Network Evidence on Credit-Risk Pricing in P2P Lending

Lennart John Baals[a,b,c,*], Jörg Osterrieder[a,c], Branka Hadji-Misheva[c], Yizhi Wang[d]

[a] *Department of Industrial Engineering and Business Information Systems, University of Twente, Building 10: Ravelijn, Drienerlolaan 5, Enschede, The Netherlands*

[b] *Department of Industrial Engineering and Operations Research, Columbia University, 500 W. 120th Street, New York, NY 10027, USA*

[c] *Bern Business School, Bern University of Applied Science, Brückenstrasse 73, 3014 Bern, Switzerland*

[d] *Cardiff Business School, Cardiff University, United Kingdom*

Abstract

We study whether posted interest rates on a leading European P2P platform, Bondora, efficiently reflect borrower default risk. Using 22,000 primary-market loans from 2012–2022, we (i) extract key risk drivers via a Random Forest, (ii) build a weighted Gower similarity network, and (iii) partition loans into risk-homogeneous communities via the Louvain algorithm. Within each community we benchmark quoted rates against realised defaults ex-post and subsequently model ex-ante implied rates through gradient-boosting models (GBMs) that replicate the platform's pricing function. Our analysis uncovers substantial mispricing: 8.8% of loans exhibit rate deviations exceeding 100 basis points, creating annual wealth transfers of €102,000 between investors and borrowers in our sample alone. Mispricing peaks during the platform's 2015-2018 transition to risk-based pricing and diminishes thereafter, consistent with a platform learning effect. We find underpriced loans to concentrate among B-rated borrowers, while C-rated borrowers face systematic overpricing despite similar risk profiles. Our network-analytical framework offers investors a tractable way to identify latent pricing inefficiencies and highlights how information frictions persist even in data-rich fintech markets.

Keywords: Peer-to-peer lending, Loan pricing, Network analysis, Machine learning, Credit risk

JEL: G14, G21, G23, C45

[*] Corresponding author: Lennart John Baals - l.j.baals@utwente.nl

1. Introduction

Quantifying credit risk remains one of the major challenges in modern finance (Kealhofer, 2003), given the inherent uncertainty around loan default events and the financial consequences associated with this risk for creditors. Across traditional and digital consumer lending markets, lenders are commonly found to face several frictions that can impede accurate risk-based pricing. Factors such as fraud risk (e.g. identity theft), borrower private information unobservable to the lender, and competitive pressure from alternative funding sources, are generally found to distort the alignment between expected default rates and loan interest (Einav et al., 2012; Iyer et al., 2016; Argyle et al., 2023). These challenges can be especially pronounced in online peer-to-peer (P2P) marketplace lending, which connects borrowers directly with investors for both secured and unsecured credit, often involving higher risk-taking and opaque information transparency for lenders (Dömötör et al., 2023). Whereas other credit markets often rely on established infrastructures or intermediary processes, lenders in centralized P2P lending must solely rely on the platform to manage these complexities directly via sufficient loan screening and credit scoring techniques. Despite the initial expectation of efficient pricing mechanisms in P2P marketplaces, due to rapid adoptions of technological innovations and advanced digital informativeness of lenders, real-world observations can reveal deviations from fair-value pricing.[1] These discrepancies, collectively referred to as mispricing, not only undermine investor confidence but also pose challenges to the efficient allocation of credit.

Over time, many marketplace lenders have shifted from an decentralized auction-based market mechanism to a centralized risk-based pricing system, setting loan prices according to credit scores and the loan size (Phillips, 2018; Wei and Lin, 2016; Balyuk and Davydenko, 2024).[2] This change in the operating practice, away from a price discovery process led by market forces towards a predetermined pricing of loans, highlights an important consequence, namely the lending platform's shift from a passive intermediary to an active role in screening and scoring loans. Assessing borrower risk can be inherently complex, with scaling loan pool diversity and platform-level growth, inducing scoring complexities, potentially leading to mispricing. Such mispriced loans can manifest in two primary forms: underpricing, where interest rates are charged below the borrower's fair-risk level, and overpricing, where borrowers are charged excessively above the equivalent fair risk level. In both cases adverse selection, higher default rates, or inefficiencies in credit allocation may appear as direct outcome.

The existing literature around price efficiency in P2P lending is still evolving. Specific works have examined the centralized role in application screening and loan pricing by P2P platforms, employing an active price quotation process, which was found to lead to limited price discovery for investors (Vallée and Zeng, 2019; Balyuk, 2023; Balyuk and Davydenko, 2024; Liu et al., 2020)

[1] See Jagtiani and Lemieux (2019) and Berg et al. (2020) for further details.

[2] See Balyuk and Davydenko (2024) for a temporal overview of this trend.

and pricing inefficiency (Wei and Lin, 2016; Caglayan et al., 2020). However, these studies do not numerically control for the heterogeneity of the loan pool or focus on the pricing mechanisms and accuracy underlying the primary issuance of loans, which we set out to address. Specifically, we shed light on the price formation of primary loans and analyze the degree of alignment between default risk and loan price immanent in the credit market of Bondora, a leading European P2P lending platform.[3]

Accordingly, we analyze a randomly sampled data set comprising 22,000 loans issued on the Bondora platform during the period of 2012 to 2022. Utilizing network-based analysis, we construct a risk similarity network based on key credit features to capture loans with similar risk profiles by matching them with their most similar counterparts and apply the Louvain algorithm (Blondel et al., 2008) to segment these loans into distinct clusters. The resulting communities form the basis of our mispricing tests: we first benchmark platform-assigned interest rates against realized default outcomes and then compare those rates with model-implied, ex-ante rates generated by gradient boosting models (GBMs). Our approach not only quantifies the extent of mispricing within each detected community but also reveals borrower segmentation patterns associated with underpricing and overpricing.

This study contributes to the growing literature on credit pricing in P2P lending in three significant ways. i) we introduce a transparent framework to evaluate pricing opacity in the primary P2P loan market by integrating network-based borrower segmentation with investor-replicable modeling approaches solely based on accessible data. We further demonstrate that systematic deviations exist across borrower segments, highlighting the potential for pricing inefficiencies and misaligned risk assessment, particularly in the presence of undisclosed third-party data. By retaining those pairwise distances, instead of collapsing them into a single propensity score or resorting to rule-based matching (e.g. Pursiainen, 2024; Di Maggio and Yao, 2021), the method captures higher-order interactions among the key credit features and yields borrower segments that are strictly comparable to every other loan in the dataset. In this sense it extends the threshold-based network idea of Baumöhl and Lyocsa (2025) to a fully data-driven pricing analysis that is replicable by market participants. ii) the use of Louvain modularity optimization ensures that the resulting borrower communities are globally coherent clusters whose internal similarity is maximized relative to the entire loan portfolio. This allows us to study pricing behaviour at the level at which the platform appears to make economically meaningful distinctions among borrowers. iii) This study provides first systematic evidence on the prevalence and magnitude of loan mispricing, thereby quantifying detected pricing inefficiencies into cash-flow transfers at the portfolio-level. By shedding light on

[3] P2P lending on Bondora has been studied by range of scholars including Dömötör et al. (2023); Caglayan et al. (2020); Liu et al. (2024a,b). Further information on Bondora's lending practices can be found in: https://help.bondora.com/hc/en-us/sections/14814527192209-General-information.

hidden inefficiencies in posted credit pricing, our work supports more informed investing and fairer lending, which is critical to the sustainable growth of fintech credit markets.

The remainder of this study is organized as follows. Section 2 reviews the related literature on mispricing in traditional consumer- and online P2P lending markets. It further introduces the development of the credit scoring system on Bondora and outlines the underlying hypotheses. Subsequently, Section 3 describes the data used in this study and Section 4 presents the methodology alongside the discussion of the empirical findings. Finally, Section 5 concludes with recommendations for future research and practice.

2. Literature Review and Hypothesis Development

2.1. Pricing Inefficiencies in Consumer Lending

The literature on consumer credit links persistent deviations of loan rates from risk-adjusted benchmarks to two major frictions: limited borrower risk assessment and asymmetric information. Building on Stiglitz and Weiss (1981), who first show that adverse selection and moral hazard can prompt lenders to ration credit rather than raise rates, empirical work finds that risk-based pricing has generally favored consumers in accessing credit but at the expense of higher risk premia. Edelberg (2006) find for U.S. consumer loans that credit scoring generally enabled high-risk borrowers access to credit but in return widened spreads between high- and low risk borrower segments. Other studies on subprime auto loans show that lenders can significantly improve their expected profit through risk-based loan pricing but consumers still hold adverse private information that remains unpriced (Einav et al., 2012, 2013). Keys et al. (2010) argue that in the context of loan securitization, information loss is immanent from reduced screening efforts of lenders that resulted in price deficiencies of assigned loan interest rates to corresponding default rates. Subsequent studies by Stango and Zinman (2016), Agarwal et al. (2018), Bhutta et al. (2020), and Argyle et al. (2020) further highlight limitations in risk-based credit pricing due to informational- and search frictions that limit both lenders and borrowers to arrive at equilibrium interest rate levels that would accurately reflect the inherent credit risk of borrowers. Stango and Zinman (2016) further find that borrowers face significant equilibrium price dispersion, resulting from a considerable cross-issuer heterogeneity in risk-based pricing models that distort uniform pricing of consumer credit across different loan issuers. In turn, Karlan and Zinman (2009) reveal evidence that hidden information and moral hazard are found to drive up to 20% of defaults in their sample of South African consumer loans, indicating that information asymmetries may explain the prevalence of credit constraints even in markets that specialize in financing high-risk borrowers at very high rates. Given this context, online P2P lending may be specifically prone to resulting deficiencies in loan price formation, stemming from such frictions, as platforms are commonly found to issue credit to riskier borrowers, thus acting as substitutes for traditional bank lenders.

2.2. Loan Price Dispersion in Online P2P Consumer Lending

Several studies have documented misalignment between ex-ante rates and ex-post defaults in marketplace lending due to heterogeneous borrower groups (Wei and Lin, 2016; Mild et al., 2015; Dömötör et al., 2023; Franks et al., 2021), systematic biases in credit-screening models (Iyer et al., 2016; Vallée and Zeng, 2019; Fuster et al., 2019), valuation errors in secondary-market trades (Caglayan et al., 2020), and price distortions linked to social-network endorsements or unverifiable borrower narratives (Freedman and Jin, 2017; Lin et al., 2013; Michels, 2012).

Two inter-related mechanisms emerge as principal drivers of these pricing anomalies and closely match the frictions of constrained borrower risk assessment and asymmetric information evidenced in traditional consumer credit. First, similar to findings in bank-based lending, where mismatches between loan prices and borrower risk profiles are well documented (Bubb and Kaufman, 2014; Keys et al., 2010; Agarwal et al., 2018; Adams et al., 2009; Argyle et al., 2023), recent P2P-lending studies show that verification constraints and screening biases significantly drive mispricing by limiting a platform's ability to differentiate risk accurately. Iyer et al. (2016) find that open-access rating systems employed by some P2P lending platforms can induce investors to over-rely on a single numeric score and to ignore "soft" information, thereby disadvantaging certain borrower types. Similarly, Michels (2012) shows that unverifiable personal narratives can systematically reduce interest rates, while Freedman and Jin (2017) and Lin et al. (2013) find that borrower "friend" endorsements also secure cheaper credit, although many such social signals fail to predict lower default. Conversely, Mild et al. (2015) document that limited borrower verification led investors to underestimate default likelihoods on the European platform 'myC4'. Nevertheless, the evidence on screening accuracy is mixed: Tang (2019) find that P2P platforms substitute for banks with comparable default performance, whereas Di Maggio and Yao (2021) report significantly higher default rates for fintech loans, implying weaker screening accuracy. However, over time P2P lending platforms are incentivized to increase their prescreening intensity to actively manage adverse selection and trade-off prescreening costs against loosing unsophisticated investors on the market place (Vallée and Zeng, 2019). Congruently, several scholars outline a learning effect of platforms to conduct better loan screening over time as the pool of financed loans increases and more data becomes available, thereby reducing screening- and informational costs (Lin and Viswanathan, 2016; Freedman and Jin, 2011; Jagtiani and Lemieux, 2019; Vallée and Zeng, 2019).

