Section 4.3.4.

The closing branch $\times$ period effects absorb the closure itself and anything else common to the borrowers of the closing branch: shocks to the bank, to the local economy or to the branch's clientele as a whole, and every characteristic of the main absorbing branch that does not depend on the industry, such as its size, its distance to the borrowers or its staff. The coefficient β therefore compares the cohorts of different industries of the same closing branch: it measures how a cohort's outcome changes after the announcement, relative to the other cohorts of its closing branch, per unit of the change in the specialization of its lending branch.

Outcomes. We measure two outcomes of each cohort, per quarter and as shares of the cohort's credit from the bank in quarter -7. *New lending* is the new medium- and long-term credit granted by the bank to the cohort's firms. The credit register records stocks of credit, not loans, so we measure the new credit to a firm in a quarter by the increase in its stock of medium- and long-term credit from the bank since the previous quarter, summed over the bank's branches so that transfers between branches do not count. The measure is a lower bound: a new loan smaller than the quarter's repayments, and the refinancing of a loan, do not show. *New lending to future defaulters* is the part of this new credit that goes to firms for which, within a year, the Banque de France records at least one significant payment incident or a worse event, such as collective proceedings (a rating of 7 or worse; Section 4.2). We measure both outcomes over the firm-quarters in which default within

a year can be observed. About 1% of new credit goes to future defaulters: 0.9 to 1.1% before the announcement and 1.2% after it (Table E.4 in Appendix E). The two outcomes separate a branch that lends more to all its borrowers from one that changes which of its borrowers it lends to.

Identification. Closures need not be random. A bank may close branches where local activity or credit demand is declining, or where online banking spreads fastest, and the borrowers of closing branches may differ from other firms. Such forces act on all the cohorts of a closing branch, and the closing branch × period effects absorb them. The design requires that, among the cohorts of a given closing branch, the differences in how the shocks behind the closure affect them are not correlated with their Δp_{cj}. Since the closing branch and the main absorbing branch are the same for all of them, Δp_{cj} differs across these cohorts only through their industry. The requirement is therefore that the industries hit hardest by the shocks behind a closure are not systematically those in which the absorbing branch is less, or more, specialized than the closing branch. It would fail, for example, if branches closed because the industries in which they specialize declined locally: their cohorts in those industries would have a negative Δp_{cj}, and their outcomes would deteriorate for reasons unrelated to the new branch.

The requirement has two testable implications. First, before the announcement, when no account has yet been transferred, Δp_{cj} should not predict the trends in the cohorts' outcomes (Section 4.3.1). Second, Δp_{cj} should not predict the outcomes of the borrowers in industry j of the branches that do not close, since a local decline of the industry would affect them as well.

Substitution regression. We also ask how a firm distributes its borrowing across its lenders when the specialization of one of them changes. Consider a firm f in industry j that borrows from bank k through branch b, which closes, and is transferred to branch b' of the same bank. We measure the change in the specialization of its lending branch by $\Delta p_{fk} = p_{b'j} - \bar{p}_r$, where $p_{b'j}$ is defined as in (9) and $\bar{p}_r$ is the p of the firm's relationship

r with the closing branch, which leaves the relationship out (Section 2.3), averaged over the fifth to eighth quarters before the announcement. Where Δp_{fk} cannot be measured, because the relationship ends before the closure or is too recent to have a p five quarters before the announcement, we set it to zero. We estimate, on firms that borrow from at least two banks,

$$\ln C_{fkt} = \eta_0 \operatorname{Post}_{fkt} + \eta_1 \operatorname{Post}_{fkt} \times \Delta p_{fk} + \gamma\, p_{fkt} + \delta' \operatorname{Post}_{fkt} \times X_{fk} + \alpha_{fk} + \alpha_{ft} + \alpha_{g(k)t} + \varepsilon_{fkt}, \quad (11)$$

where C_{fkt} is the credit (drawn credit plus undrawn credit lines) that bank k grants to firm f in quarter t, summed over the bank's branches. $\operatorname{Post}_{fkt}$ equals one for the relationships reassigned by a closure from the quarter in which the closure is announced to eight quarters after the closure, and zero otherwise. The reassigned relationships enter from 16 quarters before the closure to 8 quarters after it, so that their reference period is the ten quarters before the announcement; the other relationships enter in every quarter. p_{fkt} is the p of the firm's relationship with the bank, averaged over the fifth to eighth quarters before t, which controls for the level of specialization. X_{fk} contains the other changes brought about by the reassignment, each measured as the value after the reassignment minus the value before it and winsorized at the 1st and 99th percentiles: the change in the distance, in kilometres, between the firm and its branch, and the change in the size of the branch, the log of its total credit; it also contains an indicator of a relationship with the closing bank of at least 12 quarters at the closure. The fixed effects are for the firm–bank pair (α_{fk}), the firm and quarter (α_{ft}), and the bank's group and quarter ($\alpha_{g(k)t}$). Standard errors are clustered by the firm's urban unit and year.

The firm–quarter effects make this regression a comparison between the reassigned lender and the firm's other lenders in the same quarter. $\eta_1 > 0$ means that the firm borrows more from its bank, relative to its other lenders, when the branch that serves it becomes more specialized in its industry, either because it borrows more in total or because it shifts its borrowing towards that bank. Since $\bar{p}_r$ leaves the firm's own relationship out, the

firm's credit does not enter Δp_{fk}. The identification of η_1 rests on the requirement stated above. Comparing lenders of the same firm does not by itself separate credit supply from credit demand when banks are specialized: a shock to one of the firm's activities shifts its demand for credit towards the lenders specialized in that activity (Paravisini et al., 2023, pp. 2077–2078).

The power of p and the interpretation of the coefficients. Two properties of p established in Section 2.5 bear on β and η_1. Write $p = \mathrm{AUC}(\kappa) + u$, where κ is the degree of specialization of the branch in industry j, $\mathrm{AUC}(\kappa)$ is the expected value of p under specialized matching of degree κ (equation (7)), and the deviation u has mean zero given κ. First, in sparse cells $\mathrm{AUC}(\kappa)$ varies little with κ (Proposition 2), so most of the variation of p across branches comes from u.[31] If u is unrelated to the outcomes, it attenuates the coefficients towards zero, and more so where cells are sparser. Second, the same change in κ moves the expected value of p by different amounts in different cells: by much where $\mathrm{AUC}(\kappa)$ rises steeply, and by little where it is flat. A unit of Δp is the same change, for every cohort, in the probability that random matching would give its lending branch a lower share, but not the same change in the degree of specialization. Both properties affect the magnitude of the coefficients and their comparison across market definitions. We therefore also estimate them separately for groups of cohorts defined by the power of p for the closing and the absorbing branch, measured by the smaller of their two values of AUC(2). Within a group, the AUC curves of the two branches rise similarly between $\kappa = 1$ and $\kappa = 2$, so that near the benchmark a unit of Δp corresponds to changes in κ of similar size; and in the groups with high power the deviation u matters less. In the group with the highest power we also replace Δp by the change in $\hat{\kappa}$, the estimator of the degree of specialization of Section 2.5, equation (8), whose sampling noise is smallest there.

[31] Write P_0 and P_κ as in Section 2.5. Given κ, p equals $\frac{1}{2}(1 - P_0)$ with probability $1 - P_\kappa$ and is at least $1 - P_0$ otherwise, so by the law of total variance its variance given κ is at least $\frac{1}{4}(1 - P_0)^2 P_\kappa (1 - P_\kappa)$. Since $\mathrm{AUC}(\kappa)$ increases from $1/2$ at $\kappa = 1$ and is at most $\frac{1}{2}(1 + P_\kappa)$, the variance of $\mathrm{AUC}(\kappa)$ across branches is at most $\frac{1}{4}E[P_\kappa^2]$. The ratio of this variance to the mean variance of u is therefore at most $E[P_\kappa^2] / E[(1 - P_0)^2 P_\kappa (1 - P_\kappa)]$, which tends to zero as the probabilities P_κ tend to zero, that is, as the cell becomes sparse.

4.2 Branch closures in France

Branch networks and closures. France had a dense network of bank branches in the early 2010s. Following the financial crisis, and with the rise of online banking, several large French banks announced plans to reduce their networks. The number of branches lending to households and firms fell by 4% between January 2010 and December 2016, from 38,784 to 37,261.[32] About a third of them lend to small and medium-sized enterprises (SMEs): 12,291 at the end of 2016, according to the credit register. About two thirds of the closures recorded since 2010 follow a merger between two banks; the others result from decisions of the bank itself. We use only the latter, because only then are the accounts of the closing branch transferred to another branch of the same bank. These closures are spread over time and over the territory (Appendix E, Figures E.1 and E.2).

How a closure works. Banks must announce a closure to their customers and to the regulator before the branch ceases to operate. In the data, announcements take place 14 to 17 months before the closure. On closure, the firm's accounts are transferred automatically to another branch of the same bank, chosen by the bank. More than 90% of the firms of the closing branches are transferred to the absorbing branch designated by the bank; the others move on their own to another branch. The design of Section 4.1 does not depend on who chooses the absorbing branch, nor does it require that closures be random. It requires that the shocks behind a closure do not affect the cohorts of the closing branch in proportion to their Δp_{cj}, which differs across them only through their industry.[33]

Data. The credit data are those of Section 3.1. Closures come from the Banque de France's register of bank branches (*Fichier des Implantations Bancaires*, FIB), which records for every branch its location, opening date and closing date, and for every clo-

[32] Source: ECB structural banking indicators.

[33] Which branches close is mostly a matter of which bank they belong to: bank fixed effects explain 32% of the variance of an indicator of closure between 2010 and 2017 across branches, département and urban unit fixed effects 3% and 8% on their own (Appendix E, Table E.1). A branch that is the only one of its bank in its urban unit is less likely to close, its rank by size within the bank and within the bank and urban unit predicts closure, and a branch in the top quartile of portfolio shares in some industry is not significantly more likely to close than others (Table E.2).

sure the date of the announcement, the reason (a merger or the bank's own decision) and, when the accounts are transferred within the bank, the identity of the absorbing branch. Before 2015 the history of closed branches is available only for banks that keep permanent account identifiers (*RIB permanent*); from 2015 it is exhaustive. The sample of the closure analysis consists of SMEs (fewer than 250 employees and turnover below EUR 50 million or total assets below EUR 43 million), excluding micro-enterprises, financial and real estate firms, public entities and non-resident firms, located in an urban unit in which at least one branch closes, and present in the credit register for at least eight consecutive quarters. We leave out the closures in which one bank renumbered its whole network of branches, in the second quarter of 2012. All its branch codes end in the first quarter of 2012 and are replaced by new codes in the second, and the new codes map almost one to one onto the old ones: 139 of the 141 new codes are the main destination of exactly one old code, which sends a median of 85% of its borrowers to a single new code. These events change the identifiers of branches rather than transfer borrowers from one branch to another, and the new branches have no lending, hence no p, before the announcement. They account for 109 of the 620 closing branches in the data.

Ratings come from the Banque de France's rating of firms (*cotation*). Grades 3++ (the best) to 6 rest on the firm's accounts. Grade 7 signals at least one significant payment incident, grades 8 and 9 a threatened or compromised solvency, and grade P collective proceedings (safeguard, receivership or liquidation). A firm is a future defaulter in a quarter if it is not rated 7 or worse in that quarter and is so rated in one of the next four quarters. Since the panel ends in 2017Q3, default within a year can be observed up to 2016Q3. The regressions of Table 7 use 605 to 1,807 cohorts of 185 to 208 closing branches, depending on the industry level, and about 2,800 firms; a closing branch has a median of three or four cohorts (Table E.3 in Appendix E).

The variation in specialization induced by closures. Figure E.3 in Appendix E shows the distribution of Δp_{cj} across all the industry cohorts with a Δp_{cj}, for markets of départements at NAF levels 0 to 3: 807 to 2,081 cohorts, depending on the industry level,

of which the regressions use 605 to 1,807. Δp_{cj} is negative on average, from -0.12 at NAF level 0 to -0.26 at NAF level 3. The closing branch lends to the industry of each of its cohorts by construction, whereas the main absorbing branch does not lend to the industry of 51 of the 807 cohorts at NAF level 0 and of 824 of the 2,081 cohorts at NAF level 3 (43 and 712 in the regression sample); for these, $p_{a(c)j}$ is the p of an empty cell. Closures nonetheless move cohorts in both directions: Δp_{cj} is positive for 36% of the cohorts at NAF levels 0 and 1, 29% at NAF level 2 and 23% at NAF level 3. The changes are large. Across all cohorts the standard deviation of Δp_{cj} is 0.32 to 0.38, depending on the industry level; within closures, the variation used by regression (10), it is 0.24 to 0.26 (Table 7).

4.3 Results

4.3.1 Trends before the announcement

The identification requirement of Section 4.1 concerns the shocks behind a closure, which we do not observe. One of its implications can be tested: before the announcement, when no account has yet been transferred, Δp_{cj} should not predict the trends in the cohorts' outcomes. We estimate event-study versions of regression (10), in which Δp_{cj} is interacted with indicators of years relative to the announcement, with cohort and closing branch $\times$ year fixed effects; the year before the announcement (quarters -10 to -7) is the reference. Figures E.4 and E.5 in Appendix E report the coefficients for new lending and for new lending to future defaulters at each industry level. In none of the eight regressions is a coefficient before the reference significant at the 5% level, and the hypothesis that the two are zero is never rejected: the p-values are 0.445 to 0.736 for new lending and 0.180 to 0.930 for new lending to future defaulters.

4.3.2 New lending and new lending to future defaulters

Table 7 reports regression (10) for the two outcomes, in markets of départements at NAF levels 0 to 3. New lending rises with the change in specialization at every industry level, significantly at NAF levels 2 and 3. A cohort whose lending branch becomes more spe-

cialized by one standard deviation of Δp_{cj} within closures (0.26 at NAF level 2) receives about a fifth more new credit than the average. New lending to future defaulters falls with the change in specialization by similar amounts at NAF levels 0 to 2, and does not change at NAF level 3.

Table 7: Branch closures, the specialization of the absorbing branch, new lending and new lending to future defaulters: industry cohorts

	New lending from the bank				New lending to future defaulters			
NAF level	0	1	2	3	0	1	2	3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δp_{cj} x Post	0.0154	0.0156	0.0211**	0.0193**	-0.0015*	-0.0026**	-0.0023*	0.0002
	(0.0176)	(0.0102)	(0.0099)	(0.0089)	(0.0008)	(0.0011)	(0.0012)	(0.0007)
Mean of dependent variable	0.0279	0.0276	0.0281	0.0282	0.0004	0.0004	0.0005	0.0005
SD of Δp_{cj} within closure	0.24	0.24	0.26	0.25	0.24	0.24	0.26	0.25
Cell-periods	1,210	1,586	2,660	3,614	1,210	1,586	2,660	3,614
Closing branches (clusters)	185	200	203	208	185	200	203	208
Cohort cell FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Closing branch x period FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column is one regression of an outcome of an industry cohort cj: the firms in industry j that borrow from closing branch c in quarter -7 relative to the closure, followed wherever they borrow, in two periods, pre (quarters -16 to -7) and post (-6 to $+8$). Regressor: Δp_{cj} x Post, where $\Delta p_{cj} = p_{a(c)j} - p_{cj}$ is the p in industry j of $a(c)$, the branch that receives most of c's reassigned borrowers, minus the p of c, both in quarter -7 (intention to treat; markets of départements). Only the closing branches' cohorts enter, with cohort fixed effects and closing branch x period fixed effects, so the coefficient compares the cohorts of different industries of the same closing branch. New lending: new medium- and long-term credit from the bank in the firm-quarters in which the borrower's default within a year is observed (quarters up to 2016q3, borrowers not yet rated 7 or worse), as a share of these firm-quarters' credit from the bank in quarter -7, per quarter. New lending to future defaulters: the part of this new credit that goes to borrowers that receive a Banque de France rating of 7 or worse (7, 8, 9 or P) within a year. The ratio of the two is the share of the new credit that goes to future defaulters. Mean of dependent variable: the closing branches' cohorts in the post period. SD of Δp_{cj} within closure: the standard deviation of the cohorts' Δp_{cj} around the mean of the cohorts of the same closing branch, in the post period of the regression sample, with the regression weights. Given the closing branch x period fixed effects, only this variation in Δp_{cj} enters the coefficient. Weights: the cohort's credit from the bank in quarter -7. Standard errors clustered by closing branch. Figure 7 reports the coefficients for 16 market definitions and two other measures of specialization. In Table E.6, the outcome of columns 5–8 is net of the default rate of new loans to borrowers with the same rating, industry and size.

Table 8 estimates separate coefficients for moves to a more and to a less specialized branch. For new lending to future defaulters, the hypothesis that the two coefficients are equal is not rejected at any industry level (p-values 0.119 to 0.962), and at NAF level 1, where the pooled estimate is most precise, they are nearly identical (-0.0027 and -0.0026). For new lending (Table E.5 in Appendix E), the two coefficients are imprecise and not significantly different (p-values 0.245 to 0.795).

Table 8: New lending to future defaulters: moves to a more and to a less specialized branch

	New lending to future defaulters			
NAF level	0	1	2	3
	(1)	(2)	(3)	(4)
Δp^{+}_{cj} x Post (gain)	-0.0028	-0.0027	-0.0051	0.0052
	(0.0029)	(0.0025)	(0.0033)	(0.0034)
Δp^{-}_{cj} x Post (loss)	-0.0007	-0.0026*	-0.0014	-0.0007
	(0.0012)	(0.0014)	(0.0012)	(0.0008)
p-value, gain = loss	0.588	0.962	0.285	0.119
Mean of dependent variable	0.0004	0.0004	0.0005	0.0005
Cell-periods	1,210	1,586	2,660	3,614
Closing branches (clusters)	185	200	203	208
Cohort cell FE	Yes	Yes	Yes	Yes
Closing branch x period FE	Yes	Yes	Yes	Yes

Notes: As Table 7, columns 5–8, with separate coefficients for moves to a more specialized branch, $\Delta p^{+}_{cj} = \max(\Delta p_{cj}, 0)$ ("gain"), and to a less specialized one, $\Delta p^{-}_{cj} = \min(\Delta p_{cj}, 0)$ ("loss"). The loss term is never positive, so both coefficients measure the change in new lending to future defaulters per unit of Δp_{cj}, and they are equal when the response is symmetric. p-value, gain = loss: the test that the two coefficients are equal (standard errors clustered by closing branch).

The decline in new lending to future defaulters does not come from a shift towards borrowers that are safer on observable characteristics. In Table E.6 in Appendix E, each firm's new credit is weighted by its default indicator minus the default rate of comparable new loans: loans to borrowers with the same Banque de France rating in the same year, and in addition the same industry, the same size, or both. The coefficients do not change: at NAF level 1 they lie between -0.0025 and -0.0027 in all five columns. Nor does the decline come from less new lending in proportion to the cohort's existing credit. Excess lending to future defaulters, the new credit to future defaulters beyond their share of the cohort's credit in quarter -7, also falls with Δp_{cj} at NAF levels 0 to 2 in every geography (Figure E.6 in Appendix E). The change concerns which firms of the cohort receive new credit, not only how much they receive.

4.3.3 Substitution across lenders

Table 9 reports the substitution regression (11) for urban units at NAF level 3. Column 1 includes only the post indicator: after the announcement, the firm's credit from the reas-

signed bank falls by 0.092 log points relative to its other lenders (standard error 0.014), about 9%. We do not read this decline as the effect of closing a branch, since banks choose which branches to close; it serves only as a benchmark for the magnitudes that follow. Column 2 adds the level of specialization of the relationship, which keeps only the relationships that have a p five to eight quarters earlier; we discuss its coefficient, $\gamma = 0.036$ (standard error 0.008), in Section 4.3.4. Column 3 adds the change in specialization. Its coefficient is 0.091 (standard error 0.049), significant at the 10% level: the firm's credit from the reassigned bank, relative to its other lenders, rises with the specialization of the new branch relative to the old one. Column 4 adds the interactions of the post indicator with the change in distance, the change in branch size and a long relationship with the closing bank, and the coefficient on the change in specialization is 0.088 (standard error 0.049). Column 5 keeps only the large changes in specialization, setting the change to zero when its absolute value is below 0.25; the coefficient is 0.074 (standard error 0.050).

A firm whose new branch is 0.1 higher in p than its old one obtains about 0.9% more credit from the bank, relative to its other lenders, than a firm whose branch's specialization does not change, about a tenth of the average decline after a closure in column 1. Firms therefore shift part of their borrowing between the reassigned bank and their other lenders in the direction of the change in specialization: towards the bank when the new branch is more specialized in their industry, away from it when it is less. The direction is that of the cohort estimates for new lending (Section 4.3.2). Of the three controls, only the change in branch size enters significantly, at the 5% level: the firm obtains less credit from the bank, relative to its other lenders, when the absorbing branch is larger than the closing one (-0.087 per log point of branch credit, standard error 0.044). With these interactions, the post indicator measures the change for relationships shorter than 12 quarters whose distance and branch size do not change, and it is not significant (0.062, standard error 0.065).

Table E.7 in Appendix E estimates separate coefficients for moves to a more specialized branch and to a less specialized one. In the specification of column 4 they are 0.025 (standard error 0.083) and 0.140 (standard error 0.085, significant at the 10% level), and

Table 9: Branch closures, the specialization of the firm's lender and credit from that lender: firm–bank regression

	Total Credit$_{bft}$				
	(1)	(2)	(3)	(4)	(5)
Closure x Post	-0.092***	-0.052***	-0.049***	0.062	0.062
	(0.014)	(0.016)	(0.016)	(0.065)	(0.065)
Specialization		0.036***	0.035***	0.035***	0.035***
		(0.008)	(0.008)	(0.008)	(0.008)
Closure x Post x Δ specialization			0.091*	0.088*	
			(0.049)	(0.049)	
Closure x Post x Long relation				-0.109	-0.109
				(0.068)	(0.068)
Closure x Post x Δ distance				0.000	0.000
				(0.001)	(0.001)
Closure x Post x Δ size				-0.087**	-0.087**
				(0.044)	(0.044)
Closure x Post x Large Δ specialization					0.074
					(0.050)
Firm x Bank FE	Yes	Yes	Yes	Yes	Yes
Firm x Quarter FE	Yes	Yes	Yes	Yes	Yes
GEA x Quarter FE	Yes	Yes	Yes	Yes	Yes
Observations	3919413	2707268	2707268	2707268	2707268
R-Square	0.90	0.91	0.91	0.91	0.91

Notes: Regression (11). Dependent variable: log credit (drawn credit and undrawn credit lines) of the bank to the firm. Multi-bank firms. Urban units at NAF level 3. "Closure x Post": post indicator, from the announcement to eight quarters after the closure. "Δ Specialization": p of the absorbing branch minus p of the closing branch in the firm's industry, equation (9); the closing branch's p is the mean over the fifth to eighth quarters before the announcement, the absorbing branch's is the p of its share of lending in the firm's industry in the quarter before the announcement, with no relationship left out (the exact p of an empty cell where the branch lends in the market but not to the firm's industry, 1/2 where it does not exist or does not lend); missing changes are set to zero. "Specialization": p of the relationship, mean over the fifth to eighth quarters before. Changes in distance (km between the firm and its branch) and size (log of the branch's total credit): after minus before the reassignment. Long relation: at least 12 quarters with the closing bank at the closure. Column 5: the change in specialization set to zero when its absolute value is below 0.25. Standard errors clustered by the firm's urban unit and year.

the hypothesis that they are equal is not rejected (p-value 0.402). The point estimates place most of the response in moves to a less specialized branch, but they are too imprecise to tell the two directions apart.

4.3.4 Alternatives to our approach

We now ask how the estimates of Section 4.3.2 change when each element of our approach is replaced by an alternative: measures of specialization built from portfolio shares in place of p, and broader or narrower markets. Figure 7 reports the coefficients of regression (10) for each of the 16 market definitions of Section 3.1 (four geographies of the branch crossed with four NAF levels), and for three measures of the specialization of the closing and the absorbing branch: p; the indicator that the branch's portfolio share in the industry is in the top quartile of its market, the rule of Paravisini et al. (2023); and the indicator that the industry is the branch's favorite industry, the one with its largest excess share (Blickle et al., 2026). Since p ranges from 0 to 1 and the two indicators take the values 0 and 1, each coefficient measures the difference between the bottom and the top of its measure.

The measure of specialization. The estimates are clearest with p. With p, new lending to future defaulters falls with the change in specialization at NAF levels 1 and 2 in every geography, by 0.0022 to 0.0027 per unit, and new lending rises at NAF level 2 in every geography, by 0.018 to 0.021. With the top-quartile indicator, 2 of the 16 estimates for new lending to future defaulters and 1 of the 16 for new lending are significant at the 10% level; with the favorite-industry indicator, 5 and none. The significant estimates for new lending to future defaulters are smaller with the indicators, between -0.0003 and -0.0014. A researcher who measured specialization with either indicator would find much weaker evidence that the specialization of the lending branch matters for how new credit is allocated.