A second driver is the significant difference in the degree of risk ownership and information disclosure between P2P lending platforms and investors documented in the literature (Vallée and Zeng, 2019; Balyuk and Davydenko, 2024). Rooted in the seminal theory of financial intermediation (Diamond, 1984), the rational concern over an 'originate-to-distribute' model lies in the appropriate incentives for a lender or intermediary to screen and monitor loans properly (Holmstrom and Tirole, 1997; Keys et al., 2010). Similar to originate-to-distribute lenders, P2P lending platforms operate on an intermediary basis, selling loans to investors without bearing the default risk (Balyuk and

Davydenko, 2024) and controlling the flow of information from borrowers to investors. Platforms are frequently found to utilize additional proprietary data sources (e.g. from credit-bureaus) and behavioral metrics (e.g. digital user data) when screening and scoring credit applicants, which are not disclosed to investors (Berg et al., 2020; Fuster et al., 2019). Therefore, investors often find themselves in a position to merely accept platform grades based on trust (Thakor and Merton, 2024). Several studies further outline resulting moral-hazard incentives. Balyuk and Davydenko (2024) point out that centralized platforms, faced with volume-linked origination fees, may reduce screening stringency or information disclosure to boost loan volumes for increased profit and cost reduction, thereby stressing the accuracy of loan pricing. Vallée and Zeng (2019) show that decentralized platforms balance prescreening with strategic reductions in information disclosure, which may further inhibit investors from conducting refined pricing assessments as their screening costs increase with coarser information distribution. If platforms simultaneously decide to disclose only partial information, both sophisticated and unsophisticated investors are dependent on platform-assigned ratings as risk verification, which could further enforce informational asymmetries (Dömötör et al., 2023; Freedman and Jin, 2011). Empirical evidence supports the notion that moral hazard within P2P lending structures can lead to distortions in loan pricing. Hildebrand et al. (2017) document that P2P group leaders, incentivized by origination fees, encouraged funding of riskier loans, ultimately leading to higher default rates.

2.3. Development of Bondora Rating

A key milestone in Bondora's lending infrastructure is the introduction of Bondora Rating, a proprietary credit scoring system designed to enable risk-based pricing. From its inception until early 2015, Bondora did not apply a systematic risk-based mechanism to price loans, despite growth in loan volume. Instead, loans were screened into broad categories based on a borrower's payment history (indexed as 600–1000) or discretionary income brackets (A, B, C).[4] This approach effectively caused loans, belonging to the same group, to be assigned the same nominal interest rates, irrespective of individual risk distinctions among borrowers. Consequently, two prospective loans that appeared nearly identical in country of residence, loan size, or purpose could be priced differently if a borrower happened to fall into a higher-level bracket. Conversely, genuinely higher-risk borrowers could potentially obtain loans at disproportionately favorable rates.[5] To mitigate uncertainty, the platform retroactively assigned ratings to older loans from this era, informing investors that such ratings approximate the ex-ante risk the loans would have borne. However, from an *ex-post* perspective it is not certain that these classifications fully capture borrowers' actual risk levels,

[4] https://bondora.com/en/blog/explaining-bondora-rating/

[5] As acknowledged by Bondora, the initial categories (e.g., "600–1000" or "A, B, C") did not necessarily reflect underlying risk metrics. Instead, they rested on high-level borrower attributes, including discretionary income and a limited view of past payment problems (https://bondora.com/en/blog/explaining-bondora-rating/).

given that no robust modeling of any risk metrics (probability of default (PD), loss given default (LGD), exposure at default (EAD)) was performed at origination. In early 2015, Bondora went on to introduce Bondora Rating to each newly originated loan, taking into account a variety of proprietary data sources and statistical models.[6] Under the new system, each loan application is assigned a rating from AA (the safest) to HR (high risk), based on an internally computed expected loss (EL) metric that incorporates borrower-level attributes such as income stability, externally validated credit bureau data, and behavioral patterns identified on Bondora's website. By pricing loans in accordance with a forward-looking assessment of PD, LGD, and EAD, Bondora now aimed to charge higher rates for loans with greater expected default risk and lower rates for safer borrowers. Although early data suggest that risk-based pricing improved the alignment between interest rates and default outcomes, Bondora's own communications noted that the rating model underwent regular reviews from 2015 onward, acknowledging that predictive power could drift as borrower demographics and macroeconomic conditions evolved.[7] Over the 2016–2017 period, Bondora iterated its rating model by integrating more extensive third-party data, refining the weighting of behavioral signals, and adjusting interest rate spreads assigned to each rating category.[8]

2.4. Hypotheses

In the absence of active investor participation in price discovery, P2P platforms assume a central role in screening borrowers and assigning loan prices. Theoretically, this centralized control can introduce pricing inefficiencies when platforms face weak incentives for accurate screening or rely on opaque, proprietary scoring systems that are not fully transparent to investors. Drawing on foundational work in credit markets, such as (Stiglitz and Weiss, 1981), mispricing can emerge when asymmetric information distorts the alignment of loan prices to borrower risk. In marketplace lending, where platforms often operate under an originate-to-distribute model without bearing default risk, empirical studies have identified systematic mismatches between quoted interest rates and realized loan outcomes (Wei and Lin, 2016; Franks et al., 2021; Mild et al., 2015). These inefficiencies are further exacerbated when borrower evaluations are based on unverifiable personal narratives or coarse-grained credit scores that fail to capture nuanced risk indicators (Michels, 2012; Iyer et al., 2016). Against this backdrop, we test whether similar inefficiencies are present in the primary loan market of Bondora, where centralized pricing decisions may lead to persistent valuation errors.

H1. Mispricing exists among P2P issued loans on Bondora's primary market, characterized by over- or undervaluation with non-zero frequencies.

[6] https://bondora.com/en/blog/introducing-risk-based-pricing/

[7] https://help.bondora.com/hc/en-us/articles/14814705732881-How-are-Bondora-risk-ratings-calculated

[8] Bondora emphasizes its credit rating process to incorporate external information, including public records and behavioral data from platform interactions (https://bondora.com/en/blog/explaining-bondora-rating/).

Following Bondora's introduction of its proprietary risk-based pricing model in 2015, the platform entered a critical transition period in which the accuracy and calibration of the new screening system were still evolving. Drawing on concepts from statistical learning theory (Vapnik, 1999a,b), the early stages of model implementation are often characterized by cost–complexity trade-offs: models with greater predictive expressiveness require more training data and higher calibration costs to achieve generalizable accuracy. During such periods, platforms may prioritize rapid loan origination over costly model refinement, particularly when facing limited historical data or economic pressures to scale. This introduces a risk of pricing inefficiency, where loan rates deviate from their true underlying credit risk due to temporary miscalibration or incomplete data integration. Empirical studies on platform evolution in P2P lending similarly document a "learning effect," in which platforms gradually improve screening intensity and risk segmentation over time (Vallée and Zeng, 2019; Lin and Viswanathan, 2016; Freedman and Jin, 2011; Jagtiani and Lemieux, 2019). Based on this theoretical and empirical context, we expect the alignment between loan prices and borrower default risk to be weakest in the initial years of the risk-based model rollout.

> **H2.** Loans issued during the transition period (2016–2018) exhibit greater misalignment between default risk and assigned interest rates, reflecting early-stage inefficiencies in Bondora's risk-based pricing model.

As Bondora expanded over time, its risk assessment capabilities likely improved due to iterative refinement of its proprietary rating model, accumulation of borrower-level data, and experience gained in credit scoring and pricing. This dynamic aligns with the concept of a platform learning effect, wherein fintech lenders gradually enhance the accuracy of loan pricing through feedback loops, increased training data, and cost efficiencies in data processing and model calibration. Prior research suggests that as lending platforms scale, they tend to reduce both information acquisition costs and mispricing frequencies by improving model specification and parameter stability over time (Jagtiani and Lemieux, 2019; Vallée and Zeng, 2019; Freedman and Jin, 2011). These developments are expected to translate into more accurate alignment between borrower default risk and the assigned loan interest rates in later periods. On this basis, we formulate the following hypothesis:

> **H3.** The frequency and magnitude of mispricing diminish over time, reflecting enhanced accuracy in risk assessment and pricing calibration.

3. Data

3.1. Data Description and Preprocessing

We utilize a sub-dataset of the entire loan book obtained from Bondora with access to loan-level information, borrower characteristics, and loan performance indicators, comprising a total of 49,207 observations before our initial filtering.[9] The loans in the dataset were issued in Bondora's primary market in Estonia between 2012 – 2022, which gave us a broad temporal scope for the analysis.[10] Our target variable is the loan interest rate, which serves as a reduced-form output of the platform's internal pricing model, reflecting all available information at the time of loan issuance. Additionally, we consider the default status of loans as ex-post measure to assess credit risk outcomes.

Prior to the analysis, we performed extensive data cleaning and preprocessing to ensure the integrity and suitability of the dataset for our modeling process. This involved handling missing values, variable selection and transformation, and class balancing, leading to a final dataset composition. Variable selection involved removing irrelevant or redundant information. We excluded language indicators and date variables unrelated to the analysis, as well as variables with high multicollinearity, identified by a correlation coefficient exceeding 0.95 with other predictors. Ex-post variables, which would not be available at the time of loan issuance, were also removed to ensure that our network modeling approach relies solely on ex-ante information. We transformed certain variables to enhance interpretability and modeling efficacy. Binary variables were converted into categorical factors to appropriately capture their discrete nature. Temporal features, such as the hour of loan application and the day of the week, were extracted to capture potential time-related patterns in loan issuance and pricing.[11] Given the imbalanced nature of the default status in the dataset (11,312 defaults versus 17,811 non-defaults), we performed stratified sampling to create a final sample tailored to robust modeling. A total of 22,000 loans were randomly selected,thereby maintaining the original proportion of defaulted and non-defaulted loans to preserve the inherent class distribution.[12]

4. Empirical Analysis and Discussion

4.1. An Initial Analysis through Network Construction and Community Detection

4.1.1. Feature Selection

In our first step towards constructing a similarity network of loans based on risk profiles, we identify the most significant predictors of loan default through a Random Forest (RF) algorithm (Breiman,

[9] Bondora is a European P2P lending platform operating since 2009 in Estonia, Finland, Spain, and Slovakia. The platform has funding from approximately 1,345,873 individual lenders and has disbursed about €1.402 billion in loans. Details on the variables comprising the loan-level information, borrower characteristics, and loan performance indicators in the dataset are provided in Table A.6.

[10] We include only loans from Estonia with credit ratings A,B,C for the analysis as we find loans originating from this country within Bondoras operational geographic area to be the least risky, which is congruent with findings from other studies (Caglayan et al., 2020). Through this specification we apply a conservative scope to the detection of potential mispricing that will likely convey to be evident in more riskier segments of the market.

[11] We excluded any loans in the initial sample who's scheduled loan duration was heavily exceeded due to late payment by beyond 10 months.

[12] This amount of selected loans represented the maximum processing amount for the computing machine (MacBook Pro, 2021, M1Max, 32GB RAM configuration) to conduct the network computation and graph construction.

2001). The RF is capable at handling datasets with a large number of variables and can effectively capture complex, non-linear relationships among variables (ibid., Genuer et al. (2010)). By assessing the importance of each variable in predicting the target outcome, the model facilitates the selection of the most influential features for our subsequent analysis.

We defined the target variable as the default status of the loan.[13] This enables us to search for factors that significantly influence the default likelihood and thus provide us with a direct linkage to the credit risk associated with these loans. Subsequently, the target variable and any other added credit variables initially computed by Bondora such as credit ratings (A, B, C) and the corresponding interest rates were excluded.[14] The dataset was then prepared for the modeling step. Categorical variables were converted into factors to ensure proper handling by the RF algorithm.[15] Through this step, we identified 20 most influential risk factors contributing significantly to the classification of defaulted versus non-defaulted loans. Any subsequent features were found to marginally influence the associated credit risk of loans contained in the sample. Hence, the selected features can be deemed instrumental in capturing the multifaceted nature of credit risk in P2P lending. Depicted in Table A.6 these features encompass borrower demographics (such as age and education level), financial indicators (including debt-to-income ratio and liabilities), loan characteristics (like loan amount and monthly payment), and temporal variables (such as application time and verification level). The identification of these top features provides a focused set of variables for constructing the risk factor similarity network in the subsequent analysis. By selecting variables that significantly influence loan default and price sensitivity of lenders such as the credit terms (loan duration, payment period, loan principle, collateral), as well as personal information, we ensure that the network captures the most relevant aspects of borrower- and credit risk. Table 1 presents the summary statistics of the binary- and continuous variables among the 20 most influential risk factors.

[13] Represented by a binary indicator based on the *DefaultDate* attribute, assigning a value of 1 to loans with a non-null *DefaultDate* and a value of 0 to those without. The default flag follows the platforms definition of classifying a loan as defaulted if interest payments are overdue beyond 60 days (`https://www.bondora.com/blog/p2p-finance-association-defaults/`), which exceeds the loan default definition of the P2P Finance Association (120 days) as industry standard (`https://p2pfa.info`).

[14] This is done in order to assure that the network is not biased by variables that could potentially fall to mispricing by the platform.

[15] The dataset was split into training and testing sets using a 70:30 ratio, maintaining the class distribution of the default status to prevent any bias. The RF was trained on the training set with 500 trees ($ntree = 500$) to ensure robustness. Variable importance was measured using the Mean Decrease in Accuracy metric, which quantifies the loss of predictive power when a variable is excluded from the model.