Both indicators rank portfolio shares, so they flag branches as specialized in every market, whether or not any branch is: the top-quartile rule flags the branches with the highest shares of each industry, and every branch has a favorite industry. At NAF level 3,

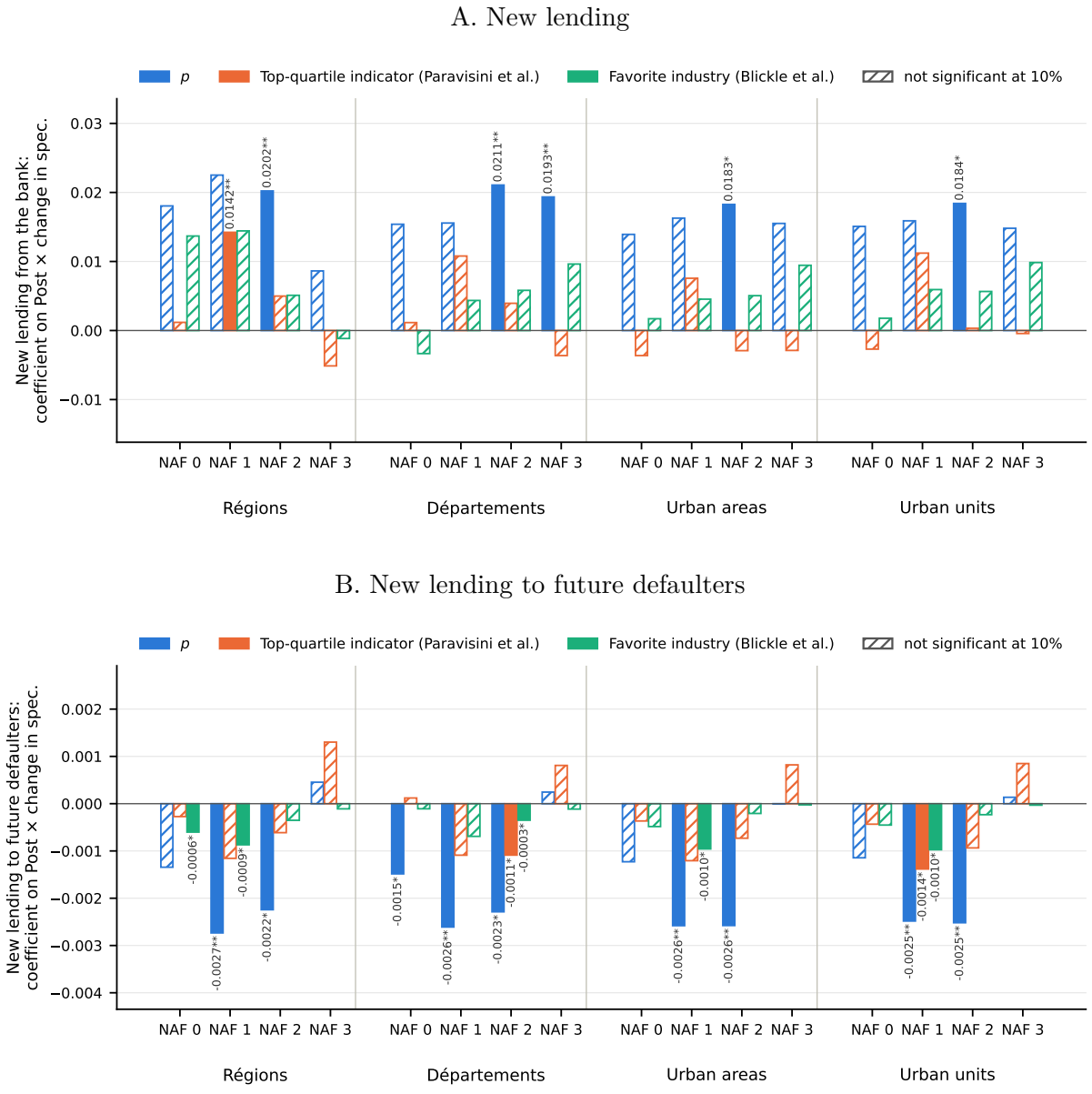


Figure 7: Change in the specialization of the lending branch, new lending and new lending to future defaulters: industry cohorts, by market definition and measure of specialization

Notes: Coefficients on the change in specialization x Post in the regression of Table 7 for 16 market definitions (geography of the branch × NAF level) and three measures of specialization: p, the top-quartile indicator of Paravisini et al. (2023) and the favorite-industry indicator of Blickle et al. (2026). Panel A: new lending (Table 7, columns 1–4). Panel B: new lending to future defaulters (columns 5–8). The bars of p in départements are the estimates of Table 7. With the two indicators, the change in specialization is the indicator of $a(c)$ in industry j minus that of c, both in quarter -7, in the markets in which p is measured; the indicator of $a(c)$ is zero where it does not lend to industry j. Same cohorts, fixed effects and weights as in Table 7. Solid bars: significant at least at the 10% level, labelled with the coefficient and its stars; hatched bars: not significant. Standard errors clustered by closing branch. Figure E.6 reports excess lending to future defaulters.

moreover, the branches selected by the top-quartile rule look no more specialized than those it selects under random matching in the typical cell of every geography: their mean relative portfolio share exceeds its random-matching value in 49% of the cells of urban units, 47% of urban areas, 37% of départements and 30% of régions (Section 3.1). A change in either indicator between the closing and the absorbing branch is then a noisy measure of the change in their specialization, which biases its coefficient towards zero.

The definition of markets. With p, the geography matters little. At a given industry level, the four geographies give estimates of similar size and significance for both outcomes (Figure 7): new lending to future defaulters falls at NAF levels 1 and 2 in every geography, and new lending rises at NAF level 2 in every geography. Geography matters little for a reason of construction: since markets are defined by the location of the branch (Section 2.3), a broader geography changes the branches against which a branch's portfolio is compared, but not its portfolio share, whereas a broader industry changes the share itself. The industry level matters more. New lending to future defaulters responds at NAF levels 1 and 2 in every geography, at NAF level 0 only with départements, and not at NAF level 3, the finest; new lending responds at NAF level 2 in every geography and at NAF level 3 with départements. The absence of an estimate for new lending to future defaulters at NAF level 3 is consistent with the attenuation by noise in p described in Section 4.1, which is strongest where cells are sparsest and p has the least power (Section 3.2.2).

The source of the variation. The substitution regression (11) also estimates γ, the coefficient on the level of specialization of the firm's current lender, p_{fkt}. This coefficient is what a regression without a research design measures: how credit co-moves with the specialization of the lender, holding fixed the relationship (firm–bank effects), everything common to the firm's lenders in a quarter (firm–quarter effects) and shocks to the bank's group. It is the kind of evidence on which studies without a source of variation in specialization rely. Blickle et al. (2026, pp. 1535, 1550), for example, regress loan terms on the bank's excess share in the borrower's industry with industry–time, bank–time and,

in some specifications, firm–time fixed effects, and interpret the estimates as correlations rather than causal effects.

References

Acharya, V. V., I. Hasan, and A. Saunders (2006). Should banks be diversified? Evidence from individual bank loan portfolios. *Journal of Business 79*(3), 1355–1412.

Armenter, R. and M. Koren (2014). A balls-and-bins model of trade. *American Economic Review 104*(7), 2127–51.

Autorité de la concurrence (2009). Décision n° 09-DCC-16 du 22 juin 2009 relative à la fusion entre les groupes Caisse d'Épargne et Banque Populaire.

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage. *The Manchester School 33*(2), 99–123.

Basel Committee on Banking Supervision (2023). Basel Framework, SRP32: Credit Risk. Bank for International Settlements. Version effective as of January 1, 2023.

Beck, T., O. De Jonghe, and K. Mulier (2022). Bank sectoral concentration and risk: Evidence from a worldwide sample of banks. *Journal of Money, Credit and Banking 54*(6), 1705–1739.

Blickle, K., C. Parlatore, and A. Saunders (2026). Specialization in banking. *The Journal of Finance 81*(3), 1531–1572.

Board of Governors of the Federal Reserve System (2025). BancFirst Corporation: Order Approving the Acquisition of a Bank Holding Company and the Merger of Bank Holding Companies. FRB Order No. 2025-15. October 3, 2025.

Böve, R., K. Düllmann, and A. Pfingsten (2010). Do specialization benefits outweigh concentration risks in credit portfolios of German banks? Discussion Paper, Series 2 10/2010, Deutsche Bundesbank.

Carrington, W. J. and K. R. Troske (1997). On measuring segregation in samples with small units. *Journal of Business & Economic Statistics 15*(4), 402–409.

Casado, A. and D. Martínez-Miera (2024). Local lending specialization and monetary policy. Working Paper 2440, Banco de España.

Casado, A. and D. Martínez-Miera (2025). Banks' specialization and private information. Working Paper 2539, Banco de España.

Cowles, 3rd, A. (1933). Can stock market forecasters forecast? *Econometrica 1*(3), 309–324.

De Jonghe, O., H. Dewachter, K. Mulier, S. Ongena, and G. Schepens (2020). Some borrowers are more equal than others: Bank funding shocks and credit reallocation. *Review of Finance 24*(1), 1–43.

Degryse, H., O. De Jonghe, L. Gambacorta, and C. Huylebroek (2024). Bank specialization and corporate innovation. Working Paper 458, National Bank of Belgium.

Di, W. and N. Pattison (2023). Industry specialization and small business lending. *Journal of Banking and Finance 149*, 106797.

Drechsler, I., A. Savov, and P. Schnabl (2017). The deposits channel of monetary policy. *Quarterly Journal of Economics 132*(4), 1819–1876.

Dursun-de Neef, H. O. (2023). Bank specialization, mortgage lending and house prices. *Journal of Banking and Finance 151*, 106836.

Ellison, G. and E. L. Glaeser (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy 105*(5), 889–927.

Federal Deposit Insurance Corporation (2026). Merger Transactions. *Federal Register 91*(182), 60196–60235. Proposed rule, September 22, 2026, RIN 3064-AG18.

Federico, S., F. Hassan, and V. Rappoport (2025). Trade shocks and credit reallocation. *American Economic Review 115*(4), 1142–1169.

Fisman, R., D. Paravisini, and V. Vig (2017, February). Cultural proximity and loan outcomes. *American Economic Review 107*(2), 457–92.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

Gaubert, C. and O. Itskhoki (2021). Granular comparative advantage. *Journal of Political Economy 129*(3), 871–939.

Giometti, M. and S. Pietrosanti (2022). Bank specialization and the design of loan contracts. Working Paper 2022-14, FDIC Center for Financial Research.

Gopal, M. (2021). How collateral affects small business lending: The role of lender specialization. Working Paper 21-22, U.S. Census Bureau, Center for Economic Studies.

INSEE (2020). Unité urbaine. Definition, https://www.insee.fr/fr/metadonnees/definition/c1501. Zonings of 2010 and 2020; last updated 9 December 2020; accessed 26 October 2026.

INSEE (2025). Aire urbaine. Definition, https://www.insee.fr/fr/metadonnees/definition/c2070. Zoning of 2010, replaced by the aires d'attraction des villes in 2020; last updated 28 July 2025; accessed 26 October 2026.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association 56*(294), 223–234.

Loutskina, E. and P. E. Strahan (2011). Informed and uninformed investment in housing: The downside of diversification. *Review of Financial Studies 24*(5), 1447–1480.

Nguyen, H.-L. Q. (2019). Are Credit Markets Still Local? Evidence from Bank Branch Closings. *American Economic Journal: Applied Economics 11*(1), 1–32.

Niepmann, F. and T. Schmidt-Eisenlohr (2017). No guarantees, no trade: How banks affect export patterns. *Journal of International Economics 108*, 338–350.

Paravisini, D., V. Rappoport, and P. Schnabl (2023). Specialization in bank lending: Evidence from exporting firms. *The Journal of Finance 78*(4), 2049–2085.

Petersen, M. A. and R. G. Rajan (1995). The effect of credit market competition on lending relationships. *Quarterly Journal of Economics 110*(2), 407–443.

Petersen, M. A. and R. G. Rajan (2002). Does distance still matter? The information revolution in small business lending. *Journal of Finance 57*(6), 2533–2570.

Sackrowitz, H. and E. Samuel-Cahn (1999). P values as random variables—expected P values. *The American Statistician 53*(4), 326–331.

Saidi, F. and D. Streitz (2021). Bank concentration and product market competition. *Review of Financial Studies 34*(10), 4999–5035.

Simoens, M. and F. Tamburrini (2027). The expert's edge? Bank lending specialization and informational advantages for credit risk assessment. *Journal of Corporate Finance 102*, 103081.

Tabak, B. M., D. M. Fazio, and D. O. Cajueiro (2011). The effects of loan portfolio concentration on Brazilian banks' return and risk. *Journal of Banking and Finance 35*(11), 3065–3076.

United States v. Philadelphia National Bank (1963). 374 U.S. 321. Supreme Court of the United States.

U.S. Census Bureau (2022). Urban Area Criteria for the 2020 Census—Final Criteria. *Federal Register 87*(57), 16706–16715. March 24, 2022, document 2022-06180.

U.S. Department of Justice (2024). 2024 Banking Addendum to 2023 Merger Guidelines. September 17, 2024.

U.S. Department of Justice and federal banking agencies (1995). Bank Merger Competitive Review: Introduction and Overview. Withdrawn by the Department of Justice on September 17, 2024.

U.S. Department of Justice and Federal Trade Commission (2023). Merger Guidelines. December 18, 2023.

A Construction of the random-matching percentile p

In this appendix we describe, step by step, how we construct the random-matching percentile p, the probability that random matching would give a branch a lower share of its lending in a firm's industry than the share it has (Section 2.4), from loan-level credit register data. The procedure requires only data that record, for each quarter, the lender (at the branch level), the borrower, the amount of credit, the borrower's industry, and the branch's location. Anyone with data of this form can replicate it.

The procedure is run separately for each geographic market and quarter. To keep notation simple, we suppress the market and quarter subscripts throughout: every object below is computed within a single market m in a single quarter t, except the size-matching rates of Steps 4 and 5, which are estimated within the branch's département in the same quarter.

A.1 Inputs

Step 1: Lending relationships. The unit of observation is a *lending relationship* $r = (f, b)$ between firm f and bank branch b in quarter t. If a firm has several loans with the same branch, we add them up and treat them as a single relationship. Let $L_r > 0$ denote the amount of credit in relationship r.

Each relationship carries the industry of the borrowing firm at four levels of the French industry classification (NAF rev. 2), from coarse to fine: $j_1(r), \ldots, j_4(r)$. The steps below are identical for every industry level, so we write $j(r)$ for the industry of r at a given level.

Step 2: Geographic markets. A market m is a geographic unit, and it contains every relationship of every branch located in that unit, regardless of where the borrowing firm is located. We use four geographic partitions of metropolitan France, from finest to coarsest: urban units (*unités urbaines*), urban areas (*aires urbaines*), départements, and the 22 pre-2016 régions. A branch's urban unit, urban area, département and région are those of the municipality where it is located. We exclude the branches located outside metropolitan France. Section 3 also treats metropolitan France as a single market, as a benchmark for the four partitions.

Let $\mathcal{B}$ be the set of branches in the market, with $B = |\mathcal{B}|$, and $\mathcal{R}$ the set of relationships, with $N = |\mathcal{R}|$. Markets with a single branch ($B = 1$) or a single relationship ($N = 1$) are dropped: in these markets any reassignment reproduces the observed data, so the counterfactual distribution is degenerate and p is not defined.

A.2 Observed portfolio shares

Step 3: Observed and leave-one-out shares. For each branch b, let $V_b = \sum_{r \in \mathcal{R}_b} L_r$ be its total lending in the market, where $\mathcal{R}_b$ is the set of relationships of branch b, and let $A_{bj} = \sum_{r \in \mathcal{R}_b : j(r) = j} L_r$ be its lending to industry j. The portfolio share of branch b in industry j is

$$S_{bj} = \frac{A_{bj}}{V_b}. \tag{A.1}$$

The measure is computed for each relationship r of branch b with a firm in industry $j = j(r)$, using the branch's portfolio share *excluding* r:

$$S_{bj}^{-r} = \frac{A_{bj} - L_r}{V_b - L_r}. \tag{A.2}$$

This leave-one-out share measures how much of branch b's lending goes to firm f's industry based only on the branch's *other* borrowers, so the firm's own borrowing does not mechanically raise the measured specialization of its lender.

A.3 Counterfactual portfolio shares

The counterfactual randomly reassigns relationships (the "balls") to branches in the same market (the "bins"). Assignment probabilities are proportional to the branches' lending volumes, adjusted for the observed tendency of large branches to hold large relationships, but they do not depend on the borrower's industry.

Step 4: Size classes and matching rates. Large branches tend to hold large relationships. The natural way to reproduce this pattern is to measure it within the market itself, by grouping the market's branches and relationships into size classes and measuring how much of each relationship class each branch class holds. Random matching with rates measured in this way keeps the structure of the market fixed: in expectation, every branch receives as much lending as it actually holds (see "Size matching and market structure" after Step 5). In sparse markets, however, size classes formed within the market hold few relationships. A market is often dominated by a single branch, which then forms a size class of its own, and the counterfactual sends back to it relationships of the sizes it actually holds. We therefore measure the pattern in each département δ and quarter, pooling all branches located in the département (Step 2) and all their relationships. The cost of this choice is that the counterfactual keeps every branch's expected lending exactly equal to its actual lending only when the market is a département. We measure the distortion in other markets after Step 5.

Relationships are grouped into size classes $q = 1, \ldots, 10$ that each hold about one tenth of the département's *lending*.[34] Rank the relationships of the département's branches

[34] We considered several ways of forming these classes. The simplest, deciles of the *number* of relationships, puts most of the lending in the top class (a median of 56% of a département's lending), which then mixes relationships of very different sizes, while the bottom classes, which barely affect portfolio shares, take up most of the resolution. Classes that hold equal shares of lending put the resolution where the lending is. We compared the two, and several variants, in simulations on the market structure of the credit register file: relationships were reassigned to the branches of their market by a known process, with or without size matching, so that the correct p is known. When size matching is present, lending classes roughly halve the bias in the mean p of the largest branch of each market relative to count classes (-0.013 against -0.027); when it is absent, both introduce a similar spurious matching ($+0.040$ against $+0.038$). Imposing at least 30 relationships per class, or splitting the top count decile at the 99th and 99.9th percentiles, gave nearly the same results. Placing each relationship with matching rates estimated without it (leave-one-out) reduced the spurious matching only partly ($+0.031$): the link between a branch's size and the sizes of its relationships is collective, since a branch that holds many large relationships is large because of all of them and not of any single one. Classifying branches by their number of relationships reduced the spurious matching more ($+0.018$, with count classes for relationships) but not the bias when size matching is present. We keep the simplest rule. Its cost is that the top class is small (a median of 9 relationships, held by a median of 3.5 branches) and is often held entirely by the top branch class (in

by amount, and let $F_\delta(\ell)$ be the share of the département's lending held in relationships of amount at most ℓ. For $k = 1, \ldots, 9$, let ℓ_k be the smallest amount at which F_δ reaches $k/10$, and set $\ell_0 = 0$ and $\ell_{10} = \infty$. Relationship r is in class q if $\ell_{q-1} < L_r \leq \ell_q$. Relationships with the same amount are in the same class. When a single relationship holds a large share of the lending, several bounds can coincide, and the département then has fewer than ten classes. The bottom class therefore contains many small relationships, and the top class a few large ones. Branches are grouped into size classes $c = 1, \ldots, 10$ given by the deciles of total branch lending V_b across the département's branches, so that each class contains a tenth of the branches (fewer classes when several deciles coincide because of ties).

Let V^δ_{cq} be the lending in relationships of size class q held by branches of size class c, and let V^δ_c, V^δ_q and V^δ be the corresponding totals by branch class, by relationship class, and overall. The matching rate between branch class c and relationship class q is

$$w_\delta(c, q) = \frac{V^\delta_{cq}/V^\delta_q}{V^\delta_c/V^\delta} : \tag{A.3}$$

the share of lending in size class q held by branches of class c, relative to the share of all lending held by those branches. It equals one for every (c, q) when relationship size is unrelated to branch size, and exceeds one for the size classes held disproportionately by branches of class c.

Because each département pools many branches and relationships, the matching rates depend little on the portfolio of any single branch (the top relationship class, discussed in the footnote above, aside), and they are the same whichever geographic partition defines the market. The opposite choice, a single set of matching rates for the whole country, would pool even more relationships, but it measures branch size against all the branches in France rather than against those of the same département.[35]

Step 5: Assignment probabilities. For branch b, let $\delta(b)$ be its département, $c(b)$ its size class in that département, and $q_{\delta(b)}(r)$ the size class of relationship r under the class bounds ℓ_k of département $\delta(b)$. In market m, relationship r is assigned to branch $b \in \mathcal{B}$ with probability

$$\pi_b(r) = \frac{V_b \; w_{\delta(b)}\Big(c(b),\, q_{\delta(b)}(r)\Big)}{\sum_{b' \in \mathcal{B}} V_{b'} \; w_{\delta(b')}\Big(c(b'),\, q_{\delta(b')}(r)\Big)}. \tag{A.4}$$

Without size matching ($w = 1$), a relationship goes to each branch in proportion to the branch's lending. The matching rates tilt these probabilities towards branches whose size

69 of 94 départements), so that the counterfactual assigns the largest relationships only to the largest branches. This may reflect a genuine constraint, if only large branches can extend the largest loans, and is in part mechanical, since these branches are large partly because they hold these relationships. A single relationship forms a class of its own in 9 départements.

[35] We considered estimating the matching rates at the national level and decided against it. With size classes that hold equal shares of lending, the largest relationship classes are then held almost exclusively by the branches in the national top decile of lending (1,999 of 19,984 branches, which hold 86% of all lending): these branches hold all the lending in the three largest classes (1,787 relationships, 30% of lending) and at least 96% of the lending in the six largest (60% of lending). The counterfactual would then assign these relationships almost only to branches that are large by national standards, and in a market with a single such branch it would send them back to that branch with near certainty. At the département level, by contrast, only the top class is held entirely by the top branch class, in 69 of 94 départements (footnote above).

class, in their département, holds more than its share of lending in r's size class. When all the branches of the market lie in one département, the size classes are common to all branches and $\pi_b(r)$ depends on r only through its size class. This is the case for départements themselves and for most urban units and urban areas (97% and 95% of those with at least two branches), but not for the largest ones, such as Paris: urban units and urban areas whose branches lie in several départements hold 35% and 50% of relationships. The probabilities depend on the branch and on the size of the relationship, but not on the borrower's industry.

Size matching and market structure. Random matching should change only which branch lends to which industry, and keep the rest of the market as it is. The probabilities (A.4) add up to one over the branches of the market for every relationship, so every relationship is allocated. Branch b's expected lending under random matching is

$$E\big[\tilde{V}_b\big] = \sum_{r\in\mathcal{R}} \pi_b(r) L_r = V_b \sum_{r\in\mathcal{R}} \frac{L_r\, w_b(r)}{\sum_{b'\in\mathcal{B}} V_{b'}\, w_{b'}(r)}, \tag{A.5}$$

where $w_b(r) = w_{\delta(b)}\Big(c(b), q_{\delta(b)}(r)\Big)$ is the matching rate of branch b for relationship r. Without size matching ($w = 1$), the sum is $\sum_r L_r / \sum_{b'} V_{b'} = 1$, and $E[\tilde{V}_b] = V_b$ for every branch.

With size matching, $E[\tilde{V}_b] = V_b$ still holds when the matching rates are estimated on the relationships of the market itself. All branches then share one set of classes, and with V_c, V_q, V_{cq} and V the market's own totals, as in (A.3), the denominator in (A.5) is the same for all relationships of class q: $\sum_c V_c\, w(c,q) = V \sum_c V_{cq}/V_q = V$. Hence $E[\tilde{V}_b] = (V_b/V) \sum_q V_q\, w(c(b),q) = V_b \sum_q V_{c(b)q}/V_{c(b)} = V_b$. The same algebra applies to a département with département rates, since the market is then the population on which the rates are estimated. In any other market the market's own totals differ from the département's. For a market m that lies within one département,

$$\frac{E\big[\tilde{V}_b\big]}{V_b} = \sum_q \frac{V_q^m}{V^m} \frac{w(c(b),q)}{\bar{w}_m(q)}, \qquad \bar{w}_m(q) = \sum_c \frac{V_c^m}{V^m}\, w(c,q),$$

where the superscript m denotes the market's totals. The ratio equals one for every branch only if the market's mix of branch and relationship size classes reproduces the département's. When the branches of a market belong to different départements, each branch's rates also come from a different table, with different class bounds, and (A.5) combines them.

Table A.1 reports the ratio $E[\tilde{V}_b]/V_b$ by market definition. In urban units and urban areas, the largest branch of a market keeps close to its lending (a median ratio of 0.99 in both; the 10th percentile is 0.82 and 0.76). The other branches deviate more: the median ratio is 0.99 in both, but one in ten receives less than 57% and 48% of its lending, and one in ten more than 145% and 143%. About a quarter of branches (26% and 29%) fall outside the band from 0.8 to 1.25, but they are small and hold 5% and 6% of lending. In régions and in France as a single market, which pool branches of many départements, the distortion is larger: the branches outside the band hold 12% and 17% of lending.

The distortion can be removed by *raking*: replace V_b in (A.4) by $\beta_b V_b$ and choose the factors β_b by iterative proportional fitting, so that every branch's expected lending equals

Table A.1: Expected lending under random matching relative to actual lending, 2016Q1

		Largest branch			Other branches			Outside $[0.8, 1.25]$	
Geography	Markets	10%	50%	90%	10%	50%	90%	Branches	Lending
France	1		0.98		0.42	0.74	1.63	71%	17%
Regions	21	0.91	0.94	0.98	0.51	0.93	1.52	60%	12%
Départements	94	1.00	1.00	1.00	1.00	1.00	1.00	0%	0%
Urban areas	704	0.76	0.99	1.02	0.48	0.99	1.43	29%	6%
Urban units	1,238	0.82	0.99	1.02	0.57	0.99	1.45	26%	5%

Notes: Ratio $E[\tilde{V}_b]/V_b$, equation (A.5), with the département matching rates of Steps 4 and 5, for every branch of every market with at least two branches. Largest branch: the branch with the most lending in its market (France has a single one). 10%, 50%, 90%: percentiles across branches. Outside $[0.8, 1.25]$: share of branches with a ratio outside this band, and their share of lending. With rates estimated within each market, the ratio is one for every branch.

V_b while the matching rates keep the same relative weight on relationships of different sizes. In our data, the iteration converges in every urban unit, in at most 82 iterations, and changes the assignment probabilities by a median of about 4%. In the simulations with a known assignment process of the footnote to Step 4, with size matching, raking brings p closer to its correct value relationship by relationship (a mean absolute difference of 0.070, against 0.084 without raking), but it does not reduce the bias in the mean p of the largest branch (-0.022, against -0.013). We do not apply it here, to keep the construction of the counterfactual simple.

Step 6: Random reassignment. For each draw $d = 1, \ldots, D$, with $D = 1{,}000$, every relationship $r \in \mathcal{R}$ is assigned independently to a branch, drawn with probabilities $\{\pi_b(r)\}_{b \in \mathcal{B}}$. Relationships are drawn with replacement: a branch can receive any number of relationships in a given draw. Each relationship keeps its amount L_r and its industry codes. The same draws are used for all four industry levels.

Step 7: Counterfactual leave-one-out shares. In draw d, let $A_{bj}^{(d)}$ and $V_b^{(d)}$ denote the lending of branch b to industry j and in total, computed over the relationships assigned to b. For relationship r with observed branch b and industry $j = j(r)$, the counterfactual share is branch b's share in industry j among the relationships *other than* r assigned to it in draw d:

$$S_{bj}^{(d),-r} = \begin{cases} \dfrac{A_{bj}^{(d)} - L_r}{V_b^{(d)} - L_r} & \text{if } r \text{ is assigned to } b \text{ in draw } d, \\ \dfrac{A_{bj}^{(d)}}{V_b^{(d)}} & \text{otherwise.} \end{cases} \tag{A.6}$$

This treats the observed share (A.2) and the counterfactual shares symmetrically: in both, relationship r is excluded from branch b's portfolio.

The counterfactual distribution for relationship r consists of the draws in which

branch b receives at least one relationship other than r:

$$\mathcal{D}_r = \Big\{ d : \; V_b^{(d)} - L_r \, \mathbb{1}\{r \text{ is assigned to } b \text{ in draw } d\} > 0 \Big\}. \tag{A.7}$$

In the other draws the counterfactual share (A.6) is 0/0, and they are discarded. This mirrors the observed share (A.2), which is defined only when branch b has relationships other than r (case 2 of Step 8). Draws in which branch b receives other relationships, but none in industry j, are *kept*. In these draws the counterfactual share is zero, a well-defined value, and the observed share can be zero too. We explain after Step 8 why this matters.

A.4 The measure p

Step 8: Percentile of the observed share. The specialization measure of relationship r is the fraction of valid counterfactual draws in which branch b's share in industry $j(r)$ is below its observed share, where draws in which the two shares are equal (ties) count one half:

$$p_r = \frac{1}{|\mathcal{D}_r|} \sum_{d \in \mathcal{D}_r} \left[\mathbb{1}\left\{ S_{bj}^{(d),-r} < S_{bj}^{-r} \right\} + \frac{1}{2} \, \mathbb{1}\left\{ S_{bj}^{(d),-r} = S_{bj}^{-r} \right\} \right], \qquad p_r \in [0,1]. \tag{A.8}$$

In words, p_r is the mid-percentile of the observed share within the counterfactual distribution. A value close to one means that the branch's share of lending in the firm's industry is higher than almost all the shares produced by random matching, given market structure; a value close to zero, that it is lower. For a branch that is not specialized, the expected value of p_r is 1/2 in every market (see below).

Three cases require comment.

1. If firm f is the only borrower of branch b in industry j, the observed leave-one-out share is zero. No draw is below it, and the ties are the valid draws in which branch b has no lending to industry j, so p_r is half the fraction of such draws. It is close to zero when random matching would almost always give branch b other borrowers in industry j, so that having none is unusual. It is close to 1/2 when random matching would rarely do so, and the observation then says little about specialization.

2. If r is the branch's only relationship ($V_b = L_r$), the observed leave-one-out share is not defined, and neither is p_r.

3. If r is the only relationship in industry j in the whole market, the observed share and every valid counterfactual share are zero, and (A.8) would give $p_r = 1/2$. The counterfactual distribution is then degenerate and carries no information about specialization, so we leave p_r undefined. Such cases become more frequent at finer geographic and industry partitions.

Why zero draws are kept and ties count one half. These two conventions give p the same benchmark in every market, for every branch and at every industry level: its expected value is 1/2 for a branch that is not specialized. Suppose that branch b is not specialized, in the sense of the counterfactual: its relationships other than r were assigned by the random matching of Steps 5 and 6. Its observed leave-one-out share is then itself a

draw of the counterfactual share, conditional on the same event as the valid draws (branch b has relationships other than r), and independent of the counterfactual draws. Denote the observed share by X and a counterfactual draw by X'. As the number of draws grows, p_r converges to $P(X' < X \mid X) + \frac{1}{2}P(X' = X \mid X)$, so

$$E[p_r] = P(X' < X) + \tfrac{1}{2}\,P(X' = X) = \tfrac{1}{2}. \tag{A.9}$$

The second equality holds because X and X' are independent and identically distributed, so that $P(X' < X) = P(X' > X)$, and the three probabilities $P(X' < X)$, $P(X' = X)$ and $P(X' > X)$ add up to one. No assumption on the shape of the counterfactual distribution is needed: (A.9) holds however sparse the market, with a mass of draws at zero, or with only a few distinct values. It does assume that the assignment probabilities $\pi_b(r)$ are those of the random matching. In practice they are estimated from the same data (branch lending V_b and the matching rates of Step 4), and the footnote to Step 4 measures the resulting shift in simulations with a known assignment process: a few hundredths for the largest branch of each market.

Both conventions are needed. Consider the alternative rule that discards the draws in which branch b has no lending to industry j and counts only the draws strictly below the observed share. An observed share of zero then gives $p_r = 0$, while a positive observed share is compared with positive draws only. Let $P(A > 0)$ be the probability that branch b's share in industry j is positive in a valid draw, and τ^+ the probability that two independent positive draws are equal. For a branch that is not specialized, the observed share is zero with probability $1 - P(A > 0)$, giving $p_r = 0$. Otherwise it is distributed as a positive draw, and by the argument above it exceeds an independent positive draw with probability $(1 - \tau^+)/2$. Under the alternative rule, therefore,

$$E[p_r] = P(A > 0)\,\frac{1 - \tau^+}{2}. \tag{A.10}$$

Both factors fall below one in sparse cells. When the branch can expect few relationships in the industry, its share is often zero, so $P(A > 0)$ is small, and it takes few distinct values, so τ^+ is large. The bias is large. In the running example of Section 2, (A.10) is 0.42 and 0.43 for the two relationships of Branch 3 in Services, 0.21 for the relationship of Branch 1 in Services, and 0.13 for its relationship in Manufacturing. In the 10-replica of the market it is between 0.48 and 0.50 for every relationship. Under (A.8), it is 0.5 in all these cases.

This bias would distort precisely the comparisons that p is meant to make.

1. It grows as markets are partitioned more finely, because finer partitions make cells sparser. Measured specialization would decline mechanically with the fineness of the partition: the mirror image of the mechanical increase in portfolio shares that motivates p (Section 2.2).

2. It differs across branches of the same market and industry. Small branches, which can expect fewer relationships in each industry, have a lower benchmark (0.21 for Branch 1 against 0.42–0.43 for Branch 3, both in Services). Fixed effects for market $\times$ industry cells would therefore not remove it.

3. In the branch-closure design of Section 4, a firm's account moves from a closing

branch to another branch of the same bank. When that branch is larger, the benchmark (A.10) rises with the move even if neither branch is specialized, and the change in p would partly reflect the change in branch size rather than in specialization.

Counting ties as one half while still discarding the draws with zero share would remove τ^+ but not $P(A > 0)$: by the same argument, the expected value would be $P(A > 0)/2$, which is 0.17 for Branch 1 in Manufacturing. Breaking ties at random, instead of counting them as one half, would make p exactly uniform for a branch that is not specialized, but it would add noise unrelated to the data. Since we read p as a probability over its whole range, and not as the outcome of a test with a cutoff, its expected value is the property we need.

The benchmark of 1/2 does not make p informative in sparse cells. When the counterfactual distribution has a large mass at zero or few distinct values, p takes few values, and even a strongly specialized branch may not have a p far above 1/2. In Appendix B we measure how much specialization p can detect in each cell.

Step 9: Diagnostics. For each relationship and industry level we also record the number of valid draws $|\mathcal{D}_r|$, the fraction of valid draws in which branch b's share in industry j is zero, and the quantiles and moments (mean, variance, skewness, kurtosis) of the counterfactual distribution $\{S_{bj}^{(d),-r}\}_{d\in\mathcal{D}_r}$, zero shares included. A small number of valid draws, a large fraction of zero shares, or a distribution concentrated on few distinct values signals that the market is too sparse at that partition for portfolio shares to be informative (see Section 2). These draws also give two of the diagnostics of Section 2.5, the probability of positive lending to the industry, $P(A > 0)$, and the probability of a tie. The AUC of p requires draws under specialized matching, which differ for every relationship (Appendix B).

Implementation. The procedure is implemented in Python, with a fixed random seed so that all draws can be reproduced exactly. It is run separately for each quarter from 2006Q1 to 2017Q3, for each of the four geographic partitions (and France as a single market) and the four industry levels.

B Diagnosing sparsity

The random-matching percentile p of Appendix A locates a branch's observed portfolio share within a counterfactual distribution obtained by random reassignment: it is the probability that random matching would give the branch a lower share than the one it has. When the market is sparse, that distribution is wide, lumpy, or concentrated at zero, and p carries little information about specialization. In this appendix we develop a diagnostic that tells, market by market, whether p can reveal specialization. We introduced these diagnostics with the running example in Section 2.5. In Section B.1 we define what the diagnostic must measure, give its exact expression, and approximate it under the assumption that portfolio shares are normally distributed. In Section B.2 we show that this assumption fails when loan amounts are skewed, and compute the diagnostic from the simulated distributions instead. In Section B.3 we apply it to French credit markets. The setting is the balls-and-bins model of Armenter and Koren (2014), in which each ball falls independently into a bin with a probability given by the bin's size. Here the balls are

relationships, the bins are branches, and a branch's size is its lending, adjusted for size matching (Appendix A, Step 5).

B.1 What we need, and the diagnostics under normality

The question. The measure p asks whether a branch lends more to an industry than random matching would make it lend. Random matching, however, produces a range of portfolio shares. When that range is wide, a branch that is specialized in an industry may still have a share often produced by random matching, and p will not single it out. The relevant question is therefore not whether a market has few firms or many industries, but how specialized a branch must be for p to reveal it. To answer it we need three things: the distribution of the branch's share when it is not specialized, which is the counterfactual distribution behind p; the distribution of its share when it is specialized to a given degree; and a criterion for when the second stands out from the first.

Notation. Fix a market and a quarter, a branch b, an industry j (at a given NAF level), and a relationship r_0 of branch b with a firm in industry j. We use the notation of Appendix A. *Leave-one-out.* By equation (A.6), the counterfactual share $S_{bj}^{(d),-r_0}$ is branch b's share in industry j computed over the relationships assigned to it in draw d *other than* r_0. Where r_0 itself is assigned does not affect it. The counterfactual distribution for r_0 is therefore the distribution of branch b's share when only the relationships in $\mathcal{R}^- = \mathcal{R} \setminus \{r_0\}$ are reassigned. Everything below is computed on $\mathcal{R}^-$, and we drop the superscript $-r_0$ to lighten notation.

Random variables. For $r \in \mathcal{R}^-$, let $X_r = \mathbb{1}\{r \text{ is assigned to } b\}$ and $\pi_r = \pi_b(r)$, the assignment probability of equation (A.4). Since relationships are assigned independently (Step 6), the X_r are independent Bernoulli variables with $P(X_r = 1) = \pi_r$. Let $\mathcal{J} = \{r \in \mathcal{R}^- : j(r) = j\}$ and $\mathcal{K} = \mathcal{R}^- \setminus \mathcal{J}$. In a draw, branch b's lending to industry j, its lending to other industries, and its total lending are

$$A = \sum_{r \in \mathcal{J}} L_r X_r, \qquad C = \sum_{r \in \mathcal{K}} L_r X_r, \qquad V = A + C, \tag{B.1}$$

and its counterfactual share is $S = A/V$.

Valid draws. Step 7 keeps the draws in which branch b receives at least one relationship, $V > 0$, including those in which it receives none in industry j, where $A = 0$ and $S = 0$. The counterfactual distribution from which p is computed is therefore the distribution of S *conditional on* $V > 0$, and from now on we write $P(\cdot)$ for probabilities conditional on a valid draw. Since $A > 0$ implies $V > 0$, the probability that branch b has positive lending to industry j in a valid draw is

$$P(A > 0) = \frac{1 - \prod_{r \in \mathcal{J}} (1 - \pi_r)}{1 - \prod_{r \in \mathcal{R}^-} (1 - \pi_r)}. \tag{B.2}$$

It is a diagnostic in its own right: when it is small, the counterfactual share is zero in most draws, and so is the observed share of a branch that is not specialized.

Specialized matching. The power analysis compares random matching with a second random counterfactual, *specialized matching* (Section 2.5), in which the branch is special-

ized to a known degree. We model specialization as a change in the composition of the branch's portfolio at an unchanged size. We fix a degree of specialization $\kappa \geq 1$, the factor by which the branch's expected lending to industry j exceeds what random matching gives it, and construct assignment probabilities that deliver it. Relationships are still assigned at random and independently, but branch b receives each relationship in $\mathcal{J}$ with higher odds and each other relationship with lower odds than under random matching, while its expected total lending is the same:

$$\pi_r^\kappa = \begin{cases} \dfrac{\theta \pi_r}{1 - \pi_r + \theta \pi_r} & r \in \mathcal{J}, \\ \dfrac{\lambda \pi_r}{1 - \pi_r + \lambda \pi_r} & r \in \mathcal{K}, \end{cases} \qquad \text{with} \quad \sum_{r \in \mathcal{R}^-} \pi_r^\kappa L_r = \sum_{r \in \mathcal{R}^-} \pi_r L_r, \quad \sum_{r \in \mathcal{J}} \pi_r^\kappa L_r = \kappa \sum_{r \in \mathcal{J}} \pi_r L_r. \tag{B.3}$$

The odds factors $\theta \geq 1$ and $\lambda \leq 1$ are common to all relationships in the industry and outside it, respectively, so specialization does not favor relationships of a particular size. The two conditions pin them down uniquely, because each sum is increasing in its factor. Since expected total lending is unchanged, κ is also the ratio of the share of industry j in the branch's expected lending, $\sum_{r \in \mathcal{J}} \pi_r^\kappa L_r / \sum_{r \in \mathcal{R}^-} \pi_r^\kappa L_r$, to the same share under random matching, the benchmark μ_{bj} of Section 2.4 computed on $\mathcal{R}^-$. It is the relative portfolio share (equation (2)) that the branch can expect under specialized matching, measured against the random-matching benchmark rather than against the market share s_j (the two benchmarks coincide when all relationships have the same assignment probability). At $\kappa = 1$, $\pi_r^\kappa = \pi_r$. Holding the branch's expected total lending at its value under random matching keeps specialized matching comparable with random matching, so that specialization means a different portfolio mix and not more lending. (With the size matching of Appendix A, this value equals the branch's actual lending exactly in départements and approximately in other markets; see "Size matching and market structure" in Appendix A.)

Maximum specialization. κ is bounded. A branch is at most fully specialized when it receives every relationship of industry j that it can receive, or, if these add up to more than its expected lending, when it lends only to industry j. Size matching rules out some relationships for some branches: when no branch of b's size class holds relationships of r's size class in b's département, the matching rate is zero and $\pi_r = 0$ (Appendix A, Step 4). The odds tilt leaves these probabilities at zero, so

$$\kappa_{\max} = \frac{\min\left(\sum_{r \in \mathcal{J}:\, \pi_r > 0} L_r,\ \sum_{r \in \mathcal{R}^-} \pi_r L_r\right)}{\sum_{r \in \mathcal{J}} \pi_r L_r}. \tag{B.4}$$

When the branch can receive no other relationship of industry j, the denominator is zero; we then set $\kappa_{\max} = 1$, since no specialization can show in its share. $\kappa_{\max}$ is close to one when industry j accounts for a large share of the branch's expected lending under random matching, which is the case for large industries and for branches that dominate their market. In such cells even complete specialization barely changes the portfolio share.

The AUC of p. Let S_0 be a draw of the branch's leave-one-out share under random matching, the counterfactual distribution of Appendix A (a valid draw), and S_κ an independent draw of the same share under specialized matching, with probabilities π_r^κ (also

given $V > 0$, the condition for the observed share to be defined). By definition (A.8), the expected value of p under specialized matching of degree κ is

$$\mathrm{AUC}(\kappa) = E_\kappa[p] = P\left(S_0 < S_\kappa\right) + \tfrac{1}{2} P\left(S_0 = S_\kappa\right). \tag{B.5}$$

$\mathrm{AUC}(\kappa)$ is the probability that the branch has a higher portfolio share under specialized matching of degree κ than under random matching, with ties counted one half: the area under the ROC curve (AUC) of p as a classifier of specialization of degree κ, as shown in Section 2.5, equation (7). At $\kappa = 1$ the two matchings coincide, S_1 and S_0 are identically distributed, so $\mathrm{AUC}(1) = 1/2$ exactly (equation (A.9)). $\mathrm{AUC}(\kappa)$ increases with κ.[36]

Summary statistics. We summarize the AUC by its value at a fixed degree of specialization, $\kappa = 2$: $\mathrm{AUC}(2)$ is the probability that p ranks the branch under specialized matching with $\kappa = 2$, where it expects to lend to industry j twice what random matching gives it, above the same branch under random matching, ties counted one half. Like any AUC, it equals $1/2$ when p does not separate the two and one when it separates them perfectly, and because it averages over all thresholds of p, it requires no cutoff defining a "specialized" branch. A doubling is a round degree of specialization, but how much of the room for specialization it uses differs across cells, so we report it with $2/\kappa_{\max}$. Because κ is the ratio of the branch's expected share in industry j to its expected share under random matching, and expected total lending does not change with κ, $2/\kappa_{\max}$ is the expected share of the branch under specialized matching with $\kappa = 2$, as a fraction of the largest it could have, at $\kappa_{\max}$, where the branch lends only to industry j or receives every relationship of the industry that it can receive (equation (B.4)). When $2/\kappa_{\max} > 1$, a doubling is infeasible and $\mathrm{AUC}(2)$ is not defined. We report the AUC at $\kappa = 1.5$ and $\kappa = 3$ as robustness checks.

Diagnostics. For each relationship (and hence each branch, industry, market and partition) we report:

1. the probability of positive lending to the industry in a valid draw, $P(A > 0)$, equation (B.2);
2. $\mathrm{AUC}(2)$, where a doubling is feasible;
3. $2/\kappa_{\max}$;
4. the *ceiling* $\mathrm{AUC}(\kappa_{\max})$, the expected p under specialized matching at $\kappa_{\max}$. Since $\mathrm{AUC}(1) = 1/2$ in every cell, the interval $[1/2, \mathrm{AUC}(\kappa_{\max})]$ is the range over which the expected p can move in the cell, and it requires no choice of parameters. It is the diagnostic of the cells in which a doubling is infeasible.

[36] As κ rises, θ rises and λ falls. Write $X_r = \mathbb{1}\{U_r < \pi_r^\kappa\}$ with U_r uniform and fixed, and compare $\kappa < \kappa'$ draw by draw. Every X_r with $r \in \mathcal{J}$ is non-decreasing in κ and every X_r with $r \in \mathcal{K}$ is non-increasing, so A rises and C falls. In a draw that is valid at both values, $S = A/(A + C)$ rises. A draw that is valid at κ' but not at κ had $A = C = 0$ at κ, so at κ' it has $C = 0$ and $A > 0$: its share is one, the largest possible value. A draw that is valid at κ but not at κ' has $A = 0$ at κ', hence also at κ: its share was zero, the smallest possible value. Moving from κ to κ' therefore raises every share and replaces draws of the smallest value by draws of the largest, so the distribution of S_κ given a valid draw is stochastically non-decreasing in κ. Since $\mathrm{AUC}(\kappa) = E\big[F_0(S_\kappa)\big]$, with $F_0(s) = P(S_0 < s) + \frac{1}{2}P(S_0 = s)$ non-decreasing, the AUC is non-decreasing in κ.

The exact AUC. All the diagnostics follow from $\mathrm{AUC}(\kappa)$, which has an exact expression. Write $P_0 = P(A > 0)$ for the probability of a positive share in a counterfactual draw, equation (B.2), and $P_\kappa = P_\kappa(A > 0)$ for the same probability with π_r^κ in place of π_r; let S_0^+ and S_κ^+ be the two shares conditional on being positive. A share of zero ties with the counterfactual draws of zero and is below every positive draw; a positive share is above every draw of zero. Conditioning on whether each share is positive therefore gives

$$\mathrm{AUC}(\kappa) = P_\kappa\,(1 - P_0) + P_\kappa\,P_0\,Q(\kappa) + \tfrac{1}{2}\,(1 - P_\kappa)(1 - P_0), \tag{B.6}$$

where

$$Q(\kappa) = P\Big(S_0^+ < S_\kappa^+\Big) + \tfrac{1}{2}\,P\Big(S_0^+ = S_\kappa^+\Big)$$

compares the positive parts of the two shares. At $\kappa = 1$, $P_\kappa = P_0$ and $Q(1) = 1/2$, and (B.6) equals $1/2$. Both S_0^+ and S_κ^+ have exact distributions. Let $\bar{P}$ denote probabilities not conditioned on a valid draw. For $s \in (0, 1)$, $S \le s$ if and only if $D_s = \sum_{r \in \mathcal{R}^-} L_r X_r \Big(\mathbb{1}\{r \in \mathcal{J}\} - s\Big) \le 0$, and $A = 0$ implies $D_s \le 0$. Since $A > 0$ implies $V > 0$,

$$P(S \le s \mid A > 0) = \frac{\bar{P}(D_s \le 0) - \bar{P}(A = 0)}{\bar{P}(A > 0)}. \tag{B.7}$$

D_s is a weighted sum of independent Bernoulli variables. Its distribution is determined by its characteristic function, $\bar{E}[e^{itD_s}] = \prod_{r \in \mathcal{R}^-} \left(1 - \pi_r + \pi_r e^{itL_r(\mathbb{1}\{r \in \mathcal{J}\} - s)}\right)$, with π_r^κ in place of π_r for specialized matching. With $F_0^+(s) = P(S_0^+ < s) + \frac{1}{2}P(S_0^+ = s)$, the comparison of positive parts is $Q(\kappa) = E\Big[F_0^+(S_\kappa^+)\Big]$. This expression is exact but has no closed form: the distribution of D_s depends on every amount and probability in the market, and must be computed numerically for every s. We approximate it below, and compute it by simulation in Section B.2.

Sparse cells. Equation (B.6) shows why sparse cells reveal little. An observed share of zero gives the same p whatever the degree of specialization, so the AUC exceeds $1/2$ only through the draws in which the branch has another borrower in industry j.