Table 1: Descriptive Statistics for the Top 20 Risk Factors

(a) Binary Variables

Feature	**Proportion (%)**	**Count**
educ.5 (Higher Education)	24.0	22,000
duration.12 (12-Month Duration)	5.84	22,000
duration.60 (60-Month Duration)	41.2	22,000
joint.ownership (Joint Ownership)	11.0	22,000
duration.24 (24-Month Duration)	4.89	22,000
duration.09 (9-Month Duration)	2.60	22,000
educ.4 (Secondary Education)	49.0	22,000
ver.4 (Verification Level 4)	55.0	22,000
duration.36 (36-Month Duration)	36.5	22,000
no.liab.01 ($\geq$ 1 Liability)	15.5	22,000

(b) Continuous Variables

Feature	**Mean (SD)**	**Min – Max (Count = 22,000)**
time (Application Time)	3.59 (0.65)	2.04 – 4.80
inc.total (Total Income in Logs)	6.98 (0.53)	2.30 – 13.1
AmountOfPreviousLoansBeforeLoan (€)	5,089 (6,062)	0 – 74,740
MonthlyPayment (€)	4.09 (0.90)	1.31 – 7.55
liab.l (Liabilities in Logs)	5.02 (2.13)	0 – 10.5
loan_amount (€)	2,221 (2,219)	106 – 10,632
Age (Years)	39.2 (11.5)	18 – 70
NoOfPreviousLoansBeforeLoan	2.59 (3.00)	0 – 26
FreeCash.l (Log-transformed Free Cash)	0.996 (2.15)	0 – 11.3
DebtToIncome	6.28 (15.0)	0 – 70.0

Note: This table provides descriptive statistics for the 10 binary (panel a) and 10 continuous (panel b) variables identified among the top 20 risk factors selected via a Random Forest model. For binary features, "Proportion (%)" is the share of loans exhibiting that attribute, and "Count" is the total sample size. For continuous features, we report mean, standard deviation (in parentheses), and the observed range of values. All data are based on 22,000 sampled loans. For the meaning of each feature name or abbreviation, refer to Table B.10 and Table B.11.

From the descriptive statistics we can observe a certain degree of variability and range within the factors, indicating different borrower profiles and loan characteristics. For instance, the average loan amount is €2,221 with a standard deviation of €2,219, pointing towards some significant variation in loan sizes among borrowers.

4.1.2. Risk Factor Similarity Network Construction

To analyze the relationships among loans based on borrower and loan characteristics, we constructed a risk factor similarity network as in Baumöhl and Lyocsa (2025), using the top risk factors identified through the feature selection process in Subsection 4.1.1. These factors, each weighted according to their variable importance in the RF classifier (see Breiman 2001), collectively capture the most relevant dimensions of credit risk in our dataset. Specifically, the RF's mean-decrease-in-accuracy scores were normalized to derive a weight vector $\boldsymbol{w} = (w_1, \ldots, w_p)$ for $p = 20$ features, ensuring the most predictive features to receive greater emphasis in the subsequent distance calculations. Given the mix of numerical and categorical variables among these risk factors, we employed a weighted Gower distance (Gower, 1971; Kaufman and Rousseeuw, 2009) to compute pairwise distances between loans. The weighted version of Gower's measure incorporates the importance scores $\boldsymbol{w}$ to modulate each dimension's influence on the overall distance (Ahmad and Dey, 2007). Formally, for loans i and j, the weighted Gower distance d_{ij} is:

$$d_{ij} \;=\; \frac{\sum_{k=1}^{p} w_k \, d_{ijk}}{\sum_{k=1}^{p} w_k}, \tag{1}$$

where d_{ijk} measures the distance between i and j on variable k. For numerical variables,

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{\text{range}_k}, \tag{2}$$

while for binary (categorical) variables,

$$d_{ijk} = \begin{cases} 0, & \text{if } x_{ik} = x_{jk}, \\ 1, & \text{if } x_{ik} \neq x_{jk}. \end{cases} \tag{3}$$

Here, range_k denotes the observed range of variable k, and w_k is the corresponding normalized importance weight. The resulting pairwise distance matrix $\mathbf{D} = [\, d_{ij} \,]$ encapsulates the multi-dimensional similarity structure between loans and reflects each factor's relative contribution to default risk. To focus on the most similar loans and reduce noise from weaker connections, we transformed $\mathbf{D}$ into an adjacency matrix $\mathbf{A}$ by imposing a threshold $\alpha = 0.1$. Specifically, we retained only those edges where the distance between loans i and j satisfies:

$$d_{ij} \leq \alpha.$$

This threshold ensures that only pairs of loans with high similarity (low distance) are connected in the network. We constructed the adjacency matrix $\mathbf{A} = [a_{ij}]$ as:

$$a_{ij} = \begin{cases} d_{ij}, & \text{if } d_{ij} \leq \alpha \text{ and } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

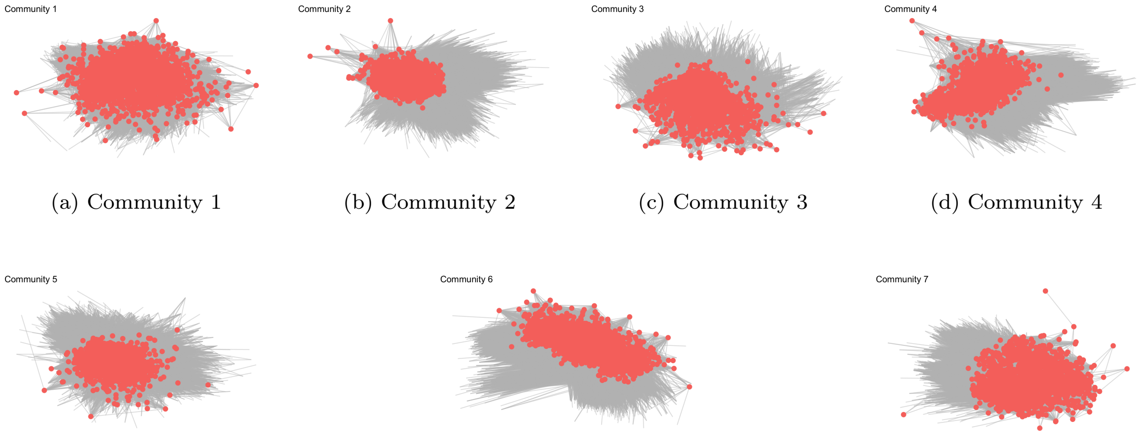
Hence, only pairs of loans whose weighted Gower distance satisfies $d_{ij} \leq \alpha$ are linked in the network, with edge weight $a_{ij} = d_{ij}$. We then remove any isolated nodes, ensuring each retained loan has at least one neighbor of sufficiently close distance. Any pairs of loans that do not share meaningful risk similarities were discarded. Figure A.8 illustrates the distribution of pairwise distances before thresholding: most values lie below 0.3, which motivated our choice of $\alpha = 0.1$ to preserve only the closest loan pairs. Conceptually, loans grouped in this network segment should share key ex-ante risk traits and hence under an ideal pricing mechanism be assigned comparable interest rates.

4.1.3. Community Detection via Louvain Algorithm

Having established the weighted Gower network of loans, we next apply the Louvain algorithm for community detection (Blondel et al., 2008), which is recognized for its computational efficiency and its ability to optimize a measure of modularity in large, weighted graphs (Lancichinetti and Fortunato, 2009; Fortunato and Hric, 2016). Formally, we can specify the Louvain procedure to maximize the modularity function

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i \, k_j}{2m} \right) \delta(c_i, c_j), \tag{5}$$

where A_{ij} denotes the weight of the edge between nodes i and j, $k_i = \sum_j A_{ij}$ is the weighted degree of node i, m is half the sum of all edge weights, and δ is the Kronecker delta, taking value 1 if $c_i = c_j$ and 0 otherwise. The algorithm iteratively assigns each node i to the community of a neighboring node j if doing so increases Q. Once no further local improvement is possible, each identified community is collapsed into a single node component and the process is repeated on the reduced graph until Q cannot be increased further. Following the Louvain partition, we employed a Fruchterman–Reingold (FR) layout to visualize the community structure in two dimensions. As depicted in Figure 1 each node in this layout was colored according to its community assignment, thus facilitating us with a direct inspection of cluster cohesion and potential outliers.

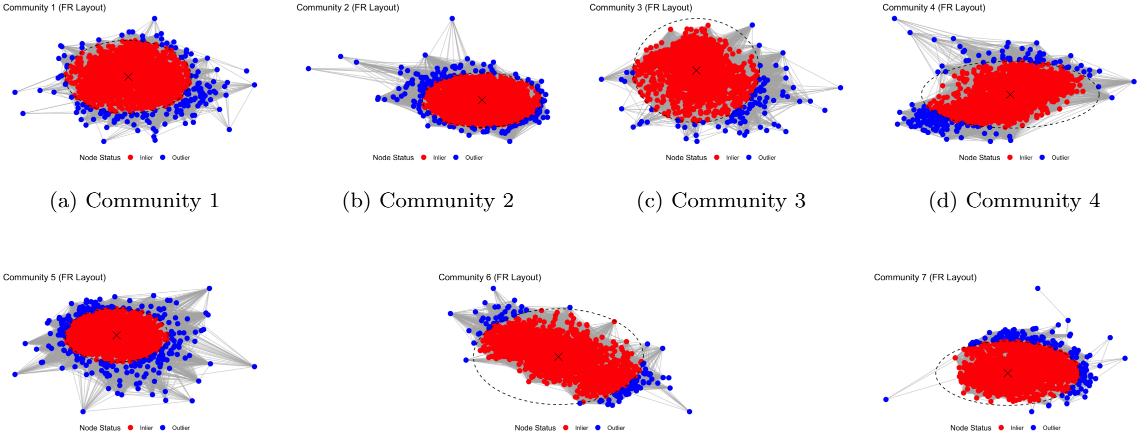


(a) Community 1 (b) Community 2 (c) Community 3 (d) Community 4

(e) Community 5 (f) Community 6 (g) Community 7

Figure 1: **Fruchterman–Reingold (FR) layout of the Louvain communities (1–7).**
The figure shows the FR layout of each detected community *before* the outlier removal. Nodes are colored by community membership.

To identify and remove loans whose network distances might artificially skew the subsequent analysis, we computed for each community (i) the centroid in the adjacency matrix space (through the column means of its submatrix), and (ii) the Euclidean distance of each loan to that centroid. We then designated as outliers those loans exceeding the 95th percentile of distances within their community. Figure 2 visualizes the resulting FR cluster layout subsequent to this approach. Through this pruning step we retrieve sufficiently dense community structures that ensure a strong cohesive loan-pair similarity for the next stages of analysis.[16]



(a) Community 1

(b) Community 2

(c) Community 3

(d) Community 4

(e) Community 5

(f) Community 6

(g) Community 7

Figure 2: **Fruchterman–Reingold (FR) layout of the Louvain communities (1–7).**
The figure shows the FR layout of each detected community *after* the centroid-based outlier removal. Red nodes are considered "inliers," and blue nodes flagged as outliers within each community.

After outlier removal, we recomputed each community's membership, obtaining a robust partition of the loan network that captures meaningful risk homogeneity. For each loan, we then appended ex-post performance indicators, such as the loan default. Table 2 summarizes the principal characteristics of the resulting communities, including total loan principal, weighted average interest rate, and weighted default rate. The detected community sizes range from 1,696 to 4,649 loans, with total principal spanning approximately €3.5 million to €10.35 million. Weighted interest rates primarily lie between 18.2% and 20.2%, while weighted default rates exhibit broader variability, from around 25% to over 50%. This dispersion suggests that the platform's assigned rates are not fully aligned with the underlying credit risk in that the default rate does not seem to be affected by loan prices and vice versa.

[16] A similar check was performed using the FR-layout distances to confirm the consistency of adjacency-based outlier detection.

Table 2: Summary Statistics of Loan Communities

Community	**Total Loan Amount (€)**	**Wtd. Avg. Interest (%)**	**Wtd. Avg. Default (%)**	**Loans**
1	4,773,348	20.2	25.1	2,938
2	10,356,141	19.7	32.1	3,154
3	3,507,685	19.2	36.5	1,696
4	8,633,074	18.2	50.9	3,451
5	6,425,380	18.3	45.6	4,649
6	6,964,902	18.5	31.6	2,085
7	5,892,359	18.8	28.2	2,924

Note: Weighted averages are computed by loan principal. "Loans" indicates the number of contracts retained in each community after excluding outliers.