Proposition 2 (Sparse cells)**.** *For the relationship r_0 of branch b in industry j and every degree of specialization $\kappa \in [1, \kappa_{\max}]$:*

(i) with probability $1 - P_\kappa$, the branch has no other borrower in industry j, and then $p_{r_0} = \frac{1}{2}(1 - P_0)$;

(ii) $\mathrm{AUC}(\kappa) \le \frac{1}{2}(1 + P_\kappa)$;

(iii) hence, for any level $a > 1/2$ of the AUC, $\mathrm{AUC}(\kappa) \ge a$ requires $P_\kappa \ge 2a - 1$: p can rank the branch under specialized matching of degree κ above the same branch under random matching with probability a only if, under specialized matching, the branch has another borrower in the industry with probability at least $2a - 1$.

We prove it in Section B.5. In a sparse cell, P_κ is small unless κ is large: it is at most the expected number of other relationships of industry j that the branch receives under specialized matching, divided by the probability that it receives any relationship (Section B.5). Without size matching, that expected number is κ times its value under random matching,

$\kappa\pi n_j$, where π is the common assignment probability and n_j the number of relationships of industry j other than r_0.

Diagnostics under normality. The approximation proceeds in three steps. Only the second and third involve assumptions. Throughout, the moments are computed with the assignment probabilities π_r for random matching (subscript 0) and with π_r^κ for specialized matching (subscript κ). The probabilities P_0 and P_κ in (B.6) are exact and need no approximation.

Step 1: moments of a positive share (exact). A and C are sums over disjoint sets of independently assigned relationships, so they are independent, and conditioning on $A > 0$ leaves the distribution of C unchanged. Since $A = 0$ outside the event $A > 0$, $E[A^k \mid A > 0] = \bar{E}[A^k]/\bar{P}(A > 0)$ for $k = 1, 2$, with $\bar{P}(A > 0) = 1 - \prod_{r\in\mathcal{J}}(1 - \rho_r)$. With assignment probabilities ρ_r (equal to π_r or π_r^κ), the mean and variance of A given $A > 0$, and of C, are therefore

$$\begin{aligned} a^+ &= \frac{\sum_{r\in\mathcal{J}} \rho_r L_r}{\bar{P}(A>0)}, \quad & v_A^+ &= \frac{\sum_{r\in\mathcal{J}} \rho_r(1-\rho_r)L_r^2 + \left(\sum_{r\in\mathcal{J}} \rho_r L_r\right)^2}{\bar{P}(A>0)} - \left(a^+\right)^2, \\ c &= \sum_{r\in\mathcal{K}} \rho_r L_r, & v_C &= \sum_{r\in\mathcal{K}} \rho_r(1-\rho_r)L_r^2. \end{aligned} \tag{B.8}$$

Step 2: linearization. Assume that, given $A > 0$, A and C stay close to their means relative to their size. This holds when the branch can expect many relationships in and outside industry j. The share $S = A/(A + C)$ is then approximately linear in A and C.[37] Its mean and variance are

$$\mu^+ = \frac{a^+}{a^+ + c}, \qquad \left(\sigma^+\right)^2 = \frac{(1-\mu^+)^2\, v_A^+ + (\mu^+)^2\, v_C}{(a^+ + c)^2}. \tag{B.9}$$

Step 3: normality. Assume that the positive parts S_0^+ and S_κ^+ are normal. The linear term of Step 2 is a sum of independent terms, one per relationship, so by the central limit theorem this holds when no single relationship accounts for a large part of its variance. This is the assumption that fails when relationship amounts are skewed (Section B.2). Ties between continuous variables have probability zero, and $Q(\kappa) \approx \Phi\left((\mu_\kappa^+ - \mu_0^+)/\sqrt{(\sigma_0^+)^2 + (\sigma_\kappa^+)^2}\right)$. Substituting into (B.6),

$$\mathrm{AUC}(\kappa) \approx P_\kappa(1-P_0) + P_\kappa P_0\, \Phi\left(\frac{\mu_\kappa^+ - \mu_0^+}{\sqrt{\left(\sigma_0^+\right)^2 + (\sigma_\kappa^+)^2}}\right) + \tfrac{1}{2}(1-P_\kappa)(1-P_0). \tag{B.10}$$

The diagnostics follow: the approximation also gives $\mathrm{AUC}(1) = 1/2$, $\mathrm{AUC}(2)$ is (B.10) at $\kappa = 2$, and the ceiling is (B.10) at $\kappa_{\max}$ from (B.4). Each evaluation needs only the relationship amounts and the probabilities π_r and π_r^κ, not simulation. Even when the positive parts are far apart ($Q = 1$), the AUC is at most $P_\kappa + \frac{1}{2}(1 - P_\kappa)(1 - P_0)$: when the branch has a high probability of receiving no lending in industry j even under specialized

[37] The first-order expansion of $S = A/(A + C)$ around (a^+, c) is $S \approx \mu^+ + \left[(1-\mu^+)(A - a^+) - \mu^+(C - c)\right]/(a^+ + c)$, since $\partial S/\partial A = (1 - S)/(A + C)$ and $\partial S/\partial C = -S/(A + C)$. Its variance follows from the independence of A and C.

matching, its p often reflects only that event, and specialization is hard to detect, whatever the dispersion of the shares.

A closed form. Two further assumptions give a closed form that shows what drives the AUC. (i) The branch is sure to receive some lending in industry j, $P(A > 0) = 1$; then also $P_\kappa(A > 0) = 1$, since $\pi_r^\kappa \geq \pi_r$ for $r \in \mathcal{J}$. (ii) Specialization shifts the distribution of the share without changing its dispersion, $\sigma_\kappa^+ = \sigma_0^+$. Under (i), the moments in (B.8)–(B.9) are unconditional, and $\mu_0^+ = m$ and $\sigma_0^+ = \sigma_0$, with

$$m = \frac{\sum_{r\in\mathcal{J}} \pi_r L_r}{\sum_{r\in\mathcal{R}^-} \pi_r L_r}, \qquad CV = \frac{\sigma_0}{m} = \frac{\sqrt{\sum_{r\in\mathcal{R}^-} \pi_r(1-\pi_r)L_r^2 \left(\mathbb{1}\{r \in \mathcal{J}\} - m\right)^2}}{\sum_{r\in\mathcal{J}} \pi_r L_r}. \tag{B.11}$$

Because specialization multiplies expected lending to industry j by κ and leaves expected total lending unchanged, $\mu_\kappa^+ = \kappa m$. With (ii), (B.10) becomes

$$\mathrm{AUC}(\kappa) \approx \Phi\left(\frac{\kappa - 1}{\sqrt{2}\, CV}\right), \qquad \mathrm{AUC}(2) \approx \Phi\left(\frac{1}{\sqrt{2}\, CV}\right). \tag{B.12}$$

The baseline is 1/2, and the AUC depends only on the dispersion of the counterfactual share relative to its mean: AUC(2) exceeds 0.9 when $CV < 0.55$, and falls towards 1/2 as the coefficient of variation grows. The general form (B.10) adds the two features omitted from the closed form: the draws in which the branch receives no lending in industry j, and the change in dispersion under specialized matching. Both matter in our data.[38]

The effective number of relationships. The closed form has a simple interpretation. Define

$$n_{\text{eff}} = \frac{\left(\sum_{r\in\mathcal{J}} \pi_r L_r\right)^2}{\sum_{r\in\mathcal{J}} \pi_r L_r^2}. \tag{B.13}$$

The same quantity computed on the relationships of the other industries,

$$n_{\text{eff}}^{\mathcal{K}} = \frac{\left(\sum_{r\in\mathcal{K}} \pi_r L_r\right)^2}{\sum_{r\in\mathcal{K}} \pi_r L_r^2},$$

enters as well. Splitting the sum in (B.11) between $\mathcal{J}$ and $\mathcal{K}$, and using $m/\sum_{r\in\mathcal{J}} \pi_r L_r = (1-m)/\sum_{r\in\mathcal{K}} \pi_r L_r$, gives the exact decomposition

$$CV^2 = (1-m)^2 \left(\frac{g_{\mathcal{J}}}{n_{\text{eff}}} + \frac{g_{\mathcal{K}}}{n_{\text{eff}}^{\mathcal{K}}}\right), \qquad g_{\mathcal{S}} = \frac{\sum_{r\in\mathcal{S}} \pi_r(1-\pi_r)L_r^2}{\sum_{r\in\mathcal{S}} \pi_r L_r^2} \in [0,1] \quad (\mathcal{S} = \mathcal{J}, \mathcal{K}). \tag{B.14}$$

The first term is the dispersion of the branch's lending to industry j, the second that of its lending to other industries, which moves the share through its denominator. Three

[38] In the French application of Section B.3, with both forms computed on the retained relationships defined there, the closed form is as close to the simulated AUC(2) as the general form at the median, but less precise where assumption (i) fails. For relationships with $P(A > 0) \geq 0.99$, where only assumption (ii) is violated, it differs from the simulated AUC(2) by a median of 0.001, with 80% of the differences between -0.04 and $+0.03$. For those with $P(A > 0) < 0.9$, about 14% of the sample, the median difference is -0.006 (between -0.07 and $+0.02$), against -0.005 (between -0.03 and 0.00) for the general form (B.10), whose rank correlation with the simulation is 0.95, against 0.87.

conditions reduce (B.14) to $CV^2 \approx 1/n_{\text{eff}}$.

(a) Industry j is a small part of the branch's expected lending, so that m is close to zero and $(1-m)^2$ close to one. When industry j makes up most of the branch's expected lending, its share varies less, since lending to j is then also most of the denominator of the share, and the factor $(1-m)^2$ records this.

(b) The relationships that carry the weight in $\sum_{r\in\mathcal{J}} \pi_r L_r^2$, the largest ones of industry j, have small assignment probabilities, so that $g_{\mathcal{J}}$ is close to one. $g_{\mathcal{J}}$ is a weighted average of $1-\pi_r$, and it falls below one when the branch is likely to receive a large relationship: a relationship it receives for sure adds no dispersion.

(c) The branch's expected lending to other industries rests on many more effective relationships than its lending to industry j, $n_{\text{eff}}^{\mathcal{K}} \gg n_{\text{eff}}$, so that the second term is small relative to the first. This fails when the other industries' lending is concentrated in a few large relationships, for instance when a single large relationship of another industry may or may not go to the branch.

Conditions (a) and (b) only remove factors no larger than one, so dropping them overstates CV^2; dropping the second term, condition (c), understates it. Under the three conditions, $CV^2 \approx 1/n_{\text{eff}}$ and

$$\text{AUC}(2) \approx \Phi\left(\sqrt{n_{\text{eff}}/2}\right), \tag{B.15}$$

which exceeds 0.9 when the branch can expect at least about three effective relationships in the industry ($n_{\text{eff}} \geq 3.3$). When all relationships have the same probability π_b, as without size matching, $n_{\text{eff}} = \pi_b/H_j$, where $H_j = \sum_{r\in\mathcal{J}} L_r^2 / \left(\sum_{r\in\mathcal{J}} L_r\right)^2$ is the Herfindahl index of relationship amounts in industry j. Since $1/H_j$ is the number of equal-sized relationships that would have the same Herfindahl index, n_{eff} is the *effective number* of industry-j relationships received by the branch under random matching: the number of equal-sized relationships that would give its counterfactual share the same dispersion. It is small when the branch can expect few relationships in the industry, and also when the amounts in the industry are concentrated in a few large relationships. Unequal amounts therefore act like sparsity. Section B.4 shows how n_{eff} changes as markets grow.

An example. Consider a market with 490 loans in three industries: Manufacturing (30 loans of 10 and 300 of 1, total 600), Services (10 loans of 10 and 50 of 1, total 150), and Construction (100 loans of 1, total 100). Three banks receive each loan independently with probabilities 0.6, 0.3 and 0.1, whatever its size and industry, so no bank is specialized. Table B.1 reports the diagnostics under normality for each bank and industry, and Figure B.1 illustrates them for the medium bank in Services. Without specialization, that bank's share in Services is centered on the industry's market share, 0.18, with a standard deviation of 0.05. When the bank is specialized with $\kappa = 1.5$, its share distribution shifts to the right and widens slightly, and the AUC is 0.88. At $\kappa = 2$ the two distributions barely overlap: AUC(2) = 0.99. Full specialization would be $\kappa_{\max} = 3.33$, so under specialized matching with $\kappa = 2$ the bank's expected share is 0.60 of the largest it could have ($2/\kappa_{\max} = 0.60$).

The table shows how the diagnostics depend on market structure. Smaller banks receive fewer loans, so their counterfactual shares are more dispersed and their AUC is lower: in Services, AUC(2) falls from 0.99 for the medium bank to 0.87 for the small

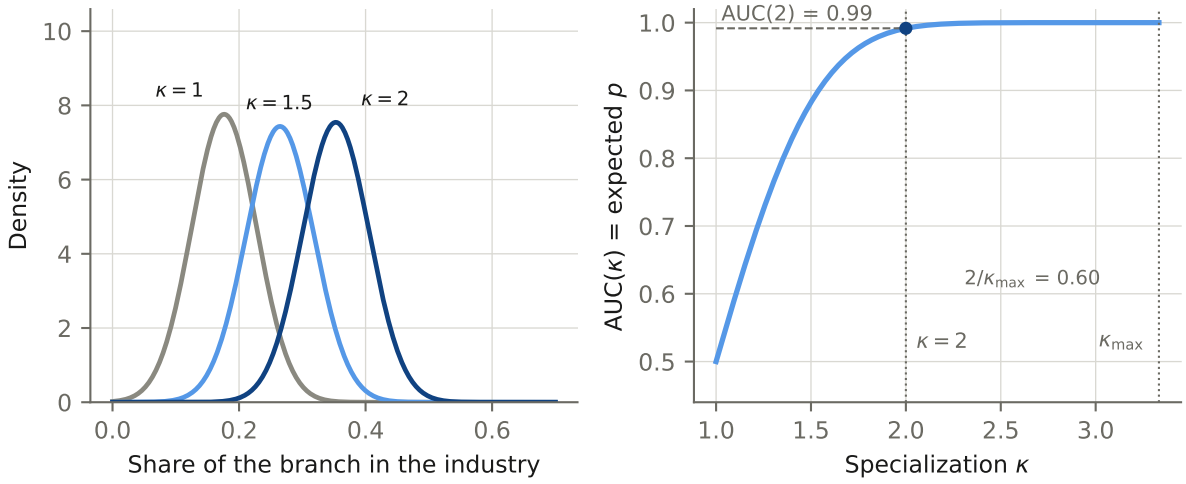


Figure B.1: The AUC of p under normality (simulated example)

Notes: Toy market of Table B.1; medium bank ($\pi = 0.3$), Services. Left: normal distributions of the bank's share in Services without specialization and with specialization of degree κ, with mean μ_κ^+ and standard deviation σ_κ^+ from equations (B.8)–(B.9). Right: AUC under normality, equation (B.10), with AUC(2) and $\kappa_{\max}$ marked.

Table B.1: Diagnostics under normality in a toy market (simulated example)

					AUC(2)		
Bank	Industry	$P(A>0)$	CV	$2/\kappa_{\max}$	General	Closed form	AUC($\kappa_{\max}$)
Large	Manuf.	1.00	0.04	1.41	–	–	1.00
($\pi = 0.6$)	Serv.	1.00	0.16	1.20	–	–	1.00
	Constr.	1.00	0.10	1.20	–	–	1.00
Medium	Manuf.	1.00	0.07	1.41	–	–	1.00
($\pi = 0.3$)	Serv.	1.00	0.29	0.60	0.99	0.99	1.00
	Constr.	1.00	0.18	0.60	1.00	1.00	1.00
Small	Manuf.	1.00	0.15	1.41	–	–	1.00
($\pi = 0.1$)	Serv.	1.00	0.57	0.35	0.87	0.89	1.00
	Constr.	1.00	0.35	0.24	0.94	0.98	1.00

Notes: Toy market of 490 loans (Manufacturing: 30 of 10 and 300 of 1; Services: 10 of 10 and 50 of 1; Construction: 100 of 1), assigned independently to three banks with probabilities π. $P(A>0)$: equation (B.2); CV: equation (B.11); $\kappa_{\max}$: equation (B.4); AUC(2), general form, and the ceiling AUC($\kappa_{\max}$): equation (B.10); closed form: equation (B.12). "–": a doubling is infeasible ($2/\kappa_{\max} > 1$).

one. Granularity matters as much as industry size: the small bank's AUC(2) is lower in Services (0.87) than in Construction (0.94), although Services is the larger industry, because two thirds of Services lending comes from ten large loans. A doubling is infeasible for the large bank in every industry, and for every bank in Manufacturing ($2/\kappa_{\max}$ of 1.20 to 1.41): the large bank already receives 60% of every industry under random matching, and Manufacturing is 71% of every bank's expected lending. For these cells the ceiling is the relevant diagnostic, and here it is close to one. In this market $P(A > 0) = 1$ in every cell, so the closed form (B.12) differs from the general form only because it ignores the change in dispersion under specialized matching. The difference is at most 0.04, for the small bank in Construction.

B.2 Skewed data: diagnostics by simulation

Why normality fails. Counterfactual shares are not normal in general. When a branch can expect to receive few relationships in an industry, its share takes few distinct values and has a mass of draws with no lending at all. When relationship amounts are unequal, a few large relationships make the distribution skewed, and a single very large relationship can split it in two: the draws in which the branch receives it, and those in which it does not. The normal diagnostics then misstate the AUC, in either direction. Credit data are of this kind: relationship amounts have tails close to a power law (Section B.3).

The procedure. We therefore compute the diagnostics from the simulated distributions themselves, without assuming a shape. The draws of S_0 can be those already generated to compute p (Step 6 of Appendix A), or new draws of the branch's portfolio under random matching when more precision is needed. For a given κ, we draw the branch's portfolio under specialized matching, with the same procedure and the probabilities π_r^κ of (B.3), and compute its p against the random-matching draws of S_0, exactly as p is computed for an observed branch. The average of these values is AUC(κ), evaluated at $\kappa = 2$ and at the ceiling $\kappa_{\max}$. Unlike the draws of S_0, the draws under specialized matching cannot be shared across relationships: they depend on the branch, the industry and κ, so they must be generated anew for every relationship and for every value of κ. The cost is therefore of a different order from that of p. The measure needs a fixed number of draws of the whole market, which serve all of its relationships; the simulated diagnostics need, for each relationship, a fixed number of draws for each value of κ. We report the cost in our data in Section B.3. The simulated AUC(1) equals 1/2 up to simulation noise, which provides a check on the computation.

The example with a giant loan. To see how skewness affects the diagnostics, we add to the toy market a single Services loan of 400, about a third of all lending. Table B.2 compares the diagnostics under normality with the simulated ones, before and after adding the loan, and Figure B.2 shows the case of the small bank in Services. In the base market the two agree: AUC(2) and the ceiling are the same to two decimals in every cell. With many loans of similar size, the shares are close to normal. With the giant loan, the normal diagnostics fail. The small bank receives the giant loan in only 10% of draws. In these draws its share in Services jumps to about 0.85 (Figure B.2, left), which inflates the standard deviation of its counterfactual share. The normal diagnostic reads this dispersion as noise and puts the small bank's AUC(2) at 0.78 in Services, 0.81 in Manufacturing and 0.68

in Construction. But p is a rank: a specialized branch's share only needs to exceed the counterfactual draws, and the rare draws with the giant loan barely matter. The simulation gives 0.92, 0.99 and 0.86. The error is mostly in this direction, understating how well p separates specialized matching from random matching, but not always: for the large bank in Construction, for which a doubling is infeasible, the normal ceiling is 0.90, against 0.82 by simulation.

Table B.2: Diagnostics under normality and simulated, with and without a giant loan (simulated example)

		Base market					With a giant Services loan				
			AUC(2)		Ceiling			AUC(2)		Ceiling	
Bank	Industry	$\frac{2}{\kappa_{\max}}$	Norm.	Sim.	Norm.	Sim.	$\frac{2}{\kappa_{\max}}$	Norm.	Sim.	Norm.	Sim.
Large	Manuf.	1.41	–	–	1.00	1.00	1.20	–	–	0.93	0.91
($\pi = 0.6$)	Serv.	1.20	–	–	1.00	1.00	1.20	–	–	0.97	1.00
	Constr.	1.20	–	–	1.00	1.00	1.20	–	–	0.90	0.82
Medium	Manuf.	1.41	–	–	1.00	1.00	0.96	0.95	0.98	0.99	1.00
($\pi = 0.3$)	Serv.	0.60	0.99	0.99	1.00	1.00	0.88	0.94	0.94	0.98	1.00
	Constr.	0.60	1.00	1.00	1.00	1.00	0.60	0.82	0.83	0.92	0.98
Small	Manuf.	1.41	–	–	1.00	1.00	0.96	0.81	0.99	0.87	1.00
($\pi = 0.1$)	Serv.	0.35	0.87	0.87	1.00	1.00	0.88	0.78	0.92	0.85	1.00
	Constr.	0.24	0.94	0.94	1.00	1.00	0.20	0.68	0.86	0.97	1.00

Notes: Base market: toy market of Table B.1. With one giant Services loan: the same market plus one Services loan of 400. Norm.: under normality, general form (B.10). Sim.: AUC(κ) computed from 100,000 draws of random matching and 50,000 of specialized matching. Ceiling: AUC($\kappa_{\max}$). "–": a doubling is infeasible ($2/\kappa_{\max} > 1$).

A remark on the relative portfolio share. Computed on a counterfactual portfolio, the relative portfolio share of equation (2) is S/s_j, with $s_j = \sum_{r\in\mathcal{J}} L_r / \sum_{r\in\mathcal{R}^-} L_r$. In the simplest case it is centered at one. Suppose that all n relationships in $\mathcal{R}^-$ have the same amount and the same assignment probability π, and that n_j of them are in industry j, so $s_j = n_j/n$. Given that the branch receives $v \geq 1$ relationships, which ones it receives is a uniformly random subset of size v, so the number of them in industry j is hypergeometric, and

$$E\left[\frac{S}{s_j}\,\middle|\, V = v\right] = 1, \qquad \mathrm{Var}\left(\frac{S}{s_j}\,\middle|\, V = v\right) = \frac{1 - s_j}{s_j}\,\frac{n - v}{v\,(n-1)}. \tag{B.16}$$

The counterfactual relative portfolio share therefore averages one over the valid draws, but its dispersion is of the order of $1/\sqrt{v\,s_j}$: large when the branch expects few relationships in the industry. A branch that receives two relationships in a market where the industry accounts for a tenth of lending has a counterfactual relative portfolio share of 0, 5 or 10. With unequal amounts or unequal probabilities the counterfactual relative portfolio share is no longer centered at one, because A/V is a ratio of random sums and because the random-matching benchmark differs from s_j. Neither feature affects p, which compares the observed share with the counterfactual distribution whatever its center and dispersion,

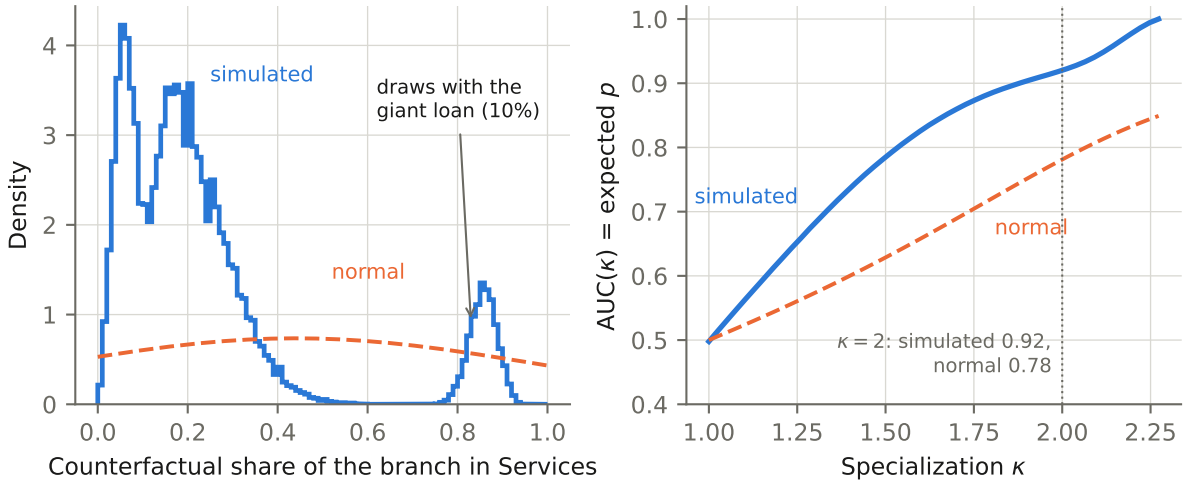


Figure B.2: A giant loan: simulated and normal diagnostics (simulated example)

Notes: Toy market with one giant Services loan (Table B.2); small bank ($\pi = 0.1$), Services. Left: simulated distribution of the bank's counterfactual share in Services (valid draws) and the normal distribution with mean μ_0^+ and standard deviation σ_0^+ of equation (B.9). Right: AUC(κ) computed from the simulated distributions and under normality (general form, equation (B.10)); the dotted line marks $\kappa = 2$.

but both are reasons why a simple benchmark such as "relative portfolio share above one" does not reveal specialization in sparse markets.

B.3 Sparsity in French local credit markets

In this section we apply the diagnostics to our data for the first quarter of 2016: 1.82 million relationships between firms and bank branches, in the eight market definitions obtained by crossing two geographic partitions (urban units and urban areas, by the location of the branch) with four levels of the industry classification. Assignment probabilities follow Step 5 of Appendix A, with département matching rates and size classes that hold equal shares of lending (Step 4).

Heavy tails. Table B.3 reports Hill estimates of the tail exponent of relationship amounts and of branch lending, and Figure B.3 plots their survival functions on a log-log scale. Both distributions have tails close to a power law with exponent near one: between 0.97 and 1.00 for relationship amounts, and between 0.82 and 1.20 for branch lending, depending on the share of the largest observations used. With an exponent below two, the variance of a sum of relationship amounts is dominated by its largest terms, which is why the normal approximation fails for them, as we now explain.

Why the normal approximation fails here. With tails this heavy, the normal approximation (B.10) fails even in markets with many relationships, for the reason illustrated by the giant loan of Section B.2. The variance of a branch's counterfactual share is then dominated by the few largest relationships of the market, which a small branch receives only rarely: for a small branch in a large market, the variance mostly measures the size and probability of these rare events, and when the tail exponent is close to or below two it

Table B.3: Tails of relationship amounts and branch lending

		Tail exponent, top share used			
	Observations	0.5%	1%	2%	5%
Relationship amounts	1,815,416	1.00	0.97	0.98	0.99
Branch lending	19,984	1.20	0.88	0.86	0.82

Notes: Hill estimator of the tail exponent ζ in $P(X > x) \propto x^{-\zeta}$, computed on the given share of the largest observations. Relationships and branches of 2016Q1.

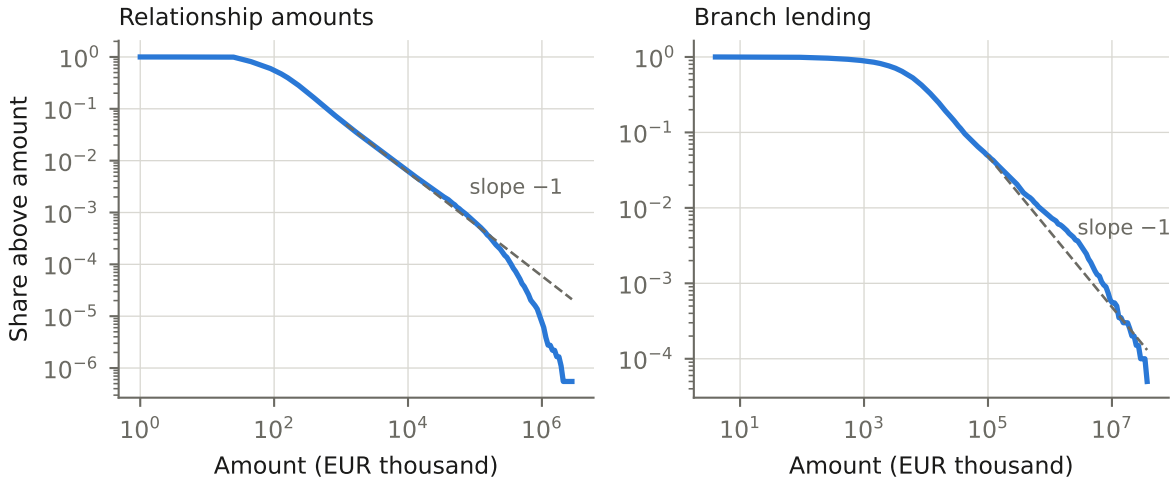


Figure B.3: Survival functions of relationship amounts and branch lending

Notes: Share of relationships (left) and of branches (right) with amount above x, 2016Q1, log-log scale. The dashed line has slope −1 (Zipf's law).

settles down slowly, if at all, as the market grows. But p is a rank. An event that occurs in a small fraction of draws barely changes whether the branch's share under specialized matching exceeds most of the random-matching draws. A rule that uses the variance of the full distribution, exact or approximate, therefore overstates how hard it is to detect specialization at small branches. Simpler summaries built on second moments, such as the variance H_j/π_b of the relative portfolio share in a simple balls-and-bins model, also rank cells poorly (see the end of this section).

A screening rule with trimming. The normal approximation remains useful for screening cells without simulation, provided its moments are computed only on the relationships that shape the bulk of the counterfactual distribution, the *retained* relationships. Sort the relationships of $\mathcal{R}^-$ by decreasing amount, and set aside the largest ones as long as their probabilities add up to at most $\alpha = 0.1$: branch b receives any of them in at most one draw in ten, too rarely to shape the bulk of the counterfactual distribution against which p ranks the branch's share. Let $\mathcal{G}$ be the remaining relationships. The rule is the normal approximation (B.10), with the probabilities of a valid draw and the moments of (B.8)–(B.9) computed on $\mathcal{G}$ instead of $\mathcal{R}^-$. The probabilities π_r^κ of specialized matching are those of (B.3), which are computed on all relationships. The rule's AUC(2) and ceiling are this approximation at $\kappa = 2$ and at $\kappa_{\max}$. It needs only the relationship amounts and the assignment probabilities.

How sparse are the markets? Table B.4 reports the diagnostics by market definition, computed by simulation for a random sample of 250 relationships per market definition (drawn uniformly among relationships for which p is defined, in all markets), next to the trimmed rule for the same relationships and for all relationships. The measure p is undefined for few relationships (at most 1.5%, at the finest industry level), because the firm is almost never the only borrower of its industry in the market. A doubling is infeasible for 14 to 22% of the sampled relationships. Where it is feasible, detectability deteriorates as industries become finer. The median simulated AUC(2) falls from 0.94 at NAF level 0 to 0.83 in urban units and 0.84 in urban areas at NAF level 3, its lower quartile from 0.85 to 0.70 in urban units, and the share of relationships with AUC(2) of at least 0.9 from about 60% to about 40%: at fine industry levels, a doubling is often not enough for p to tell specialized matching from random matching. Even full specialization leaves the expected p below 0.9 for only 0.7% of the sampled relationships. Urban areas, which are larger than urban units, give similar results. In Section 3.2.2 we report the diagnostics for all market definitions, including the coarser ones, and explain why larger geographic markets do not make specialization easier to detect.

Accuracy of the screening rule. The first two blocks of Table B.4 show that the trimmed rule reproduces the quartiles of the simulated AUC(2) in every market definition, with slightly lower values at fine industry levels: in the sparsest cells it makes specialization look harder to detect than it is. Relationship by relationship, the rule's AUC(2) differs from the simulated one by a median of -0.001 across the 1,645 sampled relationships for which a doubling is feasible, with 80% of the differences between -0.07 and $+0.01$ (a mean absolute difference of 0.02), and the rule ranks relationships much like the simulation (rank correlation 0.98). When it errs, it errs on the side of caution: it understates AUC(2) by

Table B.4: Sparsity diagnostics by market definition: simulation and trimmed rule

			Simulation (sample)			Trimmed rule (same sample)			Rule (all)	
			AUC(2) quartiles			AUC(2) quartiles				
Geography	Industry	$2/\kappa_{\max} > 1$ (%)	25%	50%	75%	25%	50%	75%	Median AUC(2)	$2/\kappa_{\max} > 1$ (%)
Urban units	NAF level 0	20.8	0.85	0.94	1.00	0.82	0.94	1.00	0.95	19.1
	NAF level 1	21.6	0.78	0.89	0.99	0.73	0.87	1.00	0.91	19.1
	NAF level 2	21.6	0.74	0.87	1.00	0.68	0.81	1.00	0.87	19.0
	NAF level 3	14.8	0.70	0.83	0.98	0.67	0.80	0.99	0.81	16.0
Urban areas	NAF level 0	14.4	0.85	0.94	1.00	0.82	0.95	1.00	0.94	16.5
	NAF level 1	18.0	0.79	0.89	1.00	0.74	0.88	1.00	0.90	16.4
	NAF level 2	16.4	0.74	0.87	0.99	0.68	0.84	0.99	0.86	16.3
	NAF level 3	14.4	0.71	0.84	0.99	0.66	0.83	0.99	0.81	13.0

Notes: 2016Q1. Markets: urban units or urban areas of the branch, excluding overseas territories, pooled rural codes and markets with a single branch. $2/\kappa_{\max} > 1$: share of sampled relationships for which a doubling is infeasible. "Simulation (sample)": AUC(2) computed by simulation (Section B.2) for a random sample of 250 relationships per market definition, drawn uniformly among relationships for which p is defined; quartiles among those for which a doubling is feasible. "Trimmed rule": normal approximation (B.10) computed on the retained relationships, $\alpha = 0.1$, for the same relationships and for all relationships of the market definition.

more than 0.05 for 15% of these relationships and never overstates it by as much. Classifying relationships by whether AUC(2) reaches 0.9, the rule agrees with the simulation for 98% of them; the 7 it places above 0.9 by mistake have a simulated AUC(2) between 0.896 and 0.900. Its ceilings differ from the simulated ones by 0.002 on average. Trimming matters most where heavy tails do. In a separate sample of 200 relationships of small branches in large markets, the same approximation without trimming understates AUC(2) by a median of 0.027, and by more than 0.05 for 29% of the relationships for which a doubling is feasible (0.008 and 19% with trimming), and it ranks relationships less like the simulation (rank correlation 0.95, against 0.97 with trimming). We therefore recommend the rule as a screen: its values can be used directly, but cells where it finds AUC(2) just below a threshold of interest, or a low ceiling, should be checked by simulation.

Comparison with a simple balls-and-bins formula. In a balls-and-bins model with a single branch probability π_b and equal treatment of all relationships, the variance of the relative portfolio share is approximately H_j/π_b, the reciprocal of the effective number of relationships (B.13) computed on all relationships. This formula ranks cells poorly. In the 2,000 sampled relationships, it orders 25% of pairs differently from the simulated AUC(2), against 5% for the trimmed rule, and it finds only 47% of the quartile of relationships where specialization is hardest to detect, against 87% for the rule. Even when comparing two industries of the same branch, it orders them differently from the simulation in 12% of pairs, against 2% for the rule.

Computation. The diagnostics cost much more to compute than p (Section 2.5). In our data for 2016Q1, with 1.8 million relationships and the 16 market definitions of Section 3.1 other than France as a single market, computing p together with $P(A > 0)$ and the tie probability from its draws takes 1.2 hours of one processor core. The trimmed rule for every relationship takes 1.5 hours. The simulated diagnostics for every relationship would take

of the order of 30,000 hours.[39] Simulation is therefore feasible for samples of relationships, as in Table B.4, and the trimmed rule for every relationship and every quarter.

The cost of the rule comes from specialized matching. For each relationship and for each value of κ, the odds factors θ and λ solve the two conditions of (B.3), which have no closed form, and every term of these conditions and of the moments (B.8) is a sum over the relationships of the market. By (A.4), $\pi_b(r)$, and hence π_r^κ, depends on r only through its size classes in the départements of the market's branches. Call a set of relationships with the same size classes a *size group*. Each sum then reduces to a sum over size groups, inside and outside industry j, of group totals of counts, amounts and squared amounts. A market whose branches all lie in one département has at most ten size groups. A market whose branches lie in several départements combines their class bounds, and has up to one group for each interval between the bounds of all the départements present: 71 groups in the urban unit of Paris, 120 in its urban area, and between 19 and 73 in the régions. Markets with more than ten groups hold 35% of relationships in urban units and 50% in urban areas. The rule takes 0.1 milliseconds per relationship and industry level in markets with at most ten groups, and 0.4 milliseconds in markets with more than 50. The cost of p does not depend on the number of groups (about 0.6 milliseconds per relationship for the four industry levels together). With national matching rates every market would have at most ten groups, so the higher cost of the rule in large markets is a consequence of estimating the rates by département (Appendix A, Step 4).

Four devices keep the rule affordable without changing its value. (i) All sums are computed from size-group totals; leaving the focal relationship out subtracts it from the totals of its group, and trimming proceeds group by group, since groups are intervals of amounts. (ii) Relationships with the same inputs (branch, industry, size group, amount, and whether they are set aside by the trimming) are computed once. (iii) θ and λ are found by Newton's method safeguarded by bisection, and the search for each value of κ starts from their values at the previous one. (iv) The relationships of the largest markets are split into blocks processed in parallel.

B.4 Market size and heavy tails

In French markets, p reveals specialization less reliably in larger markets: AUC(2) is lower for France as a single market than in urban units (Section 3.2.2). The effective number of relationships (B.13) explains why. Consider a market of N relationships whose amounts $L_1, \ldots, L_N$ are independent draws from a continuous distribution F on $(0, \infty)$, each in industry j independently with probability $s \in (0, 1)$, and a branch b whose own lending V_b does not change with N. Assignment is without size matching, so every relationship goes to the branch with probability $\pi_b = V_b / \sum_{i=1}^{N} L_i$, and $n_{\text{eff},N} = \pi_b / H_j$. The *trimmed* effective number $n_{\text{eff},N}^T$ is computed in the same way after setting aside the $\lfloor \alpha / \pi_b \rfloor$ largest relationships of the market, whose probabilities add up to at most α, as in the screening rule

[39] Timings on our computer, in hours of one processor core; only their ratios carry over to other machines. The simulation uses 10,000 draws under random matching and 5,000 draws for each value of κ at which the AUC is evaluated: $\kappa = 1.5$, 2 and 3, $\kappa_{\max}$, and $\kappa = 1$ as a check. The rule evaluates the same values except $\kappa = 1$. The total for the simulation is extrapolated from the time taken by 200 relationships sampled uniformly (25 for each geography at NAF levels 0 and 3), and is imprecise because a few relationships take much longer than the rest (median 1.0 to 2.5 seconds per relationship across geographies and industry levels, maximum 51 seconds).

of Section B.3. Let $\mu = E[L]$ when it is finite, and for a threshold ℓ let $\mu_\ell = E[L\,\mathbb{1}\{L \le \ell\}]$ and $\nu_\ell = E[L^2\,\mathbb{1}\{L \le \ell\}]$.

Proposition 3 (Market size)**.** *As $N \to \infty$:*

(i) if $E[L^2] < \infty$, then $n_{\text{eff},N} \to sV_b\,\mu/E[L^2]$ almost surely;

(ii) if amounts follow a Pareto distribution with minimum one and tail exponent $\zeta \in (0,2)$, $1 - F(x) = x^{-\zeta}$ for $x \ge 1$, then $n_{\text{eff},N} \to 0$ in probability;

(iii) if $\mu < \infty$ and $V_b > \alpha\mu$, then $n^T_{\text{eff},N} \to sV_b\,\mu_{\bar\ell}^2/(\mu\,\nu_{\bar\ell}) > 0$ almost surely, where $\bar\ell$ is the $1 - \alpha\mu/V_b$ quantile of F;

(iv) if $F(x) = 1 - 1/x$ for $x \ge 1$ (the Pareto distribution of (ii) with $\zeta = 1$), the expected number of relationships received by the branch, $N\pi_b$, satisfies $N\pi_b \log N/V_b \to 1$ in probability.

With amounts of finite variance (i), the effective number of relationships converges to a constant: larger markets neither help nor hurt. With a power-law tail of exponent below two (ii), it falls to zero, because the Herfindahl index of a market's amounts stays dominated by its largest relationships however large the market. Part (ii) is the counterpart, for a branch, of a result of Gabaix (2011, Proposition 2): with power-law sizes of exponent $\zeta \in [1,2)$, the Herfindahl index of N sizes falls like $N^{-(2-2/\zeta)}$, and like $1/(\log N)^2$ under Zipf's law, rather than like $1/N$, while the probability π_b that a branch of fixed size receives a given relationship falls like $1/N$ (like $1/(N \log N)$ when $\zeta = 1$). Their ratio, $n_{\text{eff},N}$, therefore falls to zero. The normal approximation, and any measure of dispersion computed on all relationships, then makes specialization look harder and harder to detect as the market grows. Part of this is an artifact: the largest relationships are those the branch almost never receives, and p, a rank, barely reacts to them. After they are set aside (iii), the effective number converges to a positive constant whenever the mean amount is finite. With an exponent of one, the mean is infinite, and a second channel remains (iv): the average amount in the market grows like $\log N$, so a branch of fixed size can expect fewer and fewer relationships, and its counterfactual share becomes sparser. Fewer expected relationships mean a smaller effective number, trimmed or not: by the Cauchy–Schwarz inequality, $\left(\sum_{r\in\mathcal{J}} \pi_r L_r\right)^2 \le \sum_{r\in\mathcal{J}} \pi_r \sum_{r\in\mathcal{J}} \pi_r L_r^2$, so n_{eff} never exceeds $\sum_{r\in\mathcal{J}} \pi_r$, the expected number of industry-j relationships received by the branch. Here that number is π_b times the number of industry-j relationships, about $sN\pi_b$, and by (iv) it falls like $sV_b/\log N$. Through (B.15), a smaller effective number means a lower AUC(2). In the French credit data, the estimated tail exponent of relationship amounts is close to one (Table B.3), and setting all amounts equal closes the gap between the diagnostics of urban units and of France as a single market (Section 3.2.2). Table B.5 checks each statement numerically.

B.5 Proofs of Propositions 1, 2 and 3

We use the notation of Section B.1: r_0 is the relationship of branch b in industry j for which p is computed, A, C and $V = A + C$ are the branch's counterfactual lending to industry j, to other industries and in total, over the relationships other than r_0, and $P(\cdot)$ denotes probabilities conditional on a valid draw, $V > 0$, while $\bar P$ denotes unconditional probabilities.

Table B.5: The effective number of relationships as the market grows (simulated example)

Amounts	Relationships N	n_{eff}		Expected number received, $N\pi_b$	
		All	Trimmed	Simulated	$V_b/\log N$
Exponential	1,000	10.000	10.000	199.6	–
	10,000	9.994	10.065	199.3	–
	100,000	9.998	10.112	199.9	–
	1,000,000	10.010	10.106	200.0	–
	Limit	10.000	10.104		
Pareto, $\zeta = 1.5$	1,000	2.624	2.782	68.4	–
	10,000	1.220	2.013	67.6	–
	100,000	0.722	2.056	67.5	–
	1,000,000	0.283	2.034	67.1	–
	Limit	0	2.027		
Pareto, $\zeta = 1$	1,000	0.145	0.305	19.8	29.0
	10,000	0.056	0.275	18.2	21.7
	100,000	0.005	0.252	14.5	17.4
	1,000,000	0.001	0.239	12.9	14.5
	Limit	0	–		

Notes: Markets of N relationships with independent amounts (exponential with mean one; Pareto with minimum one and tail exponent ζ), each in industry j with probability $s = 0.1$; branch lending $V_b = 200$; $\pi_b = V_b/\sum_i L_i$. n_{eff}: equation (B.13), on all relationships and after setting aside the largest relationships whose probabilities add up to at most $\alpha = 0.1$ (trimmed). $N\pi_b$: expected number of relationships received by the branch. Medians over 40 markets. Limit: Proposition 3, parts (i) to (iii). $V_b/\log N$: the approximation of $N\pi_b$ in part (iv).

Proposition 1. Fix a market and its h-replica, in which every relationship r of the market appears h times with the same amount L_r and probability $\pi_r = \pi_b(r)$, and r_0 is one copy of a relationship of branch b. Let $a = \sum_{r:\, j(r)=j} \pi_r L_r$ and $v = \sum_r \pi_r L_r > 0$, the sums running over the relationships of the original market, so that $\mu_{bj} = a/v$. In the h-replica, A_h and V_h are sums of independent terms over the h copies of every relationship except one copy of r_0, so

$$E[A_h] = ha - \pi_{r_0} L_{r_0}, \qquad \mathrm{Var}(A_h) \le h \sum_r L_r^2,$$

and likewise for V_h with v in place of a. By Chebyshev's inequality, $A_h/h \to a$ and $V_h/h \to v$ in probability, so the counterfactual share $S_h = A_h/V_h$ converges in probability to μ_{bj}, and $\bar{P}(V_h > 0) \to 1$. Conditioning on a valid draw does not change this limit: for every $\delta > 0$, $P(|S_h - \mu_{bj}| \ge \delta) \le \bar{P}(|S_h - \mu_{bj}| \ge \delta)/\bar{P}(V_h > 0) \to 0$. The observed leave-one-out share in the h-replica is $(hA_{bj} - L_{r_0})/(hV_b - L_{r_0})$, which converges to $S_{bj} = A_{bj}/V_b$, the same in every replica.

Suppose $S_{bj} > \mu_{bj}$ and let $\delta = (S_{bj} - \mu_{bj})/2$. For h large enough, the observed share exceeds $\mu_{bj} + \delta$, so by definition (5)

$$p_{r_0} \ge P\Big(S_h < \mu_{bj} + \delta\Big) \ge 1 - P\Big(|S_h - \mu_{bj}| \ge \delta\Big) \to 1.$$

If $S_{bj} < \mu_{bj}$, the same argument with the observed share below $\mu_{bj} - \delta$ gives $p_{r_0} \le P(S_h \le \mu_{bj} - \delta) \le P(|S_h - \mu_{bj}| \ge \delta) \to 0$. □

Proposition 2. (i) For a branch with specialization κ, the observed leave-one-out share is a draw of S with probabilities π_r^κ (equation (B.3)), given $V > 0$. It is zero exactly when $A = 0$, which has probability $1 - P_\kappa$. In a valid counterfactual draw, $S \ge 0$, with $S = 0$ exactly when $A = 0$, which has probability $1 - P_0$. An observed share of zero therefore has no counterfactual draw below it and ties with probability $1 - P_0$, so by (5), $p_{r_0} = \frac{1}{2}(1 - P_0)$.

(ii) Expanding the last term of (B.6), $\frac{1}{2}(1 - P_\kappa)(1 - P_0) = \frac{1}{2} - \frac{1}{2}P_\kappa - \frac{1}{2}P_0 + \frac{1}{2}P_\kappa P_0$, gives the identity

$$\mathrm{AUC}(\kappa) - \tfrac{1}{2} = \tfrac{1}{2}\,(P_\kappa - P_0) + P_\kappa P_0 \left(Q(\kappa) - \tfrac{1}{2}\right). \tag{B.17}$$

Since $Q(\kappa) \le 1$, $\mathrm{AUC}(\kappa) - \frac{1}{2} \le \frac{1}{2}(P_\kappa - P_0) + \frac{1}{2}P_\kappa P_0 = \frac{1}{2}P_\kappa - \frac{1}{2}P_0(1 - P_\kappa) \le \frac{1}{2}P_\kappa$.

(iii) If $\mathrm{AUC}(\kappa) \ge a$, then by (ii) $\frac{1}{2}(1 + P_\kappa) \ge a$, that is, $P_\kappa \ge 2a - 1$. □

A bound on P_κ. Since $A > 0$ requires at least one relationship of industry j to be assigned to the branch, the union bound gives

$$P_\kappa = \frac{\bar{P}_\kappa(A > 0)}{\bar{P}_\kappa(V > 0)} \le \frac{\sum_{r \in \mathcal{J}} \pi_r^\kappa}{\bar{P}_\kappa(V > 0)},$$

where the numerator is the expected number of other relationships of industry j that the branch receives under specialized matching. By (B.3), $\pi_r^\kappa \le \theta \pi_r$ for $r \in \mathcal{J}$, since $1 - \pi_r + \theta\pi_r \ge 1$ when $\theta \ge 1$. When all relationships in $\mathcal{J}$ have the same probability π, as without size matching, they also have the same probability under specialized matching, and the second condition in (B.3) makes it $\kappa\pi$: the numerator is then exactly $\kappa\pi n_j$, with $n_j = |\mathcal{J}|$.

Proposition 3. Write $S_N = \sum_{i \le N} L_i$, $A_N = \sum_{i \le N, i \in \mathcal{J}} L_i$, $Q_N = \sum_{i \le N, i \in \mathcal{J}} L_i^2$ and $M_N = \max_{i \le N, i \in \mathcal{J}} L_i$, so that $n_{\text{eff},N} = V_b A_N^2/(S_N Q_N)$.

(i) By the strong law of large numbers, $A_N/N \to s\mu$, $S_N/N \to \mu$ and $Q_N/N \to sE[L^2]$ almost surely, so $n_{\text{eff},N} \to V_b s^2 \mu^2/(\mu\, sE[L^2])$.

(ii) For $\zeta \in [1,2)$, this part also follows from Gabaix (2011, Proposition 2 and its proof), who derives the rate at which the square root of the Herfindahl index of power-law sizes falls with N. The argument below is self-contained and also covers $\zeta < 1$. Since $A_N \le S_N$ and $Q_N \ge M_N^2$, $n_{\text{eff},N} \le V_b A_N/Q_N \le V_b A_N/M_N^2$, and it suffices to show that $A_N/M_N^2 \to 0$ in probability. The number n of relationships in industry j is binomial with mean sN, so $n/N \to s$ almost surely, and given n their amounts are n independent draws from F.