From this insight we can identify that the distinct communities within the risk similarity network of loans highlight heterogeneity in credit pricing among borrowers and loan characteristics with similar risk attributes on Bondora. From the observed variations in default rates and interest rates we infer that the communities capture meaningful differences in the allocation of borrower risk profiles with the associated interest rates.

4.1.4. Loan Pricing and Defaults

To further investigate whether the interest rates assigned on the Bondora platform significantly reflect the given level of default risk per identified loan risk segment, we proceed with the following analysis. Similar to Di Maggio and Yao (2021) and Franks et al. (2021) we examine whether interest rates are a valid predictor of ex post loan performance measured by the occurence of loan default. Accordingly, we utilize a two-step LASSO-based variable selection procedure that allows for additional flexibility in capturing the time-varying expansion of Bondora's borrower pool.[17] Specifically, we estimate the following logistic model for each risk-based community c:

$$\mathrm{DP}_{i,c} = \beta_1 r_{i,c} + \alpha X_{i,c} + \varepsilon_{i,c}, \tag{6}$$

where $\mathrm{DP}_{i,c}$ represents the probability that loan i in community c defaults. The variable $r_{i,c}$ denotes the interest rate at issuance, and the vector $X_{i,c}$ consists of additional borrower- and loan-specific features (e.g., total income, debt-to-income ratio, loan amount, monthly payment, number of previous loans, etc.). We also include year-specific time fixed effects to capture the temporal character of the loan screening performance, changes in platform expansion, and the borrower composition over time.

[17] For a detailed discussion of LASSO and variants of LASSO, see Tibshirani (1996, 2011).

Table 3: Post-LASSO Logistic Regression Results by Community (with Year Fixed Effects)

Dependent Variable:	**Default (1 = default, 0 = no default)**						
Variable	**Community**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Interest Rate	0.015	0.086^{***}	0.029^{**}	0.053^{***}	0.029^{***}	0.092^{***}	0.049^{***}
Total Income	-0.541^{***}	-0.514^{***}	—	-0.351^{***}	-0.360^{***}	-0.612^{***}	-0.648^{***}
Monthly Payment	0.198	0.363^{**}	-0.254^{***}	-0.042	0.147	0.323^{***}	0.059
Loan Amount (Log)	0.006	-0.073	—	0.082	-0.256	—	—
No. of Prev. Loans	-0.160^{***}	-0.052^{**}	-0.058	0.075^{***}	0.035^{***}	-0.005	0.000
Debt to Income	—	—	0.007	-0.001	0.027	0.002	-0.008^{*}
Age	-0.011^{**}	0.004	-0.006	0.003	0.002	0.027^{***}	0.015^{***}
Maturity	0.065^{***}	—	0.024^{***}	—	0.066^{***}	-0.006	0.055^{***}
Free Cash Flow	—	—	-0.151^{***}	-0.035	-0.172^{**}	-0.027	-0.047
Liabilities	-0.053^{**}	-0.036^{*}	—	0.129^{***}	0.063^{**}	-0.082^{**}	0.061^{**}
Issue Year FE							
2014	—	—	—	—	—	-0.208	-0.685
2015	—	—	-0.903^{***}	-0.178	—	—	—
2016	—	—	—	—	—	0.478^{*}	-0.206
2017	0.454	1.324^{*}	0.209^{*}	0.372^{**}	0.419^{***}	1.197^{***}	0.138
2018	0.207	0.978^{***}	—	0.360^{*}	—	0.477	-0.457
2019	0.709^{***}	1.420^{***}	—	0.266	0.140^{*}	-0.004	-0.305
2020	—	—	—	-0.564^{*}	-0.242	-1.548^{***}	-1.233^{***}
2021	-0.976^{***}	-0.425^{***}	—	-1.749^{**}	-2.649^{**}	-1.628^{***}	-2.328^{***}
2022	-14.641^{***}	-16.218^{***}	—	—	—	-15.715^{***}	-13.272^{***}
McFadden's R^2	0.199	0.142	0.062	0.029	0.040	0.179	0.102
AIC	2105.93	3155.06	1935.82	4540.25	6143.37	1903.89	2863.87
BIC	2183.58	3227.51	1984.31	4632.04	6233.46	1999.27	2965.32

Note: This table presents post-LASSO logistic regression results per community. A two-step approach was applied: LASSO selection with non-standardized predictors (including year fixed effects) to preserve economic interpretability, followed by refitting an unpenalized logistic regression model without an intercept. "Interest Rate" is the nominal rate at issuance. "Loan Amount (Log)" is log(Loan Amount + 1). "Maturity" is in months. Year fixed effects (FE) reflect the loan's issuance year. Dashes (—) indicate variables not selected for the given community. Robust HC1 standard errors were used. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3 presents the final post-LASSO estimates across the seven communities. In line with other studies examining fintech lending markets (Vallée and Zeng, 2019; Franks et al., 2021; Di Maggio and Yao, 2021; Dorfleitner et al., 2016), we observe that the interest rate is generally a significant predictor of default in several communities but the magnitude and significance of its coefficients vary markedly across the different risk segments. Most communities exhibit a positive and significant effect of interest rate on the probability of default, whereas in Community 1, the coefficient remains positive but not statistically significant at conventional levels. This pattern is in line with the aspect of risk-based pricing that concludes the premium paid per unit of risk to increase with time and given level of default risk (Edelberg, 2006; Phillips, 2018). In terms of the magnitude of the β coefficients of the interest rate it becomes observable that the underlying default risk is not uniformly reflected in the price of credit across the different risk levels represented by the communities. In principle a better set pricing algorithm should yield a higher correlation between the interest rate and ex post performance of the underlying credit (Rajan et al., 2015). In communities where the coefficient is large and significant, the platform appears to incorporate default risk more systematically into the interest rates, effectively aligning credit risk with pricing.

To interpret these findings more concretely, we compute for each community the percentage change in the probability of default associated with a one standard deviation increase in the assigned interest rate.[18] From this we can infer that the assigned interest rate in communities 2 and 6 contributes to a larger proportion to the default likelihood of loans in these clusters of about 46% and 49%, which is roughly twice the elasticity reported by Wei and Lin (2016) for Prosper and Dorfleitner et al. (2016) for Smava respectively. The higher elasticities in communities 2 and 6 further imply that posted rates embed borrower risk more holistically in these segments than elsewhere on the platform. Hence, changes in the interest rate for these communities are more consequential in shifting the default outcome, resulting in a better alignment of pricing with credit risk. When contrasting this finding with the average interest rates charged per community in Table 2 and the coefficient values for the control variables of borrower liquidity (Total Income, Liabilities) and payment sensitivity (Monthly Payment, Debt-to-Income) it is observable that for both communities (2 & 6) this observation may indeed be driven by mispriced credit risk. However, we can not completely rule out at this stage that heightened borrower sensitivity in these segments is influencing the results. By contrast, Communities 3, 4, and 5 see more modest increases in default odds (ranging from 13% to 26%), suggesting that the assigned interest rates may capture less of the underlying default risk.

We can interpret these discrepancy as a sign of potential mispricing, insofar as some risk segments react more strongly than others to interest rate adjustments. Specifically, for community 3 we can observe that the interest rate for early issued loans only explains moderate increases in default odds, which aligns with H2. However, interestingly earlier issued loans as indicated by the significant time fixed effect dummy do negatively effect the default odds of loans in the respective community. One reason for this observation might link to the rather homogeneous borrower pool that Bondora hold prior to its transition to a risk-based rating approach, which was smaller in size and less diverse.[19] Hence, with increased platform scaling and more heterogeneous borrower clienteles the risk-based pricing mechanism may have been challenged in subsequent years (2015-2019) due to insufficient screening experience. We can trace this effect specifically across most detected communities where loans issued in later periods (2020-2022) are also available. Here, time-fixed effect dummies of the later years have a significant negative effect of larger magnitude on the default probability of loans, congruent with the observed learning effect of marketplace lenders to score credit risk more accurately with time and platform growth (Jagtiani and Lemieux, 2019; Vallée and Zeng, 2019; Lin and Viswanathan, 2016). This finding corroborates H3. Therefore, similar to Dömötör et al. (2023) we subsume that false or insufficient process caring at the screening or verification level of borrowers

[18] see Table A.7 for more details.

[19] Bondora significantly expanded its operation, observable from 2019 onwards in roughly five fold quantities from approximately 300, 000 to 1, 345, 873 borrowers. See (https://bondora.com/en/public-statistics/) for more details.

in early periods is likely contemplating to the observed findings. This reinforces the notion that Bondora's pricing strategies depend as expected on the credit terms and borrower characteristics but vary markedly across the risk segments.[20] Our observation also reflects on earlier evidence that fintech lenders screening efforts appear more accurate on the intensive margin (Di Maggio and Yao, 2021) but still fail to achieve information- and segment conform pricing efficiency. Specifically, the uniform price of a loan does markedly vary across communities in predictive power to explain the odds of default, thus indicating that the interest rate does not equally reflect all risk-relevant information at the time of loan issuance for all borrower segments. If the interest rate would fully incorporate all available information, then increases in the interest rate should be associated with a proportional (or close-to-one) increase in the probability of default (Franks et al., 2021).

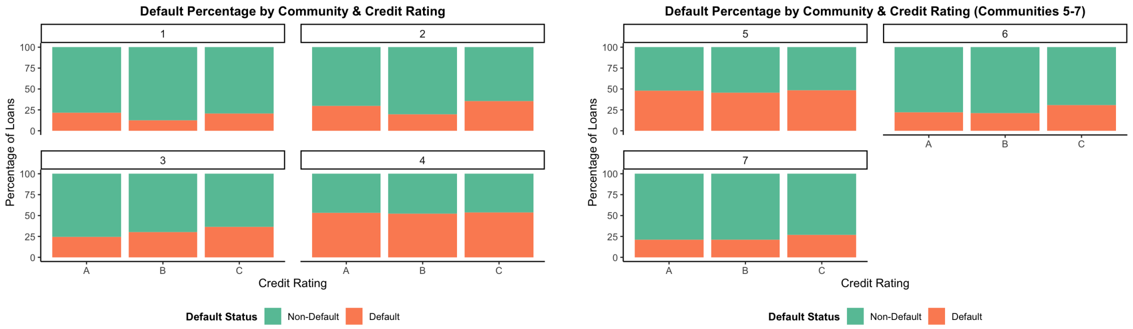
4.1.5. Initial Mispricing Identification Based on Default Outcomes

We extend our ex-post analysis to identify potential mispricing of loans based on observed default outcomes. By further examining the relationship between assigned interest rates and actual loan performance across the detected communities, we aim to uncover discrepancies that further suggest misalignment between loan inherent risk and pricing detected in the previous part of this study. We begin by exploring the relationship between default outcomes and the credit ratings applied within each identified community. Since ratings (A, B, C) reflect the platform's ex-ante assessment of default likelihood, a sound pricing mechanism should, in principle, exhibit lower default rates in higher-rated categories (e.g., A) and higher default rates in lower-rated categories (e.g., C) coherent with risk-based pricing (Duffie and Singleton, 1999; Phillips, 2018). Figure 3 presents the default rate distribution, segmented by credit rating for Communities 1–4 and 5–7, respectively. From the plots we observe that while certain patterns are broadly consistent with the rating system (e.g., a greater number of defaults among lower-rated categories), there also exist communities where default frequencies for higher-rated loans are unexpectedly elevated.

In particular, some communities show higher default rates among A-rated loans, suggesting potential underpricing of risk within these segments. Conversely, in selected communities, C-rated loans exhibit relatively modest default frequencies in relation to safer rating segments within the respective community, potentially indicating an overpricing of risk. For example, Communities 2, 4, and 5 display a substantial volume of defaults even among A-rated loans, while Communities 1 and 7 show comparatively low default rates for C-rated loans, which appears inconsistent with risk-based

[20] We run additional tests as in Di Maggio and Yao (2021) to rule out that other important credit information captured by the credit ratings might further influence the analysis as determinant of default risk. Consequently, we estimate the degree of information results in Bondora's price of credit (Interest) via the following pooled regression $r_{i,c} = \beta_1 \, \mathrm{CR}_{i,c} + \alpha \, X_{i,c} + \varepsilon_{i,c}$, where $r_{i,c}$ denotes the interest rate for loan i in community c. The term $\mathrm{CR}_{i,c}$ captures credit rating dummies (e.g., A, B, C), while the vector $X_{i,c}$ includes borrower and loan-specific characteristics. The term $\varepsilon_{i,c}$ is the idiosyncratic error. The results are displayed in Table A.8. From the regressions we observe that the credit ratings explain most of the variation in the interest rate which confirms Bondora's rating-based pricing mechanism and the degree of pricing informativeness contained by the credit ratings.

pricing. Additionally, the variation in default distributions across communities for similar credit ratings suggests that the platform's rating system may not capture borrower risk accuratly across all clusters, potentially leading to systematic mispricing in some segments. This finding provides evidence that the platform might not uniformly price loans well on the intensive margin.



(a) Communities 1–4

(b) Communities 5–7

Figure 3: **Stacked bar charts illustrating the default rate distribution by credit rating (A, B, C) per community.** The y-axis indicates the percentage of loans within each credit rating category that resulted in default (1) or non-default (0). Panel (a) presents Communities 1–4, while Panel (b) displays Communities 5–7.

By combining ex-post default outcomes with ex-ante credit ratings in the context of community membership, we also obtain a preliminary view of potential mispricing: Communities with unexpectedly high default rates among A-rated loans, for instance, may indicate underpricing of those loans, while communities where C-rated loans exhibit fewer defaults than anticipated could suggest overpricing. To systematically identify potentially mispriced loans, we implemented a classification approach that evaluates deviations from community-specific interest rate percentiles. The procedure begins by computing the 25th percentile (Q_1) and the 75th percentile (Q_3) of interest rates within each community, establishing the lower and upper bounds for expected pricing given the risk level of loans within each cluster. Loans that deviated significantly from these thresholds were flagged as mispriced based on their ex-post default status. Specifically, loans that defaulted despite carrying interest rates below the first quartile ($Interest < Q_1$ and $default = 1$) were classified as underpriced, whereas loans that did not default, yet carried interest rates above the third quartile ($Interest > Q_3$ and $default = 0$) were categorized as overpriced. To further quantify the degree of mispricing, we calculated a mispricing magnitude, denoted as Δ, which measures the difference between a loan's assigned interest rate and the relevant percentile. This ensures that the classification not only captures mispricing occurrences but also provides a continuous measure of deviation from the expected pricing range.[21] The Δ value was computed as the difference between a loan's interest rate and the lower quartile for underpriced loans, while for overpriced loans, it reflects the difference from the upper quartile:

$$\Delta = \begin{cases} Interest - Q_1, & \text{if loan is underpriced} \\ Interest - Q_3, & \text{if loan is overpriced} \end{cases} \tag{7}$$

After classifying individual loans, the proportion of mispriced loans was computed for each community, along with the average Δ values for underpriced and overpriced loans, allowing for an aggregate-level assessment of pricing inefficiencies. Table 4 summarizes the distribution of mispriced loans across communities, while Figure 4 visually presents the relationship between mispricing proportions, weighted average interest rates, and default rates.

Table 4: Potentially Mispriced Loans by Community

Comm.	**Total**	**Underp.**	**Overp.**	**Avg. Δ (Underp.)**	**Avg. Δ (Overp.)**	**% Mispriced**
1	2,938	101	580	-1.86	2.92	23.2%
2	3,154	143	478	-1.70	3.49	19.7%
3	1,696	115	236	-1.60	2.46	20.7%
4	3,451	453	395	-1.70	2.64	24.6%
5	4,649	496	573	-1.53	2.66	23.0%
6	2,085	109	337	-1.70	2.75	21.4%
7	2,924	165	510	-1.69	2.98	23.1%

Note: Mispriced loans deviate from the community-specific interest rate percentiles. Delta (Δ) measures deviation from the relevant quartile, with negative values indicating underpricing and positive values reflecting overpriced loans. Windsorization removes extreme outliers (1st and 99th percentile of Δ values). % Mispriced = (Underpriced + Overpriced)/Total × 100.

[21] Given that extreme interest rate deviations could bias the results, we applied a Windsorization process to remove outliers. Any loan whose Δ value fell below the 1st percentile or exceeded the 99th percentile of the Delta distribution was excluded from the analysis. This ensures that the identification of mispricing is robust to extreme cases that may arise due to data anomalies or rare lending behaviors.

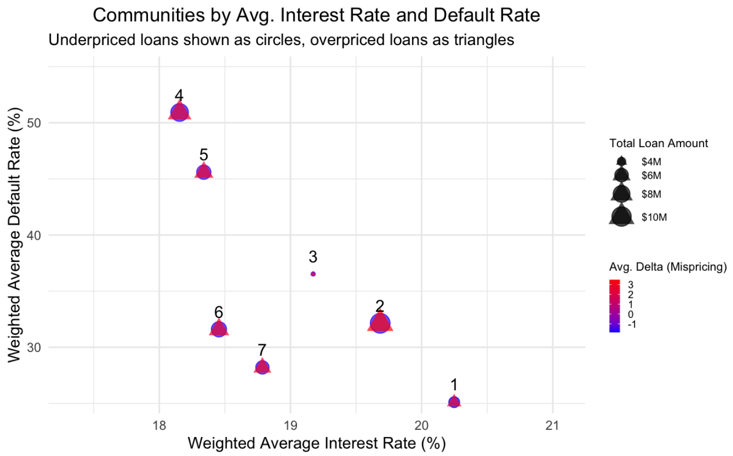


Figure 4: **Communities by Weighted Average Interest Rate, Default Rate, and Mispriced Loan Proportion.** Circle markers represent underpriced loans, triangle markers denote overpriced loans. Marker color intensity indicates the magnitude of Δ.

The results indicate substantial heterogeneity in the extent and nature of mispricing across loan communities. Communities 1 and 5 contain the highest number of overpriced loans (580 and 573, respectively), suggesting that in these segments the assigned interest rates often exceed the platform's risk-adjusted benchmark. By contrast, Communities 4 and 5 exhibit the largest counts of underpriced loans, with 453 and 496 instances, respectively. This finding indicates that, for a considerable subset of borrowers in these communities, the platform appears to underestimate credit risk, potentially giving rise to adverse selection. These patterns are further illustrated in Figure 4, which maps each community by its weighted average interest rate and default rate. Theoretically, risk-based pricing implies a monotonic relationship between default risk and interest rates, whereby higher default probabilities should be compensated with proportionally higher interest rates (Duffie and Singleton, 1999; Edelberg, 2006; Phillips, 2018). However, the figure reveals several communities that deviate from this principle. For instance, Community 1 shows a relatively high average interest rate but only moderate default risk, suggesting that borrowers may be overcharged relative to their actual credit risk. Conversely, Community 4 exhibits the highest default rate among all segments, yet its average interest rate remains comparatively low, which indicates potential underpricing. Such discrepancies are inconsistent with efficient risk-adjusted pricing and may reflect latent information asymmetries or oversights in the platform's pricing algorithm.[22] Moreover, the visual encoding of mispricing intensity reinforces this divergence: Communities with the highest average deltas (i.e., largest deviations from expected pricing) do not align with what a risk-based pricing approach

[22] Johnson et al. (2023) find similar evidence for loan data from LendingClub, Upstart, and Avant, suggesting the existence of pricing inefficiencies and arbitrary risk-return relationships in P2P lending.

would predict in the presence of full information and rational expectations, thus confirming H1.

4.2. Refined Pricing Analysis

4.2.1. Community-based Interest Rate Modeling and Price Quotation

To ensure that our initial classification of mispriced loans does not erroneously include loans that defaulted due to unforeseeable shocks, rather than from genuinely underestimated risk, we refine the identification by developing community-specific predictive models. In particular, an ex-post classification can label a loan as "underpriced" merely because the borrower defaulted, even if the default arose from an unpredictable event (for instance, sudden unemployment or a health emergency), rather than a systematic miscalculation of credit risk. To eliminate this potential hindsight, we implement a more granular, ex-ante modeling approach within each community, to better distinguish mispriced loans from those affected by idiosyncratic events. Hence, we approximate the platform's pricing mechanism by training GBMs on loans initially identified as properly priced, using borrower- and loan characteristics.[23] These community-specific models belong to the class of non-parametric ensemble learners that are found to flexibly capture nonlinear relationships among predictors (Berg et al., 2022), thereby accounting for potential heterogeneity in risk profiles and pricing strategies across borrower segments. Subsequently, we apply these trained GBMs to predict expected interest rates for loans initially classified as mispriced. Comparing these predictions with actual interest rates allows us to precisely re-assess and verify mispricing classifications within each risk-based community. Before constructing the GBMs, we employ an RF regressor to identify the most relevant features for predicting interest rates similar to the modeling approach in Subsection 4.1.1. Through this selection process we ensure that only the most informative variables contribute to the pricing models.[24] For each community c, we construct a separate GBM using the subset of properly priced loans discovered in Subsection 4.1.5 within the particular community as the training dataset. The dependent variable in each model is the observed actual interest rate (r) assigned by the platform, while the independent variables include borrower demographics (age, gender, education level), financial indicators (income, liabilities, debt-to-income ratio), loan characteristics (loan amount, loan duration, monthly payment, credit rating), and temporal attributes (application time and hour of application). Accordingly the model specification for community c is defined as:

$$\hat{r}_{i,c} = f_c(X_{ij,c}; \theta_c), \tag{8}$$

[23] This initial pricing classification is detailed in Subsection 4.1.5.

[24] Specifically, we train an RF regression model using a comprehensive set of borrower demographics, financial indicators, and loan attributes while excluding variables that capture ex-post performance. The model ranks predictor importance based on the increase in mean squared error (%IncMSE), and the top 20 features are selected for subsequent GBM estimation.

where $\hat{r}_{i,c}$ is the predicted interest rate for loan i in community c, $X_{ij,c}$ represents the borrower and loan characteristics, and θ_c denotes the set of optimized GBM hyperparameters specific to community c. The results displayed in Table A.9 indicate that borrower and loan characteristics contained in the credit application explain a substantial share of the variation in interest rates, which confirms earlier evidence that fintech P2P lenders primarily rely on salient hard-information when setting rates (Di Maggio and Yao, 2021; Balyuk, 2023; Balyuk et al., 2025).

Subsequently, we generate predicted interest rates $\hat{r}_{i,c}$ for *all* loans in community c, including those flagged as mispriced in the initial classification. We then compute the difference

$$\Delta r_{i,c} \;=\; \hat{r}_{i,c} \;-\; r_{i,c}, \tag{9}$$

where a positive $\Delta r_{i,c}$ indicates that our model-implied rate differential exceeds the platform's assigned rate (potential underpricing), while a negative value suggests potential overpricing. To avoid small discrepancies from influencing the mispricing classification, we establish thresholds within each community based on the distribution of $\Delta r_{i,c}$ among the properly priced loans in the training set:

$$\theta_{\mathrm{u},c} \;=\; \mu_{\mathrm{pos},c} \;+\; \sigma_{\mathrm{pos},c}, \quad \theta_{\mathrm{o},c} \;=\; \mu_{\mathrm{neg},c} \;-\; \sigma_{\mathrm{neg},c}, \tag{10}$$

where $\mu_{\mathrm{pos},c}$ (resp. $\mu_{\mathrm{neg},c}$) and $\sigma_{\mathrm{pos},c}$ (resp. $\sigma_{\mathrm{neg},c}$) are the mean and standard deviation of $\Delta r_{i,c}$ for positive (resp. negative) values among properly priced loans. Any loan with $\Delta r_{i,c}$ greater than $\theta_{\mathrm{u},c}$ is deemed *underpriced*, while any loan with $\Delta r_{i,c}$ less than $\theta_{\mathrm{o},c}$ is deemed *overpriced.* This refined process accounts for community-level pricing norms and produces a tighter alignment of identified mispricing solely based on the community specific interest rate distribution and the platform's plausible rate-setting function.

4.2.2. Economic Significance and Implications for Investors

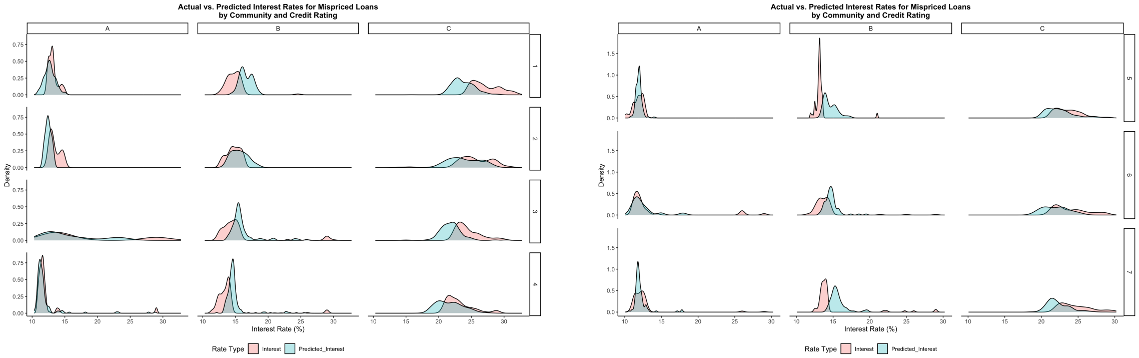
From the results of the refined pricing analysis we set out to formulate several implications for investors. Figure 5 and Table 5 summarize the results on the economic impact as well as distributions of *actual* and *model-predicted* interest rates for mispriced loans, categorized by community. Across all seven loan communities, 8.8% of contracts appear mispriced by at least one percentage point, and 8.4% by more than two points. Given an average principal of €2,230, this implies an aggregate annual transfer of €102,000 between investors and borrowers.[25] On the aggregate portfolio level this amounts to an additional 23 basis points yield on the €45.7 million loan sample which financially benefits investors at the expense of borrowers.