The maximum. For $\varepsilon > 0$ and n large enough that $\varepsilon n^{1/\zeta} \ge 1$,

$$P\big(M_N \le \varepsilon n^{1/\zeta}\big) = \left(1 - \frac{\varepsilon^{-\zeta}}{n}\right)^n \le e^{-\varepsilon^{-\zeta}},$$

so $M_N^2 \ge \varepsilon^2 n^{2/\zeta}$ with probability at least $1 - e^{-\varepsilon^{-\zeta}}$.

The sum. If $1 < \zeta < 2$, the mean $\mu = \zeta/(\zeta - 1)$ is finite, and by the strong law of large numbers $A_N \le 2\mu n$ for N large enough. If $\zeta \le 1$, fix $\delta \in (0,1)$ and let $b = n^{1/\zeta}/\delta$. By the union bound, no amount of industry j exceeds b with probability at least $1 - nb^{-\zeta} = 1 - \delta^\zeta$, and on that event A_N equals the truncated sum $T = \sum_{i \in \mathcal{J}} L_i \mathbb{1}\{L_i \le b\}$, whose expected value is

$$nE\Big[L\,\mathbb{1}\{L \le b\}\Big] = n\int_1^b \zeta x^{-\zeta}\,dx = \begin{cases} n\zeta\big(b^{1-\zeta} - 1\big)/(1-\zeta) \le \zeta\delta^{\zeta-1} n^{1/\zeta}/(1-\zeta), & \zeta < 1, \\ n \log b = n \log(n/\delta), & \zeta = 1. \end{cases}$$

By Markov's inequality, for $K > 0$, T is at most K times its expected value with probability at least $1 - 1/K$.

Combining. With probability at least $1 - e^{-\varepsilon^{-\zeta}} - \delta^\zeta - 1/K$ (or $1 - e^{-\varepsilon^{-\zeta}}$, for N large enough, when $\zeta > 1$),

$$\frac{A_N}{M_N^2} \le \begin{cases} c\, n^{1-2/\zeta}, & 1 < \zeta < 2, \\ c\, \log(n/\delta)/n, & \zeta = 1, \\ c\, n^{-1/\zeta}, & \zeta < 1, \end{cases}$$

where c depends only on ε, δ and K. Each bound tends to zero as N grows, since $2/\zeta > 1$, and the probability can be made arbitrarily close to one by taking ε and δ small and K large. Hence $A_N/M_N^2 \to 0$ in probability, and so does $n_{\text{eff},N}$; for $1 < \zeta < 2$ the rate is $N^{1-2/\zeta}$.

(iii) The number of relationships set aside is $m_N = \lfloor \alpha S_N/V_b \rfloor$, and $m_N/N \to \phi = \alpha\mu/V_b < 1$ almost surely. The relationships kept are those below the empirical $1 - m_N/N$ quantile of the amounts, which converges almost surely to $\bar{\ell}$, the $1 - \phi$ quantile of F, since F is continuous and increasing there. For every $\epsilon > 0$, the sums of L and L^2 over the kept relationships of industry j, divided by N, eventually lie between their values with thresholds $\bar{\ell} - \epsilon$ and $\bar{\ell} + \epsilon$, which converge by the strong law of large numbers to $s\mu_{\bar{\ell}\mp\epsilon}$ and $s\nu_{\bar{\ell}\mp\epsilon}$. Letting $\epsilon \to 0$, they converge to $s\mu_{\bar{\ell}}$ and $s\nu_{\bar{\ell}}$, and with $S_N/N \to \mu$, $n_{\text{eff},N}^T \to V_b s^2 \mu_{\bar{\ell}}^2/(\mu\, s\, \nu_{\bar{\ell}})$.

(iv) Let $b_N = N \log N$. By the union bound, no amount exceeds b_N with probabil-

ity at least $1 - N/b_N = 1 - 1/\log N$, and on that event S_N equals the truncated sum $T_N = \sum_{i \le N} L_i \mathbb{1}\{L_i \le b_N\}$. For $F(x) = 1 - 1/x$, $E[L\,\mathbb{1}\{L \le b\}] = \int_1^b x^{-1} dx = \log b$ and $E[L^2\,\mathbb{1}\{L \le b\}] = \int_1^b dx = b - 1$, so $E[T_N] = N \log b_N = N(\log N + \log\log N)$ and $\mathrm{Var}(T_N) \le N(b_N - 1) < N^2 \log N$. By Chebyshev's inequality, for $\eta > 0$,

$$P\Big(|T_N - E[T_N]| > \eta N \log N\Big) < \frac{N^2 \log N}{\eta^2 N^2 (\log N)^2} = \frac{1}{\eta^2 \log N} \to 0.$$

Since $E[T_N]/(N \log N) \to 1$, $T_N/(N \log N) \to 1$ in probability, and since $S_N = T_N$ with probability tending to one, $S_N/(N \log N) \to 1$ in probability. Hence, with $N\pi_b = V_b N/S_N$, $N\pi_b \log N/V_b = N \log N/S_N \to 1$. □

C Sparsity and concentration

Sparsity affects any statistic built from shares, not only measures of specialization. Concentration is an example. When the firms of a market are split into finer industries, each industry has fewer borrowers, and the few largest lenders in an industry account for more of its credit by construction; with a single borrower per industry, the three largest lenders hold all of it. Concentration measured within industries therefore rises with the fineness of the industry partition, whether or not lenders specialize. To measure how much of the rise is mechanical, we compare the observed concentration with the concentration produced by random matching in the same cells, as Ellison and Glaeser (1997) compare the geographic concentration of industries with the concentration produced by random plant locations.

Table C.1 reports this comparison for French credit markets, with the data and market definitions of Section 3.1. For each market, it reports the share of credit granted by the three largest branches, on all credit and within each industry (averaged across industries with credit weights), observed and under the random matching of Appendix A (mean over 100 draws). Random matching keeps the lending volumes of the largest branches close to their actual values in expectation (Table A.1), so it reproduces the concentration of all credit almost exactly (in the median urban unit, 0.79 observed and 0.78 under random matching). Within industries, it reproduces most of the rise in concentration in small markets. In the median urban unit, the observed top-3 share rises from 0.79 on all credit to 0.84 by industry at NAF level 0 and 0.88 at NAF level 3, and random matching alone gives 0.82 and 0.86: most of the rise is mechanical. In larger markets, observed concentration within industries clearly exceeds random matching. In the median région it is 0.41 at NAF level 0 against 0.34, and it is above the 95th percentile of the random-matching draws in 81% of régions; nationally, 0.15 against 0.09. Comparisons of concentration across industry levels, like comparisons of portfolio shares, are informative only against a counterfactual that holds sparsity fixed, which is what the random-matching percentile p does branch by branch.

Table C.1: Concentration of lending: observed and under random matching, 2016Q1

			Top-3 share (median)		Observed above 95th pct. (%)
Geography	Credit	Markets	Observed	Random matching	
France	All credit	1	0.09	0.09	0
	By industry, NAF 0		0.15	0.09	100
	By industry, NAF 1		0.15	0.09	100
	By industry, NAF 2		0.17	0.10	100
	By industry, NAF 3		0.19	0.12	100
Regions	All credit	21	0.35	0.32	0
	By industry, NAF 0		0.41	0.34	81
	By industry, NAF 1		0.42	0.34	90
	By industry, NAF 2		0.45	0.35	95
	By industry, NAF 3		0.48	0.37	100
Départements	All credit	94	0.40	0.41	0
	By industry, NAF 0		0.48	0.42	62
	By industry, NAF 1		0.49	0.43	71
	By industry, NAF 2		0.51	0.45	76
	By industry, NAF 3		0.54	0.47	82
Urban areas	All credit	537	0.77	0.75	19
	By industry, NAF 0		0.83	0.78	47
	By industry, NAF 1		0.84	0.78	51
	By industry, NAF 2		0.85	0.80	54
	By industry, NAF 3		0.87	0.83	52
Urban units	All credit	734	0.79	0.78	16
	By industry, NAF 0		0.84	0.82	41
	By industry, NAF 1		0.85	0.83	44
	By industry, NAF 2		0.87	0.84	47
	By industry, NAF 3		0.88	0.86	44

Notes: Share of a market's credit granted by its three largest branches. All credit: computed on all credit of the market. By industry: computed within each industry and averaged across industries with weights equal to their share of the market's credit. Observed: in the data. Random matching: mean over 100 draws of the random matching of Appendix A, Steps 5–6. Medians across markets with at least four branches (with three or fewer, the share is one by construction). Last column: share of markets whose observed value exceeds the 95th percentile of its 100 random-matching values.

D Specialization in France: additional figures and tables

In this appendix we collect the figures and tables that complement Sections 3.1 and 3.2.

By market definition. Tables D.1, D.2 and D.3 report, for each of the twenty market definitions, the share-based measures of specialization observed and under random matching (Section 3.1), the distribution of p (Section 3.2.1) and the sparsity diagnostics (Section 3.2.2) that Figure 3 summarizes.

Table D.1: Share-based measures of specialization: observed and under random matching, 2016Q1

		Top-quartile branches: relative portfolio share			Favorite industry: share of the portfolio	
Geography	Industry	Observed	Random matching	Cells above	Observed	Random matching
France	NAF 0	2.43	2.21	44%	0.67	0.39
	NAF 1	2.48	3.49	35%	0.67	0.35
	NAF 2	7.33	11.12	19%	0.67	0.32
	NAF 3	16.87	28.19	16%	0.63	0.29
Regions	NAF 0	2.48	2.81	44%	0.67	0.36
	NAF 1	3.22	3.84	40%	0.67	0.32
	NAF 2	8.73	12.03	29%	0.67	0.29
	NAF 3	17.84	27.45	30%	0.62	0.26
Départements	NAF 0	2.49	2.45	50%	0.67	0.41
	NAF 1	3.05	3.35	45%	0.67	0.39
	NAF 2	6.84	8.98	37%	0.66	0.36
	NAF 3	11.01	15.52	37%	0.62	0.32
Urban areas	NAF 0	2.01	1.76	62%	0.68	0.42
	NAF 1	2.28	2.04	57%	0.67	0.39
	NAF 2	3.25	2.96	50%	0.67	0.36
	NAF 3	4.00	3.70	47%	0.63	0.32
Urban units	NAF 0	1.85	1.63	61%	0.67	0.42
	NAF 1	2.06	1.85	57%	0.67	0.38
	NAF 2	2.75	2.50	51%	0.66	0.35
	NAF 3	3.28	2.97	49%	0.62	0.31

Notes: Top-quartile branches: in each market × industry cell, the branches whose portfolio share in the industry is positive and at or above the 75th percentile of the shares of all branches of the market (zero shares included); mean relative portfolio share of these branches, S_{bj}/s_j; median across cells. Favorite industry: share of the branch's portfolio in the industry with the largest excess share $S_{bj} - s_j$; median across branches. Observed: in the data. Random matching: median over 100 draws of the random matching of Appendix A, Steps 5–6, cell by cell (branch by branch for the favorite industry), then median across cells (branches). We use the median draw, the typical outcome of random matching, because the top-quartile statistic is right-skewed across draws: a small branch that receives one relationship of a small industry has a very large relative portfolio share, and such rare draws pull the mean above the typical draw. Cells above: share of cells in which the observed value exceeds the median draw.

Table D.2: Distribution of p by market definition, 2016Q1

					Quantiles					
Geography	Industry	N	Mean	S.d.	10%	25%	50%	75%	90%	Mean, largest / other branches
France	NAF 0	1,814,741	0.70	0.33	0.14	0.47	0.84	0.99	1.00	0.90 / 0.70
	NAF 1	1,814,741	0.71	0.32	0.15	0.48	0.85	0.99	1.00	0.95 / 0.71
	NAF 2	1,814,741	0.73	0.31	0.20	0.51	0.88	0.99	1.00	0.96 / 0.73
	NAF 3	1,814,740	0.74	0.30	0.26	0.49	0.88	0.99	1.00	0.96 / 0.74
Regions	NAF 0	1,814,741	0.68	0.35	0.05	0.40	0.85	0.99	1.00	0.56 / 0.69
	NAF 1	1,814,740	0.69	0.35	0.06	0.42	0.86	0.99	1.00	0.57 / 0.70
	NAF 2	1,814,713	0.71	0.34	0.09	0.46	0.88	1.00	1.00	0.57 / 0.72
	NAF 3	1,814,465	0.72	0.33	0.15	0.46	0.88	0.99	1.00	0.57 / 0.72
Départements	NAF 0	1,814,739	0.68	0.37	0.04	0.37	0.87	1.00	1.00	0.58 / 0.69
	NAF 1	1,814,731	0.69	0.36	0.04	0.38	0.88	1.00	1.00	0.56 / 0.70
	NAF 2	1,814,350	0.70	0.35	0.06	0.43	0.89	1.00	1.00	0.56 / 0.72
	NAF 3	1,812,666	0.71	0.34	0.11	0.44	0.89	1.00	1.00	0.60 / 0.72
Urban areas	NAF 0	1,715,181	0.67	0.36	0.04	0.37	0.85	0.99	1.00	0.56 / 0.70
	NAF 1	1,714,546	0.68	0.36	0.04	0.39	0.87	0.99	1.00	0.58 / 0.71
	NAF 2	1,709,172	0.70	0.35	0.07	0.44	0.88	1.00	1.00	0.59 / 0.72
	NAF 3	1,699,730	0.71	0.34	0.11	0.44	0.88	0.99	1.00	0.59 / 0.73
Urban units	NAF 0	1,727,440	0.67	0.37	0.03	0.36	0.85	0.99	1.00	0.56 / 0.70
	NAF 1	1,726,077	0.68	0.36	0.04	0.38	0.86	0.99	1.00	0.57 / 0.71
	NAF 2	1,717,380	0.70	0.35	0.06	0.42	0.87	1.00	1.00	0.58 / 0.72
	NAF 3	1,702,714	0.70	0.34	0.10	0.43	0.87	0.99	1.00	0.59 / 0.73

Notes: One observation per relationship with p defined; p from Appendix A with 1,000 draws per market. Unweighted statistics across relationships. Last column: mean p of the relationships of the largest branch of each market (by lending) and of the other branches.

Table D.3: Sparsity diagnostics by market definition, 2016Q1

		$P(A>0)$ (%)		Ties	AUC(2) quartiles						
Geography	Industry	< 0.9	< 0.5	(mean)	25%	50%	75%	AUC(2) ≥ 0.9 (%)	$2/\kappa_{\max} > 1$ (%)	Median $2/\kappa_{\max}$	Median ceiling
France	NAF 0	6	1	0.00	0.76	0.89	0.99	49	0	0.33	1.00
	NAF 1	10	3	0.01	0.71	0.86	0.99	43	0	0.21	1.00
	NAF 2	20	9	0.04	0.66	0.83	0.99	41	0	0.11	1.00
	NAF 3	35	20	0.08	0.63	0.77	0.98	36	0	0.04	1.00
Regions	NAF 0	5	1	0.00	0.78	0.92	1.00	55	10	0.36	1.00
	NAF 1	9	2	0.01	0.74	0.88	1.00	47	10	0.27	1.00
	NAF 2	18	8	0.04	0.68	0.85	0.99	45	10	0.20	1.00
	NAF 3	31	17	0.08	0.65	0.79	0.99	42	4	0.10	1.00
Départements	NAF 0	4	1	0.00	0.82	0.95	1.00	61	15	0.45	1.00
	NAF 1	7	2	0.01	0.77	0.92	1.00	53	14	0.40	1.00
	NAF 2	16	7	0.03	0.70	0.88	1.00	48	14	0.37	1.00
	NAF 3	29	15	0.07	0.66	0.82	1.00	44	11	0.26	1.00
Urban areas	NAF 0	5	1	0.00	0.80	0.94	1.00	59	17	0.51	1.00
	NAF 1	8	2	0.01	0.75	0.90	1.00	50	17	0.49	1.00
	NAF 2	16	7	0.03	0.69	0.87	1.00	46	16	0.48	1.00
	NAF 3	28	15	0.07	0.66	0.81	1.00	42	13	0.40	1.00
Urban units	NAF 0	4	1	0.00	0.81	0.95	1.00	60	19	0.58	1.00
	NAF 1	7	2	0.01	0.75	0.91	1.00	51	19	0.57	1.00
	NAF 2	16	7	0.03	0.69	0.87	1.00	47	19	0.56	1.00
	NAF 3	27	14	0.06	0.66	0.82	1.00	43	16	0.48	1.00

Notes: Relationships with p defined. $P(A > 0)$: share of valid counterfactual draws in which the branch has positive lending to the firm's industry from other borrowers; columns: share of relationships with $P(A > 0)$ below 0.9 and below 0.5. Ties: mean probability that the counterfactual share equals the observed one. On a fixed random sample of about 21,000 relationships per market definition, with the trimmed screening rule of Appendix B.3: AUC(2), equation (7) at $\kappa = 2$ (quartiles and share of at least 0.9, among relationships for which a doubling is feasible); share for which it is not, $2/\kappa_{\max} > 1$; median of $2/\kappa_{\max}$; median ceiling AUC($\kappa_{\max}$).

Robustness to the degree of specialization. Table D.4 reports the median AUC of p at $\kappa = 1.5$, 2 and 3, and the share of relationships for which each degree of specialization is infeasible (Section 3.2.2).

Table D.4: The AUC of p at three degrees of specialization, by market definition, 2016Q1

		$\kappa = 1.5$		$\kappa = 2$		$\kappa = 3$	
Geography	Industry	Median AUC	Infeasible (%)	Median AUC	Infeasible (%)	Median AUC	Infeasible (%)
France	NAF 0	0.74	0	0.89	0	0.94	39
	NAF 1	0.71	0	0.86	0	0.88	39
	NAF 2	0.69	0	0.83	0	0.80	39
	NAF 3	0.65	0	0.77	0	0.74	37
Regions	NAF 0	0.79	0	0.92	10	0.98	34
	NAF 1	0.76	0	0.88	10	0.93	34
	NAF 2	0.74	0	0.85	10	0.88	34
	NAF 3	0.69	0	0.79	4	0.79	31
Départements	NAF 0	0.83	2	0.95	15	0.99	37
	NAF 1	0.80	2	0.92	14	0.96	37
	NAF 2	0.79	2	0.88	14	0.91	36
	NAF 3	0.73	2	0.82	11	0.83	33
Urban areas	NAF 0	0.83	2	0.94	17	0.99	40
	NAF 1	0.80	2	0.90	17	0.95	40
	NAF 2	0.78	2	0.87	16	0.90	40
	NAF 3	0.73	2	0.81	13	0.83	36
Urban units	NAF 0	0.84	4	0.95	19	0.99	43
	NAF 1	0.81	4	0.91	19	0.95	43
	NAF 2	0.78	4	0.87	19	0.90	42
	NAF 3	0.74	3	0.82	16	0.83	39

Notes: On the fixed random sample of Table D.3, with the trimmed screening rule of Appendix B.3. Median AUC: median of AUC(κ), equation (7), among relationships for which $\kappa \leq \kappa_{\max}$. Infeasible: share of relationships with $\kappa > \kappa_{\max}$.

Maps. Figure D.1 maps the mean p of the relationships of each urban area at the four industry levels, and Figure D.2 at NAF level 2 under the four geographic partitions, on the color scale of Figure 4 (Section 3.2.3).

Finer partitions. Table D.5 reports the level of specialization and the variation beyond random matching uncovered by finer geographic and industry partitions, the statistics of Section 3.2.3.

Sparse cells, all geographies. Figure D.3 extends Figure 5 (panels A and B) to the five geographic partitions.

Share-based measures. Table D.6 reports the mean p and the rank correlations with p of the three share-based measures of Section 3.2.4, by market definition.

The closest-branch assumption. Table D.7 reports how often, and how far, the assumption that firms borrow from the closest branch of their bank errs (Section 3.2.4).

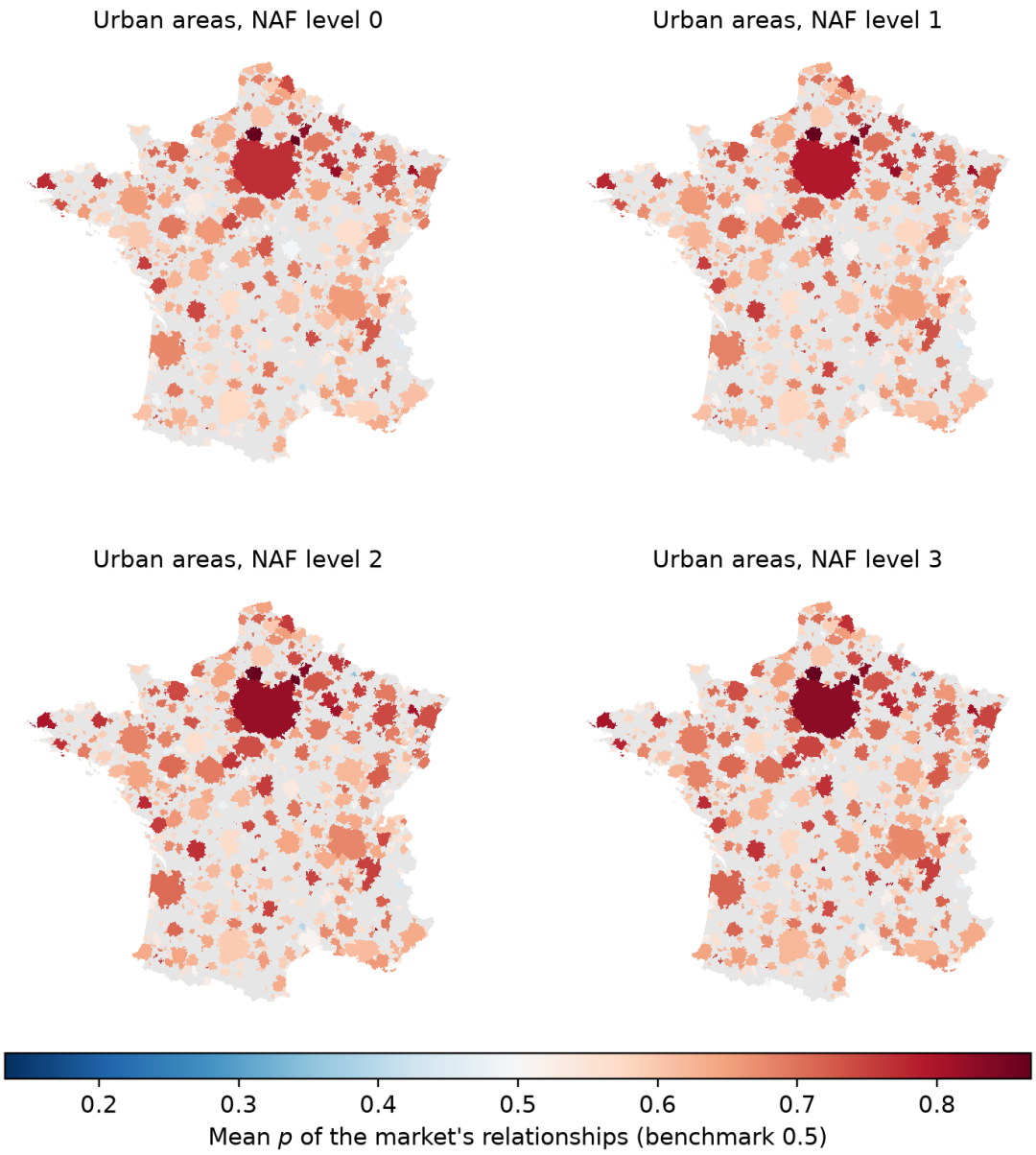


Figure D.1: Mean p by urban area at four industry levels, 2016Q1

Notes: Mean p of the relationships of the branches of each urban area, with industries defined at NAF levels 0 to 3. Common diverging color scale centered at 0.5, the expected value of p without specialization. Grey: municipalities outside urban areas, and urban areas with a single branch. Metropolitan France without Corsica.