[25] Annual transfer is computed as $\sum_i \mathrm{Principal}_i \times (\, r_{i,c} - \hat{r}_{i,c})/100$.

Table 5: Economic Impact of Mispricing by Community

Community	**Total Loans**	**Total Principal (€)**	**Avg. Principal (€)**	**Mispriced Loans**	**% Mispriced**	**≥1 pp Loans**	**% ≥1 pp**	**≥2 pp Loans**	**% ≥2 pp**	**Annual Transfer (€)**
1	2,901	4,700,430	1,620	403	13.9%	403	13.9%	403	13.9%	–21,282
2	3,093	10,137,882	3,278	280	9.1%	280	9.1%	272	8.8%	–14,648
3	1,616	3,397,260	2,102	152	9.4%	152	9.4%	150	9.3%	–6,644
4	3,359	8,453,893	2,517	217	6.5%	217	6.5%	216	6.4%	–11,983
5	4,606	6,334,052	1,375	284	6.2%	284	6.2%	221	4.8%	–8,572
6	2,020	6,806,914	3,370	153	7.6%	153	7.6%	152	7.5%	–19,826
7	2,887	5,836,945	2,022	310	10.7%	310	10.7%	301	10.4%	–18,761
Overall	20,482	45,667,376	2,230	1,799	8.8%	1,799	8.8%	1,715	8.4%	–101,716

Notes: "Mispriced Loans" counts contracts whose actual Bondora rate deviates from the GBM-predicted rate by any amount (i.e. those flagged "Underpriced" or "Overpriced"). "≥1 pp" / "≥2 pp" counts the subset whose absolute rate-gap is at least one or two percentage points, respectively. All percentages are calculated relative to the total loans in each community (column "Total Loans"). "Annual Transfer" is the net implied flow in euros, computed as $\sum_{\text{loan}}$ principal × (ActualRate − PredRate)/100, so a negative value indicates a net transfer from borrowers to investors.



(a) Communities 1-4

(b) Communities 5-8

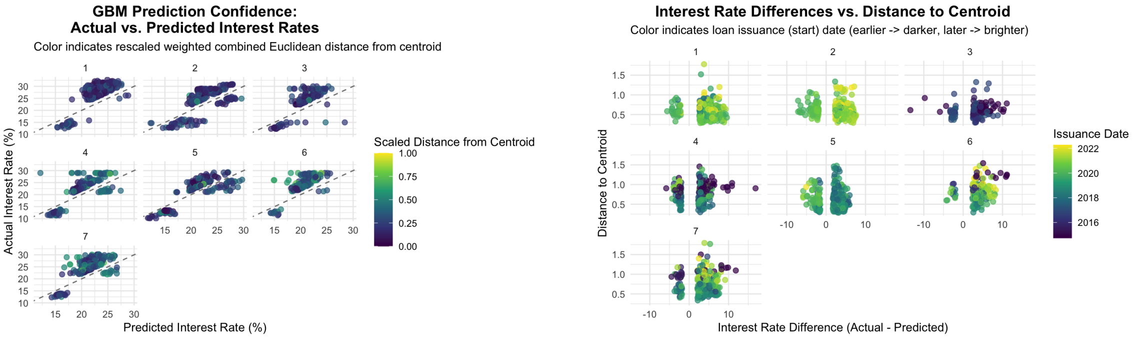
Figure 5: **Density of Actual vs. Predicted Interest Rates for Mispriced Loans by Community and Credit Rating.** *Panel (a)* displays the distributions for Communities 1-4, while *Panel (b)* presents Communities 5-8. Each density plot compares the actual interest rate distributions to the model-predicted values, revealing pricing deviations across credit ratings.

Similar to Johnson et al. (2023), we find that certain credit-rating bands appear to systematically deviate from model-implied rates, specifically for noneprime-related loans with otherwise similar risk attributes. The results are conclusive with findings in traditional consumer lending markets that find loan pricing, specifically for nonprime lenders, to not fully align with borrower credit risk (Keys et al., 2010; Argyle et al., 2020; Agarwal et al., 2018). In particular, C-rated loans exhibit higher actual interest rates than the model would predict across most communities, implying that these loans may offer a comparatively attractive returns for investors with higher overall risk preferences. Investors seeking yields above the norm could potentially benefit from the price dispersion by aligning their risk-return objectives to select C-rated loans within these segments. By contrast, the results for B-rated loans indicate that Bondora's posted interest rates may understate the true risk profile of loans in this rating band. Specifically, communities that exhibit persistent gaps between actual and model-predicted rates for B-rated loans suggest a potential underpricing of risk by the platform. In practical terms, investors funding these loans might face higher-than-anticipated default probabilities without adequate compensation. For A-rated loans, the evidence is more nuanced. Communities 1 and 2 feature densities in which the model's predicted rates are modestly below Bondora's actual rates, suggesting that A-rated borrowers here are charged a premium relative to their estimated default risk. This discrepancy would favor investors at the cost of borrowers, as the premium confers additional protection above the model-implied fair rate. Conversely, A-rated loans in Community 3 show the opposite tendency, with actual rates surpassing model-based benchmarks by a wider margin. In this segment, investors may find the interest spreads insufficiently compensatory for the implied credit risk. In turn, Communities 4–7 demonstrate narrower differences between actual and predicted rates for A-rated loans, implying fewer arbitrage opportunities in these segments.

4.3. Robustness Checks

To examine the confidence of our GBM-based pricing models and test confidence stability over time, we implement a set of controls. Initially, we visually compare the rate differentials with each mispriced loan's weighted Euclidean distance from the centroid of properly priced loans within the same community.[26] Conceptually, loans whose characteristics diverge substantially from this community centroid are presumed to be relatively riskier to the core segment of similarly risky borrowers, thus suggesting lower model confidence in the predicted rate for those observations. By contrast, loans positioned closer to the centroid are more likely to reflect subtler deviations in risk profiles to properly priced loans that the model can coherently interpret with more confidence as potentially over- or underpriced. Figure 6 visualizes the two dimensions, the residual magnitude and the centroid distance jointly.

[26] We compute the weighted Euclidean distance using the most influential predictor variables identified in Subsection 4.2.1 via the Random Forest model. Each predictor is assigned a weight proportional to its feature-importance score from the RF, ensuring that more critical risk factors exert greater influence on the distance calculation. For each community, we define the centroid of properly priced loans by taking the mean of continuous predictors and the mode of categorical variables. This provides a representative "core" profile of adequately priced credit, serving as the point of reference for subsequent distance computations.



(a) **Actual vs. Predicted Interest Rates Across Communities**

(b) **Interest Rate Difference vs. Distance to Centroid**

Figure 6: **Visualizing Mispriced Loans Relative to Their Centroid.**
Panel (a) shows a facet-wrapped scatter plot of actual versus model-predicted interest rates for mispriced loans in Communities 1–7. Each point's color encodes its rescaled, weighted Euclidean distance to the centroid of properly priced loans in the same community. Darker shades indicate loans more akin to their centroid (i.e., the model is comparatively more confident), while lighter shades mark those that deviate substantially from the prevailing patterns of properly priced loans, implying lower model confidence. The dashed 45-degree line represents the hypothetical scenario of perfect alignment between actual and predicted rates. *Panel (b)* depicts a scatter plot of the prediction error (actual rate minus predicted rate) on the horizontal axis against the centroid distance on the vertical axis. Each point's color gradient reflects the loan's issuance date, providing a temporal lens on how atypical or poorly estimated loans may emerge. Points far from the origin on the x-axis exhibit greater mispricing, whereas those high along the y-axis are especially dissimilar to the centroid. Taken together, these panels illustrate how combining mispricing magnitude with centroid distance and temporal cues facilitates deeper insight into the credit model's stability and confidence for each loan.

From the observations we infer a potential limiting factor of our results: the inherent distance for some loans positioned further away from the fairly priced community centroid, may be induced from our pricing algorithm's omission of non-accessible proprietary borrower variables, such as detailed credit-bureau records and digital behavior metrics.[27] However, as can be seen in *Panel (a)* of Figure 6, most loans classified as mispriced remain within the closer proximity of the centroid of properly priced loans within their community, indicating that their inherent risk profile closer resembles the risk characteristics of the fairly priced loans. Yet, for mispriced loans specific to communities 4, 6, and 7 we can observe that more loans appear further distant from the community centroid indicated by the brighter color gradient. This indicates that the model certainty for these predictions is weaker, which points to the aspect that loans in these community segments inherently differ from the risk profile of the fairly priced centroid. One may recall that the centroid reflects the true pricing mechanism of Bondora, where the assigned interest rate is found to match the risk profile of loans within the loan community. Thus a further distance away from the centroid reveals a clear mismatch between the assigned interest rate and the risk profile (reflected by the credit rating) of the loan. Two potential reasons for this finding become apparent: The first reason could lie in the observed implementation of Bondora Rating in 2015, which leaves loans originated prior

[27] These metrics remain non-disclosed by Bondora as indicated in the platform's guide on the credit rating bands retrievable at
https://help.bondora.com/hc/en-us/articles/14814705732881-How-are-Bondora-risk-ratings-calculated.

to be potentially affected by less accurate risk-based pricing, due to contemporaneous borrower information and ex-post risk quantification. The second reason could lie in the observed learning effect of the platform to gradually improve the scoring accuracy over time as more borrowers seek for credit at the platform (Jagtiani and Lemieux, 2019; Lin and Viswanathan, 2016; Freedman and Jin, 2011; Vallée and Zeng, 2019), thereby increasing the pool of available hard information for accurate risk assessment. We investigate this further in *Panel (b)* of Figure 6. Indeed from the plot it becomes observable that mispriced loans originating closer to the beginning of the sample period showcase further distance from the fairly priced centroid and stronger deviation from the predicted interest rate. These loans are primarily grouped in the respective communities 4, 6, and 7, which gives rise to the notion that insufficient credit scoring skills stemming from early operations of the lending platform might form the basis for the decreasing model confidence in these communities. Nevertheless, this argument might seem counterintuitive, with findings for community 3 that contains primarily loans from the early sampling period but otherwise suggests sufficient model confidence by the GBM. However, it is important to consider that all loans in community 3, including the fairly priced centroid, congruently stem from the close post-2015 adoption period of Bondora Rating, which might influence the confidence measure (distance to centroid) as the fairly priced centroid itself might be subject to coarser loan screening practices by the lending platform for this risk bracket.[28] For communities 4, 6, and 7, when contrasting the older loans (issue date closer to post-2015) with the fairly priced centroids, originating from later sampling periods, it becomes apparent that older mispriced loans deviate considerably from the inherent risk profile of the centroid, thus reducing overall model confidence in the prediction.[29] From this evidence, we acknowledge that our pricing mechanism may not fully capture the inherent risk of loans stemming from early operations of Bondora. Nevertheless, for loans originating in later periods (2018-2022) our pricing mechanism remains confident about the predicted interest rates based on the loan's underlying risk profile and the determinants of interest rate within each loan community.

5. Conclusion

Building on the extensive literature on credit risk and pricing in marketplace lending, this study analyzes the presence and determinants of pricing inefficiencies in the primary market of the European P2P lending platform Bondora. By applying a network-based loan segmentation and pricing analysis, we provide a multifaceted approach to detecting anomalies where posted prices diverge systematically from fundamental credit risk factors. Our analysis uncovers a significant fraction

[28] As displayed in Figure A.7, community 3 contains loans with the oldest issuance date.

[29] This is the case for communities 4, 6, and 7 as the issuance date of the fairly priced centroid for each community is more recent. For community 4 it is 2018-01-31. For community 6 it is the 2019-02-20. For community 7 it is the 2018-08-22.

of loans that are mispriced: some clusters exhibit persistently high default rates despite relatively low interest charges, while others feature materially elevated rates combined with low default frequencies. These outcomes challenge the premise of frictionless price discovery on P2P platforms, highlighting persistent information asymmetries and potential biases in borrower screening. Our results indicate that Bondora's risk-based credit-scoring framework failed to capture the full spectrum of borrower risk, particularly during the platform's early expansion. However, we detect that the platform's screening and scoring ability is influenced by a learning process that improves risk alignment with posted prices over time. From a methodological perspective, our multi-stage approach, combining community detection with refined price estimates, offers a new perspective for both researchers and practitioners to assess pricing anomalies in marketplace lending. We demonstrate that controlling for community-level heterogeneity provides deeper insights into how pricing errors persist in distinct borrower segments. Indeed, these patterns can hold particular significance for both retail and institutional investors who rely on automated investment processes and posted interest rates in constructing P2P loan portfolios. Beyond these immediate applications, our findings have broader implications for market design. In principle, P2P platforms aim to foster transparency and reduce transaction costs relative to traditional finance; however, we show that moral hazard and incomplete disclosure can still hinder efficient risk-based pricing. Our results support the notion for enhanced borrower screening tools and more informative disclosures at loan origination. By more closely aligning market prices with borrowers' default probabilities, P2P platforms could reinforce investor confidence, mitigate adverse selection, and promote a more stable flow of credit. Future research might validate these results using data from other lending platforms or more granular measures of borrower risk, thereby adding further clarity on the interplay between technological innovation, market microstructure, and credit allocation.