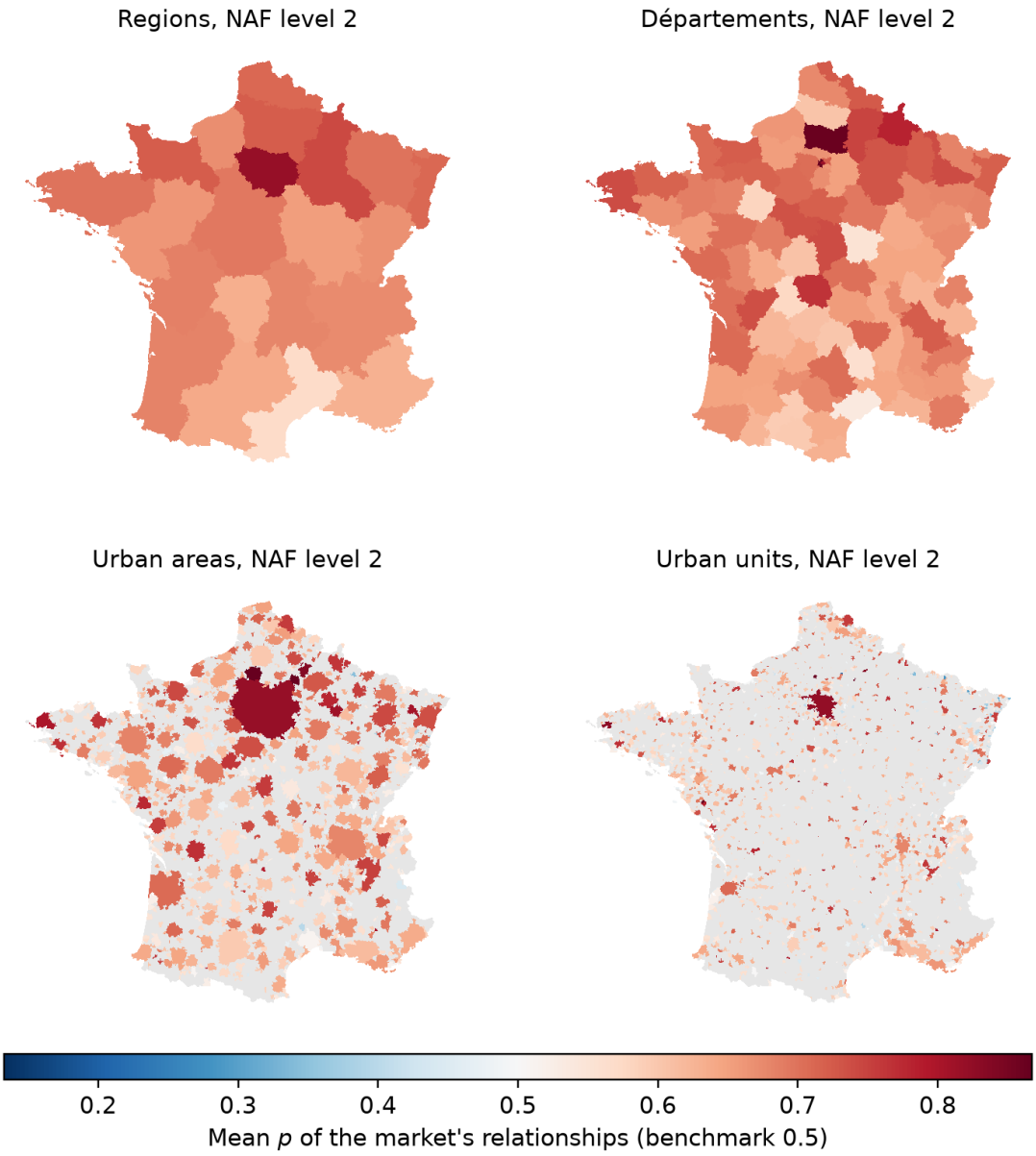


Figure D.2: Mean p under four geographic partitions, NAF level 2, 2016Q1

Notes: Mean p of the relationships of the branches of each market, with industries at NAF level 2 and markets defined as régions, départements, urban areas and urban units. Same color scale as Figure D.1. Grey: municipalities outside any market of the partition, and markets with a single branch.

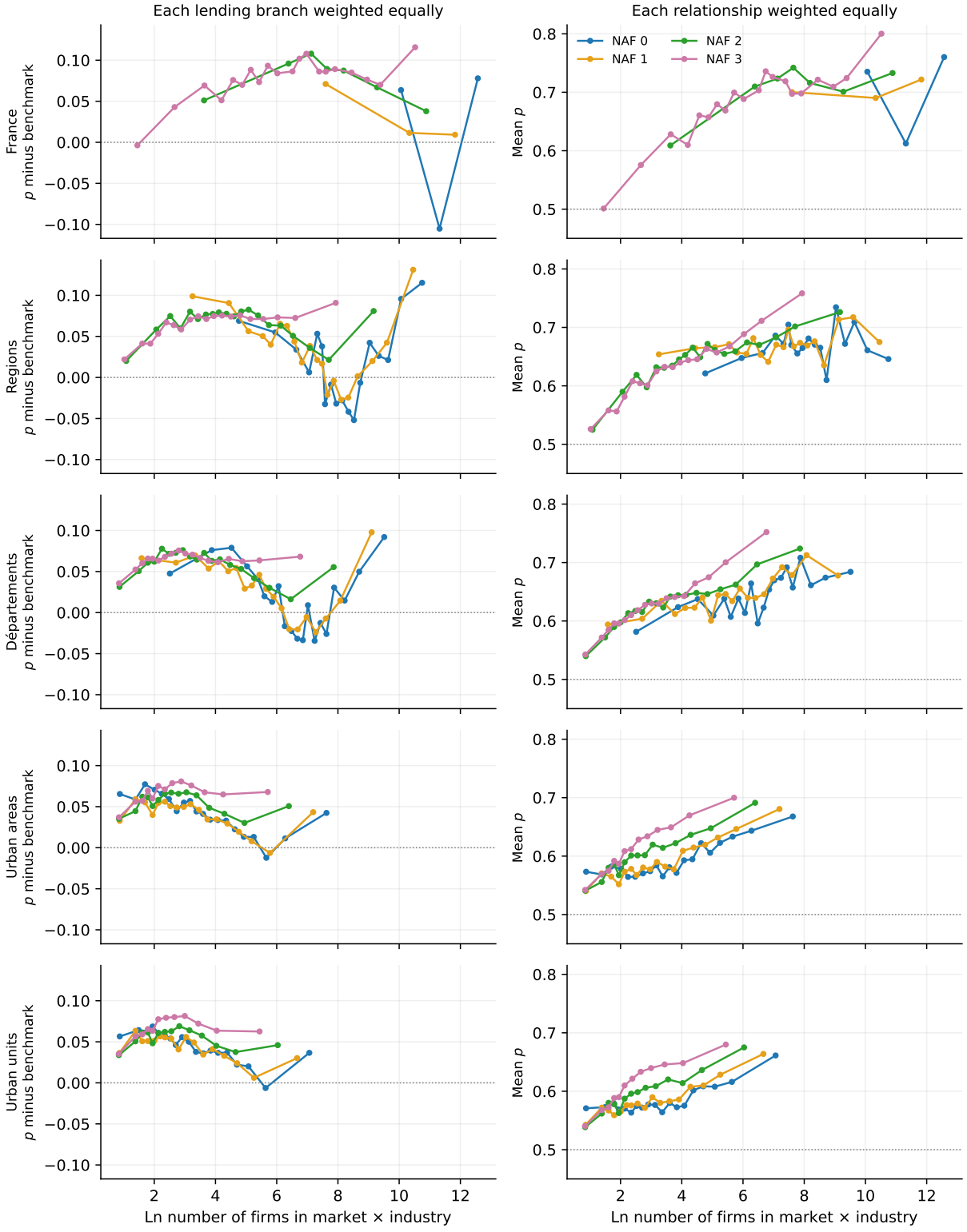


Figure D.3: Specialization and the sparsity of markets, 2016Q1

Notes: One row per geographic partition. Cells: market × industry, at NAF levels 0 to 3, with at least two borrowing firms. Right: mean p over the relationships of the cell (each relationship weighted equally); its benchmark without specialization is 1/2 (dotted line). Left: mean over the branches that lend to the industry of the branch's mean p (each lending branch weighted equally), minus the same mean on allocations of the relationships drawn by the random matching of Appendix A (Steps 5–6) and treated as data (20 allocations for France and régions, 10 for départements, 5 for urban areas and urban units); this weighting moves the benchmark below 1/2, which is why the panel is centered at zero (Section 3.2.3). Means of the cells in up to 20 equal-sized bins of the log number of borrowing firms.

Table D.5: Finer partitions: the level of specialization and the variation they uncover, 2016Q1

	Level: mean p		Variation: p			Variation: relative share
	Coarser	Finer	Beyond random matching	Within	Between	Within
Finer geography (NAF level 2)						
Urban units in urban areas	0.702	0.696	65%	0.057	0.057	0.60
Urban areas in départements	0.702	0.702	64%	0.053	0.051	0.71
Départements in régions	0.711	0.702	88%	0.051	0.042	1.09
Finer industries (urban areas)						
NAF 1 in NAF 0	0.674	0.685	19%	0.069	0.123	5.56
NAF 2 in NAF 1	0.685	0.702	27%	0.075	0.109	19.74
NAF 3 in NAF 2	0.702	0.707	16%	0.053	0.097	18.36

Notes: Each row nests smaller cells in larger units: smaller markets in the larger geographic unit holding most of their lending (industries at NAF level 2), or finer industries in the coarser industry of the same urban area. Level: mean p across relationships, in the coarser and in the finer partition. Variation, p: dispersion across the smaller cells of the same larger unit of their mean p (mean over relationships). Beyond random matching: share of the variance within larger units that exceeds the variance on an allocation drawn by random matching and treated as data. Within and Between: standard deviation beyond random matching (square root of the variance in the data minus the variance under random matching) within the larger units and between them, after removing the relation between a cell's mean p and the log of its number of relationships. Relative share: the same within statistic for the leave-one-out relative portfolio share. Panel B of Figure 4 plots the Within and Between columns and the share beyond random matching.

Table D.6: p and share-based measures of specialization: rank correlations, 2016Q1

			Rank correlation with p		
Geography	Industry	Mean p	Excess share	Share rank	Relative share
France	NAF 0	0.70	0.90	0.69	0.91
	NAF 1	0.71	0.89	0.70	0.89
	NAF 2	0.73	0.84	0.69	0.86
	NAF 3	0.74	0.81	0.71	0.86
Regions	NAF 0	0.68	0.90	0.68	0.88
	NAF 1	0.69	0.90	0.69	0.87
	NAF 2	0.71	0.86	0.68	0.84
	NAF 3	0.72	0.83	0.69	0.84
Départements	NAF 0	0.68	0.87	0.69	0.85
	NAF 1	0.69	0.86	0.69	0.84
	NAF 2	0.70	0.83	0.68	0.81
	NAF 3	0.71	0.81	0.69	0.81
Urban areas	NAF 0	0.67	0.87	0.65	0.84
	NAF 1	0.68	0.86	0.66	0.83
	NAF 2	0.70	0.83	0.65	0.81
	NAF 3	0.71	0.81	0.66	0.81
Urban units	NAF 0	0.67	0.87	0.64	0.85
	NAF 1	0.68	0.87	0.65	0.83
	NAF 2	0.70	0.84	0.64	0.82
	NAF 3	0.70	0.81	0.65	0.81

Notes: Relationships with p defined. Mean p: mean across relationships (Table D.2). Rank correlation: Spearman correlation with p across relationships. Excess share: leave-one-out portfolio share S_{bj}^{-r} minus the industry's leave-one-out share of market lending. Share rank: mid-percentile of the branch's leave-one-out share in the firm's industry among all branches of the market. Relative share: leave-one-out share divided by the industry's leave-one-out share of market lending.

Table D.7: The closest-branch assumption: how often and how far it errs, 2016Q1

				Wrong branch		When wrong			
Firm located in	Relationships	Actual not closest	Tie	Share	Credit-weighted	Km, firm to actual	Km, actual to assumed	Other urban unit	Other urban area
All	1,800,067	29%	51%	61%	49%	6	0	48%	44%
Outside urban units	479,990	54%	49%	67%	66%	30	30	81%	81%
UU without branches	55,354	40%	23%	51%	48%	18	19	78%	53%
UU Q1 (fewest branches)	48,436	38%	14%	44%	48%	21	22	86%	58%
UU Q2	63,539	33%	10%	38%	45%	26	26	88%	65%
UU Q3	97,597	33%	10%	37%	41%	31	29	89%	67%
UU Q4 (most branches)	1,055,151	16%	62%	63%	48%	0	0	26%	23%

Notes: One observation per relationship. Firms and branches are located at the centroid of their urban unit, or of the part of their département outside urban units; rows by the firm's location, with urban units in quartiles of their number of branches. Actual not closest: the actual branch is outside the closest urban unit in which the firm's bank has a branch. Tie: the bank has several branches in that urban unit. Wrong branch: the assumed branch (the closest one, ties broken at random) differs from the actual branch, in share of relationships and of credit. When wrong: median distance (km, between centroids) from the firm to the actual branch and from the actual to the assumed branch; share of wrong assignments in which the assumed branch lies in another urban unit or urban area.

E Branch closures: additional material

E.1 Closures over time and space

Figure E.1 shows the number of closures per year, and Figure E.2 where they take place (Section 4.2).

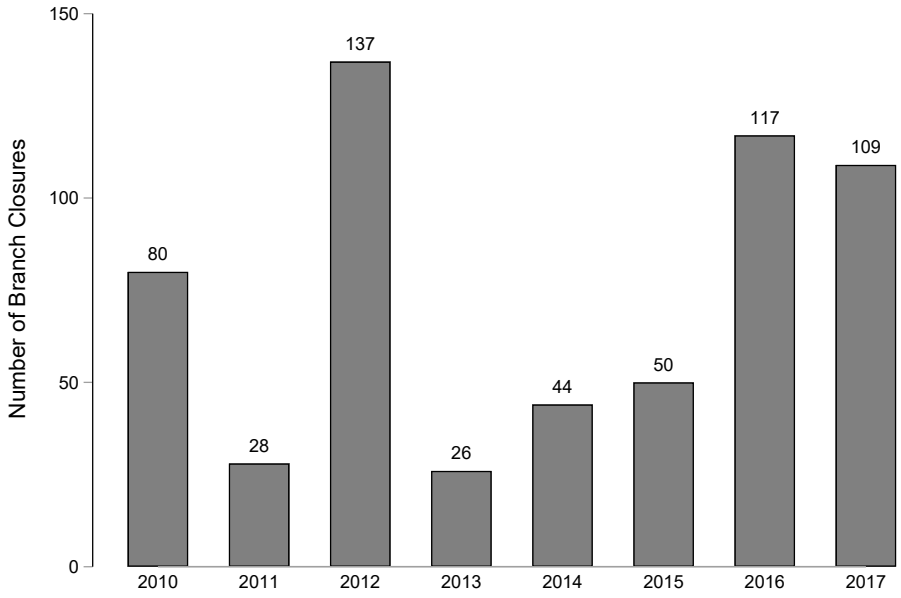


Figure E.1: Bank branch closures in France, 2010–2017

Notes: Number of branch closures per year, 2010Q1 to 2017Q4. Branches lending to SMEs according to the credit register; before 2015, branches of banks with permanent account identifiers (*RIB permanent*).

E.2 What determines branch closures?

The design of Section 4.1 does not require that closures be random. Here we describe which branches close. Banks do not disclose how they choose the branches they close; their annual reports refer to general developments such as online banking rather than to local conditions. Table E.1 decomposes the variance of an indicator of closure between 2010 and 2017 across branches. Bank fixed effects explain 32% of it; département and urban unit fixed effects explain 3% and 8% on their own, and add 0.1 and 4.2 percentage points to the bank effects. Which branches close is thus mostly a matter of which bank they belong to. Table E.2 relates the closure of a branch to its characteristics before 2010. A branch is less likely to close when it is the only branch of its bank in its urban unit, and more likely to close the larger the urban unit's population growth between 1990 and 2006, although the latter association disappears with bank fixed effects. Within the bank, a branch's rank by size in the bank and in the bank and urban unit predicts closure, with or without bank fixed effects. In this table a branch is specialized if its portfolio share in an industry is in the top quartile of its market, the rule of Paravisini et al. (2023), and specialization does not predict closure in any column.

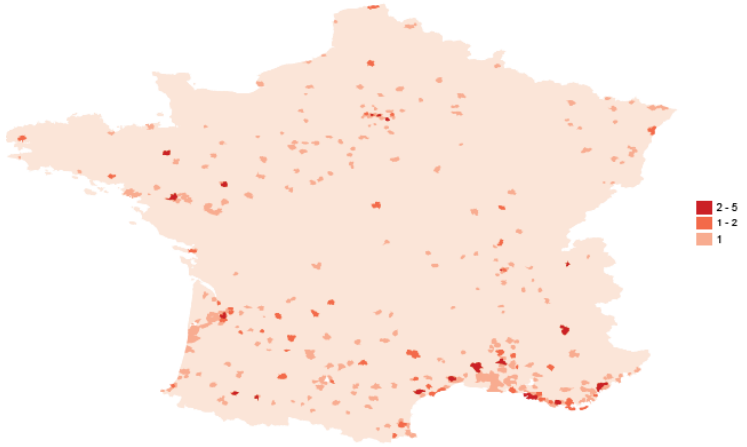


Figure E.2: The geography of bank branch closures in France, 2010–2017

Notes: Branch closures over 2010–2017, by location of the branch (postal code). Branches lending to SMEs according to the credit register; before 2015, branches of banks with permanent account identifiers.

Table E.1: What determines branch closures? Analysis of variance

	County FE (1)	Urban Unit FE (2)	Banking Group FE (3)	Bank FE (4)	(4) + County FE (5)	(4) + UU FE (6)
R-Square	0.03	0.08	0.02	0.32	0.32	0.35
Explained by Bank FE					0.296	0.276
Explained by County FE					0.001	
Explained by UU FE						0.042

Notes: Branches active in the sample. Dependent variable: indicator of a closure between 2010 and 2017. Partial analysis of variance, adding bank, banking group, département ("County") and urban unit fixed effects; R^2 and the share of the variance explained by each set of fixed effects.

Table E.2: What determines branch closures? Branch characteristics

	P(closure)				
	(1)	(2)	(3)	(4)	(5)
Specialized bank branch	0.296	0.298	0.530	0.290	0.206
	(0.388)	(0.390)	(0.382)	(0.390)	(0.388)
Bank branch size (Total Loan) in 2008	0.000	-0.076**		-0.073*	-0.059
	(0.023)	(0.038)		(0.038)	(0.049)
Market share (Branch total loan / UU total loan) in 2008	-0.320	-0.461		-0.584*	-0.575
	(0.281)	(0.304)		(0.330)	(0.361)
Dummy (Only branch of the bank in the UU in 2008)	-0.698***	-0.967***		-0.994***	-1.139***
	(0.138)	(0.171)		(0.174)	(0.180)
Branch rank in the bank x UU in 2008		0.696***		0.697***	0.526**
		(0.244)		(0.243)	(0.255)
Branch rank in the bank in 2008		-1.024***		-1.033***	-0.904***
		(0.280)		(0.280)	(0.328)
UU size (population in 2006)			0.115***	-0.028	-0.189***
			(0.034)	(0.045)	(0.048)
UU size variation: 1990-2006			0.281***	0.295***	0.138
			(0.069)	(0.066)	(0.161)
Banking Group FE	✓	✓	✓	✓	✓
Bank FE	×	×	×	×	✓
Observations	14,580	14,580	14,769	14,580	14,005
Adj. R-Square	0.09	0.09	0.08	0.09	0.16

Notes: Branches active in the sample. Dependent variable: indicator of a closure between 2010 and 2017. Explanatory variables measured before the period: 2008 for branch variables, 2006 and 1990–2006 for urban-unit population (French census). "Specialized bank branch": the branch's portfolio share in an industry is in the top quartile of its market at least one year before the closure. Rank: rank of the branch by size within the bank (or within the bank and urban unit), 1 for the largest. Credit and population in logs.

E.3 Additional results

Table E.3 describes the sample of industry cohorts of Section 4.1, and Figure E.3 shows the distribution of the change in specialization across them. Table E.4 reports the means of the two outcomes and how new lending to future defaulters is distributed across cohorts. Figures E.4 and E.5 report the event studies of new lending and of new lending to future defaulters (Section 4.3.1). Table E.5 splits the change in specialization by sign for new lending, as Table 8 does for new lending to future defaulters. Table E.6 measures new lending to future defaulters net of the default rate of comparable new loans, and Figure E.6 reports excess lending to future defaulters for the 16 market definitions and the three measures of specialization (Section 4.3.2). Table E.7 and Figure E.7 report the substitution regression by sign of the change in specialization and by market definition (Section 4.3.3).

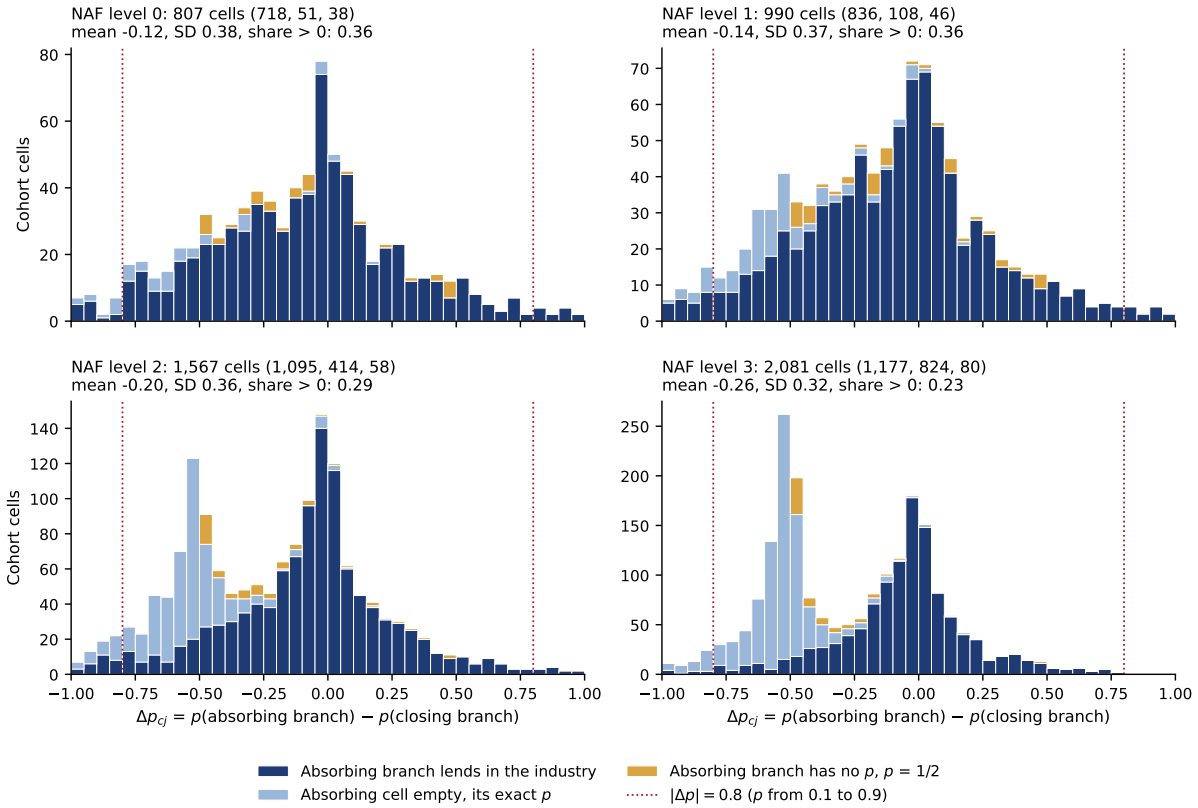


Figure E.3: Change in the specialization of the lending branch at a closure: industry cohorts

Notes: Histogram of $\Delta p_{cj} = p_{a(c)j} - p_{cj}$, the regressor of Table 7, one observation per industry cohort with a Δp_{cj}, markets of départements, NAF levels 0 to 3, bins of 0.05. Stacked by how $p_{a(c)j}$ is measured in quarter -7: $a(c)$ lends to industry j, and $p_{a(c)j}$ is the p of its share of lending in the industry, with no relationship left out; it lends in the market but not to the industry, and $p_{a(c)j}$ is the exact p of the empty cell; it has no p, and $p_{a(c)j} = 1/2$. Panel titles: the number of cohorts in total and in the three groups, the mean and standard deviation of Δp_{cj} across all cohorts, and the share of cohorts with $\Delta p_{cj} > 0$. Table 7 reports instead the standard deviation within closing branch, in its regression sample. Dotted lines: $|\Delta p_{cj}| = 0.8$, a move from $p = 0.1$ to $p = 0.9$.

Table E.3: Industry cohorts: the sample

NAF level	0	1	2	3
A. Sample				
Closing branches (clusters)	185	200	203	208
Cohorts (closing branch x industry)	605	793	1,330	1,807
Cohorts per closing branch, mean	3.3	4.0	6.6	8.7
Cohorts per closing branch, median	3	3	3	4
Cohorts per closing branch, range	2–8	2–12	2–32	2–56
Firms	2,775	2,809	2,788	2,766
Cohort-periods	1,210	1,586	2,660	3,614
B. Cohorts by the source of $p_{a(c)j}$ (mean $p_{a(c)j}$ in brackets)				
$a(c)$ lends to j: simulated p	530	660	918	1,023
	[0.58]	[0.62]	[0.72]	[0.79]
$a(c)$ does not lend to j: exact p of the empty cell	43	93	360	712
	[0.17]	[0.23]	[0.33]	[0.37]
$a(c)$ has no p at -7: set to 1/2	32	40	52	72
C. Moved firms, per closing branch				
Cohort firms that moved, share: mean	0.81	0.82	0.82	0.81
median	0.87	0.87	0.87	0.87
pooled	0.86	0.86	0.86	0.85
Cohort's credit at -7 that moved, share: mean	0.85	0.85	0.85	0.84
median	0.95	0.96	0.96	0.96
pooled	0.93	0.93	0.93	0.93
Moved cohort firms that went to $a(c)$, share: mean	0.88	0.88	0.88	0.87
median	1.00	1.00	1.00	1.00
pooled	0.90	0.90	0.90	0.90
Their credit at -7 that went to $a(c)$, share: mean	0.87	0.87	0.87	0.86
median	1.00	1.00	1.00	1.00
pooled	0.88	0.88	0.88	0.88

Notes: The regression sample of Table 7: industry cohorts of closing branches, markets of départements, NAF levels 0 to 3. Panel A: closing branches, cohorts and firms. Panel B: cohorts by how the p of the main absorbing branch $a(c)$ in the cohort's industry is measured in quarter -7, with its mean in brackets. Panel C: per closing branch, the share of the cohort's firms, and of the cohort's credit from the bank in quarter -7, that move to another branch of the bank, and the share of these firms, and of their credit, that move to $a(c)$: mean and median across closing branches, and pooled over all of them.