Acknowledgments

Lennart John Baals, Jörg Osterrieder, and Branka Hadji-Misheva acknowledge the financial support of the Swiss National Science Foundation (SNSF) under project number `100018E_205487`. We also gratefully acknowledge the support of the Marie Sklodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym: DIGITAL, Project No. 101119635.

The authors are grateful to Ali Hirsa, Viraat Singh, Lucia Gomez Teijeiro, Yiting Liu, and Julius Kooistra for valuable conversations and advice. Lennart John Baals gratefully acknowledges the hospitality of the Industrial Engineering and Operations Research Department at Columbia University during his 2025 research visit, which provided an excellent environment for developing this study. Any remaining errors are the authors' sole responsibility.

CRediT authorship contribution statement

Lennart John Baals: Conceptualisation, Investigation, Validation, Formal analysis, Methodology, Data curation, Visualisation, Project administration, Writing - Original draft, Writing - Review & Editing. **Jörg Osterrieder:** Supervision, Project administration, Methodology, Funding acquisition, Resources, Writing - Review & Editing. **Branka Hadji-Misheva:** Supervision, Project administration, Methodology, Resources, Funding acquisition, Writing - Review & Editing. **Yizhi Wang:** Validation, Visualisation, Project administration, Methodology, Resources, Writing - Review & Editing.

Data Availability

The data will be made available upon request.

Appendix A. Additional Tables and Figures

Table A.6: Top 20 Risk Factors Identified by Random Forest Model

Feature	**Type**	**Description**
time	Continuous	Time spent on application in decimal hours
inc.total	Continuous	Total income (log-transformed)
AmountOfPreviousLoansBeforeLoan	Continuous	Total amount of previous loans taken
educ.5	Binary	Higher education attainment
MonthlyPayment	Continuous	Monthly loan payment amount
liab.l	Continuous	Total liabilities (log-transformed)
loan_ amount	Continuous	Loan amount requested by borrower
Age	Continuous	Borrower's age in years
duration.12	Binary	Loan duration of 12 months
duration.60	Binary	Loan duration of 60 months
NoOfPreviousLoansBeforeLoan	Continuous	Number of previous loans taken
joint.ownership	Binary	Borrower has joint property ownership
duration.24	Binary	Loan duration of 24 months
duration.09	Binary	Loan duration of 9 months
educ.4	Binary	Secondary education attainment
ver.4	Binary	Verification level 4
FreeCash.l	Continuous	Available free cash after expenses (log-transformed)
duration.36	Binary	Loan duration of 36 months
no.liab.01	Binary	Number of liabilities ≥ 1
DebtToIncome	Continuous	Borrower's debt-to-income ratio

Note: This table lists the top 20 predictor variables contributing most to the construction of the Gower's distance-based similarity network in the Random Forest model. Binary variables indicate specific borrower attributes (e.g., education, loan duration, ownership status), while continuous variables represent financial and demographic factors. Variables are sorted in descending order of relative importance.

Table A.7: Summary of Interest Rates and One-Standard-Deviation Effects on Default Odds

Community	**n**	**Mean** r	**SD** r	**Median** r	**Min** r	**Max** r	$\boldsymbol{\beta_r}$	**%∆**Odds
1	2,901	20.4	5.18	20.2	9.90	32.8	0.0146	7.85
2	3,093	20.0	4.44	19.9	9.16	31.4	0.0858	46.3
3	1,616	19.1	4.32	18.8	10.7	35.0	0.0295	13.6
4	3,359	17.9	4.30	17.6	9.16	35.0	0.0530	25.5
5	4,606	17.5	4.41	16.9	9.88	31.0	0.0293	13.8
6	2,020	18.1	4.33	17.5	9.16	34.0	0.0916	48.8
7	2,887	18.1	4.68	17.5	9.53	34.0	0.0511	27.0

Note: This table reports, for each community, the number of loans (**n**), the mean, standard deviation, median, minimum, and maximum observed interest rates (r), and the estimated coefficient (β_r) from the logistic model. The final column (% ∆Odds) indicates the percentage change in the odds of default associated with a one-standard-deviation increase in r. Formally, the log-odds effect is computed as $\beta_r \cdot \mathrm{SD}(r)$, so that $\%\Delta\mathrm{Odds} = [\exp(\beta_r\,\mathrm{SD}(r)) - 1] \times 100$. All figures are derived from the post-LASSO community-level logistic regressions described in Table 3.

Table A.8: Pooled Linear Regression Results by Community

Dependent Variable:	**Interest Rate**						
Group	**Community**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Credit Rating Categories							
Credit Rating A	13.282***	13.233***	25.948***	16.182***	13.473***	10.940***	14.297***
	(0.782)	(0.785)	(1.325)	(0.825)	(0.457)	(0.983)	(0.631)
Credit Rating B	17.458***	17.670***	27.630***	18.550***	16.709***	14.037***	17.750***
	(0.786)	(0.780)	(1.256)	(0.802)	(0.454)	(0.960)	(0.627)
Credit Rating C	24.876***	23.719***	32.812***	23.798***	23.041***	19.482***	24.045***
	(0.790)	(0.779)	(1.235)	(0.799)	(0.454)	(0.968)	(0.621)
Credit Term Variables							
Maturity	-0.000	—	-0.058***	—	-0.039***	0.002	-0.052***
	(0.005)	—	(0.008)	—	(0.005)	(0.006)	(0.004)
Loan Amount (Log)	-0.043	-0.044	0.172	0.018	0.215***	-0.019	0.225***
	(0.069)	(0.048)	(0.107)	(0.060)	(0.043)	(0.067)	(0.056)
Borrower Characteristics							
Total Income	0.109	0.049	-0.350	-0.450***	0.066	0.410***	-0.048
	(0.110)	(0.104)	(0.437)	(0.113)	(0.052)	(0.124)	(0.085)
Debt to Income	—	—	-0.030***	-0.010**	0.008	-0.020***	-0.011**
	—	—	(0.007)	(0.004)	(0.010)	(0.005)	(0.004)
Age	-0.008$^{\cdot}$	0.001	-0.026***	0.005	-0.012***	-0.015**	-0.016***
	(0.005)	(0.005)	(0.007)	(0.004)	(0.002)	(0.006)	(0.004)
Free Cash Flow	—	—	-0.122	0.526***	-0.203***	0.225***	0.065*
	—	—	(0.099)	(0.031)	(0.055)	(0.040)	(0.028)
Liabilities	-0.190***	-0.183***	-0.859*	-0.238***	-0.248***	-0.106***	-0.130***
	(0.021)	(0.020)	(0.340)	(0.042)	(0.030)	(0.032)	(0.024)
Model Statistics							
R^2	0.984	0.983	0.978	0.977	0.988	0.978	0.985
Adj. R^2	0.983	0.983	0.978	0.977	0.988	0.978	0.985
AIC	14012.65	14932.25	8076.07	16445.90	19275.08	9820.53	13015.21
BIC	14066.41	14980.55	8135.34	16507.09	19345.86	9882.25	13080.86

Note: This table presents linear regression results with robust standard errors for each community (1–7), using the **interest rate** at issuance as the dependent variable. Estimations are conducted without an intercept term. "Credit Rating A/B/C" are categorical variables indicating borrower ratings assigned by the platform. "Loan Amount (Log)" is the log-transformed loan size. "Maturity" denotes loan duration in months. All predictors are kept in their natural scale to preserve economic interpretability. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: GBM Model Performance by Community

Community	**Validation RMSE**	**Validation MAE**	**Validation R^2**
1	1.9938	1.5153	0.7792
2	1.6950	1.2937	0.8004
3	2.0405	1.4776	0.7305
4	1.8635	1.2648	0.7639
5	1.3053	0.8836	0.8748
6	1.6475	1.2284	0.7845
7	1.7254	1.2854	0.7738

Note: For each community, the GBM was trained on the "fairly priced" loans using a 70%/30% train/validation split (default-to-non-default ratio = 0.3438). Models employ 5-fold cross-validation to tune boosting iterations, learning rate, tree depth, and minimum observations per leaf to prevent overfitting. Validation RMSE (Root Mean Squared Error) quantifies average prediction error magnitude; MAE (Mean Absolute Error) measures average absolute deviation; and R^2 indicates the share of variance in quoted interest rates explained by the model.

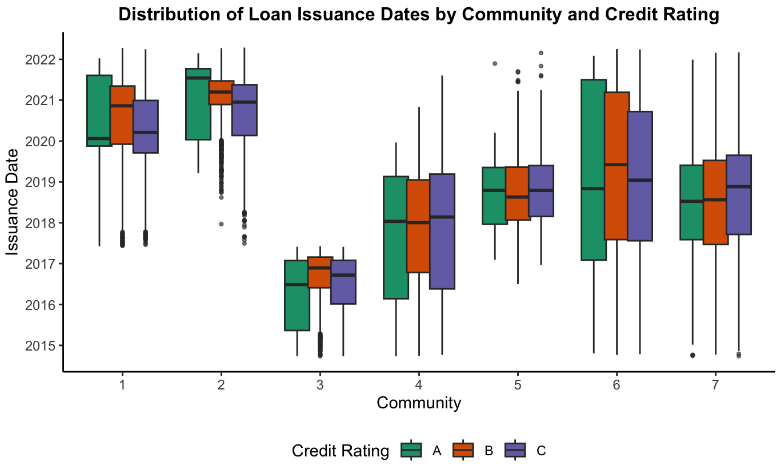


Figure A.7: **Distribution of Loan Issuance Dates by Community and Credit Rating.**
This figure visualizes the issuance date distribution of loans across different communities, stratified by their assigned credit rating. Each boxplot represents the interquartile range (IQR) within a given community, with the central line indicating the median issuance date. Outliers, represented as individual points, signify loans issued at significantly different times than the majority in their community.
The plot provides insight into whether certain credit rating bands were issued more frequently during specific timeframes, revealing potential temporal segmentation in loan pricing. Observed variations across communities may reflect shifts in platform risk assessment criteria, adjustments in pricing mechanisms, or changing macroeconomic conditions influencing borrower behavior.

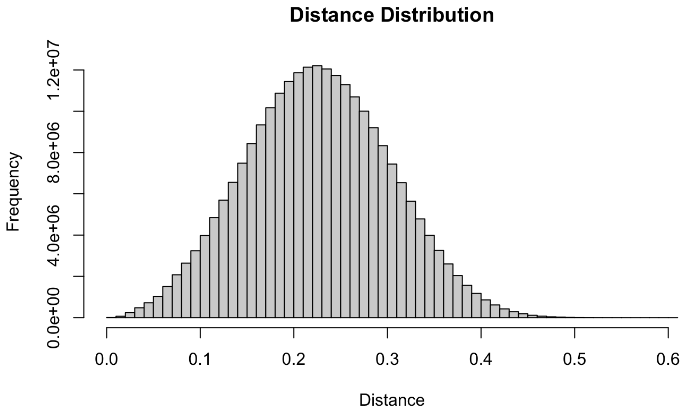


Figure A.8: Distribution of Pairwise Distances Among Loans

Appendix B. Loan Application Data Features and Descriptions

Table B.10: Loan application data features and descriptions.

Features	**Description**
date.start	Date when the loan was issued
date.end	Loan maturity date according to the latest loan schedule
default	1 - loan defaulted, 0 - otherwise
return	Nominal annual return = [(Future value of all cash-inflows + loan amount)/loan amount]$^{1/(\text{actual loan duration in days}/365)} - 1$
RR1	Modified Internal Rate of Return - 0 re-investment rate
RR2	Modified Internal Rate of Return - re-investment rate given by return on loans ending 365 days prior to the start of the loan (mean, median, weighted)
NPRP	Nominal profit in % = cash inflows / loan amount - 1
NPRA	Nominal net cash = sum of all cash inflows - loan amount
FVCI	Future value of cash inflows - loan amount, re-invested at the return (mean, median, weighted)
new	1 - it is a new customer
Age	The age of the loan applicant
Gender	1 - Woman, 0 - Male or couple of undefined
Interest	Maximum interest accepted in the loan application
MonthlyPayment	Estimated amount the borrower has to pay every month
No. Prev. Loans	Number of previous loans
Amt. Prev. Loans Bef. Loan	Value of previous loans
time	Time index in days = Current date of the loan application - Earliest date of a loan application in the dataset
time2	Square of the time index
time3	Cube of the time index
Hour	Application hour (ranging from 0 to 22)
weekday	Day of the week (1 for Friday, 2 for Monday, 3 for Saturday, 4 for Sunday, 5 for Thursday, 6 for Tuesday)
ver	Method used to verify loan application data (2 for income unverified and cross-referenced by phone, 3 for income verified, 4 for income and expenses verified)
lang	Language (1 for Estonian, 2 for English, 3 for Russian, 4 for Finnish, 6 for Spanish)
log.amount	Natural log of the loan amount
duration	Duration of the loan in months (options include 6, 9, 12, 18, 24, 36, 48, 60 months)
use	Loan use - consolidation, real estate, home improvement, business, education, travel, vehicle, other, health, not specified
educ	Loan applicant's education - basic education, vocational education, secondary education, higher education, not specified
marital	Loan applicant's marital status - married, cohabitant, single, divorced, widow
depen	Loan applicant's number of children or other dependents - 0, 1, 2, 3, 4
employ	Loan applicant's employment status - partially employed, fully employed, self-employed, entrepreneur, retiree
em.dur	Loan applicant's employment duration - more than 5 years, other, retiree, trial period, less than 1 year, less than 2 years, less than 3 years, less than 4 years, less than 5 years
exper	Loan applicant's experience - less than 2 years, less than 5 years, less than 10 years, less than 15 years, less than 25 years, more than 25 years

Table B.11: Loan application data features and descriptions continued.

Features cont.	**Description cont.**
Other	Loan applicant's occupation area
Mining	Loan applicant's occupation area
Processing	Loan applicant's occupation area
Energy	Loan applicant's occupation area
Utilities	Loan applicant's occupation area
Construction	Loan applicant's occupation area
Retail.wholesale	Loan applicant's occupation area
Transport.warehousing	Loan applicant's occupation area
Hospitality.catering	Loan applicant's occupation area
Info.telecom	Loan applicant's occupation area
Finance.insurance	Loan applicant's occupation area
Real.estate	Loan applicant's occupation area
Research	Loan applicant's occupation area
Administrative	Loan applicant's occupation area
Civil.service.military	Loan applicant's occupation area
Education	Loan applicant's occupation area
Healthcare.social.help	Loan applicant's occupation area
Art.entertainment	Loan applicant's occupation area
Agriculture.for.fish	Loan applicant's occupation area
homeless	Loan applicant's home ownership type - homeless
owner	Loan applicant's home ownership type - owner
livingw.parents	Loan applicant's home ownership type - living with parents
tenant.pfp	Loan applicant's home ownership type - tenant, pre-furnished property
council.house	Loan applicant's home ownership type - council house
joint.tenant	Loan applicant's home ownership type - tenant
joint.ownership	Loan applicant's home ownership type - joint ownership
mortgage	Loan applicant's home ownership type - mortgage
encumbrance	Loan applicant's home ownership type - owner with encumbrance
inc.princ.empl.no	1 - has income from a principal employer
inc.pension.no	1 - has income from a pension
inc.fam.all.no	1 - has income from family allowances
inc.soc.wel.no	1 - has income from social welfare
inc.leave.no	1 - has income from leave
inc.child.no	1 - has income from child support
inc.other.no	1 - has income from other sources
inc.total	Total income [log(x+1)]
no.liab	Loan applicant's number of existing liabilities (0, 1, 2, 3, 4, 5, up to 10)
liab.l	Total amount of existing liabilities [log(x+1)]
no.refin	Loan applicant's number of liabilities after refinancing (0, 1, 2, 3, 4)
inc.support	Loan applicant's income from alimony payments [log(x+1)]
FreeCash.d	1 - has free cash
FreeCash.l	Total amount of free cash [log(x+1)]
no.previous.loan	Loan applicant's number of previous loans (0, 1, 2, 3, 4, 5, 6, 7)
previous.loan.l	Total amount of loan applicant's previous loan amounts [log(x+1)]
no.previous.repay	Loan applicant's number of previous early repayments (0, more than 1)
previous.repay.l	Total amount of loan applicant's previous loan repayments [log(x+1)]
A	Bondora rating - A
AA	Bondora rating - AA
B	Bondora rating - B
C	Bondora rating - C

References

Adams, W., Einav, L., and Levin, J. (2009). Liquidity Constraints and Imperfect Information in Subprime Lending. *American Economic Review*, 99(1):49–84.

Agarwal, S., Chomsisengphet, S., Mahoney, N., and Stroebel, J. (2018). Do Banks Pass through Credit Expansions to Consumers Who want to Borrow?*. *The Quarterly Journal of Economics*, 133(1):129–190.

Ahmad, A. and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2):503–527.

Argyle, B., Nadauld, T., and Palmer, C. (2023). Real Effects of Search Frictions in Consumer Credit Markets. *The Review of Financial Studies*, 36(7):2685–2720.

Argyle, B. S., Nadauld, T. D., and Palmer, C. J. (2020). Monthly Payment Targeting and the Demand for Maturity. *The Review of Financial Studies*, 33(11):5416–5462.

Balyuk, T. (2023). FinTech Lending and Bank Credit Access for Consumers. *Management Science*, 69(1):555–575.

Balyuk, T., Berger, A. N., and Hackney, J. (2025). What Is Fueling FinTech Lending? The Role of Banking Market Structure. *The Review of Corporate Finance Studies*, 00:1–47.

Balyuk, T. and Davydenko, S. (2024). Reintermediation in FinTech: Evidence from Online Lending. *Journal of Financial and Quantitative Analysis*, 59(5):1997–2037.

Baumöhl, E. and Lyocsa, S. (2025). Alpha-threshold networks in credit risk models. *Quantitative Finance*, pages 1–23.

Berg, T., Burg, V., Gombović, A., and Puri, M. (2020). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies*, 33(7):2845–2897.

Berg, T., Fuster, A., and Puri, M. (2022). FinTech Lending. *Annual Review of Financial Economics*, 14(1):187–207.

Bhutta, N., Fuster, A., and Hizmo, A. (2020). Paying Too Much? Price Dispersion in the U.S. Mortgage Market. *Finance and Economics Discussion Series*, 2020.0(62).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Bubb, R. and Kaufman, A. (2014). Securitization and moral hazard: Evidence from credit score cutoff rules. *Journal of Monetary Economics*, 63:1–18.

Caglayan, M., Pham, T., Talavera, O., and Xiong, X. (2020). Asset mispricing in peer-to-peer loan secondary markets. *Journal of Corporate Finance*, 65:101769.

Di Maggio, M. and Yao, V. (2021). Fintech Borrowers: Lax Screening or Cream-Skimming? *The Review of Financial Studies*, 34(10):4565–4618.

Diamond, D. W. (1984). Financial Intermediation and Delegated Monitoring. *The Review of Economic Studies*, 51(3):393.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., De Castro, I., and Kammler, J. (2016). Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking & Finance*, 64:169–187.

Duffie, D. and Singleton, K. J. (1999). Modeling Term Structures of Defaultable Bonds. *Review of Financial Studies*, 12(4):687–720.

Dömötör, B., Illés, F., and Ölvedi, T. (2023). Peer-to-peer lending: Legal loan sharking or altruistic investment? Analyzing platform investments from a credit risk perspective. *Journal of International Financial Markets, Institutions and Money*, 86:101801.

Edelberg, W. (2006). Risk-based pricing of interest rates for consumer loans. *Journal of Monetary Economics*, 53(8):2283–2298.

Einav, L., Jenkins, M., and Levin, J. (2012). Contract Pricing in Consumer Credit Markets. *Econometrica*, 80(4):1387–1432.

Einav, L., Jenkins, M., and Levin, J. (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, 44(2):249–274.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Franks, J., Serrano-Velarde, N., and Sussman, O. (2021). Marketplace Lending, Information Aggregation, and Liquidity. *The Review of Financial Studies*, 34(5):2318–2361.

Freedman, S. and Jin, G. Z. (2011). Learning by Doing with Asymmetric Information: Evidence from Prosper.com. Technical Report w16855, National Bureau of Economic Research, Cambridge, MA.

Freedman, S. and Jin, G. Z. (2017). The information value of online social networks: Lessons from peer-to-peer lending. *International Journal of Industrial Organization*, 51:185–222.

Fuster, A., Plosser, M., Schnabl, P., and Vickery, J. (2019). The Role of Technology in Mortgage Lending. *The Review of Financial Studies*, 32(5):1854–1899.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4):857.

Hildebrand, T., Puri, M., and Rocholl, J. (2017). Adverse Incentives in Crowdfunding. *Management Science*, 63(3):587–608.

Holmstrom, B. and Tirole, J. (1997). Financial Intermediation, Loanable Funds, and the Real Sector. *Quarterly Journal of Economics*, 112(3):663–691.

Iyer, R., Khwaja, A. I., Luttmer, E. F., and Shue, K. (2016). Screening Peers Softly: Inferring the Quality of Small Borrowers. *Management Science*, 62(6):1554–1577.

Jagtiani, J. and Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, 48(4):1009–1029.

Johnson, M. J., Ben-David, I., Lee, J., and Yao, V. (2023). Fintech lending with lowtech pricing. *NBER Working Paper*, (31154):1–45.

Karlan, D. and Zinman, J. (2009). Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment. *Econometrica*, 77(6):1993–2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA5781.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons, New York.

Kealhofer, S. (2003). Quantifying Credit Risk I: Default Prediction. *Financial Analysts Journal*, 59(1):30–44.

Keys, B. J., Mukherjee, T., Seru, A., and Vig, V. (2010). Did Securitization Lead to Lax Screening? Evidence from Subprime Loans. *Quarterly Journal of Economics*, 125(1):307–362.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 80(5):56117–56128.

Lin, M., Prabhala, N. R., and Viswanathan, S. (2013). Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *Management Science*, 59(1):17–35.

Lin, M. and Viswanathan, S. (2016). Home Bias in Online Investments: An Empirical Study of an Online Crowdfunding Market. *Management Science*, 62(5):1393–1414.

Liu, X., Wei, Z., and Xiao, M. (2020). Platform Mispricing and Lender Learning in Peer-to-Peer Lending. *Review of Industrial Organization*, 56(2):281–314.

Liu, Y., Baals, L. J., Osterrieder, J., and Hadji-Misheva, B. (2024a). Leveraging network topology for credit risk assessment in P2P lending: A comparative study under the lens of machine learning. *Expert Systems with Applications*, 252:124100.

Liu, Y., Baals, L. J., Osterrieder, J., and Hadji-Misheva, B. (2024b). Network centrality and credit risk: A comprehensive analysis of peer-to-peer lending dynamics. *Finance Research Letters*, 63:105308.

Michels, J. (2012). Do Unverifiable Disclosures Matter? Evidence from Peer-to-Peer Lending. *The Accounting Review*, 87(4):1385–1413.

Mild, A., Waitz, M., and Wöckl, J. (2015). How low can you go? — Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research*, 68(6):1291–1305.

Phillips, R. L. (2018). *Pricing credit products*. Stanford Business Books, an imprint of Stanford University Press, Stanford, California.

Pursiainen, V. (2024). Inaccurate Borrower Information and Credit Risk: Evidence from Marketplace Loans. *The Review of Corporate Finance Studies*, 00:1–40.

Rajan, U., Seru, A., and Vig, V. (2015). The failure of models that predict failure: Distance, incentives, and defaults. *Journal of Financial Economics*, 115(2):237–260.

Stango, V. and Zinman, J. (2016). Borrowing High versus Borrowing Higher: Price Dispersion and Shopping Behavior in the U.S. Credit Card Market. *Review of Financial Studies*, 29(4):979–1006.

Stiglitz, J. E. and Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *American Economic Review*, 71(3):393–410.

Tang, H. (2019). Peer-to-Peer Lenders Versus Banks: Substitutes or Complements? *The Review of Financial Studies*, 32(5):1900–1938.

Thakor, R. T. and Merton, R. C. (2024). Trust in Lending. *The Review of Economics and Statistics*, pages 1–45.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R. (2011). Regression Shrinkage and Selection via The Lasso: A Retrospective. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3):273–282.

Vallée, B. and Zeng, Y. (2019). Marketplace Lending: A New Banking Paradigm? *The Review of Financial Studies*, 32(5):1939–1982.

Vapnik, V. (1999a). *The nature of statistical learning theory*. Springer Science & Business Media.

Vapnik, V. (1999b). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

Wei, Z. and Lin, M. (2016). Market Mechanisms in Online Peer-to-Peer Lending. *Management Science*, 63(12):4236–4257.