Table E.4: Industry cohorts: new lending and new lending to future defaulters before and after the announcement

NAF level	0	1	2	3
A. Means, weighted by the regression weight (SD in brackets)				
New lending from the bank, pre	0.0523	0.0529	0.0541	0.0557
	[0.0713]	[0.0778]	[0.0919]	[0.1030]
New lending from the bank, post	0.0286	0.0284	0.0289	0.0290
	[0.0325]	[0.0391]	[0.0573]	[0.0663]
New lending to future defaulters, pre	0.000601	0.000600	0.000658	0.000663
	[0.002991]	[0.003754]	[0.006844]	[0.010798]
New lending to future defaulters, post	0.000382	0.000394	0.000441	0.000447
	[0.003948]	[0.004344]	[0.005324]	[0.009454]
B. New lending to future defaulters				
Share of new lending, pre	0.0114	0.0113	0.0113	0.0094
Share of new lending, post	0.0123	0.0123	0.0123	0.0123
Cohort-periods with any, share, pre	0.048	0.038	0.023	0.015
Cohort-periods with any, share, post	0.063	0.053	0.033	0.025
weighted by the regression weight, pre	0.169	0.130	0.042	0.022
weighted by the regression weight, post	0.189	0.139	0.044	0.016
Cohort-periods with any, number	67	72	75	73
C. Across the cohort-periods with any: EUR thousand				
Percentile 10	7	7	7	7
Percentile 25	13	12	11	11
Median	50	48	43	41
Percentile 75	100	91	93	81
Percentile 90	296	253	241	222
Maximum	1,092	1,092	1,092	1,092
Mean	112	105	101	95
D. Across the cohort-periods with any: share of the cohort's credit at −7				
Percentile 10	0.0002	0.0002	0.0007	0.0018
Percentile 25	0.0004	0.0007	0.0051	0.0087
Median	0.0016	0.0031	0.0160	0.0238
Percentile 75	0.0119	0.0138	0.0533	0.0695
Percentile 90	0.0373	0.0373	0.1301	0.2021
Maximum	0.4697	0.4697	0.4697	0.8500
Mean	0.0222	0.0267	0.0478	0.0762
E. Table 7, for reference				
New lending: Δp_{cj} x Post	0.0154	0.0156	0.0211	0.0193
New lending to future defaulters: Δp_{cj} x Post	-0.0015	-0.0026	-0.0023	0.0002
SD of Δp_{cj} within closure	0.24	0.24	0.26	0.25
Cohort-periods	1,210	1,586	2,660	3,614

Notes: The regression sample of Table 7. Panel A: means of the two outcomes before (quarters −16 to −7) and after (quarters −6 to +8) the announcement, per quarter and as shares of the cohort's credit from the bank in quarter −7, weighted by the regression weight, with standard deviations in brackets. Panel B: new lending to future defaulters as a share of new lending, and the share of cohort-periods with any new lending to future defaulters, unweighted and weighted by the regression weight, and their number. Panels C and D: across the cohort-periods with any new lending to future defaulters, the distribution of its amount, in EUR thousand and as a share of the cohort's credit from the bank in quarter −7. Panel E: the estimates of Table 7, for reference.

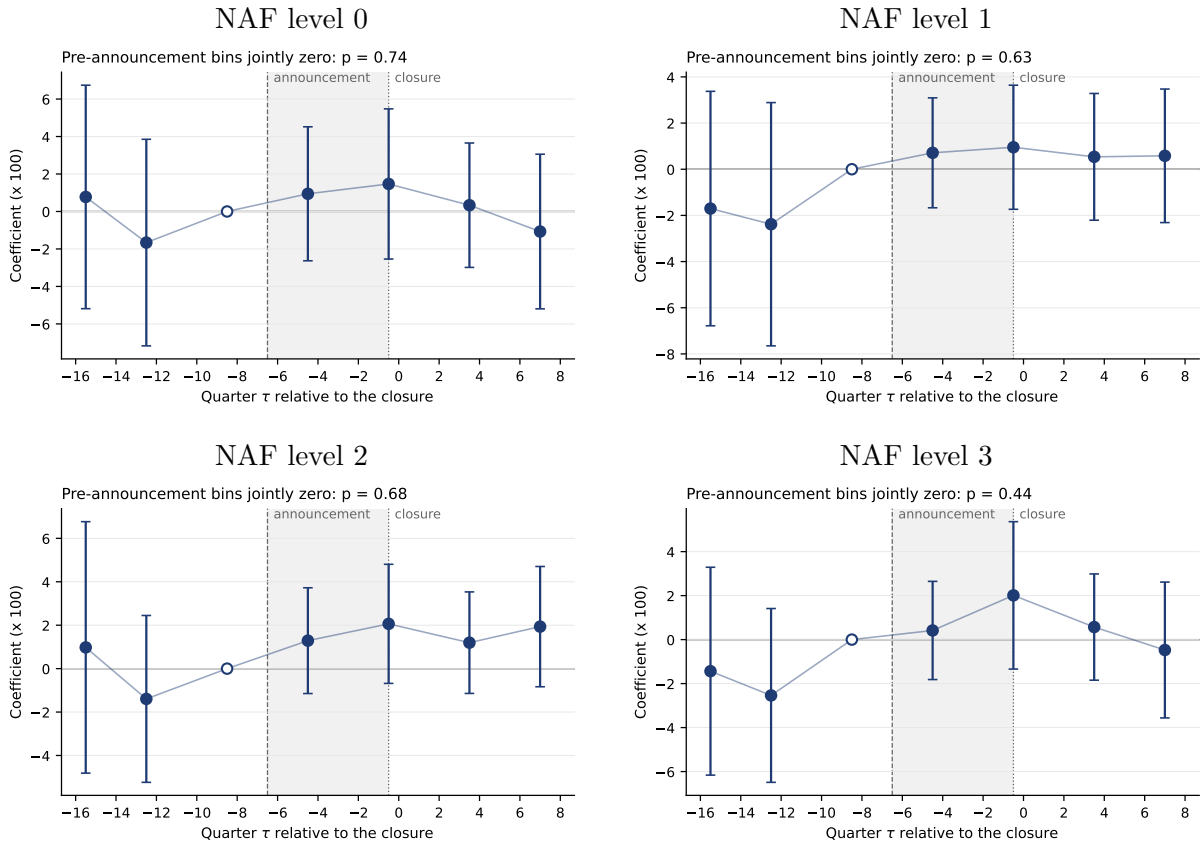


Figure E.4: Change in the specialization of the lending branch and new lending, by year relative to the announcement: industry cohorts

Notes: Coefficients on Δp_{cj} interacted with indicators of bins of quarters relative to the closure, in the regression of Table 7, columns 1–4 (new lending), where they replace Δp_{cj} x Post. The bins are years from the announcement (quarter −6): quarters −16 to −15, −14 to −11, −10 to −7 (reference, drawn hollow at 0), −6 to −3, −2 to +1, +2 to +5 and +6 to +8; markers at the centre of each bin. One observation per cohort and bin. Cohort fixed effects and closing branch x bin fixed effects, so that only the cohorts of the same closing branch are compared; weights as in Table 7. Coefficients multiplied by 100. Panel title: the p-value of the test that the two coefficients before the reference are zero. 3,303, 4,343, 7,466 and 10,207 cohort-bins and 194, 210, 214 and 220 closing branches at NAF levels 0 to 3; 95% confidence intervals, standard errors clustered by closing branch. Grey band: from the announcement (quarter −6) to the closure (quarter 0).

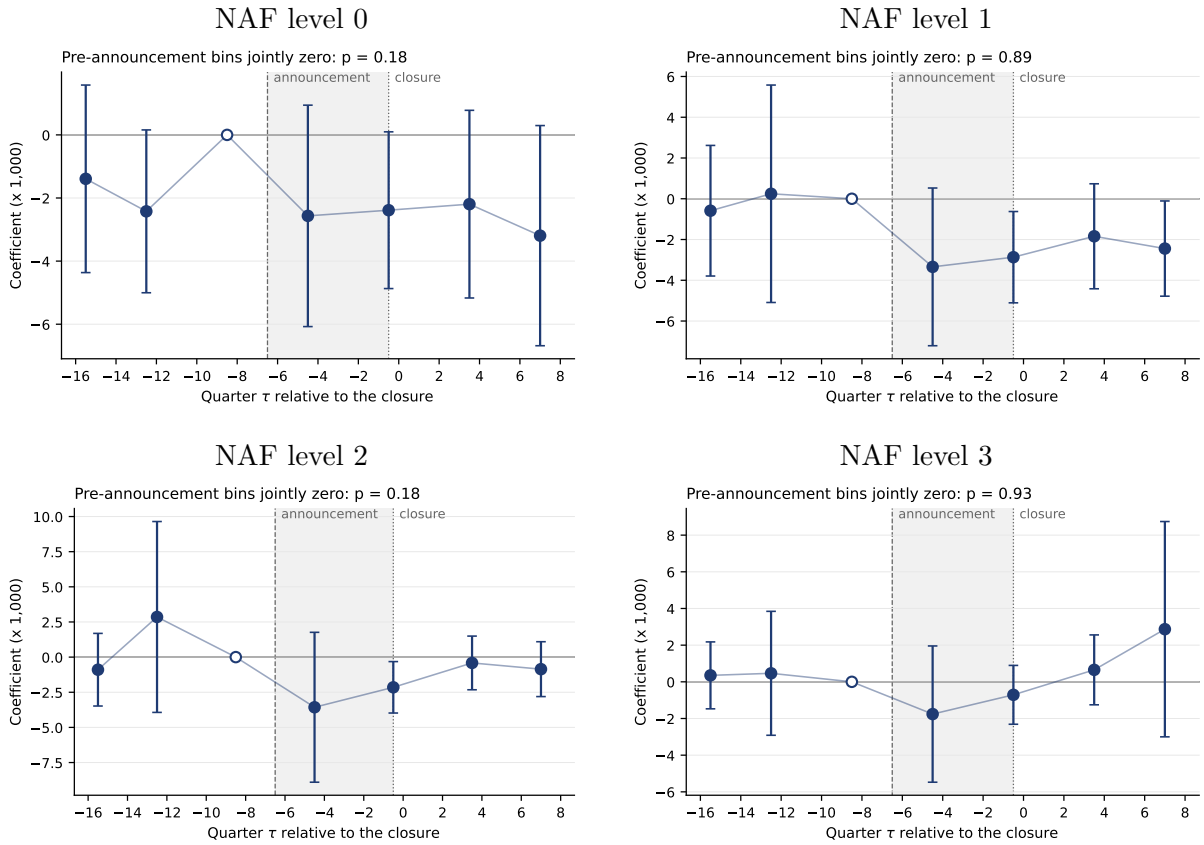


Figure E.5: Change in the specialization of the lending branch and new lending to future defaulters, by year relative to the announcement: industry cohorts

Notes: As Figure E.4, for new lending to future defaulters (Table 7, columns 5–8). Coefficients multiplied by 1,000.

Table E.5: New lending: moves to a more and to a less specialized branch

	New lending from the bank			
NAF level	0	1	2	3
	(1)	(2)	(3)	(4)
Δp^{+}_{cj} x Post (gain)	-0.0033	-0.0140	0.0121	0.1056
	(0.0252)	(0.0239)	(0.0354)	(0.0718)
Δp^{-}_{cj} x Post (loss)	0.0258	0.0301	0.0238	0.0032
	(0.0326)	(0.0186)	(0.0149)	(0.0190)
p-value, gain = loss	0.571	0.246	0.795	0.245
Mean of dependent variable	0.0279	0.0276	0.0281	0.0282
Cell-periods	1,210	1,586	2,660	3,614
Closing branches (clusters)	185	200	203	208
Cohort cell FE	Yes	Yes	Yes	Yes
Closing branch x period FE	Yes	Yes	Yes	Yes

Notes: As Table 7, columns 1–4, with separate coefficients for moves to a more specialized branch, $\Delta p^{+}_{cj} = \max(\Delta p_{cj}, 0)$ ("gain"), and to a less specialized one, $\Delta p^{-}_{cj} = \min(\Delta p_{cj}, 0)$ ("loss"), as in Table 8. The loss term is never positive, so both coefficients measure the change in new lending per unit of Δp_{cj}, and they are equal when the response is symmetric. p-value, gain = loss: the test that the two coefficients are equal (standard errors clustered by closing branch).

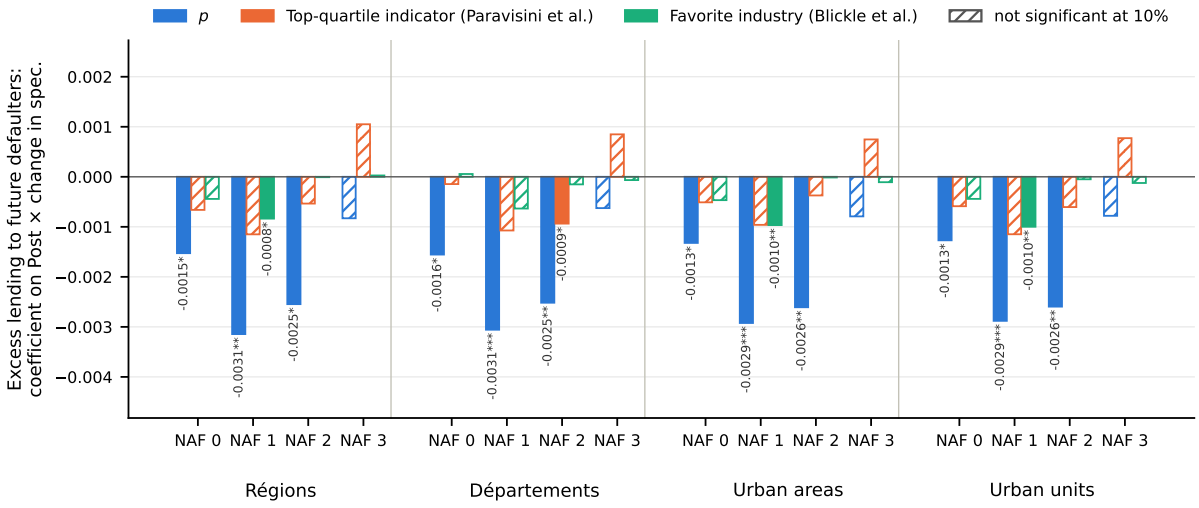


Figure E.6: Change in the specialization of the lending branch and excess lending to future defaulters: industry cohorts, by market definition and measure of specialization

Notes: As Figure 7, for excess lending to future defaulters: new lending to future defaulters minus new lending times the future defaulters' share of the cohort's credit from the bank in quarter −7, as a share of that credit, per quarter, over the firm-quarters of Table 7 in which the borrower's default within a year is observed. It is zero when the new credit is distributed across the cohort's firms in proportion to their credit from the bank in quarter −7, and positive when future defaulters receive more than this proportional share.

Table E.6: New lending to future defaulters, net of the default rate of comparable new loans: industry cohorts

		Net of the default rate of new loans by			
	Unadjusted	Rating x year	Rating x industry x year	Rating x size x year	Rating x industry x size x year
	(1)	(2)	(3)	(4)	(5)
NAF level 0					
Δp_{cj} x Post	-0.0015*	-0.0016*	-0.0014*	-0.0015*	-0.0015*
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Mean of dependent variable	0.0004	0.0002	0.0001	0.0001	0.0001
NAF level 1					
Δp_{cj} x Post	-0.0026**	-0.0027**	-0.0025**	-0.0027**	-0.0026**
	(0.0011)	(0.0011)	(0.0012)	(0.0011)	(0.0011)
Mean of dependent variable	0.0004	0.0002	0.0001	0.0001	0.0001
NAF level 2					
Δp_{cj} x Post	-0.0023*	-0.0024**	-0.0024**	-0.0023*	-0.0024*
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Mean of dependent variable	0.0005	0.0002	0.0001	0.0002	0.0002
NAF level 3					
Δp_{cj} x Post	0.0002	0.0001	0.0000	0.0002	0.0000
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0008)
Mean of dependent variable	0.0005	0.0002	0.0001	0.0002	0.0002
Cohort cell FE	Yes	Yes	Yes	Yes	Yes
Closing branch x period FE	Yes	Yes	Yes	Yes	Yes

Notes: As Table 7, columns 5–8, with one panel per NAF level; column 1 repeats them. In columns 2–5, each firm-quarter's new credit is multiplied by the borrower's default indicator minus a benchmark default rate. The outcome is then new lending to future defaulters in excess of the defaults expected at the benchmark rate. In column 2, the benchmark is the default rate within a year of all new medium- and long-term credit in the sample (firm x bank, weighted by the new credit) to borrowers with the same Banque de France rating class in the previous quarter, in the same calendar year. Unrated borrowers form a class of their own. The benchmarks of columns 3–5 are computed in finer cells, which add the borrower's industry at the panel's NAF level (column 3), its decile of total bank credit in the previous quarter (column 4; deciles by year, borrowers without credit in a class of their own), or its industry and its quintile of total bank credit (column 5). In columns 3–5, the borrower's own new credit is left out of its benchmark, and a cell with fewer than 20 quarters of new credit to other borrowers takes the rate of column 2. Across firm-quarters with new credit, weighted by the new credit, the benchmark explains 2.5% of the variance of the default indicator in column 2 and 2.8% in column 4. In columns 3 and 5 it explains 0.2 to 2.2% and 0.8 to 2.4%, less at finer NAF levels. The sample is the same in every column: 1,210, 1,586, 2,660 and 3,614 cell-periods and 185, 200, 203 and 208 closing branches (clusters) at NAF levels 0 to 3, as in Table 7.

Table E.7: Branch closures, the specialization of the firm's lender and credit from that lender: moves to a more and to a less specialized branch

	(1)	(2)	(3)	(4)	(5)
Closure x Post	-0.092***	-0.052***	-0.035*	0.072	0.067
	(0.014)	(0.016)	(0.020)	(0.067)	(0.066)
Specialization		0.036***	0.035***	0.035***	0.035***
		(0.008)	(0.008)	(0.008)	(0.008)
Closure x Post x Δ specialization, gain			0.017	0.025	
			(0.083)	(0.083)	
Closure x Post x Δ specialization, loss			0.151*	0.140*	
			(0.085)	(0.085)	
Closure x Post x Long relation				-0.107	-0.108
				(0.069)	(0.068)
Closure x Post x Δ distance				0.000	0.000
				(0.001)	(0.001)
Closure x Post x Δ size				-0.086**	-0.087**
				(0.044)	(0.044)
Closure x Post x Large Δ specialization, gain					0.032
					(0.079)
Closure x Post x Large Δ specialization, loss					0.106
					(0.081)
Gain = loss (p-value)			0.331	0.402	0.559
Firm x Bank FE	Yes	Yes	Yes	Yes	Yes
Firm x Quarter FE	Yes	Yes	Yes	Yes	Yes
GEA x Quarter FE	Yes	Yes	Yes	Yes	Yes
Observations	3919413	2707268	2707268	2707268	2707268
R-Square	0.90	0.91	0.91	0.91	0.91

Notes: As Table 9, with separate coefficients for $\Delta p_{fk} > 0$ ("gain": the new branch is more specialized in the firm's industry) and $\Delta p_{fk} < 0$ ("loss"); the second term is Δp_{fk} itself, negative, so both coefficients are positive when credit moves with specialization. "Gain = loss": p-value of the test that the two coefficients are equal.

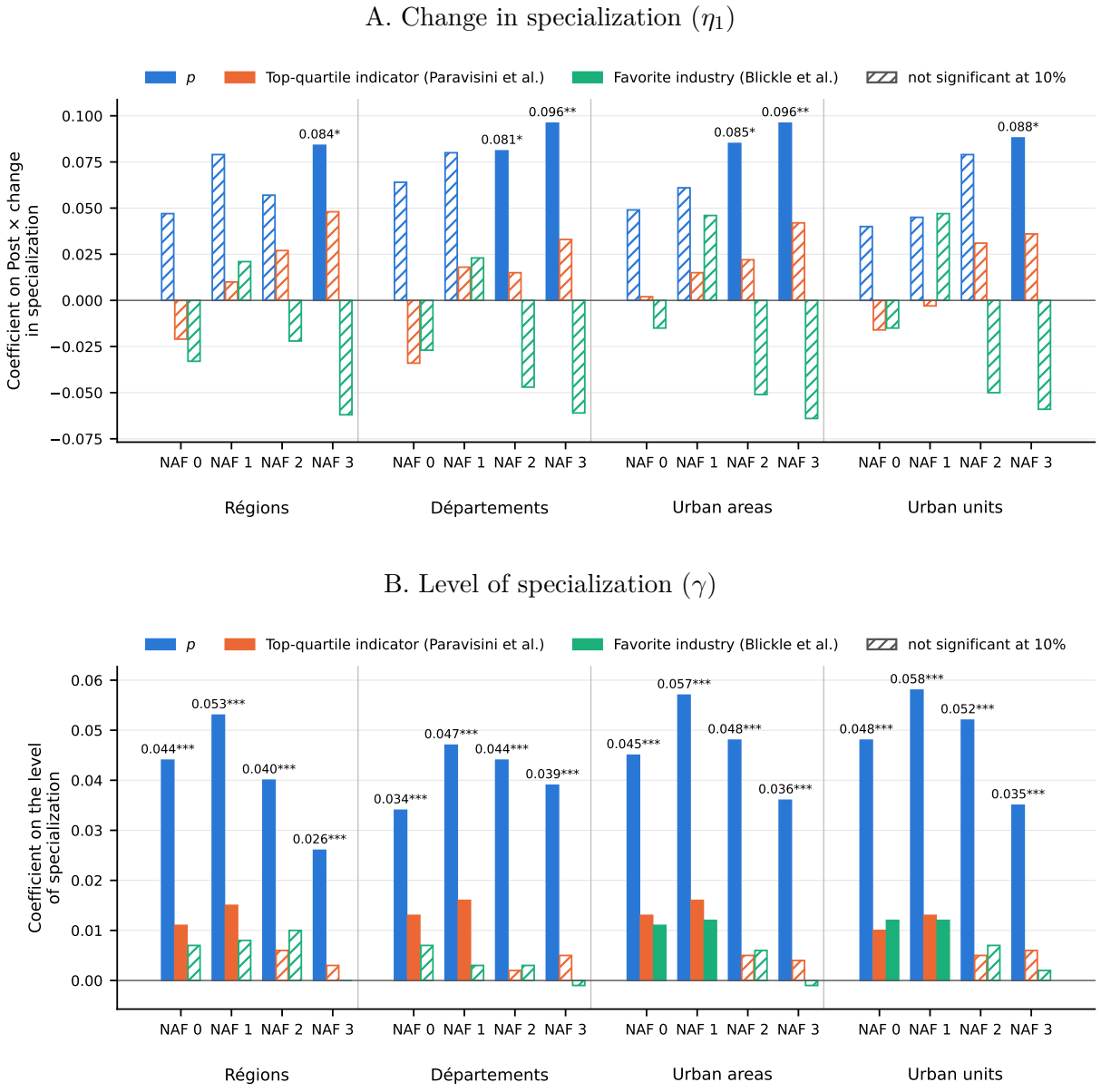


Figure E.7: Specialization of the firm's lender and credit from that lender, by market definition and measure of specialization

Notes: Coefficients of the firm–bank regression (11) (specification of Table 9, column 4) for 16 market definitions (geography of the branch × NAF level) and three measures of specialization: p, the top-quartile indicator of Paravisini et al. (2023) and the favorite-industry indicator of Blickle et al. (2026). Panel A: η_1, on the change in specialization, defined as the absorbing branch's minus the closing branch's. Panel B: γ, on the level of specialization of the firm's current lender. With p, Δp_{fk} is set as in Table 9 (absorbing branch's p in the quarter before the announcement: the p of its share of lending in the firm's industry, with no relationship left out; the exact p of an empty cell where it lends in the market but not to the firm's industry, 1/2 where it does not exist or does not lend; missing changes set to zero); with the two indicators, the absorbing branch's indicator is zero where it does not lend to the firm's industry or does not exist, and missing changes are set to zero. Solid bars: significant at least at the 10% level; hatched bars: not significant. Standard errors clustered by the firm's urban unit and year, as in Table 9.